High-throughput comparative genomics for plant breeding and its application in the tomato clade

Saulo Alves Aflitos

Thesis committee

Promotors
Prof. Dr J.H.S.G.M. de Jong
Personal chair at the Laboratory of Genetics
Wageningen University

Prof. Dr D. de Ridder
Professor of Bioinformatics
Wageningen University

Co-promotor
Dr S.A. Peters
Senior researcher, Plant Research International (PRI)
Wageningen University and Research Centre

Other members
Prof. Dr M.A.M. Groenen, Wageningen University
Prof. Dr W.J. Stiekema, University of Amsterdam
Dr J.M. de Haas, HZPC Holland, Metslawier
Dr J.P.H. Nap, Hanze University of Applied Sciences Groningen

This research was conducted under the auspices of the Graduate School Experimental Plant Sciences (EPS)

High-throughput comparative genomics for plant breeding and its application in the tomato clade

Saulo Alves Aflitos

Thesis
submitted in fulfillment of the requirements for the degree of doctor
at Wageningen University
by the authority of the Rector Magnificus
Prof. Dr A.P.J. Mol,
in the presence of the
Thesis Committee appointed by the Academic Board
to be defended in public
on Wednesday 16 December 2015
at 4 p.m. in the Aula.

Saulo Alves Aflitos

High-throughput comparative genomics for plant breeding
and its application in the tomato clade,
198 pages.

PhD thesis, Wageningen University, Wageningen, NL (2015)
With references, with summaries in English and Dutch

ISBN: 978-94-6257-613-1

Contents

Chapter 1

General Introduction

The application of next generation sequencing and other -omics technologies to crop plants and their wild relatives have created high expectations for pinpointing and dissecting complex traits, as these methods allow for complete reading of genomes, transcriptomes, proteomes and metabolomes. The utilization of wild relatives for their invaluable wealth of genetic diversity has become even more attractive for broadening the genetic base of crops by introgressive hybridization breeding, additionally it also helps to solve fundamental questions on genome evolution, especially on the genomic mechanisms underlying domestication and long term selective breeding. In particular, introgression breeding has proven an important tool by which plant breeders transfer desirable traits from related (sub)species to their crops by interspecific hybridization and repeated backcrossings. The selection of such favorable genotypes often takes many years of consecutive crossings and selections, assuming that the traits of interest can be integrated successfully into the recipient genome. However, a detailed comprehension of such genetic processes, including the sequence-level mechanisms, is still largely lacking, impeding efficient use of introgressions in breeding practices.

To speed up selection processes of elite breeding material, powerful tools and strategies for fast and accurate genome sequence analyses in comparative genomics are now being developed. However, such approaches require production of sequence resources of entire clades, the proper assembly of reference genomes and a powerful bioinformatics toolset to analyze the vast amount of genome information in light of fundamental and applied research questions as addressed above.

In this thesis, I present a large genomic resource for tomato as well as new bioinformatics methods to mine and interpret this data. Below, I will first review the state-of-the-art in tomato breeding, followed by an introduction to genome sequencing and assembly and ending with an overview of the contributions of this thesis.

Tomato breeding

Plant Breeding

Since the start of farming, humans have practiced artificial selection. By selecting individuals from a segregating population for a superior trait, that trait is enriched. While preferred traits have been selected for in several consecutive generations, unwanted traits will gradually be removed from the population. Examples of desirable traits are high starch, protein, sugar or fibres content, biomass volume, yield, pliability, size and ease of handling, while examples of unwanted traits are poisonous substances, irritants, undesirable flavours, precocious and unsynchronized maturation and early seed release (Doebley *et al.*, 2006; Vaughan *et al.*, 2007; Gross and Olsen, 2010). However, strong selection can also bring unwanted side effects, such as loss of genetic diversity due to high inbreeding rates. This process, also known as genetic erosion (van de Wouw *et al.*, 2010;

Zamir, 2001; Hammer and Teklu, 2008), is often found in domesticated species resulting for instance in reduced (a)biotic stress tolerance. Current practices of large monocultures for most cultivated crops consequently make food production and sustainability extremely vulnerable (Jump and Penuelas, 2005; Nicotra *et al.*, 2010; Garrett *et al.*, 2006; Matesanz *et al.*, 2010).

Tomato

Tomato is an outstanding model for the Solanaceae family including several of its important crops (tomato, potato, eggplant, pepper and tobacco) and ornamental plants (*Petunia*, *Datura*). The extensive knowledge on the genome structure and genetic variation of these species and that of their wild relatives makes these crops the favourite subject of study for many geneticists, taxonomists and breeders all over the world.

Tomato (*Solanum lycopersicum*) is an economically and nutritionally important crop with a worldwide production of over 161 million tonnes (http://faostat.fao.org/). Originally from South and Central America, tomato was brought to Europe in the 14th century and since then has rapidly expanded to other regions. Nowadays it is grown on more than 4.8 million hectares worldwide. In order to meet consumers' demands and to adapt to different growing systems, selection for tomato has concentrated on specific taste, uniformity and shape, self-pruning, plant height and earliness, and (a)biotic stress tolerance (Rodríguez *et al.*, 2011; Bauchet and Causse, 2012). Unfortunately and inadvertently, these improvements came with the loss of genetic diversity and variation in the genetic materials used today; a phenomenon that is considered a classic example of 'domestication syndrome' (Hammer, 1985; Doebley *et al.*, 2006; Bai and Lindhout, 2007; Bauchet and Causse, 2012). With the great advantages of introgressive hybridization in mind, breeders are now eager to integrate other qualities, such as genes conferring resistance or tolerance to biotic and abiotic stress (to meet the challenges imposed by rapidly changing climate conditions, diseases and pests) or adding unique taste and fragrance, traits that breeders try to find in the repositories of wild relatives.

Tomato exemplifies the struggle of genetics and genomics in the understanding of plant biology, as is well documented in a review entitled "Tomato paste: a concentrated review of genetic highlights from the beginnings to the advent of molecular genetics" (Rick, 1991). The history of its scientific discovery started with a detailed cytogenetic characterization of tomato chromosomes using pachytene morphology (Ramanna and Prakken, 1967) and deletion mapping (Khush and Rick, 1968). The first linkage map based on morphological traits was derived from a segregating population of an *S. lycopersicum* x *S. pennellii* cross (Khush and Rick, 1963; Rick, 1980) and later this map was supplemented with isozyme and molecular markers (Klein-Lankhorst *et al.*, 1991; Tanksley *et al.*, 1992). In 2006, the Tomato Genome consortium was created, comprising institutions from 14 countries in Europe, Asia and Latin America (www.solgenomics.net) and several

papers were published covering various aspects of tomato biology (Arumuganathan and Earle, 1991; Sherman *et al.*, 1995; Zhong *et al.*, 1996; Xu and Earle, 1996; Peterson *et al.*, 1996; Peterson *et al.*, 1999; Areshchenkova and Ganal, 1999; Budiman *et al.*, 2000; Mueller *et al.*, 2005; Wang *et al.*, 2005; Wang *et al.*, 2006; Peters *et al.*, 2006; Kahlau *et al.*, 2006; Chang *et al.*, 2007; Barone *et al.*, 2008; Todesco *et al.*, 2008; Peters *et al.*, 2009; Peters *et al.*, 2012), a genome status update (Mueller *et al.*, 2009) and finally the reference genome of tomato (Tomato Genome Consortium, 2012).

Genetics and breeding

With the advent of molecular biology plant breeders began to develop tools to study relationships between traits of interest and underlying genetic elements (Giovannoni, 2001; Rodríguez *et al.*, 2011). Assessing the genome for such traits and their corresponding genes is expected to improve both the specificity and speed of breeding programs. Historically, some of the first molecular markers for breeding purposes were based on restriction enzyme and / or PCR based fingerprinting technologies such as Restriction Fragment Length Polymorphism (RFLP, Botstein *et al.*, 1980), Randomly Amplified Polymorphic DNA (RAPD, Williams *et al.*, 1990), and Amplified Fragment Length Polymorphism (AFLP, Vos *et al.*, 1995). In general, such methods reveal DNA polymorphisms by using the random distribution of particular restriction sites or PCR primer sequences, followed by a systematic processing of isolated DNA and profiling DNA patterns in an agarose gel. Some of the polymorphic gel bands of plants from a segregating population can be correlated with a particular trait and hence can be used to trace its presence in consecutive crossing generations, a method known as marker assisted selection (Ribaut and Hoisington, 1998; Collard and Mackill, 2007). Once a detailed genetic map with large numbers of markers has been obtained, quantitative traits can then be mapped to genetic map positions, a method that is referred to as Quantitative Trait Locus analysis (QTL, reviewed by Miles and Wayne, 2008). A major step forward in the use of molecular marker selections has been the development of nextgen sequencing technologies providing huge numbers of single nucleotide polymorphisms (SNPs) at relatively low cost. The availability of a dense set of such markers in the genetic region of interest allows the identification of the causal gene or set of genes linked to a particular trait. With this information, PCR, that directly amplify the gene of interest can now be used in breeding programs as a tool for extremely precise trait mapping.

A very recent approach to genetic mapping is Genotyping by Sequencing (GBS) (Elshire *et al.*, 2011; Deschamps *et al.*, 2012), which combines the advantages of fast, undirected PCR fingerprinting with DNA sequence information and causal polymorphism identification at only a fraction of the cost compared to other DNA marker technologies. GBS works by fragmenting genomic DNA with restriction enzymes, size filtering and sequencing the restriction sites to reveal polymorphisms. In this way, causal polymor-

phisms can be identified without the need for whole genome sequencing or even a reference genome.

Bioinformatics for plant breeding

Genome sequencing

As outlined in the previous section genome sequencing and reconstruction is becoming more and more important for developing new (sequence based) breeding approaches. At present, the prominent role of genome sequencing is reflected by NCBI statistics, reporting the number of base pairs deposited in GenBank has doubled approximately every 18 months since 1982 (Figure 1). The same trend can be seen in using Google Ngram Viewer (Figure 2), which reports the occurrence of ngrams (words, phrases) in the corpus of books over the years. Not surprisingly, the same trend is also seen in the number of patents associated with genome sequencing: 24,100 filed until the year 2000 and 348,000 new patents filed between the years 2001 and 2015 (Google patent search: www.google.com/patents, May 2015). This growth in the amount of data being generated is mainly caused by a drastic reduction in sequencing costs (Figures 3 and 4), made possible by the advent of new sequencing technologies.

In the early days of automated sequencing, the inception of a genome sequencing project was the construction of a Bacterial Artificial Chromosomes (BAC) library. Each BAC contained a genomic fragment of 150-350 Kbp which was then sequenced using Sanger/454 generating reads of 500-1000 bp. The assembly complexity was smaller since each DNA pool would amount to 150-350 Kbp instead of the whole genome being sequenced. The high quality and long read length also facilitated the assembly. The drawbacks of this approach were its high cost and laborious (often manual) methodology.
Whole Genome Shotgun sequencing and assembly, although already used for small genomes for decades (Staden, 1979), became more widely used in the mid-nineties with the sequencing of Haemophilus influenzae (Fleischmann *et al.*, 1995), and several eukaryotic model species including *Saccharomyces cerevisiae* (Goffeau *et al.*, 1996), Arabidopsis (Meinke *et al.*, 1998; Tabata *et al.*, 2000; Bevan *et al.*, 2001), *Drosophila melanogaster* (Adams *et al.*, 2000) and the human genome in 2001 (Lander *et al.*, 2001; Venter *et al.*, 2001). The human genome was initially sequenced using the BAC-by-BAC methodology in 1990 by a publically funded collaboration of twenty universities from six countries. In 1998, the Celera Corporation started its own initiative to sequence the human genome using whole genome shotgun and finished the sequencing in 2000, together with the public initiative, at 10% of the cost, largely due to the new technologies it applied. In 2000 the two initiatives were effectively merged and the genome was declared finished in 2003.
From this breakthrough, BAC-by-BAC sequencing started to fade in disuse while whole genome shotgun sequencing took over as the preferred method. This was accompa-

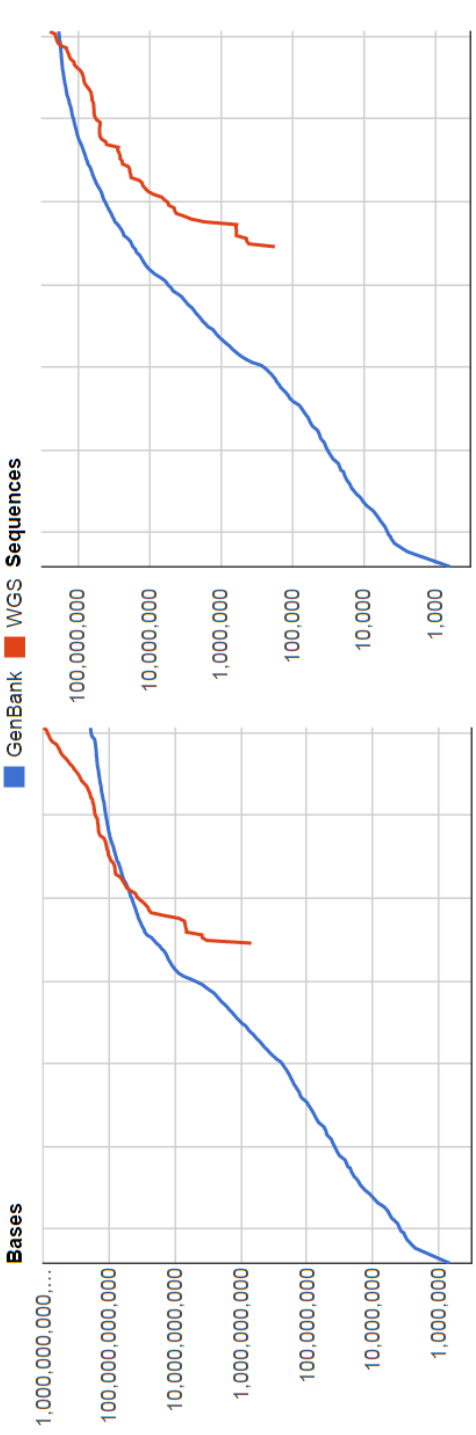


Figure 1: Growth statistics for Whole Genome Sequencing (WGS) datasets and the GenBank datasets processed by the US National Center for Biotechnology Information (NCBI), United States National Library of Medicine (NLM), a branch of the US National Institutes of Health (NIH). Available at www.ncbi.nlm.nih.gov/genbank/statistics, May 2015

nied with the replacement of Sanger sequencing technology by the first generation of Next Generation Sequencing (NGS) technology, with read lengths smaller than 50 bp, followed by the second generation NGS, with read lengths smaller than 150 bp, and now with the third generation NGS technologies with read lengths in the order of a 10-50 Kbp.

First and second generation NGS technologies have low cost as an advantage, but their short reads have to be uniquely placed in the context of a whole genome containing complex structures such as repeats, duplications and low complexity regions. Several sequencing and assembly strategies have been created such as linked reads, where a long insert, sized larger than the read length, has its insert ends sequenced so that two linked reads are obtained with a known spacing (although without knowing the actual sequence between the reads ends). This allows for less fragmented assemblies than when using non-paired or non-overlapping reads. As NGS suffers from the inability to assemble repeats larger than twice its read lengths or to link two sequences sepa-

rated by such repeats, a technique called "gap filling" has been developed, but its throughput and the ability to close large repeats is limited due to the need of several PCRs and Sanger sequencing rounds (Boetzer and Pirovano, 2012). Another approach to sequence and bridge repeats is the use of jumping libraries (mate-pair) to connect previously unlinked sequencing fragments (scaffolds) and to fill the space between the sequences with gaps of (estimated) known size. Jumping libraries are still limited in size and can only span up to 20 Kbp. Any repeat region larger than that, or containing low complexity sequences around the gap (preventing the anchoring of the jumping libraries), cannot be merged with the rest of the sequences if only NGS data are available. In practice, the best methods for genome completion still rely on BACs anchored to physical maps.

Current genome sequence practice consists of isolating highly purified DNA, fragmentation and separation of the DNA molecules, followed by PCR amplification (except for single molecule sequencing), reading the base pairs using a specialized sequencing platform and finally processing the data. An overview of the

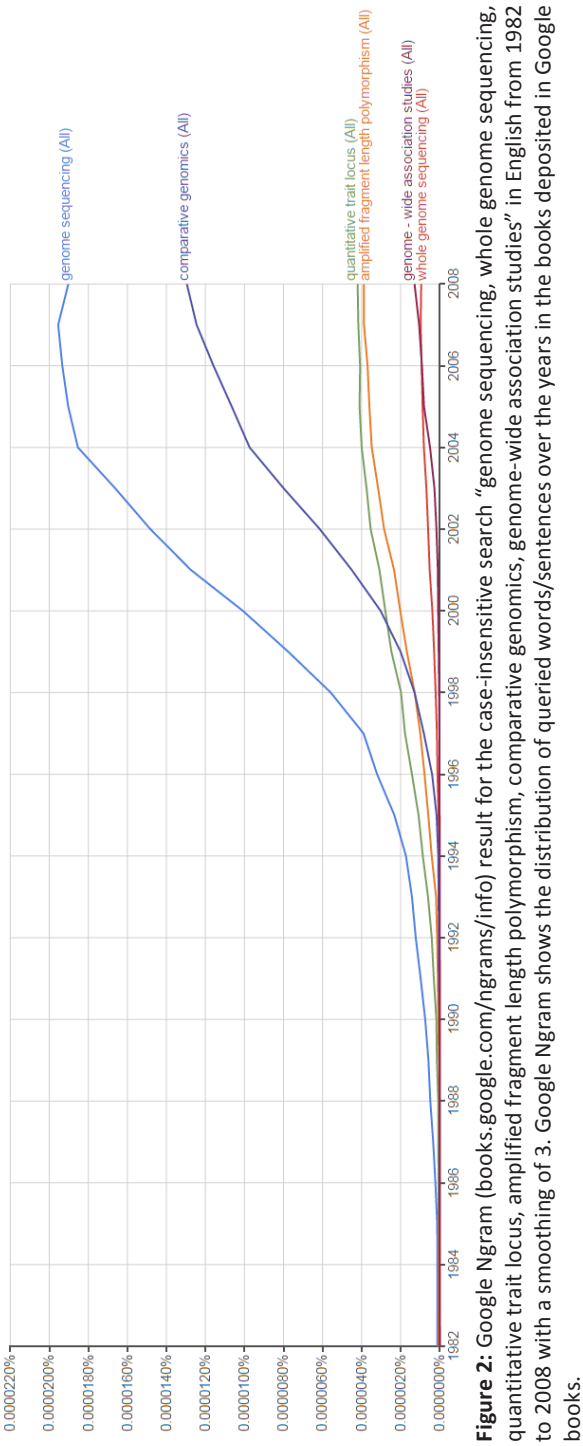


Figure 2: Google Ngram (books.google.com/ngrams/info) result for the case-insensitive search "genome sequencing, whole genome sequencing, quantitative trait locus, amplified fragment length polymorphism, comparative genomics, genome-wide association studies" in English from 1982 to 2008 with a smoothing of 3. Google Ngram shows the distribution of queried words/sentences over the years in the books deposited in Google books.

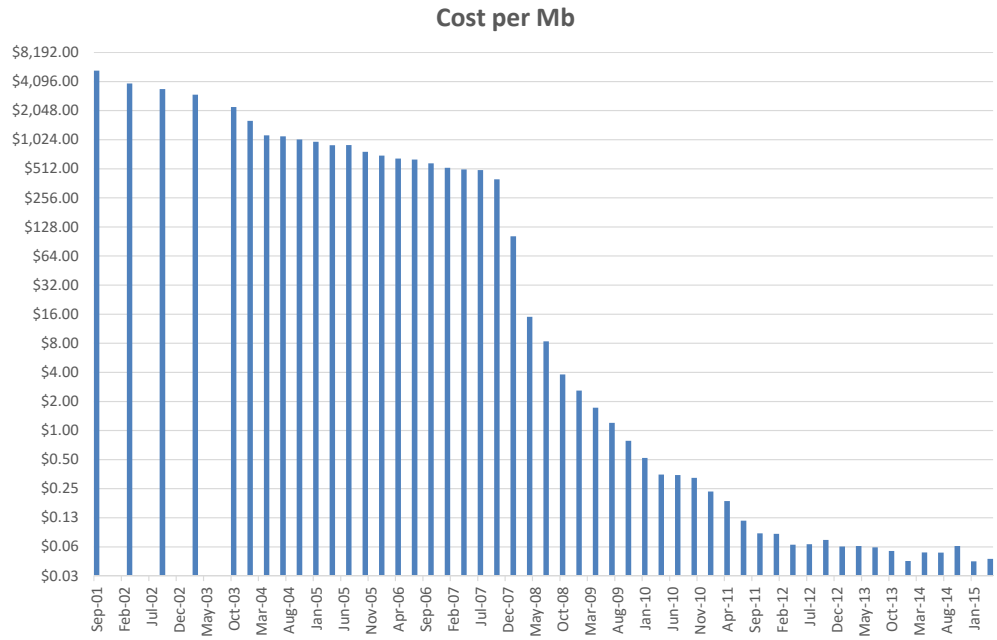


Figure 3: Estimated cost per millions of base pairs sequence. Adapted from the National Human Genome Research Institute (Wetterstrand, 2015)

most common first generation sequencing and NGS methods and their characteristics is given in Table 1 and Figure 4.

Many sample preparation methods and sequencing technologies exist and new approaches are being developed at a fast pace. Whatever technology is used for sequencing, the data has to be computationally processed. This processing generally starts with the interpretation of the signals acquired by the sequencing apparatus although, nowadays, this process is mostly done automatically.

Genome assembly

Genome assembly makes use of algorithms that find overlaps between reads of DNA fragments in order to reconstruct the original, large-scale molecules that existed before DNA fragmentation. The ultimate goal is to generate the assembled DNA of complete chromosomes (pseudo-molecules), or parts of it (scaffolds or super-contigs, contigs or consensus sequences) as long as possible. For the assembly, all reads that are found in the sample are described in terms of sequence, sequence quality, and length. Assemblers then apply algorithms that use these sequence features such that identical segments (overlaps) are identified, aligned, and reported. Methods for read overlap identification can be roughly separated in those based on comparisons of full reads or comparisons of

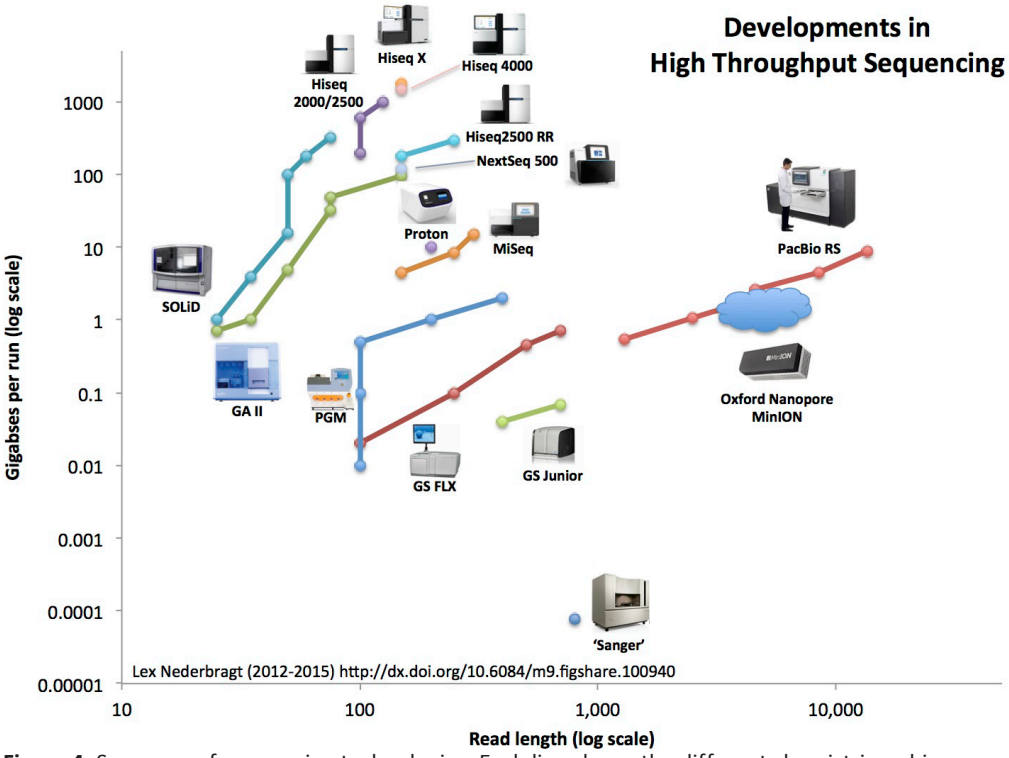


Figure 4: Summary of sequencing technologies. Each line shows the different chemistries, chips or run types (Nederbragt, 2012).

k-mers, i.e. subsequences of length k. For first generation technology such as Sanger sequencing, which delivers high quality long reads, Overlap-Layout-Consensus and string graphs (Li and Homer, 2010), were the preferred methods for full read comparison. Today, with NGS technologies that yield short reads, the most commonly used methods are *k*-mer based, employing data structures such as the de Bruijn graph (Pevzner *et al.*, 2001; Compeau *et al.*, 2011), prefix tree (Warren *et al.*, 2007) and suffix array. With the advent of the third generation sequencing methods, which generates very long reads, the Overlap-Layout-Consensus and string graph based methods are receiving renewed interest. Nevertheless, the de Bruijn graph-based methods are still by far the most successful ones, reflected by the fact of being used by six out of the nine contestants in the latest Assemblathon 2 competition of genome assemblers and pipelines (Bradnam *et al.*, 2013). A summary of these methods can be found in Table 2.

The output of an assembler is a list of consensus sequences that the algorithm was able to reconstruct. This consensus sequence is the closest to the original molecules before sequencing and, ideally, the number of consensus sequences would equal the number of chromosomes in the eukaryotic organism or the number of Bacterial Artificial Chromosome (BAC), plasmids or other vectors in prokaryotes. Although current

Technology	DNA amplification	Size Selection	Reading Method	Fragment Size	Error rate
Sanger	Y	N	F	Up to 500 bp	Very low
454	Y	N	F	Up to 800 bp	Very low
Illumina	Y	Y	P1	170 bp – 700 bp 2 Kbp – 20 Kbp	High
PacBio	N	N	P2	Up to 50 Kbp	Very high
Nanopore	N	N	H	Up to 1 Mbp	Extremely high

Table 1: An overview of the most common sequencing methods. Reading methods are classified as: F = Modified nucleotides containing a fluorophore; P1 = Pyrosequencing with alternated cycles of DNA polymerase and washing; P2 = Single molecule pyrosequencing; H = DNA Nano channels measuring pH of each base pair; Error rates are classified as very low (less than 1 per million), high (less than 1 per thousand), very high (approximately 1 per 10) and extremely high (less than 1 per 10)

short-read technology is far from delivering complete end-to-end chromosome length sequences, we are able to assemble most of the single copy regions of genomes cheaply and fast. Nevertheless, plant genomes can be very challenging to assemble due to their large size and complexity, which usually is proportional to their repeat content, ploidy level and heterozygosity. Repeats often are the main cause of our inability to assemble complex regions in the genome, leading to fragmented and incomplete assemblies (Fransz *et al.*, 2000; Hoskins *et al.*, 2007; Szinay *et al.*, 2008; Chang *et al.*, 2008). This is due to the limited length of short reads that do not span an entire repeat region. Such reads lack the unique flanking sequences, needed to anchor them unambiguously on the genome. In de assembly these repetitive reads will be stacked, leaving gaps in the assembly.

To resolve complex genome sequences that cannot be reconstructed with short reads, long read technology is applied. Nevertheless, long reads suffer from low base quality. To improve the quality, short reads are mapped to long reads for error correction. The error corrected long reads are subsequently used for a *de novo* assembly. Although the use of long read sequencing technology like PacBio has proven advantageous for genome reconstruction, overcoming many of the assembly problems that cannot be solved with short reads, many reference genomes assembled with long reads are still incomplete consisting of large sequence contigs that lack positional order and orientation. Therefore, yet other technologies to properly order and orient contigs along the genome are required.

Fluorescent *In situ* Hybridization and Genetic Maps

One of the methods to position large sequencing segments on the chromosomes is Fluorescence *In situ* Hybridization (FISH). It consists of hybridization of probe DNA obtained from labelling single copy or repetitive DNA sequences in cell spreads containing nuclei and chromosomes of the species under study. The DNA label is either a fluorescence dye,

	Read length	Number of reads	Technology
Overlap-Layout-Consensus	Long	Small	Sanger, PacBio
String graph	Long	Small	Sanger, PacBio
Prefix tree	Short/Long	Large/Very large	Illumina, PacBio
de Bruijn graph	Short/Long	Large/Very large	Illumina, PacBio

Table 2: Summary of assembly methods. Read lengths are classified as "short" (less than 500 base pairs) or "long" (larger than 500 base pairs); Number of reads is classified as "small", "large" or "very large" (thousands, millions and billions of base pairs, respectively).

like FITC and Cy3, or a reporter molecule like digoxigenin or biotin that can be detected by anti-dig and streptavidin detection systems. The fluorescent foci thus obtained on the chromosomal targets allow to determine 1) the length of the fragment; 2) the chromosome identity; 3) the chromosomal position and 4) and its relative position to other fragments. The major advantage of FISH is that assembled DNA sequences can be mapped on the chromosomes irrespective of any problems posed by repeats. Many FISH studies use probe DNA, like that of BACs, which contains markers that have been anchored to the genetic map. The great advantage is that genetic and physical map position and distances can be linked. In other words: by searching for the sequence of specific genes and markers in the sequenced genome, their relative genetic position, in centimorgans, can be used to a certain extent to order the fragments of the assembly. If a correlation between centimorgans and base pairs exists, gap sizes can be estimated accordingly (Chang *et al.*, 2007; Szinay *et al.*, 2008).

Comparative Genomics

Once plant genomes have been elucidated, the next step for plant breeders is comparative genomics. Until today the majority of genome-wide studies on collinearity of tomato and wild relatives have been studied mostly with molecular genetic maps. However, the limitations in comparative genetic study on the basis of linkage map comparisons are (1) the need for mapping populations; (2) insufficient DNA polymorphisms for simple markers that are locus specific across species; (3) deviation between the genetic and physical chromosome maps and (4) the large pericentromere regions are genetically "blind" due to lack of crossover recombination. These problems were significantly recognized within the framework of tomato genome sequencing project (Szinay *et al.*, 2008; Peters *et al.*, 2009). To avoid these limitations, comparative genetic studies are becoming more and more genome sequence based, which is reviewed and well reflected in the definition of comparative genomics provided by Touchman (2010):

> *"Comparative genomics is a field of biological research in which the genome sequences of different species - human, mouse, and a wide variety of other organisms from bacteria to chimpanzees - are compared. By comparing the sequences of genomes of different organisms, researchers can understand what, at the molecular level, distinguishes different life forms from each other. Comparative genomics also provides a powerful tool for studying evolutionary changes among organisms, helping to identify genes that are conserved or common among species, as well as genes that give each organism its unique characteristics"*

Through comparative genomics, several genomic features such as genes (e.g.: presence/absence, sequence or order), regulatory elements, repetitive sequences, and number and structure of chromosomes can be studied, allowing to trace the functional relationship between DNA and corresponding traits. Causal polymorphisms in genes from different species or genera can now be identified, and tracked in introgressed material in a much faster and more targeted manner than ever before. Such introgressions can be immortalized (made homozygous) by selfing each plant for 8-12 generations, creating lines with donor fragments in the recipient plant, while at the same time the excess of alien genomic DNA is eliminated. If the original parents are closely related and crossable (like accessions or genotypes), and the chromosomes involved in meiotic recombination are full homologues, they give rise to Recombinant Inbred Lines or RILs. If the parents are from different (sub-)species, and the meiotic pairing partners are homeologues, the resulting immortalized plants are called introgression lines or ILs. The use of RILs and ILs is very powerful for establishing QTLs. By analyzing a large population of RILs or ILs the trait of interest can be mapped on the linkage maps of all chromosomes and can be correlated with the physical position on the chromosomes. In this way it is feasible to narrow down the chromosomal segment(s) causal to the QTL phenotype (Eshed and Zamir, 1995; Schauer *et al.*, 2006; Fridman *et al.*, 2004; Causse, *et al.*, 2004; Zygier *et al.*, 2005; Lippman *et al.*, 2007; Saliba-Colombani *et al.*, 2001; Causse *et al.*, 2001). Nonetheless, these QTLs domains often harbor hundreds of genes requiring additional analysis to link a particular trait to (an) underlying gene(s).

Another method to link a genomic region to a trait of interest is the analysis of natural populations by Genome Wide Association Studies (GWAS; reviewed by Bush and Moore, 2012), which applies mass sequencing and phenotyping of a set of wild relatives of the species. From this dataset, DNA polymorphisms can be correlated to a particular phenotype. Although GWAS is less powerful than QTL analysis in RIL populations, it has the advantage of correlating specific SNPs or haplotypes very accurately to the quantitative trait (Keurentjes *et al.*, 2011; Korte and Farlow, 2013).

At the chromosome level, comparative genomics can provide insight in synteny (here defined as a set of genes found in the same order in different species), which is of practical importance to plant breeding. A break in collinearity between genes by

hom(e)ologous recombination during meiosis may lead to linkage drag, infertile gametes and even embryo lethality. Linkage drag is a huge hurdle for breeders as it can signify the impossibility of separating a gene of interest from other undesirable, linked genes. The phenomenon is in general caused by a chromosome rearrangement (mostly an inversion) containing the gene of interest and adjacent "wild" genes. Comparative genomics can then be of assistance by identifying the closest relative that possesses the polymorphism of interest but without the chromosome rearrangement.

Contributions of this thesis

In this thesis I present tools and genomic datasets with the aim to better understand the phylogenetic relationships, genomic diversity and chromosomal variation in the tomato clade and to support future sequencing projects. This thesis consists of a general introduction (this chapter), three article chapters, one manuscript chapter and a general discussion.

Chapter 2 of this thesis describes the whole genome shallow sequencing by NGS of 84 accessions of tomato from 13 different Solanum species and the deep sequencing, coupled with *de novo* assembly, of three wild tomato species (*Solanum arcanum* LA2157, *S. habrochaites* LYC4 and *S. pennellii* LA0716; Aflitos *et al.*, 2014). With this project we aim to get further insight in the diversity and phylogenetic relationships of tomato and related wild species.

Chapter 3 of this thesis describes the development of a software package (Introgression Browser; Aflitos *et al.*, 2015) for high-throughput visualization of sequence divergence between species when compared to a common reference, as well as the visualization of introgression borders and crossing barriers. Such a tool can be of great value for the identification of introgressions, sequence differences between varieties and diversity resources in wild species. For this goal, we have shallow sequenced the donor species *S. pimpinellifolium* CGN14498 and an F8 RIL population of 60 individuals with *S. lycopersicum* cv. AllRound LA2463 (previously sequence and described in chapter 2) as a receptor species.

Chapter 4 describes a new tool for *k*-mer based species identification, named Cnidaria, without the need for time and resource expensive genome assembly or reference mapping. This tool is useful to sample identification in forensics, medicine and research. Also, it can be used as quality control by sequencing facilities and to validate the data received from sequencing providers.

Chapter 5 describes our efforts to identify unique regions in the tomato genome for the development of primers, which can generate unique amplicons for PCR-FISH experiments. Although BAC-FISH is an extremely powerful method, BAC libraries are costly and laborious to create, therefore becoming less popular for sequencing projects. With PCR-FISH we are able to extend the use of FISH for genome closure without the need

for BAC, designing probes directly from the NGS data and amplifying it directly from genomic DNA. PCR-FISH can be a valuable tool to finalize genome assemblies where other techniques failed, allowing for the assembly of pseudo-molecules (chromosomes).

Chapter 6 gives a general discussion of the bioinformatics tools that were developed during my PhD study which have a potential use in many areas of basic and applied research for plants and animals. The datasets created in this thesis can serve as the base for new research projects and help to obtain both better understanding and improvements of the tomato crop. In this respect I also work out some ideas regarding the significance of this work for plant breeders and present an outlook for genomics and bioinformatics.

Chapter 2

Exploring genetic variation in the tomato (*Solanum* section *Lycopersicon*) clade by whole-genome sequencing

Aflitos, Saulo, Elio Schijlen, Hans Jong, Dick de Ridder, Sandra Smit, Richard Finkers, Jun Wang *et al*., "Exploring genetic variation in the tomato (*Solanum* section *Lycopersicon*) clade by whole-genome sequencing." *The Plant Journal* 80, no. 1 (2014): 136-148. DOI: 10.1111/tpj.12616

Summary

Genetic variation in the tomato clade was explored by sequencing a selection of 84 tomato accessions and related wild species representative for the Lycopersicon, Arcanum, Eriopersicon, and Neolycopersicon groups. We present a reconstruction of three new reference genomes in support of our comparative genome analyses. Sequence diversity in commercial breeding lines appears extremely low, indicating the dramatic genetic erosion of crop tomatoes. This is reflected by the SNP count in wild species which can exceed 10 million i.e. 20 fold higher than in crop accessions. Comparative sequence alignment reveals group, species, and accession specific polymorphisms, which explain characteristic fruit traits and growth habits in tomato accessions. Using gene models from the annotated Heinz reference genome, we observe a bias in dN/dS ratio in fruit and growth diversification genes compared to a random set of genes, which probably is the result of a positive selection. We detected highly divergent segments in wild *S. lycopersicum* species, and footprints of introgressions in crop accessions originating from a common donor accession. Phylogenetic relationships of fruit diversification and growth specific genes from crop accessions show incomplete resolution and are dependent on the introgression donor. In contrast, whole genome SNP information has sufficient power to resolve the phylogenetic placement of each accession in the four main groups in the Lycopersicon clade using Maximum Likelihood analyses. Phylogenetic relationships appear correlated with habitat and mating type and point to the occurrence of geographical races within these groups and thus are of practical importance for introgressive hybridization breeding. Our study illustrates the need for multiple reference genomes in support of tomato comparative genomics and Solanum genome evolution studies.

Introduction

The Solanaceae or Nightshade family consists of more than 3000 species covering a very large diversity in terms of habit, habitat and morphology. Its species occur worldwide growing as large forest trees in wet rain forests to annual herbs in deserts (Knapp, 2002). Solanum is the largest genus in the family and includes tomato (*Solanum lycopersicum*) and various other species of economic importance. Tomato breeding during the past decades focused on higher productivity and adaption to different growing systems. Its economic success is reflected by the fact that, on a global scale, tomato is one of the most important vegetable crops with a worldwide production of 159 million tons covering some 4,700,000 ha (www.fao.org). Yet, domestication for tomato has been clearly distinct from the species divergence by natural selection as a consequence of selecting for a limited set of traits, including attractive red fruit color and size. As a result its genetic basis was seriously narrowed, known as the 'domestication syndrome' (Hammer, 1985; Doebley *et al.*, 2006; Bai and Lindhout, 2007; Bauchet *et al.*, 2012). In more recent times tomato was adapted to different growing systems involving a small number of traits, including self-pruning, plant height, earliness, fruit morphology and non-red fruit color (Bauchet *et al.*, 2012). The relative small genetic variation became apparent in the

face of rapidly changing environmental conditions, competing claims for arable lands and new consumer requests. These challenges push tomato breeding efforts towards better biotic and abiotic stress tolerance, higher productivity and increased sensory and nutritional value. However, the reduced genetic variation that resulted from extensive inbreeding has decelerated tomato crop improvement. To enlarge the genetic basis, breeders now focus on introgression of desirable genes from the wild relatives into their elite cultivars, which so far has been quite limited (Singh, 2007; Bai and Lindhout, 2007).

The first step of introgressive hybridization includes crosses of the cultivated tomato with its heirlooms, wild relatives or more distant species of the tomato clade. Introgression breeding is practicable as cultivated tomato and related wild species are intracrossable, and most of the wild species are also intercrossable (Rick, 1979; 1986; Spooner *et al.*, 2005) despite the fact that diverse mating systems have evolved varying from allogamous self-incompatible (SI) and facultative allogamous to autogamous self-compatible (SC). Especially at the geographic margins of the distributions, interspecies changes in incompatibility systems that promote inbreeding over outcrossing have been documented (Peralta *et al.*, 2008; Grandillo *et al.*, 2011). Species boundaries and genetic diversity have been extensively studied in tomato using a wide range of molecular data (reviewed in Peralta *et al.*, 2008 and Grandillo *et al.*, 2011). For example, RFLP analysis showed that genetic diversity for SI species far exceeds that of SC species, estimated at 75% vs. 7% (Miller and Tanksley, 1990). Furthermore, 'within-accession' genetic variation was estimated at 10% of the 'between-accession' variation, in contrast to the genetic variation of the modern cultivars estimated at less than 5%. This further illustrates the dramatic erosion of genetic diversity in cultivated tomato crops.

Selection of crossing parents for interspecific hybridization requires insight in phylogenetic relationships for the tomato clade, but their tree based on morphological and molecular data has not been undisputed. Four informal species groups were proposed for the tomato clade, Lycopersicon, Arcanum, Eriopersicon and Neolycopersicon (Peralta *et al.*, 2008), which are supposed to have evolved from a most recent common ancestor approximately 7 million years ago (Nesbitt and Tanksley, 2002; Spooner *et al.*, 2005; Moyle, 2008, Peralta *et al.*, 2008). In spite of these studies, evolutionary relationships between the 13 species in the Lycopersicon clade are not fully resolved, considering the dichotomy between *Solanum pennellii* and *Solanum habrochaites* (Spooner *et al.*, 2005; Peralta *et al.*, 2005; 2008). The evolutionary history of Solanum genomes has also been investigated from the perspective of chromosome organization. The study of Szinay *et al.*, (2012) on cross-species BAC FISH painting of Solanum species revealed few large rearrangements in the short arm euchromatin of chromosomes 6, 7 and 12, whereas Anderson *et al.*, (2010) demonstrated pairing loops, multivalents and kinetochore shifts in synaptonemal complex spreads of hybrids between different members of the tomato clade, hinting at paracentric and pericentric inversions and translocations between the homeologous chromosomes. Furthermore, comparative genomics point to a Solanum ge-

nome landscape in which chromosome evolution for the majority of the 12 chromosomes has been far more dynamic than currently appreciated (Peters *et al.*, 2012). Collectively, these findings demonstrate that evolutionary relationships among the wild relatives still have to be considered provisional (Peralta *et al.*, 2008).

The availability of high-throughput sequencing technologies has provided unprecedented power to determine genome variation across entire clades, both at the structural and the genotype level. Initiatives such as the 1001 genomes project for *Arabidopsis thaliana*, the Drosophila sequence project, and the 1000 genomes project for human have been illustrative for the discovery of a vast amount of intraspecies specific polymorphic sequence features like InDels, repeats and SNPs for hundreds of genes (Weigel and Mott, 2009; Mackay *et al.*, 2012; The 1000 genomes project consortium, 2010), and have illustrated that there is no such thing as "the genome" for a particular species. Rather, the range of physiological and developmental traits appears to be reflected in the tremendous amount of sequence variants contributing to intraspecific variation. Considering the overwhelming interspecies genetic variability, tomato germplasm collections represent a gene pool with unprecedented possibilities to address new breeding demands imposed by climate change, world population increase, and consumer needs. Here we aim to reveal and study this genetic variation by genome-sequencing a selection of representative tomato accessions, which has become attainable with the recent development of the *S. lycopersicum* Heinz 1706 reference genome (The Tomato Genome Consortium, 2012). In addition to this reference genome for the Lycopersicon species, we present the construction of reference genomes of three other related species representing the Arcanum, Eriopersicon and Neolycopersicon group, respectively, providing an expanded resource for detailed comparative genomic studies in the near future. We also present results on robust/high confidence detection and identification of sequence polymorphisms, heterozygosity levels, introgressions, and assess the genetic diversity within the tomato clade from a phylogenetic and evolutionary perspective. This study provides an invaluable dataset for advanced omics studies on sequence trait relationships, the molecular mechanisms of tomato genome evolution as well as developing genotyping-by-sequencing breeding approaches.

Results

Selection of tomato accessions

We have selected 84 accessions of the Solanum clade section Lycopersicum for shallow (36 fold coverage) whole genome sequencing. A first set of 54 accessions consists of tomato landraces and heirloom cultivars of *S. lycopersicum* and *S. lycopersicum* var. cerasiforme which have been selected from the EU-SOL tomato core collection (https://www.eu-sol.wur.nl). The second set of 30 accessions comprises wild relatives of to-

mato representing the full range of expected genetic variation around *S. lycopersicum*. Their selection was based on previous usage in genetic research and previous utilization of quality or (a)biotic stress traits (reviewed in Grandillo *et al.*, 2011). We also chose *S. arcanum* LA2157, *S. habrochaites* LYC4 and *S. pennellii* LA0716 for *de novo* sequencing and whole genome reconstruction, aiming to have a reference genome available for each of the four main phylogenetic groups in the tomato clade. An important selection criterion was the self-compatibility of these accessions allowing inbreeding for several generations to minimize heterozygosity, and so reduce *de novo* genome assembly problems. A complete list of the selected accessions used in this study can be found in table 1.

De novo assembly of three wild tomato relatives and Heinz

Comparisons of molecular data have indicated relatively low DNA sequence diversity between genetically related species within the phylogenetic groups of the tomato clade (Miller and Tanksley, 1990). Furthermore, preliminary analysis indicated that SNP frequencies for *S. pimpinellifolium* and *S. pennellii* compared to *S. lycopersicum* were 1% and 10% respectively. Considering that *S. pimpinellifolium* and *S. pennellii* phylogenetically are among the closest and most distantly related species to tomato respectively, we assumed the same range for the other species. Our strategy to determine the proportion of polymorphic loci across the entire Lycopersicon clade was therefore targeted at *de novo* sequencing and assembly of three new reference genomes, followed by shallow sequencing of the bulk of the accessions and subsequently mapping them to a reference genome. For reference genome reconstruction we relied on massive parallel sequencing using Illumina HiSeq 2000 and 454FLX technology. Two paired-end libraries with insert sizes of 170 bp and 500 bp and one mate pair library of 2 kbp were sequenced using Illumina at 25, 25, and 30 fold coverage (assuming a genome size of 950 Mbp) respectively, and had at least 80% of the bases with Q-value above 30 (error rate <= 1/1,000). For the 454FLX sequencing, large insert size libraries of 8 kbp and 20kbp were created each at 0.6 coverage. *S. pennellii* LA0716 had an additional 8 kbp Illumina mate pair library at 0.4 fold. We discarded unpaired reads resulting in 205 fold coverage. For *de novo* assembly we aimed at maximizing short-range contiguity, long-range connectivity, completeness and quality by following the strategy as outlined by Gnerre *et al.*, (2010). Our assembly statistics show a total contig length for *S. arcanum*, *S. habrochaites* and *S. pennellii* reaching a plateau of approximately 760 Mb (figure 1, table 2). The unique portion is comparable in size in these genomes, which is consistent with widespread research including comparative mapping studies revealing a high level of synteny among the species of the Solanaceae (Paterson *et al.*, 2000). However, previous estimates on DNA content and flow cytometry analyses suggest a considerable variation in total genome size among species in the tomato clade (reviewed by Grandillo *et al.*, 2011). For example, the DNA content of cultivated tomato varies from 1.87 to 2.07 pg/2C indicating to a genome size of approximately 950 Mbp, whereas that of *S. pennellii* is substantially larger and corresponds to a DNA

content of 2.47 to 2.77 pg/2C corresponding to 1,200 Mbp. Furthermore, we assume that most of the estimated 35,000 genes reside on the ~220-250 Mb of DNA in the euchromatic regions (http://www.rbgkew.org.uk/cval/; Arumuganathan and Earle, 1991; Van der Hoeven *et al.*, 2002; The Tomato Genome Sequencing Consortium, 2012). The increased genome size of *S. pennellii* is likely to a greater part explained by an expansion of the repetitive portion of the genome. Repeats are known to impede genome reconstruction resulting in a more fragmented assembly and a lower N50 contig size. This is consistent with the *S. pennellii* LA0716 assembly statistics (table 2). The re-assembled *de novo* *S. lycopersicum* cv Heinz reaches the assembly size plateau more slowly and also appears more fragmented than the published reference genome, which likely is due to the use of older sequencing platforms and the BAC-by-BAC sequencing strategy that was used for this species previously.

To determine the extent of sequence diversity, read pairs from *de novo* sequenced genomes were mapped to the *S. lycopersicum* cv Heinz 1706 v2.40 reference genome. The lowest number of unmapped reads (11%), which likely consists of low quality sequences and increases for *S. arcanum* (17%), *S. pennellii* (22%) and *S. habrochaites* (25%), respectively. Given the comparable sequence quality we assume an equal percentage of low quality reads for the *de novo* sequenced genomes, while sequence diversity, introgressions and genome expansion contribute to the remainder of the unmapped reads.

Sequencing and mapping of the 84 accessions and wild species

For the 84 accessions 2.9x1012 base pairs were sequenced equaling to an average coverage of 36.7±2.3 (32.5±2.1 with Q >= 30) fold per accession. All individuals were mapped against *S. lycopersicum* cv Heinz 1706 v2.40 to assess the diversity in both crop and wild-species, resulting in 96.4%±0.88% and 52.9%±2.93% of the reads correctly mapping for crops and wild species, respectively (figure 2). These numbers improved when reads from wild species were mapped against a new reference genome from a closer relative. For *S. arcanum*, *S. habrochaites*, and *S. pennellii* 72.87%±7.87%, 78.74%±15,63%, and 55.37%±9.29% of the reads correctly mapped against the *S. arcanum* LA2157, *S. habrochaites* LYC4, and *S. pennellii* LA0716 reference genome respectively. These results illustrate the large genetic erosion within the crop tomatoes and the large sequence diversity among the wild species. Moreover, it emphasizes the need for multiple reference genomes to support interpretations of genetic variation consequences among species in the tomato clade, which would otherwise be biased toward a single reference genome that is genetically more distantly related to the wild species.

Whole genome sequence diversity

To further assess the sequence diversity in Solanum section Lycopersicon we quantified and classified the SNPs for each of the 84 accessions using read mappings against Heinz. The SNP counts for tomato cultivars are relatively low and gradually increase for

Table 1: List of selected accessions with names and culture collection IDs.

Short Name	Ref	Accession name	Accession (LA/LYC/EA/PI/T/CGN/TR/V)
S.lyc LA4345	REF	Heinz 1706	LA4345
S.lyc LA2706	RF_001	Moneymaker	LA2706/EA00840/EA02936/EA05097/EA10006/PI262996
S.lyc LA2838A	RF_002	Alisa Craig	LA2838A/EA01101/EA00240/EA01101
S.lyc PI406760	RF_003	Gardeners delight	EA06086/PI406760
S.lyc LA1090	RF_004	Rutgers	LA1090/EA00465
S.lyc EA00325	RF_005	Galina	EA00325
S.lyc EA00488	RF_006	Taxi	EA00488
S.lyc EA00375	RF_007	Katinka Cherry	EA00375
S.lyc EA00371	RF_008	John's big orange	EA00371
S.lyc LA2463	RF_011	Allround	LA2463/LYC1365/EA02617
S.lyc LYC1969	RF_012	Sonata	LYC1969/EA02724
S.lyc LYC3897	RF_013	Cross Country	LYC3897/EA03701
S.lyc LYC3476	RF_014	Iidi	LYC3476/EA03362
S.lyc TR00003	RF_015	Momatero	TR00003
S.lyc LYC11	RF_016	Rote Beere	LYC11/EA01965/CGN15464
S.lyc LYC3340	RF_017		LYC3340/EA03306/T1039
S.lyc EA01155	RF_018	Dana	EA01155
S.lyc EA01049	RF_019	Large Pink	EA01049
S.lyc LYC3153	RF_020		LYC3153/EA03221
S.lyc EA03222	RF_021		LYC3155/LYC2513/EA03222/T828
S.lyc PI129097	RF_022		PI129097/EA04710
S.lyc PI272654	RF_023		PI272654/EA05170
S.lyc EA00990	RF_024	Jersey devil	EA00990
S.cor LA0118	RF_025	S. corneliomulleri	LA0118/EA03384/T1248
S.lyc EA00157	RF_026	Polish Joe	EA00157
S.lyc EA02054	RF_027	Cal J Tm VF	EA02054/CGN20815
S.lyc PI303721	RF_028	The Dutchman	EA05581/PI303721
S.lyc LA4451	RF_029	Black Cherry	LA4451(?)/EA00027
S.lyc V710029	RF_030	Anto	EA01835/V710029
S.lyc PC11029	RF_031	Winter Tipe	PC11029
S.lyc PI093302	RF_032	Chang Li	EA04243/PI93302
S.lyc EA00892	RF_033	Belmonte	EA00892/SG16
S.lyc EA01088	RF_034	Tiffen Mennonite	EA01088
S.lyc PI203232	RF_035	Wheatleys Frost Resistant	EA04939/PI203232
S.lyc PI311117	RF_036	S. lycopersicum	EA05701/PI311117
S.lyc LA1324	RF_037	S. lycopersicum	LA1324/EA05891/PI365925
S.lyc PI158760	RF_038	Chih Mu Tao Se	EA04828/PI158760
S.lyc LA0113	RF_039	S. lycopersicum	LA0113/EA00526
S.lyc LYC1410	RF_040	ES 58 Heinz	LYC1410/EA02655
S.lyc PI169588	RF_041	S. lycopersicum Dolmalik	EA04861/PI169588
S.lyc LYC2962	RF_042	S. lycopersicum	LYC2962/EA03107/T556
S.lyc LYC2910	RF_043	S. lycopersicum	LYC2910/EA03058/T115
S.pim LYC2798	RF_044	S. pimpinellifolium	LYC2798/EA02994
S.lyc LYC2740	RF_045	S. lycopersicum	LYC2740/EA02960
S.pim LA1584	RF_046	S. pimpinellifolium	LA1584/EA00676/PI407541
S.pim LA1578	RF_047	S. pimpinellifolium	LA1578/EA00674
S.per LA1278	RF_049	S. peruvianum	LA1278/PI365941/TR00005
S.chm LA2663	RF_051	S. chmielewskii	LA2663/TR00007
S.chm LA2695	RF_052	S. chmielewskii	LA2695/EA00759
S.che LA0483	RF_053	S. cheesmaniae-f-minor / S. galapagense	LA0483/EA00581
S.lyc CGN15820	RF_054	S. lycopersicum x S. cheesmaniae	CGN15820/TR00024
S.che LA1401	RF_055	S. cheesmaniae-f-minor / S. galapagense	LA1401/EA00652/PI 365897
S.neo LA2133	RF_056	S. neorickii	LA2133/EA00729
S.neo LA0735	RF_057	S. neorickii	LA0735/CGN24193/TR00025
S.arc LA2157	RF_058	S. arcanum	LA2157/TR00008
S.arc LA2172	RF_059	S. arcanum	LA2172/TR00009
S.per LA1954	RF_060	S. peruvianum	LA1954/EA00713
S.hua LA1983	RF_062	S. huaylasense	LA1983/TR00010
S.hua LA1365	RF_063	S. huaylasense	LA1365/PI 365953/TR00011
S.chi CGN15532	RF_064	S. chilense	CGN15532/TR00012
S.chi CGN15530	RF_065	S. chilense	CGN15530/TR00013
S.hab CGN15791	RF_066	S. habrochaites F glabratum	PI 127827 (?)/CGN15791/TR00014
S.hab PI134418	RF_067	S. habrochaites F glabratum	PI134418/TR00015
S.hab CGN15792	RF_068	S. habrochaites F glabratum	CGN15792/TR00016
S.hab LA1718	RF_069	S. habrochaites F glabratum	LA1718/EA00699/PI 390663
S.hab LA1777	RF_070	S. habrochaites	LA1777/EA00703
S.hab LA0407	RF_071	S. habrochaites	LA0407/EA0558/PI 251304
S.hab LYC4	RF_072	S. habrochaites	LYC4/TR00017
S.spp LA1272	RF_073	S. sp	LA1272/LYC1831/PI 365970/EA02701
S.pen LA0716	RF_074	S. pennellii	LA0716/PI 246502/EA00585
S.hua LA1364	RF_075	S. huaylasense	LA1364/TR00030
S.lyc TR00018	RF_077	Large Red Cherry	TR00018
S.lyc EA00940	RF_078	Porter	EA00940
S.lyc TR00019	RF_088	Bloody Butcher	TR00019
S.lyc EA01019	RF_089	Brandywine	EA01019
S.lyc TR00020	RF_090	Dixie Golden Giant	TR00020
S.lyc EA01037	RF_091	Giant Belgium	EA01037
S.lyc TR00021	RF_093	Kentucky Beefsteak	TR00021
S.lyc TR00022	RF_094	Marmande	TR00022/PI647486
S.lyc TR00023	RF_096	Thessaloniki	TR00023
S.lyc EA01640	RF_097	Watermelon beefsteak	EA01640
S.lyc LA4133	RF_102	S. lycopersicum	LA4133/TR00026
S.lyc LA1421	RF_103	S. lycopersicum	LA1421/TR00027
S.gal LA1044	RF_104	S. galapagense	LA1044/TR00029
S.lyc LA1479	RF_105	S. lycopersicum	LA1479/TR00028

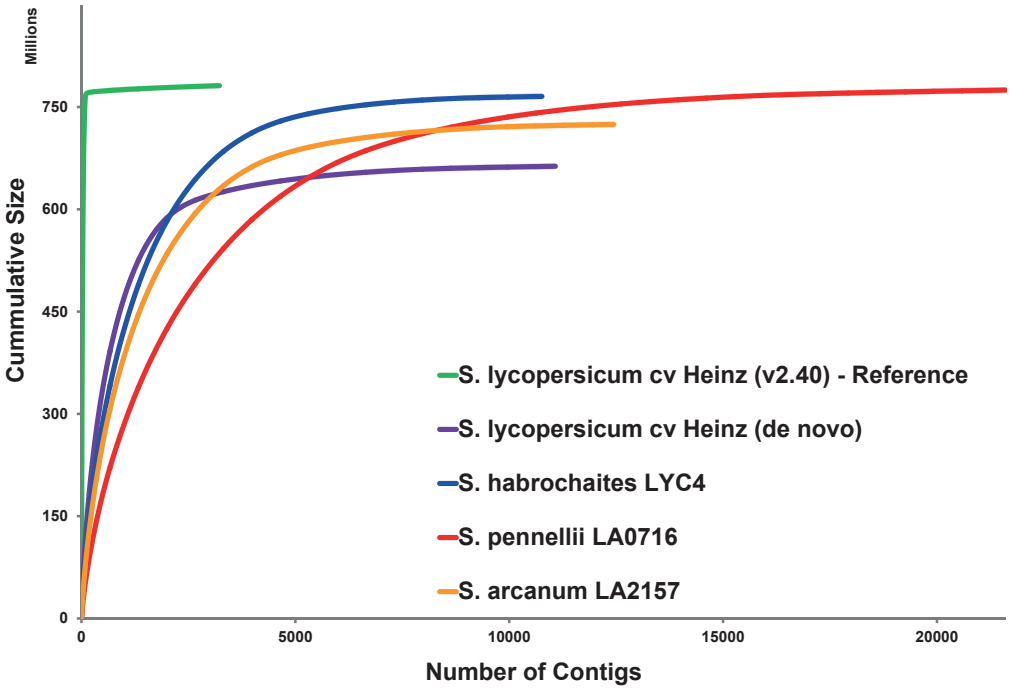


Figure 1: Evaluation of genome assemblies. *De novo* assemblies for *S. lycopersicum* Heinz 1706 (purple), *S. habrochaites* LYC4 (blue), *S. pennellii* LA0716 (red) and *S. arcanum* LA2157 (orange) were generated with the ALLPATHS-LG assembler and scaffolded with SCARPA scaffolder using the 454 data. The number of contigs (x-axis) is plotted against the cumulative contig size (y-axis) when contigs are ordered by size, largest first. The gold standard assembly *S. lycopersicum* cv Heinz 1706 v2.40 is plotted in green.

S. galapagense, *S. cheesmaniae* and *S. pimpinellifolium* accessions. Specific members of the Arcanum, Eriopersicon, and Neolycopersicon groups SNP numbers sharply increase (figures 3 and 4), which correlates with their more distant position in the phylogenetic tree in the tomato clade (Peralta *et al.*, 2008).

When compared to the Heinz annotated genome, in all accessions we consistently observed a significant higher SNP frequency in intergenic regions than in genic regions. Approximately, 89.47%±3.03% of the polymorphisms falls into intergenic regions, while 7.55%±2.19% maps to introns and 2.33%±0.68% maps to exons (figure 5). Of the polymorphisms in exons, 55.17%±11.54% is synonymous while 44.83%±21.03% is non-synonymous (figure 6).

The number of SNPs in wild species on average appears 20 times higher than in crop tomatoes. These results are consistent with the notion that crop tomato genomes are extensively genetically eroded compared to the large genetic diversity found among the wild species. A striking trend is the genome wide ratio between synonymous and non-synonymous SNPs (dN/dS). For crops, non-synonymous SNPs outnumber synonymous

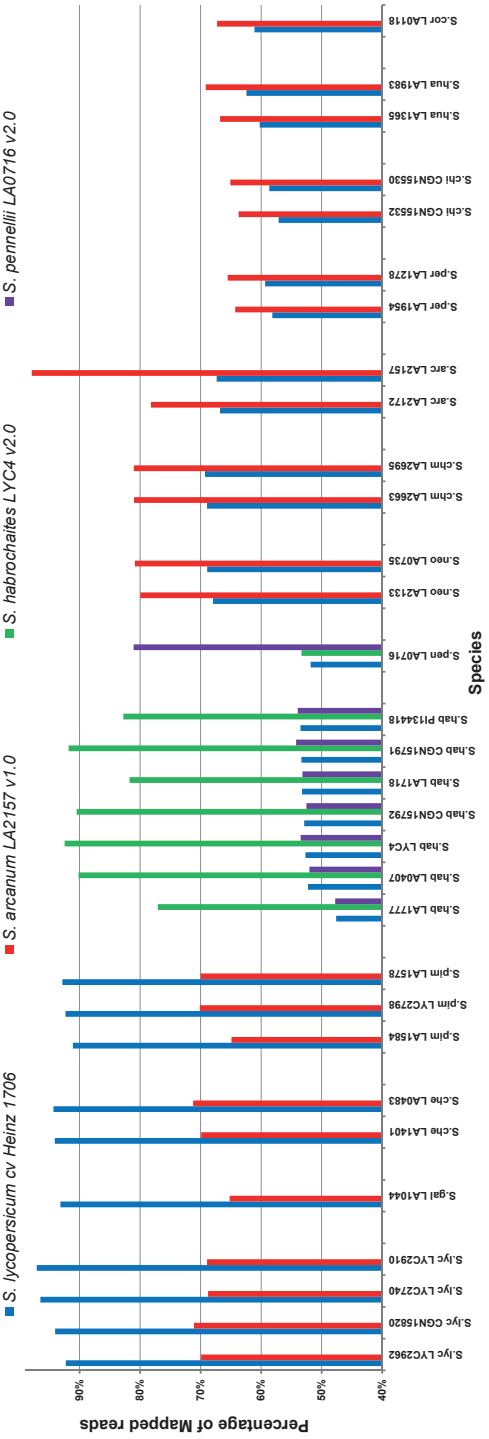


Figure 2: Percentage reads from *S. lycopersicum*, *S. arcanum*, *S. habrochaites* and *S. pennellii* accessions mapped against reference genomes. Species are indicated on the x-axis. Bar color codes correspond to the reference genome indicated in the legend that was used for mapping.

Table 2: De novo assembly statistics for *S. arcanum* (LA2157), *S. pennellii* (LA0716) and *S. habrochaites* (LYC4).

Name	N25	l25	N50	l50	N75	l75	N90	l90	Longest	Shortest	Mean	Median	Num Contigs	Total length
S. lycopersicum cv Heinz v2.40	23,291,314	7	16,467,796	17	7,023,442	35	3,041,128	57	42,121,211	2,000	242,428	2,847	3,223	781,345,411
S. lycopersicum cv Heinz de novo V2.0	165,328	609	87,131	1,944	41,078	4,574	18,768	7,902	963,611	883	35,483	14,872	17,744	629,616,014
S. lycopersicum cv Heinz de novo V3.0	711,921	154	373,293	481	165,009	1,145	46,832	2,176	2,560,154	883	59,866	6,490	11,077	663,130,306
S. habrochaites de novo V2.0	176,864	680	97,427	2,111	49,361	4,745	23,010	7,950	990,615	903	44,066	20,901	16,708	736,254,084
S. habrochaites de novo V3.0	487,032	257	253,002	819	117,458	1,922	47,819	3,402	2,330,637	903	71,129	12,705	10,763	765,557,122
S. pennellii de novo V2.0	128,631	958	70,609	2,926	33,224	6,709	14,641	11,588	627,531	887	27,883	11,066	26,421	736,687,777
S. pennellii de novo V3.0	235,771	544	127,741	1,683	58,513	3,920	19,193	7,186	1,470,620	887	35,862	7,953	21,606	774,839,444
S. arcanum de novo V1.0	35,814	3,716	16,603	12,382	4,517	36,954	1,329	83,146	241,690	200	2,869	428	290,145	832,461,203
S. arcanum de novo V2.0	159,766	725	85,931	2,203	41,599	5,035	19,217	8,546	683,045	905	36,468	16,124	18,638	679,689,580
S. arcanum de novo V3.0	420,446	292	221,078	892	101,738	2,081	39,892	3,708	1,856,562	895	58,224	9,554	12,443	724,486,902

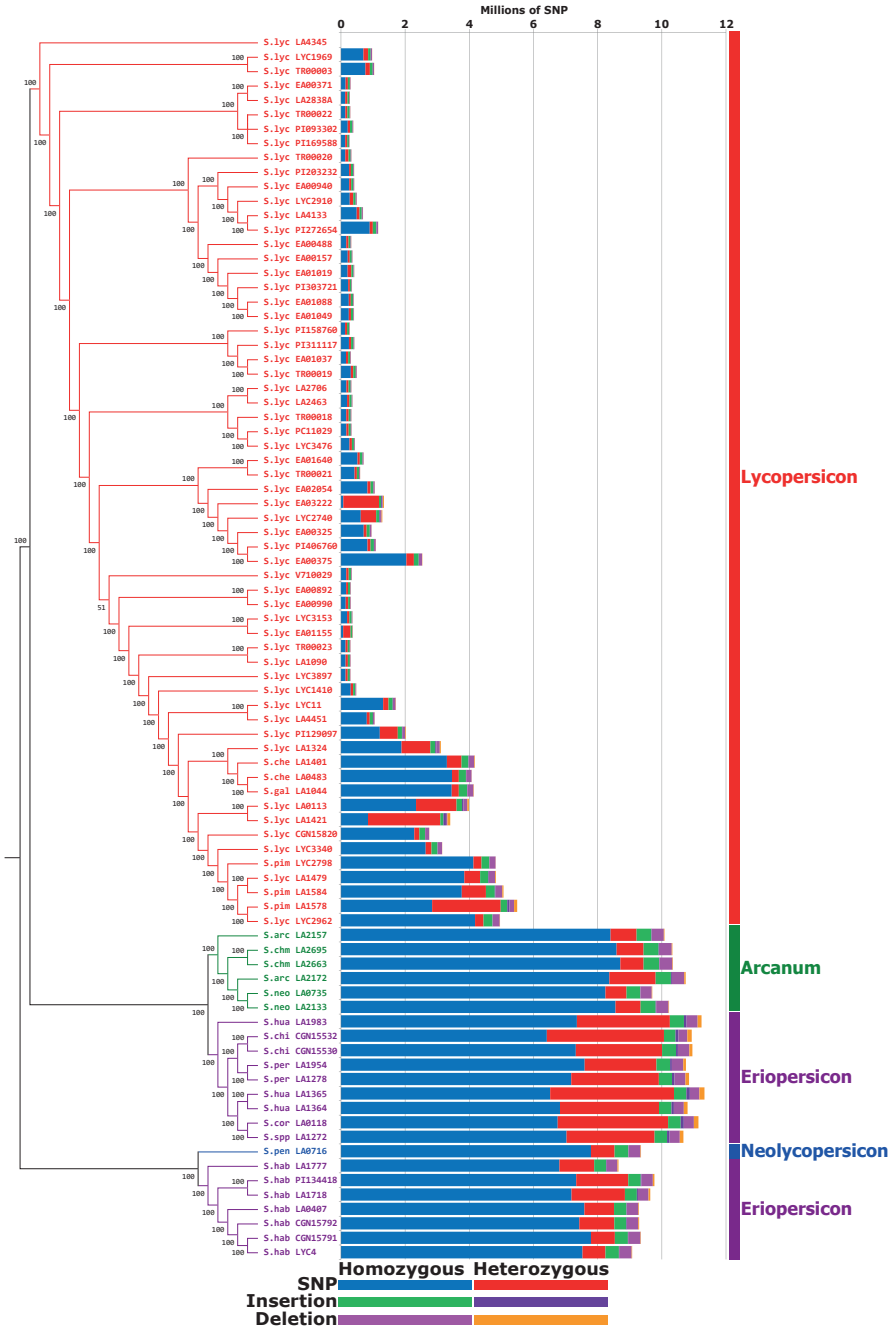


Figure 3: Strict consensus tree based on whole genome homozygous SNPs from 84 accessions with overlaid bootstrap values obtained by Maximum-Likelihood analysis. Bars show the amount of SNPs (in millions) of the different classes of polymorphisms per accession.

SNPs while the opposite is generally true for wild species (figure 6). Although we currently have no clear explanation for the higher dN frequency in crop accessions, it might partly be the result of the artificial selection pressure imposed by breeding, maintaining only a relatively small number of SNPs under positive selection while allowing the fixation of many non-synonymous SNPs as has previously been reported for tomato by comparative transcriptomics (Koenig *et al.*, 2013).

A JBrowse (Skinner *et al.*, 2009; Westesson *et al.*, 2013) supported overview of the SNP and InDel variation in the 84 accessions can be accessed in the tomato 100+ variant browser that is publicly accessible via http://www.tomatogenome.net/VariantBrowser/.

Heterozygosity and introgressions

For the lycopersicum accessions, highest heterozygosity levels were observed for beef type accessions *S. lycopersicum* EA03222 and EA01155 (Dana) as shown in figure 3. With respect to mating type, highest ratios were found for allogamous SI wild species, while facultative SC wild species display an intermediate heterozygosity ratio (figure 7). On average autogamous SC species have a slightly lower heterozygosity level compared to facultative SC species, of which the autogamous SC wild species *S. neorickii* LA0735 has the lowest (figure 7).

Surprisingly, some tomato accessions display considerable high SNP

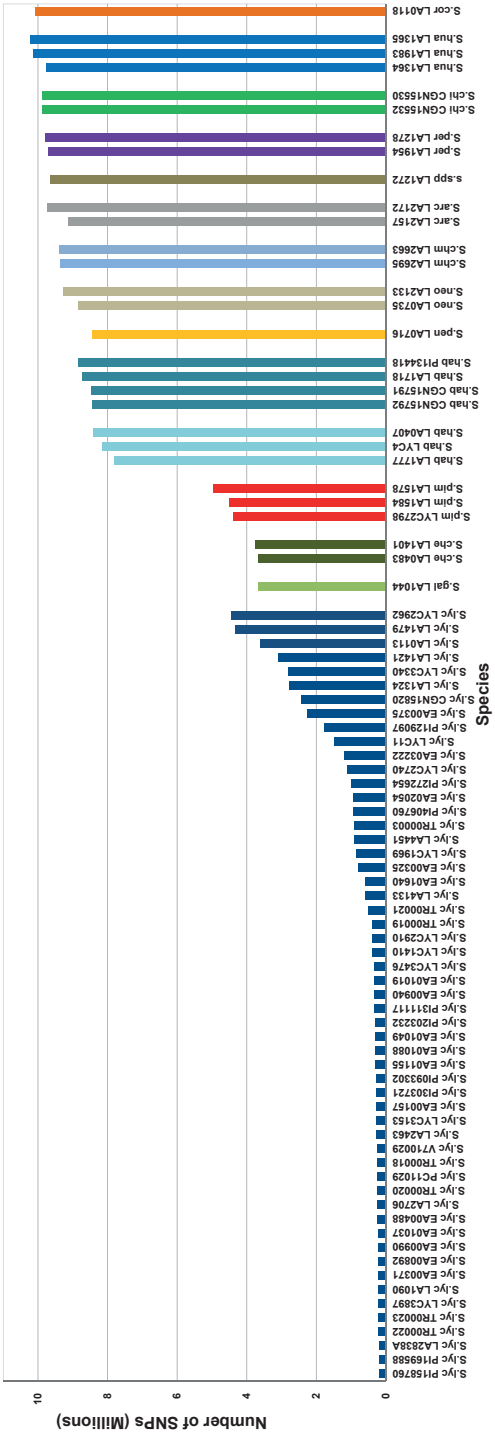


Figure 4: Genome wide SNP counts from 84 accessions versus reference genome of *S. lycopersicum* Heinz 1706. Accession names are indicated below the X-axis. Species groups are indicated by color.

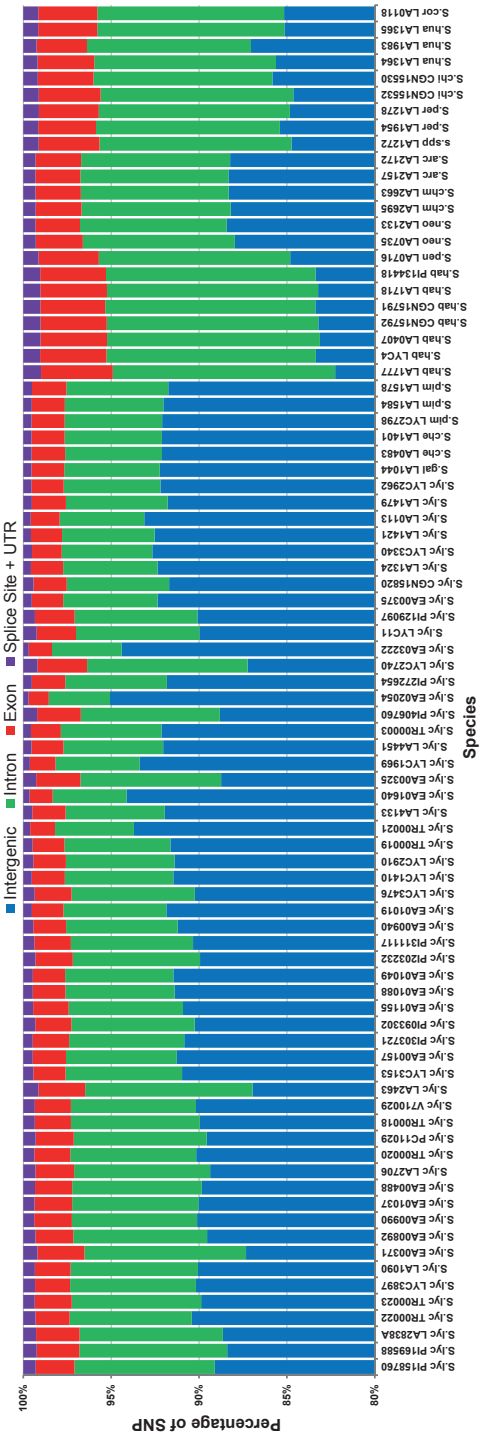


Figure 5: Genome wide SNP ratio for 84 accessions. SNP classes are color coded as indicated. Accession IDs and SNPs percentage are indicated on the x-axis and y-axis, respectively.

counts (figures 3 and 4) which might be attributed to introgressions. Indeed, additional footprints for interspecific introgressions in 54 lycopersicum accessions compared to Heinz were found by testing the SNP distributions in 1 Mbp sized bins along each of the 12 chromosomes. Bins with SNP counts deviating significantly from the average (z-test $p < 0.05$) were considered as either introgressed in crop accessions or natural highly divergent in wild *S. lycopersicum* species after subtracting the number of SNPs found when mapping Heinz reads against itself and correcting for chromosome position and species specific effects by median polish. Introgressions appeared in 5.56%±7.98% of the 767 1Mbp bins with *S. lycopersicum* PC11029 having the smallest number (0.13%), *S. lycopersicum* LA0113 having the highest number (31.42%), and *S. lycopersicum* PI272654 with 0.91% of its bins marked as introgression. Cherry, giant and beef tomatoes have a higher number of introgressions among the crops, while wild *S. lycopersicum* species are even more divergent (figure 8).

The specific chromosome locations of divergent SNP intervals are displayed in figure 8. Here, similar patterns of SNP concentrations can be observed between crop accessions which most likely are introgressions originating from the same donor accessions. In some cases the most likely source of introgressions could be deduced from the SNP identity and the phylogenetic distance inferred from the SNP alignment. Indeed, when plotting the chromosomal

SNP distribution, we found a 2.2 Mb introgressed segment in the long arm of chromosome 6 roughly between Tomato-EXPEN2000 genetic markers C2_At4g10030 (44 cM) and TG365 (50 cM) for the accessions LA2838A (Alisa Craig), LA2706 (MoneyMaker), LA2463 (All Round) and CGN15820. Phylogenetic distance analysis reveals a 2.2Mb segment in the heirloom open pollinated tomato accession MoneyMaker is most closesly related to the wild species *S. pimpinellifolium* LYC2798 (figure 9). Interestingly, the Heinz ITAG 2.4 annotation of this segment points to several loci that have been implicated in hormone induced stress responses, fruit development, flavonoid phytonutrient production, and MAPK mediated production of reactive oxygen species involved in innate immunity to Phytophtera infestance induced late blight.

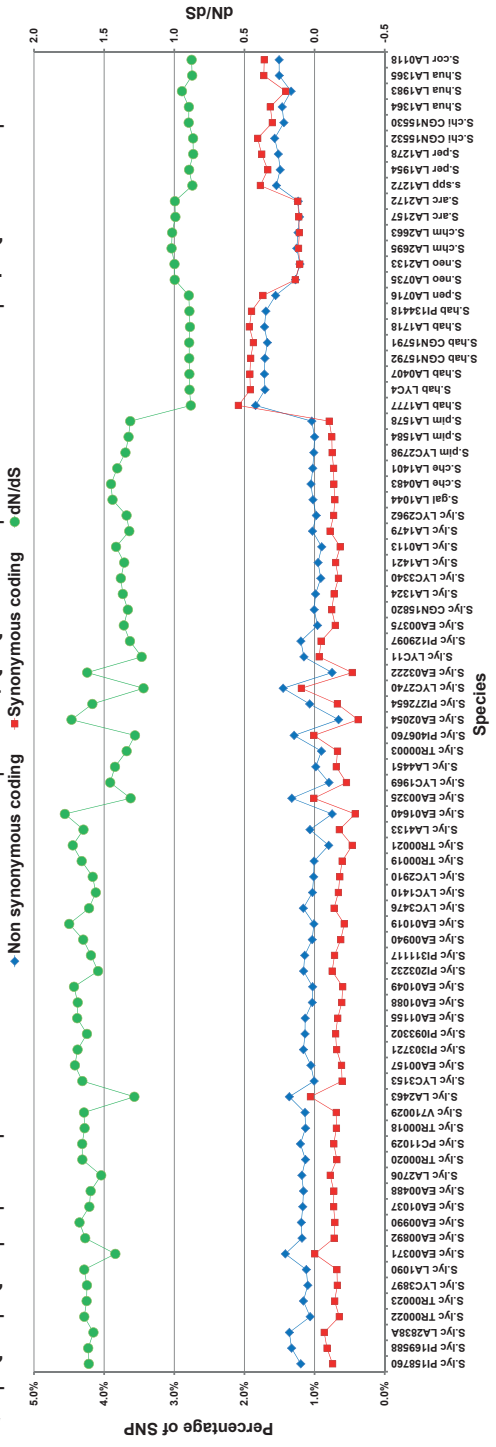
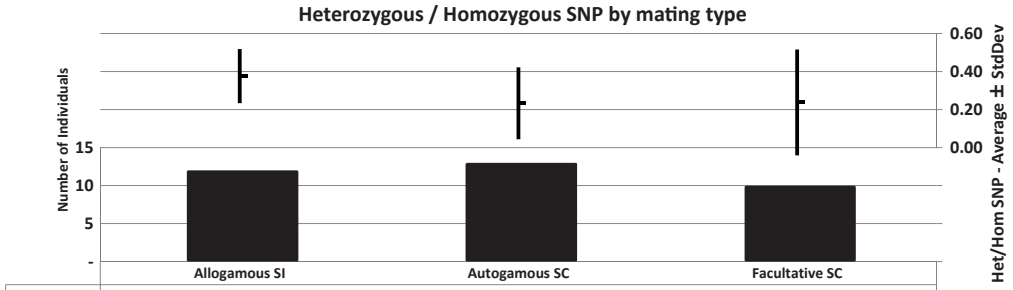


Figure 6: Non-synonymous (dN) and synonymous (dS) SNPs in tomato accessions and related wild species. The dN and dS percentage, and dN/dS ratio relative to the total number of SNPs per accession is indicated in the left and right vertical axis respectively.

Sequence diversity and phylogenetic relationships

SNPs in genes related to fruit and growth diversification

To analyze diversity in specific genes and loci which underlie a phenotypic effect on Fruit Diversification and Plant Growth (FDPG), we determined the orthologs for *ovate* (Solyc02g085500), *fw2.2* (Solyc02g090740), *ls* (Solyc07g066250), *og/beta* (Solyc06g074240), *lcy1* (Solyc04g040190), *lfy* (Solyc03g118160), *rin* (Solyc05g012020), *sp* (Solyc06g074350), *fer*



	Allogamous SI	Autogamous SC	Facultative SC
Number Indiv	12	13	10
+1δ	0.52	0.42	0.52
-1δ	0.23	0.04	-0.04
μ	0.38	0.23	0.24

Figure 7: Heterozygosity level by mating type. Bars heights indicate the number of assessed species per mating type. Average ratios between homozygous and heterozygous SNPs are indicated by horizontal line while its standard deviations are represented by vertical bars.

(Solyc06g051550), *style* (Solyc02g093580), *psy1* (Solyc03g031860), *lin5* (Solyc09g010080), and for locus *lc* (gb|JF284941). Based on the ITAG 2.4 annotation of the tomato reference genome (The Tomato Sequencing Consortium, 2012), the polymorphisms in the orthologs were classified as coding or non-coding, and as non-synonymous or synonymous (silent) SNPs to compare intra- and interspecies sequence diversity and SNP effects among the 84 accessions. FDPG genes in many cases underlie a phenotypic trait that is determined by a few SNPs or sometimes a single one (see below). While tomato breeding has primarily been directed toward selection of these traits, it is conceivable that a SNP determining a single trait went through a positive selection, whereas the bulk of the genes were subjected to a more relaxed selection. Non-synonymous SNP counts in FDPG genes from wild species are consistently higher than observed for a randomly selected set of genes. In contrast, synonymous SNP are consistently lower. Nevertheless, both counts are just within 1 standard deviation away from the average (figure 10). Perhaps this observation reflects a higher selection pressure in wild species than in crop accessions against deleterious mutations in FDPG genes.

Lycopersicon, Aracanum, Eriopersicon and Neolycopersicon specific SNPs

Several characteristic SNPs were found distinctive for the Lycopersicon, Arcanum, Eriopersicon and Neolycopersicon section. For example, the red or orange to yellow fruited Lycopersicon group accessions have a GTC codon in the *og/beta* gene of tomato chromosome 6 for the Val_{23} amino acid in the chromoplast specific lycopene beta cyclase, whereas the green-fruited Arcanum, Eriopersicon and Neolycopersicon species have a non-synonymous TTC (Phe) substitution. The *lcy1* gene on chromosome 4, which has GAG codon for the Gln_{30} residue of lycopene beta cyclase 1, has been substituted in all accessions in the Arcanum group. In particular, *S. chmielewskii* accessions have a GTG (Val), while the *S. neorickii* and *S. arcanum* accessions have a CTG (Leu) codon.

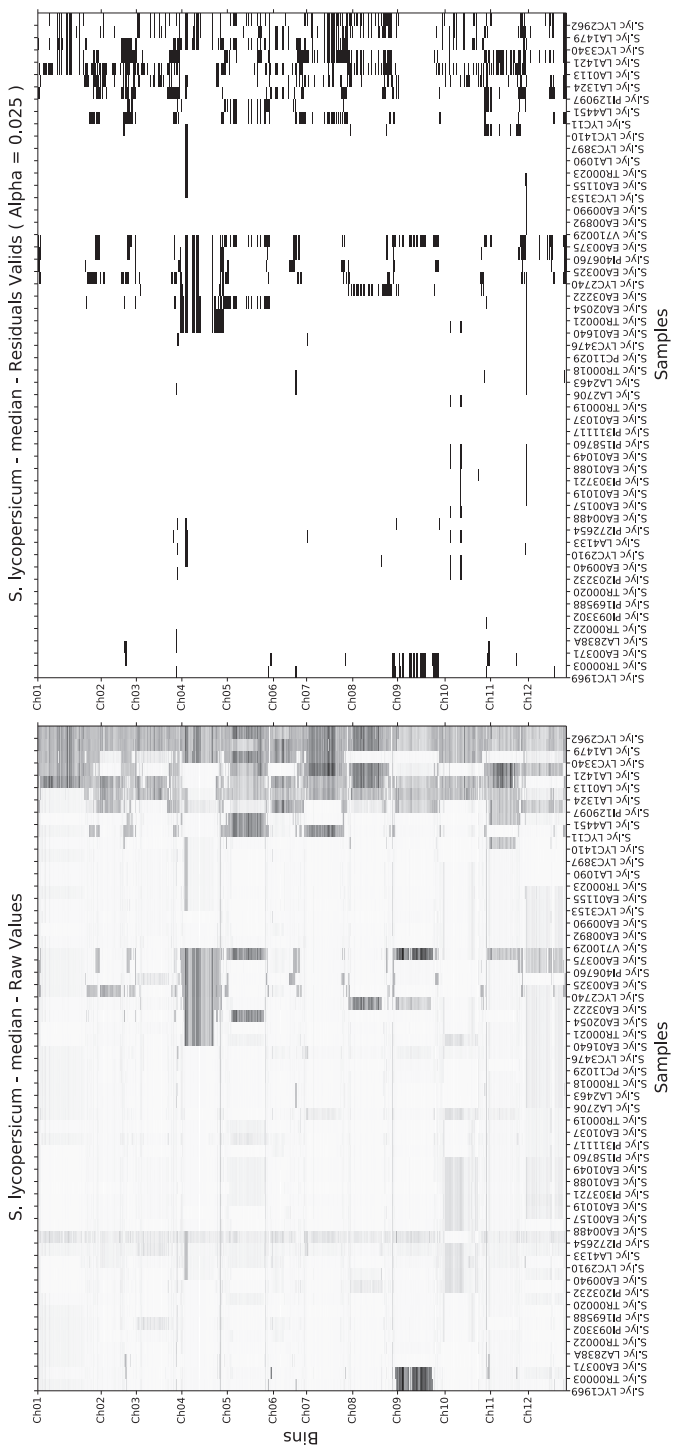


Figure 8: Heatmap showing the raw SNP counts along 12 chromosomes for 54 *S. lycopersicum* accessions in 1Mbp bins intervals (left panel) and In the right panel, regions with SNP counts above average after median polishing using α = 5% for the z-score

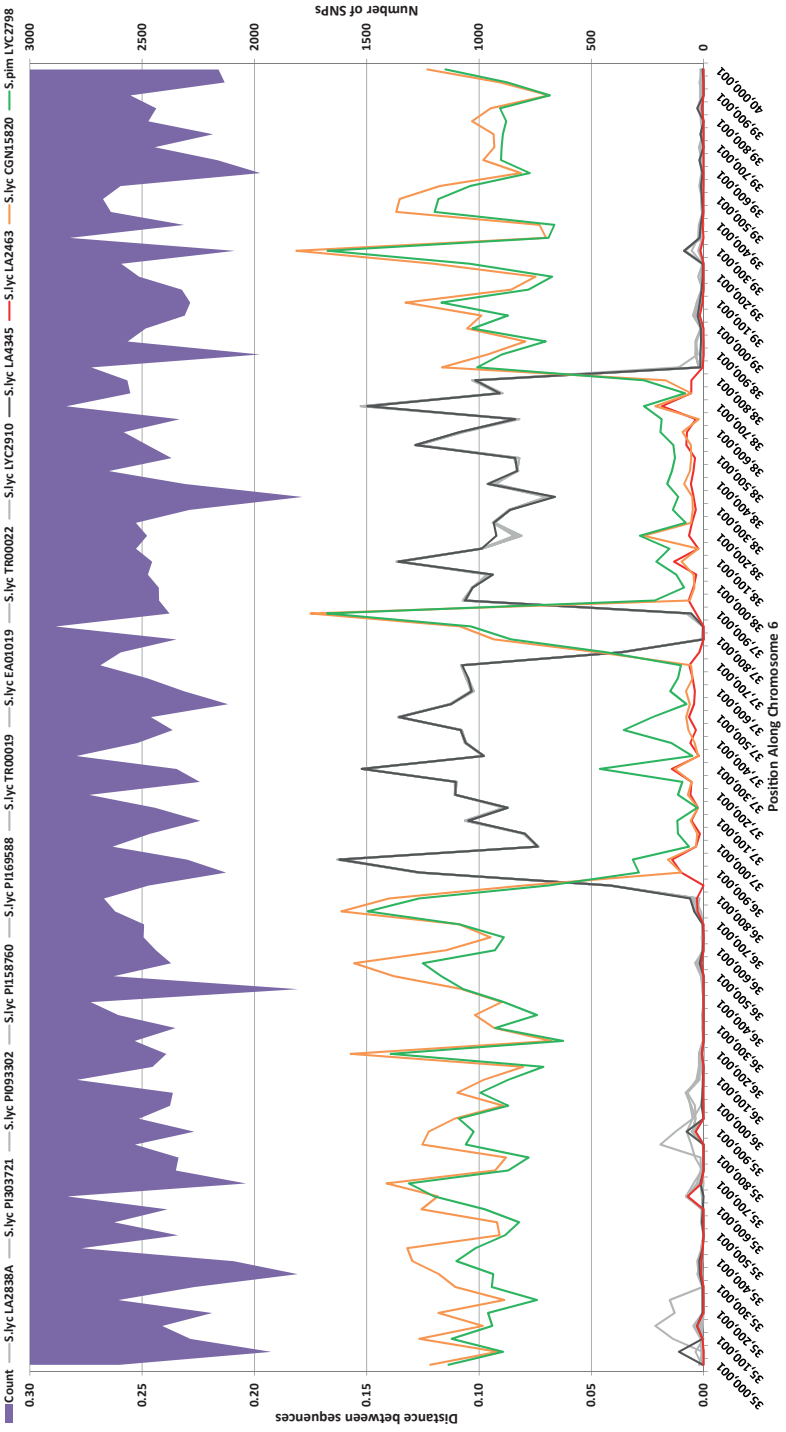


Figure 9: Sequence distance graph for tomato accessions. The top graph displays the number of SNPs (right y-axis, inverted) along chromosome 6 (x-axis) which is used to calculate the sequence distance (left y-axis). Color coded lines display the level of phylogenetic distance compared to *S. lycopersicum* LA2463 (Moneymaker). Please note that some colored lines overlap and are merged into a black line.

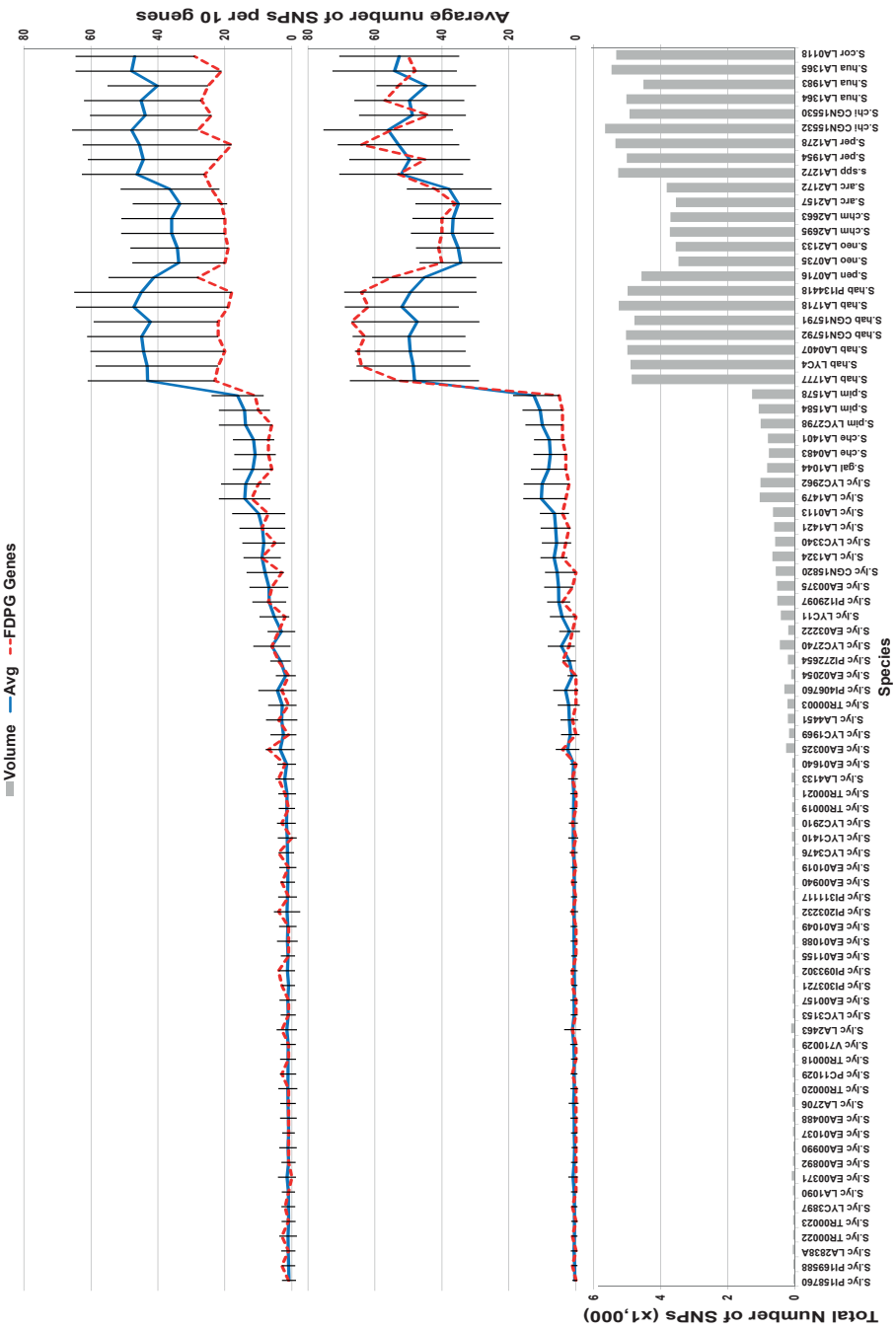


Figure 10: Average and standard deviation of non-synonymous (dN, top panel) and synonymous (dS, middle panel) SNPs in 100 random groups of 10 genes (black vertical lines), its average (blue line) plus 10 Fruit Diversification and Plant Growth (FDPG) genes (red line). The total number of SNPs per accession is indicated by grey bars (lower panel).

Accession		Color	Allele	Chr	Gene id	Mutation	Effect
	Heinz 1706	red	r	3	Psy1	wt	Lys389
	Galina	yellow	ry	3	psy1	G>del	Lys389>Ser, stop
	Taxi	orange	r	3	Psy1	wt	Lys389
	Iidi	yellow	ry	3	psy1	G>del	Lys389>Ser, stop
	RF17	yellow	ry	3	psy1	G>del	Lys389>Ser, stop
	Black Cherry	purple	ogc	6	b	A>del	Lys35>Asn, fs

Figure 11: Accession specific fruit colour traits. Fruit colour, SNP position and coding change in the Psy1 and B gene underlying the fruit colour phenotype are listed accordingly.

Species specific SNPs

In addition to group specific polymorphisms, we also found species specific SNPs. For example, further downstream in the *og/beta* orthologous gene of *S. corneliomulleri* a GCT (Ala_{437})>ACT (Thr) and a TTG (Leu_{464})>TTT (Phe) SNP occur, whereas the AAA codon for amino acid Lys_{277} in tomato is substituted into an ATA (Ile) for the *S. chmielewskii* accessions. In the *lfy* gene orthologs, a synonymous SNP TTA (Leu_{25})>CTA is shared by the *S. huaylasense* accessions, whereas a CCA (Pro_{122})>CAA (Gln) nucleotide substitution, is characteristic for *S. chmiemlewskii* accessions. In the fer gene *S. peruvianum* accessions have a CGATGA insertion (AspAsp) downstream and adjacent to (Asp_{89}). *S. arcanum* and *S. chilense* accessions share a GCC (Ala_{107})>GCA synonymous SNP in the *ovate* gene ortholog, whereas we detected several intron SNPs that are specific for *S. neorickii* in the *sp* orthologous gene. Finally, in the style gene of *S. lycopersicum* accessions a TTT (Phe_{80})>TTC substitution is characteristic for *S. chilense* accessions.

Accession specific SNPs related to fruit traits

We also observed accession specific polymorphisms related to specific fruit traits. Black Cherry has a single nucleotide deletion in the coding sequence of the chromosome 6 B gene (figure 11). This specific deletion occurs in the old-gold-crimson (*ogc*) null allele (Ronen *et al.*, 2000). The resulting frame shift causes the loss of lycopene-ß-cyclase function underlying the accumulation of lycopene and dark red/purple appearance of tomato

fruits, and thus is likely to be the cause for the characteristic dark red/purple flesh-coloured fruits of Black Cherry. Galina, Iidi and T1039 are yellow-skin cherry tomatoes, and have a single nucleotide deletion resulting in a frame shift causing a Lys_{389}>Ser substitution and a premature TGA stop codon directly downstream that would result in a truncated psy1 protein lacking the terminal 23 amino acids (figure 11). In this respect it is interesting to note that the ry mutant allele, which encodes a phytoene synthase lacking these terminal amino acids, underlies the yellow-coloured fruit skin phenotype in tomato mutants (Fray and Grierson, 1993).

Fruit shape and size in tomato is influenced by locule number. Two QTLs, lc and fas have major effects on these traits and can act synergistically leading to extreme high locule numbers (Cong *et al.*, 2008; Munos *et al.*, 2011). Fas is the major gene responsible for increasing locule numbers from 2 to more than 6, while lc has a weaker effect increasing locule numbers to 3 or 4. Two T>C and A>G SNPs are associated with the high locule number allele (*lch*), while an extreme high locule number caused by down regulation of a YABBY-like transcription factor is associated with a 6-8kb insertion in the first intron of the fas gene (Cong *et al.*, 2008). Sequence analysis revealed that all bilocule accessions have the low locule number allele (*lcl*), while accessions with 3 to 4 locules (except Cal J TM VF and Dana) have the *lch* allele. Pear-shaped tomato fruit is controlled by the quantitative trait locus *Ovate*. The allelic interactions at the *ovate* locus have been described as recessive but their expression depends on the genetic background (Ku *et al.*, 1999). Liu and co-workers (2002) showed that a GAA (Glu_{279})>TAA non-sense mutation in the second exon causes an early stop codon and a premature translation termination resulting in a 75 amino acid truncated *ovate* protein (AAN_{17752}) leads to pear-shaped fruit formation. All accessions with pear-shaped fruits have the premature stop codon, while the mutational effect is less pronounced in the *ovate*-fruited accession 'Porter' (figure 12). Hereafter, we address sequence diversity in view of the intraspecies and interspecies phylogenetic relationships.

Phylogenetic relationships

Cladistics based on molecular data resulted in the clear grouping of species within the Solanum genus section Lycopersicon (Peralta *et al.*, 2008). However, at the species level, relationships are still unresolved. For example, while *S. pennellii* was placed in its own group (Neolycopersicon) as a sister to the rest of the section Lycopersicon, it nonetheless appeared as sister to *S. habrochaites* in the main trichotomy (Spooner *et al.*, 2005; Peralta *et al.*, 2008; Grandillo *et al.*, 2011). Our SNP analysis indicates that many polymorphisms are distinct for habrochaites species, whereas *S. pennellii* LA0716 shares many SNPs with accessions of the Arcanum and Eriopersicon groups. This point to a complicated phylogenetic relationship for *S. pennellii* and *S. habrochaites*.

We applied the vast amount of multilocus molecular data to shed more light on the species and accession relationships in the tomato clade. First, we used a limited set of

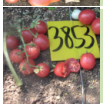
Accession		Shape	Allele	Chr	Gene id	Mutation	Effect
	Heinz 1706	round	Ovate	2	459212746	wt	Glu279
	Cross Country	pear, ovate	ovate	2	459212746	G>T	Glu279>stop
	Iidi	pear	ovate	2	459212746	G>T	Glu279>stop
	Anto	pear, ox	ovate	2	459212746	G>T	Glu279>stop
	WFR	pear	ovate	2	459212746	G>T	Glu279>stop
	RF36	pear	ovate	2	459212746	G>T	Glu279>stop
	RF43	pear	ovate	2	459212746	G>T	Glu279>stop
	Porter	ovate	ovate	2	459212746	G>T	Glu279>stop

Figure 12: Accession specific fruit shape traits. Fruit shape, SNP position and coding change in the *Ovate* gene orthologs underlying fruit shape phenotype are listed accordingly.

polymorphisms to assess the species boundaries and relationships within the tomatoes and wild relatives. The strict consensus tree for ten concatenated genes (figure 13) revealed that all *S. habrochaites* species cluster into a monophyletic group, while *S. pennellii* LA0716 is sister to *S. habrochaites*. The *S. chilense* accessions also group together and cluster with *S. corneliomulleri* and *S. peruvianum* accessions, which are representatives of the former *S. peruvianum* 'southern group', and with accession LA2172. The green-fruited self-compatible (SC) *S. chmielewskii*, and *S. neorickii* species, which are representatives of the Arcanum group (Peralta *et al.*, 2008), are resolved into two monophyletic groups and cluster with two *S. arcanum* species into a larger clade. Furthermore, all red or orange-fruited SC species of the Lycopersicon group (*S. cheesmaniae*, *S. galapagense*, *S. lycopersicum*, *S, pimpinellifolium*) form a well-supported clade. In particular, the orange-fruited *S. cheesmaniae* and *S. galapagense* cluster into a subgroup 4 species. These relations are in agreement with previously presented phylogenetic studies (reviewed by Grandillo *et al.*, 2011).

Next we excluded heterozygous SNPs from the analysis, as they are arbitrarily converted into a single nucleotide call for FASTA converted sequences thereby introducing noise and a possible bias in the data. SNPs in introns were also excluded as they are likely to be under less selective pressure than exon SNPs and probably carry less phylogenetic information and introduce more noise. Figure 14 shows the homozygous SNPs in the FDPG genes have sufficient power to resolve the phylogenetic placement consistent with the grouping at the sectional level as previously described (Peralta *et al.*, 2008). We noticed a slight increase in resolution when comparing the gene tree based on unfiltered and filtered SNPS respectively (figures 13 and 14). Nevertheless, at the species level the placement of lycopersicum, pimpinellifolium, galapagense and cheesmaniae accessions appeared largely unresolved. In this analysis, *S. pennellii* is a sister species from the Arcanum group. We therefore also assessed the clustering using genome wide homozygous SNPs. The whole genome SNP cladogram in figure 3 shows a complete resolution into separate branches with high bootstrap values for each of the Lycopersicon accessions and wild species. Although phylogenetic relationships might be influenced by SNPs that arise from introgressions, the genome wide SNP information generates sufficient resolution power and enables the interspecies and intraspecies identification of all 84 individuals in monophyletic groups. Based on our phylogenetic analysis and SNP sequences we propose to type accession LYC2740 as an *S. lycopersicum* species instead of a *S. pimpinellifolium*. We also observed several *S. lycopersicum* accessions grouping with *S. pimpinellifolium*, *S. galapagense* and *S. cheesmaniae*. Those *S. lycopersicum* accessions likely are hybrids or carry substantial S. pimpinellifoulium introgressions.

Additional analysis should be performed to substantiate this hypothesis. In addition, *S. pennellii* appears a sister species to *S. habrochaites* species group in the whole genome SNP tree, suggesting *S. pennellii* can be considered an intermediate species between *S. habrochaites* and *S. arcanum* which would coincide with its intermediate geographical distribution.

Discussion

Multiple reference genomes and sequence diversity

Our study has yielded a huge amount of precious data on sequence diversity in wild species of the tomato clade. The reads for *S. habrochaites* (78%), *S. arcanum* (73%) and *S. pennellii* (53%) were mapped onto the corresponding species reference genome illustrating the large interspecies sequence variation in the Lycopersicon clade. We also demonstrated dramatic genetic erosion in cultivated tomatoes. As there is an increasing demand for broadening the genetic base of this crop we believe that our study provide pivotal information for future tomato breeding programs. The Heinz reference genome is not only partly representative for the genetic and structural information in the related wild

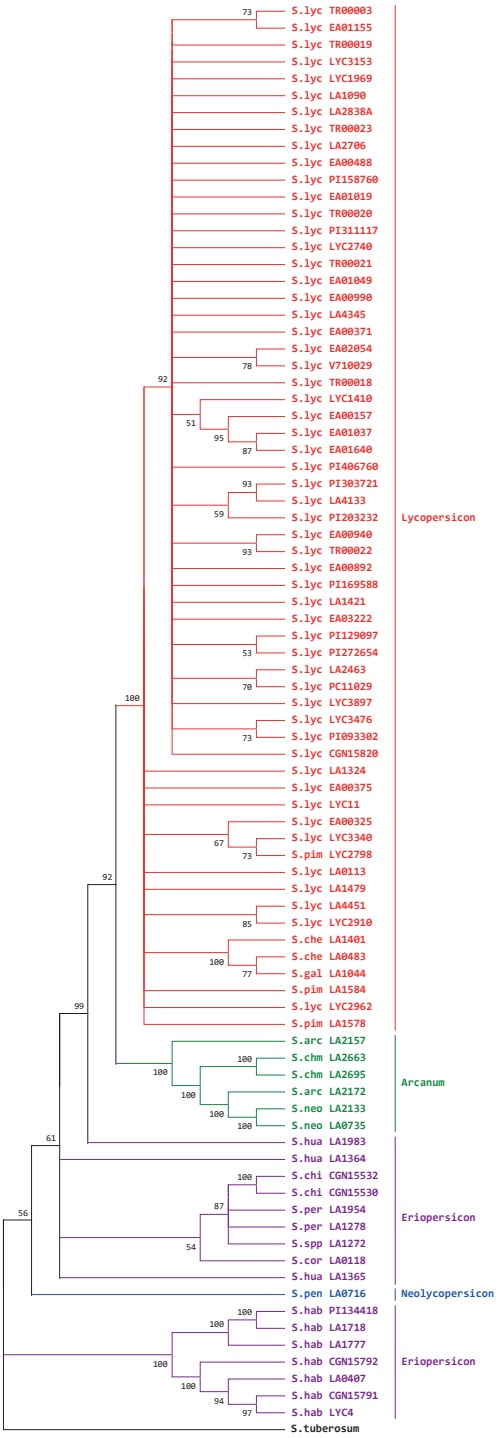


Figure 13: Strict consensus tree based on 10 FDPG genes sequences from 84 accessions and *S. tuberosum* as an outgroup with overlaid bootstrap values obtained by Maximum-Likelihood analysis. Species names are indicated and combined with accession numbers.

species but it also emphasises the need to reconstruct additional reference genomes. The three *de novo* sequenced genomes presented here thus constitute a valuable additional resource to the currently available genomic tools in support of studies on evolution, domestication and genetic bases underlying important traits like disease resistance and abiotic stress tolerance.

Our sequencing and mapping strategy effectively supports the detection and identification of high-confidence sequence polymorphisms and its application to explain the rich phenotypic diversity among a large set of cultivated tomato accessions and its wild relatives. We observed group, species, and accession specific polymorphisms some of which can be attributed to economically important fruit and growth traits. Such information can easily be translated into array or PCR based assays to genotype genetic variants across extensive populations as well as a population of progeny. Provided that gene models from the Heinz annotated reference genome are also applicable for the other species, we observe 8% to 10% of all sequence polymorphisms to be located in the genic portion of the genome. Non-synonymous and synonymous SNPs each occur at 1% of the total number of SNPs. As a conserved estimate for wild species it would equal to about 1x 105 of non-synonymous SNPs, but little is known how much of the phenotypic

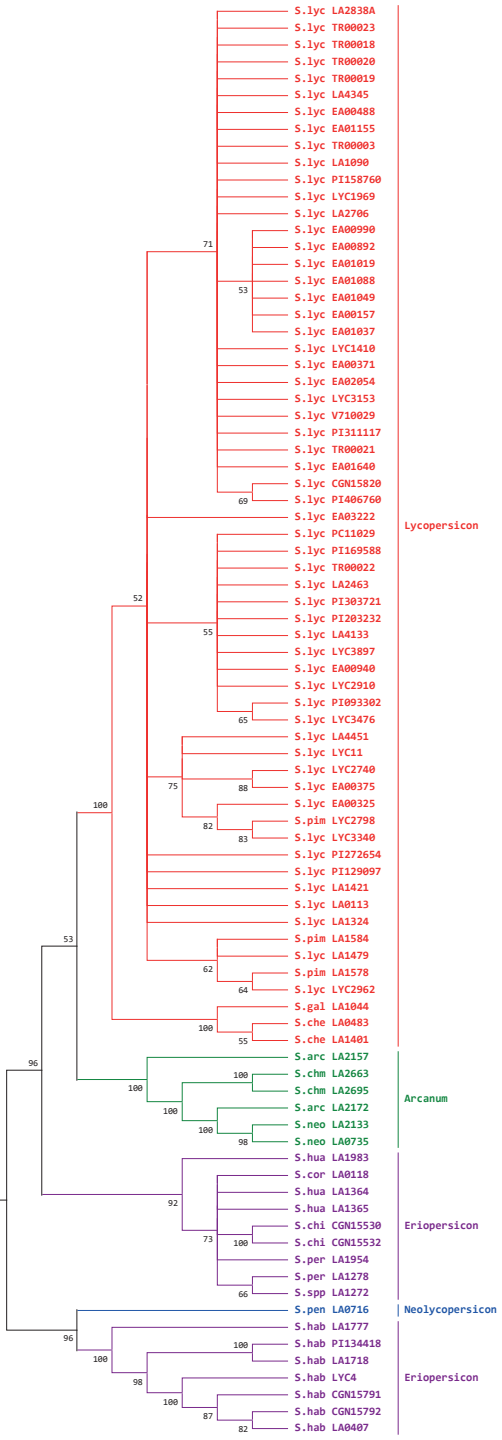


Figure 14: Strict consensus tree based on homozygous SNPs of the exons of 10 FDPG genes sequences from 84 accessions with overlaid bootstrap values obtained by Maximum-Likelihood analysis. Species names are indicated and combined with accession numbers.

diversity can be attributed to this. Considering that traits like fruit colour and shape are determined by a single SNP, the total number of SNPs most likely represents a wealth of diversity that waits exploration. We are nevertheless at the beginning of elucidating their biological relevance.

Relationships of tomatoes and wild species relatives

The past few decades have seen the publication of several phylogenetic studies of Solanum species in the Lycopersicon section, but usage of phenotypic characters, markers and sequencing data resulted in dissimilar trees, with provisional species groupings lacking fully resolved relationships (Grandillo *et al.*, 2011). In this study we reconstructed intra and interspecies relationships for a large number of tomato accessions and related wild species, taking advantage of whole genome sequence data to maximize tree resolution. Initially our phylogenetic analysis focused on genes controlling economically important traits that have been subject to interspecific hybridization breeding. Since a subset of these genes originated from wild species, cladistics potentially may result in skewed relationships. Yet, our Maximum Likelihood consensus cladograms for the targeted genes and for the whole genome SNP sets show a comparable tree topology down to the sub sectional (species group)

level, suggesting that phylogenetic relationships between fruit and growth diversification genes are not particularly biased. While, the strict consensus cladogram for the concatenated fruit and growth diversification genes displays unresolved relationships at the species level for some of the cultivated tomato and *S. peruvianum* accessions, the use of whole genome SNP data allowed increased tree resolution. Indeed, the whole genome SNP set supports the placement of taxa into separate branches with high bootstrap values for each of the accessions and wild species, including corrected placement of several previously putative typed accessions.

Ecological differences probably have resulted in dramatic genome evolutionary consequences. Moreover, there is evidence that mating system shifts have a large impact on complex multigene based traits such as floral and fruit development (Moyle, 2008), which might further account for a large intraspecies variation. Large intraspecies variation trends have been observed for *S. chilense*, which have been grouped into geographic races that can be distinguished both morphologically and genetically (reviewed in Grandillo *et al.*, 2011). Other examples involve remarkable levels of morphological and genetic diversity found in *S. peruvianum* populations (Rick, 1986; Städler *et al.*, 2005 and references therein), which might explain the distinct phylogenetic positions for *S. arcanum* LA2157 and LA2172. Here, we placed both accessions into the Arcanum group with northern species of the peruvianum complex. Accession LA2172, which is allogamous SI, appears sister to the monophyletic *S. neorickii* clade, while LA2157 is facultative SC and sister to the monophyletic *S. chmielewskii* clade. Furthermore, it is important to note that AFLP cladistics previously resulted in the grouping of *S. arcanum* LA1984 with southern *S. peruvianum* species, while the other *S. arcanum* accessions grouped with *S. huaylasense* (Spooner *et al.*, 2005). Interestingly, it has been speculated that accessions such as LA1984 could represent a 'crossing bridge' between morphologically and genetically distinct populations (Rick, 1986; Grandillo *et al.*, 2011).

Detection of introgressions in crop accessions and genome structure

While marker assisted introgressions focuses on the relations between traits and allelic variants, it is mostly used for the indirect selection of genetic determinants for a trait of interest and is restricted to alleles that can be diagnosed. Based on genome wide SNP data, introgressions from *S. pimpinellifolium* into chromosomes 4, 9, 11 and 12 of *S. lycopersicum* Heinz1706 were previously reported (The Tomato Genome Consortium, 2012). Following the same strategy, here the bulk of our introgression detection was based on SNP distributions divergent from the reference genome targeting introgressions not present in Heinz. Our approach shows that both location and size of introgressed segments can be inferred from the SNP distribution. Furthermore, based on the phylogenetic distance we assigned a closely related wild species *S. pimpinellifolium* as the most close donor species among the 84 accessions that we have tested. These results put a new perspective to future introgression hybridization breeding.

The success of introgressive hybridization breeding depends, among others, on the proper identification of colinear chromosome segments in donor and recipient genomes, which in turn is dependent on the consistency and completeness of their assembled genomes. The genome structure of the parental species influence crossing success and a difference in genomic colinearity has a direct effect on chromosome pairing at meiosis and hence determines the rate of alien chromatin transfer into a recipient crop. However, the proper ordering and orientation of contigs into megabase sized scaffolds depends on the availability of genetic and physical maps, which are currently lacking for the three *de novo* sequenced genomes. Furthermore, the N50 contig sizes for *de novo* assemblies of *S. arcanum*, *S. habrochaites* and *S. pennellii* do not exceed 400kb. Although the advances in next-generation sequencing technology for the use of extant germplasm resources now allow relatively fast and cheap assembly of large numbers of complex genomes, it does not yet allow a full genome reconstruction of the Solanaceae family and hence is yet of limited use for introgression breeding. The identification of compatible genomes for introgression breeding, the rearrangement phylogeny within the Solanaceae, and reconstruction of the ancestral Solanum karyotype all require additional physical mapping information on top of the genome sequence information to properly order contigs along the chromosome arms. Therefore, we believe there is room, in the near future, to pursue the integration of NGS and new technological platforms to advance the Solanum genome reconstruction.

Experimental procedures

Selection of tomato accessions

We genotyped the 7000 accessions in the EU-SOL project (https://www.eu-sol.wur.nl) on the basis of 20 traits and markers, followed by a denser genotyping of a subset of 1000 accessions using 384 SNP markers, and a final selection of 200 accessions covering the full genetic diversity of the crop. We also included a set of old cultivars that were selected on the basis of previously documented trait identifications of wild tomato relatives (reviewed in Grandillo *et al.*, 2011).

DNA isolation

Young leaves were collected from the first plant of each plot (self-compatible accessions) or from the pollen acceptor (self-incompatible accessions) for DNA extraction. Approximately 100 mg frozen leaf material was grinded using the Retch Mixer Mill M300. Subsequently, genomic DNA was extracted with a standard DNA isolation protocol (Bernatzky and Tanksley, 1986), using a nuclear lysis buffer with sarkosyl. The DNA was quantified using the Qubit 2.0 Fluorometer (Invitrogen). Per accessions 1,5 – 2,0 µg DNA was used for library construction.

Illumina and 454 libraries sequencing and read mapping

Shallow sequencing of 500bp inserts was carried out using Illumina HiSeq 2000 to generate a 100bp paired end library at an average of 36 fold coverage. Bases with Q < 20 were trimmed before read mapping with BWA (Li and Durbin, 2009, Li and Durbin, 2010) against *S. lycopersicum* cv. Heinz v2.40 with a maximum insert size of 750bp (50% deviation), reporting at most 30 hits and removing PCR duplicates. SAMTOOLS (Li *et al.*, 2009) was used for variant calling without skipping InDels, a minimum gap distance of 5bp, a minimum alignment quality of 20, a minimum depth of 4 and otherwise the default parameters. The same protocol was used to map the wild species to their closest *de novo* version 1.0 assembled counterpart. Contamination with *Escherichia coli*, human, insects, mouse, Phi X 174, yeast and phytoviral genomes (Adams and Antoniw, 2006) was checked with BOWTIE (LaNGMead *et al.*, 2009).

De novo assembly of the three wild species genomes and Heinz

For *de novo* sequencing of *S. arcanum* LA2157, *S. habrochaites* LYC4 and *S. pennellii* LA0716 we sequenced an overlapping paired end library with 170bp insert size, at 93.2, 76.4 and 80.8 fold coverage; re-used the 500bp insert size paired end library at 35.7, 35.6 and 28.2 fold coverage, using 100bp Illumina HiSeq 2000 reads; and a mate pair library with 2kbp insert size at 33.8, 38.0 and 31.2 fold coverage, respectively. Using 454 FLX a long mate pair library of 8kbp insert size and an extra-long mate pairs library of 20kbp insert size was sequenced at 0.55±0.10 and 0.47±0.07 fold coverage, respectively. For *S. pennellii* LA0716 we sequenced an additional short mate pair library of 3kbp insert size at 0.4 fold coverage. On average, reads produced from 454 libraries contained 35%±7% of adaptamers. *S. lycopersicum* cv Heinz 1706 used a reduced set of its original set or reads with 14.78 fold 250 PE, 17.54 fold 300 PE, 37.42 fold 500 PE, 6.25 fold 2kb MP, 6.51 fold 3kb MP, 5.94 fold 4kb MP and 6.02 fold 5kb MP in a total of 69.74 fold coverage for PE libraries, 24.73 fold coverage for MP libraries and 94.47 fold coverage overall.

The *S. pennellii* and *S. habrochaites* data were assembled with AllPaths-LG (assembly version 2.0) according to Gnerre *et al.*, (2010) with ploidy of 2, while *S. arcanum* was assembled using CLC Genomics Workbench v7 (CLC Inc, Aarhus, Denmark) with a bubble size of 300, a minimum contig length of 200 and a word size of 64 (assembly version 1.0). Subsequently, *S. arcanum* was assembled with Allpaths-LG (assembly version 2.0). AllPaths-LG generated scaffolds were further scaffolded using the 454 FLX data in the Scarpa scaffolder (Donmez and Brudno, 2013). Subsequently, the *de novo* assembly statistics were compared to the tomato reference genome *S. lycopersicum* cv. Heinz version SL2.40 (table 2). The CLC, Allpaths-LG, and the AllPaths-LG plus Scarpa assembly is referred to as Version 1.0, 2.0 and 3.0, respectively. *S. arcanum* V1.0, *S. habrochaites* V2.0 and *S. pennellii* V2.0 were used for the mapping of the 84 accessions. Version 3.0 was used to assess genome sizes and rearrangements for all species.

Sequence diversity and phylogenetic relationships of 84 accessions

To assess sequence diversity in domestication syndrome genes, orthologs in 84 accessions were obtained from reciprocal best BLASTN hits of CLC assembled contigs and tomato ITAG 2.4 annotated sequences (http://solgenomics.net/gbrowse/bin/gbrowse/ITAG2.3_genomic) and aligned with ClustalW (Thompson *et al.*, 1994). SNPs were then called using the quality-based variant detection algorithm in CLC. Optimal substitution models for ClustalW-aligned gene and concatenated gene sequences were calculated in Mega5.1 (Tamura *et al.*, 2011). Maximum Likelihood trees for each individual gene as well as concatenated gene sequences were inferred using a Neighbour-Joining initial tree (NJ) followed by Nearest-Neighbour Interchange (NNI). Phylogenies were tested using 1000 random genes separated in 100 sets of 10 genes (figure 10). Finally, strict consensus trees for individual genes and concatenated genes were calculated using a cut-off value of 50%.

For each species we used a concatenation of all homozygous non-unique SNPs (Van Gent *et al.*, 2011) with quality above 20, which were obtained from the VCF files generated by Bwa and SamTools. Multiple Nucleotide Polymorphisms (MNP) and Insertion-Deletion events (InDels) were disregarded due to their low frequency and the low aligment speed. We used ITAG v2.3 gene models with the FastTree 2.1.7 software (Price *et al.*, 2010) for a heuristic neighbor-joining as input to the approximately-maximum-likelihood algorithm, thus reducing the number of trees with a mix of nearest-neighbor interchange (NNIs) and subtree-prune-regraft. A Jukes-Cantor generalized time-reversible model, bio neighbor joining (BioNJ) weighted joins, 100 bootstrap resamples and gamma fitting for reported likelihood were used in the analysis.

Annotation of SNP calls

All VCF files from the mapped individuals were processed using SNPeff 3.4 (Cingolani *et al.*, 2012) base on ITAG 3.1 annotation and default parameters. SNPeff annotates SNP in the VCF files, based on their position and the reference annotation, with their effects and reports statistics such as rates of synonymous and non-synonymous SNPs (figures 6 and 10), heterozygosity levels (figures 3 and 7), the number of SNPs per 1 Mbp bins (figure 8) and location of the SNP (figure 5).

Introgression estimation in *S. lycopersicum*

To estimate the level of introgression in the *S. lycopersicum* species, we used the median polish procedure (Mosteller and Tukey, 1977 and Xie *et al.*, 2009) on the table of SNP count for each accession per 1Mbp bin along each chromosome (figure 8, left panel) to remove species and bin specific effects (species or bins naturally having a higher number of SNPs). The residuals were tested using a z-test and bins from crop accessions with residuals significantly different from the average ($p < 0.05$) were labeled as introgressions (figure 8, right panel). Note that in wild *S. lycopersicum* we cannot discriminate between natural variance or interspecific crossings.

Variant browser

JBrowse 1.10.12 (Skinner *et al.*, 2009) was set-up to visualize the detected structural variants. The SL2.40 genome assembly and ITAG 2.31 genome annotation was loaded together with the VCF files of the 84 accessions.

Sequence repository

Sequence reads and associated analyses are deposited at the European Nucleotide Archive (http://www.ebi.ac.uk/ena/) under PRJEB5226 (*S. arcanum* LA2157), PRJEB5227 (*S. habrochaites* LYC4), PREB5228 (*S. pennellii* LA0716), and PRJEB5235 (reseq accessions).

Acknowledgements

This research was supported by The Technological Top Institute Green Genetics (TTI GG), financial aid from the Dutch Ministry of Economic Affairs, Agriculture and Innovation, Centre for BioSystems Genomics (CBSG).

Chapter 3

Introgression Browser: High-throughput whole-genome SNP visualization

Aflitos, Saulo Alves, Gabino Sanchez-Perez, Dick Ridder, Paul Fransz, Michael E. Schranz, Hans Jong, and Sander A. Peters. "Introgression browser: high-throughput whole-genome SNP visualization." ***The Plant Journal*** **82, no. 1 (2015): 174-182. DOI: 10.1111/tpj.12800**

Summary

Breeding by introgressive hybridization is a pivotal strategy to broaden the genetic basis of crops. Usually, the desired traits are monitored in consecutive crossing generations by marker-assisted selection, but their analyses fail in chromosome regions where crossover recombinants are rare or not viable. Here, we present the Introgression Browser (IBROWSER), a bioinformatics tool aimed at visualizing introgressions at nucleotide or SNP accuracy. The software selects homozygous SNPs from Variant Call Format (VCF) information and filters out heterozygous SNPs, Multi-Nucleotide Polymorphisms (MNPs) and insertion-deletions (InDels). For data analysis IBROWSER makes use of sliding windows, but if needed it can generate any desired fragmentation pattern through General Feature Format (GFF) information. In an example of tomato (*Solanum lycopersicum*) accessions we visualize SNP patterns and elucidate both position and boundaries of the introgressions. We also show that our tool is capable of identifying alien DNA in a panel of the closely related *S. pimpinellifolium* by examining phylogenetic relationships of the introgressed segments in tomato. In a third example, we demonstrate the power of the IBROWSER in a panel of 597 Arabidopsis accessions, detecting the boundaries of a SNP-free region around a polymorphic 1.17 Mbp inverted segment on the short arm of chromosome 4. The architecture and functionality of IBROWSER makes the software appropriate for a broad set of analyses including SNP mining, genome structure analysis, and pedigree analysis. Its functionality, together with the capability to process large data sets and efficient visualization of sequence variation, makes IBROWSER a valuable breeding tool.

Introduction

Plant breeders apply introgressive hybridization for incorporating valuable traits from related species via interspecific hybridization and repeated backcrossings (Anderson, 1953). During meiotic prophase of these hybrids, crossover recombination may occur between the donor species and its homeologous counterpart of the recipient crop, thus creating novel allele combinations. Recombination usually involves the exchange of large fragments (up to complete chromosome arms), and can therefore result in the introgression of additional unwanted genes besides the target gene (Zamir 2001; Qi *et al.*, 2007). Breeders aim at maintaining the chromosomal regions of interest while removing the unwanted traits from the alien relative by recurrent backcrossing with the crop, followed by trait selection. The detection of introgressed segments harboring genes associated with economically important traits, whether or not linked to potentially deleterious or unwanted genes / alleles from the donor species, is therefore of eminent interest to breeders. Tracing of such introgressions in offspring families may also provide insight into homeologous recombination leading to incorporation of the desired genes from the alien donor into the recipient crop, while eventually losing genes controlling unwanted traits.

Detection of introgressed regions has previously relied extensively on diverse molecular marker technologies including Simple Sequence Repeats (SSR), Restriction Fragment Length Polymorphisms (RFLP), Amplified Fragment Length Polymorphisms (AFLP) and DNA microarrays (Rieseberg and Ellstrand, 1993; Powell *et al.*, 1996; Dekkers

and Hospital, 2002; Víquez-Zamora *et al.*, 2013). However, the power of marker-assisted technologies to detect and delineate introgressions and chromosome rearrangements is limited due to low marker placement accuracy and even lack of markers (Kumar *et al.*, 2012; Víquez-Zamora *et al.*, 2013; Anderson *et al.*, 2011). In addition, sequence duplications, heterozygosity, and discrepancies between genetic and physical maps can seriously hamper data interpretation. Such limitations can partly be overcome through the use of *in situ* hybridization techniques such as Genomic *in situ* Hybridization (GISH) and Fluorescent *in situ* Hybridization (FISH). GISH technology can be used to obtain information on the size and number of alien chromosomes or chromosome segments, interspecific and intergeneric translocations resulting from homeologous recombination, and the presence and approximate location of introgressed genes (Schwarzacher *et al.*, 1992; Thomas *et al.*, 1994; Chang and de Jong, 2005). In a combined chromosome painting with GISH followed by FISH using BAC DNA as probes, Dong and coworkers (2001) even revealed genetic identity of alien chromosomes and segments in potato breeding lines. Such strategies, however, are not sufficient for unravelling complex rearrangements and identification of chromosome breakpoints at nucleotide accuracy. To this end we introduced the combined usage of BAC FISH, genetic markers and reference genome sequence information to elucidate the complex topology of tomato and potato chromosomes as well as detection of introgressed regions (Tang *et al.*, 2008; Peters *et al.*, 2009; Peters *et al.*, 2012; Aflitos *et al.*, 2014).

Other methods for introgression detection include Restriction site Associated DNA (RAD) and Genotyping by Sequencing (GBS), which both generate restriction fragments that can be subsequently sequenced for posterior SNP calling (Baird *et al.*, 2008). GBS allows high-throughput detection of thousands of SNPs along the genome, producing a polymorphism density several orders of magnitude higher than genetic marker based technologies, but is also two orders of magnitude less sensitive than next-generation sequencing (NGS). Furthermore, the resolution depends on the occurrence of restriction sites, which are not evenly distributed along chromosomes. Due to the dependency on enzymatic activity, there is a high rate of non-calls and relatively low reproducibility that makes GBS less suitable for introgression detection (Galvão *et al.*, 2012).

The examination of Genome-wide Single Nucleotide Polymorphisms (SNPs) is an alternative strategy that becomes increasingly attractive for disclosing genome organization and topological context of target genes underlying agronomically important traits. Several methods to visualize such whole-genome SNP (wgSNP) data have been presented (Posada, 2002; Martin *et al.*, 2011; Lechat *et al.*, 2013; Kim *et al.*, 2014) including software packages such as VisRD (Strimmer *et al.*, 2003), SHOREmap (Schneeberger *et al.*, 2009), ReCombine (Anderson *et al.*, 2011), NGM (Austin *et al.*, 2011), PartFinder (Prasad *et al.*, 2013) and PhyloNet-HMM (Liu *et al.*, 2014). SHOREmap and NGM focus on identifying causal SNPs for phenotypes in segregating populations. VisRD, PartFinder and PhyloNet-HMM are based on a sliding window approach by mapping different genomes

to a reference genome and searching for inconsistent phylogenetic relationships that are used as markers for introgressions. However, VisRD and ReCombine only handle relatively small genomes, whereas SHOREmap, NGL and PhyloNet-HMM are targeting inbred lines. PartFinder analyses several large genomes and creates global phylogenetic trees, however, it does not identify introgressed segments. In this paper we present the Introgression Browser (iBrowser), a tool that delineates introgressed segments, identifies donor parents, and is able to handle a large number of genomes with virtually no genome size limitation. We have separated the computationally intensive database calculations from the visualization and use open standards, allowing the reuse of the data and data exchange with other supporting programs. iBrowser fills the technological gap between high-throughput sequencing and sequence-based introgression detection, and is applicable for large or industrial scale introgression hybridization breeding programs.

Experimental procedures

iBrowser consists of a back-end which calculates and stores SNPs in a database and a front-end that enables the visualization of SNPs. iBrowser takes SNP data from any variant calling algorithm provided in VCF format, and can import distance matrix information from multiple programs such as FastTree (Price *et al.*, 2010) or SnpPhylo (Lee *et al.*, 2014). This provides the user flexibility to operate and test combinations of parameters for proper visualization. iBrowser is able to efficiently process and filter data with excessive noise, although we do advise the use of high quality SNP calls and repeat masking. The set of scripts and programs necessary to run the iBrowser can either be downloaded as a pre-configured virtualized disk bundle or can be installed directly in a pre-existing system. A list of the accessions used in this study with supporting information, can be found in supplementary tables 1 to 3.

Data Generation

To visualize introgressions in wild Solanum species and crop tomatoes, we created three datasets “84-10k”, “84-50k” and “84-Genes” from the genome sequence information produced for 84 tomato accessions (Aflitos *et al.*, 2014). The 84-10k and 84-50k datasets contain consecutive segments of 10 kbp and 50 Kbp, respectively. The 84-Genes dataset consists of segments generated from the *S. lycopersicum* cv. Heinz v2.40 using gene coordinates from the v2.31 annotation (Tomato Genome Consortium, 2012). Approximately 5 hours using a 20 core Intel(R) Xeon(R) CPU E7- 4840 @ 2.00GHz was required to generate each set. The visualization takes up to 20 seconds for the largest dataset.

A dataset, hereafter referred as “60-IL” dataset, was created consisting of tomato, *Solanum lycopersicum* cv. MoneyMaker LYC 1365, *S. pimpinellifollium* CGN14498, 60 offspring F6 Inbred Line (IL) individuals, and four *S. pimpinellifolium* related accessions (LYC2798, LYC2740, LA1584 and LA1578). The tomato MoneyMaker parent, 60 ILs and

the pimpinellifolium related accessions were obtained from Aflitos *et al.*, (2014). The *S. pimpinellifollium* CGN14498 parent was sequenced using a 500 bp paired-end library with the Illumina HiSeq 2000 platform at 38 folds coverage (assuming a genome size of 950 Mbp). The 60 F6 IL individuals were likewise sequenced at 4.80±1.70 fold coverage. *S. pimpinellifollium* CGN14498 and the 60 IL individuals were mapped against *S. lycopersicum* cv. Heinz 1706 v2.4 as described in (Aflitos *et al.*, 2014) using Bwa (Li *et al.*, 2009a) and SNPs were called using SamTools (Li *et al.*, 2009b). The raw data of the 60 RILs and *S. pimpinellifollium* CGN14498 were deposited at The European Bioinformatics Institute (http://www.ebi.ac.uk/) unde identification number PRJEB6659. The resulting VCF file was split into non-overlapping 50 Kbp fragments.

The back-end

The back-end consists of a database creation tool that takes Variant Call Format (VCF) files as input generated from samples that are mapped to a reference genome, and a FASTA file containing the reference genome sequence. It converts VCF and FASTA data into an intermediate file format containing the genotype for each coordinate per sample. Alternatively, this can also be achieved by using snp-search (Al-Shahib and Underwood , 2013). Coordinates are excluded if any of the following constraints are not satisfied; (i) at least two individuals show polymorphism; (ii) any individual shows heterozygosity; or (iii) any individual contains an InDel, a protocol modified from Lee and coworkers (2014). We filter out InDels because they are difficult to align and score (Anderson *et al.*, 2011). The reason is that they are generally not fixed in the population and are usually of little value for marker-assisted breeding. Furthermore, they cannot be expressed in canonical FASTA format and most phylogeny programs do not accept IUPAC codes for ambiguous nucleotides. The filtering of polymorphisms usually reduces the input data size up to 30-40%.

The remaining coordinates are then ordered and connected head to tail into consecutive fragments. Alternatively, iBrowser can operate in a sliding window mode where each fragment overlaps with the previous. The latter operating mode allows for more accurate introgression boundary detection. However, it will generate significantly more data compared to the consecutive operating mode, and the overlapping mode therefore should be used exclusively for a limited number of small segments.

Any General Feature Format (GFF) file containing segmentation information can be used, such as coordinates of CDSs, genes, exons or QTLs. If no particular segmentation pattern is specified, the reference genome FASTA file may be used to create a GFF file containing a user-defined, evenly spaced, segmentation pattern (Huang *et al.*, 2009; Anderson *et al.*, 2011; Prasad *et al.*, 2013; Chen *et al.*, 2014; Kim *et al.*, 2014, Liu *et al.*, 2014). The resulting GFF is then used to split the VCF files into the desired segments, as consecutive windows, sliding windows, or any arbitrary filtering from the annotation file of the ref-

erence genome such as CDS, genes, exons, introns, etc. Finally, for each segment a column is produced containing the polymorphism count for each sample.

All SNPs for each segment are then concatenated per species and stored as FASTA files. Next, a Jukes-Cantor distance matrix and a Newick formatted Maximum Likelihood phylogenetic Bio-Neighbor Joining tree are constructed per coordinate using FastTree2 (Price *et al.*, 2010). The distance matrices and Newick formatted phylogenetic trees are then parsed and stored into a portable database. Alternatively, the process of filtering, exporting and processing the VCF files can be performed using SNPphylo (Lee *et al.*, 2014).

In the case of Introgression Lines (ILs) we assume the genotype of an offspring individual originates either from the donor or acceptor parent and, following this paradigm, we therefore assign each SNP to either parent. To achieve this, we first calculate the average number of SNPs per sample; subsequently test whether the number of SNPs is above or below the average and then convert the sequence of the sample to the sequence of the closest parent according to the method of Kim *et al.*, (2014) and Huang *et al.*, (2009). In this way miscalls due to low genomic coverage are repaired, resulting in improved accuracy for delineated introgressed segments and a more reliable donor species identification.

Finally, we store the phylogenetic files (Newick tree and distance matrix) either as a Python memory dump, which allows the information to be stored in RAM memory for fast access and high request loads, or as an SQLite database. The latter takes a few milliseconds longer to serve the results, but has a smaller memory footprint. The database file independently can be copied to any other computer running the front end.

The front-end

Using the project database, the front-end generates the UI in the web browser. The required Python web server can run in any OS, which is able to run Python. It can be accessed locally through the intranet, or via the internet, depending on the configuration of the host computer. As the front-end employs HTML5 technology to display the User Interface (UI), no plugin installation is required. An optional user access control system is available providing security at the login level.

For each fragment the iBrowser UI (figure 1) displays the pairwise distance values between the reference and each query sample as colors in a heat map, composed of lines and columns representing samples and genomic segments respectively. Any sample from a panel of genomes can be selected as the query sample. The first line of the heat map for each segment shows the number of SNPs using a contrasting color scale, giving insight into the distance.

Using the project database, the front-end generates the UI in the web browser. The required Python web server can run in any OS, which is able to run Python. It can be

accessed locally through the intranet, or via the internet, depending on the configuration of the host computer.

The color scale for the pairwise distances as well as the color scale for the number of SNPs is displayed at the top of the graph. The heat map color scheme and row identifiers (samples) can be customized for better discrimination between samples. Clicking in the heat map shows a tooltip box for the selected cell including its statistics on SNP counts and distances. Double clicking in the heat map or selecting a gene/fragment by name in the main menu invokes an overlay panel showing the complete data for the selected column. This includes the respective fragment identifier, coordinate, phylogenetic tree, nucleotide sequence and distance matrix, all of which can be downloaded in their native formats. The full heat map can be downloaded as an PNG/JPEG image. The web server has a JSON API allowing developers to retrieve the data in an interchangeable format, facilitating cross-program usage.

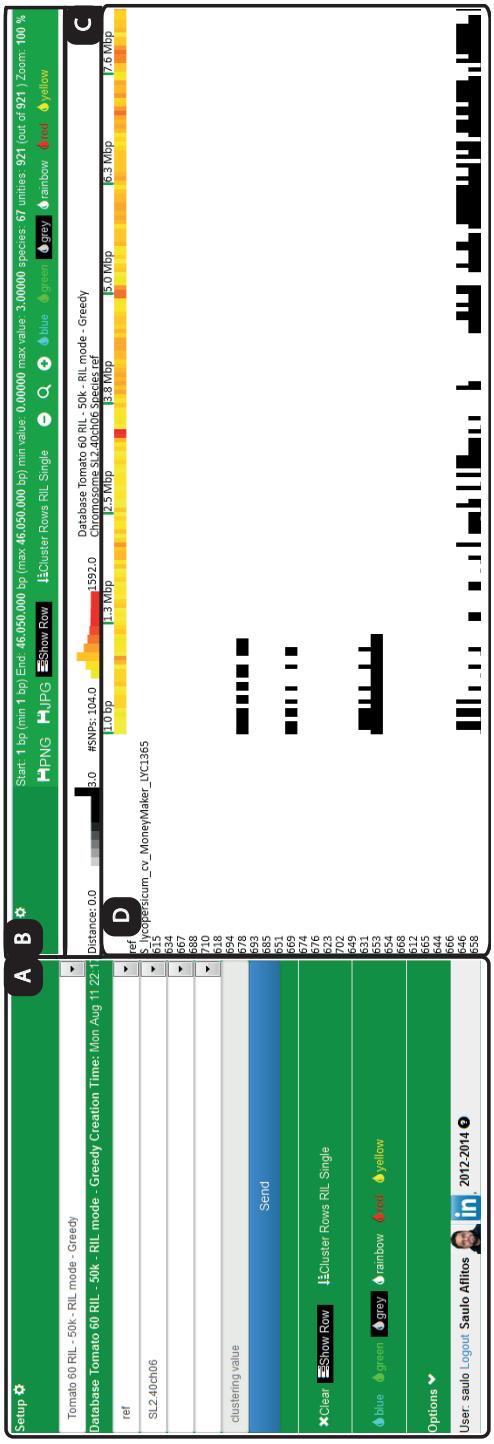


Figure 1: User interface of the IBROWSER. Panel A shows the menu for database, reference species, chromosome, gene and grouping selection. In panel B the first line shows the alignment statistics, and the second line presents interactive buttons for download, toggling of row names, clustering method selection, zoom level and heat map color scheme selection. In panel C the first line displays the color scales used for the phylogenetic distance, the number of SNP per segment, and the information of the selected database and filters. In panel D the heat map is shown in black and white, with the sample names on the left side, a ruler at the top and a yellow-red scale representing the number of SNPs in each segment.

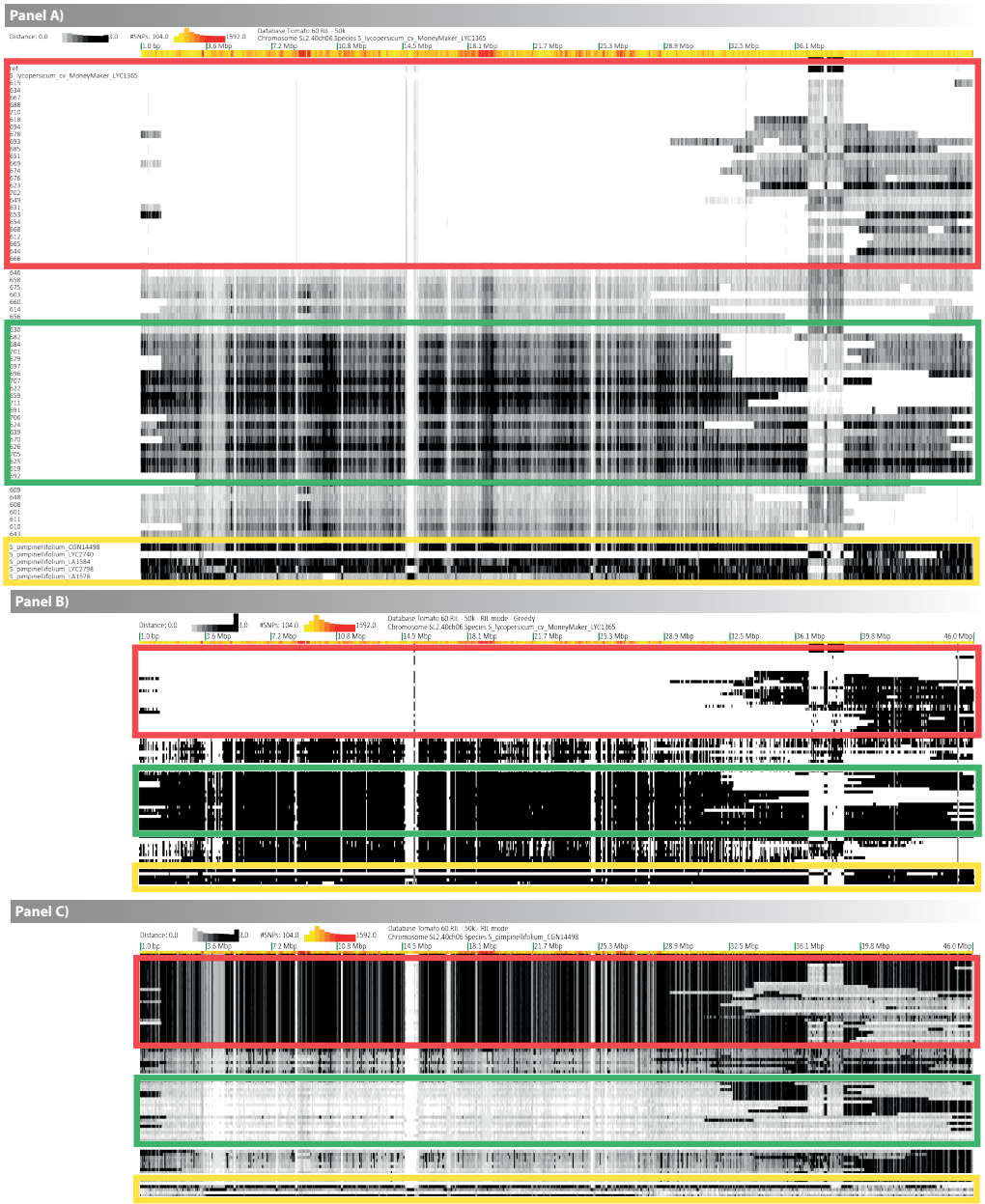
Results

iBrowser shows homeologous recombination sites in introgression lines

To analyse the chromosome structure of the 60 introgression lines (ILs) we aligned the sequence reads to both the *S. lycopersicum* Moneymaker and the *S. pimpinellifolium* CGN14498 parent and display the sequence distance to either parent in a heat map. Figure 2 displays the tomato chromosome 6 heat map for all ILs. Crossover sites denote positions where chromosome origin changes from the donor (CGN14498) to the acceptor parent (MoneyMaker). The change of chromosome origin is displayed by a color change in the heat map. Clearly, about half of the individuals hardly differ from the MoneyMaker parent across most of the chromosome (figure 2, red boxes), and hence have a high sequence similarity almost without any introgression from *S. pimpinellifolium*. One other quarter shows a high distance (figure 2, green boxes) to MoneyMaker; the SNP pattern suggests that almost the entire chromosome 6 was replaced by its homeologous part of *S. pimpinellifolium*. The heat map using the *S. pimpinellifolium* parent as a reference (figure 2, panel C) confirms that regions with a high distance to MoneyMaker are identical to the *S. pimpinellifolium* parent CGN14498, generating an almost inverted image compared to the heat map of figure 2, panel A. These results illustrate the capability of iBrowser to delineate the size of introgressions.

We also determined the sequence distance of additional *S. pimpinellifolium* accessions to assess their relationship with the CGN14498 and MoneyMaker parent. The sequence distance between the parent *S. pimpinellifolium* CGN14498 and its *S. pimpinellifolium* relatives (LYC2798 and LA1584 is very low, with LYC2798 having the closest distance (figure 2, panel C yellow box). LA1578 shows a high distance in the heatmap for a chromosome 6 segment between coordinates 4 and 24Mbp. The highest distance spanning almost the entire chromosome region was observed for LYC2740 (figure 2, panel C), whereas it shows a low distance to the MoneyMaker parent (figure 2, panel A), which together with the observed distance pattern suggests that LYC2740 is more closely related to the

Figure 2 (right): Heat maps for 60 RIL individuals, RIL parents and non-parental accessions. RILs have been generated from a cross between *S. lycopersicum* cv. MoneyMaker LA2706 (LYC1365)and *S. pimpinellifolium* CGN14498. The RILs, RIL parents, and non-parental *S. pimpinellifolium* accessions have been mapped against *S. lycopersicum* cv. Heinz 1706 chromosome 6 (x-axis) labeled "ref". Labels displayed at the left y-axis in panel A correspond to accession identifiers and RIL numbers. RILs are indicated by a 3 digit number. Panel A to C have the same order of RILs and accessions. Distances in the heatmaps are displayed per 50 kbp window size. Panel A displays distances to *S. lycopersicum* cv. MoneyMaker LA2706, panel B displays the same data as shown in panel A (though vertically compressed) using a RIL-specific data filter, and panel C displays the distances to *S. pimpinellifolium* CGN14498. Red and green boxes highlight RILs that have a large introgression from *S. lycopersicum* cv. MoneyMaker or *S. pimpinellifolium* CGN14498, respectively. Distances for additional *S. pimpinellifolium* accessions are displayed in rows highlighted by yellow boxes. Rows without highlighting display intermediate distances for RILs with a low SNP coverage.


Panel A)
Panel B)
Panel C)

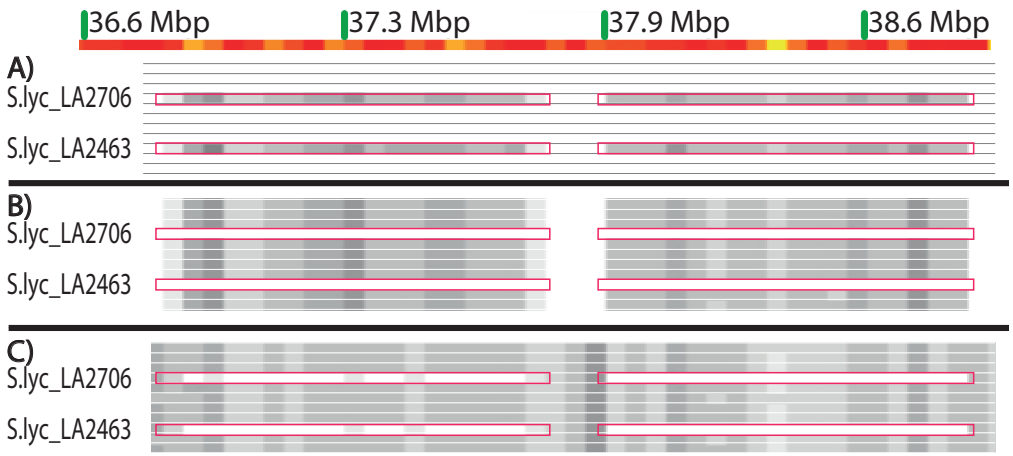
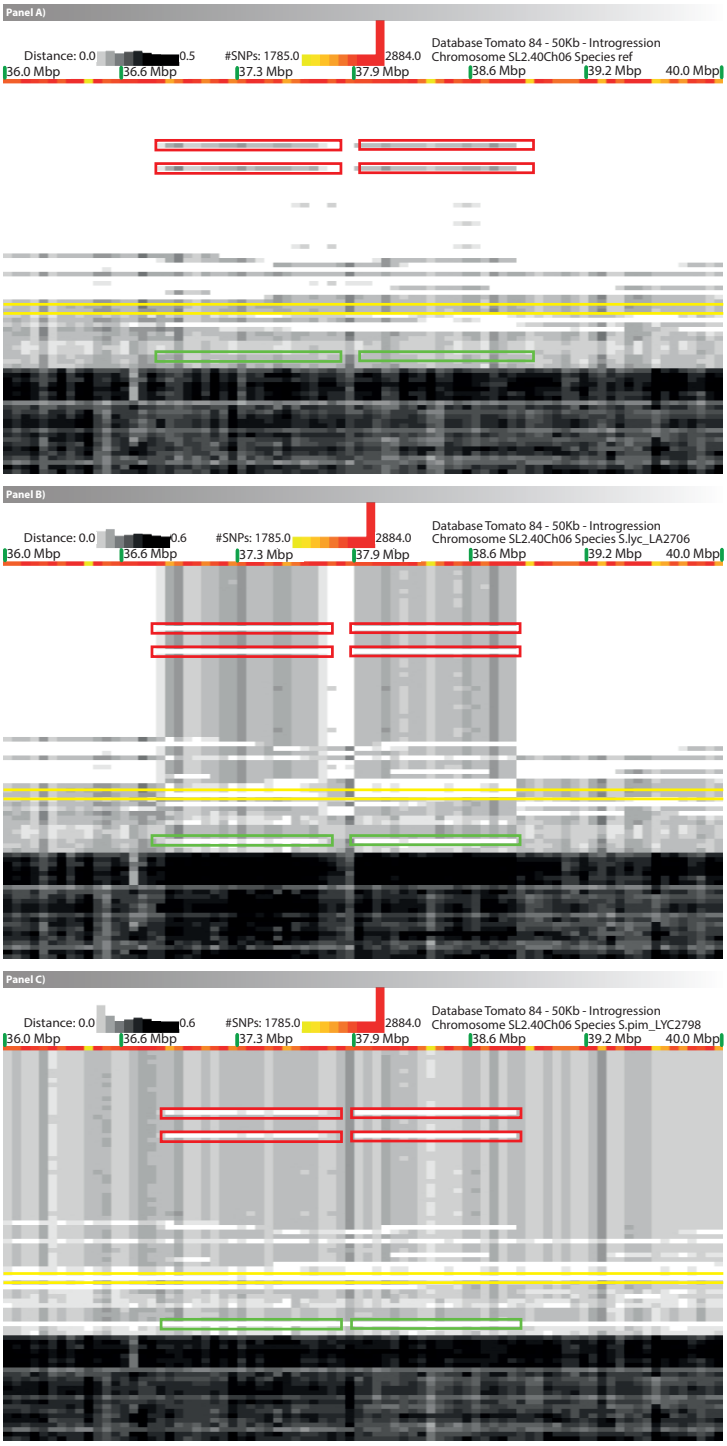


Figure 3: Heat maps for tomato chromosome 6 (v2.40) between coordinates 36 Mbp and 40 Mpb, using *S. lycopersicum* cv. Heinz 1706 (panel A), *S. lycopersicum* cv. MoneyMaker LA2706 (panel B) and *S. pimpinellifolium* LYC2798 (panel C) as a reference, respectively. *S. lycopersicum* accession IDs are indicated at the left side in each panel. Introgressed segments in LA2706 (MoneyMaker) and LA2463 (AllRound) are highlighted by red boxes. A full length chromosome 6 heatmap for all accessions is shown in figures 4 and 5.

MoneyMaker parent. Indeed, based on our previous phylogenetic and whole genome SNP analyses (Aflitos *et al.*, 2014), and the sequence distance presented here, we propose accession LYC2740 should be considered as a *S. lycopersicum* species rather than a *S. pimpinellifolium* species. Although here breeding parents for the RIL individuals are known, in general a heatmap to a single reference is not sufficient to identify the probable source of an introgression from a panel of closely related species. In addition this requires comparison of heat map distance patterns using also the other *S. pimpinellifolium* species as a reference, or alternatively by a phylogenetic analysis (see below).

The remaining samples show intermediate distances to MoneyMaker (outside the boxes). Most likely this is caused by the extremely low coverage of sequencing (on average 6x), yielding insufficient evidence for SNP calling. Normally, whenever lack of coverage occurs, the reference sequence is assumed. Because such cases hamper the detection of introgressions (Canady *et al.*, 2006; Anderson *et al.*, 2011), we included a filter specifically designed to handle ILs with low coverage sequencing. It assumes that the genotype

Figure 4 (right): Heat maps for tomato chromosome 6 (v2.40) between coordinates 36 Mbp and 40 Mpb, using *S. lycopersicum* cv. Heinz 1706 (panel A), *S. lycopersicum* cv. MoneyMaker LA2706 (panel B) and *S. pimpinellifolium* LYC2798 (panel C) as a reference, respectively. Rows are in the same order as for figure 5 and 8. The introgressed segments in LA2706 and LA2463 are highlighted in red. The introgression in *S. lycopersicum* CGN15820 is highlighted in yellow. Highlighted in green is the segment from the donor parent *S. pimpinellifolium* LYC2798 that is introgressed into LA2706, LA2463 and CGN15820 as previously has been described in Aflitos *et al.*, (2014).


Panel A)
Distance: 0.0 0.5
#SNPs: 1785.0 2884.0
Database Tomato 84 - 50Kb - Introgression
Chromosome SL2.40Ch06 Species ref
36.0 Mbp
36.6 Mbp
37.3 Mbp
37.9 Mbp
38.6 Mbp
39.2 Mbp
40.0 Mbp
Panel B)
Distance: 0.0 0.6
#SNPs: 1785.0 2884.0
Database Tomato 84 - 50Kb - Introgression
Chromosome SL2.40Ch06 Species S.lyc_LA2706
36.0 Mbp
36.6 Mbp
37.3 Mbp
37.9 Mbp
38.6 Mbp
39.2 Mbp
40.0 Mbp
Panel C)
Distance: 0.0 0.6
#SNPs: 1785.0 2884.0
Database Tomato 84 - 50Kb - Introgression
Chromosome SL2.40Ch06 Species S.pim_LYC2798
36.0 Mbp
36.6 Mbp
37.3 Mbp
37.9 Mbp
38.6 Mbp
39.2 Mbp
40.0 Mbp

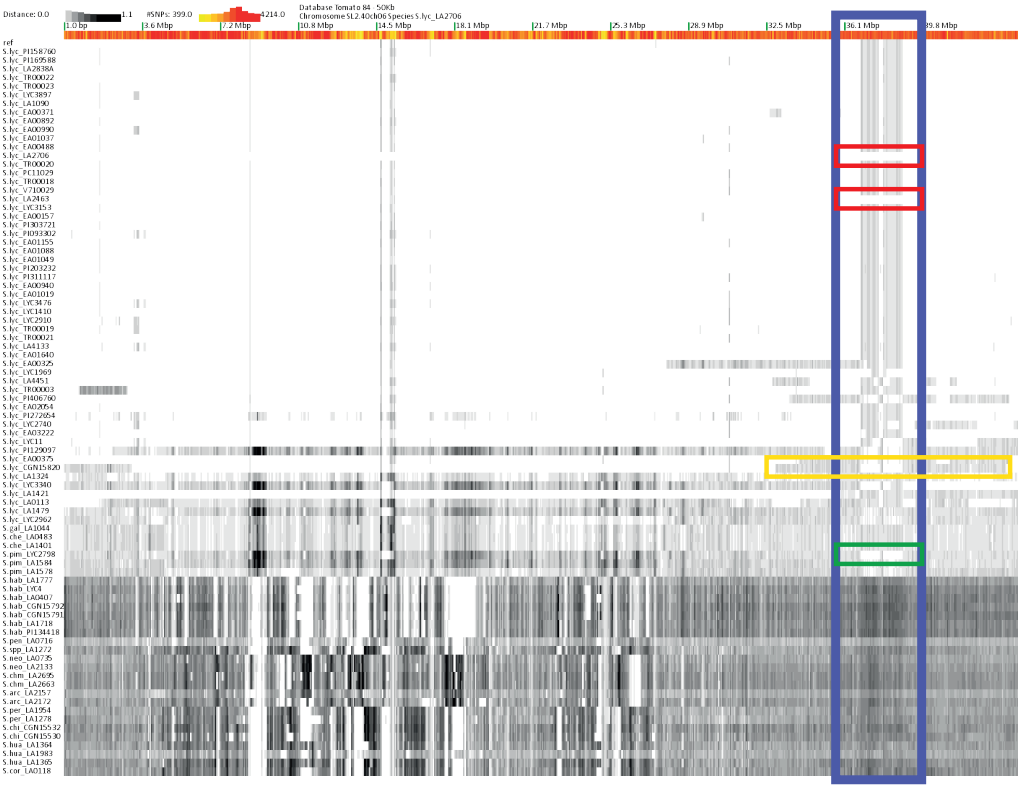


Figure 5: Heat map for tomato chromosome 6 from 84 Solanum accessions using *S. lycopersicum* cv. MoneyMaker LA2706 as reference. Highlighted in blue is the heatmap fragment displayed in figures 3 and 4. Green and red highlight boxes correspond to the donor segment in *S. pimpinellifolium* LYC2798, and introgressed segments in MoneyMaker and *S. lycopersicum* cv. AllRound LA2463, respectively. The yellow highlight box corresponds to an introgression in *S. lycopersicum* CGN15820. Distances in the heatmaps are displayed per 50 kbp window size.

of the offspring is identical to either the recipient or donor parent (see methods). Figure 2, panel B shows the result for the 60-IL dataset after such an IL-specific pre-treatment of the data where each segment was assigned to either parent, without taking intermediate values. To delineate an introgression at base pair level accuracy, the SNP alignment for the segment containing the crossover has to be analyzed in detail in order to identify the crossover site.

IBROWSER reveals positions of alien introgressions

Previously we manually traced and established source identity of introgressions in tomato (*Solanum lycopersicum*) crop accessions from related wild species (Aflitos *et al.*, 2014). Here we used the IBROWSER pipeline to test the automated detection and identification of introgressions. Figures 3, 4 and 5 visualize the introgressions from

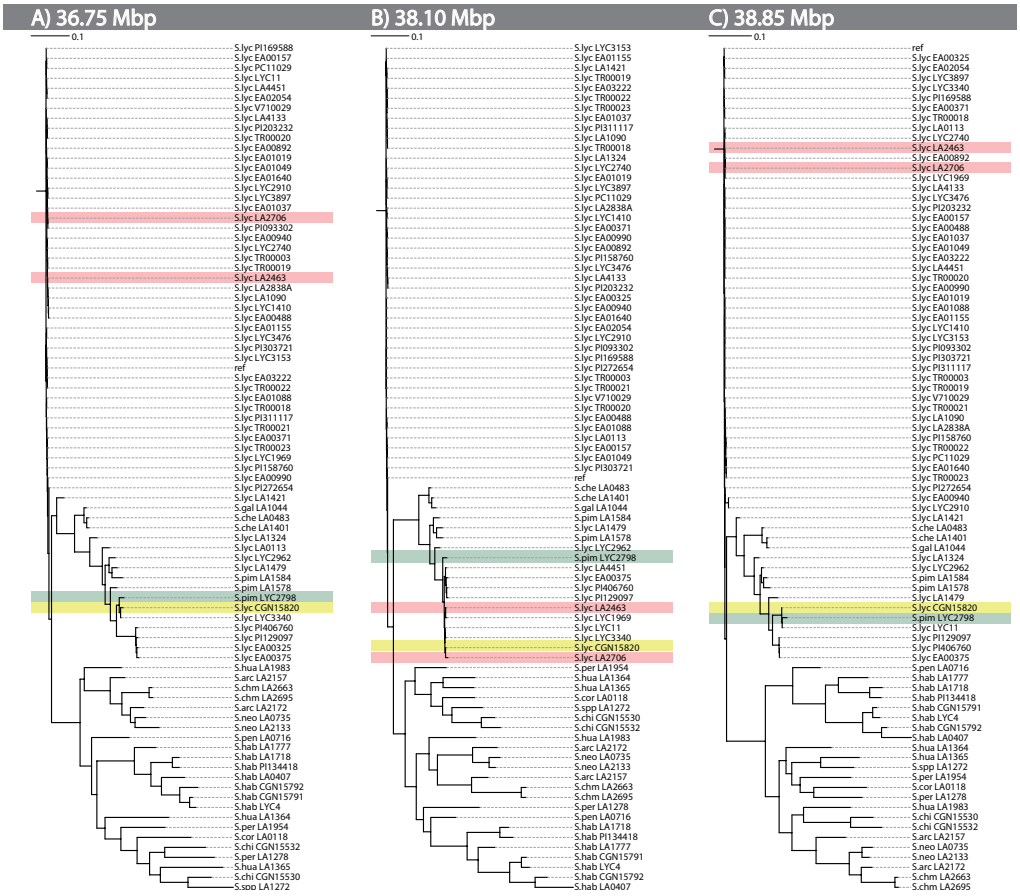


Figure 6: Maximum Likelihood phylogenetic trees for 50 Kbp segments from a chromosome 6 introgression positioned at the left side (panel A), inside (panel B), and at the right side (panel C) of the introgressed segment. Phylogenetic trees for 50 kbp segments correspond to heatmap coordinates indicated at the top of each panel. The phylogenetic tree position of *S. lycopersicum* cv. MoneyMaker LA2706 and *S. lycopersicum* cv. AllRound LA2463 among other *S. lycopersicum* group accessions are highlighted in red and change their position to the closest relative donor species *S. pimpinellifolium* LYC2798 which is highlighted in green. Positions for *S. lycopersicum* CGN15820 are highlighted in yellow. Trees have been drawn using Interactive Tree of Life v2.2.2 (Letunic and Bork, 2011).

S. pimpinellifolium LYC2798 in chromosome 6 of *S. lycopersicum* cv. MoneyMaker LA2706 and *S. lycopersicum* cv. AllRound LA2463. The phylogenetic tree for at position 38.1 Mbp in a 2.0 Mbp sized segment reveals the clustering of both Moneymaker and AllRound with the *S. pimpinellifolium* clade (figure 6, panel B). Furthermore, the vast majority of the 50kb blocks in the 2.0 Mbp segment display a phylogenetic position closest to *S. pimpinellifolium* accession LYC2798, which can be considered the likely source of the introgression. In contrast, Moneymaker and AllRound group with the other

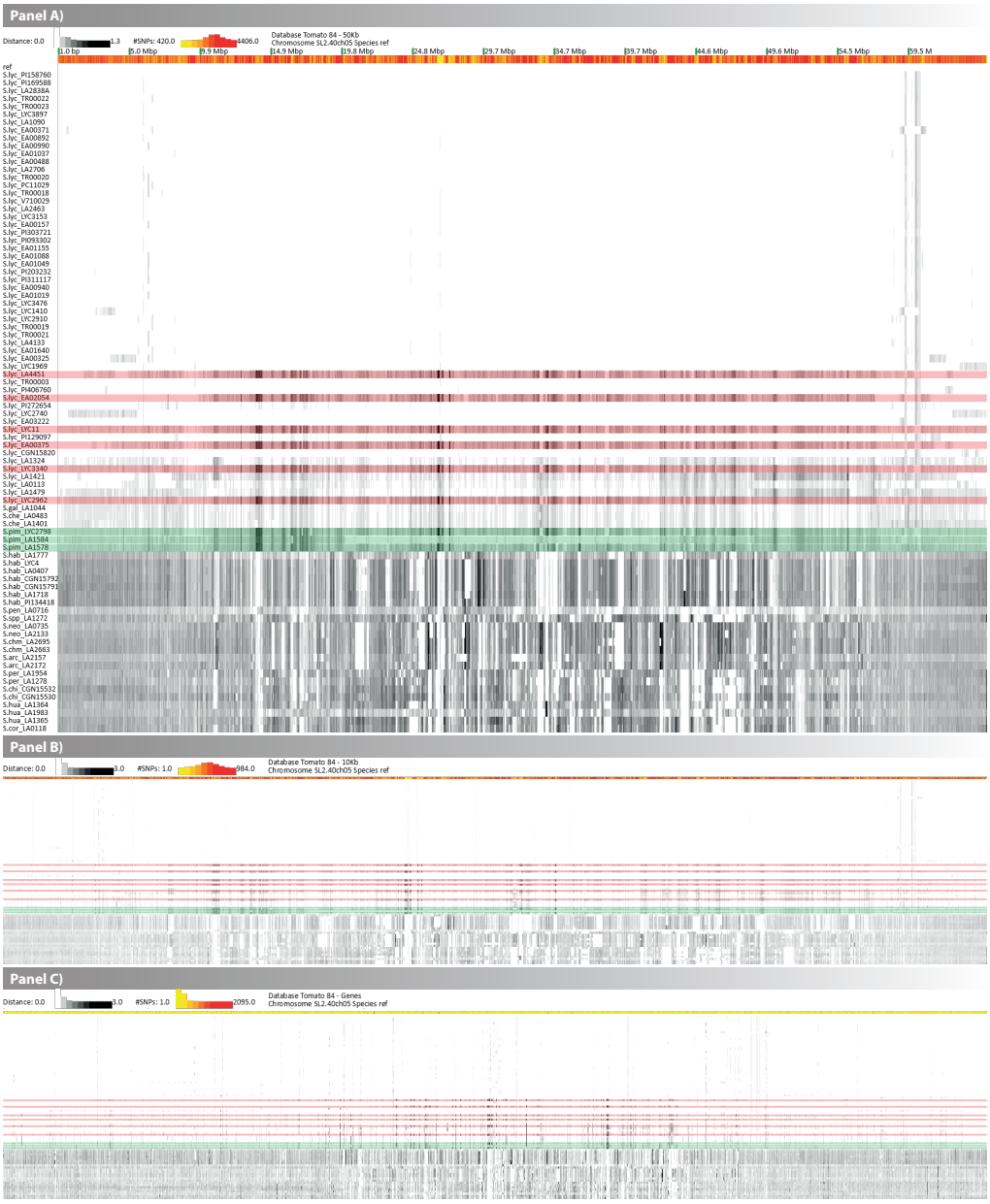


Figure 7: Heat maps for tomato chromosome 5 from 84 Solanum accessions with *S. lycopersicum* cv. Heinz 1706 as reference, using the 84-50kb (panel A), 84-10kb (panel B) and 84-Genes (panel C) datasets, respectively. Labels displayed at the left side in panel A correspond to accession identifiers. Panel A to C have the same order of accessions but panel B and C are vertically compressed. Highlighted in red are cherry tomato varieties while *S. pimpinellifolium* accessions showing similar patterns are highlighted in green.

S. lycopersicum accessions at position 36.75Mbp and 38.85 Mbp that flank the right and left borders of the 2.0 Mbp segment (figure 6). This finding is consistent with our previous SNP analysis and phylogenetic results (Aflitos *et al.*, 2014). Also other introgressions events can be detected. For example *S. lycopersicum* CGN15820 shows a 10Mbp segment located between chromosome 6 heatmap position 33Mbp to 43Mbp, spanning the 2.0Mbp introgressed segment in MoneyMaker and AllRound (figures 4 and 5, yellow boxes). The clustering suggests *S. pimpinellifolium* LYC2798 as the likely donor source and extends to position 36.75 and 38.85 Mbp (figure 6, panel A to C) and probably beyond, suggesting that a larger segment was introduced into CGN15820 by another introgression event. These results illustrate the capability of iBrowser to identify the most likely source of an introgression from a panel of very closely related species.

The average gene size in tomato is approximately 3.7kb (Aoki *et al.*, 2010), and the number of SNPs in genes between tomato cultivars is generally smaller than that in 10 kbp or 50 kbp windows. The lower SNP content from genic sequences obtained by exome capture nevertheless has been shown to contain sufficient information for solid phylogenetic analysis (Austin *et al.*, 2011; Galvão *et al.*, 2012). In addition, our previous findings indicate that SNPs in coding sequences of Solanum accessions are under higher selective pressure than SNPs in non-coding regions and contain more phylogenetic information than non-coding SNPs (Aflitos *et al.*, 2014). Figure 7 shows the analysis on the 84-10k, 84-50k and 84-Genes datasets for chromosome 5. The heatmaps display regions of conservation and introgressions from 50kbp to gene level resolution (figure 7, panel A to C) that are shared exclusively between cherry varieties and *S. pimpinellifolium* accessions. These results show that iBrowser is able to detect accession specific introgressions that can be used for pedigree analysis.

iBrowser for revealing aberrant SNP landscapes in and around inversions

The short arm of *Arabidopsis thaliana* chromosome 4 in the Col-0 and few other accessions contains a 1.17 Mb paracentric inversion (Fransz *et al.*, 2000; Fransz *et al.*, manuscript in prep). This chromosomal rearrangement coincides with almost a complete absence of SNPs in Col-0 and other accessions that have the inversion. In contrast, syntenic segments in *Landsberg erecta* (Ler) and most other accessions that do not have the inversion, contain a large number of SNPs. This is shown by sequence analysis of 597 *Arabidopsis thaliana* accessions from the 1001 Arabidopsis database (Cao *et al.*, 2011; Schmitz *et al.*, 2013; Wang *et al.*, 2013a) and clearly illustrated by the heatmap of this segment in iBrowser (Figures 8 and 9). Accession hybrids heterozygous for the inversion will not produce viable gametes with recombinations in the inverted region and do not display recombinants in the genetic map (Drouaud *et al.*, 2006; Wang *et al.*, 2013b). Apparently, the genetic rearrangement in the top arm of chromosome 4 causes a reduced cross-over frequency for the inversion and its flanking regions. Our SNP detection results are consistent with these previous observations and illustrate that the parallel vis-

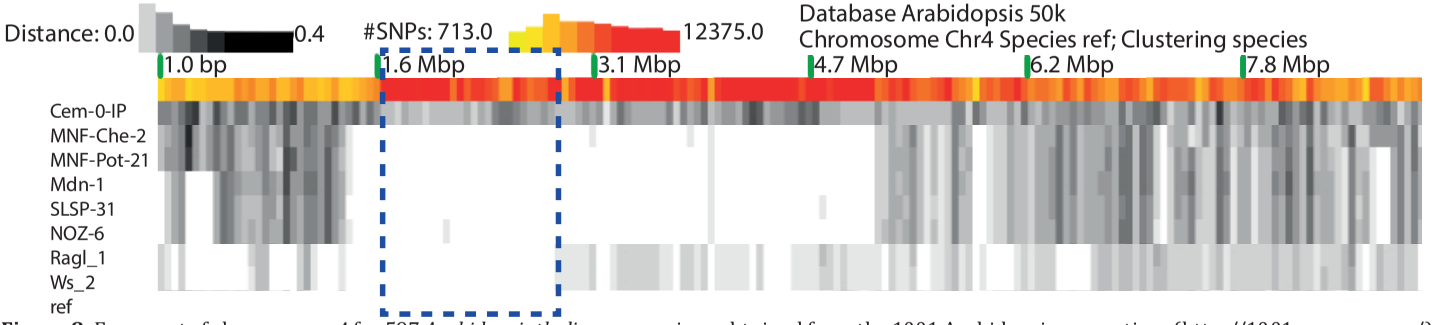


Figure 8: Fragment of chromosome 4 for 597 *Arabidopsis thaliana* accessions obtained from the 1001 Arabidopsis consortium (http://1001genomes.org/). White blocks show reduction in the number of SNPs compared to accession Col-0 TAIR v10 that was used as a reference. White blocks highlighted by a blue dashed block coincide with an inverted segment implicated in absence of recombination. Chromosome 4 positions downstream from the inversion (right side of the highlighted region) extend into the pericentromere and centromere regions. The few SNPs that are still present may have occurred from mutation events. A full length chromosome 4 heatmap for all accessions is shown in figure 9.

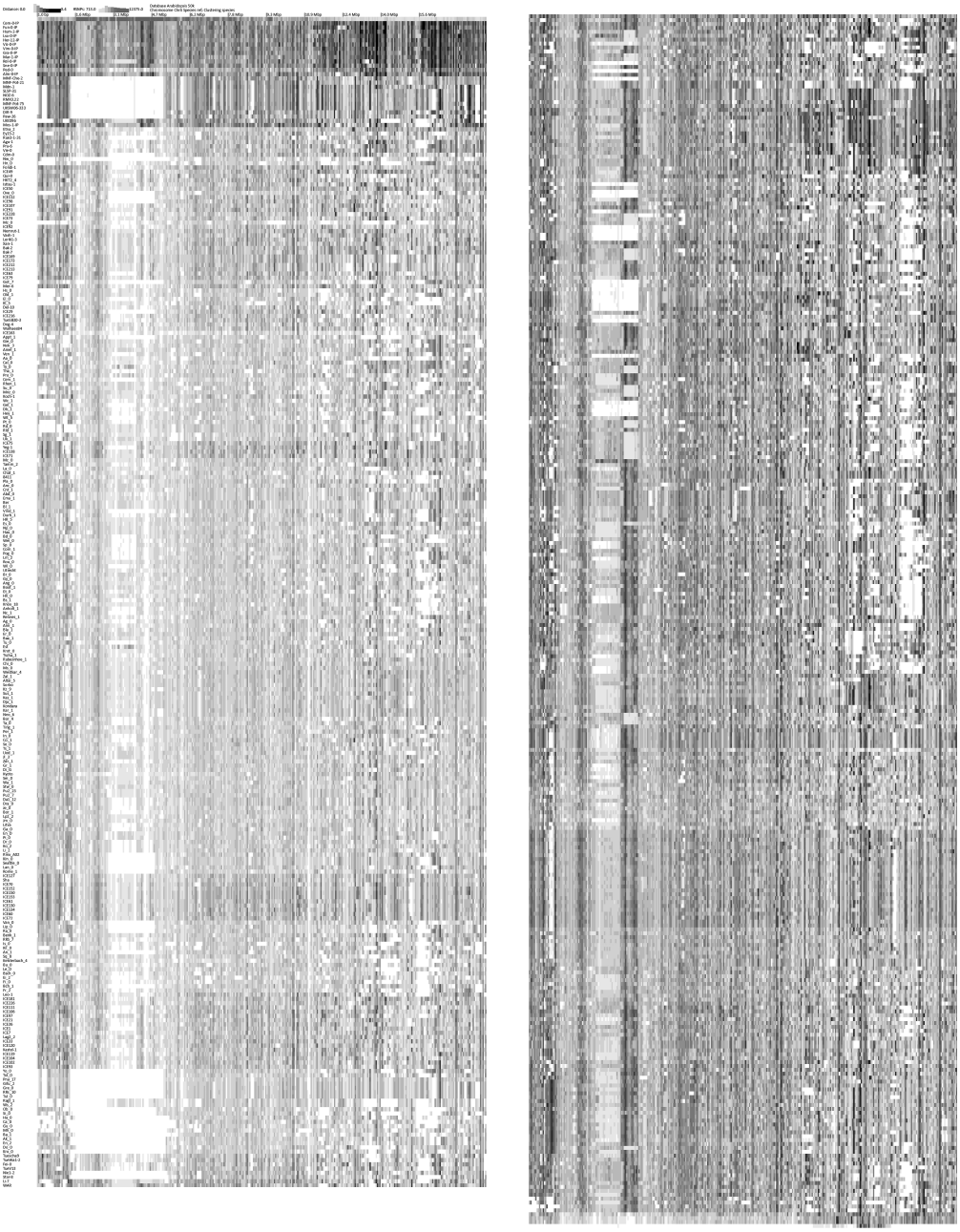


Figure 9: Heat map of chromosome 4 from 597 *Arabidopsis thaliana* accessions, obtained from the 1001 Arabidopsis consortium (http://1001genomes.org/), mapped against Col-0 TAIR v10 (ref) using a 50 Kbp window size.

ualization of introgression patterns along chromosomes in IBROWSER may help breeders to identify regions which are problematic for introgression hybridization breeding.

Discussion

SNPs are becoming increasingly important in establishing source identity of introgressions (Huang *et al.*, 2009; Chen *et al.*, 2014; Paraskevis *et al.*, 2005; Anderson *et al.*, 2011; Austin *et al.*, 2011; Víquez-Zamora *et al.*, 2013; Prasad *et al.*, 2013). Plant genomes, also those of crop species, contain massive amounts of SNPs in their gene pools. Our study on genetic variation shows the number of SNPs in tomato (*S. lycopersicum*) and related wild species can exceed 10 million (Aflitos *et al.*, 2014). Previously, we used genome-wide SNP information and phylogenetic analyses to detect the source of introgressed segments in crop tomato. However, open source software tools that integrate SNP detection, visualization and phylogenetic analysis in a single workflow aiming at delineation and source identification of introgressed segments on a genome wide scale were lacking. Here, we have developed IBROWSER that overcomes this challenge. The software requires only a small memory and disk footprint, permitting it to run on any modern computer platform, while its databases can be distributed easily. It is also highly portable and can be used in stand-alone mode or over a network connection to facilitate project sharing. IBROWSER can be accessed through its JavaScript Object Notation (JSON) Application Programming Interface (API), allowing it to combine with programs such as MUMMERPLOT (Kurtz *et al.*, 2004) and (G/J)BROWSE (Stein *et al.*, 2002; Skinner *et al.*, 2009). This opens the possibility to integrate additional information on chromosome structure, which can help to assist in breeding parent selection. The software is freely distributed at www.ab.wur.nl/ibrowser. Processed data used in this study can be provided upon request.

We implemented a computational approach to process sequence information from a large panel of genomes and retrieve information on the location, size, and source of introgressed segments. We show that IBROWSER is capable of detecting characteristic introgressions in cherry tomato accessions. In addition, by including the ITAG 2.31 annotation data for tomato into IBROWSER, the genic content of introgressed segments can be retrieved, which may help to find additional leads on gene traits relationships. Furthermore, this bioinformatics tool identified the most likely donor of a 2.0 Mbp introgressed segment into the *S. lycopersicum* cv. Moneymaker recipient from a panel of closely related *S. pimpinellifolium* accessions. This functionality makes IBROWSER an applicable tool for pedigree analysis.

We also used IBROWSER to analyze a large panel of Introgression Lines. By applying a specific filter we removed erroneous SNP calls that cannot be assigned to either breeding parent. The filtering improves the distinctness and donor source identification of introgressed segments even in genomes that have been sequenced at 4X coverage. This ca-

pability provides an alternative perspective for molecular breeding and allows studying genome features involved in crossover recombination in more detail. The information on size and sequence of introgressed segments, and frequency of crossover recombination that can be inferred from the SNP distribution, may be used to search for signatures that can either promote or prevent crossover recombination. Such features might also include information on genome structure either promoting or preventing crossover recombination. Using a comparative analysis on a panel of 597 Arabidopsis accessions, we could pinpoint a 1.17 Mb inverted segment in the top arm of chromosome 4S that has been implicated in absence of crossover recombination in inversion heterozygotes (Fransz *et al.*, 2000). In addition to this large inversion, IBrowser revealed many other segments that were devoid of SNPs. It will be interesting to investigate the relationship between the SNP distribution and topology of these segments. These results illustrate the ability of IBrowser as a tool to assist in comparative genomics studies.

Acknowledgements

We thank Suzanne Hoogstrate and Luca Santuari for comments in the manuscript; this research was supported by the Centre for BioSystems Genomics (CBSG).

Supplementary Materials

Supplementary Table 1: List of species and RIL identifiers. The order of accession names and color identifiers in the list correspond to the order of the accessions and highlight color respectively as is displayed in figure 2.

Species Name/ID	Color Used in Figure 1	Classification in Figure 1
ref	Red	S. lycopersivum cv. Heinz (Reference)
S lycopersicum cv MoneyMaker LYC1365	Red	MoneyMaker Parent
615	Red	Mostly MoneyMaker
634	Red	Mostly MoneyMaker
667	Red	Mostly MoneyMaker
688	Red	Mostly MoneyMaker
710	Red	Mostly MoneyMaker
618	Red	Mostly MoneyMaker
694	Red	Mostly MoneyMaker
678	Red	Mostly MoneyMaker
693	Red	Mostly MoneyMaker
685	Red	Mostly MoneyMaker
651	Red	Mostly MoneyMaker
669	Red	Mostly MoneyMaker
674	Red	Mostly MoneyMaker
676	Red	Mostly MoneyMaker
623	Red	Mostly MoneyMaker
702	Red	Mostly MoneyMaker
649	Red	Mostly MoneyMaker
631	Red	Mostly MoneyMaker
653	Red	Mostly MoneyMaker
654	Red	Mostly MoneyMaker
668	Red	Mostly MoneyMaker
612	Red	Mostly MoneyMaker
665	Red	Mostly MoneyMaker
644	Red	Mostly MoneyMaker
666		
646		
658		
675		
603		
660		
614		
656		
630		
682	Green	Mostly Pimpinnllifolium
684	Green	Mostly Pimpinnllifolium
701	Green	Mostly Pimpinnllifolium
679	Green	Mostly Pimpinnllifolium
697	Green	Mostly Pimpinnllifolium
696	Green	Mostly Pimpinnllifolium
707	Green	Mostly Pimpinnllifolium
622	Green	Mostly Pimpinnllifolium
659	Green	Mostly Pimpinnllifolium
711	Green	Mostly Pimpinnllifolium
691	Green	Mostly Pimpinnllifolium
706	Green	Mostly Pimpinnllifolium
624	Green	Mostly Pimpinnllifolium
639	Green	Mostly Pimpinnllifolium
670	Green	Mostly Pimpinnllifolium
626	Green	Mostly Pimpinnllifolium
705	Green	Mostly Pimpinnllifolium
625	Green	Mostly Pimpinnllifolium
619	Green	Mostly Pimpinnllifolium
692		
609		
648		
608		
601		
611		
610		
643		
S pimpinellifolium CGN14498	Yellow	Pimpinellifolium Parent
S pimpinellifolium LYC2740	Yellow	Pimpinellifolium Relative
S pimpinellifolium LA1584	Yellow	Pimpinellifolium Relative
S pimpinellifolium LYC2798	Yellow	Pimpinellifolium Relative
S pimpinellifolium LA1578	Yellow	Pimpinellifolium Relative

Supplementary Table 2: List of 84 *Solanum sp.* accessions. The order of accession names and color identifier in the list correspond to the order of the accessions and highlight color respectively as is displayed in figure 4, 5 and 8.

Accession Code	Accession Name	Figures 4 and 5	Figure 7
ref	S. lycopersicum cv Heinz 1706		
S.lyc PI158760	Chih-Mu-Tao-Se (038)		
S.lyc PI169588	PI169588 (041)		
S.lyc LA2838A	Alisa Craig (002)		
S.lyc TR00022	Marmande VFA (094)		
S.lyc TR00023	Thessaloniki (096)		
S.lyc LYC3897	Cross Country (013)		
S.lyc LA1090	Rutgers (004)		
S.lyc EA00371	John_s big orange (008)		
S.lyc EA00892	Belmonte (033)		
S.lyc EA00990	Jersey Devil (024)		
S.lyc EA01037	Giant Belgium (091)		
S.lyc EA00488	Taxi (006)		
S.lyc LA2706	Moneymaker (001)	RED	
S.lyc TR00020	Dixy Golden Giant (090)		
S.lyc PC11029	Winter Tipe (031)		
S.lyc TR00018	Large Red Cherry (077)		
S.lyc V710029	Anto (030)		
S.lyc LA2463	All Round (011)	RED	
S.lyc LYC3153	Lycopersicon esculentum 825 (020)		
S.lyc EA00157	Polish Joe (026)		
S.lyc PI303721	The Dutchman (028)		
S.lyc PI093302	Chag Li Lycopersicon esculentum (032)		
S.lyc EA01155	Dana (018)		
S.lyc EA01088	Tiffen Mennonite (034)		
S.lyc EA01049	Large Pink (019)		
S.lyc PI203232	Wheatley_s Frost Resistant (035)		
S.lyc PI311117	PI311117 (036)		
S.lyc EA00940	Porter (078)		
S.lyc EA01019	Brandywine (089)		
S.lyc LYC3476	Lidi (014)		
S.lyc LYC1410	Heinz conver infiniens var pluriloculare (040)		
S.lyc LYC2910	S pimpinellifolium unc (043)		
S.lyc TR00019	Bloodt Butcher (088)		
S.lyc TR00021	Kentucky Beefsteak (093)		
S.lyc LA4133	TR00026 (102)		
S.lyc EA01640	Watermelon Beefsteak (097)		
S.lyc EA00325	Galina (005)		
S.lyc LYC1969	Sonato (012)		
S.lyc LA4451	Black Cherry (029)		RED
S.lyc TR00003	Momatero (015)		
S.lyc PI406760	Gardeners Delight (003)		
S.lyc EA02054	Cal J TM VF (027)		RED
S.lyc PI272654	PI272654 (023)		
S.lyc LYC2740	S pimpinellifolium unc (045)		
S.lyc EA03222	Lycopersicon esculentum 828 (021)		
S.lyc LYC11	Trote Beere (016)		RED

S.lyc PI129097	PI129097 (022)		
S.lyc EA00375	Katinka Cherry (007)		RED
S.lyc CGN15820	S lycopersicum (054)	YELLOW	
S.lyc LA1324	PI365925 (037)		
S.lyc LYC3340	T1039 (017)		RED
S.lyc LA1421	TR00027 (103)		
S.lyc LA0113	LA0113 (039)		
S.lyc LA1479	TR00028_LA1479 (105)		
S.lyc LYC2962	Lycopersicon sp (042)		RED
S.gal LA1044	S galapagense (104)		
S.che LA0483	S cheesemaniae (053)		
S.che LA1401	S cheesemaniae (055)		
S.pim LYC2798	S pimpinellifolium (044)	GREEN	GREEN
S.pim LA1584	S pimpinellifolium (046)		GREEN
S.pim LA1578	S pimpinellifolium (047)		GREEN
S.hab LA1777	S habrochaites (070)		
S.hab LYC4	S habrochaites (072)		
S.hab LA0407	S habrochaites (071)		
S.hab CGN15792	S habrochaites glabratum (068)		
S.hab CGN15791	S habrochaites glabratum (066)		
S.hab LA1718	S habrochaites glabratum (069)		
S.hab PI134418	S habrochaites glabratum (067)		
S.pen LA0716	S pennellii (074)		
S.spp LA1272	S pennellii (073)		
S.neo LA0735	S neorickii (057)		
S.neo LA2133	S neorickii (056)		
S.chm LA2695	S chiemliewskii (052)		
S.chm LA2663	S chiemliewskii (051)		
S.arc LA2157	S arcanum (058)		
S.arc LA2172	S arcanum (059)		
S.per LA1954	S peruvianum (060)		
S.per LA1278	S peruvianum new (049)		
S.chi CGN15532	S chilense (064)		
S.chi CGN15530	S chilense (065)		
S.hua LA1364	S arcanum new (075)		
S.hua LA1983	S huaylasense (062)		
S.hua LA1365	S huaylasense (063)		
S.cor LA0118	Lycopersicon sp (025)		

Supplementary Table 3: List of 597 Arabidopsis accessions. The order of accession names correspond to the order of the accessions as is displayed in figure 9. The identifiers for accessions in figure 7 are indicated in right most row of the list

Figure 9 Order	File Name	Presence Figure 8
Cem-0-IP	MPICWang2013_quality_variant_vcf_9533_TAIR10	Y
Fun-0-IP	MPICWang2013_quality_variant_vcf_9542_TAIR10	
Hum-2-IP	MPICWang2013_quality_variant_vcf_9549_TAIR10	
Lso-0-IP	MPICWang2013_quality_variant_vcf_9554_TAIR10	
Her-12-IP	MPICWang2013_quality_variant_vcf_9545_TAIR10	
Vis-0-IP	MPICWang2013_quality_variant_vcf_9600_TAIR10	
Vim-0-IP	MPICWang2013_quality_variant_vcf_9598_TAIR10	
Gra-0-IP	MPICWang2013_quality_variant_vcf_9543_TAIR10	
Mar-1-IP	MPICWang2013_quality_variant_vcf_9555_TAIR10	
Rel-0-IP	MPICWang2013_quality_variant_vcf_9574_TAIR10	
Sne-0-IP	MPICWang2013_quality_variant_vcf_9583_TAIR10	
Ped-0	MPICao2010_Ped-0_TAIR10_filtered_variant	
Alm-0-IP	MPICWang2013_quality_variant_vcf_9518_TAIR10	
MNF-Che-2	MPICWang2013_quality_variant_vcf_1925_TAIR10	Y
MNF-Pot-21	MPICWang2013_quality_variant_vcf_1853_TAIR10	Y
Mdn-1	MPICWang2013_quality_variant_vcf_1829_TAIR10	Y
SLSP-31	MPICWang2013_quality_variant_vcf_2276_TAIR10	Y
NOZ-6	MPICWang2013_quality_variant_vcf_9932_TAIR10	Y
RMX3.22	MPICWang2013_quality_variant_vcf_8132_TAIR10	
MNF-Pot-75	MPICWang2013_quality_variant_vcf_1872_TAIR10	
UKSW06-333	MPICWang2013_quality_variant_vcf_4931_TAIR10	
DIR-9	MPICWang2013_quality_variant_vcf_9920_TAIR10	
Paw-26	MPICWang2013_quality_variant_vcf_2171_TAIR10	
UKID96	MPICWang2013_quality_variant_vcf_5800_TAIR10	
Mos-1-IP	MPICWang2013_quality_variant_vcf_9508_TAIR10	
Etna_2	Salk_quality_variant_filtered_Etna_2	
Ey15-2	MPICao2010_Ey15-2_TAIR10_filtered_variant	
Rue3-1-31	MPICao2010_Rue3-1-31_TAIR10_filtered_variant	
Agu-1	MPICao2010_Agu-1_TAIR10_filtered_variant	
Pra-6	MPICao2010_Pra-6_TAIR10_filtered_variant	
Vie-0	MPICao2010_Vie-0_TAIR10_filtered_variant	
Cdm-0	MPICao2010_Cdm-0_TAIR10_filtered_variant	
Nw_0	Salk_quality_variant_filtered_Nw_0	
Hn_0	Salk_quality_variant_filtered_Hn_0	
Fondi-1	MPICWang2013_quality_variant_vcf_9652_TAIR10	
ICE49	MPICao2010_ICE49_TAIR10_filtered_variant	
Qui-0	MPICao2010_Qui-0_TAIR10_filtered_variant	
HKT2_4	MPICao2010_HKT2_4_TAIR10_filtered_variant	
Istisu-1	MPICao2010_Istisu-1_TAIR10_filtered_variant	
ICE50	MPICao2010_ICE50_TAIR10_filtered_variant	
Ove_0	Salk_quality_variant_filtered_Ove_0	
ICE112	MPICao2010_ICE112_TAIR10_filtered_variant	
ICE98	MPICao2010_ICE98_TAIR10_filtered_variant	
ICE107	MPICao2010_ICE107_TAIR10_filtered_variant	
ICE91	MPICao2010_ICE91_TAIR10_filtered_variant	
ICE228	MPICao2010_ICE228_TAIR10_filtered_variant	
ICE73	MPICao2010_ICE73_TAIR10_filtered_variant	
Mz_0	Salk_quality_variant_filtered_Mz_0	
ICE92	MPICao2010_ICE92_TAIR10_filtered_variant	
Nemrut-1	MPICao2010_Nemrut-1_TAIR10_filtered_variant	
Vash-1	MPICao2010_Vash-1_TAIR10_filtered_variant	
Lerik1-3	MPICao2010_Lerik1-3_TAIR10_filtered_variant	
Xan-1	MPICao2010_Xan-1_TAIR10_filtered_variant	
Bak-2	MPICao2010_Bak-2_TAIR10_filtered_variant	
Bak-7	MPICao2010_Bak-7_TAIR10_filtered_variant	
ICE169	MPICao2010_ICE169_TAIR10_filtered_variant	
ICE173	MPICao2010_ICE173_TAIR10_filtered_variant	

ICE212	MPICao2010_ICE212_TAIR10_filtered_variant
ICE213	MPICao2010_ICE213_TAIR10_filtered_variant
ICE63	MPICao2010_ICE63_TAIR10_filtered_variant
ICE79	MPICao2010_ICE79_TAIR10_filtered_variant
Got_7	Salk_quality_variant_filtered_Got_7
Mer-6	MPICao2010_Mer-6_TAIR10_filtered_variant
Hs_0	Salk_quality_variant_filtered_Hs_0
Old_1	Salk_quality_variant_filtered_Old_1
El_0	Salk_quality_variant_filtered_El_0
Kl_5	Salk_quality_variant_filtered_Kl_5
Del-10	MPICao2010_Del-10_TAIR10_filtered_variant
ICE29	MPICao2010_ICE29_TAIR10_filtered_variant
ICE216	MPICao2010_ICE216_TAIR10_filtered_variant
TueSB30-3	MPICao2010_TueSB30-3_TAIR10_filtered_variant
Dog-4	MPICao2010_Dog-4_TAIR10_filtered_variant
WalhaesB4	MPICao2010_WalhaesB4_TAIR10_filtered_variant
ICE163	MPICao2010_ICE163_TAIR10_filtered_variant
Appt_1	Salk_quality_variant_filtered_Appt_1
Gie_0	Salk_quality_variant_filtered_Gie_0
Nok_3	Salk_quality_variant_filtered_Nok_3
Amel_1	Salk_quality_variant_filtered_Amel_1
Ven_1	Salk_quality_variant_filtered_Ven_1
Aa_0	Salk_quality_variant_filtered_Aa_0
Cal_0	Salk_quality_variant_filtered_Cal_0
Ty_0	Salk_quality_variant_filtered_Ty_0
Tha_1	Salk_quality_variant_filtered_Tha_1
Pro_0	Salk_quality_variant_filtered_Pro_0
Cerv_1	Salk_quality_variant_filtered_Cerv_1
Rhen_1	Salk_quality_variant_filtered_Rhen_1
Su_0	Salk_quality_variant_filtered_Su_0
Mnz_0	Salk_quality_variant_filtered_Mnz_0
Koch-1	MPICao2010_Koch-1_TAIR10_filtered_variant
Wc_1	Salk_quality_variant_filtered_Wc_1
Gel_1	Salk_quality_variant_filtered_Gel_1
Db_1	Salk_quality_variant_filtered_Db_1
Hey_1	Salk_quality_variant_filtered_Hey_1
Wt_5	Salk_quality_variant_filtered_Wt_5
Pt_0	Salk_quality_variant_filtered_Pt_0
Rd_0	Salk_quality_variant_filtered_Rd_0
Rld_1	Salk_quality_variant_filtered_Rld_1
Sg_1	Salk_quality_variant_filtered_Sg_1
Uk_1	Salk_quality_variant_filtered_Uk_1
ICE75	MPICao2010_ICE75_TAIR10_filtered_variant
Yeg-1	MPICao2010_Yeg-1_TAIR10_filtered_variant
ICE138	MPICao2010_ICE138_TAIR10_filtered_variant
ICE71	MPICao2010_ICE71_TAIR10_filtered_variant
Mc_0	Salk_quality_variant_filtered_Mc_0
Tamm_2	Salk_quality_variant_filtered_Tamm_2
La_0	Salk_quality_variant_filtered_La_0
Chat_1	Salk_quality_variant_filtered_Chat_1
8411	MPICWang2013_quality_variant_vcf_8411_TAIR10
Pla_0	Salk_quality_variant_filtered_Pla_0
Anz_0	Salk_quality_variant_filtered_Anz_0
Cnt_1	Salk_quality_variant_filtered_Cnt_1
Abd_0	Salk_quality_variant_filtered_Abd_0
Ema_1	Salk_quality_variant_filtered_Ema_1
Ber	Salk_quality_variant_filtered_Ber
Bl_1	Salk_quality_variant_filtered_Bl_1
Vind_1	Salk_quality_variant_filtered_Vind_1
Durh_1	Salk_quality_variant_filtered_Durh_1
HR_5	Salk_quality_variant_filtered_HR_5
Es_0	Salk_quality_variant_filtered_Es_0
Np_0	Salk_quality_variant_filtered_Np_0
Hau_0	Salk_quality_variant_filtered_Hau_0

Bd_0	Salk_quality_variant_filtered_Bd_0
Wei_0	Salk_quality_variant_filtered_Wei_0
Sp_0	Salk_quality_variant_filtered_Sp_0
Com_1	Salk_quality_variant_filtered_Com_1
Pog_0	Salk_quality_variant_filtered_Pog_0
Lm_2	Salk_quality_variant_filtered_Lm_2
Rou_0	Salk_quality_variant_filtered_Rou_0
Wl_0	Salk_quality_variant_filtered_Wl_0
Utrecht	Salk_quality_variant_filtered_Utrecht
Br_0	Salk_quality_variant_filtered_Br_0
Gy_0	Salk_quality_variant_filtered_Gy_0
Ang_0	Salk_quality_variant_filtered_Ang_0
Boot_1	Salk_quality_variant_filtered_Boot_1
Et_0	Salk_quality_variant_filtered_Et_0
Hh_0	Salk_quality_variant_filtered_Hh_0
Bs_1	Salk_quality_variant_filtered_Bs_1
Knox_18	Salk_quality_variant_filtered_Knox_18
Anholt_1	Salk_quality_variant_filtered_Anholt_1
Nc_1	Salk_quality_variant_filtered_Nc_1
Rennes_1	Salk_quality_variant_filtered_Rennes_1
Ag_0	Salk_quality_variant_filtered_Ag_0
Ann_1	Salk_quality_variant_filtered_Ann_1
Bla_1	Salk_quality_variant_filtered_Bla_1
Er_0	Salk_quality_variant_filtered_Er_0
Baa_1	Salk_quality_variant_filtered_Baa_1
Tu_0	Salk_quality_variant_filtered_Tu_0
Est	Salk_quality_variant_filtered_Est
Krot_0	Salk_quality_variant_filtered_Krot_0
Tscha_1	Salk_quality_variant_filtered_Tscha_1
Rubeznhoe_1	Salk_quality_variant_filtered_Rubeznhoe_1
Chi_0	Salk_quality_variant_filtered_Chi_0
Ms_0	Salk_quality_variant_filtered_Ms_0
Westkar_4	Salk_quality_variant_filtered_Westkar_4
Zal_1	Salk_quality_variant_filtered_Zal_1
Altai_5	Salk_quality_variant_filtered_Altai_5
Sorbo	Salk_quality_variant_filtered_Sorbo
Kz_9	Salk_quality_variant_filtered_Kz_9
Sus_1	Salk_quality_variant_filtered_Sus_1
Kas_1	Salk_quality_variant_filtered_Kas_1
Dja_1	Salk_quality_variant_filtered_Dja_1
Kondara	Salk_quality_variant_filtered_Kondara
Kar_1	Salk_quality_variant_filtered_Kar_1
Neo_6	Salk_quality_variant_filtered_Neo_6
Bor_4	Salk_quality_variant_filtered_Bor_4
Ta_0	Salk_quality_variant_filtered_Ta_0
Ting_1	Salk_quality_variant_filtered_Ting_1
Per_1	Salk_quality_variant_filtered_Per_1
In_0	Salk_quality_variant_filtered_In_0
Co_1	Salk_quality_variant_filtered_Co_1
Se_0	Salk_quality_variant_filtered_Se_0
Ts_1	Salk_quality_variant_filtered_Ts_1
Uod_1	Salk_quality_variant_filtered_Uod_1
Jl_3	Salk_quality_variant_filtered_Jl_3
Zdr_1	Salk_quality_variant_filtered_Zdr_1
Gr_1	Salk_quality_variant_filtered_Gr_1
Di_G	Salk_quality_variant_filtered_Di_G
Kyoto	Salk_quality_variant_filtered_Kyoto
Sei_0	Salk_quality_variant_filtered_Sei_0
Wa_1	Salk_quality_variant_filtered_Wa_1
Stw_0	Salk_quality_variant_filtered_Stw_0
Pu2_23	Salk_quality_variant_filtered_Pu2_23
Pu2_7	Salk_quality_variant_filtered_Pu2_7
Da1_12	Salk_quality_variant_filtered_Da1_12
Dra_0	Salk_quality_variant_filtered_Dra_0

Je_0	Salk_quality_variant_filtered_Je_0
Bor_1	Salk_quality_variant_filtered_Bor_1
Lp2_2	Salk_quality_variant_filtered_Lp2_2
Jm_0	Salk_quality_variant_filtered_Jm_0
Litva	Salk_quality_variant_filtered_Litva
Ga_0	Salk_quality_variant_filtered_Ga_0
En_D	Salk_quality_variant_filtered_En_D
Pi_0	Salk_quality_variant_filtered_Pi_0
Dr_0	Salk_quality_variant_filtered_Dr_0
Ko_2	Salk_quality_variant_filtered_Ko_2
Li_2:1	Salk_quality_variant_filtered_Li_2:1
Rmx_A02	Salk_quality_variant_filtered_Rmx_A02
Kin_0	Salk_quality_variant_filtered_Kin_0
Seattle_0	Salk_quality_variant_filtered_Seattle_0
Lan_0	Salk_quality_variant_filtered_Lan_0
Rome_1	Salk_quality_variant_filtered_Rome_1
ICE127	MPICao2010_ICE127_TAIR10_filtered_variant
Sha	MPICao2010_Sha_TAIR10_filtered_variant
ICE70	MPICao2010_ICE70_TAIR10_filtered_variant
ICE152	MPICao2010_ICE152_TAIR10_filtered_variant
ICE150	MPICao2010_ICE150_TAIR10_filtered_variant
ICE153	MPICao2010_ICE153_TAIR10_filtered_variant
ICE61	MPICao2010_ICE61_TAIR10_filtered_variant
ICE130	MPICao2010_ICE130_TAIR10_filtered_variant
ICE134	MPICao2010_ICE134_TAIR10_filtered_variant
ICE60	MPICao2010_ICE60_TAIR10_filtered_variant
ICE72	MPICao2010_ICE72_TAIR10_filtered_variant
Van_0	Salk_quality_variant_filtered_Van_0
Lip_0	Salk_quality_variant_filtered_Lip_0
Ra_0	Salk_quality_variant_filtered_Ra_0
Benk_1	Salk_quality_variant_filtered_Benk_1
RRS_7	Salk_quality_variant_filtered_RRS_7
Is_0	Salk_quality_variant_filtered_Is_0
Kil_0	Salk_quality_variant_filtered_Kil_0
An_1	Salk_quality_variant_filtered_An_1
Sq_8	Salk_quality_variant_filtered_Sq_8
Kelsterbach_4	Salk_quality_variant_filtered_Kelsterbach_4
Bu_0	Salk_quality_variant_filtered_Bu_0
Le_0	Salk_quality_variant_filtered_Le_0
Bsch_0	Salk_quality_variant_filtered_Bsch_0
Ei_2	Salk_quality_variant_filtered_Ei_2
Fi_0	Salk_quality_variant_filtered_Fi_0
Bch_1	Salk_quality_variant_filtered_Bch_1
Fr_2	Salk_quality_variant_filtered_Fr_2
Leo-1	MPICao2010_Leo-1_TAIR10_filtered_variant
ICE181	MPICao2010_ICE181_TAIR10_filtered_variant
ICE226	MPICao2010_ICE226_TAIR10_filtered_variant
ICE111	MPICao2010_ICE111_TAIR10_filtered_variant
ICE106	MPICao2010_ICE106_TAIR10_filtered_variant
ICE97	MPICao2010_ICE97_TAIR10_filtered_variant
ICE21	MPICao2010_ICE21_TAIR10_filtered_variant
ICE36	MPICao2010_ICE36_TAIR10_filtered_variant
ICE1	MPICao2010_ICE1_TAIR10_filtered_variant
ICE7	MPICao2010_ICE7_TAIR10_filtered_variant
Lag2_2	MPICao2010_Lag2_2_TAIR10_filtered_variant
ICE33	MPICao2010_ICE33_TAIR10_filtered_variant
ICE120	MPICao2010_ICE120_TAIR10_filtered_variant
Kastel-1	MPICao2010_Kastel-1_TAIR10_filtered_variant
ICE119	MPICao2010_ICE119_TAIR10_filtered_variant
ICE104	MPICao2010_ICE104_TAIR10_filtered_variant
ICE102	MPICao2010_ICE102_TAIR10_filtered_variant
ICE93	MPICao2010_ICE93_TAIR10_filtered_variant
Yo_0	Salk_quality_variant_filtered_Yo_0
Tol_0	Salk_quality_variant_filtered_Tol_0

Pna_17	Salk_quality_variant_filtered_Pna_17	
Gifu_2	Salk_quality_variant_filtered_Gifu_2	
Gre_0	Salk_quality_variant_filtered_Gre_0	
RRs_10	Salk_quality_variant_filtered_RRs_10	
Tul_0	Salk_quality_variant_filtered_Tul_0	
Ragl_1	Salk_quality_variant_filtered_Ragl_1	Y
Ws_2	Salk_quality_variant_filtered_Ws_2	Y
Ob_0	Salk_quality_variant_filtered_Ob_0	
Si_0	Salk_quality_variant_filtered_Si_0	
Ha_0	Salk_quality_variant_filtered_Ha_0	
Ca_0	Salk_quality_variant_filtered_Ca_0	
Gu_0	Salk_quality_variant_filtered_Gu_0	
Mh_0	Salk_quality_variant_filtered_Mh_0	
Ba_1	Salk_quality_variant_filtered_Ba_1	
Ak_1	Salk_quality_variant_filtered_Ak_1	
En_2	Salk_quality_variant_filtered_En_2	
Do_0	Salk_quality_variant_filtered_Do_0	
Kro_0	Salk_quality_variant_filtered_Kro_0	
Tuescha9	MPICao2010_Tuescha9_TAIR10_filtered_variant	
TueWa1-2	MPICao2010_TueWa1-2_TAIR10_filtered_variant	
Fei-0	MPICao2010_Fei-0_TAIR10_filtered_variant	
TueV13	MPICao2010_TueV13_TAIR10_filtered_variant	
Nie1-2	MPICao2010_Nie1-2_TAIR10_filtered_variant	
Star-8	MPICao2010_Star-8_TAIR10_filtered_variant	
Li-7	MPICWang2013_quality_variant_vcf_7231_TAIR10	
WAR	MPICWang2013_quality_variant_vcf_7477_TAIR10	
Cap-1-IP	MPICWang2013_quality_variant_vcf_9529_TAIR10	
Car-1-IP	MPICWang2013_quality_variant_vcf_9530_TAIR10	
UKNW06-481	MPICWang2013_quality_variant_vcf_5644_TAIR10	
Hom-4-IP	MPICWang2013_quality_variant_vcf_9546_TAIR10	
Iso-4-IP	MPICWang2013_quality_variant_vcf_9550_TAIR10	
ENC-2-1	MPICWang2013_quality_variant_vcf_9907_TAIR10	
Ty-1	MPICWang2013_quality_variant_vcf_5784_TAIR10	
San-10-IP	MPICWang2013_quality_variant_vcf_9579_TAIR10	
LI-OF-065	MPICWang2013_quality_variant_vcf_630_TAIR10	
KYC-33	MPICWang2013_quality_variant_vcf_801_TAIR10	
PLY-20	MPICWang2013_quality_variant_vcf_9924_TAIR10	
Men-2-IP	MPICWang2013_quality_variant_vcf_9556_TAIR10	
ARR-17	MPICWang2013_quality_variant_vcf_9927_TAIR10	
MIC-31	MPICWang2013_quality_variant_vcf_870_TAIR10	
UKNW06-003	MPICWang2013_quality_variant_vcf_5353_TAIR10	
Ha-HBT1-2	MPICWang2013_quality_variant_vcf_9785_TAIR10	
Orb-10-IP	MPICWang2013_quality_variant_vcf_9565_TAIR10	
UKSE06-500	MPICWang2013_quality_variant_vcf_5253_TAIR10	
Rev-0-IP	MPICWang2013_quality_variant_vcf_9576_TAIR10	
Scm-0-IP	MPICWang2013_quality_variant_vcf_9580_TAIR10	
Pob-0-IP	MPICWang2013_quality_variant_vcf_9570_TAIR10	
Mur-0-IP	MPICWang2013_quality_variant_vcf_9562_TAIR10	
Oso-0-IP	MPICWang2013_quality_variant_vcf_9566_TAIR10	
Ber-0-IP	MPICWang2013_quality_variant_vcf_9524_TAIR10	
Pan-0-IP	MPICWang2013_quality_variant_vcf_9568_TAIR10	
Cal-0-IP	MPICWang2013_quality_variant_vcf_9528_TAIR10	
Cor-0-IP	MPICWang2013_quality_variant_vcf_9536_TAIR10	
Vaz-0-IP	MPICWang2013_quality_variant_vcf_9593_TAIR10	
Moa-0-IP	MPICWang2013_quality_variant_vcf_9557_TAIR10	
Nog-17-IP	MPICWang2013_quality_variant_vcf_9564_TAIR10	
Bis-0-IP	MPICWang2013_quality_variant_vcf_9525_TAIR10	
Ria-0-IP	MPICWang2013_quality_variant_vcf_9577_TAIR10	
Pal-0-IP	MPICWang2013_quality_variant_vcf_9567_TAIR10	
Vdm-0-IP	MPICWang2013_quality_variant_vcf_9594_TAIR10	
Coc-1-IP	MPICWang2013_quality_variant_vcf_9535_TAIR10	
Ses-0-IP	MPICWang2013_quality_variant_vcf_9582_TAIR10	
Tol-7-IP	MPICWang2013_quality_variant_vcf_9588_TAIR10	
11C1	MPICWang2013_quality_variant_vcf_9503_TAIR10	

LDV-18	MPICWang2013_quality_variant_vcf_108_TAIR10
LDV-46	MPICWang2013_quality_variant_vcf_139_TAIR10
PYL-6	MPICWang2013_quality_variant_vcf_265_TAIR10
Tu-PK-7	MPICWang2013_quality_variant_vcf_9783_TAIR10
Alo-0-IP	MPICWang2013_quality_variant_vcf_9506_TAIR10
PT2.21	MPICWang2013_quality_variant_vcf_8077_TAIR10
Balan-1	MPICWang2013_quality_variant_vcf_9613_TAIR10
SLSP-35	MPICWang2013_quality_variant_vcf_2278_TAIR10
Schip-1	MPICWang2013_quality_variant_vcf_9721_TAIR10
Vav-0-IP	MPICWang2013_quality_variant_vcf_9511_TAIR10
Tu-KB-6	MPICWang2013_quality_variant_vcf_9809_TAIR10
UKID63	MPICWang2013_quality_variant_vcf_5768_TAIR10
Pro-0-IP	MPICWang2013_quality_variant_vcf_9571_TAIR10
LIN-S-5	MPICWang2013_quality_variant_vcf_915_TAIR10
Set-1	MPICWang2013_quality_variant_vcf_5772_TAIR10
Tu-KS-7	MPICWang2013_quality_variant_vcf_9810_TAIR10
Tu-NK-12	MPICWang2013_quality_variant_vcf_9811_TAIR10
PNA3.40	MPICWang2013_quality_variant_vcf_7947_TAIR10
UKSW06-179	MPICWang2013_quality_variant_vcf_4779_TAIR10
Gol-2	MPICWang2013_quality_variant_vcf_9314_TAIR10
LP3413.41	MPICWang2013_quality_variant_vcf_8472_TAIR10
Mun-0-IP	MPICWang2013_quality_variant_vcf_9561_TAIR10
Vpa-1-IP	MPICWang2013_quality_variant_vcf_9602_TAIR10
Fell3-7	MPICWang2013_quality_variant_vcf_9776_TAIR10
Erg2-6	MPICWang2013_quality_variant_vcf_9784_TAIR10
HE-1	MPICWang2013_quality_variant_vcf_9769_TAIR10
Svi-0-IP	MPICWang2013_quality_variant_vcf_9585_TAIR10
Ren-6-IP	MPICWang2013_quality_variant_vcf_9575_TAIR10
Vig-1-IP	MPICWang2013_quality_variant_vcf_9597_TAIR10
CHA-41	MPICWang2013_quality_variant_vcf_932_TAIR10
UKSW06-226	MPICWang2013_quality_variant_vcf_4826_TAIR10
SAUL-24	MPICWang2013_quality_variant_vcf_9918_TAIR10
Ha-HBT3-11	MPICWang2013_quality_variant_vcf_9815_TAIR10
Pigna-1	MPICWang2013_quality_variant_vcf_9659_TAIR10
Mot-0-IP	MPICWang2013_quality_variant_vcf_9560_TAIR10
Cum-1-IP	MPICWang2013_quality_variant_vcf_9537_TAIR10
Gua-1-IP	MPICWang2013_quality_variant_vcf_9544_TAIR10
Adm-0-IP	MPICWang2013_quality_variant_vcf_9514_TAIR10
Rei-0-IP	MPICWang2013_quality_variant_vcf_9510_TAIR10
TOU-A1-89	MPICWang2013_quality_variant_vcf_351_TAIR10
Bea-0-IP	MPICWang2013_quality_variant_vcf_9522_TAIR10
Coa-0-IP	MPICWang2013_quality_variant_vcf_9507_TAIR10
Ver-5-IP	MPICWang2013_quality_variant_vcf_9596_TAIR10
Lag1-4	MPICWang2013_quality_variant_vcf_9102_TAIR10
Lag1-2	MPICWang2013_quality_variant_vcf_9100_TAIR10
Lag1-6	MPICWang2013_quality_variant_vcf_9104_TAIR10
Ben-0-IP	MPICWang2013_quality_variant_vcf_9523_TAIR10
Mon-5-IP	MPICWang2013_quality_variant_vcf_9559_TAIR10
Ala-0-IP	MPICWang2013_quality_variant_vcf_9515_TAIR10
Vid-1-IP	MPICWang2013_quality_variant_vcf_9512_TAIR10
Vdt-0-IP	MPICWang2013_quality_variant_vcf_9595_TAIR10
Deh-1-IP	MPICWang2013_quality_variant_vcf_9539_TAIR10
Tu-B1-2	MPICWang2013_quality_variant_vcf_9794_TAIR10
CON-7	MPICWang2013_quality_variant_vcf_9913_TAIR10
UKSE06-470	MPICWang2013_quality_variant_vcf_5236_TAIR10
Obh-13	MPICWang2013_quality_variant_vcf_9789_TAIR10
Ha-S-B	MPICWang2013_quality_variant_vcf_9800_TAIR10
UKID74	MPICWang2013_quality_variant_vcf_5779_TAIR10
BEZ-9	MPICWang2013_quality_variant_vcf_9928_TAIR10
MAR2-3	MPICWang2013_quality_variant_vcf_159_TAIR10
Vae-2-IP	MPICWang2013_quality_variant_vcf_9592_TAIR10
Trs-0-IP	MPICWang2013_quality_variant_vcf_9590_TAIR10
Ara-4-IP	MPICWang2013_quality_variant_vcf_9520_TAIR10
Bach2-1	MPICWang2013_quality_variant_vcf_9796_TAIR10

Ha-P-13	MPICWang2013_quality_variant_vcf_9786_TAIR10
Tdc-0-IP	MPICWang2013_quality_variant_vcf_9587_TAIR10
UKSE06-118	MPICWang2013_quality_variant_vcf_5023_TAIR10
Alt-1	MPICWang2013_quality_variant_vcf_9774_TAIR10
Fue-2-IP	MPICWang2013_quality_variant_vcf_9541_TAIR10
Tu-WH	MPICWang2013_quality_variant_vcf_9816_TAIR10
Pfn-N2.2-6	MPICWang2013_quality_variant_vcf_9771_TAIR10
Ru-N2	MPICWang2013_quality_variant_vcf_9793_TAIR10
TOU-A1-88	MPICWang2013_quality_variant_vcf_350_TAIR10
Ha-HBT2-10	MPICWang2013_quality_variant_vcf_9797_TAIR10
Pfn-10	MPICWang2013_quality_variant_vcf_9805_TAIR10
Berg-1	MPICWang2013_quality_variant_vcf_9775_TAIR10
Haes-1	MPICWang2013_quality_variant_vcf_9791_TAIR10
BRE-14	MPICWang2013_quality_variant_vcf_9919_TAIR10
Ullapool-8	MPICWang2013_quality_variant_vcf_9312_TAIR10
Yeg-5	MPICWang2013_quality_variant_vcf_9131_TAIR10
Liri-1	MPICWang2013_quality_variant_vcf_9654_TAIR10
Sever-1	MPICWang2013_quality_variant_vcf_9643_TAIR10
Sac-0-IP	MPICWang2013_quality_variant_vcf_9578_TAIR10
UKSW06-360	MPICWang2013_quality_variant_vcf_4958_TAIR10
UKSE06-325	MPICWang2013_quality_variant_vcf_5151_TAIR10
UKSE06-533	MPICWang2013_quality_variant_vcf_5276_TAIR10
UKNW06-403	MPICWang2013_quality_variant_vcf_5577_TAIR10
CATS-6	MPICWang2013_quality_variant_vcf_9937_TAIR10
Cnt-1	MPICWang2013_quality_variant_vcf_5726_TAIR10
UKSW06-285	MPICWang2013_quality_variant_vcf_4884_TAIR10
GEN-8	MPICWang2013_quality_variant_vcf_9909_TAIR10
UKSW06-302	MPICWang2013_quality_variant_vcf_4900_TAIR10
RUM-20	MPICWang2013_quality_variant_vcf_9925_TAIR10
Cur-4-IP	MPICWang2013_quality_variant_vcf_9538_TAIR10
Hart-2	MPICWang2013_quality_variant_vcf_9799_TAIR10
Tor-1-IP	MPICWang2013_quality_variant_vcf_9589_TAIR10
CYR	MPICWang2013_quality_variant_vcf_88_TAIR10
UKSW06-207	MPICWang2013_quality_variant_vcf_4807_TAIR10
KBG1-14	MPICWang2013_quality_variant_vcf_9788_TAIR10
Ha-P2-1	MPICWang2013_quality_variant_vcf_9798_TAIR10
Ha-SP-2	MPICWang2013_quality_variant_vcf_9801_TAIR10
UKSE06-432	MPICWang2013_quality_variant_vcf_5210_TAIR10
Schl-7	MPICWang2013_quality_variant_vcf_9807_TAIR10
Gn-1	MPICWang2013_quality_variant_vcf_9777_TAIR10
Ru-2	MPICWang2013_quality_variant_vcf_9806_TAIR10
RAD-21	MPICWang2013_quality_variant_vcf_9917_TAIR10
Muh-2	MPICWang2013_quality_variant_vcf_9803_TAIR10
Bar-1-IP	MPICWang2013_quality_variant_vcf_9521_TAIR10
Tu-B2-3	MPICWang2013_quality_variant_vcf_9808_TAIR10
Ru4-16	MPICWang2013_quality_variant_vcf_9768_TAIR10
Hof-1	MPICWang2013_quality_variant_vcf_9772_TAIR10
Bach-7	MPICWang2013_quality_variant_vcf_9778_TAIR10
Bai-10	MPICWang2013_quality_variant_vcf_9779_TAIR10
Fell2-4	MPICWang2013_quality_variant_vcf_9780_TAIR10
Obe1-15	MPICWang2013_quality_variant_vcf_9804_TAIR10
Lu3-30	MPICWang2013_quality_variant_vcf_9782_TAIR10
Lu4-2	MPICWang2013_quality_variant_vcf_9792_TAIR10
TRE-1	MPICWang2013_quality_variant_vcf_9926_TAIR10
VED-10	MPICWang2013_quality_variant_vcf_9933_TAIR10
Cad-0-IP	MPICWang2013_quality_variant_vcf_9527_TAIR10
Moc-11-IP	MPICWang2013_quality_variant_vcf_9558_TAIR10
All-0-IP	MPICWang2013_quality_variant_vcf_9517_TAIR10
Jim-1-IP	MPICWang2013_quality_variant_vcf_9551_TAIR10
Pds-1-IP	MPICWang2013_quality_variant_vcf_9569_TAIR10
Sdv-3-IP	MPICWang2013_quality_variant_vcf_9581_TAIR10
Hor-0-IP	MPICWang2013_quality_variant_vcf_9547_TAIR10
WAV-8	MPICWang2013_quality_variant_vcf_9938_TAIR10
KBG2-13	MPICWang2013_quality_variant_vcf_9770_TAIR10

Sarno-1	MPICWang2013_quality_variant_vcf_9660_TAIR10
Zu-1	MPICWang2013_quality_variant_vcf_7418_TAIR10
BI-4	MPICWang2013_quality_variant_vcf_9813_TAIR10
Fell1-10	MPICWang2013_quality_variant_vcf_9814_TAIR10
Aitba-1	MPICWang2013_quality_variant_vcf_9606_TAIR10
Mitterberg-3-187	MPICWang2013_quality_variant_vcf_9671_TAIR10
Mitterberg-2-185	MPICWang2013_quality_variant_vcf_9669_TAIR10
Mitterberg-2-184	MPICWang2013_quality_variant_vcf_9668_TAIR10
Mitterberg-1-180	MPICWang2013_quality_variant_vcf_9665_TAIR10
Mitterberg-1-182	MPICWang2013_quality_variant_vcf_9666_TAIR10
Mitterberg-1-183	MPICWang2013_quality_variant_vcf_9667_TAIR10
Olympia-2	MPICWang2013_quality_variant_vcf_9727_TAIR10
CSHL-5	MPICWang2013_quality_variant_vcf_6744_TAIR10
ISS-20	MPICWang2013_quality_variant_vcf_9929_TAIR10
MNF-Jac-12	MPICWang2013_quality_variant_vcf_1954_TAIR10
Toc-1	MPICWang2013_quality_variant_vcf_9739_TAIR10
Petergof	MPICWang2013_quality_variant_vcf_7296_TAIR10
Grivo-1	MPICWang2013_quality_variant_vcf_9714_TAIR10
Geg-14	MPICWang2013_quality_variant_vcf_9125_TAIR10
Yeg-2	MPICWang2013_quality_variant_vcf_9128_TAIR10
Yeg-4	MPICWang2013_quality_variant_vcf_9130_TAIR10
Yeg-7	MPICWang2013_quality_variant_vcf_9133_TAIR10
Yeg-8	MPICWang2013_quality_variant_vcf_9134_TAIR10
Pue-0-IP	MPICWang2013_quality_variant_vcf_9572_TAIR10
Iasi-1	MPICWang2013_quality_variant_vcf_9744_TAIR10
Bela-2	MPICWang2013_quality_variant_vcf_9733_TAIR10
Dolen-1	MPICWang2013_quality_variant_vcf_9697_TAIR10
Goced-1	MPICWang2013_quality_variant_vcf_9698_TAIR10
Teiu-2	MPICWang2013_quality_variant_vcf_9736_TAIR10
Staro-1	MPICWang2013_quality_variant_vcf_9757_TAIR10
Dospa-1	MPICWang2013_quality_variant_vcf_9706_TAIR10
Slavi-2	MPICWang2013_quality_variant_vcf_9723_TAIR10
Melni-2	MPICWang2013_quality_variant_vcf_9704_TAIR10
Dolna-1	MPICWang2013_quality_variant_vcf_9712_TAIR10
Smolj-1	MPICWang2013_quality_variant_vcf_9718_TAIR10
Giffo-1	MPICWang2013_quality_variant_vcf_9653_TAIR10
Teano-1	MPICWang2013_quality_variant_vcf_9663_TAIR10
Leska-1	MPICWang2013_quality_variant_vcf_9716_TAIR10
Stara-1	MPICWang2013_quality_variant_vcf_9713_TAIR10
Kolar-2	MPICWang2013_quality_variant_vcf_9702_TAIR10
Corig-1	MPICWang2013_quality_variant_vcf_9650_TAIR10
Melic-1	MPICWang2013_quality_variant_vcf_9657_TAIR10
Gr-5	MPICWang2013_quality_variant_vcf_7158_TAIR10
Zagub-1	MPICWang2013_quality_variant_vcf_9748_TAIR10
Lesno-2	MPICWang2013_quality_variant_vcf_9612_TAIR10
Panik-1	MPICWang2013_quality_variant_vcf_9607_TAIR10
MNF-Riv-21	MPICWang2013_quality_variant_vcf_1890_TAIR10
Basta-1	MPICWang2013_quality_variant_vcf_9619_TAIR10
Lebja-2	MPICWang2013_quality_variant_vcf_9632_TAIR10
Sij-1/96	MPICWang2013_quality_variant_vcf_9745_TAIR10
Chaba-2	MPICWang2013_quality_variant_vcf_9624_TAIR10
Kolyv-2	MPICWang2013_quality_variant_vcf_9625_TAIR10
Lesno-1	MPICWang2013_quality_variant_vcf_9611_TAIR10
Basta-2	MPICWang2013_quality_variant_vcf_9620_TAIR10
Noveg-3	MPICWang2013_quality_variant_vcf_9638_TAIR10
Rakit-3	MPICWang2013_quality_variant_vcf_9642_TAIR10
Adam-1	MPICWang2013_quality_variant_vcf_9609_TAIR10
K-oze-1	MPICWang2013_quality_variant_vcf_9629_TAIR10
Kolyv-3	MPICWang2013_quality_variant_vcf_9626_TAIR10
Noveg-1	MPICWang2013_quality_variant_vcf_9636_TAIR10
K-oze-3	MPICWang2013_quality_variant_vcf_9630_TAIR10
Lebja-4	MPICWang2013_quality_variant_vcf_9633_TAIR10
Lesno-4	MPICWang2013_quality_variant_vcf_9610_TAIR10
Parti-1	MPICWang2013_quality_variant_vcf_9615_TAIR10

Karag-1	MPICWang2013_quality_variant_vcf_9617_TAIR10
Kolyv-6	MPICWang2013_quality_variant_vcf_9628_TAIR10
Stilo-1	MPICWang2013_quality_variant_vcf_9662_TAIR10
Noveg-2	MPICWang2013_quality_variant_vcf_9637_TAIR10
Rakit-1	MPICWang2013_quality_variant_vcf_9640_TAIR10
Ulies-1	MPICWang2013_quality_variant_vcf_9737_TAIR10
Orast-1	MPICWang2013_quality_variant_vcf_9741_TAIR10
Furni-1	MPICWang2013_quality_variant_vcf_9743_TAIR10
Bijisk-4	MPICWang2013_quality_variant_vcf_9622_TAIR10
Panke-1	MPICWang2013_quality_variant_vcf_9639_TAIR10
Kolyv-5	MPICWang2013_quality_variant_vcf_9627_TAIR10
Karag-2	MPICWang2013_quality_variant_vcf_9608_TAIR10
Masl-1	MPICWang2013_quality_variant_vcf_9634_TAIR10
Castelfed-4-211	MPICWang2013_quality_variant_vcf_9695_TAIR10
Castelfed-4-214	MPICWang2013_quality_variant_vcf_9696_TAIR10
IST-29	MPICWang2013_quality_variant_vcf_9914_TAIR10
Faneronemi-3	MPICWang2013_quality_variant_vcf_9726_TAIR10
Koren-1	MPICWang2013_quality_variant_vcf_9719_TAIR10
Zupan-1	MPICWang2013_quality_variant_vcf_9644_TAIR10
Gradi-1	MPICWang2013_quality_variant_vcf_9645_TAIR10
Bela-1	MPICWang2013_quality_variant_vcf_9730_TAIR10
Knjas-1	MPICWang2013_quality_variant_vcf_9749_TAIR10
Castelfed-1-197	MPICWang2013_quality_variant_vcf_9681_TAIR10
Epidauros-1	MPICWang2013_quality_variant_vcf_9725_TAIR10
MOL-1	MPICWang2013_quality_variant_vcf_9916_TAIR10
Lab-7-IP	MPICWang2013_quality_variant_vcf_9552_TAIR10
Vad-0-IP	MPICWang2013_quality_variant_vcf_9591_TAIR10
Nav-0-IP	MPICWang2013_quality_variant_vcf_9563_TAIR10
Voz-0-IP	MPICWang2013_quality_variant_vcf_9601_TAIR10
Kolar-1	MPICWang2013_quality_variant_vcf_9699_TAIR10
Ang-0-IP	MPICWang2013_quality_variant_vcf_9519_TAIR10
BRI-2	MPICWang2013_quality_variant_vcf_9910_TAIR10
Stp-0-IP	MPICWang2013_quality_variant_vcf_9584_TAIR10
Nicas-1	MPICWang2013_quality_variant_vcf_9658_TAIR10
ARGE-1-15	MPICWang2013_quality_variant_vcf_9911_TAIR10
ESP-1-11	MPICWang2013_quality_variant_vcf_9908_TAIR10
Bivio-1	MPICWang2013_quality_variant_vcf_9649_TAIR10
Aiell-1	MPICWang2013_quality_variant_vcf_9646_TAIR10
Marce-1	MPICWang2013_quality_variant_vcf_9655_TAIR10
Ste-0	MPICWang2013_quality_variant_vcf_7346_TAIR10
Gn2-3	MPICWang2013_quality_variant_vcf_9790_TAIR10
Stiav-1	MPICWang2013_quality_variant_vcf_9728_TAIR10
Kus2-2	MPICWang2013_quality_variant_vcf_9781_TAIR10
MNF-Pin-39	MPICWang2013_quality_variant_vcf_2016_TAIR10
Zdrl-2-9	MPICWang2013_quality_variant_vcf_6434_TAIR10
Zdarec3	MPICWang2013_quality_variant_vcf_403_TAIR10
DraIV-6-13	MPICWang2013_quality_variant_vcf_5984_TAIR10
DraIV-5-12	MPICWang2013_quality_variant_vcf_5950_TAIR10
Draha2	MPICWang2013_quality_variant_vcf_424_TAIR10
UduI-3-36	MPICWang2013_quality_variant_vcf_6390_TAIR10
MAR-4-16	MPICWang2013_quality_variant_vcf_9915_TAIR10
Doubravnik7	MPICWang2013_quality_variant_vcf_410_TAIR10
DraIV-6-22	MPICWang2013_quality_variant_vcf_5993_TAIR10
UduI-1-11	MPICWang2013_quality_variant_vcf_6296_TAIR10
DraIV-1-8	MPICWang2013_quality_variant_vcf_5890_TAIR10
DraII-6	MPICWang2013_quality_variant_vcf_5874_TAIR10
UduI-4-9	MPICWang2013_quality_variant_vcf_6396_TAIR10
Borky1	MPICWang2013_quality_variant_vcf_428_TAIR10
DraIV-2-9	MPICWang2013_quality_variant_vcf_5907_TAIR10
DraIV-1-11	MPICWang2013_quality_variant_vcf_5893_TAIR10
LEC-25	MPICWang2013_quality_variant_vcf_9930_TAIR10
Zdrl-1-23	MPICWang2013_quality_variant_vcf_6424_TAIR10
Zdrl-2-21	MPICWang2013_quality_variant_vcf_6445_TAIR10
Cimin-1	MPICWang2013_quality_variant_vcf_9661_TAIR10

Filet-1	MPICWang2013_quality_variant_vcf_9651_TAIR10	
Ldd-0-IP	MPICWang2013_quality_variant_vcf_9553_TAIR10	
Cab-3-IP	MPICWang2013_quality_variant_vcf_9526_TAIR10	
Tam-0-IP	MPICWang2013_quality_variant_vcf_9586_TAIR10	
MOU2-25	MPICWang2013_quality_variant_vcf_9931_TAIR10	
Rds-0-IP	MPICWang2013_quality_variant_vcf_9573_TAIR10	
Cdc-3-IP	MPICWang2013_quality_variant_vcf_9531_TAIR10	
Cdo-0-IP	MPICWang2013_quality_variant_vcf_9532_TAIR10	
Elb-0-IP	MPICWang2013_quality_variant_vcf_9540_TAIR10	
Vin-0-IP	MPICWang2013_quality_variant_vcf_9599_TAIR10	
Uod-2	MPICWang2013_quality_variant_vcf_8428_TAIR10	
UKID107	MPICWang2013_quality_variant_vcf_5811_TAIR10	
MIL-2	MPICWang2013_quality_variant_vcf_9922_TAIR10	
UKSE06-362	MPICWang2013_quality_variant_vcf_5165_TAIR10	
Don-0	MPICao2010_Don-0_TAIR10_filtered_variant	
Bik_1	Salk_quality_variant_filtered_Bik_1	
Qar_8a	Salk_quality_variant_filtered_Qar_8a	
PLO-1	MPICWang2013_quality_variant_vcf_9923_TAIR10	
Alst_1	Salk_quality_variant_filtered_Alst_1	
Cmo-3-IP	MPICWang2013_quality_variant_vcf_9534_TAIR10	
Nosov-1	MPICWang2013_quality_variant_vcf_9635_TAIR10	
UKSE06-252	MPICWang2013_quality_variant_vcf_5104_TAIR10	
ref	Col-0	Y
7015	MPICWang2013_quality_variant_vcf_7015_TAIR10	
UKID114	MPICWang2013_quality_variant_vcf_5818_TAIR10	
Malii-1	MPICWang2013_quality_variant_vcf_9746_TAIR10	
Boo2-1	MPICWang2013_quality_variant_vcf_8266_TAIR10	
QUI-8	MPICWang2013_quality_variant_vcf_9934_TAIR10	

Chapter 4

CNIDARIA: fast, reference-free clustering of raw and assembled genome and transcriptome NGS data

Aflitos, Saulo Alves, Edouard Severing, Gabino Sanchez-Perez, Sander Peters, Hans de Jong, Dick de Ridder. CNIDARIA fast, reference-free clustering of raw and assembled genome and transcriptome NGS data, accepted for publication by *BMC Bioinformatics*

Summary

Background: Identification of biological specimens is a major requirement for a range of applications. Reference-free methods analyse unprocessed sequencing data without relying on prior knowledge, but generally do not scale to arbitrarily large genomes and arbitrarily large phylogenetic distances.
Results: We present Cnidaria, a practical tool for clustering genomic and transcriptomic data with no limitation on genome size or phylogenetic distances. We successfully simultaneously clustered 169 genomic and transcriptomic datasets from 4 kingdoms, achieving 100% identification accuracy at supra-species level and 78% accuracy for species level.
Discussion: Cnidaria allows for fast, resource-efficient comparison and identification of both raw and assembled genome and transcriptome data. This can help answer both fundamental (e.g. in phylogeny, ecological diversity analysis) and practical questions (e.g. sequencing quality control, primer design).

Background

Unequivocal identification of biological specimens is a major requirement for reliable and reproducible (bio)medical research, control of intellectual property by biological patent holders, regulating the flow of biological specimen across national borders, enforcing the Nagoya protocol (Diversity, 2010) and verifying the authenticity of claims of the biological source of products by customs authority.

Several methods for species identification have been developed based on DNA analysis, that can be classified as probe-based and nucleotide sequencing based methods. Probe-based technologies include microarrays, PCR probes, DNA fingerprinting and immunoassays involving the hybridization of DNA samples with predetermined sets of probes or primers. Such methods are cheap and allow precise identification, but may fail in cases where target DNA is not precisely matched by the probes or primers. Alternatively, nucleotide sequencing methods have been developed to increase accuracy, flexibility and throughput. These can be separated into complete or targeted approaches. Targeted identification of short and highly variable genomic regions by exome capture, Expressed Sequence Tag (EST), DNA barcoding and ribosomal DNA (rDNA) sequencing has been used for many years. Targeted DNA sequencing can be done iteratively for taxonomic identification at subspecies, accession and cultivar levels. Whole Genome Sequencing (WGS) and RNA-seq using Next Generation Sequencing (NGS) technology, examples of complete sequencing methods, have the highest information content of all methods, although its high cost has prevented it from being adopted massively. However, with the recent reduction of costs and increase in throughput, NGS starts to become more prevalent, making it a feasible alternative method for species identification. This calls for the creation of a new a set of tools to comprehensively analyse the deluge of data. Methods for species identification based on NGS data can be separated into two main classes: reference-based and reference-free methods (reviewed in Pettengill *et al.*, 2014). Reference-based methods usually map the sequence reads to the genome of a close rel-

ative and infer the phylogeny by aligning the observed polymorphisms. This technology requires quality control (cleaning) of the data, mapping the data to the genomic sequence of a close relative, and detection and comparison of polymorphisms (Bertels *et al.*, 2014). In contrast, reference-free methods (RFMs) are designed to analyse unprocessed sequencing data without any previous knowledge of its identity. The data can be compared against other datasets of unknown samples, in the case of metagenomics comparing population structures (Chan *et al.*, 2008a; Chan *et al.*, 2008b; Diaz *et al.*, 2009; Greenblum *et al.*, 2015; Hurwitz *et al.*, 2014; McHardy and Rigoutsos, 2007; Schloissnig *et al.*, 2013; Smits *et al.*, 2014; Wood and Salzberg, 2014; Yang *et al.*, 2010) or against a panel of known species. In the latter case, it can identify previously unknown samples, giving it an approximate position relative to the known species.

RFM methods can be based on the Discrete Fourier Transform (DFT), compression and *k*-mers. DFT methods, such as in (Hoang *et al.*, 2015), transform nucleotide sequences into frequency statistics and compare these for species classification. Although remarkably fast, this method is not able to store the differences between the genomes for further enquiry, yielding no insight into sequence composition. Compression based methods calculate the distance between pairs of sequences by analysing the reduction in computer memory usage when both sequences are compressed together (Tran and Chen, 2014). However, compression-based methods are time and resource intensive for large genomes or large datasets. *K*-mer based methods split the nucleotide sequence information in the form of NGS reads (.fastq files) or assembled data (.fasta files) into all its constituent substrings of size k, which are then used to calculate the similarity between the sequences of the samples. Several implementations of *k*-mer based RFMs exists, such as FFP (Sims *et al.*, 2009), Co-Phylog (Yi and Jin 2013), NextABP (Roychowdhury *et al.*, 2013), MultiAlignFree (Ren *et al.*, 2013), kSNP (Gardner and Hall, 2013) and Spaced Words/KMACS (Horwege *et al.*, 2014). Although their underlying principles are generally useful for the analysis of large data collections, most implementations are designed for either analysis of a limited portion of the data, such as organelles or ribosomal DNA, or analysis of closely related species (such as bacteria, in metagenomics applications). As a consequence, it is not feasible to apply these tools on large amounts of whole-genome sequencing data or to analyse data that spans large phylogenetic distances. One exception is the tool proposed in (Cannon *et al.*, 2010) that is able to find polymorphisms shared by subsets of the data by counting and merging sets of *k*-mers. This tool was successfully used in (Kua *et al.*, 2012) to compare 174 chloroplast genomes and has a similar approach to ours.

Here we present Cnidaria, an algorithm that employs a novel RFM strategy for species identification based on *k*-mer counting, designed from the ground up to allow analysis of very large collections of genome, transcriptome and raw NGS data using minimal resources. Cnidaria improves over previous methods and overcomes their limitations on size and phylogenetic distance by allowing fast analysis of complete NGS data. To this

end, it can export a database with pre-processed data so that new samples can be quickly compared against a large database of references, without the need to re-process all the data. In contrast to the method used by ReferenceFree (Cannon *et al.,* 2010), Cnidaria is much faster, produces smaller files, is able to produce phylogenetic trees and uses the popular and fast *k*-mer count software Jellyfish (Marcais and Kingsford, 2011), allowing for easy integration in existing NGS quality checking pipelines. We demonstrate the performance and capabilities of Cnidaria by analysing 169 samples, achieving excellent identification accuracy.

Implementation

Cnidaria works with both raw sequencing data and assembled data, both from WGS and RNA-seq sources, in any combination. It uses *k*-mers extracted by Jellyfish (Marcais and Kingsford, 2011), a fast *k*-mer counting tool that produces a database of all *k*-mers present in a query sequence. The advantage of Jellyfish over comparable software is its ability to create a sparse, compressed database in which the *k*-mers are ordered according to a deterministic hashing algorithm, thus allowing for the parallel and efficient merging/processing of the databases once all *k*-mers are in the same predictable order across different databases.

After creating a database containing all *k*-mers in each sample by running Jellyfish, Cnidaria efficiently merges these databases and extracts the number of *k*-mers shared between the samples. Cnidaria then converts this data into a matrix containing the number of *k*-mers shared between each pair of samples. This information is then used to calculate the *Jaccard* distance, as in Co-Phylog (Yi and Jin, 2013), between samples:

$$D_{Jaccard} = 1 - \frac{V_{ab}}{V_a + V_b - V_{ab}}$$

Here, V_{ab} is the number of *k*-mers shared by both samples A and B, Va is the number of valid *k*-mers in sample A and Vb is the number of valid *k*-mers in sample B. When A is equal to B, the distance is 0. The resulting *Jaccard* distance matrix is then processed by Pycogent v.1.5.3 (Knight *et al.*, 2007), which clusters the data using Neighbour-Joining and creates a phylogenetic tree in Newick format, aiding in the identification of the unknown samples in the dataset. For easy visualization of the data, the summary database can also be converted to a standalone HTML page for (dynamic) display of the phylogenetic tree and plotting any statistics of the analysis directly in the tree. A graphical representation of these steps can be found in Figure 1.

Cnidaria can be run in two modes: Database Creation Mode and Sample Analysis Mode. The latter is an order of magnitude faster than the former, generating only a Cnidaria Summary Database (CSD); Database Creation Mode takes longer to run as it

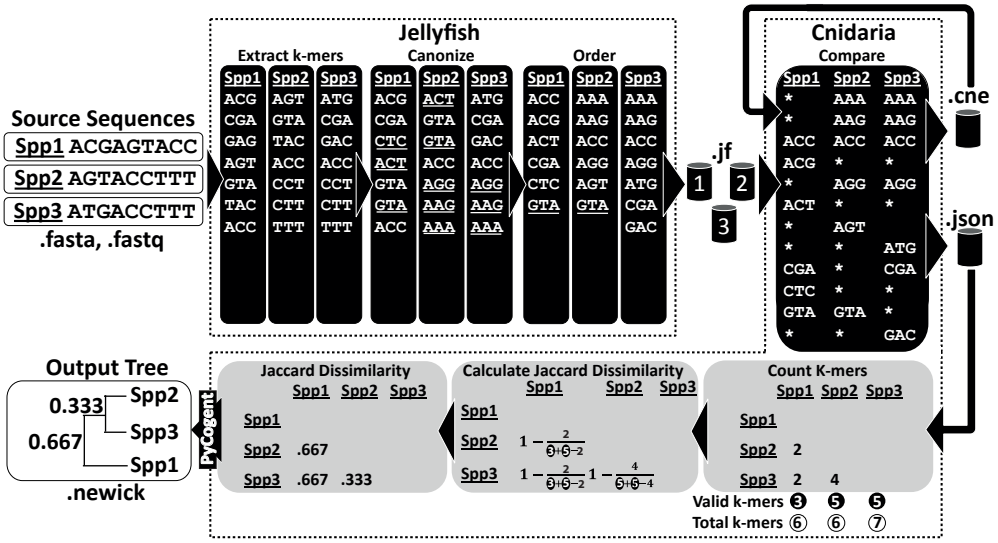


Figure 1: Cnidaria analysis summary. The Jellyfish software reads each of the source sequence files (in Fasta or Fastq formats), extracts their *k*-mers (*k* = 3 in this example), canonizes them (by generating the reverse complement of each *k*-mer and storing only the *k*-mer which appears first lexicographically), orders them according to a deterministic hashing algorithm (in this example, alphabetically) and then saves each dataset in a separated database file (.jf). Cnidaria subsequently reads these databases and compares them, side-by-side, by counting the total number of *k*-mers (white circles), the number of valid *k*-mers (*k*-mers shared by at least two samples, black circles) and the number of shared *k*-mers for each pair of samples as a matrix. Those values are exported to a Cnidaria Summary Database (CSD, a .json file) that is then used to construct a matrix of Jaccard distances between the samples (Formula 1). This dissimilarity matrix is then used for clustering by Neighbour-Joining and exported as a Newick tree. Alternatively, Cnidaria can export a Cnidaria Complete Database (CCD, a .cne file) containing all *k*-mers and a linked list describing their presence/absence in the samples. This second database can be used as an input dataset together with other .cne or .jf files.

exports both a CSD file and a Cnidaria Complete Database (CCD). The CSD contains the total number of *k*-mers for each sample, the number of *k*-mers shared by at least two samples (valid *k*-mers), and the pairwise number of shared *k*-mers. The CCD file contains all *k*-mers present in the datasets analysed, stored in the database using two bits per nucleotide encoding, and their respective presence/absence list describing which set of samples contains each *k*-mer. The CCD can be used as an input to Cnidaria itself, allowing new samples to be directly compared against a pre-calculated larger dataset, speeding up the analysis significantly since the speed of Cnidaria is directly correlated to the number and size of the input files. Hence, the software permits a shorter run time for the comparison of a new sample, using Sample Analysis Mode, against a large reference panel stored in a single CCD file.

Besides the source code, a precompiled version of Cnidaria is available that runs on most 64-bit Linux distributions. To use this, NGS and genome files should first be

converted to a Jellyfish database using Jellyfish (a precompiled version 2.13 is also included with Cnidaria, along with auxiliary scripts to facilitate the conversion). Then, Cnidaria should be run on as many CCD databases as needed, either in Database Creation Mode (producing both a CSD and CCD file) or in Sample Analysis Mode (producing only a CSD file). Either way, an auxiliary python script can be run on the CSD file to generate a *k*-mer count matrix in CSV format, a *Jaccard* distance matrix in CSV format and a phylogenetic tree in Newick format. Helper scripts are available to visualize the Newick tree as a PNG file.

Results and Discussion

Data set

To validate the performance of Cnidaria, we gathered a collection of 135 genomic, transcriptomic and raw NGS datasets covering a wide range of organisms. 84 samples were of tomatoes, part of a large dataset recently published (Aflitos *et al.*, 2014). A list of all samples can be found in Supplementary Table 1. All datasets were analysed using Jellyfish counting canonized *k*-mers. Canonization is the process of only storing the lexicographically smallest between a *k*-mer and its reverse complement. This step is required as both molecules are technically the same: the existence of one implies the existence of the other on the complementary DNA strand. The datasets were then split in 50 pieces and divided over 20 threads on an 80 core Intel(R) Xeon(R) CPU E7- 4850 @ 2.00 GHz machine, speeding up the analysis approximately 40 times compared to single-thread analysis on the same CPU. We then created a Cnidaria Complete Database (CCD) containing all 135 samples.

Influence of *k*-mer size

To investigate the influence of the *k*-mer size in the accuracy of the phylogenetic inference of Cnidaria, we analysed the panel of 135 samples with $k = 11, 15, 17, 21$ and 31 (predefined hash sizes of 128 million, 256 million, 512 million, 1 billion and 4 billion, respectively). The resulting statistics can also be found in Supplementary Table 1.

Due to the low complexity of 11-mers, all possible *k*-mer of this size were found in the datasets and all *k*-mers were shared between at least two samples (Table 1). This carries little clustering information and generates many zero distances (minimum dissimilarity) as shown in Figures 2, 3 and 4. Phylogenetic distances increase with *k*-mer size and 31-mers have most distances equal to 1, i.e. maximum dissimilarity (except for highly related species), which does not allow clustering of distant species.

Identification accuracy

To classify samples, we used the 1-nearest neighbour algorithm on 30 samples for suprasspecies level analysis (8 genus, 7 families, 7 orders, 4 phylum and 3 kingdoms, summa-

Table 1: Summary of search space per *k*-mer size and number of *k*-mers found in datasets. The second column contains the total number of possible *k*-mers, calculated as ($4^{k\text{-mer size}}$ / 2), where the division by two is due to canonization. The third column is the median and the Median Absolute Deviation (MAD) of the total number of *k*-mers found in the samples divided by the number of possible *k*-mers (second column), showing the percentage of combinations actually found and, consequently, the saturation of the search space; the fourth column gives the median and MAD of the percentage of valid *k*-mers (shared between at least two samples).

k-mer size	# Canonical *k*-mer combinations	% of *k*-mers found per sample		% of *k*-mers shared by at least two samples	
		Median	MAD	Median	MAD
11-mer	$2.1x10^{06}$	100.00%	1.58%	100.00%	0.00%
15-mer	$5.4x10^{08}$	53.59%	17.07%	100.00%	0.00%
17-mer	$8.6x10^{09}$	8.90%	4.03%	98.37%	0.99%
21-mer	$2.2x10^{12}$	0.05%	0.03%	81.45%	20.55%
31-mer	$2.3x10^{18}$	0.000000061%	0.000000032%	67.05%	24.14%

rized in Table 2) and on 33 samples for species level analysis (11 species of the Solanum clade, described in Supplementary Tables 2 and 3, summarized in Table 2). The 1-nearest neighbour classifier assigns to each sample, for each phylogenetic level (species, genus, family, order, phylum and kingdom), the phylogenetic class of the sample with the smallest distance. We report the percentage of samples correctly classified. 15-mers and 17-mers yielded accuracy above 70% and 90%, respectively, at the supra-species level but accuracy below 75% at the species level. Both 21- and 31-mers allowed us to correctly classify 100% of the samples at the supra-species level and 78% of the samples at the species level (Figure 5). The lower accuracy for species level classification in the tomato clade can be attributed to introgressions and sympatric speciation in tomato, which is reflected in the clustering distance data (Figure 6) and is in agreement with the clustering obtained by (Aflitos *et al.*, 2014) which used whole genome SNP analysis to construct trees. The 31-mers dataset, when compared to the 21-mers dataset, resulted in an increased run time and disk usage without giving a discernible higher discriminative power (Figures 2 ,3 and 4). This suggests 21 is a good *k*-mer size for general purpose clustering. Conversely, 31-mers are frequently used for NGS data quality checking (reviewed in Leggett *et al.*, 2013) and the same Jellyfish database can be used for species identification.

Table 2: Percentage of correct classification for each k-mer size and each taxonomic level. Sample names are found in Supplementary Tables 2 and 3.

K-mer size	Species	Genus	Family	Order	Phylum	Kingdom
11-mer	0%	57%	63%	63%	77%	77%
15-mer	55%	73%	80%	80%	80%	80%
17-mer	73%	93%	97%	97%	97%	97%
21-mer	79%	100%	100%	100%	100%	100%
31-mer	79%	100%	100%	100%	100%	100%

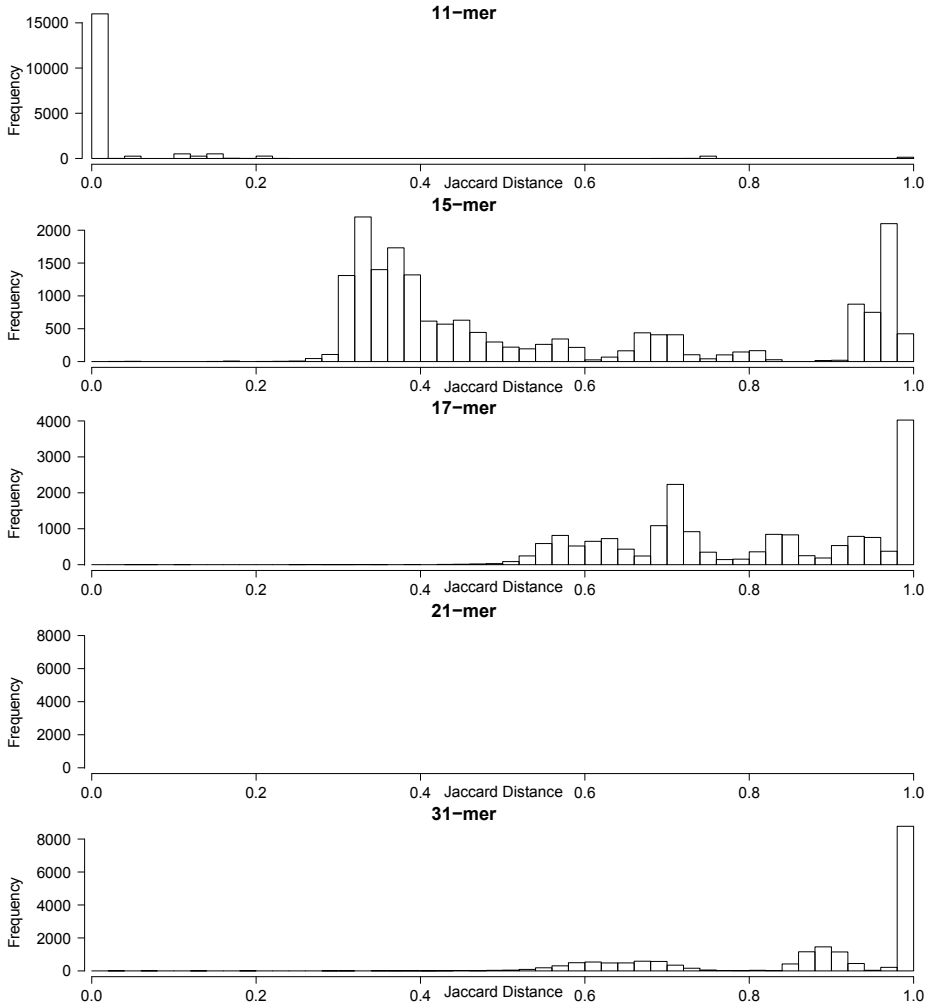


Figure 2: Histogram of Jaccard distances for each k-mer size of the 135 samples. A distance of 0 means identity while a distance of 1 means no similarity. Using 11-mers most samples are identical to each other. For 31-mers, most samples share no similarity with any other sample except for phylogenetically closely related samples. 17 and 21-mers show higher similarity between groups.

3A

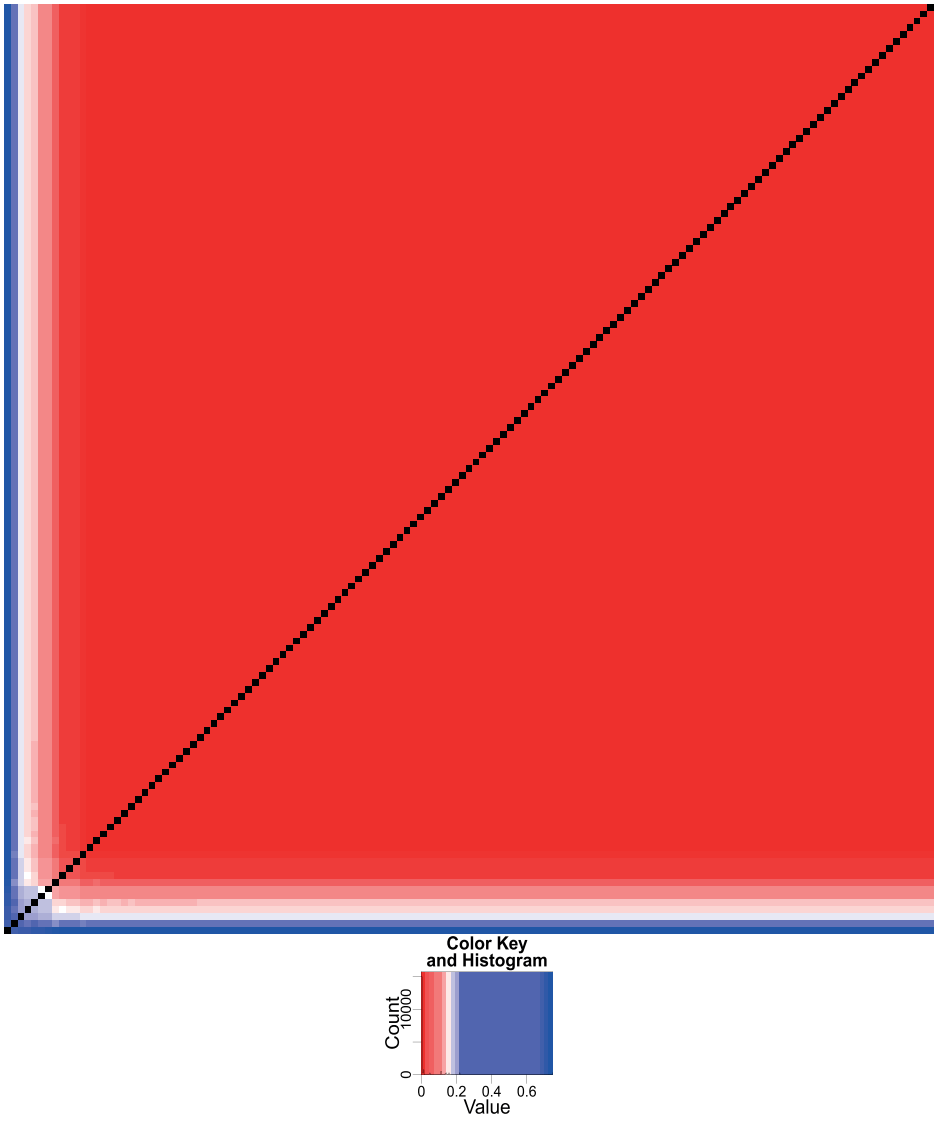
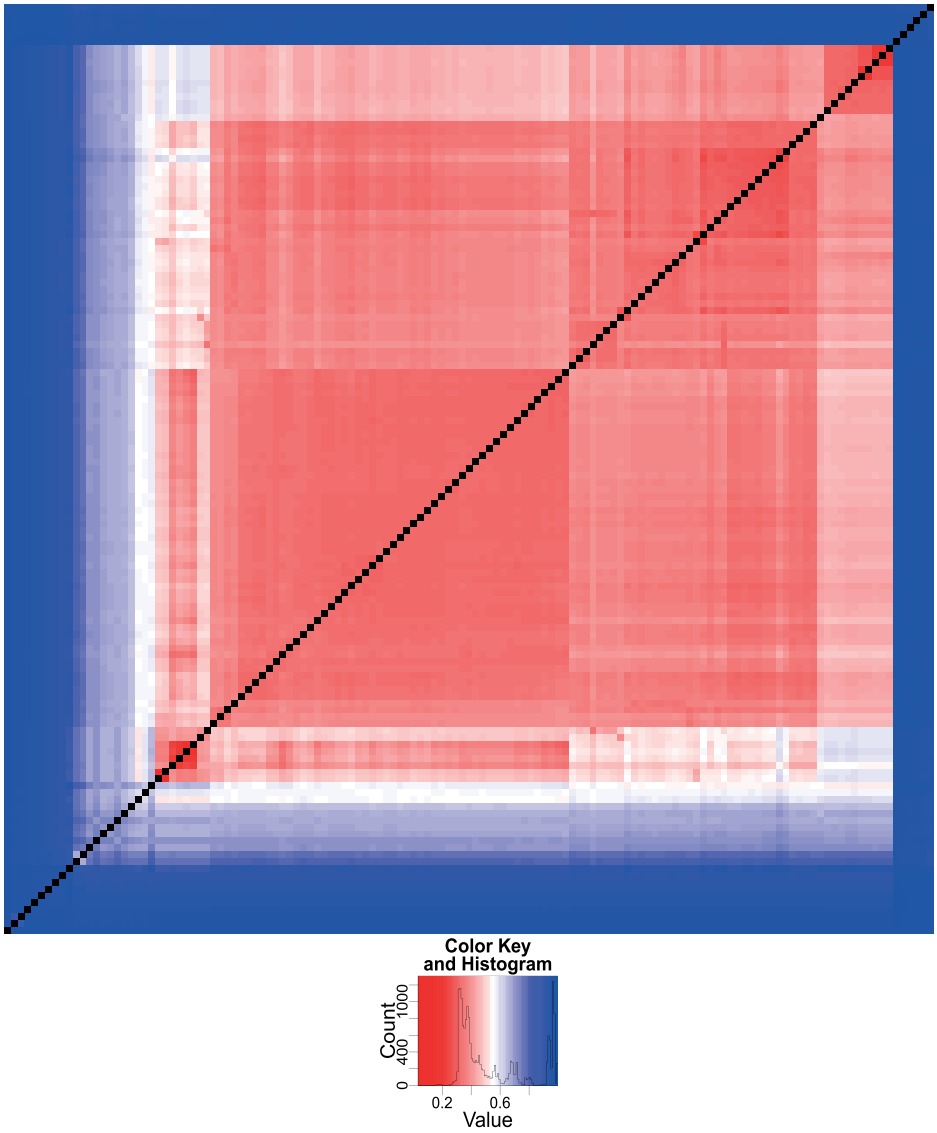
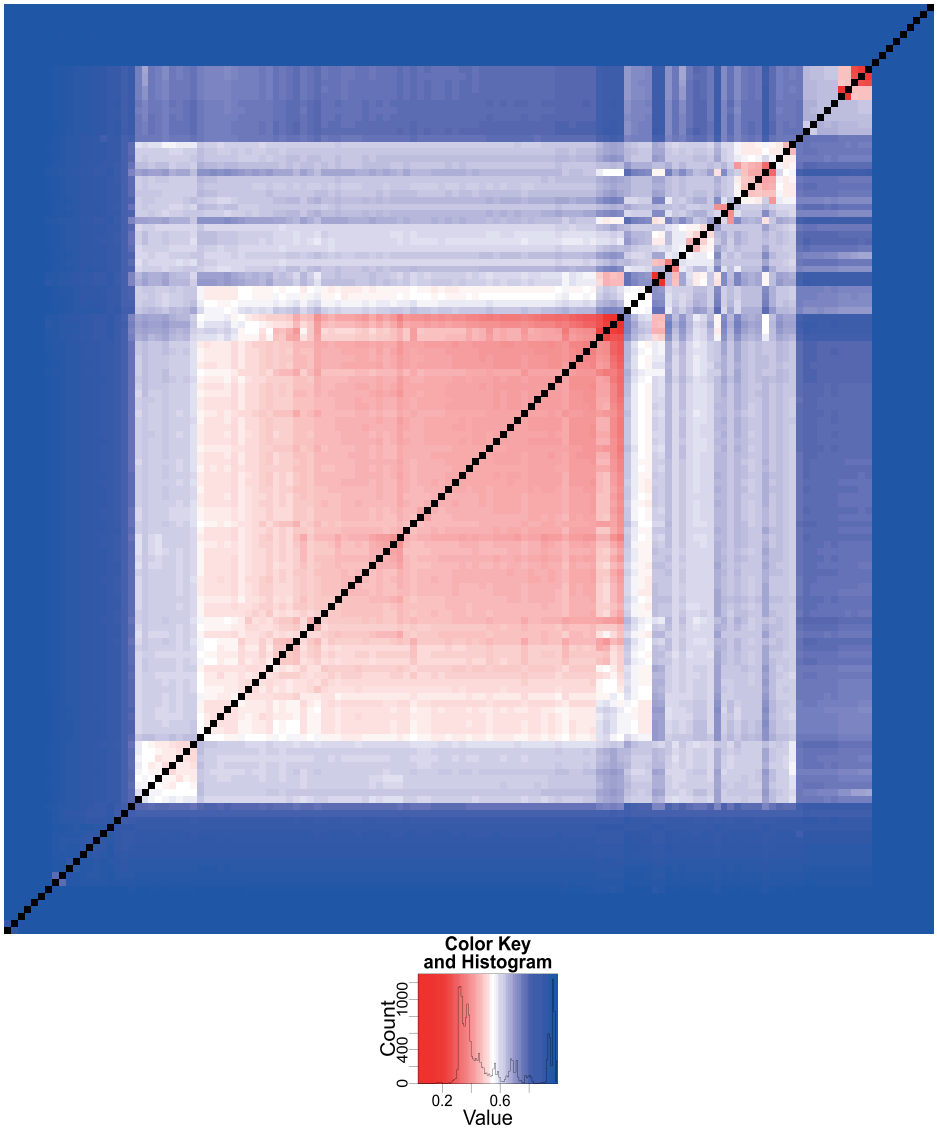


Figure 3: Heatmaps of Jaccard distance of 135 samples using 11-, 15-, 17-, 21- and 31-mers are shown in graphs A, B, C, D and E, respectively. Here, 0 (red) means identity between samples while 1 (blue) means no identity. Generally, closely related species show high similarity with closely related species and no similarity with outgroups. This leads to strong clustering inside groups but loose coupling between groups.

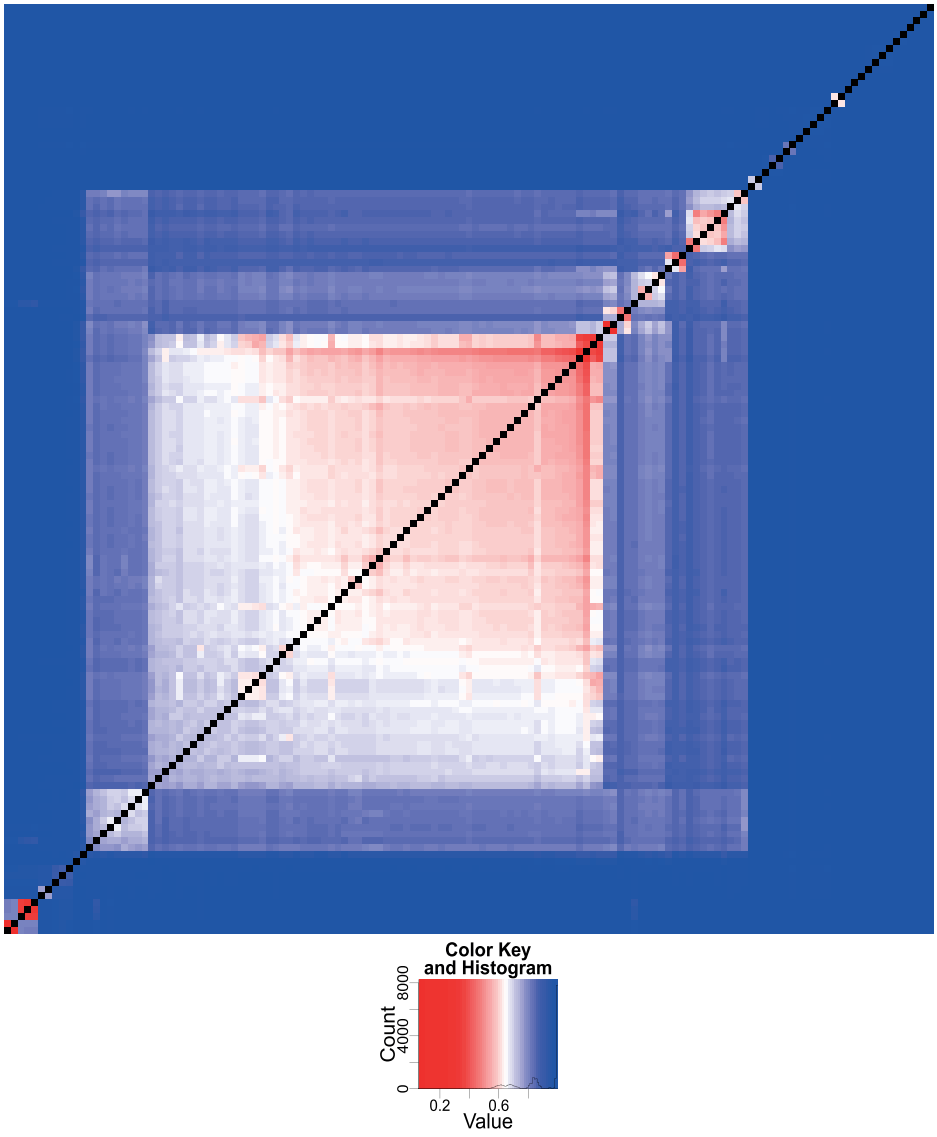
3B



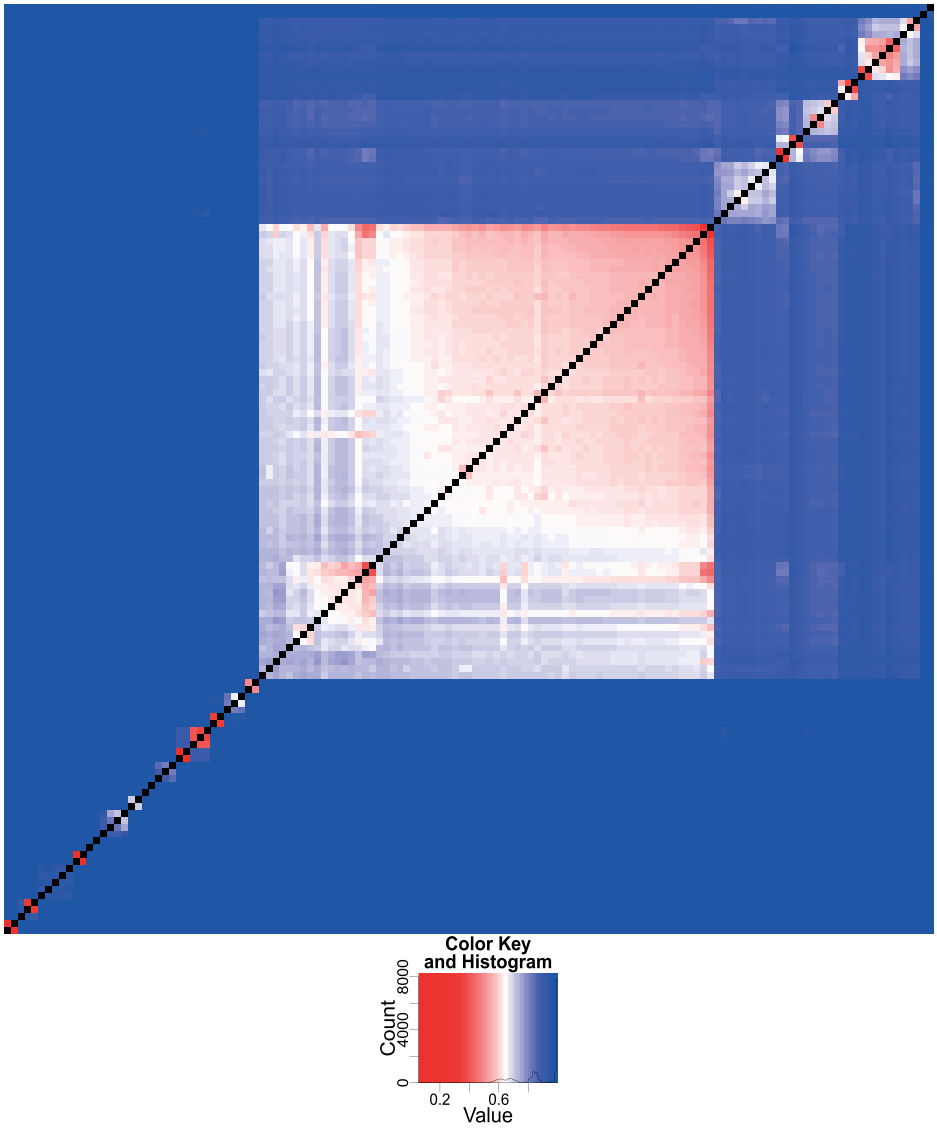
3C



3D



3E



4.A.I

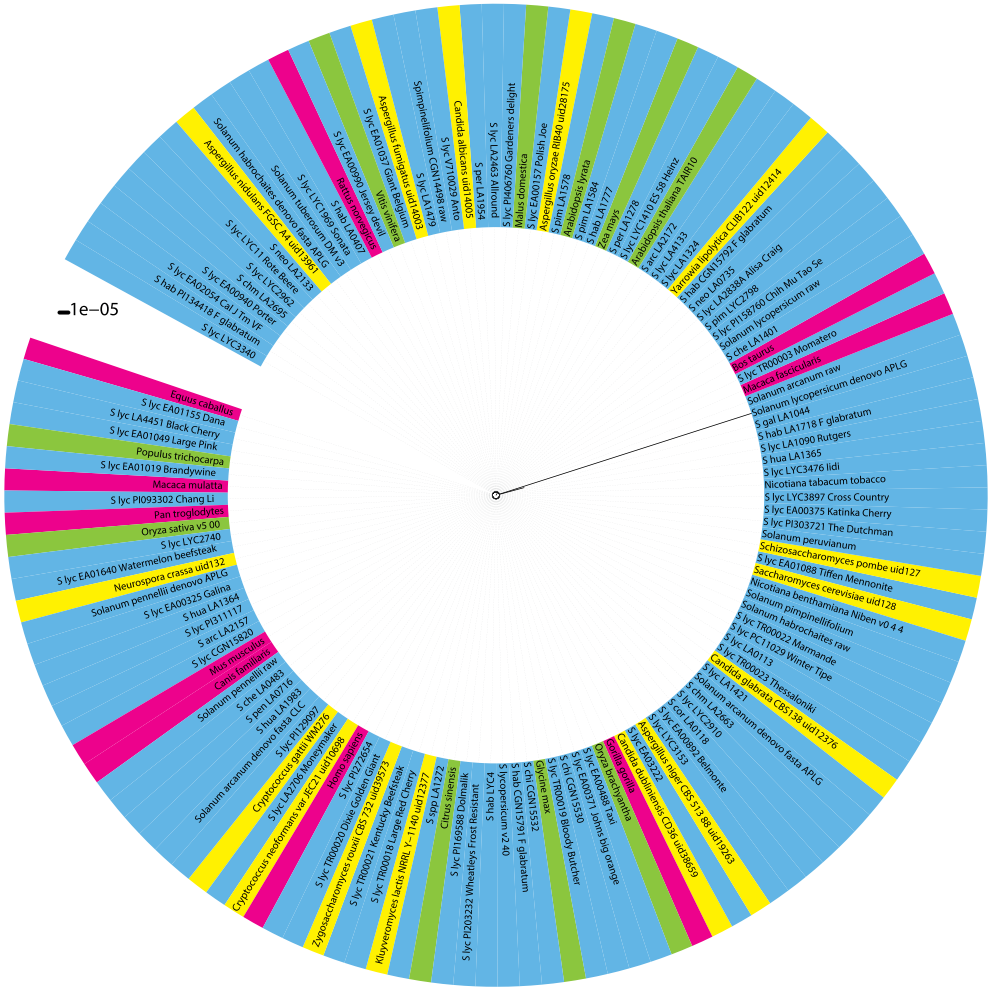
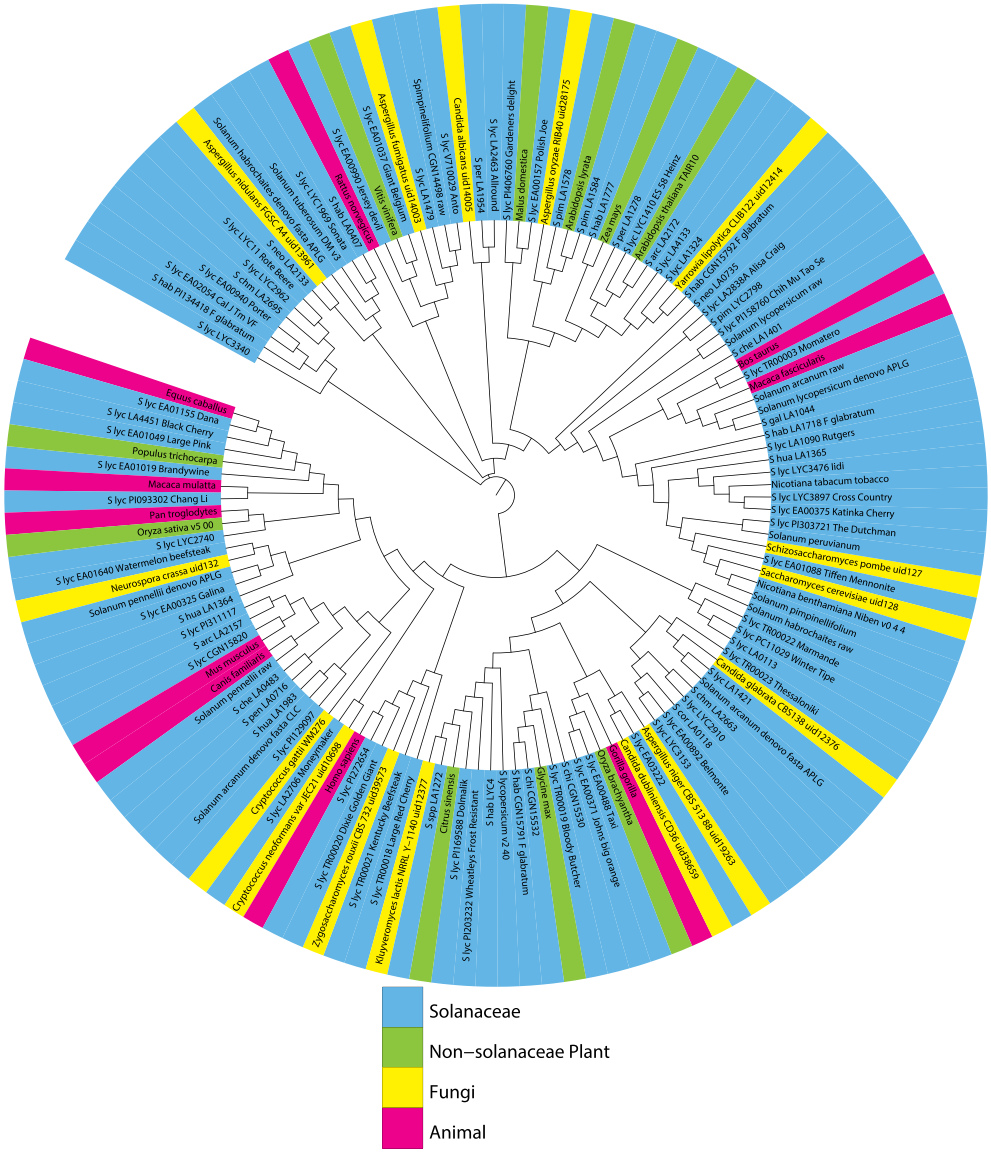
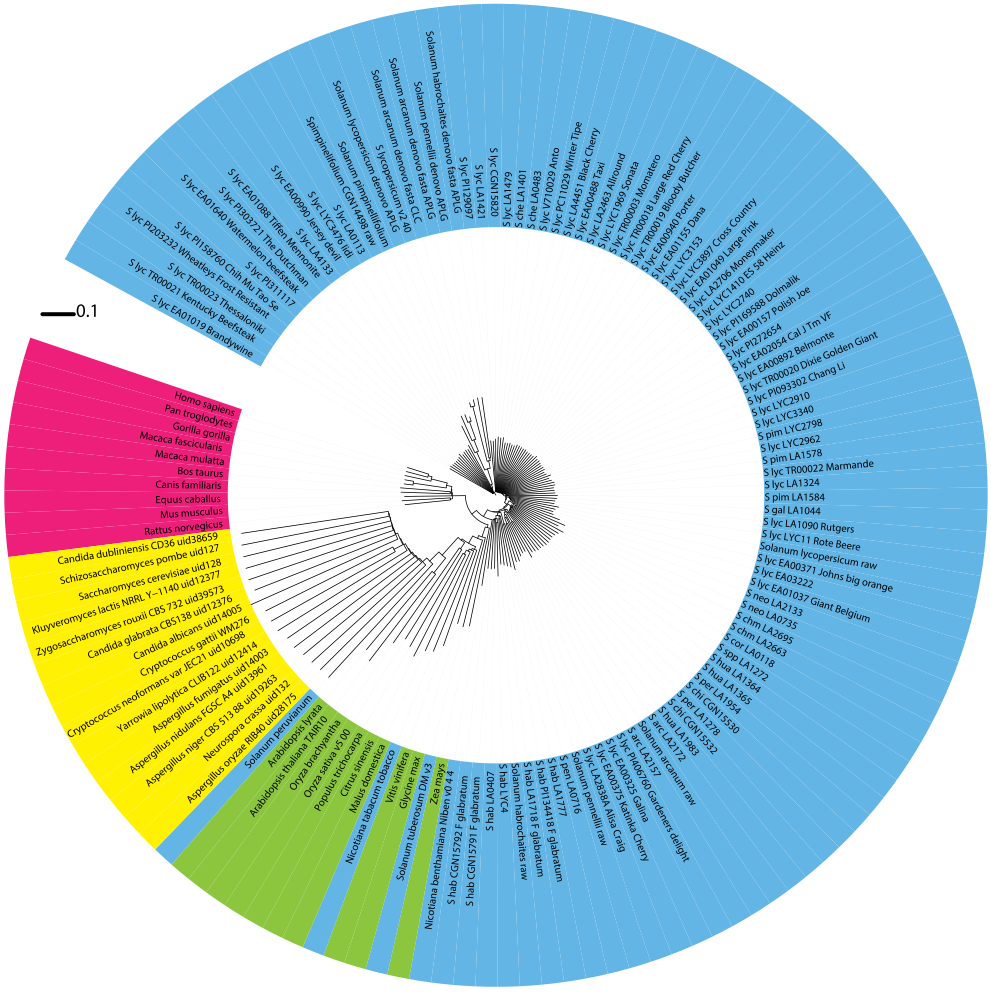


Figure 4: Phylogenetic trees from 135 samples. 11-, 15-, 17-, 21- and 31-mers are shown in graphs A, B, C, D and E, respectively. I) tree with phylogenetic distances; II) tree without phylogenetic distances, showing the clustering more clearly; trees plotted using iTOL (Letunic and Bork, 2007).

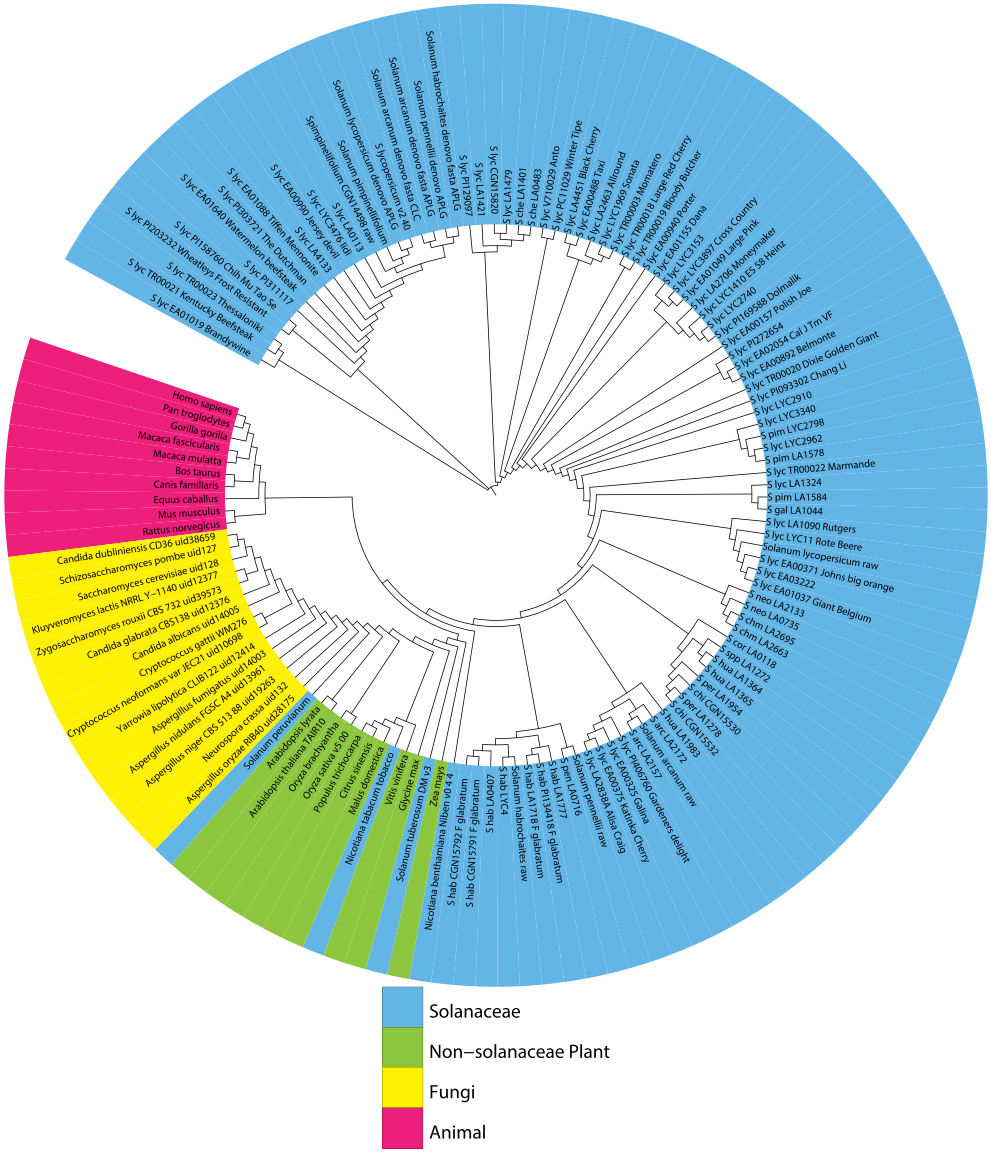
4.A.II



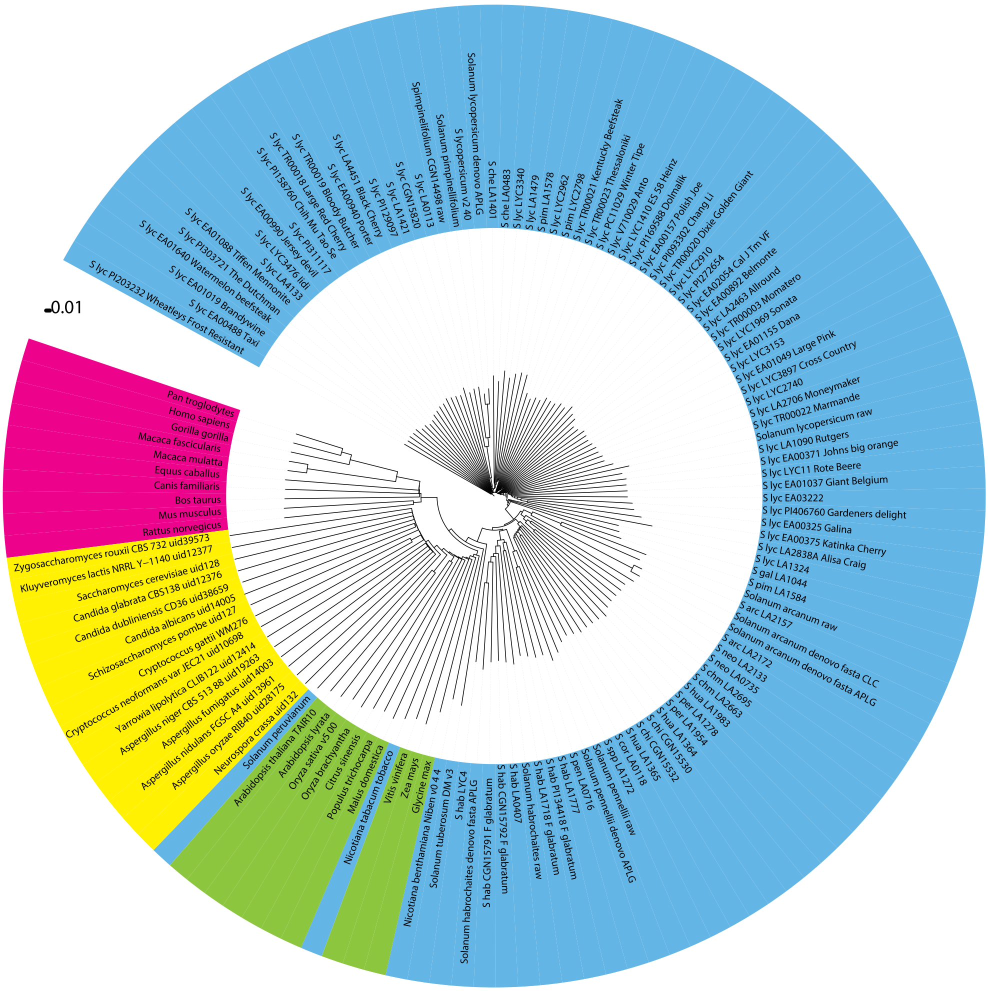
4.B.I


0.1
Homo sapiens
Pan troglodytes
Gorilla gorilla
Macaca fascicularis
Macaca mulatta
Bos taurus
Canis familiaris
Equus caballus
Mus musculus
Rattus norvegicus
Candida dubliniensis CD36 uid38659
Schizosaccharomyces pombe uid128
Saccharomyces cerevisiae uid128
Kluyveromyces lactis NRRL Y-1140 uid12377
Zygosaccharomyces rouxii CBS 732 uid39573
Candida glabrata CBS138 uid12376
Candida albicans WM276 uid14005
Cryptococcus gattii uid10698
Cryptococcus neoformans var. JEC21 uid12414
Yarrowia lipolytica CLIB122 uid14003
Aspergillus fumigatus FGSC A4 uid15961
Aspergillus nidulans FGSC A4 uid19263
Aspergillus niger CBS 513 88 uid28575
Aspergillus oryzae RIB40 uid28175
Neurospora crassa uid1132
Solanum pennellii
Arabidopsis lyrata
Arabidopsis thaliana TAIR10
Oryza brachyantha
Oryza sativa v5.00
Populus trichocarpa
Citrus sinensis
Malus domestica
Nicotiana tabacum tobacco
Vitis vinifera
Glycine max
Solanum tuberosum DM v3.3
Zea mays
Nicotiana benthamiana Niben v0.4.4
S hab CGN15792 F glabratum
S hab CGN15791 F glabratum
S hab LA0407
S hab LYC4
Solanum habrochaites raw
S hab LA1718 F glabratum
S hab PI134418 F glabratum
S hab LA0777
S pen LA0716
Solanum pennellii raw
S lyc LA2838A Alisa Craig
S lyc EA00375 Sasaka Cherry
S lyc EA00325 Golina
S lyc PI406760 Gardeners delight
Solanum arcanum raw
S arc LA2172
S chi LA1983
S chi CGN15332
S chi CGN15330
S per LA1954
S per LA1365
S hua LA1364
S spp LA1272
S cor LA0118
S chm LA2695
S chm LA2663
S neo LA0735
S neo LA2133
S lyc EA01037 Giant Belgium
S lyc EA03222
S lyc EA00371 Johns big orange
Solanum lycopersicum raw
S lyc EA00893 Belmonte
S lyc EA00020 Dixie Golden Giant
S lyc TR00020 Chang Li
S lyc PI093302 Chang Li
S lyc LYC2910
S lyc LYC3340
S pim LYC2798
S lyc LYC2962
S pim LA1578
S lyc TR00022 Marmande
S lyc LA1324
S pim LA1584
S gal LA1044
S lyc LA1090 Rutgers
S lyc LYC11 Rote Beere
S lyc EA00371 Johns big orange
S lyc EA03222
S lyc EA01037 Giant Belgium
S neo LA2133
S neo LA0735
S chm LA2663
S chm LA2695
S cor LA0118
S spp LA1272
S hua LA1364
S per LA1365
S per LA1954
S chi CGN15330
S chi CGN15332
S chi LA1983
S arc LA2172
Solanum arcanum raw
S lyc LA0157
S lyc PI406760 Gardeners delight
S lyc EA00325 Golina
S lyc EA00375 Sasaka Cherry
S lyc LA2838A Alisa Craig
Solanum pennellii raw
S pen LA0716
S hab LA1777
S hab PI134418 F glabratum
S hab LA1718 F glabratum
Solanum habrochaites raw
S hab LA0407
S hab LYC4
Zea mays
Nicotiana benthamiana Niben v0.4.4
Solanum tuberosum DM v3.3
Glycine max
Vitis vinifera
Nicotiana tabacum tobacco
Malus domestica
Citrus sinensis
Populus trichocarpa
Oryza sativa v5.00
Oryza brachyantha
Arabidopsis thaliana TAIR10
Arabidopsis lyrata
Solanum pennellii
Neurospora crassa uid1132
Aspergillus oryzae RIB40 uid28175
Aspergillus niger CBS 513 88 uid28575
Aspergillus nidulans FGSC A4 uid19263
Aspergillus fumigatus FGSC A4 uid15961
Yarrowia lipolytica CLIB122 uid14003
Cryptococcus neoformans var. JEC21 uid12414
Cryptococcus gattii uid10698
Candida albicans WM276 uid14005
Candida glabrata CBS138 uid12376
Zygosaccharomyces rouxii CBS 732 uid39573
Kluyveromyces lactis NRRL Y-1140 uid12377
Saccharomyces cerevisiae uid128
Schizosaccharomyces pombe uid128
Candida dubliniensis CD36 uid38659
Rattus norvegicus
Mus musculus
Equus caballus
Canis familiaris
Bos taurus
Macaca mulatta
Macaca fascicularis
Gorilla gorilla
Pan troglodytes
Homo sapiens
S lyc EA01019 Brandywine
S lyc TR00021 Kentucky Beefsteak
S lyc TR00023 Thessaloniki
S lyc PI203232 Wheatleys Frost Resistant
S lyc PI158760 Chih Hu Tao Se
S lyc PI261177
S lyc EA01640 Watermelon Beefsteak
S lyc EA01088 Tiffen Mentonite
S lyc PI303271 The Dutchman
S lyc EA00990 Jersey Devil
S lyc LA4132
S lyc LYC4970 Ukl
S lyc LA0713
Solanum lycopersicum CGN14988 raw
Solanum pimpinellifolium
Spiniperiellifolium CGN14988 raw
S lycopersicum denovo APLG
Solanum arcanum denovo fasta CLC
Solanum arcanum denovo fasta APLG
Solanum pimpinellifolium denovo fasta APLG
Solanum habrochaites denovo fasta APLG
S lyc LA1479
S lyc LA1401
S che LA0483
S lyc V710029 Anto
S lyc PCI 10139 Winter Tipe
S lyc LA4451 Black Cherry
S lyc EA00488 Taxi
S lyc LA2463 Allround
S lyc LYC1909 Sonata
S lyc TR00003 Monostero
S lyc TR00018 Large Red Cherry
S lyc EA00940 Porter
S lyc EA01155 Dana
S lyc LYC3153
S lyc LYC3887 Cross Country
S lyc EA01040 Large Pink
S lyc LA4706 Moneymaker
S lyc LYC1410 ES 58 Heinz
S lyc LYC2740
S lyc PI169588 Dolmalik
S lyc LYC2740
S lyc EA00157 Polish Joe
S lyc PI272654
S lyc EA02054 Cal J Tm VF
S lyc EA00893 Belmonte
S lyc EA00020 Dixie Golden Giant
S lyc TR00020 Chang Li
S lyc PI093302 Chang Li
S lyc LYC2910
S lyc LYC3340
S pim LYC2798
S lyc LYC2962
S pim LA1578
S lyc TR00022 Marmande
S lyc LA1324
S pim LA1584
S gal LA1044
S lyc LA1090 Rutgers
S lyc LYC11 Rote Beere
Solanum lycopersicum raw
S lyc EA00371 Johns big orange
S lyc EA03222
S lyc EA01037 Giant Belgium
S neo LA2133
S neo LA0735
S chm LA2663
S chm LA2695
S cor LA0118
S spp LA1272
S hua LA1364
S per LA1365
S per LA1954
S chi CGN15330
S chi CGN15332
S chi LA1983
S arc LA2172
Solanum arcanum raw
S lyc LA0157
S lyc PI406760 Gardeners delight
S lyc EA00325 Golina
S lyc EA00375 Sasaka Cherry
S lyc LA2838A Alisa Craig
Solanum pennellii raw
S pen LA0716
S hab LA1777
S hab PI134418 F glabratum
S hab LA1718 F glabratum
Solanum habrochaites raw
S hab LA0407
S hab LYC4
Zea mays
Nicotiana benthamiana Niben v0.4.4
Solanum tuberosum DM v3.3
Glycine max
Vitis vinifera
Nicotiana tabacum tobacco
Malus domestica
Citrus sinensis
Populus trichocarpa
Oryza sativa v5.00
Oryza brachyantha
Arabidopsis thaliana TAIR10
Arabidopsis lyrata
Solanum pennellii

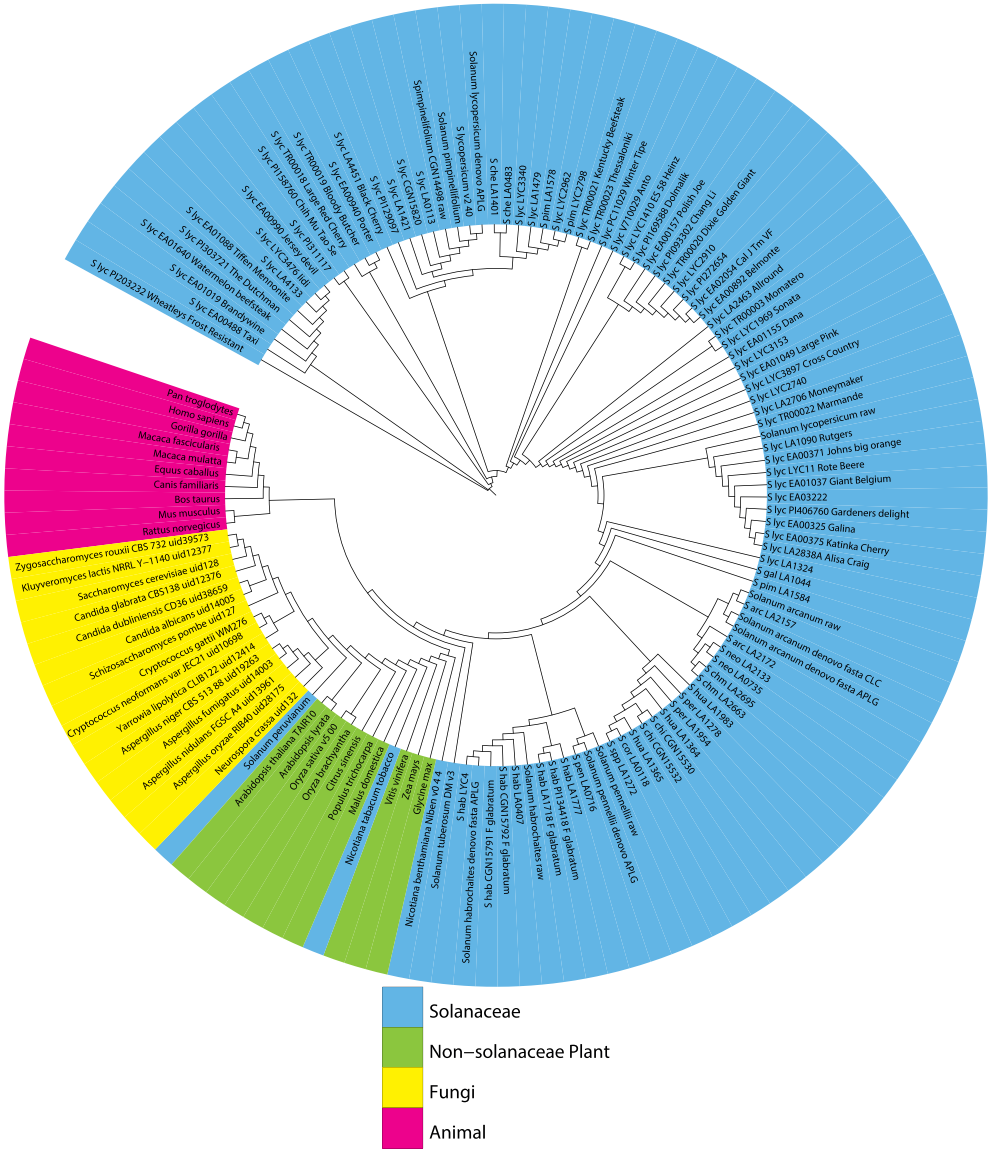
4.B.II



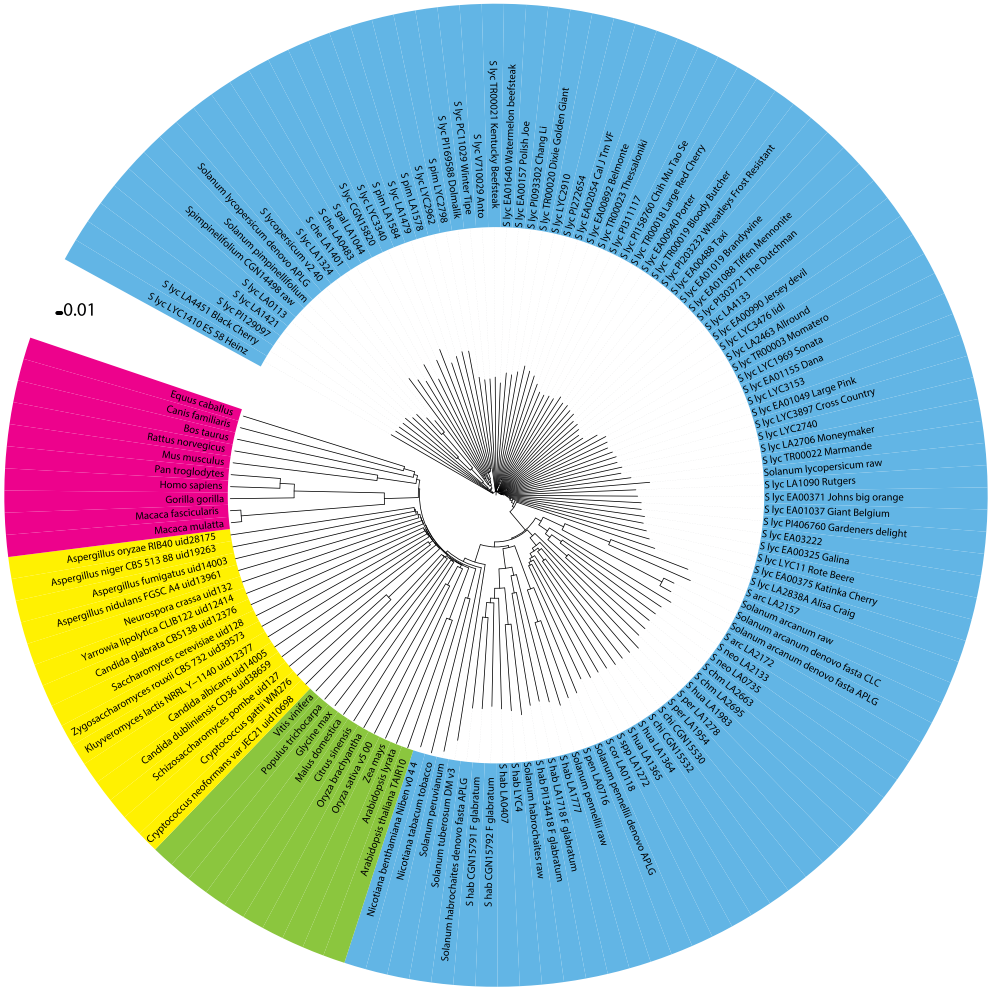
4.C.I


0.01
S lyc PI203232 Wheatleys Frost Resistant
S lyc EA01640 Watermelon beefsteak
S lyc EA01019 Brandywine
S lyc EA00488 Taxi
S lyc PI303721 The Dutchman
S lyc EA01088 Tiffen Mennonite
S lyc LA4133
S lyc EA00990 Jersey devil
S lyc LYC3476 Iidi
S lyc PI158760 Chih Mu Tao Se
S lyc PI311117
S lyc TR00018 Large Red Cherry
S lyc TR00019 Bloody Butcher
S lyc EA00940 Porter
S lyc LA4451 Black Cherry
S lyc PI129097
S lyc LA1421
S lyc CGN15820
S lyc LA0113
Spimpinellifolium CGN14498 raw
Solanum pimpinellifolium
S lycopersicum v2 40
Solanum lycopersicum denovo APLG
S che LA1401
S che LA0483
S lyc LYC3340
S lyc LA1479
S pim LA1578
S lyc LYC2962
S pim LYC2798
S lyc TR00021 Kentucky Beefsteak
S lyc TR00023 Thessaloniki
S lyc PC11029 Winter Tipe
S lyc V710029 Anto
S lyc LYC1410 ES 58 Heinz
S lyc PI169588 Dolmalik
S lyc EA00157 Polish Joe
S lyc PI093302 Chang Li
S lyc TR00020 Dixie Golden Giant
S lyc LYC2910
S lyc PI272654
S lyc EA02054 Cal J Tm VF
S lyc EA00892 Belmonte
S lyc LA2463 Allround
S lyc TR00003 Momatero
S lyc LYC1969 Sonata
S lyc EA01155 Dana
S lyc LYC3153
S lyc EA01049 Large Pink
S lyc LYC3897 Cross Country
S lyc LYC2740
S lyc LA2706 Moneymaker
S lyc TR00022 Marmande
Solanum lycopersicum raw
S lyc LA1090 Rutgers
S lyc EA00371 Johns big orange
S lyc LYC11 Rote Beere
S lyc EA01037 Giant Belgium
S lyc EA03222
S lyc PI406760 Gardeners delight
S lyc EA00325 Galina
S lyc EA00375 Katinka Cherry
S lyc LA2838A Alisa Craig
S lyc LA1324
S gal LA1044
S pim LA1584
Solanum arcanum raw
S arc LA2157
Solanum arcanum denovo fasta CLC
Solanum arcanum denovo fasta APLG
S arc LA2172
S neo LA2133
S neo LA0735
S chm LA2695
S chm LA2663
S hua LA1983
S per LA1278
S per LA1954
S hua LA1364
S chi CGN15532
S chi CGN15530
S hua LA1365
S cor LA0118
S spp LA1272
Solanum pennellii raw
Solanum pennellii denovo APLG
S pen LA0716
S hab LA1777
S hab PI134418 F glabratum
S hab LA1718 F glabratum
Solanum habrochaites raw
S hab LA0407
S hab CGN15792 F glabratum
S hab CGN15791 F glabratum
Solanum habrochaites denovo fasta APLG
S hab LYC4
Solanum tuberosum DM v3
Nicotiana benthamiana Niben v0 4 4
Glycine max
Zea mays
Vitis vinifera
Nicotiana tabacum tobacco
Malus domestica
Populus trichocarpa
Citrus sinensis
Oryza brachyantha
Oryza sativa v5 00
Arabidopsis lyrata
Arabidopsis thaliana TAIR10
Solanum peruvianum
Neurospora crassa uid132
Aspergillus oryzae RIB40 uid28175
Aspergillus nidulans FGSC A4 uid13961
Aspergillus fumigatus uid14003
Aspergillus niger CBS 513 88 uid19263
Yarrowia lipolytica CLIB122 uid12414
Cryptococcus neoformans var JEC21 uid10698
Cryptococcus gattii WM276
Schizosaccharomyces pombe uid127
Candida albicans uid14005
Candida dubliniensis CD36 uid38659
Candida glabrata CBS138 uid12376
Saccharomyces cerevisiae uid128
Kluyveromyces lactis NRRL Y−1140 uid12377
Zygosaccharomyces rouxii CBS 732 uid39573
Rattus norvegicus
Mus musculus
Bos taurus
Canis familiaris
Equus caballus
Macaca mulatta
Macaca fascicularis
Gorilla gorilla
Homo sapiens
Pan troglodytes

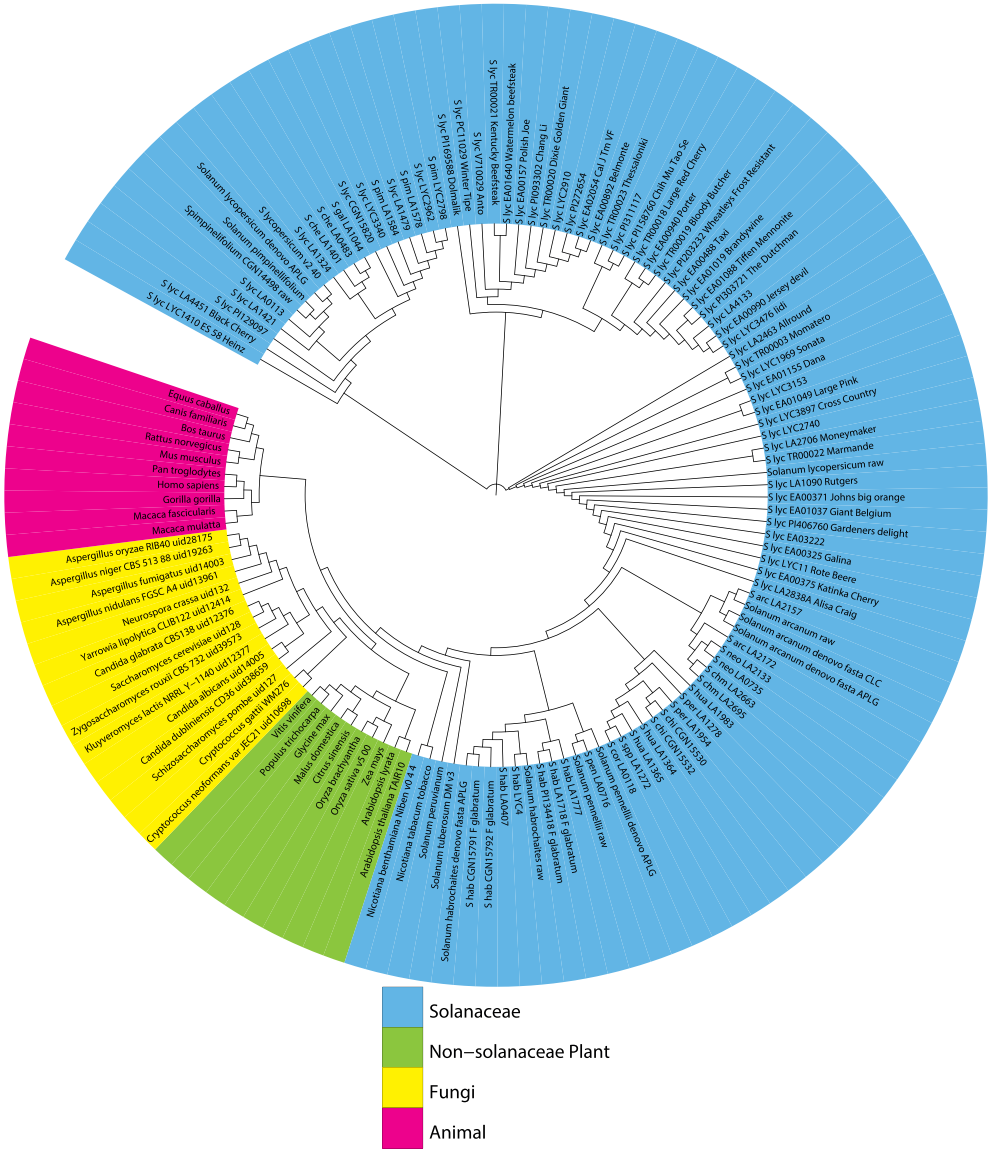
4.C.II



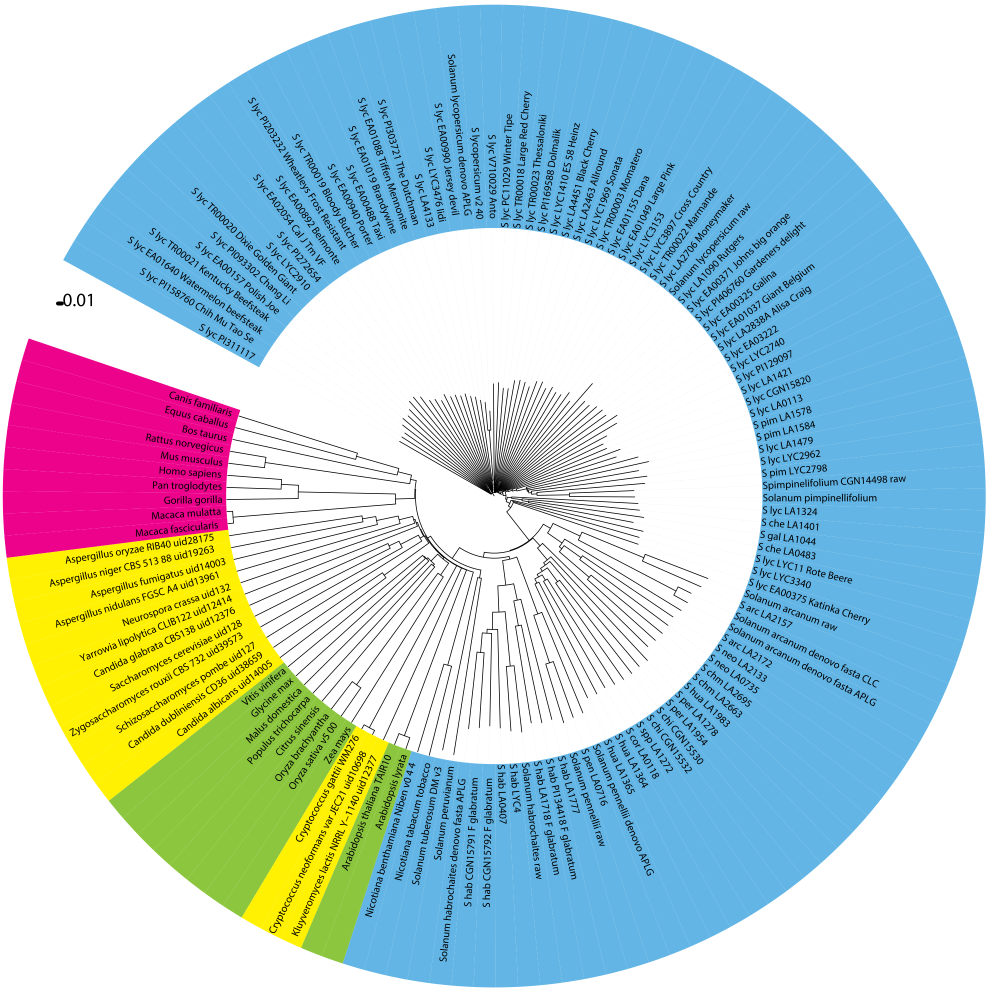
4.D.I


0.01
Equus caballus
Canis familiaris
Bos taurus
Rattus norvegicus
Mus musculus
Pan troglodytes
Homo sapiens
Gorilla gorilla
Macaca fascicularis
Macaca mulatta
Aspergillus oryzae RIB40 uid28175
Aspergillus niger CBS 513 88 uid14003
Aspergillus fumigatus uid13961
Aspergillus nidulans FGSC A4 uid13132
Neurospora crassa uid12414
Yarrowia lipolytica CLIB122 uid12376
Candida glabrata CBS138 uid12328
Saccharomyces cerevisiae uid39573
Zygosaccharomyces rouxii CBS 732 uid12377
Kluyveromyces lactis NRRL Y–1140 uid14005
Candida albicans SC5314 uid38659
Candida dubliniensis CD36 uid38658
Schizosaccharomyces pombe uid127
Cryptococcus gattii WM276
Cryptococcus neoformans var JEC21 uid10698
Vitis vinifera
Populus trichocarpa
Glycine max
Malus domestica
Citrus sinensis
Oryza brachyantha
Oryza sativa cv 09
Zea mays
Arabidopsis lyrata
Arabidopsis thaliana TAIR10
Nicotiana benthamiana Nitben v0.4.4
Nicotiana tabacum tobacco
Solanum peruvianum
Solanum tuberosum DM v3
Solanum habrochaites denovo fasta APLG
S hab GCN15791 F glabratum
S hab GCN15792 F glabratum
S hab LA0407
S hab LA1777
S hab PI134418 F glabratum
S hab LA1718 F glabratum
Solanum habrochaites raw
Solanum pennellii denovo APLG
Solanum pennellii raw
S pen A0716
S cor LA0118
S lyc LA1272
S hua LA1365
S chi GCN15532
S pen LA1954
S hua LA1278
S chm LA2663
S neo LA0735
S arc LA2133
Solanum arcanum denovo fasta APLG
Solanum arcanum raw
S arc LA2172
S arc LA2157
S lyc LA2838A Alisa Craig
S lyc EA00375 Katinka Cherry
S lyc LYC11 Rote Beere
S lyc EA00325 Galina
S lyc EA03222
S lyc PI406760 Gardeners delight
S lyc LA1090 Rutgers
S lyc EA00371 Johns big orange
S lyc EA01037 Giant Belgium
Solanum lycopersicum raw
S lyc TR00022 Marmande
S lyc LA2706 Moneymaker
S lyc LYC2740
S lyc LYC3897 Cross Country
S lyc EA01049 Large Pink
S lyc LYC3153
S lyc EA01155 Dana
S lyc LYC1968 Sonata
S lyc TR00003 Momatero
S lyc LA2463 Allround
S lyc LYC3476 Idi
S lyc EA00990 Jersey devil
S lyc LA4133
S lyc PI303721 The Dutchman
S lyc EA01088 Tiffen Mennonite
S lyc EA00488 Taxi
S lyc EA01010 Brandywine
S lyc PI203232 Wheatleys Frost Resistant
S lyc EA00940 Porter
S lyc TR00018 Large Red Cherry
S lyc EA00919 Bloody Butcher
S lyc EA00893 Belmonte
S lyc TR00023 Thessaloniki
S lyc PI311117
S lyc PI126726 Chih Mu Tao Se
S lyc EA00265 Cal J Trn VF
S lyc PI272654
S lyc LYC2910
S lyc TR00020 Dixie Golden Giant
S lyc PI093302 Chang Li
S lyc EA01157 Polish Joe
S lyc EA01640 Watermelon beefsteak
S lyc TR00027 Kentucky Beefsteak
S lyc LYC1070 Ailsa
S lyc PI169588 Dombito
S lyc PC11029 Winter Tank
S pim LYC2398
S lyc LYC2862
S pim LA1578
S lyc LA1479
S pim LA1384
S lyc LYC3340
S lyc LYC3680
S gal LA1044
S chc LA0483
S chc LA1401
S chc LA1324
S lyc LA1240
Solanum lycopersicum denovo APLG
Solanum pimpinellifolium
Spimpinellifolium CGN14488 raw
S lyc LA0113
S lyc LA1421
S lyc PI129097
S lyc LA4451 Black Cherry
S lyc LYC1410 ES 58 Heinz

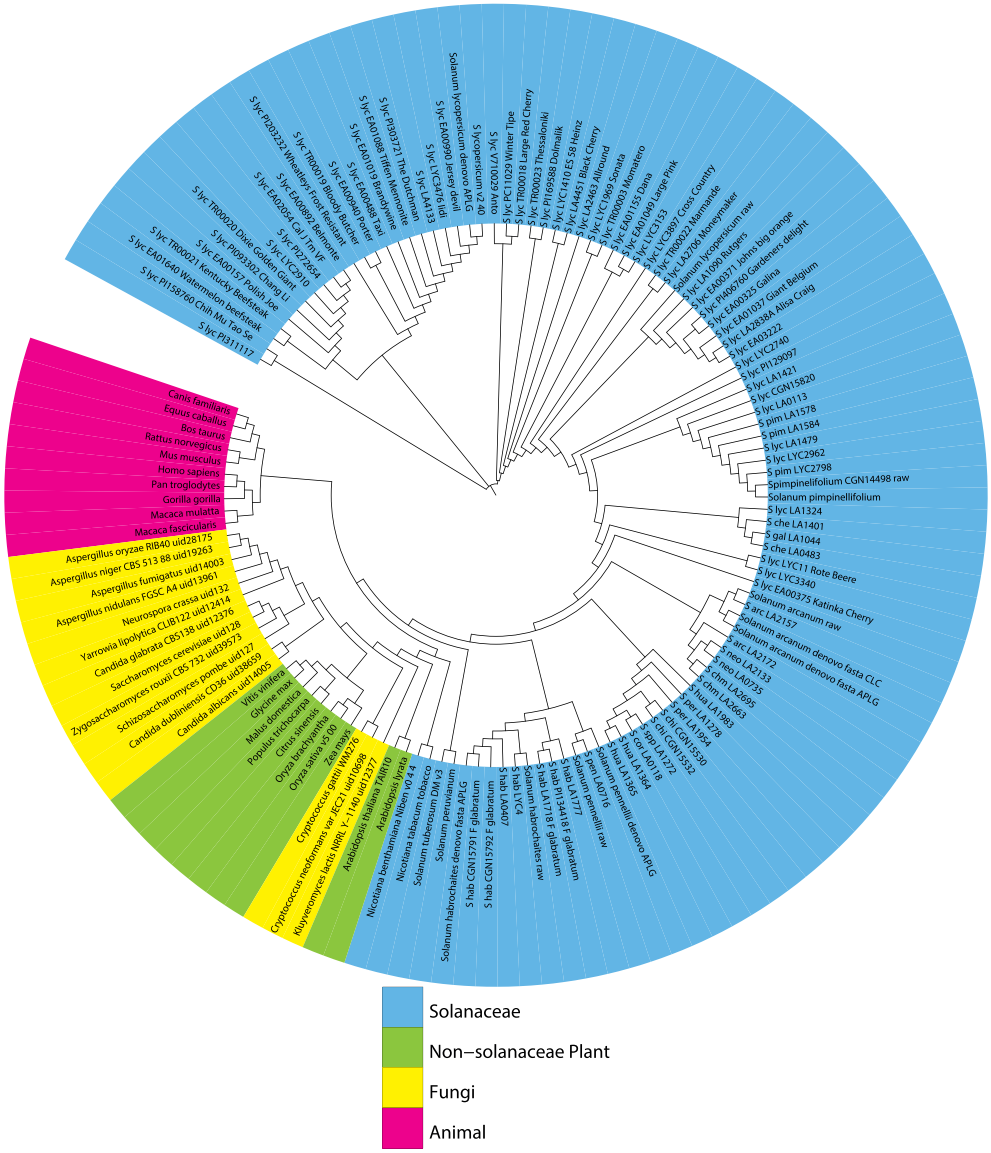
4.D.II



4.E.I


0.01
Canis familiaris
Equus caballus
Bos taurus
Rattus norvegicus
Mus musculus
Homo sapiens
Pan troglodytes
Gorilla gorilla
Macaca mulatta
Macaca fascicularis
Aspergillus oryzae RIB40 uid28175
Aspergillus niger CBS 513 88 uid19263
Aspergillus fumigatus uid14003
Aspergillus nidulans FGSC A4 uid13961
Neurospora crassa uid132
Yarrowia lipolytica CLIB122 uid12414
Candida glabrata CBS138 uid12376
Saccharomyces cerevisiae uid128
Zygosaccharomyces rouxii CBS 732 uid39573
Schizosaccharomyces pombe uid127
Candida dubliniensis CD36 uid38659
Candida albicans uid14005
Vitis vinifera
Glycine max
Malus domestica
Populus trichocarpa
Citrus sinensis
Oryza brachyantha
Oryza sativa v5 00
Zea mays
Cryptococcus gattii WM276
Cryptococcus neoformans var JEC21 uid10698
Kluyveromyces lactis NRRL Y−1140 uid12377
Arabidopsis thaliana TAIR10
Arabidopsis lyrata
Nicotiana benthamiana Niben v0 4 4
Nicotiana tabacum tobacco
Solanum tuberosum DM v3
Solanum peruvianum
Solanum habrochaites denovo fasta APLG
S hab CGN15791 F glabratum
S hab CGN15792 F glabratum
S hab LA0407
S hab LYC4
Solanum habrochaites raw
S hab PI134418 F glabratum
S hab LA1718 F glabratum
S hab LA1777
Solanum pennellii raw
S per LA0716
Solanum pennellii denovo APLG
S hua LA1364
S hua LA1365
S cor LA0118
S spp LA1272
S chi CGN15532
S chi CGN15530
S per LA1278
S per LA1954
S hua LA1983
S chm LA2663
S chm LA2695
S neo LA0735
S neo LA2133
S arc LA2172
Solanum arcanum denovo fasta APLG
Solanum arcanum denovo fasta CLC
S arc LA2157
Solanum arcanum raw
S lyc EA00375 Katinka Cherry
S lyc LYC3340
S lyc LYC11 Rote Beere
S che LA0483
S gal LA1044
S che LA1401
S lyc LA1324
Solanum pimpinellifolium
Spimpinelifolium CGN14498 raw
S pim LYC2798
S lyc LYC2962
S lyc LA1479
S pim LA1584
S pim LA1578
S lyc LA0113
S lyc CGN15820
S lyc LA1421
S lyc PI129097
S lyc LYC2740
S lyc EA03222
S lyc LA2838A Alisa Craig
S lyc EA01037 Giant Belgium
S lyc EA00325 Galina
S lyc PI406760 Gardeners delight
S lyc EA00371 Johns big orange
S lyc LA1090 Rutgers
Solanum lycopersicum raw
S lyc LA2706 Moneymaker
S lyc TR00022 Marmande
S lyc LYC3897 Cross Country
S lyc LYC3153
S lyc EA01049 Large Pink
S lyc EA01155 Dana
S lyc TR00003 Momatero
S lyc LYC1969 Sonata
S lyc LA2463 Allround
S lyc LA4451 Black Cherry
S lyc LYC1410 ES 58 Heinz
S lyc PI169588 Dolmalik
S lyc TR00023 Thessaloniki
S lyc TR00018 Large Red Cherry
S lyc PC11029 Winter Tipe
S lyc V710029 Anto
S lycopersicum v2 40
Solanum lycopersicum denovo APLG
S lyc EA00990 Jersey devil
S lyc LYC3476 Ildi
S lyc LA4133
S lyc PI303721 The Dutchman
S lyc EA00488 Taxi
S lyc EA01088 Tiffen Mennonite
S lyc EA01019 Brandywine
S lyc EA00940 Porter
S lyc TR00019 Bloody Butcher
S lyc EA00892 Belmonte
S lyc PI203232 Wheatleys Frost Resistant
S lyc EA02054 Cal J Tm VF
S lyc PI272654
S lyc LYC2910
S lyc TR00020 Dixie Golden Giant
S lyc PI093302 Chang Li
S lyc EA00157 Polish Joe
S lyc TR00021 Kentucky Beefsteak
S lyc EA01640 Watermelon beefsteak
S lyc PI158760 Chih Mu Tao Se
S lyc PI311117

4.E.II



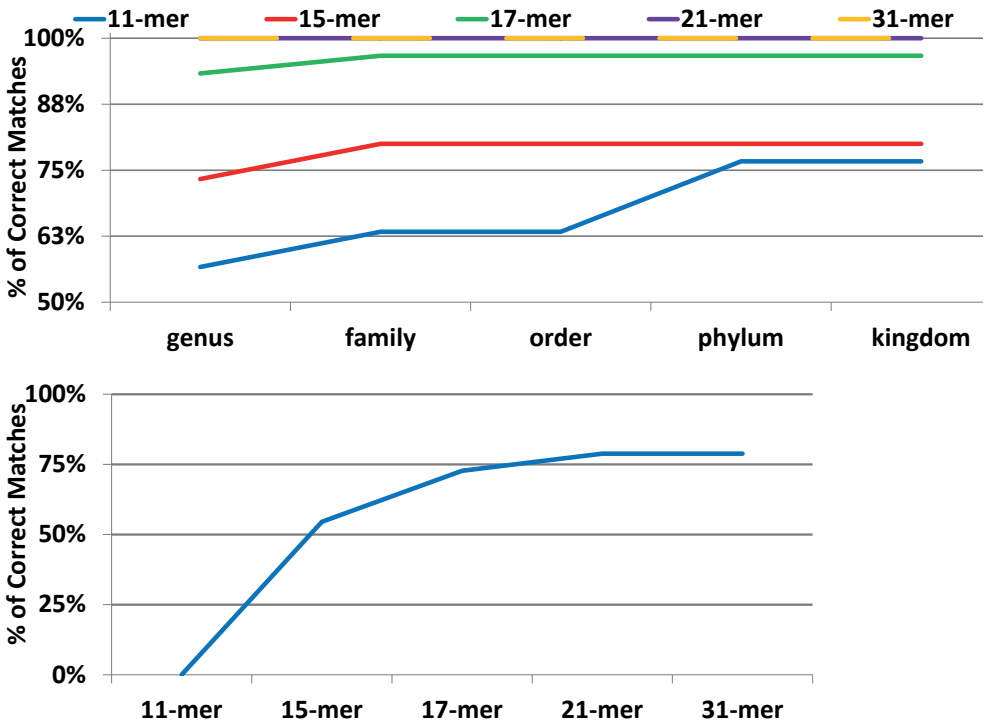


Figure 5: 1-nearest-neighbour analysis for the supra-species level and species level. Top shows supra-species level analysis of 30 samples, 8 genus, 7 families, 7 orders, 4 phylum and 3 kingdoms. Bottom shows species level analysis of 33 samples from 11 species of the Solanum clade. Classification reports the Leave-One-Out Cross-Validation error estimate (LOOCV), showing the improvement in accuracy with increased k-mer size and flattening of precision for k-mers above 21. Sample names and classes can be found in supplementary tables 3 and 4, respectively.

Speedup by subsampling

To test the influence of data set size (and possibility of speedup) and since 21-mers showed the best trade-off between speed and discriminating power (consistent with Cannon *et al.*, 2010), we sampled 2% of the dataset by analysing just 1 of the 50 pieces the data was originally split into. Figure 7 shows the phylogenetic placement of species in the trees constructed using this dataset. The tree is indistinguishable from the one generated on the full dataset, illustrating the ability of Cnidaria to correctly classify samples even at very low sequencing coverage. This suggests that Cnidaria should be able to correctly cluster and identify samples using small and affordable NGS sequencing technology such as Illumina MiSeq nano runs (500 Mbp in 2x250bp reads, Illumina, 2015).

Joint analysis of DNA and rna-seq data

Next, we expanded the 135 sample dataset (built using Database Creation Mode) with 34 extra samples, 26 genomic and 8 RNA-seq (Supplementary Table 1), using 21-mers and the faster Sample Analysis Mode. RNA-seq samples were added to verify whether transcriptome data would cluster with their genomic NGS counterparts, despite their small coverage of the genome length. Results are shown in Figure 8. The clustering of the original 135 samples is not changed and new samples cluster correctly according to their phylogeny. The consistent clustering observed for the RNA-seq dataset illustrates the ability of Cnidaria to use such data for accurate species identification.

Comparison with ReferenceFree

To demonstrate the advantages of Cnidaria, we compare it to a state-of-the-art alternative tool, proposed in (Cannon *et al.*, 2010). The latest version of the software introduced (version 1.1.3, hereafter referred to simply as ReferenceFree) was downloaded and run in conjunction with ABYSS (Simpson *et al.*, 2009). We used ABYSS version 1.3.3 rather than the latest version (1.9.0) since that was the last version tested with ReferenceFree. ReferenceFree was run single threaded on an Intel(R) Xeon(R) CPU E7- 4850 @ 2.00 GHz with a *k*-mer size of 21, a minimum frequency of 0 (i.e. using all *k*-mers appearing 1 or more times), no complexity filter and no sampling of *k*-mers. The list of shared *k*-mers generated was then parsed using the Cnidaria scripts in order to generate the phylogenetic tree.

Using a subset of our data (41 assembled genomes, Supplementary Table 1) containing 40 Gbp and 20 billion *k*-mers, ReferenceFree (Supplementary Table 1) and Jellyfish have a comparable speed for *k*-mer counting, taking 4 hours to count 445 million *k*-mers (2% of the total). ReferenceFree then took 60% more time than Cnidaria in single threaded Sample Analysis Mode for merging and summarizing the results (70 hours vs. 44 hours, respectively). Note that the databases created by Cnidaria can be re-used in subsequent comparisons, whereas ReferenceFree requires all the *k*-mer count files to be merged again when re-run. Moreover, Cnidaria has the important advantage of being highly parallelizable while ReferenceFree can only be run single threaded. The phylogenetic tree created by ReferenceFree can be found in figure 9.

Besides speed, Cnidaria (and Jellyfish) use significantly less space, due to their binary formats. They generate files which are smaller than the equivalent files created by ReferenceFree with median sizes of 9.2 Gb vs. 42.2 Gb (MADs of 2.5 Gb and 11.0 Gb, respectively) for the *k*-mer count file and 227 Gb vs. 2.1 Tb for the merged *k*-mer count file, despite the merged *k*-mer count file created by ReferenceFree contain only 2% of the total number of *k*-mers.

6A

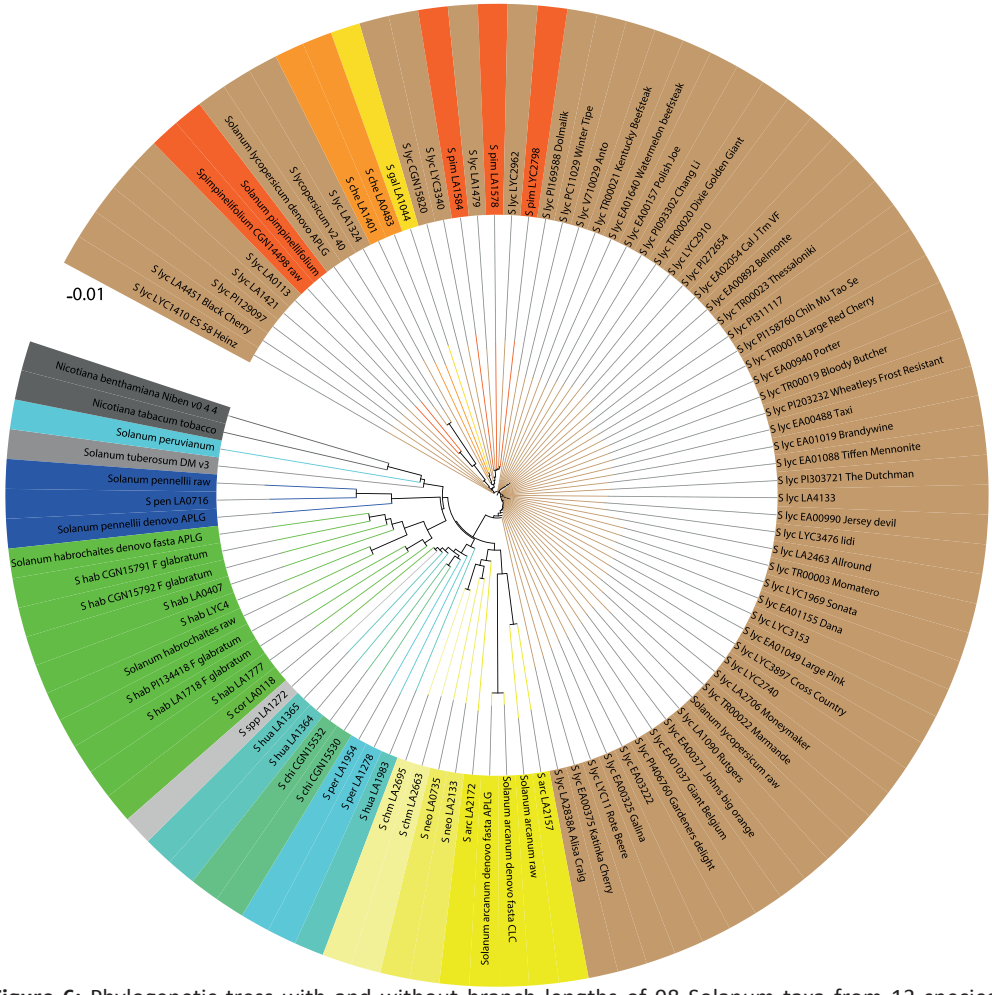
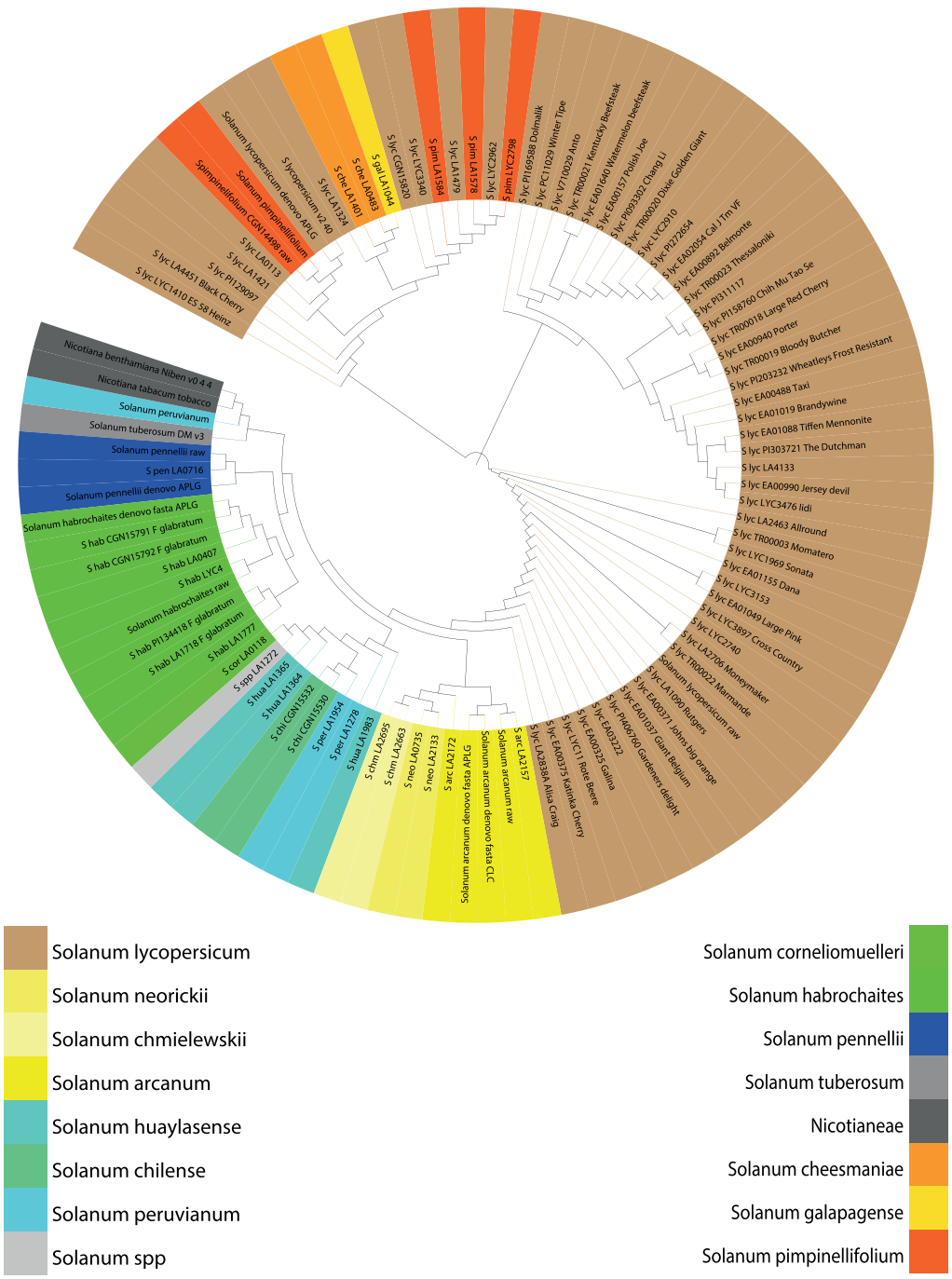


Figure 6: Phylogenetic tress with and without branch lengths of 98 Solanum taxa from 13 species. The Lycopersicon group (comprised of *Solanum lycopersicum*, *S. pimpinellifolium*, *S. cheesmaniae* and *S. galapagense*) clusters as a monophyletic group. Sometimes the non-*S. lycopersicum* species cluster inside the *S. lycopersicum* clade. We speculate these are *S. lycopersicum* varieties containing introgression clustering with the donor species, consistently with the findings of Aflitos *et al.* (2014). The Arcanum group (comprised of *S. arcanum*, *S. chmielewskii* and *S. neorikii*) also clusters monophyletically, closer to the Eriopersicon group, its sister group. The North Eriopersicon group (comprised of *S. huaylasense*, *S. chilense*, *S. peruvianum* and *S. corneliomulleri*) groups with the South Eriopersicon group (comprised of *S. habrochaites*, its only member) and its sister group, Neolycopersicon (comprised of *S. pennelli*, its only member). *S. tuberosum* and *S. nicotiana* were added as outgroups. Sample names ending in RAW are raw genomic data; names ending in APLG and CLC are assembled genomes. A) tree with phylogenetic distances; B) tree withtout phylogenetic distances, showing the clustering more clearly; Trees were plotted using iTOL (Letunic and Bork, 2007).

6B


S lyc LA4451 Black Cherry
S lyc LYC1410 ES 58 Heinz
S lyc LA0113
S lyc LA1421
S lyc PI129097
Solanum pimpinellifolium CGN14498 raw
Spimpinellifolium denovo APLG
Solanum lycopersicum denovo APLG
S lycopersicum LA0716
S lyc LA1324
S che LA1401
S gal LA1044
S lyc CGN15820
S lyc LYC3340
S pim LA1584
S lyc LA1479
S pim LA1578
S pim LYC2662
S pim LYC2798
S lyc PI169588 Dolmalik
S lyc PC-T110029 Winter Tipe
S lyc V7110029 Anto
S lyc TR00021 Kentucky Beefsteak
S lyc EA001640 Watermelon beefsteak
S lyc EA001157 Polish Joe
S lyc PI093302 Chang Li
S lyc TR00020 Dixie Golden Giant
S lyc LYC2910
S lyc PI272654
S lyc EA02054 Cal J Tm VF
S lyc EA00892 Belmonte
S lyc TR00023 Thessaloniki
S lyc PI311117
S lyc PI158760 Chih Mu Tao Se
S lyc TR00018 Large Red Cherry
S lyc EA00940 Porter
S lyc TR00019 Bloody Butcher
S lyc PI203232 Wheatleys Frost Resistant
S lyc EA00488 Taxi
S lyc EA01019 Brandywine
S lyc EA01088 Tiffen Mennonite
S lyc PI303721 The Dutchman
S lyc LA4133
S lyc EA00990 Jersey devil
S lyc LYC3476 Iidi
S lyc LA2463 Allround
S lyc TR00003 Momatero
S lyc LYC1969 Sonata
S lyc EA01155 Dana
S lyc LYC3153
S lyc EA01049 Large Pink
S lyc LYC3897 Cross Country
S lyc LYC2740
S lyc LA2706 Moneymaker
S lyc TR00022 Marmande
Solanum lycopersicum raw
S lyc LA1090 Rutgers
S lyc EA00371 Johns big orange
S lyc EA01037 Giant Belgium
S lyc Pi460760 Gartenees delight
S lyc EA00325 Galina
S lyc LYC1 Rote Beere
S lyc EA00375 Katinka Cherry
S lyc LA2838A Alisa Craig
S arc LA2157
Solanum arcanum raw
Solanum arcanum denovo fasta APLG
Solanum arcanum denovo fasta CLC
S arc LA2172
S neo LA2133
S neo LA0735
S chm LA2663
S chm LA2695
S hua LA1983
S per LA1278
S per LA1954
S chi CGN15530
S chi CGN15532
S hua LA1364
S hua LA1385
S spp LA1272
S cor LA0118
S hab LA1777
S hab LA1718 F glabratum
S hab PI134418 F glabratum
Solanum habrochaites raw
S hab LYC4
S hab LA0407
S hab CGN15792 F glabratum
S hab CGN15791 F glabratum
Solanum habrochaites denovo fasta APLG
Solanum pennellii denovo APLG
S pen LA0716
Solanum pennellii raw
Solanum tuberosum DM v3
Solanum peruvianum
Nicotiana tabacum tobacco
Nicotiana benthamiana Niben v0.4.4
Solanum lycopersicum
Solanum neorickii
Solanum chmielewskii
Solanum arcanum
Solanum huaylasense
Solanum chilense
Solanum peruvianum
Solanum spp
Solanum corneliomuelleri
Solanum habrochaites
Solanum pennellii
Solanum tuberosum
Nicotianeae
Solanum cheesmaniae
Solanum galapagense
Solanum pimpinellifolium

7A

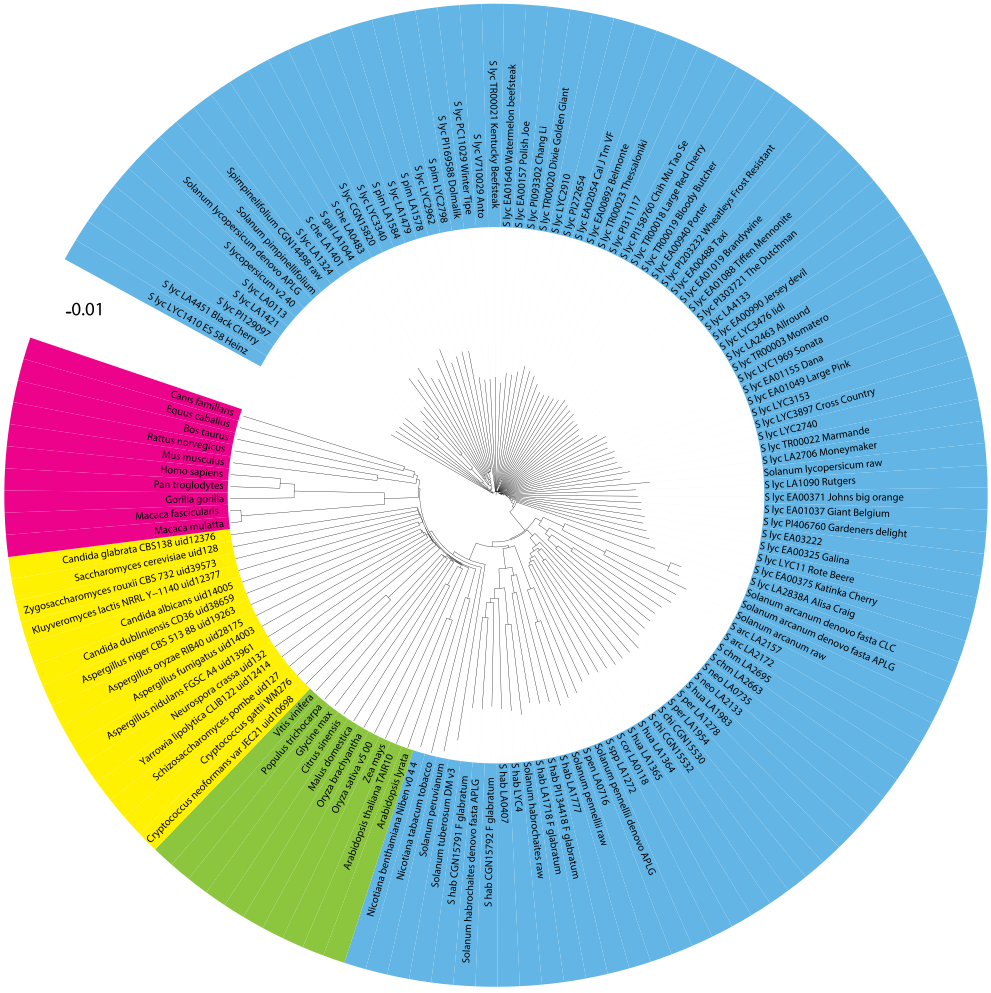
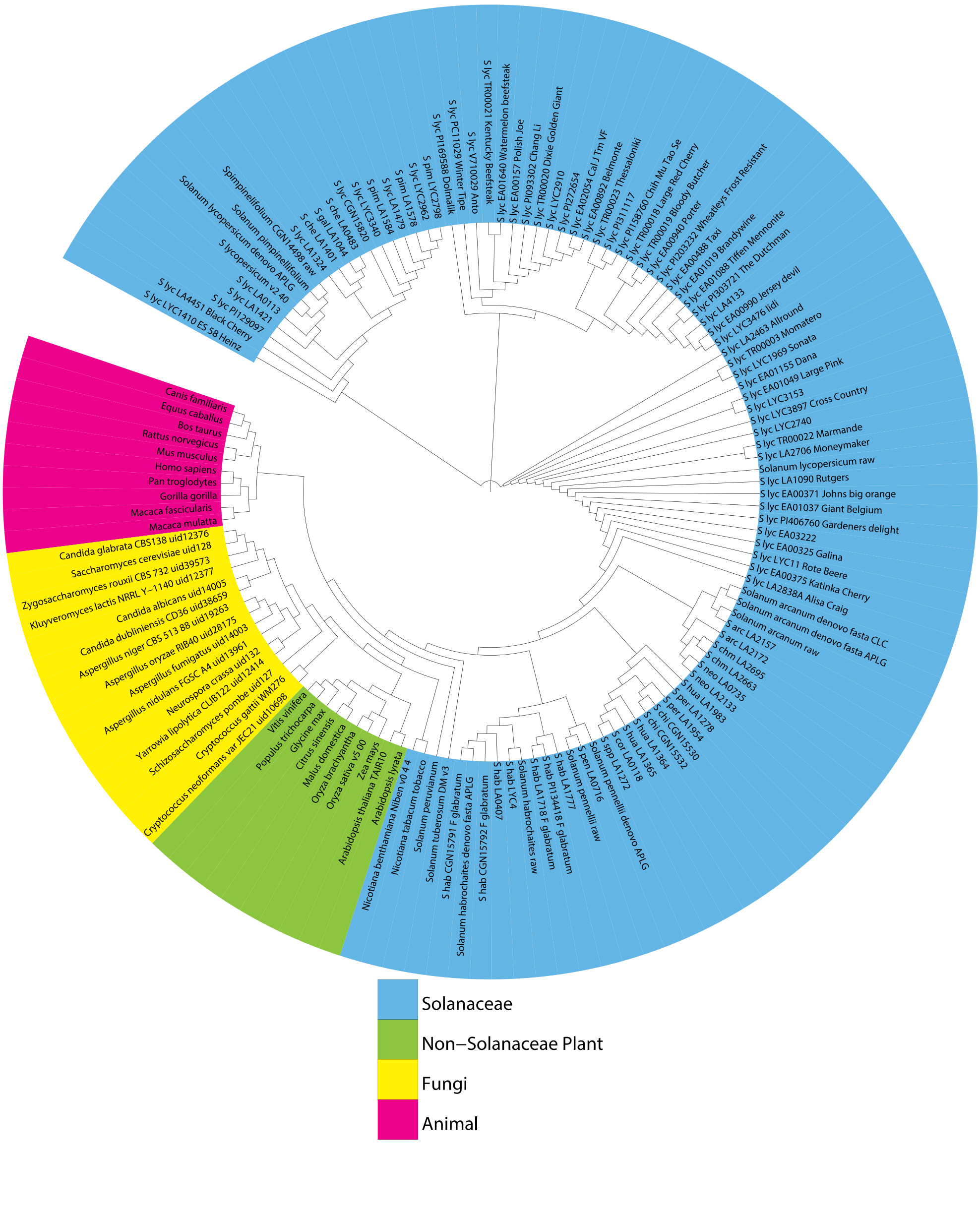
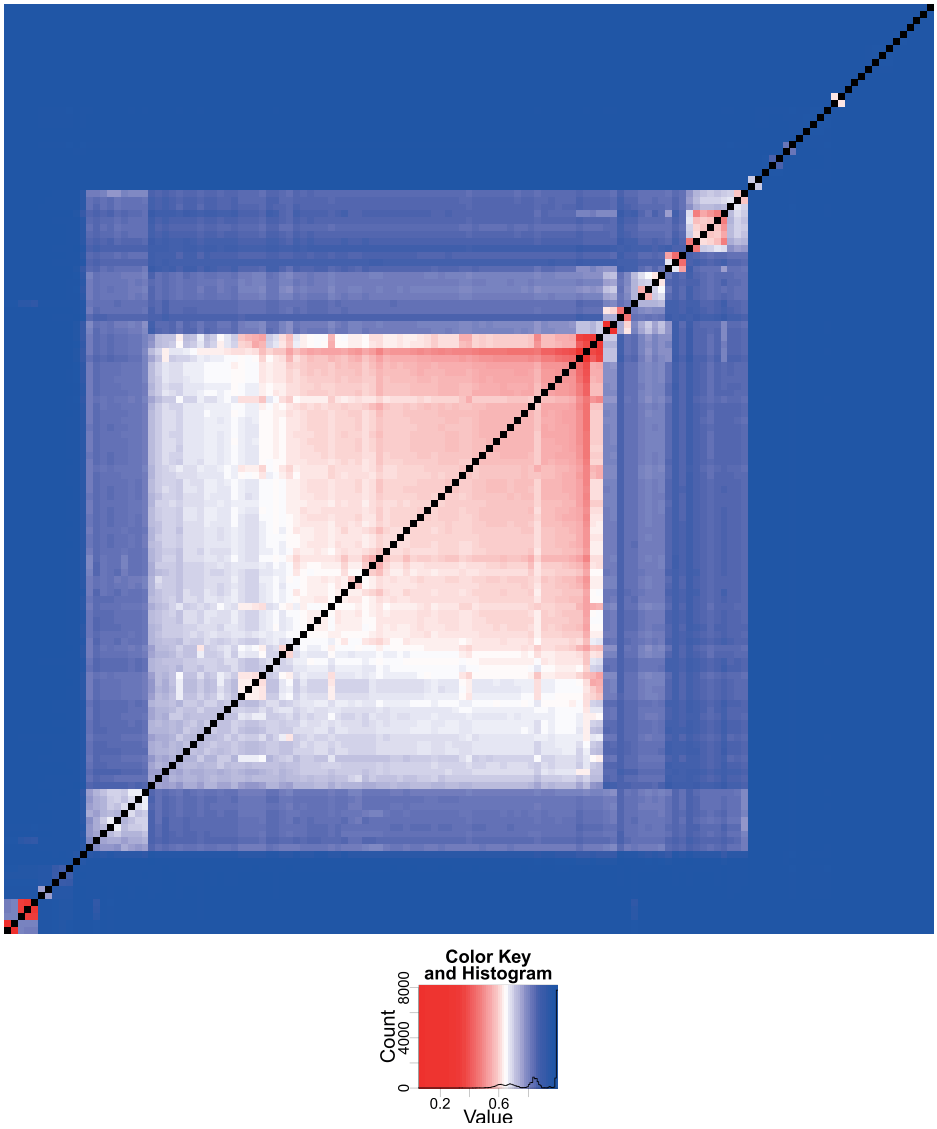


Figure 7: Results for 2% of the 21-mer dataset. A) phylogenetic tree with distance; B) phylogenetic tree without distance; C) heatmap of phylogenetic distances showing low inter-group similarity and high intra-group similarity; D) histogram of Jaccard distances showing the same feature of low inter-group similarity and high intra-group similarity. Trees were plotted using iTOL (Letunic and Bork, 2007).

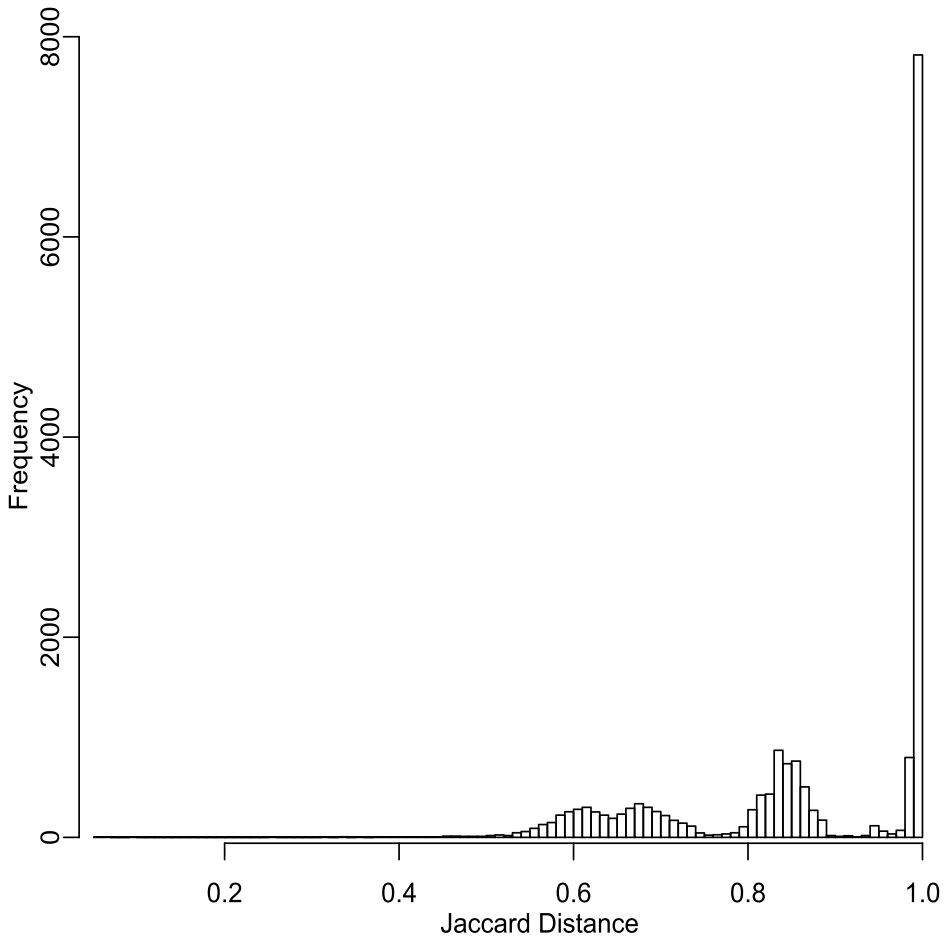
7B


S lyc LYC1410 ES 58 Heinz
S lyc LA4451 Black Cherry
S lyc PI129097
S lyc LA1421
S lyc LA0113
S lycopersicum v2 40
Solanum lycopersicum denovo APLG
Solanum pimpinellifolium
Spimpinellifolium CGN14498 raw
S lyc LA1324
S gal LA1401
S che LA1044
S che LA0483
S lyc CGN15820
S lyc LYC3340
S pim LA1584
S lyc LA1479
S pim LA1578
S lyc LYC2962
S pim LYC2798
S lyc PI169588 Dolnalik
S lyc PC11029 Winter Tipe
S lyc V710029 Anto
S lyc TR00021 Kentucky Beefsteak
S lyc EA01640 Watermelon beefsteak
S lyc EA00157 Polish Joe
S lyc PI093302 Chang Li
S lyc TR00020 Dixie Golden Giant
S lyc LYC2910
S lyc PI272654
S lyc EA02054 Cal J Tm VF
S lyc EA00892 Belmonte
S lyc TR00023 Thessaloniki
S lyc PI311117
S lyc PI158760 Chih Mu Tao Se
S lyc TR00018 Large Red Cherry
S lyc TR00019 Bloody Butcher
S lyc EA00940 Porter
S lyc PI203232 Wheatleys Frost Resistant
S lyc EA00488 Taxi
S lyc EA01019 Brandywine
S lyc EA01088 Tiffen Mennonite
S lyc PI303721 The Dutchman
S lyc LA4133
S lyc EA00990 Jersey devil
S lyc LYC3476 Iidi
S lyc LA2463 Allround
S lyc TR00003 Momatero
S lyc LYC1969 Sonata
S lyc EA01155 Dana
S lyc EA01049 Large Pink
S lyc LYC3153
S lyc LYC3897 Cross Country
S lyc LYC2740
S lyc TR00022 Marmande
S lyc LA2706 Moneymaker
Solanum lycopersicum raw
S lyc LA1090 Rutgers
S lyc EA00371 Johns big orange
S lyc EA01037 Giant Belgium
S lyc PI406760 Gardeners delight
S lyc EA03222
S lyc EA00325 Galina
S lyc LYC11 Rote Beere
S lyc EA00375 Katinka Cherry
S lyc LA2838A Alisa Craig
Solanum arcanum denovo fasta CLC
Solanum arcanum denovo fasta APLG
Solanum arcanum raw
S arc LA2157
S arc LA2172
S chm LA2695
S chm LA2663
S neo LA0735
S neo LA2133
S hua LA1983
S per LA1278
S per LA1954
S chi CGN15530
S chi CGN15532
S hua LA1364
S hua LA1365
S cor LA0118
S spp LA1272
Solanum pennellii denovo APLG
S pen LA0716
Solanum pennellii raw
S hab LA1777
S hab PI134418 F glabratum
S hab LA1718 F glabratum
Solanum habrochaites raw
S hab LYC4
S hab LA0407
S hab CGN15792 F glabratum
Solanum habrochaites denovo fasta APLG
S hab CGN15791 F glabratum
Solanum tuberosum DM v3
Solanum peruvianum
Nicotiana tabacum tobacco
Nicotiana benthamiana Niben v0 4 4
Arabidopsis lyrata
Arabidopsis thaliana TAIR10
Zea mays
Oryza sativa v5 00
Oryza brachyantha
Malus domestica
Citrus sinensis
Glycine max
Populus trichocarpa
Vitis vinifera
Cryptococcus neoformans var JEC21 uid10698
Cryptococcus gattii WM276
Schizosaccharomyces pombe uid127
Yarrowia lipolytica CLIB122 uid12414
Neurospora crassa uid132
Aspergillus nidulans FGSC A4 uid13961
Aspergillus fumigatus uid14003
Aspergillus oryzae RIB40 uid28175
Aspergillus niger CBS 513 88 uid19263
Candida dubliniensis CD36 uid38659
Candida albicans uid14005
Kluyveromyces lactis NRRL Y−1140 uid12377
Zygosaccharomyces rouxii CBS 732 uid39573
Saccharomyces cerevisiae uid128
Candida glabrata CBS138 uid12376
Macaca mulatta
Macaca fascicularis
Gorilla gorilla
Pan troglodytes
Homo sapiens
Mus musculus
Rattus norvegicus
Bos taurus
Equus caballus
Canis familiaris
Solanaceae
Non–Solanaceae Plant
Fungi
Animal

7C



7D


Frequency
0
2000
4000
6000
8000
0.2
0.4
0.6
0.8
1.0
Jaccard Distance

8A

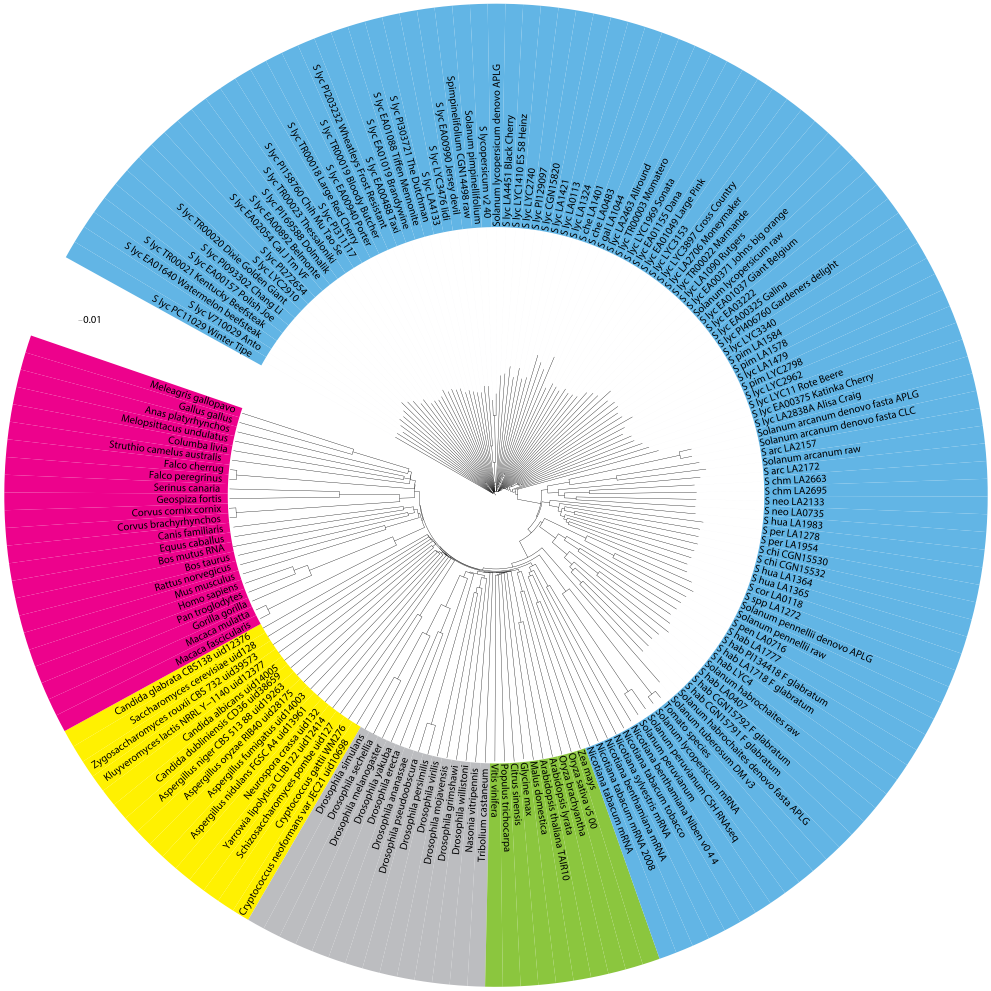
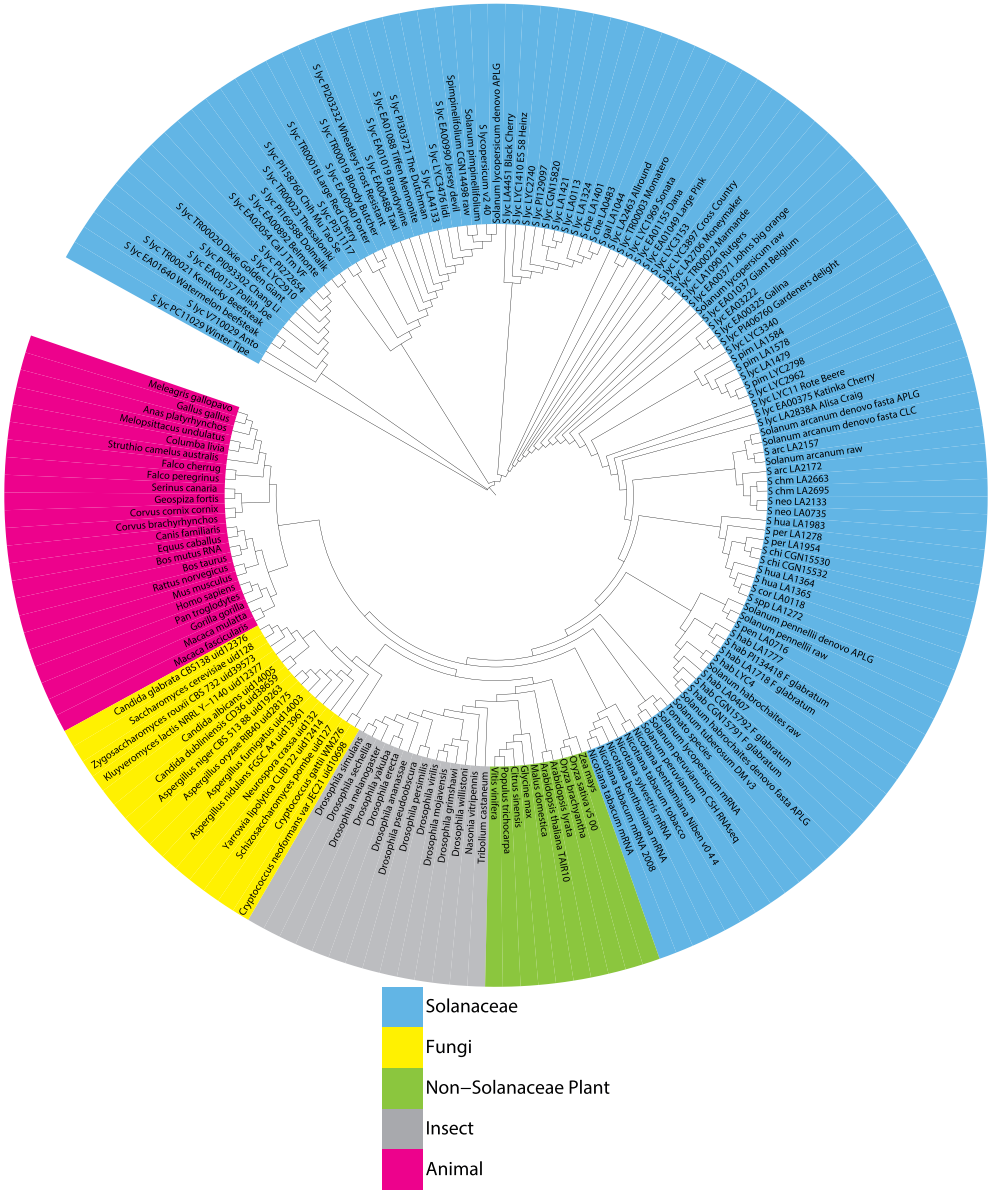
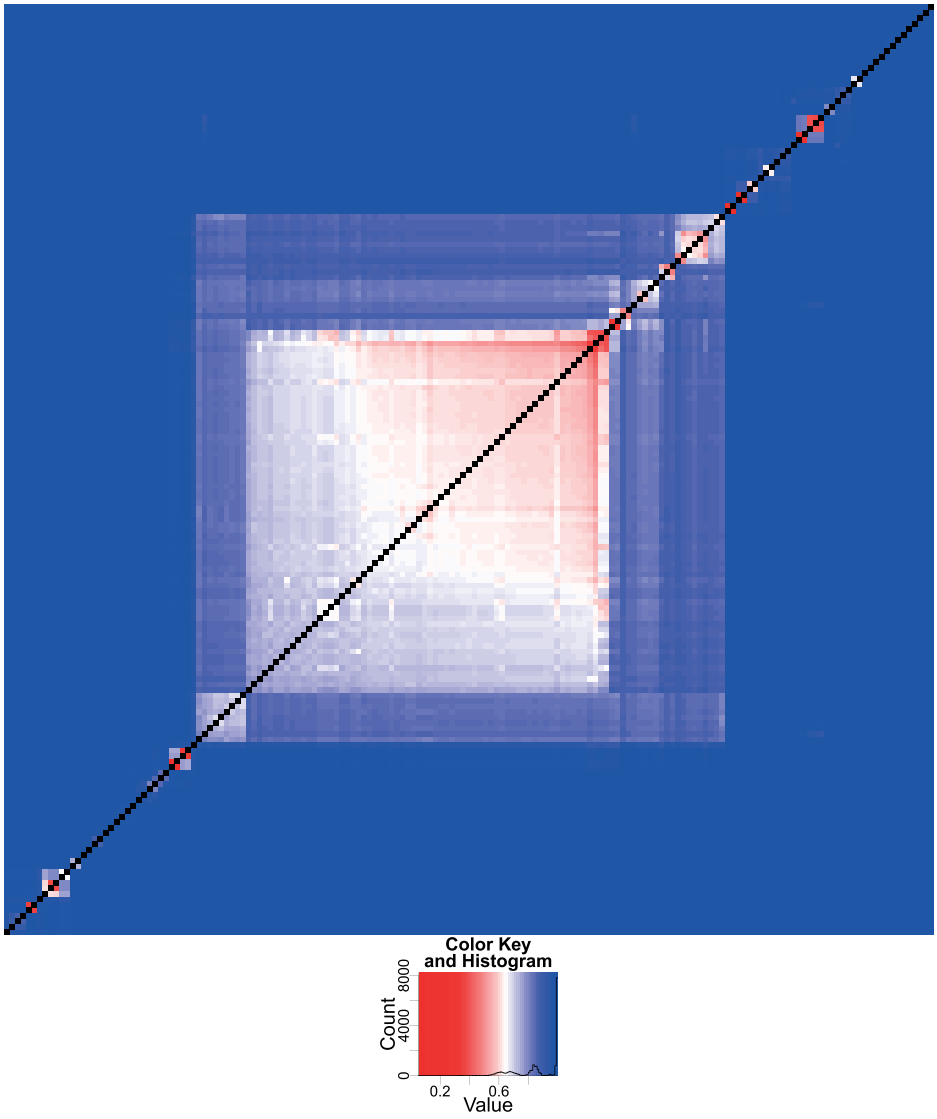


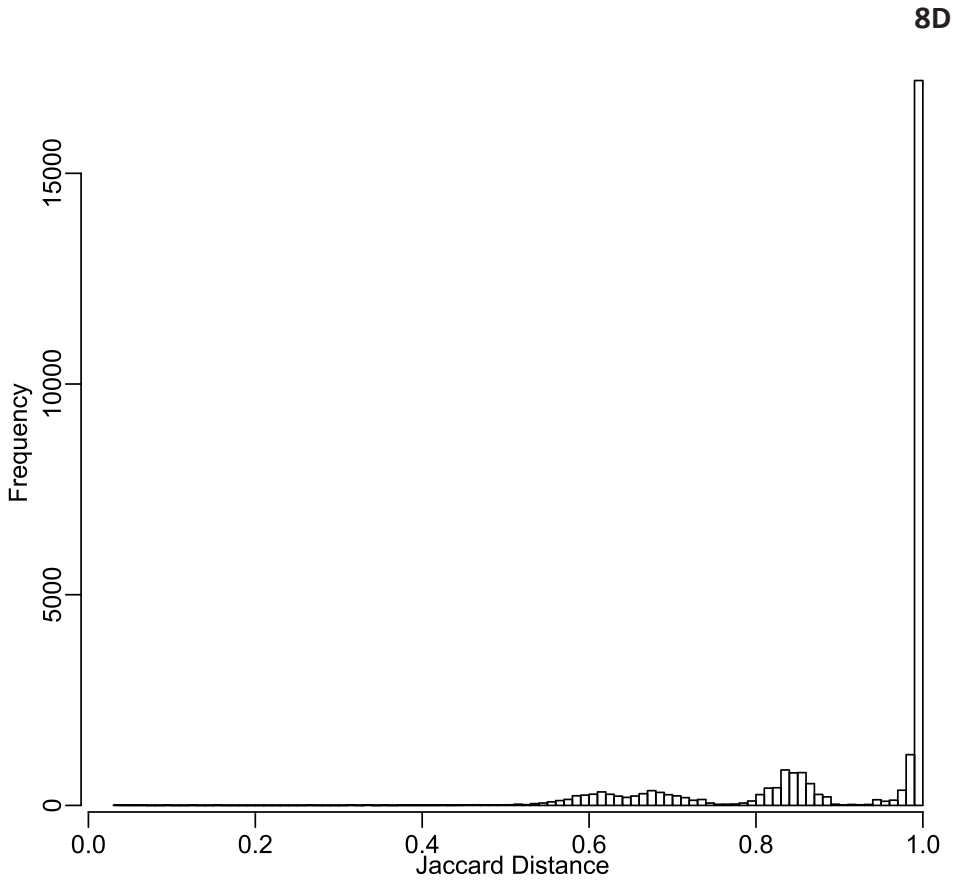
Figure 8: Results for the 21-mer dataset of 169 individuals using Jaccard distance and Neighbour-Joining. A) phylogenetic tree with distance; B) phylogenetic tree without distance (tree branch length); C) heatmap of phylogenetic distances showing low inter-group similarity and high intra-group similarity; D) histogram of Jaccard distances showing the same feature of low inter-group similarity and high intra-group similarity. Sample names ending in RAW are raw genomic data; names ending in APLG and CLC are assembled genomes; names ending in RNA, RNAseq and mRNA are RNA-seq datasets. Trees were plotted using iTOL (Letunic and Bork, 2007).

8B



8C




8D
Frequency
0
5000
10000
15000
0.0
0.2
0.4
0.6
0.8
1.0
Jaccard Distance

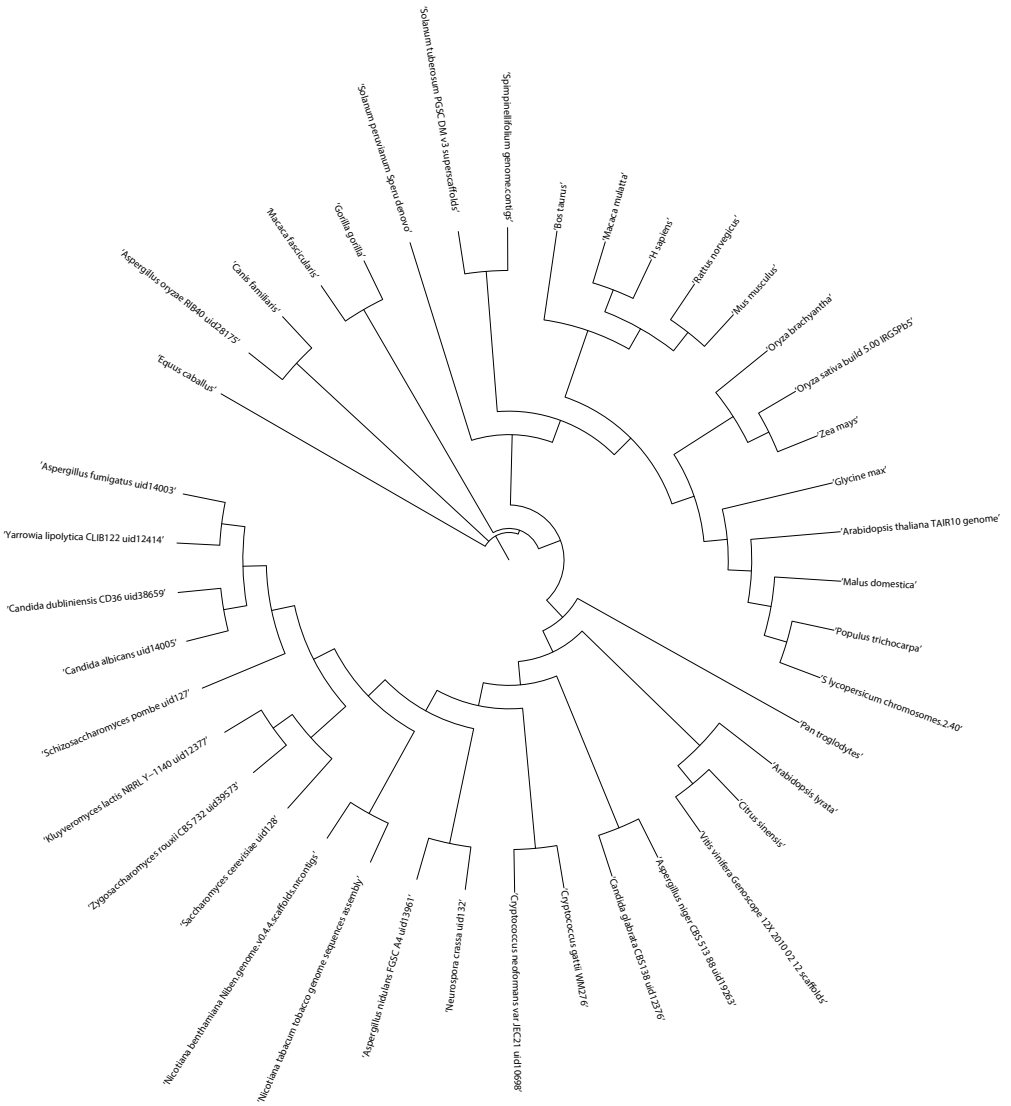


Figure 9: Phylogenetic tree created using ReferenceFree for 41 genomes. Due to filtering and sampling performed by ReferenceFree, clustering is impaired and inconsistent as exemplified by the positions of *Pan troglodites*, Nicotiana and Aspergillus.

Conclusion

We have introduced Cnidaria, a tool to quickly and reliably analyse WGS and RNA-seq samples from both assembled and unassembled NGS data, offering significant advantages in terms of time and space requirements compared to a state-of-the-art tool. By clustering in total 169 eukaryotic samples from 78 species (42 genus, 32 families, 27 orders, 5 phyla, 6 divisions and 3 kingdoms from the Eukaryota superkingdom) we have demonstrated that Cnidaria can handle a large number of samples from very distant phylogenetic origins, producing a reliable tree with up to 100% classification accuracy at the supra species level and 78% accuracy at the species level, the later value being low mostly due to interspecific crossings. As Cnidaria is also able to analyse RNA-seq data, researchers can acquire, besides the species information, physiological state information such as pathogenicity and stress response of the sample for downstream analysis.

A database created in Database Creation Mode allows querying directly for *k*-mers shared by a specified set of samples, enabling comparisons useful in several applications. Examples include identifying and quantifying polymorphisms between closely related samples, quantifying sequence diversity in the setup phase of large sequencing projects for sample selection, and ecological diversity analysis. In addition, *k*-mers shared exclusively by a set of samples can be used for diagnostic primer design, supporting the detection of target genes. Furthermore, mismatching *k*-mers between a sample and a close relative can be used to identify the source of contamination or introgressions, as performed by (Byrd *et al.*, 2014).

Availability and requirements

Project name:	Cnidaria
Project home page:	http://www.ab.wur.nl/cnidaria
Operating system(s):	64-bit Linux
Programming language:	C++ x11 and Python 2.7
Other requirements:	None to run; GCC 4.8 or higher for compiling
License:	MIT
Restrictions to use by non-academics:	No

List of Abbreviations Used

CCD	= Cnidaria Complete Database
CSD	= Cnidaria Summary Database
CSV	= Comma-Separated Values
DFT	= Discrete Fourier Transform
EST	= Expressed Sequence Tag
K-mer	= Substring of size k
MAD	= Median Absolute Deviation
NGS	= Next Generation Sequencing
PNG	= Portable Network Graphics
rDNA	= ribosomal DNA
RFM	= Reference-Free Methods
RNA-seq	= RNA sequencing

WGS	= Whole Genome Sequencing
.fastq	= Raw assembly file
.fasta	= Assembled sequence
.Newick	= Phylogenetic tree file format
HTML	= HyperText Markup Language
Read	= Contiguous sequence outputted by sequencing machine

Competing interests

The author(s) declare that they have no competing interests

Authors' contributions

SAA designed, wrote and tested the software; DR participated in the design of the software and validated the algorithm; ES, GSP, SP and HJ participated in the design of the software; All authors contributed in the confection of the manuscript, read it and approved the final manuscript.

Authors' information

Saulo Alves Aflitos	PhD candidate in Bioinformatics
Edouard Severing	PhD in Bioinformatics, Post-doc in Max-Planck-Institut für Pflanzenzüchtungsforschung - Köln
Gabino Sanchez-Perez	PhD in Bioinformatics - Senior Researcher Bioinformatics at Wageningen University, Cluster leader Cluster Bioinformatics – Plant Research International – Wageningen University
Sander Peters	PhD in Bioinformatics - Senior scientist / Bioinformatician at Plant Research International – Wageningen University
Hans de Jong	Professor of Cytogenetics at Wageningen University
Dick de Ridder	Professor of Bioinformatics at Wageningen University

Acknowledgements

This project was funded by Centre for BioSystems Genomics (CBSG) under the grant number TO09.

Supplementary materials

Supplementary Table 1: Sample description. Intermediate headers show database size and analysis time running with 1 thread (1x) or 20 threads (20x) for each Cnidaria database group. Each line contains a list of the names of the samples used, sequence ID, source type, source name, reference, size of Jellyfish database, size of input data, GC content, percentage of Ns, number of sequences in the input data and presence/absence list of a sample in the ReferenceFree analysis. For each *k*-mer size (11, 15, 17, 21 and 31bp): number of distinct *k*-mers, total number of *k*-mers, number of *k*-mers occurring only once, number of shared *k*-mers, percentage of *k*-mers shared. Input data is in the form of assembled genome (genomic - fasta files), raw genomic data (raw - fastq or BAM), filtered genomic data (raw filtered - BAM) or RNA-seq. Samples with values in the "shared" column are used for analysis. The 34 samples of the extended dataset were used exclusively against the 21-mer dataset. Analysis time and database sizes are calculated to the analysis of the dataset and do not correspond to the sum of the partial times.

Number samples	Count Group	Count Run	Count Total	Species Name	RefSeq/ID	Source Type	Source
Animals							
1	1	1	1	Bos taurus UMD_3.1	GCF_000003055.4	Genomic	NCBI
1	2	2	2	Canis lupus familiaris 3.1	GCF_000002285.3	Genomic	NCBI
1	3	3	3	Equus caballus 2.0	GCF_000002305.2	Genomic	NCBI
1	4	4	4	Gorilla gorilla 3.1	GCF_000151905.1	Genomic	NCBI
1	5	5	5	Homo sapiens GRCh38	GCF_000001405.26	Genomic	NCBI
1	6	6	6	Macaca fascicularis 5.0	GCF_000364345.1	Genomic	NCBI
1	7	7	7	Macaca mulatta 051212	GCF_000002255.3	Genomic	NCBI
1	8	8	8	Mus musculus GRCm38.p2	GCF_000001635.22	Genomic	NCBI
1	9	9	9	Pan troglodytes 2.1.4	GCF_000001515.6	Genomic	NCBI
1	10	10	10	Rattus norvegicus 6.0	GCF_000001895.5	Genomic	NCBI
Fungi							
1	1	11	11	Aspergillus fumigatus Af293 ASM265v1	GCF_000002655.1	Genomic	NCBI
1	2	12	12	Aspergillus nidulans FGSC A4	GCF_000149205.1	Genomic	NCBI
1	3	13	13	Aspergillus niger CBS 513.88	GCF_000002855.3	Genomic	NCBI
1	4	14	14	Aspergillus oryzae RIB40	GCF_000184455.1	Genomic	NCBI
1	5	15	15	Candida albicans SC5314	GCF_000182965.2	Genomic	NCBI
1	6	16	16	Candida dubliniensis CD36	GCF_000026945.1	Genomic	NCBI
1	7	17	17	Candida glabrata CBS 138	GCF_000002545.3	Genomic	NCBI
1	8	18	18	Cryptococcus gattii WM276	GCF_000185945.1	Genomic	NCBI
1	9	19	19	Cryptococcus neoformans var. neoformans JEC21	GCF_000091045.1	Genomic	NCBI
1	10	20	20	Kluyveromyces lactis NRRL Y-114	GCF_000002515.2	Genomic	NCBI
1	11	21	21	Neurospora crassa OR74A	GCF_000182925.1	Genomic	NCBI
1	12	22	22	Saccharomyces cerevisiae S288c	GCF_000146045.2	Genomic	NCBI
1	13	23	23	Schizosaccharomyces pombe 972h-	GCF_000002945.1	Genomic	NCBI
1	14	24	24	Yarrowia lipolytica CLIB122	GCF_000002525.2	Genomic	NCBI
1	15	25	25	Zygosaccharomyces rouxii CBS 732	GCF_000026365.1	Genomic	NCBI
Plants							
1	1	26	26	Arabidopsis lyrata	GCF_000004255.1	Genomic	NCBI
1	2	27	27	Arabidopsis thaliana 10		Genomic	solgenomics.net
1	3	28	28	Citrus sinensis cv valencia	GCF_000317415.1	Genomic	NCBI
1	4	29	29	Glycine max w82	GCF_000004515.3	Genomic	NCBI
1	5	30	30	Malus domestica	GCF_000148775.1	Genomic	NCBI
1	6	31	31	Nicotiana benthamiana	Niben.v0.4.4	Genomic	solgenomics.net
1	7	32	32	Nicotiana tabacum	2008	Genomic	solgenomics.net
1	8	33	33	Oryza brachyantha	GCF_000231095.1	Genomic	NCBI
1	9	34	34	Oryza sativa 5.0	5	Genomic	solgenomics.net
1	10	35	35	Populus trichocarpa	GCF_000002775.3	Genomic	NCBI
1	11	36	36	Solanum arcanum LA2157	PRJEB5226	Genomic	EBI
1	12	37	37			Raw Filter	
1	13	38	38			Genomic	
1	14	39	39	Solanum habrochaites LYC4	PRJEB5227	Genomic	EBI
1	15	40	40			Raw Filter	
1	16	41	41	Solanum lycopersicum 2.40		Genomic	NCBI
1	17	42	42			Genomic	
1	18	43	43			Raw Filter	
1	19	44	44	Solanum pennellii LA716	PRJEB5228	Genomic	EBI
1	20	45	45			Raw Filter	
1	21	46	46	Solanum peruvianum	de novo transc	Genomic	solgenomics.net
1	22	47	47	Solanum pimpinellifolium	A-1.0	Genomic	solgenomics.net
1	23	48	48	Solanum pimpinellifolium CGN14498	PRJEB6659	Raw Filter	EBI
1	24	49	49	Solanum tuberosum	PGSC DM v3	Genomic	solgenomics.net
1	25	50	50	Vitis vinifera	12x	Genomic	solgenomics.net
1	26	51	51	Zea mays	GCF_000005005.1	Genomic	NCBI
84	110	135	135	84 tomatoes reseq	PRJEB5235	Raw	EBI
All References							

Species Name	Reference	Analysis Time d:hh:mm (1x/20x)
		2d:01:20 / 3:25
Bos taurus UMD_3.1	Zimin et al., 2009	
Canis lupus familiaris 3.1	Lindblad-Toh et al., 2005	
Equus caballus 2.0	Wade et al., 2009	
Gorilla gorilla 3.1	Scally et al., 2012	
Homo sapiens GRCh38	Lander et al., 2001	
Macaca fascicularis 5.0	Ebeling et al., 2011	
Macaca mulatta 051212	Gibbs et al., 2007	
Mus musculus GRCm38.p2	Church et al., 2009	
Pan troglodytes 2.1.4	The Chimpanzee Sequencing and Analysis Consortium, 2005	
Rattus norvegicus 6.0	Gibbs et al., 2004	
		15:03 / 1:36
Aspergillus fumigatus Af293 ASM265v1	Nierman et al., 2005	
Aspergillus nidulans FGSC A4	Galagan et al., 2005	
Aspergillus niger CBS 513.88	Pel et al., 2007	
Aspergillus oryzae RIB40	Machida et al., 2005	
Candida albicans SC5314	Chibana et al., 2005	
Candida dubliniensis CD36	Jackson et al., 2009	
Candida glabrata CBS 138	Dujon et al., 2004	
Cryptococcus gattii WM276	D'Souza et al., 2011	
Cryptococcus neoformans var. neoformans JEC21	Loftus et al., 2005	
Kluyveromyces lactis NRRL Y-114	Dujon et al., 2004	
Neurospora crassa OR74A	Galagan et al., 2003	
Saccharomyces cerevisiae S288c	Goffeau et al., 1996	
Schizosaccharomyces pombe 972h-	Wood et al., 2002	
Yarrowia lipolytica CLIB122	Dujon et al., 2004	
Zygosaccharomyces rouxii CBS 732	Souciet et al., 2009	
		19d:08:51 / 1d:07:00
Arabidopsis lyrata	Hu et al., 2011	
Arabidopsis thaliana 10	Tabata et al., 2000	
Citrus sinensis cv valencia	Xu et al., 2013	
Glycine max w82	Schmutz et al., 2010	
Malus domestica	Velasco et al., 2010	
Nicotiana benthamiana	Bombarely et al., 2012	
Nicotiana tabacum	Sierro et al., 2014	
Oryza brachyantha	Chen et al., 2013	
Oryza sativa 5.0	Yamamoto et at., 2010	
Populus trichocarpa	Tuskan et al., 2006	
Solanum arcanum LA2157	Aflitos et al., 2014	
Solanum habrochaites LYC4	Aflitos et al., 2014	
Solanum lycopersicum 2.40	Tomato genome Consortium, 2012	
Solanum pennellii LA716	Aflitos et al., 2014	
Solanum peruvianum	Park et al., 2012	
Solanum pimpinellifolium	Ware et al., unpublished, 2011	
Solanum pimpinellifolium CGN14498	Aflitos et al., 2015	
Solanum tuberosum	Xu et al., 2011	
Vitis vinifera	Jaillon et al., 2007	
Zea mays	Schnable et al., 2009	
84 tomatoes reseq	Aflitos et al., 2014	
		6d:06:30 / 23:02

Species Name	Database Size (Gb)	Data Size	GC%	N%	Count sequences	ReferenceFree
	181.0					
Bos taurus UMD_3.1	19.0	2.7 Gbp	41.50	0.78	3.3 K	y
Canis lupus familiaris 3.1	19.0	2.4 Gbp	40.99	0.76	3.3 K	y
Equus caballus 2.0	19.0	2.5 Gbp	40.73	1.86	9.6 K	y
Gorilla gorilla 3.1	20.0	3.0 Gbp	37.77	6.81	50.2 K	y
Homo sapiens GRCh38	21.0	3.2 Gbp	38.95	4.98	455	y
Macaca fascicularis 5.0	21.0	2.9 Gbp	38.92	4.85	7.6 K	y
Macaca mulatta 051212	21.0	3.1 Gbp	37.88	7.31	122.1 K	y
Mus musculus GRCm38.p2	19.0	2.8 Gbp	40.54	2.83	179	y
Pan troglodytes 2.1.4	21.0	3.3 Gbp	35.59	12.67	24.1 K	y
Rattus norvegicus 6.0	20.0	2.9 Gbp	39.88	4.89	955	y
	2.8					
Aspergillus fumigatus Af293 ASM265v1	0.2	29.4 Mbp	48.82	1.96	8	y
Aspergillus nidulans FGSC A4	0.2	29.7 Mbp	50.08	0.59	17	y
Aspergillus niger CBS 513.88	0.3	34.0 Mbp	50.28	0.13	20	y
Aspergillus oryzae RIB40	0.3	37.1 Mbp	48.26	0.00	27	y
Candida albicans SC5314	0.0	949.6 Kbp	33.51	0.00	1	y
Candida dubliniensis CD36	0.1	14.6 Mbp	33.25	0.00	8	y
Candida glabrata CBS 138	0.1	12.3 Mbp	38.62	0.00	14	y
Cryptococcus gattii WM276	0.1	18.4 Mbp	47.85	0.07	14	y
Cryptococcus neoformans var. neoformans JEC21	0.2	19.1 Mbp	48.54	0.01	14	y
Kluyveromyces lactis NRRL Y-114	0.1	10.7 Mbp	38.72	0.00	7	y
Neurospora crassa OR74A	0.3	38.0 Mbp	49.87	0.00	822	y
Saccharomyces cerevisiae S288c	0.1	12.2 Mbp	38.15	0.00	17	y
Schizosaccharomyces pombe 972h-	0.1	12.6 Mbp	36.05	0.00	4	y
Yarrowia lipolytica CLIB122	0.2	20.6 Mbp	48.99	0.01	7	y
Zygosaccharomyces rouxii CBS 732	0.1	9.8 Mbp	39.13	0.01	7	y
	1393.8					
Arabidopsis lyrata	1.3	206.7 Mbp	32.07	11.11	695	y
Arabidopsis thaliana 10	0.9	119.7 Mbp	36.00	0.16	7	y
Citrus sinensis cv valencia	2.0	327.8 Mbp	31.30	8.11	4.8 K	y
Glycine max w82	5.7	973.8 Mbp	34.10	1.88	1.1 K	y
Malus domestica	2.3	524.3 Mbp	27.75	26.48	18	y
Nicotiana benthamiana	17.0	2.6 Gbp	35.64	6.26	140.9 K	y
Nicotiana tabacum	2.5	382.0 Mbp	36.45	0.00	300.2 K	y
Oryza brachyantha	2.0	259.9 Mbp	37.75	6.71	2.5 K	y
Oryza sativa 5.0	2.5	382.8 Mbp	42.43	2.61	12	y
Populus trichocarpa	2.3	307.8 Mbp	30.74	7.66	19	y
Solanum arcanum LA2157	5.1	665.1 Mbp	33.55	0.00	46.6 K	
	31.0	96.3 Gbp	34.81	0.02		
	5.5	832.5 Mbp	30.89	9.63	290.1 K	
Solanum habrochaites LYC4	5.5	724.2 Mbp	33.61	0.00	43.0 K	
	27.0	92.3 Gbp	34.95	0.01		
Solanum lycopersicum 2.40	5.3	781.7 Mbp	32.13	5.63	13	y
	4.7	599.7 Mbp	33.87	0.00	56.7 K	
	18.0	29.4 Gbp	36.41	0.13		
Solanum pennellii LA716	5.4	720.4 Mbp	34.00	0.00	57.2 K	
		76.2 Gbp	35.07	0.00		
Solanum peruvianum	0.2	24.5 Mbp	41.22	0.00	13.8 K	y
Solanum pimpinellifolium	5.1	688.9 Mbp	33.75	0.00	309.7 K	y
Solanum pimpinellifolium CGN14498	6.4	11.1 Gbp	34.25	0.00		
Solanum tuberosum	4.6	727.4 Mbp	32.66	6.15	66.3 K	y
Vitis vinifera	3.2	485.2 Mbp	33.48	3.09	2.1 K	y
Zea mays	6.9	2.1 Gbp	46.59	0.63	523	y
84 tomatoes reseq	1298.0	2784.1 G	36.1±0.6	0.01±0.02	0	
	2355.2					

Species Name	11-mers			15-mers		
	Distinct	Total	Unique	Distinct	Total	Unique
Bos taurus UMD_3.1	2.1 M	426.1 M	103.0	331.2 M	2.2 G	90.2 M
Canis lupus familiaris 3.1	2.1 M	426.5 M	52.0	335.3 M	2.2 G	90.7 M
Equus caballus 2.0	2.1 M	448.9 M	24.0	354.9 M	2.3 G	94.0 M
Gorilla gorilla 3.1	2.1 M	427.5 M	240.0	330.8 M	2.4 G	84.6 M
Homo sapiens GRCh38	2.1 M	434.7 M	186.0	334.2 M	2.6 G	83.7 M
Macaca fascicularis 5.0	2.1 M	431.3 M	171.0	337.4 M	2.5 G	87.8 M
Macaca mulatta 051212	2.1 M	432.7 M	124.0	337.0 M	2.5 G	87.0 M
Mus musculus GRCm38.p2	2.1 M	424.4 M	322.0	330.1 M	2.4 G	85.0 M
Pan troglodytes 2.1.4	2.1 M	430.8 M	159.0	333.9 M	2.5 G	85.3 M
Rattus norvegicus 6.0	2.1 M	446.7 M	20.0	349.5 M	2.5 G	88.5 M
Aspergillus fumigatus Af293 ASM265v1	2.1 M	28.8 M	13.4 K	26.9 M	28.8 M	25.6 M
Aspergillus nidulans FGSC A4	2.1 M	29.5 M	7.7 K	27.8 M	29.5 M	26.6 M
Aspergillus niger CBS 513.88	2.1 M	33.9 M	5.6 K	32.0 M	33.9 M	30.4 M
Aspergillus oryzae RIB40	2.1 M	37.0 M	1.9 K	34.9 M	37.1 M	33.1 M
Candida albicans SC5314	528.3 K	948.5 K	334.5 K	915.4 K	949.6 K	890.1 K
Candida dubliniensis CD36	1.6 M	14.4 M	339.9 K	12.3 M	14.5 M	11.0 M
Candida glabrata CBS 138	1.9 M	12.3 M	291.6 K	11.6 M	12.3 M	11.1 M
Cryptococcus gattii WM276	2.1 M	18.4 M	73.0 K	17.0 M	18.4 M	16.3 M
Cryptococcus neoformans var. neoformans JEC21	2.1 M	19.0 M	67.9 K	17.5 M	19.0 M	16.6 M
Kluyveromyces lactis NRRL Y-114	1.8 M	10.7 M	316.2 K	10.2 M	10.7 M	9.8 M
Neurospora crassa OR74A	2.1 M	37.8 M	6.1 K	34.1 M	38.0 M	31.6 M
Saccharomyces cerevisiae S288c	1.9 M	12.1 M	301.7 K	11.0 M	12.2 M	10.4 M
Schizosaccharomyces pombe 972h-	1.8 M	12.6 M	311.8 K	11.5 M	12.6 M	10.8 M
Yarrowia lipolytica CLIB122	2.1 M	20.5 M	79.7 K	19.0 M	20.5 M	18.1 M
Zygosaccharomyces rouxii CBS 732	1.8 M	9.8 M	343.5 K	9.3 M	9.8 M	8.9 M
Arabidopsis lyrata	2.1 M	151.0 M	11.6 K	87.6 M	181.9 M	58.4 M
Arabidopsis thaliana 10	2.1 M	108.3 M	21.5 K	75.9 M	119.2 M	55.9 M
Citrus sinensis cv valencia	2.1 M	203.4 M	3.0 K	115.1 M	296.0 M	67.6 M
Glycine max w82	2.1 M	336.6 M	232.0	194.6 M	845.7 M	86.3 M
Malus domestica	2.1 M	247.7 M	409.0	137.1 M	348.3 M	79.1 M
Nicotiana benthamiana	2.1 M	493.9 M	0.0	337.6 M	2.3 G	105.0 M
Nicotiana tabacum	2.1 M	249.9 M	62.0	138.7 M	364.7 M	77.3 M
Oryza brachyantha	2.1 M	202.6 M	13.0	136.0 M	239.4 M	93.8 M
Oryza sativa 5.0	2.1 M	277.8 M	1.0	159.7 M	362.3 M	101.2 M
Populus trichocarpa	2.1 M	192.8 M	6.4 K	124.2 M	278.4 M	75.5 M
Solanum arcanum LA2157	2.1 M	297.1 M	1.7 K	182.7 M	650.2 M	88.7 M
	2.1 M	534.6 M	0.0	394.5 M	34.9 G	82.9 M
	2.1 M	313.0 M	1.1 K	188.0 M	720.5 M	89.2 M
Solanum habrochaites LYC4	2.1 M	306.1 M	1.7 K	186.9 M	704.8 M	88.9 M
	2.1 M	534.5 M	0.0	380.0 M	33.9 G	84.7 M
Solanum lycopersicum 2.40	2.1 M	308.2 M	1.2 K	185.8 M	703.9 M	89.0 M
	2.1 M	287.4 M	2.1 K	178.8 M	589.2 M	89.0 M
	2.1 M	533.8 M	0.0	362.0 M	17.7 G	91.8 M
Solanum pennellii LA716	2.1 M	307.3 M	1.3 K	187.9 M	699.6 M	89.8 M
	2.1 M	534.5 M	0.0	379.7 M	30.4 G	84.3 M
Solanum peruvianum	2.0 M	24.4 M	157.2 K	22.5 M	24.3 M	21.0 M
Solanum pimpinellifolium	2.1 M	302.1 M	2.3 K	184.0 M	672.9 M	87.5 M
Solanum pimpinellifolium CGN14498	2.1 M	517.0 M	59.0	209.7 M	7.3 G	25.6 M
Solanum tuberosum	2.1 M	309.4 M	1.6 K	175.8 M	661.2 M	84.8 M
Vitis vinifera	2.1 M	255.5 M	3.7 K	145.3 M	455.0 M	77.5 M
Zea mays	2.1 M	488.3 M	0.0	262.4 M	1.4 G	119.9 M
84 tomatoes reseq	533.1 M±426.6 K		0.1±0.5	315.4 M±33.5 M	16.7 G±916.7 M	73.9 M±8.7 M	06.0 M±201.8 M

Species Name	17-mers			21-mers		
	Distinct	Total	Unique	Distinct	Total	Unique
Bos taurus UMD_3.1	1.3 G	2.3 G	904.2 M	2.0 G	2.3 G	1.9 G
Canis lupus familiaris 3.1	1.3 G	2.2 G	929.3 M	2.0 G	2.3 G	1.9 G
Equus caballus 2.0	1.4 G	2.3 G	979.2 M	2.1 G	2.4 G	2.0 G
Gorilla gorilla 3.1	1.4 G	2.5 G	903.7 M	2.2 G	2.6 G	2.0 G
Homo sapiens GRCh38	1.4 G	2.7 G	919.4 M	2.3 G	2.8 G	2.1 G
Macaca fascicularis 5.0	1.4 G	2.5 G	950.8 M	2.2 G	2.6 G	2.1 G
Macaca mulatta 051212	1.4 G	2.5 G	936.9 M	2.2 G	2.6 G	2.1 G
Mus musculus GRCm38.p2	1.3 G	2.4 G	917.4 M	2.0 G	2.5 G	1.9 G
Pan troglodytes 2.1.4	1.4 G	2.6 G	923.7 M	2.2 G	2.7 G	2.1 G
Rattus norvegicus 6.0	1.4 G	2.5 G	947.4 M	2.1 G	2.6 G	1.9 G
Aspergillus fumigatus Af293 ASM265v1	28.0 M	28.8 M	27.6 M	28.2 M	28.8 M	27.9 M
Aspergillus nidulans FGSC A4	28.9 M	29.5 M	28.6 M	29.0 M	29.5 M	28.8 M
Aspergillus niger CBS 513.88	33.5 M	33.9 M	33.3 M	33.7 M	33.9 M	33.7 M
Aspergillus oryzae RIB40	36.6 M	37.1 M	36.3 M	36.8 M	37.1 M	36.6 M
Candida albicans SC5314	928.2 K	949.6 K	911.7 K	934.7 K	949.6 K	922.0 K
Candida dubliniensis CD36	13.5 M	14.6 M	13.1 M	13.9 M	14.6 M	13.6 M
Candida glabrata CBS 138	12.1 M	12.3 M	11.9 M	12.1 M	12.3 M	12.0 M
Cryptococcus gattii WM276	17.5 M	18.4 M	17.3 M	17.6 M	18.4 M	17.5 M
Cryptococcus neoformans var. neoformans JEC21	18.1 M	19.0 M	17.7 M	18.2 M	19.0 M	17.9 M
Kluyveromyces lactis NRRL Y-114	10.5 M	10.7 M	10.4 M	10.6 M	10.7 M	10.5 M
Neurospora crassa OR74A	36.7 M	38.0 M	35.9 M	37.3 M	38.0 M	37.0 M
Saccharomyces cerevisiae S288c	11.4 M	12.2 M	11.2 M	11.5 M	12.2 M	11.3 M
Schizosaccharomyces pombe 972h-	12.1 M	12.6 M	11.9 M	12.2 M	12.6 M	12.0 M
Yarrowia lipolytica CLIB122	19.9 M	20.5 M	19.6 M	20.0 M	20.5 M	19.9 M
Zygosaccharomyces rouxii CBS 732	9.6 M	9.8 M	9.5 M	9.6 M	9.8 M	9.6 M
Arabidopsis lyrata	122.4 M	182.5 M	103.5 M	135.3 M	183.0 M	121.3 M
Arabidopsis thaliana 10	101.4 M	119.3 M	92.7 M	108.8 M	119.4 M	104.4 M
Citrus sinensis cv valencia	182.2 M	297.9 M	144.3 M	214.3 M	299.1 M	184.1 M
Glycine max w82	419.0 M	868.5 M	302.3 M	581.6 M	888.2 M	505.1 M
Malus domestica	213.7 M	348.1 M	170.7 M	246.7 M	345.4 M	215.0 M
Nicotiana benthamiana	971.6 M	2.3 G	644.7 M	1.5 G	2.4 G	1.3 G
Nicotiana tabacum	226.0 M	365.2 M	176.7 M	266.4 M	364.4 M	229.6 M
Oryza brachyantha	194.2 M	240.4 M	175.2 M	215.3 M	241.1 M	205.5 M
Oryza sativa 5.0	239.4 M	364.9 M	206.5 M	271.7 M	367.5 M	249.9 M
Populus trichocarpa	204.5 M	281.7 M	172.9 M	244.8 M	283.2 M	228.0 M
Solanum arcanum LA2157	388.9 M	656.2 M	301.6 M	524.6 M	659.9 M	479.9 M
	1.5 G	57.2 G	796.3 M	2.7 G	63.8 G	2.0 G
	405.2 M	729.4 M	309.4 M	553.1 M	736.4 M	499.6 M
Solanum habrochaites LYC4	406.0 M	712.4 M	311.1 M	555.9 M	717.5 M	505.4 M
	1.4 G	54.8 G	721.7 M	2.4 G	60.8 G	1.7 G
Solanum lycopersicum 2.40	398.7 M	713.0 M	306.5 M	541.3 M	721.4 M	491.4 M
	373.1 M	593.3 M	295.6 M	493.1 M	595.8 M	457.9 M
	1.1 G	20.6 G	600.4 M	1.7 G	20.7 G	1.1 G
Solanum pennellii LA716	403.6 M	707.6 M	310.3 M	549.4 M	713.4 M	498.7 M
	1.4 G	45.9 G	704.3 M	2.4 G	50.2 G	1.6 G
Solanum peruvianum	24.1 M	24.3 M	23.8 M	24.2 M	24.2 M	24.2 M
Solanum pimpinellifolium	394.2 M	678.0 M	301.0 M	531.6 M	680.6 M	479.7 M
Solanum pimpinellifolium CGN14498	465.6 M	7.9 G	80.9 M	642.9 M	7.9 G	129.5 M
Solanum tuberosum	350.9 M	669.0 M	264.2 M	461.8 M	674.8 M	401.0 M
Vitis vinifera	262.4 M	461.4 M	207.4 M	328.3 M	465.2 M	290.3 M
Zea mays	478.7 M	1.5 G	350.1 M	611.4 M	1.6 G	488.9 M
84 tomatoes reseq	20.3 G±1.3 G	04.7 M±143.8 M	1.4 G±382.9 M	21.0 G±1.4 G	79.8 M±353.1 M	1.8 G±457.3 M

Species Name	31-mers		
	Distinct	Total	Unique
Bos taurus UMD_3.1	2.2 G	2.5 G	2.1 G
Canis lupus familiaris 3.1	2.2 G	2.3 G	2.1 G
Equus caballus 2.0	2.3 G	2.4 G	2.2 G
Gorilla gorilla 3.1	2.4 G	2.7 G	2.3 G
Homo sapiens GRCh38	2.5 G	2.9 G	2.4 G
Macaca fascicularis 5.0	2.5 G	2.7 G	2.4 G
Macaca mulatta 051212	2.5 G	2.7 G	2.4 G
Mus musculus GRCm38.p2	2.2 G	2.6 G	2.1 G
Pan troglodytes 2.1.4	2.5 G	2.8 G	2.4 G
Rattus norvegicus 6.0	2.3 G	2.6 G	2.1 G
Aspergillus fumigatus Af293 ASM265v1	28.3 M	28.8 M	28.1 M
Aspergillus nidulans FGSC A4	29.1 M	29.5 M	28.9 M
Aspergillus niger CBS 513.88	33.8 M	33.9 M	33.7 M
Aspergillus oryzae RIB40	36.8 M	37.1 M	36.7 M
Candida albicans SC5314	939.3 K	949.6 K	929.9 K
Candida dubliniensis CD36	14.1 M	14.6 M	13.9 M
Candida glabrata CBS 138	12.2 M	12.3 M	12.1 M
Cryptococcus gattii WM276	17.7 M	18.4 M	17.5 M
Cryptococcus neoformans var. neoformans JEC21	18.3 M	19.0 M	18.0 M
Kluyveromyces lactis NRRL Y-114	10.6 M	10.7 M	10.6 M
Neurospora crassa OR74A	37.7 M	38.0 M	37.5 M
Saccharomyces cerevisiae S288c	11.6 M	12.2 M	11.4 M
Schizosaccharomyces pombe 972h-	12.2 M	12.6 M	12.0 M
Yarrowia lipolytica CLIB122	20.1 M	20.5 M	20.0 M
Zygosaccharomyces rouxii CBS 732	9.7 M	9.8 M	9.6 M
Arabidopsis lyrata	143.4 M	183.4 M	130.8 M
Arabidopsis thaliana 10	111.9 M	119.5 M	108.6 M
Citrus sinensis cv valencia	234.6 M	299.8 M	206.3 M
Glycine max w82	677.5 M	915.7 M	621.3 M
Malus domestica	269.3 M	336.3 M	244.9 M
Nicotiana benthamiana	1.9 G	2.4 G	1.8 G
Nicotiana tabacum	293.8 M	360.7 M	264.0 M
Oryza brachyantha	227.4 M	241.5 M	221.0 M
Oryza sativa 5.0	296.6 M	370.3 M	278.6 M
Populus trichocarpa	262.6 M	283.7 M	251.6 M
Solanum arcanum LA2157	601.4 M	661.5 M	576.2 M
	3.7 G	61.7 G	2.8 G
	646.0 M	740.1 M	612.1 M
Solanum habrochaites LYC4	644.8 M	720.2 M	615.2 M
	3.2 G	59.1 G	2.4 G
Solanum lycopersicum 2.40	627.4 M	730.7 M	595.3 M
	555.8 M	596.3 M	538.4 M
	2.1 G	18.4 G	1.4 G
Solanum pennellii LA716	642.1 M	716.2 M	612.4 M
	3.1 G	48.7 G	2.3 G
Solanum peruvianum	24.1 M	24.1 M	24.1 M
Solanum pimpinellifolium	608.0 M	679.4 M	574.7 M
Solanum pimpinellifolium CGN14498	762.0 M	7.1 G	173.9 M
Solanum tuberosum	546.6 M	677.7 M	498.6 M
Vitis vinifera	372.6 M	467.8 M	342.4 M
Zea mays	813.8 M	1.7 G	675.8 M
84 tomatoes reseq	20.0 G±1.4 G	1.1 G±427.2 M	1.0±0.0

Number samples	Count Group	Count Run	Count Total	Species Name	RefSeq/ID	Source Type	Source
Fast Comparison Dataset							
1	1	1	136	Anas platyrhynchos	GCF_000355885.1	Genomic	NCBI
1	2	2	137	Gallus gallus 4.0	GCF_000002315.3	Genomic	NCBI
1	3	3	138	Columba livia	GCF_000337935.1	Genomic	NCBI
1	4	4	139	Corvus brachyrhynchos	GCF_000691975.1	Genomic	NCBI
1	5	5	140	Corvus cornix cornix	GCF_000738735.1	Genomic	NCBI
1	6	6	141	Drosophila ananassae	dana_r1.3_FB2014_03	Genomic	FlyBase
1	7	7	142	Drosophila erecta	dere_r1.3_FB2014_03	Genomic	FlyBase
1	8	8	143	Drosophila grimshawi	dgri_r1.3_FB2014_03	Genomic	FlyBase
1	9	9	144	Drosophila melanogaster	GCF_000001215.4	Genomic	NCBI
1	10	10	145	Drosophila mojavensis	dmoj_r1.3_FB2014_03	Genomic	FlyBase
1	11	11	146	Drosophila persimilis	dper_r1.3_FB2014_03	Genomic	FlyBase
1	12	12	147	Drosophila pseudoobscura	dpse_r3.2_FB2014_04	Genomic	FlyBase
1	13	13	148	Drosophila sechellia	dsec_r1.3_FB2014_03	Genomic	FlyBase
1	14	14	149	Drosophila simulans	dsim_r1.4_FB2014_03	Genomic	FlyBase
1	15	15	150	Drosophila virilis	dvir_r1.2_FB2014_03	Genomic	FlyBase
1	16	16	151	Drosophila willistoni	dwil_r1.3_FB2014_03	Genomic	FlyBase
1	17	17	152	Drosophila yakuba	dyak_r1.3_FB2014_03	Genomic	FlyBase
1	18	18	153	Falco cherrug	GCF_000337975.1	Genomic	NCBI
1	19	19	154	Falco peregrinus	GCF_000337955.1	Genomic	NCBI
1	20	20	155	Geospiza fortis	GCF_000277835.1	Genomic	NCBI
1	21	21	156	Meleagris gallopavo	GCF_000146605.2	Genomic	NCBI
1	22	22	157	Melopsittacus undulatus	GCF_000238935.1	Genomic	NCBI
1	23	23	158	Nasonia vitripennis	GCF_000002325.3	Genomic	NCBI
1	24	24	159	Serinus canaria	GCF_000534875.1	Genomic	NCBI
1	25	25	160	Struthio camelus australis	GCF_000698965.1	Genomic	NCBI
1	26	26	161	Tribolium castaneum	GCF_000002335.2	Genomic	NCBI
Transcriptome							
1	1	27	162	Bos mutus	PRJNA221623	mRNA	NCBI
1	2	28	163	Nicotiana benthamiana	2008_08_25	mRNA	solgenomics.net
1	3	29	164	Nicotiana sylvestris	2008_08_25	mRNA	solgenomics.net
1	4	30	165	Nicotiana tabacum	2008_07_14	mRNA	solgenomics.net
1	5	31	166	Nicotiana tabacum	2008_07_14	mRNA	solgenomics.net
1	6	32	167	solanum lycopersicum	2008_10_21	mRNA	solgenomics.net
1	7	33	168	Solanum peruvianum	CSH transc.	mRNA	solgenomics.net
1	8	34	169	tomato species	2008_10_21	mRNA	solgenomics.net
All							

Species Name	Reference	Analysis Time d:hh:mm (1x/20x)
		3:30 / 0:20
Anas platyrhynchos	Huang et al., 2013	
Gallus gallus 4.0	International Chicken Genome Sequencing Consortium, 2005	
Columba livia	Shapido et al., 2013	
Corvus brachyrhynchos	Zhang and Li, unpublished, 2014	
Corvus cornix cornix	Poelstra et al., 2014	
Drosophila ananassae	Clark et al., 2007; St Pierre et al., 2014	
Drosophila erecta	Clark et al., 2007; St Pierre et al., 2014	
Drosophila grimshawi	Clark et al., 2007; St Pierre et al., 2014	
Drosophila melanogaster	Celniker, 2002	
Drosophila mojavensis	Clark et al., 2007; St Pierre et al., 2014	
Drosophila persimilis	Clark et al., 2007; St Pierre et al., 2014	
Drosophila pseudoobscura	Clark et al., 2007; St Pierre et al., 2014	
Drosophila sechellia	Clark et al., 2007; St Pierre et al., 2014	
Drosophila simulans	Clark et al., 2007; St Pierre et al., 2014	
Drosophila virilis	Clark et al., 2007; St Pierre et al., 2014	
Drosophila willistoni	Clark et al., 2007; St Pierre et al., 2014	
Drosophila yakuba	Clark et al., 2007; St Pierre et al., 2014	
Falco cherrug	Zhan et al., 2013	
Falco peregrinus	Zhan et al., 2013	
Geospiza fortis	Zhang G, unpublished, 2010	
Meleagris gallopavo	Dalloul et at., 2010	
Melopsittacus undulatus	Jarvis et al., unpublished, 2011	
Nasonia vitripennis	Werren et al., 2010	
Serinus canaria	Frankl et al., unpublished, 2013	
Struthio camelus australis	Zhang and Li, unpublished, 2014	
Tribolium castaneum	Richards et al., 2008	
		1d:01:07 / 2:58
Bos mutus	Qiu et al., 2012	
Nicotiana benthamiana		
Nicotiana sylvestris		
Nicotiana tabacum		
Nicotiana tabacum		
solanum lycopersicum		
Solanum peruvianum	Park et al., 2012	
tomato species		
		6d:10:00 / 23:22

Species Name	Database Size (Gb)	Data Size	GC%	N%	Count sequences	ReferenceFree
	0.0					
Anas platyrhynchos	0.7	1.1 Gbp	39.693264	3.17452147	78.5 K	
Gallus gallus 4.0	8.3	1.0 Gbp	41.2698016	1.34462292	15.9 K	
Columba livia	0.7	1.1 Gbp	40.6835306	1.90106002	14.9 K	
Corvus brachyrhynchos	0.6	1.1 Gbp	40.3299598	3.6223274	10.5 K	
Corvus cornix cornix	0.6	1.0 Gbp	40.4454388	2.61908937	1.3 K	
Drosophila ananassae	0.1	231.0 Mbp	38.8632237	7.39165001	13.7 K	
Drosophila erecta	0.1	152.7 Mbp	40.1542739	4.99509797	5.1 K	
Drosophila grimshawi	0.1	200.5 Mbp	35.2531201	7.17179948	17.4 K	
Drosophila melanogaster	1.1	149.5 Mbp	40.7422127	2.90279696	2.5 K	
Drosophila mojavensis	0.1	193.8 Mbp	36.7051738	7.02612509	6.8 K	
Drosophila persimilis	0.1	188.4 Mbp	41.8908118	6.78995914	12.8 K	
Drosophila pseudoobscura	0.1	152.7 Mbp	44.1143217	2.40298094	4.8 K	
Drosophila sechellia	0.1	166.6 Mbp	39.731609	5.60615323	14.7 K	
Drosophila simulans	0.1	137.8 Mbp	39.2199213	7.68326684	10.0 K	
Drosophila virilis	0.1	206.0 Mbp	36.7200101	8.16439532	13.5 K	
Drosophila willistoni	0.1	235.5 Mbp	35.3634211	5.05541127	14.8 K	
Drosophila yakuba	0.1	165.7 Mbp	41.4874403	1.87820501	8.1 K	
Falco cherrug	0.7	1.2 Gbp	40.8600885	2.02738411	5.9 K	
Falco peregrinus	0.7	1.2 Gbp	41.1044409	1.58445233	7.0 K	
Geospiza fortis	0.6	1.1 Gbp	40.6648067	2.25348054	27.2 K	
Meleagris gallopavo	0.6	1.0 Gbp	35.6272877	11.8243463	33	
Melopsittacus undulatus	0.7	1.1 Gbp	40.1111804	2.75282182	25.2 K	
Nasonia vitripennis	0.1	295.8 Mbp	33.6501375	19.3266605	6.1 K	
Serinus canaria	0.7	1.2 Gbp	41.540296	2.15828334	304.4 K	
Struthio camelus australis	0.7	1.2 Gbp	39.698985	3.28590438	6.9 K	
Tribolium castaneum	0.1	210.3 Mbp	24.3695347	28.0277322	2.2 K	
	1.8					
Bos mutus	0.3	53.3 Mbp	51.07	0.04	25.7 K	
Nicotiana benthamiana	0.1	31.8 Mbp	42.89	0.18	55.0 K	
Nicotiana sylvestris	0.0	2.7 Mbp	43.40	0.58	7.9 K	
Nicotiana tabacum	0.1	148.8 Mbp	42.10	0.05	239.8 K	
Nicotiana tabacum	0.6	10.2 Mbp	41.21	0.01	16.1 K	
solanum lycopersicum	0.3	130.4 Mbp	41.10	0.03	223.4 K	
Solanum peruvianum	0.2	47.6 Mbp	41.91	0.15	31.3 K	
tomato species	0.3	139.2 Mbp	41.02	0.03	239.6 K	
	2355.2					

Species Name	11-mers			15-mers		
	Distinct	Total	Unique	Distinct	Total	Unique
Anas platyrhynchos	2.1 M	360.4 M	1.1 K	270.8 M	1.1 G	94.1 M
Gallus gallus 4.0	2.1 M	359.5 M	803.0	264.8 M	1.0 G	94.3 M
Columba livia	2.1 M	365.5 M	678.0	275.5 M	1.1 G	96.8 M
Corvus brachyrhynchos	2.1 M	341.1 M	3.6 K	253.7 M	1.0 G	85.6 M
Corvus cornix cornix	2.1 M	334.5 M	5.1 K	249.4 M	1.0 G	84.3 M
Drosophila ananassae	2.1 M	168.4 M	86.0	98.3 M	197.4 M	73.3 M
Drosophila erecta	2.1 M	128.5 M	250.0	91.6 M	139.1 M	72.0 M
Drosophila grimshawi	2.1 M	133.8 M	3.0 K	88.6 M	158.8 M	62.8 M
Drosophila melanogaster	2.1 M	132.0 M	320.0	89.1 M	143.3 M	69.1 M
Drosophila mojavensis	2.1 M	141.6 M	1.4 K	93.8 M	168.6 M	69.4 M
Drosophila persimilis	2.1 M	154.5 M	81.0	96.3 M	169.0 M	73.2 M
Drosophila pseudoobscura	2.1 M	136.2 M	88.0	95.8 M	147.0 M	73.3 M
Drosophila sechellia	2.1 M	137.5 M	272.0	89.7 M	150.0 M	69.0 M
Drosophila simulans	2.1 M	117.4 M	404.0	86.1 M	125.4 M	67.8 M
Drosophila virilis	2.1 M	149.1 M	491.0	98.4 M	175.2 M	72.7 M
Drosophila willistoni	2.1 M	168.4 M	1.2 K	98.7 M	213.9 M	68.3 M
Drosophila yakuba	2.1 M	145.4 M	192.0	94.1 M	159.0 M	68.6 M
Falco cherrug	2.1 M	369.1 M	984.0	281.8 M	1.1 G	95.0 M
Falco peregrinus	2.1 M	369.8 M	925.0	282.2 M	1.1 G	95.2 M
Geospiza fortis	2.1 M	325.4 M	6.9 K	239.9 M	1.0 G	78.2 M
Meleagris gallopavo	2.1 M	333.5 M	3.0 K	247.9 M	904.8 M	90.3 M
Melopsittacus undulatus	2.1 M	334.4 M	4.6 K	252.0 M	1.1 G	81.7 M
Nasonia vitripennis	2.1 M	197.7 M	19.0	128.9 M	234.0 M	90.6 M
Serinus canaria	2.1 M	334.1 M	3.8 K	240.8 M	1.1 G	77.5 M
Struthio camelus australis	2.1 M	389.6 M	111.0	300.6 M	1.2 G	103.3 M
Tribolium castaneum	2.1 M	118.4 M	1.2 K	81.7 M	149.7 M	59.8 M
Bos mutus	2.1 M	52.7 M	67.1 K	35.3 M	52.9 M	25.8 M
Nicotiana benthamiana	1.9 M	29.6 M	145.4 K	10.2 M	29.6 M	4.4 M
Nicotiana sylvestris	1.1 M	2.5 M	482.6 K	2.0 M	2.4 M	1.7 M
Nicotiana tabacum	2.1 M	136.9 M	13.8 K	51.3 M	143.7 M	29.3 M
Nicotiana tabacum	1.7 M	10.1 M	321.8 K	7.2 M	10.0 M	5.4 M
solanum lycopersicum	2.0 M	116.9 M	58.3 K	27.0 M	124.1 M	11.4 M
Solanum peruvianum	2.0 M	44.9 M	55.8 K	25.8 M	44.1 M	10.6 M
tomato species	2.0 M	123.7 M	54.6 K	28.2 M	132.5 M	11.9 M

Species Name	17-mers			21-mers		
	Distinct	Total	Unique	Distinct	Total	Unique
Anas platyrhynchos	793.8 M	1.1 G	632.2 M	1.0 G	1.1 G	994.0 M
Gallus gallus 4.0	754.6 M	1.0 G	605.8 M	1.0 G	1.1 G	1.0 G
Columba livia	808.7 M	1.1 G	648.6 M	1.0 G	1.0 G	985.6 M
Corvus brachyrhynchos	772.5 M	1.0 G	609.7 M	985.6 M	1.0 G	974.1 M
Corvus cornix cornix	760.9 M	1.0 G	604.1 M	142.1 M	201.4 M	133.0 M
Drosophila ananassae	129.3 M	199.1 M	116.2 M	122.3 M	140.2 M	118.8 M
Drosophila erecta	115.9 M	139.6 M	109.1 M	134.5 M	162.7 M	124.5 M
Drosophila grimshawi	121.7 M	160.5 M	106.6 M	119.3 M	144.0 M	114.1 M
Drosophila melanogaster	113.0 M	143.6 M	104.8 M	143.8 M	172.5 M	138.2 M
Drosophila mojavensis	128.8 M	170.5 M	117.5 M	132.5 M	170.6 M	125.5 M
Drosophila persimilis	123.4 M	169.8 M	112.8 M	131.7 M	148.1 M	126.0 M
Drosophila pseudoobscura	122.9 M	147.6 M	113.2 M	120.0 M	151.5 M	113.8 M
Drosophila sechellia	113.6 M	150.6 M	104.4 M	113.9 M	125.8 M	110.4 M
Drosophila simulans	108.3 M	125.6 M	101.6 M	147.7 M	178.9 M	142.0 M
Drosophila virilis	134.3 M	176.8 M	123.1 M	162.9 M	218.1 M	152.4 M
Drosophila willistoni	143.2 M	215.9 M	125.4 M	128.9 M	159.9 M	117.5 M
Drosophila yakuba	120.8 M	159.4 M	106.2 M	1.1 G	1.1 G	1.1 G
Falco cherrug	863.6 M	1.1 G	685.6 M	1.1 G	1.2 G	1.1 G
Falco peregrinus	865.2 M	1.1 G	686.9 M	983.0 M	1.0 G	953.2 M
Geospiza fortis	745.6 M	1.0 G	574.5 M	883.9 M	911.2 M	874.6 M
Meleagris gallopavo	701.1 M	908.5 M	571.9 M	1.0 G	1.1 G	992.6 M
Melopsittacus undulatus	785.8 M	1.1 G	617.9 M	1.0 G	1.1 G	992.6 M
Nasonia vitripennis	181.0 M	234.8 M	163.6 M	198.0 M	235.8 M	188.2 M
Serinus canaria	756.5 M	1.1 G	580.1 M	1.0 G	1.1 G	974.4 M
Struthio camelus australis	902.9 M	1.2 G	724.3 M	1.2 G	1.2 G	1.1 G
Tribolium castaneum	116.0 M	150.2 M	103.5 M	131.7 M	150.6 M	125.8 M
Bos mutus	39.5 M	52.8 M	31.8 M	40.1 M	52.7 M	32.8 M
Nicotiana benthamiana	10.7 M	29.5 M	4.8 M	11.0 M	29.3 M	5.1 M
Nicotiana sylvestris	2.0 M	2.4 M	1.7 M	2.0 M	2.4 M	1.7 M
Nicotiana tabacum	59.7 M	143.3 M	38.9 M	62.9 M	142.4 M	42.4 M
Nicotiana tabacum	7.4 M	10.0 M	5.8 M	7.5 M	9.9 M	5.9 M
solanum lycopersicum	29.9 M	123.9 M	13.9 M	31.3 M	123.2 M	15.4 M
Solanum peruvianum	27.7 M	43.7 M	12.5 M	27.9 M	42.9 M	13.3 M
tomato species	31.5 M	132.2 M	14.7 M	33.2 M	131.5 M	16.4 M

Species Name	31-mers		
	Distinct	Total	Unique
Anas platyrhynchos	1.0 G	1.1 G	1.0 G
Gallus gallus 4.0	1.1 G	1.1 G	1.0 G
Columba livia	1.0 G	1.0 G	1.0 G
Corvus brachyrhynchos	1.0 G	1.0 G	995.7 M
Corvus cornix cornix	151.6 M	205.0 M	142.1 M
Drosophila ananassae	125.5 M	141.3 M	122.1 M
Drosophila erecta	140.3 M	166.4 M	131.3 M
Drosophila grimshawi	121.8 M	144.4 M	116.1 M
Drosophila melanogaster	150.9 M	174.7 M	146.3 M
Drosophila mojavensis	138.1 M	171.6 M	131.2 M
Drosophila persimilis	135.9 M	148.6 M	131.1 M
Drosophila pseudoobscura	123.2 M	153.1 M	116.6 M
Drosophila sechellia	115.8 M	126.0 M	112.7 M
Drosophila simulans	155.1 M	181.3 M	149.7 M
Drosophila virilis	173.9 M	220.4 M	163.8 M
Drosophila willistoni	133.7 M	160.6 M	122.8 M
Drosophila yakuba	1.1 G	1.1 G	1.1 G
Falco cherrug	1.1 G	1.2 G	1.1 G
Falco peregrinus	1.0 G	1.0 G	978.6 M
Geospiza fortis	901.0 M	911.7 M	897.0 M
Meleagris gallopavo	1.0 G	1.1 G	1.0 G
Melopsittacus undulatus	1.0 G	1.1 G	1.0 G
Nasonia vitripennis	208.7 M	236.8 M	200.4 M
Serinus canaria	1.0 G	1.1 G	1.0 G
Struthio camelus australis	1.2 G	1.2 G	1.2 G
Tribolium castaneum	138.2 M	150.8 M	133.8 M
Bos mutus	40.2 M	52.5 M	33.1 M
Nicotiana benthamiana	11.4 M	28.8 M	5.6 M
Nicotiana sylvestris	1.9 M	2.2 M	1.7 M
Nicotiana tabacum	66.1 M	139.8 M	46.3 M
Nicotiana tabacum	7.6 M	9.8 M	6.1 M
solanum lycopersicum	33.3 M	121.1 M	17.6 M
Solanum peruvianum	27.6 M	41.0 M	14.3 M
tomato species	35.5 M	129.3 M	18.9 M

Supplementary Table 2: List of samples used for the 1-nearest-neighbour analysis for supra-species classification, including supra-species taxonomy.

Sample	Genus	Family	Order	Phylum	Kingdom
Aspergillus fumigatus uid14003	Aspergillus	Aspergillaceae	Eurotiales	Ascomycota	Fungi
Aspergillus nidulans FGSC A4 uid13961	Aspergillus	Aspergillaceae	Eurotiales	Ascomycota	Fungi
Aspergillus niger CBS 513 88 uid19263	Aspergillus	Aspergillaceae	Eurotiales	Ascomycota	Fungi
Aspergillus oryzae RIB40 uid28175	Aspergillus	Aspergillaceae	Eurotiales	Ascomycota	Fungi
Candida albicans uid14005	Candida	Debaryomycetaceae	Saccharomycetales	Ascomycota	Fungi
Candida dubliniensis CD36 uid38659	Candida	Debaryomycetaceae	Saccharomycetales	Ascomycota	Fungi
Cryptococcus gattii WM276	Filobasidiella	Tremellaceae	Tremellales	Basidiomycota	Fungi
Cryptococcus neoformans var JEC21 uid10698	Filobasidiella	Tremellaceae	Tremellales	Basidiomycota	Fungi
Macaca fascicularis	Macaca	Cercopithecidae	Primates	Chordata	Metazoa
Macaca mulatta	Macaca	Cercopithecidae	Primates	Chordata	Metazoa
Arabidopsis lyrata	Arabidopsis	Brassicaceae	Brassicales	Streptophyta	Viridiplantae
Arabidopsis thaliana TAIR10	Arabidopsis	Brassicaceae	Brassicales	Streptophyta	Viridiplantae
Nicotiana benthamiana Niben v0.4.4	Nicotiana	Solanaceae	Solanales	Streptophyta	Viridiplantae
Nicotiana tabacum tobacco	Nicotiana	Solanaceae	Solanales	Streptophyta	Viridiplantae
Oryza brachyantha	Oryza	Poaceae	Poales	Streptophyta	Viridiplantae
Oryza sativa v5.00	Oryza	Poaceae	Poales	Streptophyta	Viridiplantae
Solanum arcanum LA2157 raw	Solanum	Solanaceae	Solanales	Streptophyta	Viridiplantae
Solanum cheesmaniae LA1401	Solanum	Solanaceae	Solanales	Streptophyta	Viridiplantae
Solanum chilense CGN15532	Solanum	Solanaceae	Solanales	Streptophyta	Viridiplantae
Solanum chmielewskii LA2663	Solanum	Solanaceae	Solanales	Streptophyta	Viridiplantae
Solanum corneliomuelleri LA0118	Solanum	Solanaceae	Solanales	Streptophyta	Viridiplantae
Solanum galapagense LA1044	Solanum	Solanaceae	Solanales	Streptophyta	Viridiplantae
Solanum habrochaites LYC4 raw	Solanum	Solanaceae	Solanales	Streptophyta	Viridiplantae
Solanum huaylasense LA1983	Solanum	Solanaceae	Solanales	Streptophyta	Viridiplantae
Solanum lycopersicum cv Heinz v2.40 raw	Solanum	Solanaceae	Solanales	Streptophyta	Viridiplantae
Solanum neorickii LA2133	Solanum	Solanaceae	Solanales	Streptophyta	Viridiplantae
Solanum pennellii LA0716 raw	Solanum	Solanaceae	Solanales	Streptophyta	Viridiplantae
Solanum peruvianum	Solanum	Solanaceae	Solanales	Streptophyta	Viridiplantae
Solanum pimpinelifolium CGN14498 raw	Solanum	Solanaceae	Solanales	Streptophyta	Viridiplantae
Solanum tuberosum DM v3	Solanum	Solanaceae	Solanales	Streptophyta	Viridiplantae

Supplementary Table 3: List of samples used for the 1-nearest-neighbour analysis for species classification, including species taxonomy.

Sample	Species
Solanum arcanum LA2157	Solanum arcanum
Solanum arcanum LA2172	Solanum arcanum
Solanum arcanum LA2157 denovo fasta APLG	Solanum arcanum
Solanum arcanum LA2157 raw	Solanum arcanum
Solanum cheesmaniae LA0483	Solanum cheesmaniae
Solanum cheesmaniae LA1401	Solanum cheesmaniae
Solanum chilense CGN15530	Solanum chilense
Solanum chilense CGN15532	Solanum chilense
Solanum chmielewskii LA2663	Solanum chmielewskii
Solanum chmielewskii LA2695	Solanum chmielewskii
Solanum habrochaites CGN15791 F glabratum	Solanum habrochaites
Solanum habrochaites LA0407	Solanum habrochaites
Solanum habrochaites LYC4 denovo fasta APLG	Solanum habrochaites
Solanum habrochaites LYC4 raw	Solanum habrochaites
Solanum huaylasense LA1364	Solanum huaylasense
Solanum huaylasense LA1365	Solanum huaylasense
Solanum huaylasense LA1983	Solanum huaylasense
Solanum lycopersicum cv Heinz v2.40	Solanum lycopersicum
Solanum lycopersicum LA2463 Allround	Solanum lycopersicum
Solanum lycopersicum LA2706 Moneymaker	Solanum lycopersicum
Solanum lycopersicum cv Heinz raw	Solanum lycopersicum
Solanum neorickii LA0735	Solanum neorickii
Solanum neorickii LA2133	Solanum neorickii
Solanum pennellii LA0716	Solanum pennellii
Solanum pennellii LA0716 raw	Solanum pennellii
Solanum peruvianum LA1278	Solanum peruvianum
Solanum peruvianum LA1954	Solanum peruvianum
Solanum peruvianum	Solanum peruvianum
Solanum pimpinellifolium LA1578	Solanum pimpinellifolium
Solanum pimpinellifolium LA1584	Solanum pimpinellifolium
Solanum pimpinellifolium LYC2798	Solanum pimpinellifolium
Solanum pimpinellifolium	Solanum pimpinellifolium
Solanum pimpinelifolium CGN14498 raw	Solanum pimpinellifolium

Supplementary Table 4: ReferenceFree datasets. Datasets used in the ReferenceFree analysis with the respective number of sequences, number of *k*-mers, number of valid *k*-mers (present in at least two samples) and percentage of *k*-mers considered valid for each dataset. On average, 0.016% ± 0.023% of the data is used.

Species name	# sequences	# total k-mers
Arabidopsis lyrata	695	206,654,035
Arabidopsis thaliana TAIR10 genome	7	119,667,610
Aspergillus fumigatus uid14003	8	29,384,798
Aspergillus nidulans FGSC A4 uid13961	17	29,710,946
Aspergillus niger CBS 513 88 uid19263	20	34,006,271
Aspergillus oryzae RIB40 uid28175	27	37,088,042
Bos taurus	3,317	2,670,355,959
Candida albicans uid14005	1	949,606
Candida dubliniensis CD36 uid38659	8	14,618,262
Candida glabrata CBS138 uid12376	14	12,338,028
Canis familiaris	3,268	2,410,911,515
Citrus sinensis	4,844	327,732,660
Cryptococcus gattii WM276	14	18,374,480
Cryptococcus neoformans var JEC21 uid10698	14	19,051,642
Equus caballus	9,637	2,474,736,322
Glycine max	1,147	973,756,350
Gorilla gorilla	50,196	3,034,656,224
H sapiens	455	3,209,277,005
Kluyveromyces lactis NRRL Y-1140 uid12377	7	10,729,307
Macaca fascicularis	7,601	2,946,691,717
Macaca mulatta	122,145	3,094,944,391
Malus domestica	18	526,594,476
Mus musculus	179	2,798,781,944
Neurospora crassa uid132	822	38,031,484
Nicotiana benthamiana Niben.genome.v0.4.4.scaffolds.nrcontigs	140,890	2,590,822,236
Nicotiana tabacum tobacco genome sequences assembly	300,158	376,868,190
Oryza brachyantha	2,491	259,857,775
Oryza sativa build 5.00 IRGSPb5	12	382,787,888
Pan troglodytes	24,129	3,322,785,342
Populus trichocarpa	19	307,840,388
Rattus norvegicus	955	2,870,165,093
S lycopersicum chromosomes.2.40	13	781,666,151
Saccharomyces cerevisiae uid128	17	12,156,765
Schizosaccharomyces pombe uid127	4	12,591,171
Solanum peruvianum Speru denovo	13,840	24,245,039
Solanum tuberosum PGSC DM v3 superscaffolds	66,254	726,099,466
Spimpinellifolium genome.contigs	309,695	682,664,352
Vitis vinifera Genoscope 12X 2010 02 12 scaffolds	2,059	485,144,450
Yarrowia lipolytica CLIB122 uid12414	7	20,550,757
Zea mays	523	2,067,611,843
Zygosaccharomyces rouxii CBS 732 uid39573	7	9,764,495
Average±Standard Deviation		974,943,036±1,206,232,257

Species name	# filtered k-mers	% filtered k-mers
Arabidopsis lyrata	23,017	0.011%
Arabidopsis thaliana TAIR10 genome	108,798,727	90.917%
Aspergillus fumigatus uid14003	28,202,227	95.976%
Aspergillus nidulans FGSC A4 uid13961	1,380,403	4.646%
Aspergillus niger CBS 513 88 uid19263	31,024	0.091%
Aspergillus oryzae RIB40 uid28175	3,023	0.008%
Bos taurus	84,913,103	3.180%
Candida albicans uid14005	934,662	98.426%
Candida dubliniensis CD36 uid38659	3,174,370	21.715%
Candida glabrata CBS138 uid12376	19,835	0.161%
Canis familiaris	4,329	0.000%
Citrus sinensis	166,728	0.051%
Cryptococcus gattii WM276	2,206,325	12.008%
Cryptococcus neoformans var JEC21 uid10698	1,052,619	5.525%
Equus caballus	9,434	0.000%
Glycine max	38,727,735	3.977%
Gorilla gorilla	19,985	0.001%
H sapiens	2,257,251,769	70.335%
Kluyveromyces lactis NRRL Y-1140 uid12377	1,747,364	16.286%
Macaca fascicularis	3,190	0.000%
Macaca mulatta	564,983,154	18.255%
Malus domestica	246,673,025	46.843%
Mus musculus	2,016,630,284	72.054%
Neurospora crassa uid132	685,879	1.803%
Nicotiana benthamiana Niben.genome.v0.4.4.scaffolds.nrcontigs	881,990	0.034%
Nicotiana tabacum tobacco genome sequences assembly	1,144,609	0.304%
Oryza brachyantha	26,817,832	10.320%
Oryza sativa build 5.00 IRGSPb5	271,676,231	70.973%
Pan troglodytes	13,189	0.000%
Populus trichocarpa	244,830,902	79.532%
Rattus norvegicus	2,095,384,468	73.006%
S lycopersicum chromosomes.2.40	541,281,299	69.247%
Saccharomyces cerevisiae uid128	564,367	4.642%
Schizosaccharomyces pombe uid127	12,175,174	96.696%
Solanum peruvianum Speru denovo	975,396	4.023%
Solanum tuberosum PGSC DM v3 superscaffolds	867,187	0.119%
Spimpinellifolium genome.contigs	959,514	0.141%
Vitis vinifera Genoscope 12X 2010 02 12 scaffolds	20,804	0.004%
Yarrowia lipolytica CLIB122 uid12414	20,047,446	97.551%
Zea mays	611,377,663	29.569%
Zygosaccharomyces rouxii CBS 732 uid39573	862,423	8.832%
Average±Standard Deviation	224,085,920±556,492,428	27.006%±35.730%

Species name	# valid k-mers	% valid k-mers on total	% valid k-mers on filtered
Arabidopsis lyrata	2,117	0.001%	9.198%
Arabidopsis thaliana TAIR10 genome	5,216,206	4.359%	4.794%
Aspergillus fumigatus uid14003	298,192	1.015%	1.057%
Aspergillus nidulans FGSC A4 uid13961	14,457	0.049%	1.047%
Aspergillus niger CBS 513 88 uid19263	3,245	0.010%	10.460%
Aspergillus oryzae RIB40 uid28175	329	0.001%	10.883%
Bos taurus	6,008,109	0.225%	7.076%
Candida albicans uid14005	55,239	5.817%	5.910%
Candida dubliniensis CD36 uid38659	217,847	1.490%	6.863%
Candida glabrata CBS138 uid12376	5,993	0.049%	30.214%
Canis familiaris	24	0.000%	0.554%
Citrus sinensis	9,296	0.003%	5.576%
Cryptococcus gattii WM276	62,303	0.339%	2.824%
Cryptococcus neoformans var JEC21 uid10698	46,324	0.243%	4.401%
Equus caballus	83	0.000%	0.880%
Glycine max	2,795,539	0.287%	7.218%
Gorilla gorilla	6,566	0.000%	32.855%
H sapiens	250,550,896	7.807%	11.100%
Kluyveromyces lactis NRRL Y-1140 uid12377	42,224	0.394%	2.416%
Macaca fascicularis	1,782	0.000%	55.862%
Macaca mulatta	188,249,253	6.082%	33.319%
Malus domestica	10,410,042	1.977%	4.220%
Mus musculus	201,653,910	7.205%	10.000%
Neurospora crassa uid132	18,013	0.047%	2.626%
Nicotiana benthamiana Niben.genome.v0.4.4.scaffolds.nrcontigs	44,691	0.002%	5.067%
Nicotiana tabacum tobacco genome sequences assembly	78,105	0.021%	6.824%
Oryza brachyantha	2,230,475	0.858%	8.317%
Oryza sativa build 5.00 IRGSPb5	11,576,302	3.024%	4.261%
Pan troglodytes	9,787	0.000%	74.206%
Populus trichocarpa	16,275,860	5.287%	6.648%
Rattus norvegicus	200,211,211	6.976%	9.555%
S lycopersicum chromosomes.2.40	26,676,385	3.413%	4.928%
Saccharomyces cerevisiae uid128	17,474	0.144%	3.096%
Schizosaccharomyces pombe uid127	422,017	3.352%	3.466%
Solanum peruvianum Speru denovo	722,297	2.979%	74.052%
Solanum tuberosum PGSC DM v3 superscaffolds	221,105	0.030%	25.497%
Spimpinellifolium genome.contigs	832,047	0.122%	86.715%
Vitis vinifera Genoscope 12X 2010 02 12 scaffolds	1,222	0.000%	5.874%
Yarrowia lipolytica CLIB122 uid12414	322,471	1.569%	1.609%
Zea mays	15,472,673	0.748%	2.531%
Zygosaccharomyces rouxii CBS 732 uid39573	24,625	0.252%	2.855%
Average±Standard Deviation	22,946,506±62,256,403	1.614%±2.342%	14.314%±21.075%

Chapter 5

Detection of PCR amplified single copy tomato sequences on chromosomal targets by Fluorescent *in situ* Hybridization

Saulo Alves Aflitos, José van de Belt, Sander Peters, Hans de Jong. Will be published in a chromosome research oriented journal.

Abstract

Fluorescent *in situ* Hybridization (FISH) is one of the physical mapping strategies to validate the quality of genome sequence assembly in tomato. Long stretches of genomic DNA in Bacterial Artificial Chromosomes (BACs) are used as probes for hybridization on the long and well-differentiated pachytene chromosomes. This so called BAC FISH painting is a multistep, time consuming technology, in which the presence of repetitive sequences in the probes often complicate straightforward interpretation of the cytogenetic mapping. However, with the greater part of the tomato genome being assembled, we are now able to select short single copy DNA sequences from the tomato genome database. In this study we show how bioinformatics filtering of non-repetitive chromosome region in chromosome 12 can be used to select single copy sequences for PCR amplification. To this end we selected four scaffolds (4,057; 4,878; 5,380 and 5,611) from chromosome 12 of *Solanum lycopersicum* cv Heinz v2.40 genome database. The sequences were first filtered for exact repeated sequences (using *k*-mers) and then filtered for non-exact matches, followed by a gapped alignment check. Based on the remaining sequences, we designed flanking primers for short (< 5 Kbp) amplicons that were obtained from PCR on genomic and BAC DNA templates. Subsequently we tested the PCR products on agarose gel electrophoresis and in FISH experiments under standard conditions without blocking of repeats. The single copy DNA from PCR on BAC templates produced the expected small fluorescent targets on chromosome 12. However, the PCR products amplified from genomic DNA produced various DNA bands on gel, and FISH demonstrated multiple fluorescent foci on chromosomes, suggesting that different amplicons were formed. In such cases we used gel extracted DNA for an additional nested PCR, whose probes indeed produce single fluorescent signals on the tomato chromosomes. Finally, we discuss the potential of PCR-FISH technology in support of genome assembly projects and compare with the recently developed oligo-based chromosome painting technologies.

Introduction

One of the biggest challenges in current genomics is the complete sequencing of the nuclear DNA and assembly into contigs, ultimately generating a number of pseudo-chromosomes that equals the haploid number of chromosomes in a cell complement (Koren and Phillippy, 2015). The obstacles for proper reconstruction into pseudo-chromosomes are intrinsically connected with the complex structure of the chromosome itself, which features islands of single copy sequences embedded in long stretches of tandem and dispersed repetitive sequences, including transposons, retroelements and tandem repeats. Most of such repeats are in large heterochromatin domains located in the centromeres, pericentromeres and nucleolar organizer regions (NORs), whereas several eukaryotes also have heterochromatic blocks at distal and interstitial locations of the chromosomes. Several plant and animal species display a very clear differentiation of euchromatin and heterochromatin (Stack *et al.*, 1974; de Jong *et al.*, 1999, Bennetzen, 2000; Fransz *et al.*, 2003), whereas other eukaryotes have dispersed or gradual patterns of heterochromatin. This pattern of genome organization in the chromosome is highly indicative for the repeat organization and presents problems for the sequencing and assembly of its genome.

In this study we focus on tomato, which is one of the best-studied plant model species. Its genome size is 1.03 pg (1007 MB, slightly more than the generally accepted 980 MB). It has a detailed genetic map and its 12 chromosomes are highly differentiated with clear patterns of euchromatin blocks (de Jong, 1999). The genome sequencing and assembly of tomato started with the BAC-by-BAC sequencing technology (Mueller *et al.*, 2005; 2009), in which seed BACs anchored on the genetic map were positioned on the chromosomes, whereas extension BACs with a minimal tiling path were selected to cover as much as possible the euchromatic parts of the chromosomes (Szinay *et al.*, 2008). In spite of its accuracy and robustness, the strategy was extremely time consuming and laborious. This forced the participating laboratories to resort to the alternative short read shotgun sequencing technology, which allowed lowering per-base cost at the expense of read length (The Tomato Genome consortium, 2012). The fragmented assemblies thus obtained produced large numbers of smaller contigs interrupted by an equally high number of gaps, most of which are likely filled with stretches of repetitive motifs that are almost impossible to assemble (Peters *et al.*, 2009).

In support of the genome assemblies of BACs and larger contigs, high-resolution Fluorescent *in situ* Hybridization is being used to map BAC sequences on acetic acid - ethanol fixed pachytene chromosomes (Chang *et al.*, 2007; de Jong *et al.*, 1999, Szinay *et al.*, 2008). Pachytene chromosomes are about 10 times longer than their mitotic metaphase counterparts and have well differentiated heterochromatin banding (de Jong *et al.*, 1999). An alternative technology for high-resolution BAC-FISH painting in tomato is the use of synaptonemal complex (SC) spread preparations obtained from hypotonically burst microsporocytes at pachytene, fixed at alkaline formaldehyde (Chang *et al.*, 2007, Stack *et al.*, 2009, Shearer *et al.*, 2014). The lengths of the SCs thus obtained are two times shorter than their corresponding acetic acid fixed pachytene used and the heterochromatin less pronounced, but the SC complements show less overlap and are more suitable for high-throughput BAC-FISH mapping.

As many of the BACs and contigs contain variable amounts of repetitive sequences in their DNA, hybridization of their probes may produce multiple fluorescing foci on the chromosome complements, which hampers unequivocal cytogenetic mapping of the single copy sequences. However, a pool of repeats can be isolated from genomic DNA using a Cot reannealing of denatured genomic DNA. This method has been tested and applied extensively on tomato by Peterson *et al.*, (1995), Chang *et al.*, (2008) and Szinay *et al.*, (2008) showing that the Cot100 fraction of the genome has all the repetitive sequences to paint the heterochromatin regions of all chromosomes. Complementary to this is the use of the same pool of repeats for effectively blocking off these repeats in the BAC probes, preventing hybridization to the chromosomes when single copy sequences have to be positioned on the chromosomes (Szinay *et al.*, 2008).

With the completion of the tomato genome (http://solgenomics.net/, *Solanum lycopersicum* cv Heinz v2.40), we reasoned that complicated BAC FISH painting with

Cot100 blocking can now effectively be circumvented using the sequence information as a source for directly selecting single copy sequence stretches and producing its DNA by PCR amplification using the primers that flank the selected regions. Here we propose the use of short (less than 5 Kbp) labeled amplicons as bases for FISH probes produced from BAC or genomic templates to analyze chromosomal organization and features. Compared to a large sized genomic template, a relatively small sized BAC template is less complex and has less chance of producing amplicons resulting from aspecific PCR primer amplification. As the amplicon usually has a smaller size than the BAC itself, this method allows for a higher resolution. The advantage is that PCR amplicons can be produced relatively fast and cheap in a high throughput setting. The disadvantage when using a BAC template is that this method requires a BAC library and the identification of BACs anchored to the region of interest. The most important aspect for FISH is the sensitivity and specificity of the PCR probe. Detection of smaller labeled PCR amplicons in general is less sensitive compared to a larger labeled BAC probe and therefore will be more difficult to visualize in the microscope. Furthermore, unique primer sequences in a whole genome template or BAC template background are required in order to produce a specific PCR amplicon.

Over the last decade there has been an increasing interest in alternative strategies for FISH studies using synthetic oligo sequences as probes. Zhong *et al.*, (2001) used Rolling Circle Amplification (RCA) as a surface-anchored DNA replication reaction that can be exploited to visualize single target DNA sequences as small as 50 bp in peripheral blood lymphocytes or in stretched DNA fibers. The specificity of the RCA method was high enough to use allele-discriminating oligonucleotide probes for distinguishing between wildtype and mutant alleles. Beliveau *et al.*, (2012) developed an oligo-painting technology. The oligonucleotide probes are renewable and highly efficient, and are able to robustly label chromosomes in cell culture, fixed tissues, and metaphase spreads. In Beliveau *et al.*, (2015) two advances in FISH microscopy were further exploited for the *in situ* visualization of single-copy regions of the genome using two single-molecule super-resolution methodologies. The developed system was robust and reliable enough to detect single-nucleotide polymorphisms (SNPs) that distinguish the maternal and paternal homologous chromosomes in mammalian and insect systems. This technology is enabled by computationally designed, oligonucleotide-based oligopaint probes, which is, in contrast to classical FISH probes, produced from segments of purified genomic DNA amplified in bacterial vectors or PCR reactions. Oligopaints belong to a new generation of probes that are derived entirely from synthetic DNA oligonucleotides. As such probes have their sequences chosen computationally; they can be made for any organism whose genome has been sequenced. Bienko *et al.*, (2012) described a cost-effective genome-scale PCR-based method for high-definition DNA FISH, using human and mouse genomic libraries of PCR primer pairs with optimal thermodynamic features, delimiting amplicons 200–220 nucleotides (nt) in length. After filtering out primer pairs that amplify multiple

targets and cross-hybridizing amplicons, databases of 4,823,784 and 4,387,601 unique primer pairs were created for the human and mouse genome, respectively.

Han *et al.*, (2015) developed an oligonucleotide (oligo)-based chromosome painting technique in cucumber (*Cucumis sativus*) that is specific to a single chromosome of the crop. In the first step all repetitive sequences were filtered using RepeatMasker (http://www.repeatmasker.org, Smit *et al.*, 2013). Oligos were identified using a newly developed bioinformatic pipeline called Chorus, which includes discarding oligos with homopolymers, removing remaining repetitive oligos and hairpin sequences, and massively synthesizing the remaining oligos *de novo*. These oligos were finally amplified and labeled as single stranded end-labeled probes for use in FISH. The specificity of the probes was high enough to differentially painting homeologues in interspecific Cucumis hybrids, and so can follow the process of meiotic chromosome pairing and recombination in detail. The potential of this technology is unprecedented, as it can be applied for any crop or model species whose genomes have been fully sequenced.

Our strategy can best be compared with the repeat-free sequence PCR that are amplified from four non-overlapping barley BACs (Ma *et al.*, 2010). For repeat masking, a Mathematically Defined Repeat (MDR) analysis was carried out. The barley sequences were chopped into overlapping 20-mers, and each 20-mer was assessed for its frequency in the barley MDR index generated from a 574 Mb Illumina/Solexa genomic database. As an extra control, the regions identified as unique based on the MDR plots were blasted additionally against the TIGR Gramineae repeats database. Poursarebani *et al.*, (2014) further refined this strategy, using repeat-free FISH probes corresponding to 8 previously anchored BAC contigs that were specifically allocated to barley chromosome 2H.
We have created a two tier repeat detection system for the detection of unique regions. The genomic sequence was first filtered for exact repeated sequences (using *k*-mers) and then filtered for non-exact matches, followed by a gapped alignment check. This approach significantly reduces the analysis run time compared to doing a full genome mapping as well as the identification of inexact matches. The unique regions detected where then used to PCR amplification the and PCR product the used as FISH probe to identify their physical position in the genome.

Materials and Methods

Single copy DNA sequence selection

A modified Mathematically Defined Repeats (MDR) method (Wang *et al.,* 2002) was used to define suitable regions for DNA amplification consisting of three steps: exact *k*-mer matching, inexact *k*-mer matching and gapped amplicon matching. Canonical *k*-mers of length 23 (a suitable size according to Cannon *et al.*, 2010) were extracted from the whole genome using Jellyfish (Marçais *et al.*, 2011) and the unique *k*-mers were saved in

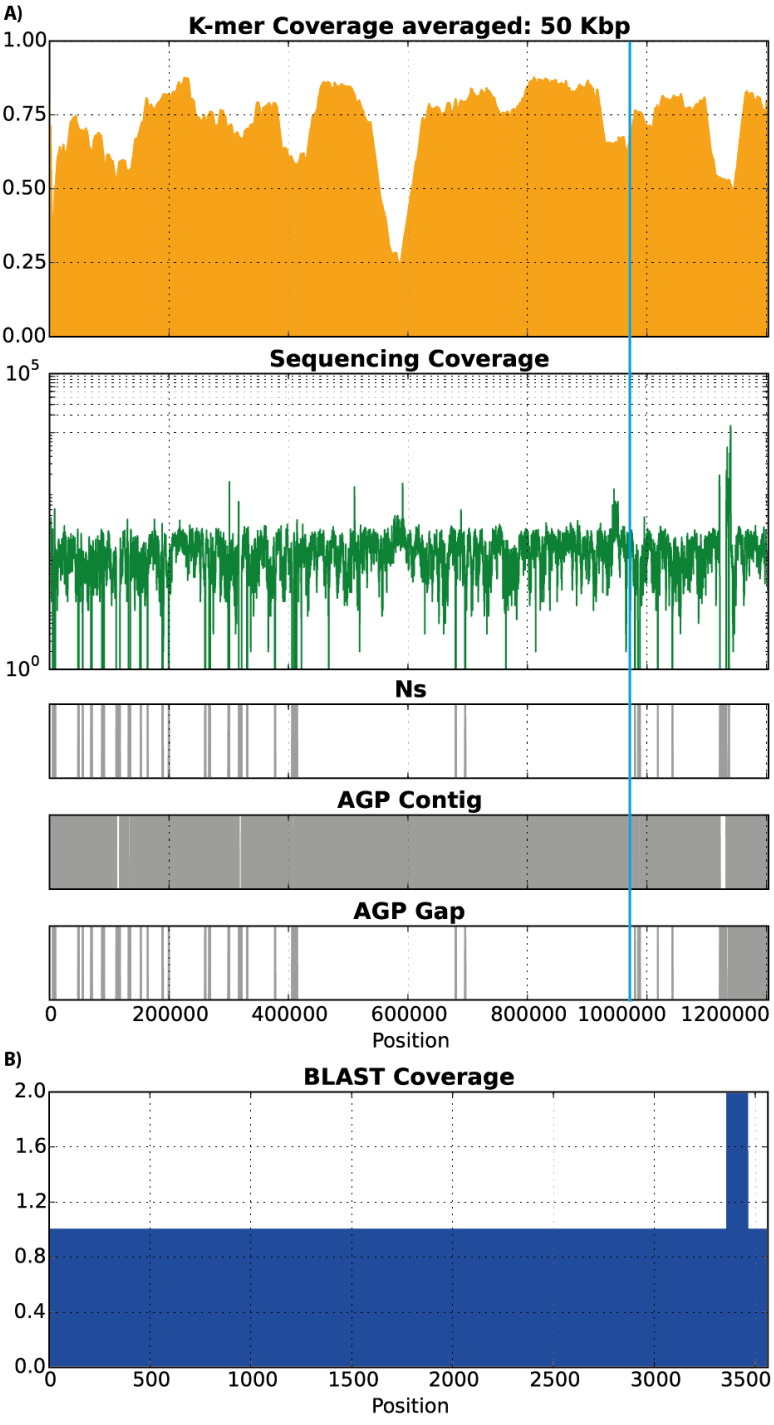


Figure 1: DNA sequence information for scaffold 5,611 on chromosome 12 of tomato, *Solanum lycopersicum* cv. Heinz 1706. Panel A) Scaffold wide information. "K-mer coverage": coverage index ranging from 0 (duplicated) to 1 (unique), smoothed by averaging in a sliding window of 50 Kbp. "Sequencing coverage": depth of coverage of genome sequencing. "Ns": stretches of N in the genome (gaps). "AGP contig": contig positions (grey) and gaps between contigs (white). "APG gap": gaps of known size in the assembly. The blue vertical line shows the position represented panel B. Panel B) BLAST information for the region selected for primer design between coordinates 967,164 bp and 970,727 bp (3.5 Kbp). "BLAST coverage": number of bits (sequence matches, weighted using Kimura, 1980, substitution matrix) at each position.

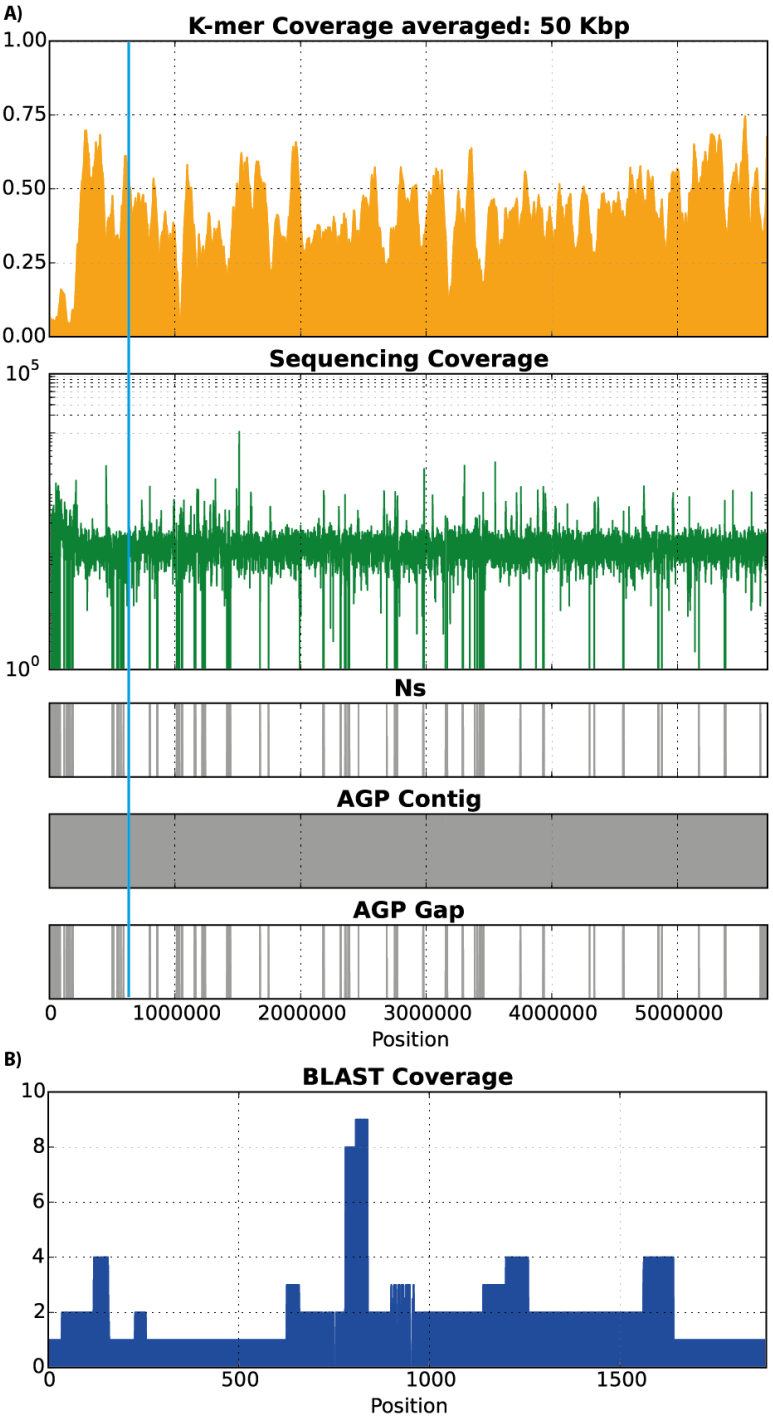


Figure 2: DNA sequence information for scaffold 4,878 on chromosome 12 of tomato, *Solanum lycopersicum* cv. Heinz 1706. Panel A) Scaffold wide information. "K-mer coverage": coverage index ranging from 0 (duplicated) to 1 (unique), smoothed by averaging in a sliding window of 50 Kbp. "Sequencing coverage": depth of coverage of genome sequencing. "Ns": stretches of N in the genome (gaps). "AGP contig": contig positions (grey) and gaps between contigs (white). "APG gap": gaps of known size in the assembly. The blue vertical line shows the position represented panel B. Panel B) BLAST information for the region selected for primer design between coordinates 622,710 bp and 624,594 bp (1.8 Kbp). "BLAST coverage": number of bits (sequence matches, weighted using Kimura, 1980, substitution matrix) at each position.

Table 1: Selected scaffolds used in this study, reference BACs, used FISH detection system and sequence length of the produced PCR amplicons. See "Materials and methods" for details.

Scaffold	Primer Name	Genomic multi foci	Genomic (gel extr/nested)	BAC template	Stringency Genomic	BAC	Scaffold v2.30	Bac Start	Bac End	Length	Chr Start	Chr End
Sc. 2	Primer_sc2-1	✓ multi	not used	✓	✓ multi	H036P06	SL2.40sc04878	622,710	624,594	1,884	16,714,760	16,716,664
	Primer_sc2-2			✓		H045M19	SL2.40sc04878	3,080,728	3,082,538	1,810	19,172,778	19,174,610
	Primer_sc2-3			✓		H210I02	SL2.40sc04878	4,093,931	4,095,019	1,088	20,185,981	20,187,092
Sc. 3	Primer_sc3-1	✓ multi	not used	✓	✓ multi	H132J16	SL2.40sc04057	3,198,454	3,200,936	2,482	25,008,367	25,010,869
	Primer_sc3-2			PCR failed		H070I24	SL2.40sc04057	7,835,794	7,838,398	2,604	29,645,707	29,648,335
	Primer_sc3-3			✓		H128N12	SL2.40sc04057	13,401,063	13,403,730	2,667	35,210,976	35,213,666
	Primer_sc3-4			PCR failed		H091D21	SL2.40sc04057	22,823,181	22,827,553	4,372	44,633,094	44,637,488
	Primer_sc3-5			✓		RH166N20	SL2.40sc04057	23,797,776	23,801,944	4,168	45,607,689	45,611,878
Sc. 9	Primer_sc9-1	✓ multi	not used	✓	✓ multi	H070I08	SL2.40sc05611	353,222	355,021	1,799	62,088,511	62,090,332
	Primer_sc9-2			✓		H170A10	SL2.40sc05611	528,593	530,597	2,004	62,263,882	62,265,906
	Primer_sc9-3		✓	✓		H234M08	SL2.40sc05611					
	Primer_sc9-4					H234M08	SL2.40sc05611	967,164	970,727	3,563	62,702,453	62,706,035
Sc. 10	Primer_sc10-1	✓ multi	not used	PCR failed	✓ multi	H204O07	SL2.40sc05380	1,864,775	1,867,879	3,104	63,618,355	63,621,478
	Primer_sc10-2			✓		H066C10	SL2.40sc05380	1,423,814	1,427,393	3,579	64,058,838	64,062,439
	Primer_sc10-3			PCR failed		H028I07	SL2.40sc05380	927,120	930,266	3,146	64,555,966	64,559,133
	Primer_sc10-4			✓		H041K16	SL2.40sc05380	369,372	371,958	2,586	65,114,271	65,116,881

a FASTA file, excluding any exact duplicate in the genome. This FASTA file was then mapped back against four selected scaffolds (4,057, 4,878, 5,380 and 5,611) in chromosome 12 of *Solanum lycopersicum* cv Heinz v2.40 genome (annotation v2.30, http://solgenomics.net) using BWA (Li *et al.*, 2009) with relaxed settings which allow non-exact matches to occur (maximum edit distance of 18, maximum of 5 gap opens, maximum of 5 gap extends, penalty of 1 for each gap extension and maximum edit distance in the seed of 3). The resulting SAM file, containing all the exact and non-exact matches in the scaffolds for each (unique) *k*-mer, was then converted to a coverage file describing how many *k*-mers map at every coordinate in the scaffold. A coverage index was then calculated for each scaffold coordinate by dividing the number of *k*-mers mapping there by the length of the *k*-mer. For a unique region, the coverage index is equal to 1, since the number of *k*-mers mapping is equal to number of *k*-mers extracted (e.g.: 23 unique *k*-mers mapping of 23 *k*-mers extracted, or 23/23 = 1), while a repeat region will have an index below 1 as the number of *k*-mers mapping is smaller than the number of *k*-mers extracted due to the filtering performed in the JELLYFISH step (e.g., 5 unique *k*-mers mapping of 23 *k*-mers extracted, or 5/23 = 0.217). Non-exact matches returned by BWA may generate a coverage index larger than 1, given that the number of *k*-mers returned may be greater than the number of *k*-mers extracted (e.g., 30 *k*-mers mapping of 23 *k*-mers extracted, or 30/23 = 1.304). Coverage indexes above 1 are converted to 0, thus creating a plot that ranges from 0 (repeats with exact and inexact matches) to 1 (unique regions with a single exact match) making it easier to identify the average uniqueness of a region by taking the average uniqueness of all the coordinates inside its borders.

In each of the five chosen scaffolds, unique target regions were selected for primer design. The following selection criteria were applied for primer amplification: (i) lengths ranging between 1 Kbp and 4 Kbp;

(ii) average coverage index as close as possible to 1; (iii) at least 1 Kbp from the border of the scaffold, but as close as possible to it; (iv) at least 1 Kbp away from any intra-scaffold contig border; and (v) not containing gaps (denoted as stretches of "N"s in the FASTA file, Figures 1 and 2). All selected regions were BLASTed (Altschul *et al.*, 1997) against the *S. lycopersicum* cv Heinz v2.40 genome without soft- or dust-masking of repeats and returning at most 100 matches for the elimination of gapped inexact repeats. The segments containing the lowest number of gapped matches elsewhere in the genome are considered candidates, which are then used for primer design with PRIMER-BLAST (Ye *et al.*, 2012). We selected the primer pairs with unique PCR products against all Solanum (NCBI taxid: 4107) genomes and with an amplicon length greater than 800 bp. The primers (cloning primers HPLC, HPLC purified) were obtained from Fisher Scientific NL (Tables 1 and 2).

DNA isolation

Genomic DNA was isolated following the protocol of Rogers and Bendich (1988) with a few modifications. Leaves of young Heinz 1706 tomato plants were collected, snap frozen and stored immediately at -80°C until use. More than 5 grams of frozen leaves per aliquot were transferred to a pre-cooled mortar, filled with liquid nitrogen and ground with a pestle until a light green powder was obtained. Then 5 grams of the ground leaves were put into a pre-cooled 50mL tube to which 5 mL hot (65°C) 2% CTAB extraction buffer was added. After shaking, the mixture was incubated for 1 hour in a 65°C water bath. An equal volume of (10 mL) chloroform/ isoamylalcohol (24:1) was added after which the mixture was rotated at 20°C for 20 minutes. After centrifugation at 3600 rpm for 30 min the aqueous upper phase was pipetted to a conical 50 mL tube and 0.1 x volume (approx. 600 µL) of hot 10% CTAB was added and mixed. Then a second chloroform/ isoamylalcohol 24:1 purification was applied and the aqueous layer transferred to a new 50 mL tube to which now an equal amount of 1% CTAB precipitation buffer

Table 2: Primers names and sequences.

Primer Name	Primer Sequence
ch12-sc02-1f	CCCTTGTTAGTGGGTATGTTGGC
ch12-sc02-1r	AGTGGCTTGTTGGGTGGTTAG
ch12-sc02-2f	AGAACTCTCTTCGGGTGGAG
ch12-sc02-2r	GACGATAGGATGTGCCTTTATGC
ch12-sc02-3f	ACCCTCAATACAAACATGAAAGTCAC
ch12-sc02-3r	TTGGAAGTTCAATAGGGAATCCAC
ch12-sc03-1f	AGGAAGCAAAGGTAGTTTGGGATG
ch12-sc03-1r	TCATTGCCTAACGTGGCCTTG
ch12-sc03-2f	ACCTGTCCCTTCCATACCTAAGTC
ch12-sc03-2r	AGTGCTGTAACCAAGTTCATAGTTC
ch12-sc03-3f	GCTCCCATGATTGTAGAGCATCC
ch12-sc03-3r	AGATAGGTGCCATCAAATCTCCAG
ch12-sc03-4f	TGGCTAACCCAATCGGAAACC
ch12-sc03-4r	GAACAGTGATGGGCTCTTAATGC
ch12-sc03-5f	GAGTCCGACAAGCCCTGAATG
ch12-sc03-5r	GCACCATTTGTTTAGAGGCAGC
ch12-sc09-1f	AGTTGGACAGTGATTGGTGGG
ch12-sc09-1r	AGTGAGGAAGTAGCTCTGAAGTG
ch12-sc09-2f	TTGCTGAGCTGGACAACAGAG
ch12-sc09-2r	TTGGTGGGCCATGCTATTCAC
ch12-sc09-4f	ACGCTCGCCTTCTTCATTACC
ch12-sc09-4r	GTGTCCCGAAGGCTAAGCAC
ch12-sc10-1f	ATGAGCTGCGTGGAGTCTTG
ch12-sc10-1r	ACCACCTCCGGTCTACCATC
ch12-sc10-2f	GGGGGAAAGAGCATGAAAGAAAG
ch12-sc10-2r	TTTTGAGATGGGCTGCTCCG
ch12-sc10-3f	TGAGACAGGAGGAGAAGACGAG
ch12-sc10-3r	GTGGGTCTGCACGCTATACTC
ch12-sc10-4f	TCTTCTAATTTCAGCATCCAGCTAC
ch12-sc10-4r	AAGTACATGGTGTGTAACATCCTC

was added. The tube was slowly inverted a few times and then spun down for 3 min at 3600 rpm at 20°C. The pellet was re-suspended in 1.5 mL High-Salt-TE buffer in a 65°C water bath for 20 minutes. After a 2x volume of ice-cold 96% ethanol was added, the mix was spun down and the pellet washed in 3 mL 80% ethanol and spun down again. After discarding the supernatant, the pellet was dried at 20°C and then resuspended in 500 µL TE buffer, to which 1 µL RNAse (10 mg/mL) was added for a 1 hr incubation at 65°C. Finally, DNA concentration and quality check was performed with the NanoDrop micro-volume system (Thermo Fisher Scientific).

PCR amplification

The PCR mix (20 µL reaction) for the genomic DNA contained 1) Sterile Milli-water 10.2 µL, 2) 5x Phire Hotstart reaction buffer 4 µL, 3) 10 mM dNTP's 0,4 µL, 4) forward and reverse primers (5 µM) 1 or 2 µL each, 5) template DNA (100x diluted) 1 µL, 6) Phire Hotstart DNA polymerase 0.4 µL and 7) optional: DMSO (3% of total reaction volume) 0.6 µL. The PCR mix (20 µL reaction) for the BAC template contained 1) Sterile MQ 10.2 µL, 2) 5x Phire Hotstart reaction buffer 4 µL, 3) 10 mM dNTP's 0.4 µL, 4) forward and reverse primers (5 µM) 1 or 2 µL, each, 5) template DNA (glycerol stock) tooth stick, 6) Phire Hotstart DNA polymerase 0.4 µL, 7) optional: DMSO (3% of total reaction volume) 0.6 µL. PCR cycling conditions are 1) 98°C for 2', 2) 98°C for 10", 3) 67°C for 20" and 30-35x, 4) 72°C for 20", 5) 75°C for 5' and 6) finally keep at 4°C. For checking the PCR products we made 1% agarose gel (0.4 gram agarose, 40 mL TBE, 2 µL ethidium bromide), for which we applied 2 µL of 1 Kb marker and 1 µL of DNA sample with 1 µL 6x loading buffer. The PCR product was purified with the Invitrogen PureLink® PCR purification Kit (K3100-01) and DNA concentration was quantified with the NanoDrop system. Finally 30 µL of the purified PCR samples were labeled with the Nick Translation kit (Roche) following the instructions of the manufacturer. For indirect labeling we used biotin-16-dUTP or digoxigenin-11-dUTP, and visualized the probe using streptavidin Cy5 and anti-digoxigenin–FITC detection protocols (Chang *et al.*, 2008), respectively. For direct labeling in the multi-color FISH, we labeled BAC DNA with dUTP-Cy3 (Amersham, http://www5.amershambiosciences.com) and dCTP-Cy3.5 (Amersham).

Slide preparation and Fluorescent *in situ* Hybridization

Fast growing flower buds were collected in the late morning and transferred directly into freshly prepared ethanol 96%: glacial acetic acid (3:1) fixative. If not processed directly the material was transferred to 70% ethanol and stored at +4°C until use. On the day of slide preparation we put a couple of flower buds in Petri dish filled with water. Buds were staged by dissecting a single anther, squashing it in a drop of 1% aceto-carmine and staining under an 18x18 mm coverslip. If the anther contains pollen mother cells at pachytene we transferred the remaining anthers of that bud to a 1.5 mL Eppendorf tube in the fridge until use. Prior to enzymatic digestion the anthers were rinsed

in 10 mM sodium citric Buffer (pH 4.5) 2 times for 5 minutes. We incubated the material in a small tube with the pectolytic enzyme mix containing 1% pectolyase Y23 (Sigma P-3026), 1% cellulase RS (Yakult 203033, Yakult Pharmaceutical, Tokyo, Japan) and 1% cytohelicase (Bio Sepra 24970-014) in the sodium citrate buffer for 2-3 h at 37°C. After the enzymatic digestion anthers were carefully rinsed in MQ water and kept on ice until use to prevent further digestion. For each slide we used only a single anther, which was put in a small Eppendorf tube to which 20 µL 60% acetic acid was added. The cell mixture was sucked with a pipet few times to release the cells from the anther tissues. Then the 8-10 µL of the anther cells were dropped onto a wet and grease-free slide on a 55°C hotplate for 2 minutes while adding 2-3 drops of 60% acetic acid. Finally a few drops of freshly prepared ethanol acetic acid (3:1) next and on top of the cell mixture improves the quality of the chromatin contrast and the spreading of the cells on the slide. After air-drying we checked each slide under the phase contrast microscope. Only slides with well spread pachytene complements, hardly any cytoplasm, and well-differentiated heterochromatin were selected for FISH.

Slides were incubated with 1 µL RNaseA (10mg/mL) in 100 µL of 2×SSC (1:100) at 37°C for 1h. In the case of slides that still contain remnants of cytoplasm on the chromosomes we continued with a pepsin treatment (1 µL Pepsin (500 µg/mL) in 100 µL of 0.01 M HCL (1:100) for 10 minutes at 37°C (times were adjusted if needed), followed by a 2 min. wash in MQ for 2 min., two times in 2×SSC for 5 min. and a post-fixation in 1% formaldehyde in PBS, pH 6.8. We then washed the slides in 2×SSC, dehydrated them through an ethanol series 70%-90%-100% for 3 min of finally let them air dry. The hybridization mixture (20 µL per slide, containing 50% formamide, 2×SSC, 10% sodium dextran sulfate, 50 mM phosphate buffer pH 7.0, 1-2 ng/µL probe DNA and 50-100 ng/µL salmon sperm DNA) was added onto the pretreated chromosome preparations and heated to 80°C for 2 minutes to denature the probe DNA and the chromosomal DNA. The hybridization on the slide was allowed to proceed at 37°C overnight, followed by post hybridization washes for 3×5 minutes in 50% formamide, 2×SSC pH 7.0 at 42°C, 5 minutes in 2×SSC at room temperature, 3×5 minutes in 0.1×SSC at 56°C and 5 minutes in 2×SSC at room temperature. Detection and amplification was according to the manufacturers protocols (Boehringer Mannheim). Digoxigenin-labeled probes were detected with anti-digoxigenin-fluorescein and amplified with rabbit-anti-sheep-fluorescein (F135, Nordic). Biotin-labeled probes were detected with Streptavidin Cy5 and amplified with Biotinylated Anti-Streptavidin and Streptavidin Cy5.

FISH chromosomes were counterstained in 5 µg/mL DAPI in Vectashield anti-fade (Vector Laboratories, http://www.vectorlabs.com). Slides were examined under a Zeiss Axioplan 2 imaging photomicroscope (http://www.zeiss.com) equipped with epifluorescence illumination, and appropriate small band filters for all fluorescence dyes used in this study. Selected images were captured using a Photometrics Sensys 1305×1024 pixel CCD camera (Photometrics, http://www.photomet.com). Image process-

Table 3: PCR conditions for all FISH experiments

	FISH on Genomic template multiple foci	FISH on BAC template	Stringency genomic DNA	FISH on Genomic template (gel extracted & Nested)
PCR	98°C 2min	98°C 2min	98°C 2min	98°C 2min
	98°C 10sec	98°C 10sec	98°C 10sec	98°C 10sec
	67°C 10sec	67°C 10sec	58°C 20sec	67°C 20sec
	72°C 20sec	72°C 20sec	72°C 20sec	72°C 20sec
	75°C 5min	75°C 5min	75°C 5min	75°C 5min
	4°C ∞	4°C ∞	4°C ∞	4°C ∞
cycles	35x cycles	35x cycles	35x cycles	35x cycles
PCR mix	10.2µL sterile MQ	10.2µL sterile MQ	10.2µL sterile MQ	25.5µL sterile MQ
	4 µL 5x buffer	4 µL 5x buffer	4 µL 5x buffer	10 µL 5x buffer
	0.4 µL 10mM dNTP's	0.4 µL 10mM dNTP's	0.4 µL 10mM dNTP's	1 µL 10mM dNTP's
	2 µL primer A 5 µM (1st set of primers!)	2 µL primer A 5 µM (1st set of primers!)	2 µL primer A 5 µM (1st set of primers!)	5 µL primer A 5 µM (1st set of primers!)
	2 µL primer B 5 µM (1st set of primers!)	2 µL primer B 5 µM (1st set of primers!)	2 µL primer B 5 µM (1st set of primers!)	5 µL primer B 5 µM (1st set of primers!)
	1 µL template DNA	toothstick template DNA	1 µL DNA	2.5 µL DNA
	0.6 µL DMSO (3%)	0.6 µL DMSO (3%)	0.6 µL DMSO (3%)	1.5 µL DMSO (3%)
	0.4 µL Polymerase	0.4 µL Polymerase	0.4 µL Polymerase	1 µL Polymerase
Obs	**) all samples annealing 67°C, except 1 at 64°C. PCR products purified with Invitrogen K3100-01 kit*	*) toothstick put into the glycerol stock, no BAC DNA isolation performed!		**) all samples annealing 67°C, except 1 at 64°C PCR products purified with Invitrogen K3100-01 kit*
	***)Phire Hot Start II DNA polymerase – Thermo Scientific*	**)Phire Hot Start II DNA polymerase – Thermo Scientific		***)Phire Hot Start II DNA polymerase – Thermo Scientific*

ing and thresholding were done with the Genus image analysis software (Applied Imaging Corporation, http://www.aicorp.com) and Photoshop (Adobe Inc., http://www.adobe.com).

Results

We developed primer pairs for chromosome 12 of *Solanum lycopersicum* based on the assembly version 2.40. Chromosome 12 was chosen for having the highest frequency of repeats, with a total of 62,363 repeats amounting to 47.4 Mbp out of the 64.5 Mbp length of the chromosome (i.e. 72.3% of the length, while the median for all chromosomes is 69.3% ± 3%, calculated by dividing the total length of annotated repeats by the length of the chromosome), consistent with The Tomato Genome Consortium (2012). Most of the repeats are Long Terminal Repeats (LTR) covering a median of 37.9% ± 0.9% of the genome (37,079 repeats amounting to 26.1 Mbp), LTR/Gypsy (22.8% ± 0.4%; 15,983 repeats amounting to 16.4 Mbp) and LTR/Copia (6.3% ± 0.3%; 6,901 repeats amounting to 4.4 Mbp), the remaining being SINE and other repeats.

We tested the primers both in BAC and genomic templates, with and without PCR nesting as described in Table 1. Figures 1 and 2 show the genomic landscape for which we designed the two successful primer pairs, indicating the regions chosen as suitable for primer design given their desirable characteristics, based on distance from contig and scaffold borders, lack of gaps and uniqueness. Table 1 shows the primers designed, their target coordinates and amplicon sizes, table 2 shows the primer sequences and ta-

Primer Name	Sequence
ch12-sc02-1f	CAACTAACCTTCTGGCTNGTGCTTTCTCCTCCCCTCAGTCTTTTGGTGCTTTTATTGCAGTTGCAAGATATNCAAGTGCC ACTCCCATGCCACCATCCATACTCAGCCGTGGCGTAGGTTGGCAGGTAAACAGATGCTGCCCCCCCTGTAAACTCCCCTCC CTAGCCCAACTTTCTAATACAATGACTTGAAAGGGAAAACGCAAACTTTTTTGTTATTTTATTGAATTTGAACGCAGAAT CAAATACATAGGGTGGTTTGAAAGGTTTATATGGAGAGTCATGACATTGTTTTCAGTCTTCCCAGATTAGCCTTTTCCAC CAGAAGATTTCTTCAGCTCATACCCCAAAAGTTCTTCAGCTAGTACCCCAAGTCTGTCACATACATGCCTCTGTTTTTAT TTATTTAGGGGAGGTGGGTGGGGGTGGGGTGTATATGGGCATATATGGAAAGACAATTAATTTGGAATTCTCAGTTTGAT TCACGAGAAGAATGATTGTTCTTGCACTTCCTCATTCCGGTCTGTTACCAACATGCTCCATTATTATCATCATATGACGA ACCTTTTTGCTGTATTTACTTTCCAAATTCTTGTTATGTCAGGGAATTGGCATTTTACTGTCGGGTTTGTTCGGCACCGG GAATGGATCTTCTGTATCTGTGTAAGTACAGGCGAAATCTTCTTCAGCACAATAGGATGGTAGAAATTAGTTATCTAAGA CGAATTAACTCTCTTCATACACATGCAGTGAAAATGCTGGTCTTTTAGCACTAACACGTGTTGGCAGCCGAAGAGTTGTT CAGATATCAGCAGGGTTTATGATTNNCTTTNCAATTCTTGGTAATTTNCCTGTCCAAGTATTTNGTNGAGTAAATTNNCT TGATTTNCGCTGATATTTAATCCCTCCCCCANCTCCCTCTTNNGCATGTTATNNATCTAAATGCACNTTGGTNTCTTTCA GGAAAAATTNGGGNGCCGNCTTNGCTTCGATTCCANCACCCATNGTAGGNGCCCCCTATNGCCTATTCNTTNGGCTTATG TAGGNNNCNTAAANATANNAANGGNAANCCCTGNNGTNANGAT
ch12-sc02-1r	ATGCGACTAGGAACCAAAGTTGCTATATTNCNACTCAAATTCACATATCCATGAATTTACAACAGAAAAACAATTTACTA AGTAACTTCAAAATGGTCAGACTAGTTATACTTTCCTTTCTTTTGCATACATATAAAAGAAAATATGTGTAGCAGAAGTG AGGCCCCAATGAAGAATTCATTGACAAGGGGTTTTATCGTGTATGAACCGTTTTGTTTTGTTCCTCAATATGATCTCATA CTGACGGGAAATACTTGTTGAGATTGAACGGAAGAGAGTAGAATTCCTCACTTCTTGTGTCACTCTTGAAGGACCGGAAC TTGTTCCACCACGGCTTGCCTCTGTCTTTCTTTGTCGAACTCTCCTTTTTGTGCATTGTATTGTCCAGGAAATAAGCCAA TATGCCAGCAACAAAAGCTTCTGATGAGAAGGGCACGTTCACCATGTCGTTGAACTGAAGAATAACAACAAAGCACAAGT AACTCTTTTCAGCCACCTCTCAAAAATGAAAATGTTAGACATGTATTAAGCTAGTGTTAATTATTGCTATACCCATCGTC CATGTGTGTGAACAGGTCCGTAACCAGCAATCACAGTGTATTCATTGAAGTACTGCGGTACTGATAAACCCAAAAAGATT GAAAAACCTAATATAAACTTGTTACGGAAACTATTAAGATTGCAGAACTGAAGAAAGCTTAAGCCTCCAACCCCTACAAA CAAAGAGAACAAGTAATCACTAACTCCTCATAACGTGAACCACATATTGGTAATGATCATCACACAGGGCTTCCCTTGTA TATTTACGTACCTACATAAGCAAAGAATAGGCAATAGAGAGCGCCTACAATGGGTGCTGGAATCGAAGCAAAGACGGCAC CAAATTTNCCTGAAAAAGAAACAAAAGNGCATTTAGATAAATAACATGCAAAAGAGGAGCTGGGGAGGGATAAAATATCA NCGAAAATCAAGAAAATTNACTCAACAAAATACTNGGACCGGGAAAATTACCNNNAATNGNAAAGAAANTCNTAAACCCN GCTGANTNTGNACNANTCTTCGGGNNGGCCAACACGNGTTN
ch12-sc02-2f	CACTAANCTTCTTGGCTNGTGCTTTCTCCTGCCCTCAGTCTTTTGGTGCTTTTATTGCAGTTGCAAGATATGCAAGTGCC ACTCCCATGCCACCATCCATACTCAGCCGTGGCGTAGGTTGGCAGGTAAACAGATGCTGCCCCCCCTGTAAACTCCCCTCC CTAGCCCAACTTTCTAATACAATGACTTGAAAGGGAAAACGCAAACTTTTTTGTTATTTTATTGAATTTGAACGCAGAAT CAAATACATAGGGTGGTTTGAAAGGTTTATATGGAGAGTCATGACATTGTTTTCAGTCTTCCCAGATTAGCCTTTTCCAC CAGAAGATTTCTTCAGCTCATACCCCAAAAGTTCTTCAGCTAGTACCCCAAGTCTGTCACATACATGCCTCTGTTTTTAT TTATTTAGGGGAGGTGGGTGGGGGTGGGGTGTATATGGGCATATATGGAAAGACAATTAATTTGGAATTCTCAGTTTGAT TCACGAGAAGAATGATTGTTCTTGCACTTCCTCATTCCGGTCTGTTACCAACATGCTCCATTATTATCATCATATGACGA ACCTTTTTGCTGTATTTACTTTCCAAATTCTTGTTATGTCAGGGAATTGGCATTTTACTGTCGGGTTTGTTCGGCACCGG GAATGGATCTTCTGTATCTGTGTAAGTACAGGCGAAATCTTCTTCAGCACAATAGGATGGTAGAAATTAGTTATCTAAGA CGAATTAACTCTCTTCATACACATGCAGTGAAAATGCTGGTCTTTTAGCACTAACACGTGTTGGCAGCCGAAGAGTTGTT CAGATATCAGCAGGGTTTATGATTTTCTTTTCAATTCTTGGTAATTTTCCTGTCCAAGTATTTTGTTGAGTAAATTTNCT TGATTNNCGCTGATATTTNATCCCTCCCCANNTCCTCTTTNGNATGTNATTTAATCTAAATGCACTNGTTTCTTTNCAGG AAAATNNGGTGNCGNCTTGGNTTCGATTCCANCACCCATNGTAGGGGCNCTCTATN
ch12-sc02-2r	ATGCGACTAGGAACCAAAGTTGCTATATTGCCCACTCAAATTTTCNATATCCATGAATTTACAACAGAAAAATNCTTTAC TAAGTAACTTCAAAATGGTCAGACTAGTTATACTTTCCTTTCTTTTGCATACATATAAAAGAAAATATGTGTAGCAGAAG TGAGGCCCCAATGAAGAATTCATTGACAAGGGGTTTTATCGTGTATGAACCGTTTTGTTTTGTTCCTCAATATGATCTCA TACTGACGGGAAATACTTGTTGAGATTGAACGGAAGAGAGTAGAATTCCTCACTTCTTGTGTCACTCTTGAAGGACCGGA ACTTGTTCCACCACGGCTTGCCTCTGTCTTTCTTTGTCGAACTCTCCTTTTTGTGCATTGTATTGTCCAGGAAATAAGCC AATATGCCAGCAACAAAAGCTTCTGATGAGAAGGGCACGTTCACCATGTCGTTGAACTGAAGAATAACAACAAAGCACAA GTAACTCTTTTCAGCCACCTCTCAAAAATGAAAATGTTAGACATGTATTAAGCTAGTGTTAATTATTGCTATACCCATCG TCCATGTGTGTGAACAGGTCCGTAACCAGCAATCACAGTGTATTCATTGAAGTACTGCGGTACTGATAAACCCAAAAAGA TTGAAAAACCTAATATAAACTTGTTACGGAAACTATTAAGATTGCAGAACTGAAGAAAGCTTAAGCCTCCAACCCCTACA AACAAAGAGAACAAGTAATCACTAACTCCTCATAACGTGAACCACATATTGGTAATGATCATCACACAGGGCTTCCCTTG GTATATTTACGTACCTACATAAGCAAAGAATAGGCAATAGAGAGCGCCTACAATGGGTGCTGGAATCGAAGCAAAGACGG CACCAAATTTCCTGAAAAGAAACAAAGTGCATTTAGATAAATAACATGCAAAAGAGGAGCTGGGGAGGGATAAAATATCA NCGAAAATCAGAAAATTTACTCCANCAAAAATACTGGGACNGGAAAATTNCCCNGAATTGGANAGAAAATCNTT

Table 4: Sequenced amplicons and primer names.

ble 3 describes which FISH condition was used to hybridize the primer pairs. From the six primer pairs developed, only two showed unique FISH signals in both BAC templates and genomic templates (with and without nested PCR). The amplicons from some of the primer pairs were sequenced using Sanger and random primers (Table 4).

In this manuscript we analyze in depth these two scaffolds, SL2.40sc04878 and SL2.40sc05611. Scaffold SL2.40sc04878 is a 5.7 Mbp long scaffold starting at position 22.5 Mbp of chromosome 12 with a median *k*-mer coverage of 0.348 ± 0.356, of which 12.67% of

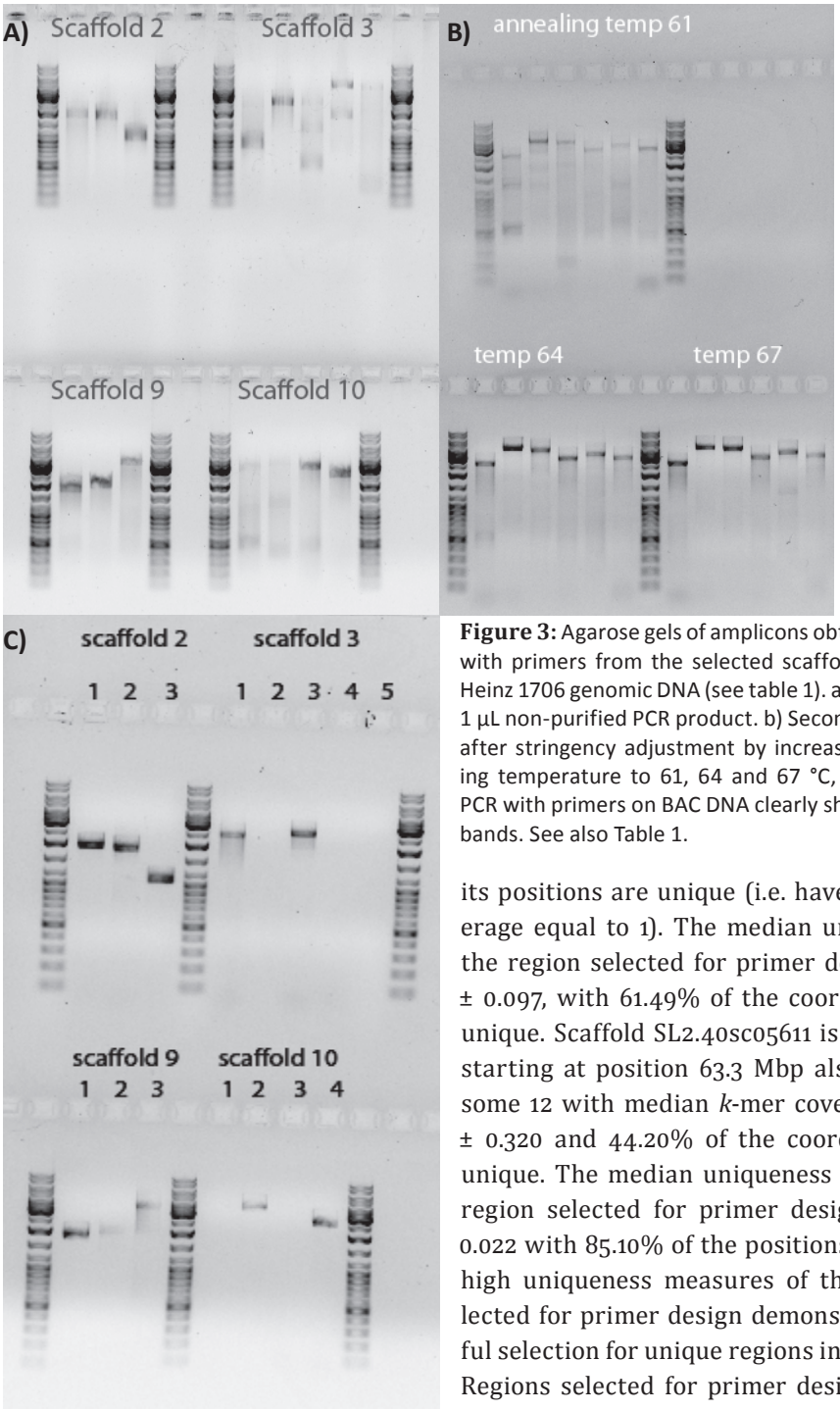


Figure 3: Agarose gels of amplicons obtained from PCR with primers from the selected scaffolds on genomic Heinz 1706 genomic DNA (see table 1). a) First PCR from 1 µL non-purified PCR product. b) Second PCR obtained after stringency adjustment by increasing the annealing temperature to 61, 64 and 67 °C, respectively. c) PCR with primers on BAC DNA clearly shows single DNA bands. See also Table 1.

its positions are unique (i.e. have a *k*-mer coverage equal to 1). The median uniqueness for the region selected for primer design is 1.000 ± 0.097, with 61.49% of the coordinates 100% unique. Scaffold SL2.40sc05611 is 1.2 Mbp long, starting at position 63.3 Mbp also of chromosome 12 with median *k*-mer coverage of 0.957 ± 0.320 and 44.20% of the coordinates 100% unique. The median uniqueness index for the region selected for primer design is 1.000 ± 0.022 with 85.10% of the positions unique. The high uniqueness measures of the regions selected for primer design demonstrate successful selection for unique regions in the scaffolds. Regions selected for primer design had a me-

dian size of 2,604 bp ± 119 bp since the theoretical minimum size detectable by FISH is around 2 Kbp (Lamb, *et al.*, 2006). Amplicons for scaffolds SL2.40sc04878 and SL2.40sc05611 were, respectively, 1,848 bp and 3,543 bp long.

Table 5: Nested PCR primers designed to reduce smear.

Primer Name	Primer Sequence
ch12-sc02-4f	ACAATTTCCCTTGTTAGTGGGTATGTTGGC
ch12-sc02-4r	GTGCTGGAATCGAAGCAAAGACGGCAC
ch12-sc02-5f	ACGTTTCCCCTCACAAGCATAGAATTTCC
ch12-sc02-5r	CCATATAGTTCTCGCCCGTTCGGC
ch12-sc02-5f	TGCACAACCCTCAATACAAACATGAAAGTC
ch12-sc02-5r	TGAACAAAAAGTGCAGTACGATTGGAAGCC
ch12-sc03-6f	AGGAAGCAAAGGTAGTTTGGGATGAGTAGG
ch12-sc03-6r	TCCCTTGTACCTTGTCTTCAAACCTCTTC
ch12-sc03-7f	AGAAAGCTCCCATGATTGTAGAGCATCC
ch12-sc03-7r	AGCCTGTTTAGTTTTCTGATAGTCCACCC
ch12-sc09-5f	GTTGGACAGTGATTGGTGGGCAACGG
ch12-sc09-5r	TGCCCTACCAGTTCATTCAGACTTGCATCC
ch12-sc09-6f	TGCTGAGCTGGACAACAGAGCTACTATGC
ch12-sc09-6r	TCCTGAGGCACTAAAGAAAGCTCAAAACG
ch12-sc10-5f	ACTGATTTCTATCCTGAAGCAGCAGAGTGC
ch12-sc10-5r	CTCATGCTTTTCCTGCTAATGTTTCACCTG
ch12-sc10-6f	AATTTCACATGCTGGTGTACTGATGAGAGC
ch12-sc10-6r	GCTTGAGCCTTGACATTGCGGAGGAG

Genomic DNA obtained from Heinz leaves was first tested on 1% and 2% agarose gels. High-quality genomic DNA appeared as distinct bands at 10-20 kb without smear. All genomic DNA samples gave 260/280 ratios of 1.9 and 260/230 of 2.2. Scaffolds and primers that are used in the experiments are described in Tables 1 and 2. In the first series of PCR reactions, we tested various primers from different scaffolds on genomic DNA and set the annealing temperature at 55 °C. Gel electrophoresis of the PCR products showed multiple DNA bands, suggesting that different amplicons were formed (Figure 3a). When increasing PCR stringency with the annealing temperature elevated up to 67 °C, respectively, we obtained the same number of multiple bands on the gel (Figure 3b). In a second series of PCR amplifications we used the corresponding BAC DNA as target for the PCR reaction. The BACs have been selected to contain the amplified DNA sequences in the corresponding PCR reaction (Table 1). The PCR products now obtained show single bands on agarose gels with sizes corresponding to the expected amplicon size (Figure 3c, Table 1).

Hybridization of the probes on chromosomes gives clear fluorescent foci on the chromosomes with little or no background signals. Amplicons obtained from PCR on genomic DNA revealed multiple foci in several experiments with FISH on pachytene complements in cell spreads of Heinz 1706 tomato pollen mother cells (Figures 4a) and correspond to the multiple DNA bands seen on gel (Figure 3a). Even raising the annealing temperature up to 67°C did not result in the expected single foci on the chromosomes.

Under similar conditions, we also extracted DNA from single bands in the gel and probed the DNA biotin and digoxigenin and detected them with streptavidin – FITC and anti-digoxigenin – Cy5. Figure 4b shows a detail of a pachytene complement in which a strong reduction of foci is apparent. In a next step we continued with gel extracted DNA, which was used for an additional nested PCR (Table 5). The green fluorescent FITC signals were obtained from labeled SC2-1 amplicons. The red Cy3.5 is from the control BAC

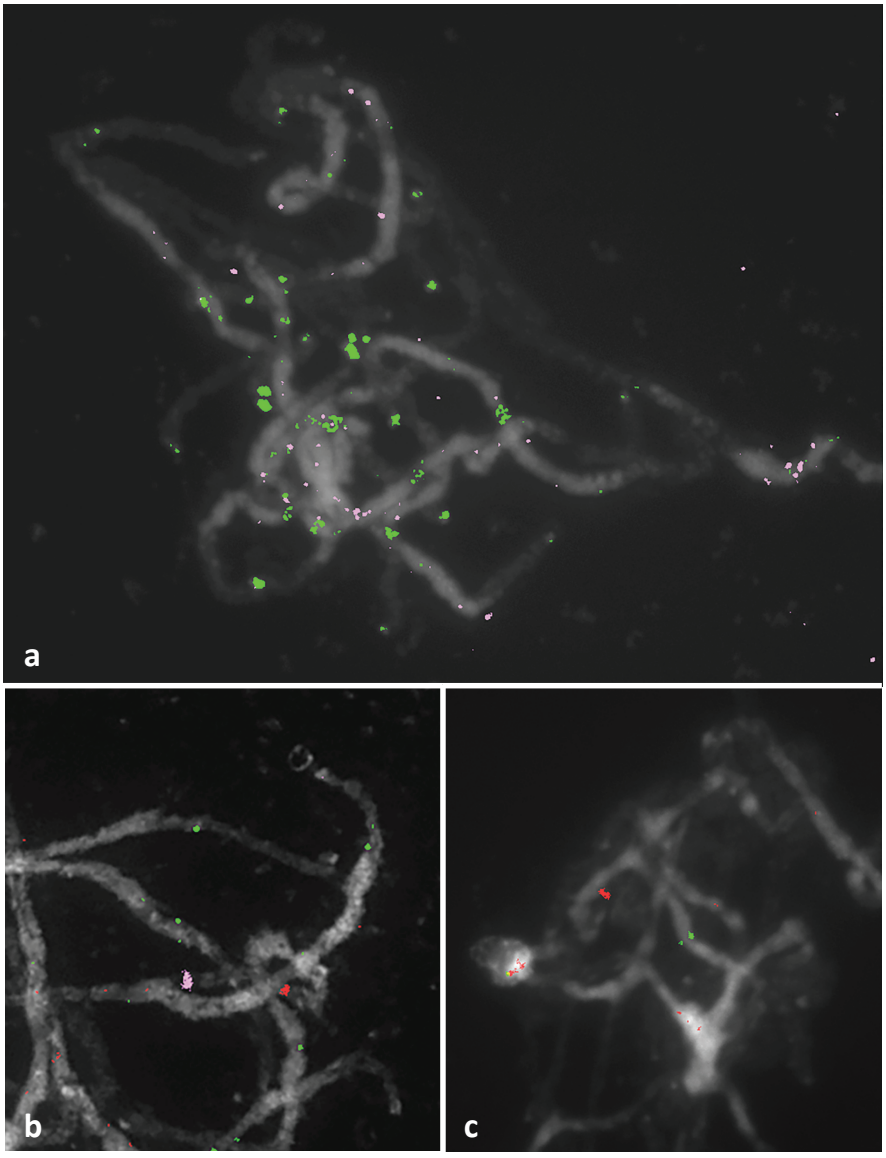


Figure 4: FISH of PCR amplicons on Heinz 1706 pollen mother cells at pachytene. a) Hybridization of pooled PCR amplicons from primers SC2-1, SC2-2 and SC2-3 (Cy5, purple fluorescence) and from primers SC3-1, SC3-2, SC3-3, SC3-4 and SC3-5 (FITC, green fluorescence) on genomic Heinz 1706 genomic DNAs. The amplicons were obtained by PCR at high stringency with annealing temperature of 67ºC, giving multiple DNA bands on the agarose gel. b) Same hybridization, but DNA was gel extracted. FITC foci from primers SC2-1, Cy5 foci from primers SC9-4. The Cy3.5 signal from BAC Mbo_011_A16 is used as an internal control for chromosome 12. c) Same hybridization, but probe DNA was obtained from gel extracted and nested PCR. See materials and methods for details. The green fluorescent FITC signals come from SC2-1. The red Cy3.5 is from the control BAC Mbo_011_A16.

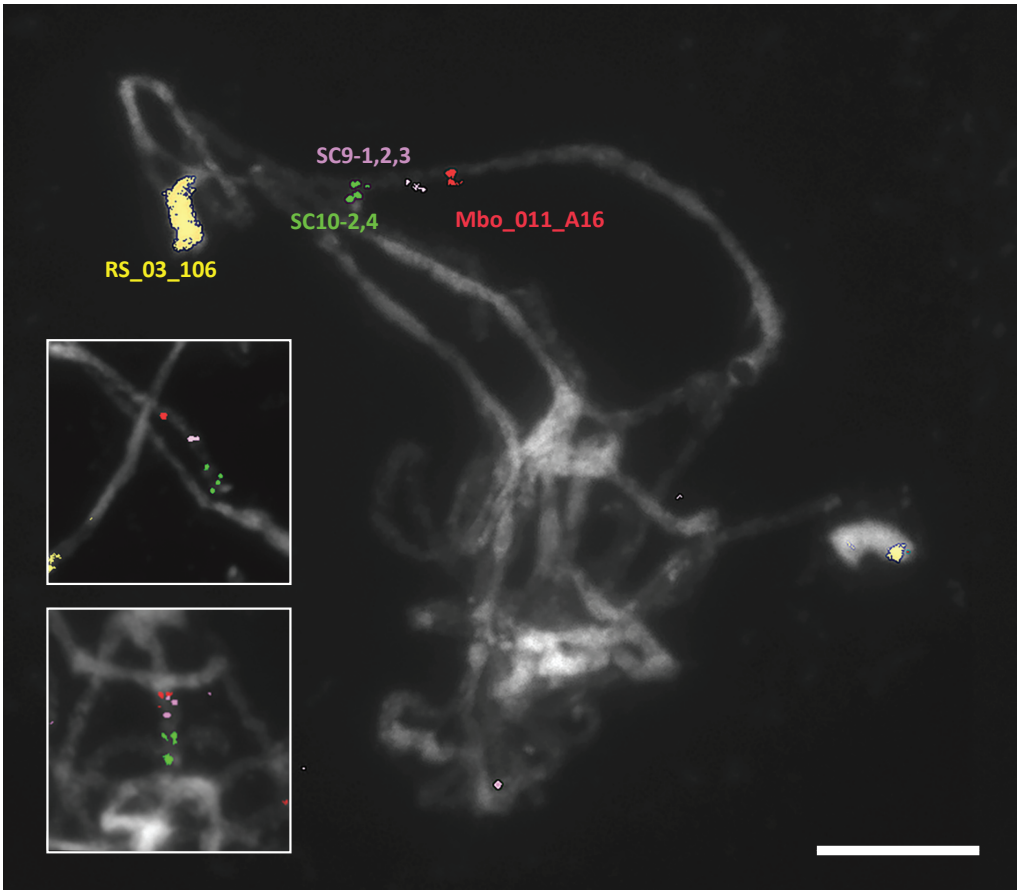


Figure 5: PCR-FISH painting on tomato pollen mother cells at pachytene. The probes were obtained from PCR amplicons with single copy primer sequences on their corresponding BAC templates of chromosome 12 (see Table 1). The purple foci denote the Cy5 labeled amplicons of SC9-1, SC9-2 and SC9-3, the green FITC foci are hybridizations of the amplicons SC10-2 and SC10-4, the orange fluorescent probe is from BAC RS_063_I06, which unexpectedly hybridizes with the complete nucleolar organizer region of chromosome 2. The red fluorescent Cy3.5 signal is from BAC Mbo_011_A16 and is used as a control for chromosome 12. The two insets show the same pattern of the three amplified scaffolds from two other pollen mother cell spreads at pachytene.

Mbo_011_A16. The yellow Cy3 of the second reference BAC (RS_063_I06) unexpectedly painted the NOR and was not used in further experiments.

Description of PCR amplicons with single copy primer sequences on their corresponding BAC templates of chromosome 12 is shown in Table 1. The purple foci denote the Cy5 labeled amplicons of SC9-1, SC9-2 and SC9-3, the green FITC foci are hybridizations of the amplicons SC10-2 and SC10-4, the orange fluorescent probe is from BAC RS_063_I06, which unexpectedly hybridizes with the complete NOR of chromosome 2. The red fluo-

rescing Cy3.5 signal is from BAC Mbo_011_A16 and is used as a control for chromosome 12. The two insets show the same pattern of the three amplified scaffolds from two other pollen mother cell spreads at pachytene.

Discussion

In this study we have shown that selection of single copy sequences in the assembled genome of tomato enables production of amplicons that can be labeled as probes for microscopic detection of such sequences on a chromosomal target. The method is relatively fast, simple and provides a powerful alternative to the BAC FISH painting technique we used previously in similar studies. The major drawback of the technique is that amplicons obtained from PCR on genomic DNA templates produce multiple PCR fragments on gel and, as expected, exhibit various fluorescent foci on different chromosomes in the cell complement. In contrast, if BAC DNA was used as a PCR template, the expected single chromosomal foci was observed. An explanation for the multiple bands obtained from PCR on total genomic DNA is that some smaller parts of the tomato genome, especially the satellite and nucleolar organizer region of chromosome 2, are simply not complete enough and hence may contain DNA sequences that were not deleted during the computational filtering. The possibility of such unknown small undetected repetitive sequences in the amplicons may be higher in the larger PCR products of several kb that we have used for our painting study.

Furthermore, considering that the assembled reference genome sequence is still incomplete, missing approximately 200 to 250 Mb (20-30% of the total length) of highly complex unassembled sequence, a false non-repetitive assignment to *k*-mers, especially for those *k*-mers that are mapped to inherently problematic regions such as contig ends, is possible to occur. However, due to the incompleteness and assembly errors in the reference genome such errors cannot be completely avoided at the moment. Although we did not test it, an alternative approach might be assessing the *k*-mer frequencies from sequence read data instead of the assembled genome sequence as well as using a more robust method for calculating the uniqueness index. Ma *et al.*, (2010) Finger Printed (FP) contig-associated genomic sequences were analyzed for *k*-mer sequence statistics, leading to repeat identification and subsequent masking called Kmasker (Schmutzer *et al.*, 2014) for the development of unique FISH probes using *k*-mers, a tool which claims to have a 98.7% success rate which differed from ours by targeting exclusively genic regions, which requires a genome annotation but seems to be an effective approach for avoiding repeats.

One of the most striking results was the fact that one of the scaffolds (RS_063_I06, Figure 5) shows a full staining of the NOR region of chromosome 2. The short arm of chromosome 2 is almost completely missing from the genome assembly of tomato and can be responsible for the large fraction of the genome size discrepancy mentioned earlier and the relative large size of the chromosome 0, the collection of un-scaffolded con-

tigs. This led us to believe that the short arm of chromosome 2, and specifically the NOR region in it, might contain representatives of several classes of repeats, including those found in our scaffold.

In conclusion, our method successfully selected unique regions in the scaffolds of interest, but its success rate in the hybridization shows that our method was not sufficient to ensure uniqueness relative to the unknown parts of the genome (unsequenced or collapsed repeats) but the probes that did work were able to be amplified directly from gDNA without the need for BAC templates. It is our belief that with the improved bioinformatics methods now available it is possible to create unique probes for genome scaffolding completely without BAC libraries using exclusively NGS-WGS assembled sequences.

Chapter 6

General Dicussion

This thesis encompasses the generation of a large dataset of tomato genomes which will serve the plant science and breeding communities to better understand this crop, fleshy fruits, and other plants in general. In order to analyse this data many pipelines, computational methods were developed and reported upon. While the work presented in this thesis can help researchers and breeders to understand and study plant diversity and improve plants by breeding, in order to arrive at true "breeding by design" a number of technological and methodological innovations will be needed. Specifically, improvements are possible in the technology we employ to measure the natural genetic and structural variation in genomes and the computational methods that we use to process, analyse and interpret the large amounts of data that this technology produces. Below, I will discuss these potential improvements in more depth.

Importance of plant breeding and natural variation information

Plant breeders face many challenges in their line of work. Among those is the constant need to innovate. New customer preferences, higher nutritional value, the need for resistant crops and increased (a)biotic stress tolerance are some of the reasons for the need to develop innovative crops with new traits. Here the knowledge of the natural variation and the exploitation of the genetic diversity of available germplasm is the most powerful tool at the disposition of breeders. Even directed genetic modification (cis-, and specially trans-genesis) requires information from natural diversity.

Until recently, information on the genomic diversity of plants was restricted to a few model species. This changed with new technological developments achieved in the last decade, permitting high-throughput sequencing at a fraction of the costs of traditional Sanger sequencing. Today it is possible to sequence many varieties of a species in search of all available diversity. Having a genome sequence, combined with phenotypical information, permits the development of molecular markers associated to traits. Such markers are indispensable in modern plant breeding where a simple and inexpensive set of PCRs in young plantlets is sufficient to identify the underlying genes of a trait without the need to grow the plant to maturity.

The high SNP density in a dataset of common genetic variants created through Whole Genome Sequencing (WGS) by Next Generation Sequencing (NGS-WGS) will increase the resolution of statistical methods created to analyse sparse sets of genetic markers created by technologies such as Genotyping by Sequencing (GBS). These powerful statistical methods, when allied to high density SNP maps, can help in the identification of hot and cold spots of recombination with high accuracy. This information can help in the development of more efficient breeding strategies, reducing the time and cost of such efforts. Nevertheless, fuelled by the recently NGS-based technological advances made and similar advances expected in the near future, whole genome re-sequencing is a competitive alternative and eventually will outcompete approaches like GBS. Moreover, non-

targeted NGS approaches are much more suited for comparative studies involving large-scale sampling of non-model species than targeted, labour intensive, approaches such as GBS (Rossetto and Henry, 2014).

A special group of challenging organisms are the polyploid species, which despite their economic importance, cannot be studied with most of the tools that are used for diploid organisms due to their added genomic complexity. With the very long read technologies becoming available and more affordable, it becomes possible to generate large haplotypes, increasing our capacity to study and understand polyploidy, creating a demand for new tools which can process and visualize more complex genome structures.

Analysis of epigenetics faces similar problems, lacking high throughput biochemical/biophysical methods of sequencing. Because of the small number of available datasets, there is a proportionally low number of programs to analyse this type of data. Fourth generation sequencing techniques (such as Oxford Nanopore) promise to change the data acquisition problem by reporting the nucleotide sequence and its methylation state at the same time and, with this, data analysis tools should be developed to accommodate this extra level of information.

With the spreading usage of NGS for routine genotyping in breeding programs, the use of specifically designed immortal populations should decline as GWAS statistical methods could be used to identify new QTLs and causal SNPs in regular F1 populations due to the high number of individuals being sequenced and the diminished problems caused by heterozygosity by the usage of longer, higher quality, reads. When this dense genotyping data is associated with mass haplotyping and phenotyping, functional understanding of SNPs will extend to explain epistatic interactions and heterosis at unprecedented levels; crucial to a true breeding-by-design program in which all necessary breeding steps (parental, intermediate and final genotypes) can be defined prior to the start of the breeding program.

For this to happen, high throughput methods of phenotyping have to be created to measure the large amount of parameters important to the final product. Allied to measurement devices, ontologies will have to be created to standardize the comparison of measured features among experiments and institutions. New algorithms have to be developed to convert all available data into information and to convert this information into a hypothesis which can be tested to validate the mathematical model.

Chromosome structure

Knowledge on the chromosome structure of a species, its morphological variations and potential *de novo* rearrangements is essential for a successful breeding program. There are several methods to identify structural differences between genomes such as genome and Bacterial Artificial Chromosome (BAC) FISH painting, and study of chromosome pairing in interspecific hybrids. In the context of genome sequencing, BAC FISH is the

most useful since the same library can be used for sequencing and for FISH, facilitating the combination of both sources of information. Once a BAC library is created, BACs are hybridized against the genome in a process called minimum tiling path (Cviková *et al.*, 2015) which tries to determine the minimum set of BACs which can cover most of the genome having some overlap between them. Once such a set of BACs are chosen, they are either individually sequenced (BAC-by-BAC sequencing) or have their 3' and 5' ends sequenced (BAC-end sequencing).

However, BAC-by-BAC sequencing has fallen into disuse due to its costs. BAC-end sequencing is also decreasing in use because the costs of creating and maintaining a BAC library can be as large as the cost of sequencing. Recently, Hi-C has been presented as an alternative to BAC-end sequencing for scaffolding genomes (Burton *et al.*, 2013). It is based on cross-linking the genome, trimming the free DNA and sequencing the remaining DNA. This remaining DNA represents the parts of the genome that are bound together by proteins at the time of crosslinking. By performing a statistical analysis of the set of segments that are sequenced together, it is possible to infer the proximity between these segments in the genome. When mapping such segments to the unfinished genome assembly, this information can be used to further order the scaffolds into super scaffolds. Hi-C does not exactly provide scaffold orientation and, in the absence of physical markers, it is not able to identify centromeres, Nucleolus Organizer Regions (NORs) or telomeres, all important chromosomal features for breeding.

A finer grained scaffolding method which is able to identify chromosomal structure is PCR-FISH on pachytene chromosomes, which is able to give order and orientation to scaffolds without the need for a BAC library. As a bridge between sequencing and BAC-FISH, we have developed a PCR-FISH method for tomato in which labelled amplicons are hybridized against the nuclear DNA. We designed unique PCR primers at the borders of assembly scaffolds, subsequently label the unique PCR product with fluorophores and then visualize them on chromosome preparations with FISH. Furthermore, markers for centromere, Nucleolar Regions and telomeres can be added and visualized by FISH, creating a more complete understanding of the genome organization.

With the recent advent of extra-long reads (PacBio, Eid *et al.*, 2019; Illumina TruSeq, Li *et al.*, 2015; and Oxford Nanopore, Eisenstein, 2012) and genome mapping technologies such as Bionano genome mapping (Levy-Sakin and Ebenstein, 2013), the number of scaffold breaks can be reduced drastically (Faino *et al.*, 2015). This more manageable number of discontinuities can be ordered and oriented in relation to each other (and to the chromosome features) using simple primers without the added cost of a BAC library. By using an assembly strategy that aims to create pseudo-molecules with a combination of long reads, genome mapping, Hi-C and PCR-FISH, it is possible to envision the routine creation of high quality assemblies with completeness comparable to what is considered a reference genome. Combined with methylation analysis and single cell sequencing, this

will generate a deeper understanding on chromosome structure variation between cells, organs, organisms, species and cancerous tissues.

Genome analysis tools

The high throughput generation of data brings challenges both in data analysis and data visualization. Data analysis has to be adapted to the constraints of computational resources and time. Data visualization has to be adapted to relay the result of the analysis while allowing the comparison of a large number of samples and dimensions at once.

Data analysis has experienced a revival with programs being re-written to yield a higher throughput, such as NCBI which released a faster version of its BLAST software suit (BLAST+, Camacho *et al.*, 2009). At the same time, efficient storage methods have been developed specifically for biological data. One of such methods is the CRAM file format (Fritz et. al., 2011) created to store mapped NGS data with up to 60% compression rate. This trend is expected to grow but, more importantly, programs will need to be designed with high and diverse parallelization methods available, able to take advantage of several computer cores as well as several computer nodes (Cochrane *et al.*, 2012).

With the increase of the data volume, the time and costs of data transfer has grown to be an important factor in project design. Some research groups and governments are already attempting to solve this problem by moving both data and computation to the cloud (Stein *et al.*, 2015). Although the concept of sharing computer power is not a new one, shared data repositories were problematic due to the privacy and secrecy needed to the projects before publication. The solution is to create fine grained permission for data access for authorized personnel. As the data is co-located with the data centre, sometimes placed there directly by the sequencing facility, access to the data becomes quicker, less redundant and cheaper. After publication, the data can also be made available without need to transport it between organisations, saving time and resources. Even private cloud computing companies are already making public biological data available within their data centres to facilitate the usage by clients such as Amazon Web Services, one of the largest cloud computing companies, making the 1000 human genomes available to its clients (1000 Genomes Project and AWS, 2015). In the future we will see that shared, federated, computer power and storage becomes more ubiquitous, while reducing the time to large analysis by using a large pool of computational resources. The Netherlands have already created a network of computer clusters and storage accessible to public and private entities (SURFsara, 2015) and such services should increase, merge and interlink to create a large network of collaboration.

Moreover, visualization has not caught up with the increase in the amount of data generated. The substantial size of the data and the large number of samples requires summarization of the data in biologically meaningful ways, allowing for a general view of the data as well as for the inspection of details, preferably in a single tool. A prob-

lem with such a dynamic visualization program is the difficulty of translating all those visualizations to a bi-dimensional paper for publishing and sharing, limiting its usability (Gehlenborg and Wong, 2012).

Based on this principle of data reduction iBrowser was created (described in chapter 3), which, instead of showing SNPs across the genome, it creates a phylogenetic tree for each consecutive block along a chromosome, displaying a heatmap summarizing the distances encountered from each sample to a reference.

The natural progression of visualization of high dimensional data, such as genomic information, will likely bring a heterogeneous approach where data exploration can be done in three dimensions with multi-depth visualizations interlinked, allowing the user to switch seamlessly between them and to extract meaning. For reporting of the findings the user would have to reduce this data to two dimensions with auxiliary graphics showing all ancillary data side-by-side for small, selected genomic areas. This would allow faster analysis while not hampering science communication.

As for scientific communication, several journals are now allowing for the embedding of dynamic content such as movies and hyperlinks, impractical to be converted to a print media. This trend can be expanded to allow interactive graphics, enabling the reader to explore the data instead of reading the authors interpretation of it. And, as web services become more ubiquitous, the language of the internet is becoming the language of reporting scientific findings. More programs will be adapted to be used in the web browser and their reports web friendly. This would allow for the long sought platform-independent distribution of programs with built-in, feature rich, reporting.

Sample comparison and identification

The exponential growth of NGS projects creates a logistic problem. If not accompanied by automation, the manual handling of hundreds of biological samples, from the greenhouse/environment until they reach the bioinformatician, may lead to sample swaps. Added to that, different plant collections have different database formats, identification codes, nomenclature standards and in some cases lack updates in case of nomenclature changes. Wild species provide us with the added problem of gene flow in sympatric speciation (Lin *et al.*, 2014), which complicates species identification.

Reference-free methods with powerful statistics have been developed which are able to identify introgressions out of the raw sequencing data, which can be pipelined with cheap and high throughput sequencing apparatus for quick species identification. It is also possible to identify pathogens from samples without growing them separately, which is crucial for several fields. For medicine it can be used to identify diseases while it can be used in plant research to identify insects and pathogens. Also, governmental customs can use such methods in an attempt to diminish illegal entering of forbidden, contaminated or infected species. Programs such as Cnidaria are viable tools to be used

as a first step in the identification of the genus of a species found, while a more specific, reference based method and databases can be used for sub-species/strain identification. The new Cnidaria tool which is able to cluster and identify genome and transcriptome data from raw and assembled data is able to cluster at genus level with 100% accuracy and at species level at 78% accuracy even when handling data as phylogenetic distant as fungi, animals and plants in a single analysis.

With the commoditization of sequencing, and even the possible portability of sequencing devices, reference free methods promises the ability of gathering meaningful information for any sample, regardless of whether there is a close relative gold standard. This can possible allow us to sequence organisms from all phylogenetic branches, including never seen ones.

Even for organisms having references to compare with, reference-free methods can serve to speed the comparison and for quick verification of known traits such as the presence of SNPs. This can allow machines to sample, analyse and automatically take pre-determined actions according to the genotype of the sample, all without ever going to the lab or requiring large computer hardware and complicated software pipelines to analyse the data.

Conclusion

What has become apparent in this study and from large genome sequencing projects and cancer-genomics projects is that there is no such thing as 'the genome' for a particular species or organism. Given the technological advances, genome sequencing and determining genetic variation across entire populations and clades using economically feasible, non-targeted NGS sequencing approaches will become a reality in the very near future. When we take this enormous genetic diversity into account many challenges lie ahead of us. From a data production and data processing point of view we will have to come up with new bioinformatic concepts to deal with the vast amount of information it contains. Methods will also have to be developed to provide biological explanations for key biological processes in plant breeding, such as for example the transfer of hereditary traits via meiotic recombination. This phenomenon, which we know exist already from the time of Darwin and Mendel and even before that, when men started to select from wild populations, we still do not fully understand. Genomics provides only one level of understanding with which we try to explain the biological processes on a species level and to a lesser extent on the population level.

Several layers of information, such as coming from the epigenetics level, molecular biology, metabolomics and proteomics, combined with population genetics, mathematics and statistics, and information technology need to be integrated in order to explain and fully appreciate the diversity of life as we know it. Not only the integration itself, but also the communication and understanding between different fields of expertise that

bioinformatics encompasses, has been problematic already from the time that bioinformatics started to emerge, and still is a problem today. This puts a new perspective to how we should translate and communicate information between fields of expertise, educate and train young bioinformaticians in order to tackle the scientific, economic, and societal challenges that lie ahead of us.

References

(2015a) **1000 Genomes Project and AWS, https://aws.amazon.com/1000genomes/**

(2015b) **SURFsara, https://www.surf.nl/en**.

Adams, M.D. (2000) The Genome Sequence of *Drosophila melanogaster*. Science, 287, 2185-2195.

Adams, M.J. and Antoniw, J.F. (2006) DPVweb: a comprehensive database of plant and fungal virus genes and genomes. *Nucleic Acids Res.* 34, D382-D385.

Aflitos, S., Schijlen, E., de Jong, H., de Ridder, D., Smit, S., Finkers, R., Wang, J., Zhang, G., Li, N., Mao, L., Bakker, F., Dirks, R., Breit, T., Gravendeel, B., Huits, H., Struss, D., Swanson-Wagner, R., van Leeuwen, H., van Ham, R.C.H.J., Fito, L., Guignier, L., Sevilla, M., Ellul, P., Ganko, E., Kapur, A., Reclus, E., de Geus, B., van de Geest, H., te Lintel Hekkert, B., van Haarst, J., Smits, L., Koops, A., Sanchez-Perez, G., van Heusden, A.W., Visser, R., Quan, Z., Min, J., Liao, L., Wang, X., Wang, G., Yue, Z., Yang, X., Xu, N., Schranz, E., Smets, E., Vos, R., Rauwerda, J., Ursem, R., Schuit, C., Kerns, M., van den Berg, J., Vriezen, W., Janssen, A., Datema, E., Jahrman, T., Moquet, F., Bonnet, J. and Peters, S. (2014) Exploring genetic variation in the tomato (Solanum section Lycopersicon) clade by whole-genome sequencing. *The Plant Journal*, **80**, 136-148.

Aflitos, S., Schijlen, E., de Jong, H., de Ridder, D., Smit, S., Finkers, R., Wang, J., Zhang, G., Li, N., Mao, L., Bakker, F., Dirks, R., Breit, T., Gravendeel, B., Huits, H., Struss, D., Swanson-Wagner, R., van Leeuwen, H., van Ham, R.C., Fito, L., Guignier, L., Sevilla, M., Ellul, P., Ganko, E., Kapur, A., Reclus, E., de Geus, B., van de Geest, H., Te Lintel Hekkert, B., van Haarst, J., Smits, L., Koops, A., Sanchez-Perez, G., van Heusden, A.W., Visser, R., Quan, Z., Min, J., Liao, L., Wang, X., Wang, G., Yue, Z., Yang, X., Xu, N., Schranz, E., Smets, E., Vos, R., Rauwerda, J., Ursem, R., Schuit, C., Kerns, M., van den Berg, J., Vriezen, W., Janssen, A., Datema, E., Jahrman, T., Moquet, F., Bonnet, J. and Peters, S. (2014) Exploring genetic variation in the tomato (Solanum section Lycopersicon) clade by whole-genome sequencing. *The Plant journal : for cell and molecular biology*, **80**, 136-148.

Aflitos, S.A., Sanchez-Perez, G., de Ridder, D., Fransz, P., Schranz, M.E., de Jong, H. and Peters, S.A. (2015) Introgression browser: high-throughput whole-genome SNP visualization. *The Plant journal : for cell and molecular biology*, **82**, 174-182.

Al-Shahib, A. and Underwood, A. (2013) snp-search: simple processing, manipulation and searching of SNPs from high-throughput sequencing. *BMC bioinformatics,* **14,** 326.

Altschul, S. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, **25**, 3389-3402.

Anderson, C.M., Chen, S.Y., Dimon, M.T., Oke, A., DeRisi, J.L. and Fung, J.C. (2011) ReCombine: a suite of programs for detection and analysis of meiotic recombination in whole-genome datasets. *PloS ONE,* **6**, (10):e25509.

Anderson, E. (1953) Introgressive Hybridization. *Biol. Rev.*, **28**, 280-307.

Anderson, L.K., Covey, P.A., Larsen, L.R., Bedinger, P., Stack, S.M. (2010) Structural differences in chromosomes distinguish species in the tomato clade. *Cytogenet Genome Res*. 129**,** 24-34

Aoki, K., Yano, K., Suzuki, A., Kawamura, S., Sakurai, N., Suda, K., Kurabayashi, A., Suzuki, T., Tsugane, T., Watanabe, M., Ooga, K., Torii, M., Narita, T., Shin, I.T., Kohara, Y., Yamamoto, N., Takahashi, H., Watanabe, Y., Egusa, M., Kodama, M., Ichinose, Y., Kikuchi, M., Fukushima, S., Okabe, A., Arie, T., Sato, Y., Yazawa, K., Satoh, S., Omura, T., Ezura, H. and Shibata, D. (2010) Large-scale analysis of full-length cDNAs from the tomato (*Solanum lycopersicum*) cultivar Micro-Tom, a reference system for the *Solanaceae* genomics. *BMC Genomics,* **11**, 210.

Areshchenkova, T. and Ganal, M.W. (1999) Long tomato microsatellites are predominantly associated with centromeric regions. *Genome*, **42**, 536-544.

Arumuganathan, K. and Earle, E.D. (1991) Estimation of nuclear DNA content of plants by flow cytometry. *Plant Molecular Biology Reporter*, **9**, 229-241.

Austin, R.S., Vidaurre, D., Stamatiou, G., Breit, R., Provart, N.J., Bonetta, D., Zhang, J., Fung, P., Gong, Y., Wang, P.W., McCourt, P. and Guttman, D.S. (2011) Next-generation mapping of *Arabidopsis* genes. *Plant J.*, **67**, 715-725.

Bai, Y. and Lindhout, P. (2007) Domestication and Breeding of Tomatoes: What have We Gained and What Can We Gain in the Future? *Annals of Botany*, **100**, 1085-1094.

Baird, N.A., Etter, P.D., Atwood, T.S., Currey, M.C., Shiver, A.L., Lewis, Z.A., Selker, E.U., Cresko, W.A. and Johnson, E.A. (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PloS ONE,* **3**, e3376.

Barone, A., Chiusano, M.L., Ercolano, M.R., Giuliano, G., Grandillo, S. and Frusciante, L. (2008) Structural and Functional Genomics of Tomato. *International Journal of Plant Genomics*, **2008**, 1-12.

Bauchet, G. and Causse, M. (2012) Genetic diversity in tomato (*Solanum lycopersicum*) and its wild relatives. In *Environmental Sciences* (Calişkan, M. ed). Rijeka, Croatia: InTechOpen, pp 133-162.

Beliveau, B.J., Boettiger, A.N., Avendaño, M.S., JuNGMann, R., McCole, R.B., Joyce, E.F., Kim-Kiselak, C., Bantignies, F., Fonseka, C.Y., Erceg, J., Hannan, M.A., Hoang, H.G., Colognori, D., Lee, J.T., Shih, W.M., Yin, P., Zhuang, X. and Wu, C.-t. (2015) Single-molecule super-resolution imaging of chromosomes and *in situ* haplotype visualization using Oligopaint FISH probes. *Nature Communications*, **6**, 7147.

Beliveau, B.J., Joyce, E.F., Apostolopoulos, N., Yilmaz, F., Fonseka, C.Y., McCole, R.B., Chang, Y., Li, J.B., Senaratne, T.N., Williams, B.R., Rouillard, J.M. and Wu, C.t. (2012) Versatile design and synthesis platform for visualizing genomes with Oligopaint FISH probes. *Proceedings of the National Academy of Sciences*, **109**, 21301-21306.

Bennetzen, J.L. (2000) *Genome Biol,* **1**, reviews107.101.

Bernatzky R., Tanksley S.D. (1986) Genetics of actin-related sequences in tomato. *Theor Appl Genet*. Jun;72(3):314-21.

Bertels, F., Silander, O.K., Pachkov, M., Rainey, P.B. and van Nimwegen, E. (2014) Automated reconstruction of whole-genome phylogenies from short-sequence reads. *Molecular biology and evolution*, **31**, 1077-1088.

Bevan, M. (2001) Sequence and analysis of the Arabidopsis genome. *Current Opinion in Plant Biology*, **4**, 105-110.

Bienko, M., Crosetto, N., Teytelman, L., Klemm, S., Itzkovitz, S. and van Oudenaarden, A. (2012) A versatile genome-scale PCR-based pipeline for high-definition DNA FISH. *Nature Methods*, **10**, 122-124.

Boetzer, M. and Pirovano, W. (2012) Toward almost closed genomes with GapFiller. *Genome Biol*, **13**, R56.

Bombarely, A., Rosli, H.G., Vrebalov, J., Moffett, P., Mueller, L.A. and Martin, G.B. (2012) A draft genome sequence of Nicotiana benthamiana to enhance molecular plant-microbe biology research. *Molecular plant-microbe interactions : MPMI*, **25**, 1523-1530.

Botstein, D., White, R.L., Skolnick, M. and Davis, R.W. (1980) Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *American Journal of Human Genetics*, **32**, 314-331.

Bradnam, K.R., Fass, J.N., Alexandrov, A., Baranay, P., Bechner, M., Birol, I., Boisvert, S., Chapman, J.A., Chapuis, G., Chikhi, R., Chitsaz, H., Chou, W.-C., Corbeil, J., Del Fabbro, C., Docking, T., Durbin, R., Earl, D., Emrich, S., Fedotov, P., Fonseca, N.A., Ganapathy, G., Gibbs, R.A., Gnerre, S., Godzaridis, E., Goldstein, S., Haimel, M., Hall, G., Haussler, D., Hiatt, J.B., Ho, I.Y., Howard, J., Hunt, M., Jackman, S.D., Jaffe, D.B., Jarvis, E.D., Jiang, H., Kazakov, S., Kersey, P.J., Kitzman, J.O., Knight, J.R., Koren, S., Lam, T.-W., Lavenier, D., Laviolette, F., Li, Y., Li, Z., Liu, B., Liu, Y., Luo, R., MacCallum, I., MacManes, M.D., Maillet, N., Melnikov, S., Naquin, D., Ning, Z., Otto, T.D., Paten, B., Paulo, O.S., Phillippy, A.M., Pina-Martins, F., Place, M., Przybylski, D., Qin, X., Qu, C., Ribeiro, F.J., Richards, S., Rokhsar, D.S., Ruby, J., Scalabrin, S., Schatz, M.C., Schwartz, D.C., Sergushichev, A., Sharpe, T., Shaw, T.I., Shendure, J., Shi, Y., Simpson, J.T., Song, H., Tsarev, F., Vezzi, F., Vicedomini, R., Vieira, B.M., Wang, J., Worley, K.C., Yin, S., Yiu, S.-M., Yuan, J., Zhang, G., Zhang, H., Zhou, S. and Korf, I.F. (2013) Assemblathon 2: evaluating *De novo* methods of genome assembly in three vertebrate species. *GigaScience*, **2**, 10.

Budiman, M.A., Mao, L., Wood, T.C. and Wing, R.A. (2000) A Deep-Coverage Tomato BAC Library and Prospects Toward Development of an STC Framework for Genome Sequencing. *Genome Research*, **10**, 129-136.

Burton, J.N., Adey, A., Patwardhan, R.P., Qiu, R., Kitzman, J.O. and Shendure, J. (2013) Chromosome-scale scaffolding of *De novo* genome assemblies based on chromatin interactions. *Nat Biotechnol*, **31**, 1119-1125.

Bush, W.S. and Moore, J.H. (2012) Chapter 11: Genome-Wide Association Studies. *PLoS Comput Biol*, **8**, e1002822.

Byrd, A.L., Perez-Rogers, J.F., Manimaran, S., Castro-Nallar, E., Toma, I., McCaffrey, T., Siegel, M., Benson, G., Crandall, K.A. and Johnson, W.E. (2014) Clinical PathoScope: rapid alignment and filtration for accurate pathogen identification in clinical samples using unassembled sequencing data. *BMC bioinformatics*, **15**, 262.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. and Madden, T.L. (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421.

Canady, M.A., Ji, Y. and Chetelat, R.T. (2006) Homeologous recombination in *Solanum lycopersicoides* introgression lines of cultivated tomato. *Genetics*, **174**, 1775-1788.

Cannon, C.H., Kua, C.S., Zhang, D. and Harting, J.R. (2010) Assembly free comparative genomics of short-read sequence data discovers the needles in the haystack. *Molecular ecology*, **19 Suppl 1**, 147-161.

Cao, J., Schneeberger, K., Ossowski, S., Gunther, T., Bender, S., Fitz, J., Koenig, D., Lanz, C., Stegle, O., Lippert, C., Wang, X., Ott, F., Muller, J., Alonso-Blanco, C., Borgwardt, K., Schmid, K.J. and Weigel, D. (2011) Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. *Nature Genetics*, **43**, 956-963.

Causse, M. (2004) A genetic map of candidate genes and QTLs involved in tomato fruit size and composition. *Journal of Experimental Botany*, **55**, 1671-1685.

Causse, M., Saliba-Colombani, V., Lesschaeve, I. and Buret, M. (2001) Genetic analysis of organoleptic quality in fresh market tomato. 2. Mapping QTLs for sensory attributes. *TAG Theoretical and Applied Genetics*, **102**, 273-283.

Celniker, S.E., Wheeler, D.A., Kronmiller, B., Carlson, J.W., Halpern, A., Patel, S., Adams, M., Champe, M., Dugan, S.P., Frise, E., Hodgson, A., George, R.A., Hoskins, R.A., Laverty, T., Muzny, D.M., Nelson, C.R., Pacleb, J.M., Park, S., Pfeiffer, B.D., Richards, S., Sodergren, E.J., Svirskas, R., Tabor, P.E., Wan, K., Stapleton, M., Sutton, G.G., Venter, C., Weinstock, G., Scherer, S.E., Myers, E.W., Gibbs, R.A. and Rubin, G.M. (2002) Finishing a whole-genome shotgun: release 3 of the *Drosophila melanogaster* euchromatic genome sequence. *Genome biology*, **3**, Research0079.

Chan, C.K., Hsu, A.L., Halgamuge, S.K. and Tang, S.L. (2008a) Binning sequences using very sparse labels within a metagenome. *BMC bioinformatics*, **9**, 215.

Chan, C.K., Hsu, A.L., Tang, S.L. and Halgamuge, S.K. (2008b) Using growing self-organising maps to improve the binning process in environmental whole-genome shotgun sequencing. *Journal of biomedicine & biotechnology*, **2008**, 513701.

Chang, S.-B. and de Jong, H. (2005) Production of alien chromosome additions and their utility in plant genetics. *Cytogenet. Genome Res*. **109**, 335-343.

Chang, S.B., Anderson, L.K., Sherman, J.D., Royer, S.M. and Stack, S.M. (2007) Predicting and Testing Physical Locations of Genetically Mapped Loci on Tomato Pachytene Chromosome 1. *Genetics*, **176**, 2131-2138.

Chang, S.-B., Yang, T.-J., Datema, E., van Vugt, J., Vosman, B., Kuipers, A., Meznikova, M., Szinay, D., Lankhorst, R.K., Jacobsen, E. and de Jong, H. (2008) FISH mapping and molecular organization of the major repetitive sequences of tomato. *Chromosome Research*, **16**, 919-933.

Chen, J., Huang, Q., Gao, D., Wang, J., Lang, Y., Liu, T., Li, B., Bai, Z., Luis Goicoechea, J., Liang, C., Chen, C., Zhang, W., Sun, S., Liao, Y., Zhang, X., Yang, L., Song, C., Wang, M., Shi, J., Liu, G., Liu, J., Zhou, H., Zhou, W., Yu, Q., An, N., Chen, Y., Cai, Q., Wang, B., Liu, B., Min, J., Huang, Y., Wu, H., Li, Z., Zhang, Y., Yin, Y., Song, W., Jiang, J., Jackson, S.A., Wing, R.A., Wang, J. and Chen, M. (2013) Whole-genome sequencing of Oryza brachyantha reveals mechanisms underlying Oryza genome evolution. *Nature communications*, **4**, 1595.

Chen, Z., Wang, B., Dong, X., Liu, H., Ren, L., Chen, J., Hauck, A., Song, W. and Lai, J. (2014) An ultra-high density bin-map for rapid QTL mapping for tassel and ear architecture in a large F(2) maize population. *BMC Genomics*, **15**, 433.

Chibana, H., Oka, N., Nakayama, H., Aoyama, T., Magee, B.B., Magee, P.T. and Mikami, Y. (2005) Sequence finishing and gene mapping for Candida albicans chromosome 7 and syntenic analysis against the *Saccharomyces cerevisiae* genome. *Genetics*, **170**, 1525-1537.

Church, D.M., Goodstadt, L., Hillier, L.W., Zody, M.C., Goldstein, S., She, X., Bult, C.J., Agarwala, R., Cherry, J.L., DiCuccio, M., Hlavina, W., Kapustin, Y., Meric, P., Maglott, D., Birtle, Z., Marques, A.C., Graves, T., Zhou, S., Teague, B., Potamousis, K., Churas, C., Place, M., Herschleb, J., Runnheim, R., Forrest, D., Amos-Landgraf, J., Schwartz, D.C., Cheng, Z., Lindblad-Toh, K., Eichler, E.E. and Ponting, C.P. (2009) Lineage-specific biology revealed by a finished genome assembly of the mouse. *PLoS biology*, **7**, e1000112.

Cingolani, P., Platts, A., Wang le, L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X. and Ruden, D.M. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. Fly (Austin). 2012 Apr-Jun;6(2):80-92. doi: 10.4161/fly.19695.

Clark, A.G., Eisen, M.B., Smith, D.R., Bergman, C.M., Oliver, B., Markow, T.A., Kaufman, T.C., Kellis, M., Gelbart, W., Iyer, V.N., Pollard, D.A., Sackton, T.B., Larracuente, A.M., Singh, N.D., Abad, J.P., Abt, D.N., Adryan, B., Aguade, M., Akashi, H., Anderson, W.W., Aquadro, C.F., Ardell, D.H., Arguello, R., Artieri, C.G., Barbash, D.A., Barker, D., Barsanti, P., Batterham, P., Batzoglou, S., Begun, D., Bhutkar, A., Blanco, E., Bosak, S.A., Bradley, R.K., Brand, A.D., Brent, M.R., Brooks, A.N., Brown, R.H., Butlin, R.K., Caggese, C., Calvi, B.R., Bernardo de Carvalho, A., Caspi, A., Castrezana, S., Celniker, S.E., Chang, J.L., Chapple, C., Chatterji, S., Chinwalla, A., Civetta, A., Clifton, S.W., Comeron, J.M., Costello, J.C., Coyne, J.A., Daub, J., David, R.G., Delcher, A.L., Delehaunty, K., Do, C.B., Ebling, H., Edwards, K., Eickbush, T., Evans, J.D., Filipski, A., Findeiss, S., Freyhult, E., Fulton, L., Fulton, R., Garcia, A.C., Gardiner, A., Garfield, D.A., Garvin, B.E., Gibson, G., Gilbert, D., Gnerre, S., Godfrey, J., Good, R., Gotea, V., Gravely, B., Greenberg, A.J., Griffiths-Jones, S., Gross, S., Guigo, R., Gustafson, E.A., Haerty, W., Hahn, M.W., Halligan, D.L., Halpern, A.L., Halter, G.M., Han, M.V., Heger, A., Hillier, L., Hinrichs, A.S., Holmes, I., Hoskins, R.A., Hubisz, M.J., Hultmark, D., Huntley, M.A., Jaffe, D.B., Jagadeeshan, S., Jeck, W.R., Johnson, J., Jones, C.D., Jordan, W.C., Karpen, G.H., Kataoka, E., Keightley, P.D., Kheradpour, P., Kirkness, E.F., Koerich, L.B., Kristiansen, K., Kudrna, D., Kulathinal, R.J., Kumar, S., Kwok, R., Lander, E., Langley, C.H., Lapoint, R., Lazzaro, B.P., Lee, S.J., Levesque, L., Li, R., Lin, C.F., Lin, M.F., Lindblad-Toh, K., Llopart, A., Long, M., Low, L., Lozovsky, E., Lu, J., Luo, M., Machado, C.A., Makalowski, W., Marzo, M., Matsuda, M., Matzkin, L., McAllister, B., McBride, C.S., McKernan, B., McKernan, K., Mendez-Lago, M., Minx, P., Mollenhauer, M.U., Montooth, K., Mount, S.M., Mu, X., Myers, E., Negre, B., Newfeld, S., Nielsen, R., Noor, M.A., O'Grady, P., Pachter, L., Papaceit, M., Parisi, M.J., Parisi, M., Parts, L., Pedersen, J.S., Pesole, G., Phillippy, A.M., Ponting, C.P., Pop, M., Porcelli, D., Powell, J.R., Prohaska, S., Pruitt, K., Puig, M., Quesneville, H., Ram, K.R., Rand, D., Rasmussen, M.D., Reed, L.K., Reenan, R., Reily, A., Remington, K.A., Rieger, T.T., Ritchie, M.G., Robin, C., Rogers, Y.H., Rohde, C., Rozas, J., Rubenfield, M.J., Ruiz, A., Russo, S., Salzberg, S.L., Sanchez-Gracia, A., Saranga, D.J., Sato, H., Schaeffer, S.W., Schatz, M.C., Schlenke, T., Schwartz, R., Segarra, C., Singh, R.S., Sirot, L., Sirota, M., Sisneros, N.B., Smith, C.D., Smith, T.F., Spieth, J., Stage, D.E., Stark, A., Stephan, W., Strausberg, R.L., Strempel, S., Sturgill, D., Sutton, G., Sutton, G.G., Tao, W., Teichmann, S., Tobari, Y.N., Tomimura, Y., Tsolas, J.M., Valente, V.L., Venter, E., Venter, J.C., Vicario, S., Vieira, F.G., Vilella, A.J., Villasante, A., Walenz, B., Wang, J., Wasserman, M., Watts, T., Wilson, D., Wilson, R.K., Wing, R.A., Wolfner, M.F., Wong, A., Wong, G.K., Wu, C.I., Wu, G., Yamamoto, D., Yang, H.P., Yang, S.P., Yorke, J.A., Yoshida, K., Zdobnov, E., Zhang, P., Zhang, Y., Zimin, A.V., Baldwin, J., Abdouelleil, A., Abdulkadir, J., Abebe, A., Abera, B., Abreu, J., Acer, S.C., Aftuck, L., Alexander, A., An, P., Anderson, E., Anderson, S., Arachi, H., Azer, M., Bachantsang, P., Barry, A., Bayul, T., Berlin, A., Bessette, D., Bloom, T., Blye, J., Boguslavskiy, L., Bonnet, C., Boukhgalter, B., Bourzgui, I., Brown, A., Cahill, P., Channer, S., Cheshatsang, Y., Chuda, L., Citroen, M., Collymore, A., Cooke, P., Costello, M., D'Aco, K., Daza, R., De Haan, G., DeGray, S., DeMaso, C., Dhargay, N., Dooley, K., Dooley, E., Doricent, M., Dorje, P., Dorjee, K., Dupes, A., Elong, R., Falk, J., Farina, A., Faro, S., Ferguson, D., Fisher, S., Foley, C.D., Franke, A., Friedrich, D., Gadbois, L., Gearin, G., Gearin, C.R., Giannoukos, G., Goode, T., Graham, J., Grandbois, E., Grewal, S., Gyaltsen, K., Hafez, N., Hagos, B., Hall, J., Henson, C., Hollinger, A., Honan, T., Huard, M.D., Hughes, L., Hurhula, B., Husby, M.E., Kamat, A., Kanga, B., Kashin, S., Khazanovich, D., Kisner, P., Lance, K., Lara, M., Lee, W., Lennon, N., Letendre, F., LeVine, R., Lipovsky, A., Liu, X., Liu, J., Liu, S., Lokyitsang, T., Lokyitsang, Y., Lubonja, R., Lui, A., MacDonald, P., Magnisalis, V., Maru, K., Matthews, C., McCusker, W., McDonough, S., Mehta, T., Meldrim, J., Meneus, L., Mihai, O., Mihalev, A., Mihova, T., Mittelman, R., Mlenga, V., Montmayeur, A., Mulrain, L., Navidi, A., Naylor, J., Negash, T., Nguyen, T., Nguyen, N., Nicol, R., Norbu, C., Norbu, N., Novod, N., O'Neill, B., Osman, S., Markiewicz, E., Oyono, O.L., Patti, C., Phunkhang, P., Pierre, F., Priest, M., Raghuraman, S., Rege, F., Reyes, R., Rise, C., Rogov, P., Ross, K., Ryan, E., Settipalli, S., Shea, T., Sherpa, N., Shi, L., Shih, D., Sparrow, T., Spaulding, J., Stalker, J., Stange-Thomann, N., Stavropoulos, S., Stone, C., Strader, C., Tesfaye, S., Thomson, T., Thoulutsang, Y., Thoulutsang, D., Topham, K., Topping, I., Tsamla, T., Vassiliev, H., Vo, A., Wangchuk, T., Wangdi, T., Weiand, M., Wilkinson, J., Wilson, A., Yadav, S., Young, G., Yu, Q., Zembek, L., Zhong, D., Zimmer, A., Zwirko, Z., Jaffe, D.B., Alvarez, P., Brockman, W., Butler, J., Chin, C., Gnerre, S., Grabherr, M., Kleber, M., Mauceli, E. and MacCallum, I. (2007) Evolution of genes and genomes on the Drosophila phylogeny. *Nature*, **450**, 203-218.

Cochrane, G., Cook, C.E. and Birney, E. (2012) The future of DNA sequence archiving. *GigaScience*, **1**, 2.

Collard, B.C.Y. and Mackill, D.J. (2008) Marker-assisted selection: an approach for precision plant breeding in the twenty-first century. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **363**, 557-572.

Compeau, P.E.C., Pevzner, P.A. and Tesler, G. (2011) How to apply de Bruijn graphs to genome assembly. *Nat Biotechnol*, **29**, 987-991.

Cong, B., Barrero, L.S. and Tanksley, S.D. (2008) Regulatory change in YABBY-like transcription factor led to evolution of extreme fruit size during tomato domestication. *Nature Gen*. 40, 800-804.

Consortium*, I.C.G.S. (2004) Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature*, **432**, 695-716.

Consortium, T.C.S.a.A. (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*, **437**, 69-87.

Cviková, K., Cattonaro, F., Alaux, M., Stein, N., Mayer, K.F.X., Doležel, J. and Bartoš, J. (2015) High-throughput physical map anchoring via BAC-pool sequencing. *BMC Plant Biol*, **15**.

Dalloul, R.A., Long, J.A., Zimin, A.V., Aslam, L., Beal, K., Blomberg Le, A., Bouffard, P., Burt, D.W., Crasta, O., Crooijmans, R.P., Cooper, K., Coulombe, R.A., De, S., Delany, M.E., Dodgson, J.B., Dong, J.J., Evans, C., Frederickson, K.M., Flicek, P., Florea, L., Folkerts, O., Groenen, M.A., Harkins, T.T., Herrero, J., Hoffmann, S., Megens, H.J., Jiang, A., de Jong, P., Kaiser, P., Kim, H., Kim, K.W., Kim, S., Langenberger, D., Lee, M.K., Lee, T., Mane, S., Marcais, G., Marz, M., McElroy, A.P., Modise, T., Nefedov, M., Notredame, C., Paton, I.R., Payne, W.S., Pertea, G., Prickett, D., Puiu,

D., Qioa, D., Raineri, E., Ruffier, M., Salzberg, S.L., Schatz, M.C., Scheuring, C., Schmidt, C.J., Schroeder, S., Searle, S.M., Smith, E.J., Smith, J., Sonstegard, T.S., Stadler, P.F., Tafer, H., Tu, Z.J., Van Tassell, C.P., Vilella, A.J., Williams, K.P., Yorke, J.A., Zhang, L., Zhang, H.B., Zhang, X., Zhang, Y. and Reed, K.M. (2010) Multi-platform next-generation sequencing of the domestic turkey (Meleagris gallopavo): genome assembly and analysis. *PLoS biology*, **8**.

Dekkers, J.C., and Hospital, F. (2002) The use of molecular genetics in the improvement of agricultural populations. *Nature Rev. Gen*. **3**, 22-32.

Deschamps, S., Llaca, V. and May, G.D. (2012) Genotyping-by-Sequencing in Plants. *Biology*, **1**, 460-483.

Diaz, N.N., Krause, L., Goesmann, A., Niehaus, K. and Nattkemper, T.W. (2009) TACOA: taxonomic classification of environmental genomic fragments using a kernelized nearest neighbor approach. *BMC bioinformatics*, **10**, 56.

Diversity, C.o.B. (2010) Nagoya Protocol on Access to Genetic Resources and the Fair and Equitable Sharing of Benefits Arising from their Utilization to the Convention on Biological Diversity.

Doebley, J.F., Gaut, B.S. and Smith, B.D. (2006) The Molecular Genetics of Crop Domestication. *Cell*, **127**, 1309-1321.

Dong, F., McGrath, J.M., Helgeson, J.P. and Jiang, J. (2001) The genetic identity of alien chromosomes in potato breeding lines revealed by sequential GISH and FISH analyses using chromosome-specific cytogenetic DNA markers. *Genome*, **44,** 729-734.

Donmez, N. and Brudno, M. SCARPA: scaffolding reads with practical algorithms. (2013) *Bioinformatics*, 29, 428-434.

Drouaud, J., Camilleri, C., Bourguignon, P.Y., Canaguier, A., Berard, A., Vezon, D., Giancola, S., Brunel, D., Colot, V., Prum, B., Quesneville, H. and Mezard, C. (2006) Variation in crossing-over rates across chromosome 4 of *Arabidopsis thaliana* reveals the presence of meiotic recombination "hot spots". *Genome Res*. **16**, 106-114.

D'Souza, C.A., Kronstad, J.W., Taylor, G., Warren, R., Yuen, M., Hu, G., Jung, W.H., Sham, A., Kidd, S.E., Tangen, K., Lee, N., Zeilmaker, T., Sawkins, J., McVicker, G., Shah, S., Gnerre, S., Griggs, A., Zeng, Q., Bartlett, K., Li, W., Wang, X., Heitman, J., Stajich, J.E., Fraser, J.A., Meyer, W., Carter, D., Schein, J., Krzywinski, M., Kwon-Chung, K.J., Varma, A., Wang, J., Brunham, R., Fyfe, M., Ouellette, B.F., Siddiqui, A., Marra, M., Jones, S., Holt, R., Birren, B.W., Galagan, J.E. and Cuomo, C.A. (2011) Genome variation in Cryptococcus gattii, an emerging pathogen of immunocompetent hosts. *mBio*, **2**, e00342-00310.

Dujon, B., Sherman, D., Fischer, G., Durrens, P., Casaregola, S., Lafontaine, I., De Montigny, J., Marck, C., Neuveglise, C., Talla, E., Goffard, N., Frangeul, L., Aigle, M., Anthouard, V., Babour, A., Barbe, V., Barnay, S., Blanchin, S., Beckerich, J.M., Beyne, E., Bleykasten, C., Boisrame, A., Boyer, J., Cattolico, L., Confanioleri, F., De Daruvar, A., Despons, L., Fabre, E., Fairhead, C., Ferry-Dumazet, H., Groppi, A., Hantraye, F., Hennequin, C., Jauniaux, N., Joyet, P., Kachouri, R., Kerrest, A., Koszul, R., Lemaire, M., Lesur, I., Ma, L., Muller, H., Nicaud, J.M., Nikolski, M., Oztas, S., Ozier-Kalogeropoulos, O., Pellenz, S., Potier, S., Richard, G.F., Straub, M.L., Suleau, A., Swennen, D., Tekaia, F., Wesolowski-Louvel, M., Westhof, E., Wirth, B., Zeniou-Meyer, M., Zivanovic, I., Bolotin-Fukuhara, M., Thierry, A., Bouchier, C., Caudron, B., Scarpelli, C., Gaillardin, C., Weissenbach, J., Wincker, P. and Souciet, J.L. (2004) Genome evolution in yeasts. *Nature*, **430**, 35-44.

Ebeling, M., Kung, E., See, A., Broger, C., Steiner, G., Berrera, M., Heckel, T., Iniguez, L., Albert, T., Schmucki, R., Biller, H., Singer, T. and Certa, U. (2011) Genome-based analysis of the nonhuman primate Macaca fascicularis as a model for drug safety assessment. *Genome research*, **21**, 1746-1756.

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., Bibillo, A., Bjornson, K., Chaudhuri, B., Christians, F., Cicero, R., Clark, S., Dalal, R., deWinter, A., Dixon, J., Foquet, M., Gaertner, A., Hardenbol, P., Heiner, C., Hester, K., Holden, D., Kearns, G., Kong, X., Kuse, R., Lacroix, Y., Lin, S., Lundquist, P., Ma, C., Marks, P., Maxham, M., Murphy, D., Park, I., Pham, T., Phillips, M., Roy, J., Sebra, R., Shen, G., Sorenson, J., Tomaney, A., Travers, K., Trulson, M., Vieceli, J., Wegener, J., Wu, D., Yang, A., Zaccarin, D., Zhao, P., Zhong, F., Korlach, J. and Turner, S. (2009) Real-Time DNA Sequencing from Single Polymerase Molecules. *Science*, **323**, 133-138.

Eisenstein, M. (2012) Oxford Nanopore announcement sets sequencing sector abuzz. *Nat Biotechnol*, **30**, 295-296.

Elshire, R.J., Glaubitz, J.C., Sun, Q., Poland, J.A., Kawamoto, K., Buckler, E.S. and Mitchell, S.E. (2011) A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. *PLoS ONE*, **6**, e19379.

Eshed, Y. and Zamir, D. (1995) An introgression line population of Lycopersicon pennellii in the cultivated tomato enables the identification and fine mapping of yield-associated QTL. *Genetics*, **141**, 1147-1162.

Faino, L., Seidl, M.F., Datema, E., van den Berg, G.C.M., Janssen, A., Wittenberg, A.H.J. and Thomma, B.P.H.J. (2015) Single-Molecule Real-Time Sequencing Combined with Optical Mapping Yields Completely Finished Fungal Genome. *mBio*, **6**, e00936-00915.

Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R., Bult, C.J., Tomb, J.F., Dougherty, B.A., Merrick, J.M. and *et al.*, (1995) Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science*, **269**, 496-512.

Fransz, P., Soppe, W., Soppe, W. and Schubert, I. (2003) *Chromosome Research*, **11**, 227-240.

Fransz, P.F., Armstrong, S., de Jong, J.H., Parnell, L.D., van Drunen, C., Dean, C., Zabel, P., Bisseling, T. and Jones, G.H. (2000) Integrated Cytogenetic Map of Chromosome Arm 4S of A. thaliana: Structural Organization of Heterochromatic Knob and Centromere Region. *Cell*, **100**, 367-376.

Fray, R.G. and Grierson, D. (1993) Identification and genetic analysis of normal and mutant phytoene synthase genes of tomato by sequencing, complementation and co-suppression. *Plant Mol. Biol*. 22, 589-602.

Fridman, E. (2004) Zooming In on a Quantitative Trait for Tomato Yield Using Interspecific Introgressions. *Science*, **305**, 1786-1789.

Galagan, J.E., Calvo, S.E., Cuomo, C., Ma, L.J., Wortman, J.R., Batzoglou, S., Lee, S.I., Basturkmen, M., Spevak, C.C., Clutterbuck, J., Kapitonov, V., Jurka, J., Scazzocchio, C., Farman, M., Butler, J., Purcell, S., Harris, S., Braus, G.H., Draht, O., Busch, S., D'Enfert, C., Bouchier, C., Goldman, G.H., Bell-Pedersen, D., Griffiths-Jones, S., Doonan, J.H., Yu, J., Vienken, K., Pain, A., Freitag, M., Selker, E.U., Archer, D.B., Penalva, M.A., Oakley, B.R., Momany, M.,

Tanaka, T., Kumagai, T., Asai, K., Machida, M., Nierman, W.C., Denning, D.W., Caddick, M., Hynes, M., Paoletti, M., Fischer, R., Miller, B., Dyer, P., Sachs, M.S., Osmani, S.A. and Birren, B.W. (2005) Sequencing of Aspergillus nidulans and comparative analysis with A. fumigatus and A. oryzae. *Nature*, **438**, 1105-1115.

Galvao, V.C., Nordstrom, K.J., Lanz, C., Sulz, P., Mathieu, J., Pose, D., Schmid, M., Weigel, D. and Schneeberger, K. (2012) Synteny-based mapping-by-sequencing enabled by targeted enrichment. *Plant J.* **71**, 517-526.

Gardner, S.N. and Hall, B.G. (2013) When whole-genome alignments just won't work: kSNP v2 software for alignment-free SNP discovery and phylogenetics of hundreds of microbial genomes. *PloS one*, **8**, e81760.

Garrett, K.A., Dendy, S.P., Frank, E.E., Rouse, M.N. and Travers, S.E. (2006) Climate Change Effects on Plant Disease: Genomes to Ecosystems. *Annu. Rev. Phytopathol.*, **44**, 489-509.

Gehlenborg, N. and Wong, B. (2012) Points of view: Into the third dimension. *Nature Methods*, **9**, 851-851.

Gibbs, R.A., Rogers, J., Katze, M.G., Bumgarner, R., Weinstock, G.M., Mardis, E.R., Remington, K.A., Strausberg, R.L., Venter, J.C., Wilson, R.K., Batzer, M.A., Bustamante, C.D., Eichler, E.E., Hahn, M.W., Hardison, R.C., Makova, K.D., Miller, W., Milosavljevic, A., Palermo, R.E., Siepel, A., Sikela, J.M., Attaway, T., Bell, S., Bernard, K.E., Buhay, C.J., Chandrabose, M.N., Dao, M., Davis, C., Delehaunty, K.D., Ding, Y., Dinh, H.H., Dugan-Rocha, S., Fulton, L.A., Gabisi, R.A., Garner, T.T., Godfrey, J., Hawes, A.C., Hernandez, J., Hines, S., Holder, M., Hume, J., Jhangiani, S.N., Joshi, V., Khan, Z.M., Kirkness, E.F., Cree, A., Fowler, R.G., Lee, S., Lewis, L.R., Li, Z., Liu, Y.S., Moore, S.M., Muzny, D., Nazareth, L.V., Ngo, D.N., Okwuonu, G.O., Pai, G., Parker, D., Paul, H.A., Pfannkoch, C., Pohl, C.S., Rogers, Y.H., Ruiz, S.J., Sabo, A., Santibanez, J., Schneider, B.W., Smith, S.M., Sodergren, E., Svatek, A.F., Utterback, T.R., Vattathil, S., Warren, W., White, C.S., Chinwalla, A.T., Feng, Y., Halpern, A.L., Hillier, L.W., Huang, X., Minx, P., Nelson, J.O., Pepin, K.H., Qin, X., Sutton, G.G., Venter, E., Walenz, B.P., Wallis, J.W., Worley, K.C., Yang, S.P., Jones, S.M., Marra, M.A., Rocchi, M., Schein, J.E., Baertsch, R., Clarke, L., Csuros, M., Glasscock, J., Harris, R.A., Havlak, P., Jackson, A.R., Jiang, H., Liu, Y., Messina, D.N., Shen, Y., Song, H.X., Wylie, T., Zhang, L., Birney, E., Han, K., Konkel, M.K., Lee, J., Smit, A.F., Ullmer, B., Wang, H., Xing, J., Burhans, R., Cheng, Z., Karro, J.E., Ma, J., Raney, B., She, X., Cox, M.J., Demuth, J.P., Dumas, L.J., Han, S.G., Hopkins, J., Karimpour-Fard, A., Kim, Y.H., Pollack, J.R., Vinar, T., Addo-Quaye, C., Degenhardt, J., Denby, A., Hubisz, M.J., Indap, A., Kosiol, C., Lahn, B.T., Lawson, H.A., Marklein, A., Nielsen, R., Vallender, E.J., Clark, A.G., Ferguson, B., Hernandez, R.D., Hirani, K., Kehrer-Sawatzki, H., Kolb, J., Patil, S., Pu, L.L., Ren, Y., Smith, D.G., Wheeler, D.A., Schenck, I., Ball, E.V., Chen, R., Cooper, D.N., Giardine, B., Hsu, F., Kent, W.J., Lesk, A., Nelson, D.L., O'Brien W, E., Prufer, K., Stenson, P.D., Wallace, J.C., Ke, H., Liu, X.M., Wang, P., Xiang, A.P., Yang, F., Barber, G.P., Haussler, D., Karolchik, D., Kern, A.D., Kuhn, R.M., Smith, K.E. and Zwieg, A.S. (2007) Evolutionary and biomedical insights from the rhesus macaque genome. *Science (New York, N.Y.)*, **316**, 222-234.

Gibbs, R.A., Weinstock, G.M., Metzker, M.L., Muzny, D.M., Sodergren, E.J., Scherer, S., Scott, G., Steffen, D., Worley, K.C., Burch, P.E., Okwuonu, G., Hines, S., Lewis, L., DeRamo, C., Delgado, O., Dugan-Rocha, S., Miner, G., Morgan, M., Hawes, A., Gill, R., Celera, Holt, R.A., Adams, M.D., Amanatides, P.G., Baden-Tillson, H., Barnstead, M., Chin, S., Evans, C.A., Ferriera, S., Fosler, C., Glodek, A., Gu, Z., Jennings, D., Kraft, C.L., Nguyen, T., Pfannkoch, C.M., Sitter, C., Sutton, G.G., Venter, J.C., Woodage, T., Smith, D., Lee, H.M., Gustafson, E., Cahill, P., Kana, A., Doucette-Stamm, L., Weinstock, K., Fechtel, K., Weiss, R.B., Dunn, D.M., Green, E.D., Blakesley, R.W., Bouffard, G.G., De Jong, P.J., Osoegawa, K., Zhu, B., Marra, M., Schein, J., Bosdet, I., Fjell, C., Jones, S., Krzywinski, M., Mathewson, C., Siddiqui, A., Wye, N., McPherson, J., Zhao, S., Fraser, C.M., Shetty, J., Shatsman, S., Geer, K., Chen, Y., Abramzon, S., Nierman, W.C., Havlak, P.H., Chen, R., Durbin, K.J., Egan, A., Ren, Y., Song, X.Z., Li, B., Liu, Y., Qin, X., Cawley, S., Worley, K.C., Cooney, A.J., D'Souza, L.M., Martin, K., Wu, J.Q., Gonzalez-Garay, M.L., Jackson, A.R., Kalafus, K.J., McLeod, M.P., Milosavljevic, A., Virk, D., Volkov, A., Wheeler, D.A., Zhang, Z., Bailey, J.A., Eichler, E.E., Tuzun, E., Birney, E., Mongin, E., Ureta-Vidal, A., Woodwark, C., Zdobnov, E., Bork, P., Suyama, M., Torrents, D., Alexandersson, M., Trask, B.J., Young, J.M., Huang, H., Wang, H., Xing, H., Daniels, S., Gietzen, D., Schmidt, J., Stevens, K., Vitt, U., Wingrove, J., Camara, F., Mar Alba, M., Abril, J.F., Guigo, R., Smit, A., Dubchak, I., Rubin, E.M., Couronne, O., Poliakov, A., Hubner, N., Ganten, D., Goesele, C., Hummel, O., Kreitler, T., Lee, Y.A., Monti, J., Schulz, H., Zimdahl, H., Himmelbauer, H., Lehrach, H., Jacob, H.J., Bromberg, S., Gullings-Handley, J., Jensen-Seaman, M.I., Kwitek, A.E., Lazar, J., Pasko, D., Tonellato, P.J., Twigger, S., Ponting, C.P., Duarte, J.M., Rice, S., Goodstadt, L., Beatson, S.A., Emes, R.D., Winter, E.E., Webber, C., Brandt, P., Nyakatura, G., Adetobi, M., Chiaromonte, F., Elnitski, L., Eswara, P., Hardison, R.C., Hou, M., Kolbe, D., Makova, K., Miller, W., Nekrutenko, A., Riemer, C., Schwartz, S., Taylor, J., Yang, S., Zhang, Y., Lindpaintner, K., Andrews, T.D., Caccamo, M., Clamp, M., Clarke, L., Curwen, V., Durbin, R., Eyras, E., Searle, S.M., Cooper, G.M., Batzoglou, S., Brudno, M., Sidow, A., Stone, E.A., Venter, J.C., Payseur, B.A., Bourque, G., Lopez-Otin, C., Puente, X.S., Chakrabarti, K., Chatterji, S., Dewey, C., Pachter, L., Bray, N., Yap, V.B., Caspi, A., Tesler, G., Pevzner, P.A., Haussler, D., Roskin, K.M., Baertsch, R., Clawson, H., Furey, T.S., Hinrichs, A.S., Karolchik, D., Kent, W.J., Rosenbloom, K.R., Trumbower, H., Weirauch, M., Cooper, D.N., Stenson, P.D., Ma, B., Brent, M., Arumugam, M., Shteynberg, D., Copley, R.R., Taylor, M.S., Riethman, H., Mudunuri, U., Peterson, J., Guyer, M., Felsenfeld, A., Old, S., Mockrin, S. and Collins, F. (2004) Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature*, **428**, 493-521.

Giovannoni, J. (2001) M OLECULAR B IOLOGY OF F RUIT M ATURATION AND R IPENING. *Annual Review of Plant Physiology and Plant Molecular Biology*, **52**, 725-749.

Gnerre, S., MacCallum, I., Przybylski, D., Ribeiro, F.J., Burton, J.N., Walker, B.J., Sharpe, T., Hall, G., Shea, T.P., Sykes, S., Berlin, A.M., Aird, D., Costello, M., Daza, R., Williams, L., Nicola, R., Gnirke, A., Nusbaum, C., Lander, E.S. and Jaffea, D.B. (2010) High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc. Natl. Acad. Sci. USA.* doi:10.1073/pnas.1017351108.

Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J.D., Jacq, C., Johnston, M., Louis, E.J., Mewes, H.W., Murakami, Y., Philippsen, P., Tettelin, H. and Oliver, S.G. (1996) Life with 6000 Genes. *Science*, **274**, 546-567.

Grandillo, S., Chetelat, R., Knapp, S., Spooner, D., Peralta, I., Cammareri, M., Perez, O., Termolino, P., Tripodi, P, Chiusano, M.L., Ercolano, M.R., Frusciante, L., Monti, L. and Pignone, D. (2011) *Solanum* sect. Lycopersicum. In *Wild crop relatives: genomic and breeding resources* (Kole, C. ed). Berlin, Heidelberg, Germany: Springer, pp 129-215.

Greenblum, S., Carr, R. and Borenstein, E. (2015) Extensive strain-level copy-number variation across human gut microbiome species. *Cell*, **160**, 583-594.

Gross, B.L. and Olsen, K.M. (2010) Genetic perspectives on crop domestication. *Trends in Plant Science*, **15**, 529-537.

Hammer, K. (1984) Das Domestikationssyndrom. *Die Kulturpflanze*, **32**, 11-34.

Hammer, K. and Teklu, Y. (2008) Plant genetic resources: selected issues from genetic erosion to genetic engineering. *Journal of Agriculture and Rural Development in the Tropics and Subtropics (JARTS)*, **109**, 15-50.

Han, Y., Zhang, T., Thammapichai, P., Weng, Y. and Jiang, J. (2015) Chromosome-Specific Painting in Cucumis Species Using Bulked Oligonucleotides. *Genetics*, **200**, 771-779.

Hans de Jong, J., Fransz, P. and Zabel, P. (1999) High resolution FISH in plants – techniques and applications. *Trends in Plant Science*, **4**, 258-263.

Hoang, T., Yin, C., Zheng, H., Yu, C., Lucy He, R. and Yau, S.S. (2015) A new method to cluster DNA sequences using Fourier power spectrum. *Journal of theoretical biology*, **372**, 135-145.

Horwege, S., Lindner, S., Boden, M., Hatje, K., Kollmar, M., Leimeister, C.A. and Morgenstern, B. (2014) Spaced words and kmacs: fast alignment-free sequence comparison based on inexact word matches. *Nucleic acids research*, **42**, W7-11.

Hoskins, R.A., Carlson, J.W., Kennedy, C., Acevedo, D., Evans-Holm, M., Frise, E., Wan, K.H., Park, S., Mendez-Lago, M., Rossi, F., Villasante, A., Dimitri, P., Karpen, G.H. and Celniker, S.E. (2007) Sequence Finishing and Mapping of *Drosophila melanogaster* Heterochromatin. *Science*, **316**, 1625-1628.

Hsi-Yang Fritz, M., Leinonen, R., Cochrane, G. and Birney, E. (2011) Efficient storage of high throughput DNA sequencing data using reference-based compression. *Genome Research*, **21**, 734-740.

Hu, T.T., Pattyn, P., Bakker, E.G., Cao, J., Cheng, J.F., Clark, R.M., Fahlgren, N., Fawcett, J.A., Grimwood, J., Gundlach, H., Haberer, G., Hollister, J.D., Ossowski, S., Ottilar, R.P., Salamov, A.A., Schneeberger, K., Spannagl, M., Wang, X., Yang, L., Nasrallah, M.E., Bergelson, J., Carrington, J.C., Gaut, B.S., Schmutz, J., Mayer, K.F., Van de Peer, Y., Grigoriev, I.V., Nordborg, M., Weigel, D. and Guo, Y.L. (2011) The Arabidopsis lyrata genome sequence and the basis of rapid genome size change. *Nature genetics*, **43**, 476-481.

Huang, X., Feng, Q., Qian, Q., Zhao, Q., Wang, L., Wang, A., Guan, J., Fan, D., Weng, Q., Huang, T., Dong, G., Sang, T. and Han, B (2009) High-throughput genotyping by whole-genome resequencing. *Genome Res*. **19**, 1068-1076.

Huang, Y., Li, Y., Burt, D.W., Chen, H., Zhang, Y., Qian, W., Kim, H., Gan, S., Zhao, Y., Li, J., Yi, K., Feng, H., Zhu, P., Li, B., Liu, Q., Fairley, S., Magor, K.E., Du, Z., Hu, X., Goodman, L., Tafer, H., Vignal, A., Lee, T., Kim, K.W., Sheng, Z., An, Y., Searle, S., Herrero, J., Groenen, M.A., Crooijmans, R.P., Faraut, T., Cai, Q., Webster, R.G., Aldridge, J.R., Warren, W.C., Bartschat, S., Kehr, S., Marz, M., Stadler, P.F., Smith, J., Kraus, R.H., Zhao, Y., Ren, L., Fei, J., Morisson, M., Kaiser, P., Griffin, D.K., Rao, M., Pitel, F., Wang, J. and Li, N. (2013) The duck genome and transcriptome provide insight into an avian influenza virus reservoir species. *Nature genetics*, **45**, 776-783.

Hurwitz, B.L., Westveld, A.H., Brum, J.R. and Sullivan, M.B. (2014) Modeling ecological drivers in marine viral communities using comparative metagenomics and network analyses. *Proceedings of the National Academy of Sciences of the United States of America*, **111**, 10714-10719.

inc., I. (2015) Illumina inc.

Jackson, A.P., Gamble, J.A., Yeomans, T., Moran, G.P., Saunders, D., Harris, D., Aslett, M., Barrell, J.F., Butler, G., Citiulo, F., Coleman, D.C., de Groot, P.W., Goodwin, T.J., Quail, M.A., McQuillan, J., Munro, C.A., Pain, A., Poulter, R.T., Rajandream, M.A., Renauld, H., Spiering, M.J., Tivey, A., Gow, N.A., Barrell, B., Sullivan, D.J. and Berriman, M. (2009) Comparative genomics of the fungal pathogens Candida dubliniensis and Candida albicans. *Genome research*, **19**, 2231-2244.

Jaillon, O., Aury, J.M., Noel, B., Policriti, A., Clepet, C., Casagrande, A., Choisne, N., Aubourg, S., Vitulo, N., Jubin, C., Vezzi, A., Legeai, F., Hugueney, P., Dasilva, C., Horner, D., Mica, E., Jublot, D., Poulain, J., Bruyere, C., Billault, A., Segurens, B., Gouyvenoux, M., Ugarte, E., Cattonaro, F., Anthouard, V., Vico, V., Del Fabbro, C., Alaux, M., Di Gaspero, G., Dumas, V., Felice, N., Paillard, S., Juman, I., Moroldo, M., Scalabrin, S., Canaguier, A., Le Clainche, I., Malacrida, G., Durand, E., Pesole, G., Laucou, V., Chatelet, P., Merdinoglu, D., Delledonne, M., Pezzotti, M., Lecharny, A., Scarpelli, C., Artiguenave, F., Pe, M.E., Valle, G., Morgante, M., Caboche, M., Adam-Blondon, A.F., Weissenbach, J., Quetier, F. and Wincker, P. (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature*, **449**, 463-467.

Jump, A.S. and Penuelas, J. (2005) Running to stand still: adaptation and the response of plants to rapid climate change. *Ecology Letters*, **8**, 1010-1020.

Kahlau, S., Aspinall, S., Gray, J.C. and Bock, R. (2006) Sequence of the Tomato Chloroplast DNA and Evolutionary Comparison of Solanaceous Plastid Genomes. *J Mol Evol*, **63**, 194-207.

Keurentjes, J.J.B., Willems, G., van Eeuwijk, F., Nordborg, M. and Koornneef, M. (2011) A comparison of population types used for QTL mapping in *Arabidopsis thaliana*. *Plant Genetic Resources*, **9**, 185-188.

Khush, G.S. and Rick, C.M. (1963) Meiosis in hybrids betweenLycopersicon esculentum and*Solanum pennellii*. *Genetica*, **33**, 167-183.

Khush, G.S. and Rick, C.M. (1968) Cytogenetic analysis of the tomato genome by means of induced deficiencies. *Chromosoma*, **23**, 452-484.

Kim, Y.H., Park, H.M., Hwang, T.Y., Lee, S.K., Choi, M.S., Jho, S., Hwang, S., Kim, H.M., Lee, D., Kim, B.C., Hong, C.P., Cho, Y.S., Kim, H., Jeong, K.H., Seo, M.J., Yun, H.T., Kim, S.L., Kwon, Y.U., Kim, W.H., Chun, H.K., Lim, S.J., Shin, Y.A., Choi, I.Y., Kim, Y.S., Yoon, H.S., Lee, S.H. and Lee, S. (2014) Variation block-based genomics method for crop plants. *BMC Genomics*, **15**, 477.

Kimura, M. (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol*, **16**, 111-120.

Klein-Lankhorst, R.M., Vermunt, A., Weide, R., Liharska, T. and Zabel, P. (1991) Isolation of molecular markers for tomato (L. esculentum) using random amplified polymorphic DNA (RAPD). *Theoret. Appl. Genetics*, **83**.

Knapp, S. (2002) Tobacco to tomatoes: a phylogenetic perspective on fruit diversity in the *Solanaceae*. *J. Exp. Botany*, 53, 2001-2022.

Knight, R., Maxwell, P., Birmingham, A., Carnes, J., Caporaso, J.G., Easton, B.C., Eaton, M., Hamady, M., Lindsay, H., Liu, Z., Lozupone, C., McDonald, D., Robeson, M., Sammut, R., Smit, S., Wakefield, M.J., Widmann, J., Wikman, S., Wilson, S., Ying, H. and Huttley, G.A. (2007) PyCogent: a toolkit for making sense from sequence. *Genome biology*, **8**, R171.

Koenig D., Jiménez-Gómez J.M., Kimura S., Fulop D., Chitwood D.H., Headland L.R., Kumar R., Covington M.F., Devisetty U.K., Tat A.V., Tohge T., Bolger A., Schneeberger K., Ossowski S., Lanz C., Xiong G., Taylor-Teeples M., Brady S.M., Pauly M., Weigel D., Usadel B., Fernie A.R., Peng J., Sinha N.R., Maloof J.N.. (2013) Comparative transcriptomics reveals patterns of selection in domesticated and wild tomato. *Proc Natl Acad Sci* Jul 9;110(28):E2655-62

Koren, S. and Phillippy, A.M. (2015) One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly. *Current Opinion in Microbiology*, **23**, 110-120.

Korte, A. and Farlow, A. (2013) The advantages and limitations of trait analysis with GWAS: a review. *Plant Methods*, **9**, 29.

Ku, H. M., Doganlar, S., Chen, K.Y., and Tanksley, S.D. (1999) The genetic basis of per-shaped tomato fruit. *Theor. Appl. Genet*. 9, 844-850.

Kua, C.S., Ruan, J., Harting, J., Ye, C.X., Helmus, M.R., Yu, J. and Cannon, C.H. (2012) Reference-free comparative genomics of 174 chloroplasts. *PloS one*, **7**, e48995.

Kumar, S., Banks, T.W. and Cloutier, S. (2012) SNP Discovery through Next-Generation Sequencing and Its Applications. *International Journal of Plant Genomics*, **2012**, 831460.

Kurtz, S., Phillippy, A., Delcher, A.L., Smoot, M., Shumway, M., Antonescu, C. and Salzberg, S.L. (2004) Versatile and open software for comparing large genomes. *Genome Biol*. **5**, R12.

Lamb, J.C., Danilova, T., Bauer, M.J., Meyer, J.M., Holland, J.J., Jensen, M.D. and Birchler, J.A. (2006) Single-Gene Detection and Karyotyping Using Small-Target Fluorescence *in situ* Hybridization on Maize Somatic Chromosomes. *Genetics*, **175**, 1047-1058.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J.P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J.C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R.H., Wilson, R.K., Hillier, L.W., McPherson, J.D., Marra, M.A., Mardis, E.R., Fulton, L.A., Chinwalla, A.T., Pepin, K.H., Gish, W.R., Chissoe, S.L., Wendl, M.C., Delehaunty, K.D., Miner, T.L., Delehaunty, A., Kramer, J.B., Cook, L.L., Fulton, R.S., Johnson, D.L., Minx, P.J., Clifton, S.W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J.F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R.A., Muzny, D.M., Scherer, S.E., Bouck, J.B., Sodergren, E.J., Worley, K.C., Rives, C.M., Gorrell, J.H., Metzker, M.L., Naylor, S.L., Kucherlapati, R.S., Nelson, D.L., Weinstock, G.M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Smith, D.R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H.M., Dubois, J., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R.W., Federspiel, N.A., Abola, A.P., Proctor, M.J., Myers, R.M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D.R., Olson, M.V., Kaul, R., Raymond, C., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G.A., Athanasiou, M., Schultz, R., Roe, B.A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W.R., de la Bastide, M., Dedhia, N., Blocker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J.A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D.G., Burge, C.B., Cerutti, L., Chen, H.C., Church, D., Clamp, M., Copley, R.R., Doerks, T., Eddy, S.R., Eichler, E.E., Furey, T.S., Galagan, J., Gilbert, J.G., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L.S., Jones, T.A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W.J., Kitts, P., Koonin, E.V., Korf, I., Kulp, D., Lancet, D., Lowe, T.M., McLysaght, A., Mikkelsen, T., Moran, J.V., Mulder, N., Pollara, V.J., Ponting, C.P., Schuler, G., Schultz, J., Slater, G., Smit, A.F., Stupka, E., Szustakowski, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y.I., Wolfe, K.H., Yang, S.P., Yeh, R.F., Collins, F., Guyer, M.S., Peterson, J., Felsenfeld, A., Wetterstrand, K.A., Patrinos, A., Morgan, M.J., de Jong, P., Catanese, J.J., Osoegawa, K., Shizuya, H., Choi, S., Chen, Y.J. and International Human Genome Sequencing, C. (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860-921.

LaNGMead, B., Trapnell, C., Pop, M. and Salzberg, S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* 10, R25.

Lechat, P., Souche, E. and Moszer, I. (2013) SynTView - an interactive multi-view genome browser for next-generation comparative microorganism genomics. *BMC Bioinformatics*, **14**, 277.

Lee, T.H., Guo, H., Wang, X., Kim, C. and Paterson, A.H. (2014) SNPhylo: a pipeline to construct a phylogenetic tree from huge SNP data. *BMC Genomics*, **15**, 162.

Leggett, R.M., Ramirez-Gonzalez, R.H., Clavijo, B.J., Waite, D. and Davey, R.P. (2013) Sequencing quality assessment tools to enable data-driven informatics for high throughput genomics. *Frontiers in genetics*, **4**, 288.

Letunic, I. and Bork, P. (2007) Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics*, **23**, 127-128.

Levy-Sakin, M. and Ebenstein, Y. (2013) Beyond sequencing: optical mapping of DNA in the age of nanotechnology and nanoscopy. *Current Opinion in Biotechnology*, **24**, 690-698.

Li, H. and Durbin, R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754-1760.

Li, H. and Homer, N. (2010) A survey of sequence alignment algorithms for next-generation sequencing. *Briefings in Bioinformatics*, **11**, 473-483.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. and Durbin, R. (2009) 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMTOOLS. *Bioinformatics*, 25, 2078-2079.

Li, R., Hsieh, C.-L., Young, A., Zhang, Z., Ren, X. and Zhao, Z. (2015) Illumina Synthetic Long Read Sequencing Allows Recovery of Missing Sequences even in the "Finished" C. elegans Genome. *Sci. Rep.*, **5**, 10814.

Lin, T., Zhu, G., Zhang, J., Xu, X., Yu, Q., Zheng, Z., Zhang, Z., Lun, Y., Li, S., Wang, X., Huang, Z., Li, J., Zhang, C., Wang, T., Zhang, Y., Wang, A., Zhang, Y., Lin, K., Li, C., Xiong, G., Xue, Y., Mazzucato, A., Causse, M., Fei, Z., Giovannoni, J.J., Chetelat, R.T., Zamir, D., Städler, T., Li, J., Ye, Z., Du, Y. and Huang, S. (2014) Genomic analyses provide insights into the history of tomato breeding. *Nature Genetics*, **46**, 1220-1226.

Lindblad-Toh, K., Wade, C.M., Mikkelsen, T.S., Karlsson, E.K., Jaffe, D.B., Kamal, M., Clamp, M., Chang, J.L., Kulbokas, E.J., 3rd, Zody, M.C., Mauceli, E., Xie, X., Breen, M., Wayne, R.K., Ostrander, E.A., Ponting, C.P., Galibert, F., Smith, D.R., DeJong, P.J., Kirkness, E., Alvarez, P., Biagi, T., Brockman, W., Butler, J., Chin, C.W., Cook, A., Cuff, J., Daly, M.J., DeCaprio, D., Gnerre, S., Grabherr, M., Kellis, M., Kleber, M., Bardeleben, C., Goodstadt, L., Heger, A., Hitte, C., Kim, L., Koepfli, K.P., Parker, H.G., Pollinger, J.P., Searle, S.M., Sutter, N.B., Thomas, R., Webber, C., Baldwin, J., Abebe, A., Abouelleil, A., Aftuck, L., Ait-Zahra, M., Aldredge, T., Allen, N., An, P., Anderson, S., Antoine, C., Arachchi, H., Aslam, A., Ayotte, L., Bachantsang, P., Barry, A., Bayul, T., Benamara, M., Berlin, A., Bessette, D., Blitshteyn, B., Bloom, T., Blye, J., Boguslavskiy, L., Bonnet, C., Boukhgalter, B., Brown, A., Cahill, P., Calixte, N., Camarata, J., Cheshatsang, Y., Chu, J., Citroen, M., Collymore, A., Cooke, P., Dawoe, T., Daza, R., Decktor, K., DeGray, S., Dhargay, N., Dooley, K., Dooley, K., Dorje, P., Dorjee, K., Dorris, L., Duffey, N., Dupes, A., Egbiremolen, O., Elong, R., Falk, J., Farina, A., Faro, S., Ferguson, D., Ferreira, P., Fisher, S., FitzGerald, M., Foley, K., Foley, C., Franke, A., Friedrich, D., Gage, D., Garber, M., Gearin, G., Giannoukos, G., Goode, T., Goyette, A., Graham, J., Grandbois, E., Gyaltsen, K., Hafez, N., Hagopian, D., Hagos, B., Hall, J., Healy, C., Hegarty, R., Honan, T., Horn, A., Houde, N., Hughes, L., Hunnicutt, L., Husby, M., Jester, B., Jones, C., Kamat, A., Kanga, B., Kells, C., Khazanovich, D., Kieu, A.C., Kisner, P., Kumar, M., Lance, K., Landers, T., Lara, M., Lee, W., Leger, J.P., Lennon, N., Leuper, L., LeVine, S., Liu, J., Liu, X., Lokyitsang, Y., Lokyitsang, T., Lui, A., Macdonald, J., Major, J., Marabella, R., Maru, K., Matthews, C., McDonough, S., Mehta, T., Meldrim, J., Melnikov, A., Meneus, L., Mihalev, A., Mihova, T., Miller, K., Mittelman, R., Mlenga, V., Mulrain, L., Munson, G., Navidi, A., Naylor, J., Nguyen, T., Nguyen, N., Nguyen, C., Nguyen, T., Nicol, R., Norbu, N., Norbu, C., Novod, N., Nyima, T., Olandt, P., O'Neill, B., O'Neill, K., Osman, S., Oyono, L., Patti, C., Perrin, D., Phunkhang, P., Pierre, F., Priest, M., Rachupka, A., Raghuraman, S., Rameau, R., Ray, V., Raymond, C., Rege, F., Rise, C., Rogers, J., Rogov, P., Sahalie, J., Settipalli, S., Sharpe, T., Shea, T., Sheehan, M., Sherpa, N., Shi, J., Shih, D., Sloan, J., Smith, C., Sparrow, T., Stalker, J., Stange-Thomann, N., Stavropoulos, S., Stone, C., Stone, S., Sykes, S., Tchuinga, P., Tenzing, P., Tesfaye, S., Thoulutsang, D., Thoulutsang, Y., Topham, K., Topping, I., Tsamla, T., Vassiliev, H., Venkataraman, V., Vo, A., Wangchuk, T., Wangdi, T., Weiand, M., Wilkinson, J., Wilson, A., Yadav, S., Yang, S., Yang, X., Young, G., Yu, Q., Zainoun, J., Zembek, L., Zimmer, A. and Lander, E.S. (2005) Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature*, **438**, 803-819.

Lippman, Z.B., Semel, Y. and Zamir, D. (2007) An integrated view of quantitative trait variation using tomato interspecific introgression lines. *Current Opinion in Genetics & Development*, **17**, 545-552.

Liu, J., Van Eck, J, Cong, B. and Tanksley, S. (2002) A new class of regulatory genes underlying the cause of pear-shaped tomato fruit. *Proc. Natl. Acad. Sci*. 99, 13302-13306.

Liu, K.J., Dai, J., Truong, K., Song, Y., Kohn, M.H. and Nakhleh, L. (2014) An HMM-based comparative genomic framework for detecting introgression in eukaryotes. *PLoS Computational Biology*, **10**, e1003649.

Loftus, B.J., Fung, E., Roncaglia, P., Rowley, D., Amedeo, P., Bruno, D., Vamathevan, J., Miranda, M., Anderson, I.J., Fraser, J.A., Allen, J.E., Bosdet, I.E., Brent, M.R., Chiu, R., Doering, T.L., Donlin, M.J., D'Souza, C.A., Fox, D.S., Grinberg, V., Fu, J., Fukushima, M., Haas, B.J., Huang, J.C., Janbon, G., Jones, S.J., Koo, H.L., Krzywinski, M.I., Kwon-Chung, J.K., Lengeler, K.B., Maiti, R., Marra, M.A., Marra, R.E., Mathewson, C.A., Mitchell, T.G., Pertea, M., Riggs, F.R., Salzberg, S.L., Schein, J.E., Shvartsbeyn, A., Shin, H., Shumway, M., Specht, C.A., Suh, B.B., Tenney, A., Utterback, T.R., Wickes, B.L., Wortman, J.R., Wye, N.H., Kronstad, J.W., Lodge, J.K., Heitman, J., Davis, R.W., Fraser, C.M. and Hyman, R.W. (2005) The genome of the basidiomycetous yeast and human pathogen Cryptococcus neoformans. *Science (New York, N.Y.)*, **307**, 1321-1324.

Ma, L., Vu, G.T.H., Schubert, V., Watanabe, K., Stein, N., Houben, A. and Schubert, I. (2010) Synteny between Brachypodium distachyon and Hordeum vulgare as revealed by FISH. *Chromosome Research*, **18**, 841-850.

Machida, M., Asai, K., Sano, M., Tanaka, T., Kumagai, T., Terai, G., Kusumoto, K., Arima, T., Akita, O., Kashiwagi, Y., Abe, K., Gomi, K., Horiuchi, H., Kitamoto, K., Kobayashi, T., Takeuchi, M., Denning, D.W., Galagan, J.E., Nierman, W.C., Yu, J., Archer, D.B., Bennett, J.W., Bhatnagar, D., Cleveland, T.E., Fedorova, N.D., Gotoh, O., Horikawa, H., Hosoyama, A., Ichinomiya, M., Igarashi, R., Iwashita, K., Juvvadi, P.R., Kato, M., Kato, Y., Kin, T., Kokubun, A., Maeda, H., Maeyama, N., Maruyama, J., Nagasaki, H., Nakajima, T., Oda, K., Okada, K., Paulsen, I., Sakamoto, K., Sawano, T., Takahashi, M., Takase, K., Terabayashi, Y., Wortman, J.R., Yamada, O., Yamagata, Y., Anazawa, H., Hata, Y., Koide, Y., Komori, T., Koyama, Y., Minetoki, T., Suharnan, S., Tanaka, A., Isono, K., Kuhara, S., Ogasawara, N. and Kikuchi, H. (2005) Genome sequencing and analysis of Aspergillus oryzae. *Nature*, **438**, 1157-1161.

MacKay, T.F.C., Richards, S., Stone, E.A. ***Barbadilla, A., Ayroles, J.F., Zhu, D., Casillas, S., Han, Y., Magwire, M.M., Cridland, J.M., Richardson, M.F., Anholt, R.R., Barrón, M., Bess, C., Blankenburg, K.P., Carbone, M.A., Castellano, D., Chaboub, L., Duncan, L., Harris, Z., Javaid, M., Jayaseelan, J.C., Jhangiani, S.N., Jordan, K.W., Lara, F., Lawrence, F., Lee, S.L., Librado, P., Linheiro, R.S., Lyman, R.F., Mackey, A.J., Munidasa, M., Muzny, D.M., Nazareth, L., Newsham, I., Perales, L., Pu, L.L., Qu, C., Ràmia, M., Reid, J.G., Rollmann, S.M., Rozas, J., Saada, N., Turlapati, L., Worley, K.C., Wu, Y.Q., Yamamoto, A., Zhu, Y., Bergman, C.M., Thornton, K.R., Mittelman, D., Gibbs, R.A.*** (2012) The *Drosophila melanogaster* genetic reference panel. *Nature*, 482, 173-178.

Marcais, G. and Kingsford, C. (2011) A fast, lock-free approach for efficient parallel counting of occurrences of *k*-mers. *Bioinformatics*, **27**, 764-770.

Martin, D.P., Lemey, P. and Posada, D. (2011) Analysing recombination in nucleotide sequences. *Molecular Ecology Resources*, **11**, 943-955.

Matesanz, S., Gianoli, E. and Valladares, F. (2010) Global change and the evolution of phenotypic plasticity in plants. *Annals of the New York Academy of Sciences*, **1206**, 35-55.

McHardy, A.C. and Rigoutsos, I. (2007) What's in the mix: phylogenetic classification of metagenome sequence samples. *Current opinion in microbiology*, **10**, 499-503.

Meinke, D.W. (1998) *Arabidopsis thaliana*: A Model Plant for Genome Analysis. *Science*, **282**, 662-682.

Miles, C. and Wayne, M. (2008) Quantitative trait locus (QTL) analysis. *Nature Education*, **1**, 208.

Miller, J.C. and Tanksley, S.D. (1990) RFLP analysis of phylogenetic relationships and genetic variation in the genus *Lycopersicon*. *Theor. Appl. Genet*. 80, 437-448.

Mosteller F. and Tukey J. (1977). Data Analysis and Regression. Reading, MA: Addison-Wesley. ISBN 0-201-04854-X

Moyle, L.C. (2008) Ecological and evolutionary genomics in the wild tomatoes (*Solanum* sect. *Lycopersicon*). *Evolution* 62, 2995-3013.

Mueller, L.A. (2005) The SOL Genomics Network. A Comparative Resource for Solanaceae Biology and Beyond. *PLANT PHYSIOLOGY*, **138**, 1310-1317.

Mueller, L.A., Lankhorst, R.K., Tanksley, S.D., Giovannoni, J.J., White, R., Vrebalov, J., Fei, Z., van Eck, J., Buels, R., Mills, A.A., Menda, N., Tecle, I.Y., Bombarely, A., Stack, S., Royer, S.M., Chang, S.B., Shearer, L.A., Kim, B.D., Jo, S.H., Hur, C.G., Choi, D., Li, C.B., Zhao, J., Jiang, H., Geng, Y., Dai, Y., Fan, H., Chen, J., Lu, F., Shi, J., Sun, S., Chen, J., Yang, X., Lu, C., Chen, M., Cheng, Z., Li, C., Ling, H., Xue, Y., Wang, Y., Seymour, G.B., Bishop, G.J., Bryan, G., Rogers, J., Sims, S., Butcher, S., Buchan, D., Abbott, J., Beasley, H., Nicholson, C., Riddle, C., Humphray, S., McLaren, K., Mathur, S., Vyas, S., Solanke, A.U., Kumar, R., Gupta, V., Sharma, A.K., Khurana, P., Khurana, J.P., Tyagi, A., Sarita, Chowdhury, P., Shridhar, S., Chattopadhyay, D., Pandit, A., Singh, P., Kumar, A., Dixit, R., Singh, A., Praveen, S., Dalal, V., Yadav, M., Ghazi, I.A., Gaikwad, K., Sharma, T.R., Mohapatra, T., Singh, N.K., Szinay, D., de Jong, H., Peters, S., van Staveren, M., Datema, E., Fiers, M.W.E.J., van Ham, R.C.H.J., Lindhout, P., Philippot, M., Frasse, P., Regad, F., Zouine, M., Bouzayen, M., Asamizu, E., Sato, S., Fukuoka, H., Tabata, S., Shibata, D., Botella, M.A., Perez-Alonso, M., Fernandez-Pedrosa, V., Osorio, S., Mico, A., Granell, A., Zhang, Z., He, J., Huang, S., Du, Y., Qu, D., Liu, L., Liu, D., Wang, J., Ye, Z., Yang, W., Wang, G., Vezzi, A., Todesco, S., Valle, G., Falcone, G., Pietrella, M., Giuliano, G., Grandillo, S., Traini, A., D'Agostino, N., Chiusano, M.L., Ercolano, M., Barone, A., Frusciante, L., Schoof, H., Jocker, A., Bruggmann, R., Spannagl, M., Mayer, K.X.F., Guigo, R., Camara, F., Rombauts, S., Fawcett, J.A., Van de Peer, Y., Knapp, S., Zamir, D. and Stiekema, W. (2009) A Snapshot of the Emerging Tomato Genome Sequence. *The Plant Genome*, **2**, 78-92.

Mueller, L.A., Tanksley, S.D., Giovannoni, J.J., van Eck, J., Stack, S., Choi, D., Kim, B.D., Chen, M., Cheng, Z., Li, C., Ling, H., Xue, Y., Seymour, G., Bishop, G., Bryan, G., Sharma, R., Khurana, J., Tyagi, A., Chattopadhyay, D., Singh, N.K., Stiekema, W., Lindhout, P., Jesse, T., Lankhorst, R.K., Bouzayen, M., Shibata, D., Tabata, S., Granell, A., Botella, M.A., Giuliano, G., Frusciante, L., Causse, M. and Zamir, D. (2005) The Tomato Sequencing Project, the First Cornerstone of the International Solanaceae Project (SOL). *Comparative and Functional Genomics*, **6**, 153-158.

Munos, S., Ranc, N., Botton, E., Bérard, A., Rolland, S., Duffé, P., Carretero, Y., Le Paslier, M-C., Delalande, C., Bouzayen, M., Brunel, D. and Causse, M. (2011) Increase in tomato locule number is controlled by two single-nucleotide polymorphisms located near *WUSCHEL*. *Plant Phys*. 156, 2244-2254.

Nederbragt, L. developments in NGS.

Nesbitt, T.C. and Tanksley, S.D. (2002) Comparative sequencing in the genus *Lycopersicon*: Implications for the evolution of fruit size in the domestication of cultivated tomatoes. *Genetics*, 162, 365-379.

Nicotra, A.B., Atkin, O.K., Bonser, S.P., Davidson, A.M., Finnegan, E.J., Mathesius, U., Poot, P., Purugganan, M.D., Richards, C.L., Valladares, F. and van Kleunen, M. (2010) Plant phenotypic plasticity in a changing climate. *Trends in Plant Science*, **15**, 684-692.

Nierman, W.C., Pain, A., Anderson, M.J., Wortman, J.R., Kim, H.S., Arroyo, J., Berriman, M., Abe, K., Archer, D.B., Bermejo, C., Bennett, J., Bowyer, P., Chen, D., Collins, M., Coulsen, R., Davies, R., Dyer, P.S., Farman, M., Fedorova, N., Fedorova, N., Feldblyum, T.V., Fischer, R., Fosker, N., Fraser, A., Garcia, J.L., Garcia, M.J., Goble, A., Goldman, G.H., Gomi, K., Griffith-Jones, S., Gwilliam, R., Haas, B., Haas, H., Harris, D., Horiuchi, H., Huang, J., Humphray, S., Jimenez, J., Keller, N., Khouri, H., Kitamoto, K., Kobayashi, T., Konzack, S., Kulkarni, R., Kumagai, T., Lafon, A., Latge, J.P., Li, W., Lord, A., Lu, C., Majoros, W.H., May, G.S., Miller, B.L., Mohamoud, Y., Molina, M., Monod, M., Mouyna, I., Mulligan, S., Murphy, L., O'Neil, S., Paulsen, I., Penalva, M.A., Pertea, M., Price, C., Pritchard, B.L., Quail, M.A., Rabbinowitsch, E., Rawlins, N., Rajandream, M.A., Reichard, U., Renauld, H., Robson, G.D., Rodriguez de Cordoba, S., Rodriguez-Pena, J.M., Ronning, C.M., Rutter, S., Salzberg, S.L., Sanchez, M., Sanchez-Ferrero, J.C., Saunders, D., Seeger, K., Squares, R., Squares, S., Takeuchi, M., Tekaia, F., Turner, G., Vazquez de Aldana, C.R., Weidman, J., White, O., Woodward, J., Yu, J.H., Fraser, C., Galagan, J.E., Asai, K., Machida, M., Hall, N., Barrell, B. and Denning, D.W. (2005) Genomic sequence of the pathogenic and allergenic filamentous fungus Aspergillus fumigatus. *Nature*, **438**, 1151-1156.

Nybom, H., Weising, K. and Rotter, B. (2014) DNA fingerprinting in botany: past, present, future. *Invest Genet*, **5**, 1.

Paraskevis, D., Deforche, K., Lemey, P., Magiorkinis, G., Hatzakis, A. and Vandamme, A.M. (2005) SlidingBayes: exploring recombination using a sliding window approach based on Bayesian phylogenetic inference. *Bioinformatics*, **21**, 1274-1275.

Park, S.J., Jiang, K., Schatz, M.C. and Lippman, Z.B. (2012) Rate of meristem maturation determines inflorescence architecture in tomato. *Proceedings of the National Academy of Sciences of the United States of America*, **109**, 639-644.

Paterson, A.H., Bowers, J.E., Burow, M.D., Draye, X., Elsik, C.G., Jiang, C-X., Katsar, C.S., Lan, T-H., Lin Y-R., Ming, R. and Wright, R.J. (2000) Comparative genomics of plant chromosomes. *Plant Cell*, 12, 1523-1539.

Pel, H.J., de Winde, J.H., Archer, D.B., Dyer, P.S., Hofmann, G., Schaap, P.J., Turner, G., de Vries, R.P., Albang, R., Albermann, K., Andersen, M.R., Bendtsen, J.D., Benen, J.A., van den Berg, M., Breestraat, S., Caddick, M.X., Contreras, R., Cornell, M., Coutinho, P.M., Danchin, E.G., Debets, A.J., Dekker, P., van Dijck, P.W., van Dijk, A., Dijkhuizen, L., Driessen, A.J., d'Enfert, C., Geysens, S., Goosen, C., Groot, G.S., de Groot, P.W., Guillemette, T., Henrissat, B., Herweijer, M., van den Hombergh, J.P., van den Hondel, C.A., van der Heijden, R.T., van der Kaaij, R.M., Klis, F.M., Kools, H.J., Kubicek, C.P., van Kuyk, P.A., Lauber, J., Lu, X., van der Maarel, M.J., Meulenberg, R., Menke, H., Mortimer, M.A., Nielsen, J., Oliver, S.G., Olsthoorn, M., Pal, K., van Peij, N.N., Ram, A.F., Rinas, U., Roubos, J.A., Sagt, C.M., Schmoll, M., Sun, J., Ussery, D., Varga, J., Vervecken, W., van de Vondervoort, P.J., Wedler, H., Wosten, H.A., Zeng, A.P., van Ooyen, A.J., Visser, J. and Stam, H. (2007) Genome sequencing and analysis of the versatile cell factory Aspergillus niger CBS 513.88. *Nature biotechnology*, **25**, 221-231.

Peralta, I.E., Knapp, S. and Spooner, D.M. (2005) New Species of Wild Tomatoes (*Solanum* Section *Lycopersicon*: Solanaceae) from Northern Peru. *Syst. Bot*., 30, 424-434.

Peralta, I.E., Spooner, D.M. and Knapp, S. (2008) Taxonomy of wild tomatoes and their relatives (*Solanum* sect. *Lycopersicoides*, sect *Jugandifolia*, sect. *Lycopersicon*; *Solanaceae*). In *Systematic botany monographs*, 84 (Anderson, C. ed.). USA: The American Society of Plant Taxonomists ISBN 987-0-912861-84-5, pp 1-186.

Peters, S.A., Bargsten, J.W., Szinay, D., van de Belt, J., Visser, R.G.F., Bai, Y. and de Jong, H. (2012) Structural homology in the Solanaceae: analysis of genomic regions in support of synteny studies in tomato, potato and pepper. *The Plant Journal*, **71**, 602-614.

Peters, S.A., Datema, E., Szinay, D., van Staveren, M.J., Schijlen, E.G.W.M., van Haarst, J.C., Hesselink, T., Abma-Henkens, M.H.C., Bai, Y., de Jong, H., Stiekema, W.J., Klein Lankhorst, R.M. and van Ham, R.C.H.J. (2009) *Solanum lycopersicum* cv. Heinz 1706 chromosome 6: distribution and abundance of genes and retrotransposable elements. *The Plant Journal*, **58**, 857-869.

Peters, S.A., van Haarst, J.C., Jesse, T.P., Woltinge, D., Jansen, K., Hesselink, T., van Staveren, M.J., Abma-Henkens, M.H.C. and Klein-Lankhorst, R.M. (2006) TOPAAS, a Tomato and Potato Assembly Assistance System for Selection and Finishing of Bacterial Artificial Chromosomes. *PLANT PHYSIOLOGY*, **140**, 805-817.

Peterson, D.G., Lapitan, N.L. and Stack, S.M. (1999) Localization of single- and low-copy sequences on tomato synaptonemal complex spreads using fluorescence *in situ* hybridization (FISH). *Genetics*, **152**, 427-439.

Peterson, D.G., Stack, S.M., Price, H.J. and Johnston, J.S. (1996) DNA content of heterochromatin and euchromatin in tomato (Lycopersicon esculentum) pachytene chromosomes. *Genome*, **39**, 77-82.

Pettengill, J.B., Luo, Y., Davis, S., Chen, Y., Gonzalez-Escalona, N., Ottesen, A., Rand, H., Allard, M.W. and Strain, E. (2014) An evaluation of alternative methods for constructing phylogenies from whole genome sequence data: a case study with Salmonella. *PeerJ*, **2**, e620.

Pevzner, P.A., Tang, H. and Waterman, M.S. (2001) An Eulerian path approach to DNA fragment assembly. *Proceedings of the National Academy of Sciences*, **98**, 9748-9753.

Pigozzi, M.I. (2007) Localization of single-copy sequences on chicken synaptonemal complex spreads using fluorescence *in situ* hybridization (FISH). *Cytogenetic and Genome Research*, **119**, 105-112.

Poelstra, J.W., Vijay, N., Bossu, C.M., Lantz, H., Ryll, B., Muller, I., Baglione, V., Unneberg, P., Wikelski, M., Grabherr, M.G. and Wolf, J.B. (2014) The genomic landscape underlying phenotypic integrity in the face of gene flow in crows. *Science (New York, N.Y.)*, **344**, 1410-1414.

Posada, D. (2002) Evaluation of methods for detecting recombination from DNA sequences: empirical data. *Mol. Biol. Evol*. **19**, 708-717.

Poursarebani, N., Ma, L., Schmutzer, T., Houben, A. and Stein, N. (2014) FISH Mapping for Physical Map Improvement in the Large Genome of Barley: A Case Study on Chromosome 2H. *Cytogenetic and Genome Research*.

Powell, W., Morgante, M., Andre, C., Hanafey, M., Vogel, J., Tingey, S. and Rafalski, A. (1996) The comparison of RFLP, RAPD, AFLP and SSR (microsatellite) markers for germplasm analysis. *Mol. Breed*. **2**, 225-238.

Prasad, A.B., Mullikin, J.C., Program, N.C.S. and Green, E.D. (2013) A scalable and flexible approach for investigating the genomic landscapes of phylogenetic incongruence. *Mol. Phyl. Evol*. **66**, 1067-1074.

Price, M.N., Dehal, P.S. and Arkin, A.P. (2010) FastTree 2 – Approximately Maximum-Likelihood trees for large alignments. *PLoS ONE* 5, 3.

Qi, L., Friebe, B., Zhang, P. and Gill, B.S. (2007) Homoeologous recombination, chromosome engineering and crop improvement. *Chromosome Res*. **15,** 3-19.

Qiu, Q., Zhang, G., Ma, T., Qian, W., Wang, J., Ye, Z., Cao, C., Hu, Q., Kim, J., Larkin, D.M., Auvil, L., Capitanu, B., Ma, J., Lewin, H.A., Qian, X., Lang, Y., Zhou, R., Wang, L., Wang, K., Xia, J., Liao, S., Pan, S., Lu, X., Hou, H., Wang, Y., Zang, X., Yin, Y., Ma, H., Zhang, J., Wang, Z., Zhang, Y., Zhang, D., Yonezawa, T., Hasegawa, M., Zhong, Y., Liu, W., Zhang, Y., Huang, Z., Zhang, S., Long, R., Yang, H., Wang, J., Lenstra, J.A., Cooper, D.N., Wu, Y., Wang, J., Shi, P., Wang, J. and Liu, J. (2012) The yak genome and adaptation to life at high altitude. *Nature genetics*, **44**, 946-949.

Ramanna, M.S. and Prakken, R. (1967) Structure of and homology between pachytene and somatic metaphase chromosomes of the tomato. *Genetica*, **38**, 115-133.

Ren, J., Song, K., Sun, F., Deng, M. and Reinert, G. (2013) Multiple alignment-free sequence comparison. *Bioinformatics*, **29**, 2690-2698.

Ribaut, J.-M. and Hoisington, D. (1998) Marker-assisted selection: new tools and strategies. *Trends in Plant Science*, **3**, 236-239.

Richards, S., Gibbs, R.A., Weinstock, G.M., Brown, S.J., Denell, R., Beeman, R.W., Gibbs, R., Beeman, R.W., Brown, S.J., Bucher, G., Friedrich, M., Grimmelikhuijzen, C.J., Klingler, M., Lorenzen, M., Richards, S., Roth, S., Schroder, R., Tautz, D., Zdobnov, E.M., Muzny, D., Gibbs, R.A., Weinstock, G.M., Attaway, T., Bell, S., Buhay, C.J., Chandrabose,

M.N., Chavez, D., Clerk-Blankenburg, K.P., Cree, A., Dao, M., Davis, C., Chacko, J., Dinh, H., Dugan-Rocha, S., Fowler, G., Garner, T.T., Garnes, J., Gnirke, A., Hawes, A., Hernandez, J., Hines, S., Holder, M., Hume, J., Jhangiani, S.N., Joshi, V., Khan, Z.M., Jackson, L., Kovar, C., Kowis, A., Lee, S., Lewis, L.R., Margolis, J., Morgan, M., Nazareth, L.V., Nguyen, N., Okwuonu, G., Parker, D., Richards, S., Ruiz, S.J., Santibanez, J., Savard, J., Scherer, S.E., Schneider, B., Sodergren, E., Tautz, D., Vattahil, S., Villasana, D., White, C.S., Wright, R., Park, Y., Beeman, R.W., Lord, J., Oppert, B., Lorenzen, M., Brown, S., Wang, L., Savard, J., Tautz, D., Richards, S., Weinstock, G., Gibbs, R.A., Liu, Y., Worley, K., Weinstock, G., Elsik, C.G., Reese, J.T., Elhaik, E., Landan, G., Graur, D., Arensburger, P., Atkinson, P., Beeman, R.W., Beidler, J., Brown, S.J., Demuth, J.P., Drury, D.W., Du, Y.Z., Fujiwara, H., Lorenzen, M., Maselli, V., Osanai, M., Park, Y., Robertson, H.M., Tu, Z., Wang, J.J., Wang, S., Richards, S., Song, H., Zhang, L., Sodergren, E., Werner, D., Stanke, M., Morgenstern, B., Solovyev, V., Kosarev, P., Brown, G., Chen, H.C., Ermolaeva, O., Hlavina, W., Kapustin, Y., Kiryutin, B., Kitts, P., Maglott, D., Pruitt, K., Sapojnikov, V., Souvorov, A., Mackey, A.J., Waterhouse, R.M., Wyder, S., Zdobnov, E.M., Zdobnov, E.M., Wyder, S., Kriventseva, E.V., Kadowaki, T., Bork, P., Aranda, M., Bao, R., Beermann, A., Berns, N., Bolognesi, R., Bonneton, F., Bopp, D., Brown, S.J., Bucher, G., Butts, T., Chaumot, A., Denell, R.E., Ferrier, D.E., Friedrich, M., Gordon, C.M., Jindra, M., Klingler, M., Lan, Q., Lattorff, H.M., Laudet, V., von Levetsow, C., Liu, Z., Lutz, R., Lynch, J.A., da Fonseca, R.N., Posnien, N., Reuter, R., Roth, S., Savard, J., Schinko, J.B., Schmitt, C., Schoppmeier, M., Schroder, R., Shippy, T.D., Simonnet, F., Marques-Souza, H., Tautz, D., Tomoyasu, Y., Trauner, J., Van der Zee, M., Vervoort, M., Wittkopp, N., Wimmer, E.A., Yang, X., Jones, A.K., Sattelle, D.B., Ebert, P.R., Nelson, D., Scott, J.G., Beeman, R.W., Muthukrishnan, S., Kramer, K.J., Arakane, Y., Beeman, R.W., Zhu, Q., Hogenkamp, D., Dixit, R., Oppert, B., Jiang, H., Zou, Z., Marshall, J., Elpidina, E., Vinokurov, K., Oppert, C., Zou, Z., Evans, J., Lu, Z., Zhao, P., Sumathipala, N., Altincicek, B., Vilcinskas, A., Williams, M., Hultmark, D., Hetru, C., Jiang, H., Grimmelikhuijzen, C.J., Hauser, F., Cazzamali, G., Williamson, M., Park, Y., Li, B., Tanaka, Y., Predel, R., Neupert, S., Schachtner, J., Verleyen, P., Raible, F., Bork, P., Friedrich, M., Walden, K.K., Robertson, H.M., Angeli, S., Foret, S., Bucher, G., Schuetz, S., Maleszka, R., Wimmer, E.A., Beeman, R.W., Lorenzen, M., Tomoyasu, Y., Miller, S.C., Grossmann, D. and Bucher, G. (2008) The genome of the model beetle and pest Tribolium castaneum. *Nature*, **452**, 949-955.

Rick, C.M. (1979) Biosystematic studies in *Lycopersicon* and closely related species of *Solanum*. In *The biology and taxonomy of Solanacae*. (Hawkes, J.G., Lester, R.N., and Skelding, A.D. eds). New York, USA: Academic Press, pp 667-677.

Rick, C.M. (1980) Tomato linkage survey. *Rep Tomato Genet Coop*, **30**, 17.

Rick, C.M. (1986) Reproductive isolation in the *Lycopersicum peruvianum* complex. In *Solanaceae, biology and systematics.* (D'Arcy, W.G.D. ed). New York, USA: Columbia University Press, pp 477-495.

Rick, C.M. (1991) Tomato paste: a concentrated review of genetic highlights from the beginnings to the advent of molecular genetics. *Genetics*, **128**, 1-5.

Rieseberg, L.H. and Ellstrand, N.C. (1993) What Can Molecular and Morphological Markers Tell Us About Plant Hybridization. *Crit. Rev. Plant. Sci.* **12**, 213-241.

Rodriguez, G.R., Munos, S., Anderson, C., Sim, S.C., Michel, A., Causse, M., Gardener, B.B.M., Francis, D. and van der Knaap, E. (2011) Distribution of SUN, *Ovate*, LC, and FAS in the Tomato Germplasm and the Relationship to Fruit Shape Diversity. *PLANT PHYSIOLOGY*, **156**, 275-285.

Rogers, S.O. and Bendich, A.J. (1988) Extraction of DNA from plant tissues. In *Plant Molecular Biology Manual*: Springer Science + Business Media, pp. 89-99.

Ronen, G., Carmel-Goren, L., Zamir, D. and Hirschberg, J. (2000) An alternative pathway to ß-carotene formation in plant chromoplasts discovered by mapped based cloning of *Beta* and *old-gold color* mutations in tomato. *Proc. Natl. Acad. Sci*. 97, 11102-11107.

Rossetto, M. and Henry, R.J. (2014) Escape from the laboratory: new horizons for plant genetics. *Trends in Plant Science*, **19**, 554-555.

Roychowdhury, T., Vishnoi, A. and Bhattacharya, A. (2013) Next-Generation Anchor Based Phylogeny (NexABP): constructing phylogeny from next-generation sequencing data. *Scientific reports*, **3**, 2634.

Salem, M., Vallejo, R.L., Leeds, T.D., Palti, Y., Liu, S., Sabbagh, A., Rexroad, C.E. and Yao, J. (2012) RNA-Seq Identifies SNP Markers for Growth Traits in Rainbow Trout. *PLoS ONE*, **7**, e36264.

Saliba-Colombani, V., Causse, M., Langlois, D., Philouze, J. and Buret, M. (2001) Genetic analysis of organoleptic quality in fresh market tomato. 1. Mapping QTLs for physical and chemical traits. *TAG Theoretical and Applied Genetics*, **102**, 259-272.

Sato, S., Tabata, S., Hirakawa, H., Asamizu, E., Shirasawa, K., Isobe, S., Kaneko, T., Nakamura, Y., Shibata, D., Aoki, K., Egholm, M., Knight, J., Bogden, R., Li, C., Shuang, Y., Xu, X., Pan, S., Cheng, S., Liu, X., Ren, Y., Wang, J., Albiero, A., Dal Pero, F., Todesco, S., Van Eck, J., Buels, R.M., Bombarely, A., Gosselin, J.R., Huang, M., Leto, J.A., Menda, N., Strickler, S., Mao, L., Gao, S., Tecle, I.Y., York, T., Zheng, Y., Vrebalov, J.T., Lee, J., Zhong, S., Mueller, L.A., Stiekema, W.J., Ribeca, P., Alioto, T., Yang, W., Huang, S., Du, Y., Zhang, Z., Gao, J., Guo, Y., Wang, X., Li, Y., He, J., Li, C., Cheng, Z., Zuo, J., Ren, J., Zhao, J., Yan, L., Jiang, H., Wang, B., Li, H., Li, Z., Fu, F., Chen, B., Han, B., Feng, Q., Fan, D., Wang, Y., Ling, H., Xue, Y., Ware, D., Richard McCombie, W., Lippman, Z.B., Chia, J.-M., Jiang, K., Pasternak, S., Gelley, L., Kramer, M., Anderson, L.K., Chang, S.-B., Royer, S.M., Shearer, L.A., Stack, S.M., Rose, J.K.C., Xu, Y., Eannetta, N., Matas, A.J., McQuinn, R., Tanksley, S.D., Camara, F., Guigó, R., Rombauts, S., Fawcett, J., Van de Peer, Y., Zamir, D., Liang, C., Spannagl, M., Gundlach, H., Bruggmann, R., Mayer, K., Jia, Z., Zhang, J., Ye, Z., Bishop, G.J., Butcher, S., Lopez-Cobollo, R., Buchan, D., Filippis, I., Abbott, J., Dixit, R., Singh, M., Singh, A., Kumar Pal, J., Pandit, A., Kumar Singh, P., Kumar Mahato, A., Dogra, V., Gaikwad, K., Raj Sharma, T., Mohapatra, T., Kumar Singh, N., Causse, M., Rothan, C., Schiex, T., Noirot, C., Bellec, A., Klopp, C., Delalande, C., Berges, H., Mariette, J., Frasse, P., Vautrin, S., Zouine, M., Latché, A., Rousseau, C., Regad, F., Pech, J.-C., Philippot, M., Bouzayen, M., Pericard, P., Osorio, S., Fernandez del Carmen, A., Monforte, A., Granell, A., Fernandez-Muñoz, R., Conte, M., Lichtenstein, G., Carrari, F., De Bellis, G., Fuligni, F., Peano, C., Grandillo, S., Termolino, P., Pietrella, M., Fantini, E., Falcone, G., Fiore, A., Giuliano, G., Lopez, L., Facella, P., Perrotta, G., Daddiego, L., Bryan, G., Orozco, M., Pastor, X., Torrents, D., van Schriek, M.G.M., Feron, R.M.C., van Oeveren, J., de Heer, P., daPonte, L., Jacobs-Oomen, S., Cariaso, M., Prins, M., van Eijk, M.J.T., Janssen, A., van Haaren, M.J.J., Jo, S.-H., Kim, J., Kwon, S.-Y., Kim, S., Koo,

D.-H., Lee, S., Hur, C.-G., Clouser, C., Rico, A., Hallab, A., Gebhardt, C., Klee, K., Jöcker, A., Warfsmann, J., Göbel, U., Kawamura, S., Yano, K., Sherman, J.D., Fukuoka, H., Negoro, S., Bhutty, S., Chowdhury, P., Chattopadhyay, D., Datema, E., Smit, S., Schijlen, E.G.W.M., van de Belt, J., van Haarst, J.C., Peters, S.A., van Staveren, M.J., Henkens, M.H.C., Mooyman, P.J.W., Hesselink, T., van Ham, R.C.H.J., Jiang, G., Droege, M., Choi, D., Kang, B.-C., Dong Kim, B., Park, M., Kim, S., Yeom, S.-I., Lee, Y.-H., Choi, Y.-D., Li, G., Gao, J., Liu, Y., Huang, S., Fernandez-Pedrosa, V., Collado, C., Zuñiga, S., Wang, G., Cade, R., Dietrich, R.A., Rogers, J., Knapp, S., Fei, Z., White, R.A., Thannhauser, T.W., Giovannoni, J.J., Angel Botella, M., Gilbert, L., Gonzalez, R., Luis Goicoechea, J., Yu, Y., Kudrna, D., Collura, K., Wissotski, M., Wing, R., Schoof, H., Meyers, B.C., Bala Gurazada, A., Green, P.J., Mathur, S., Vyas, S., Solanke, A.U., Kumar, R., Gupta, V., Sharma, A.K., Khurana, P., Khurana, J.P., Tyagi, A.K., Dalmay, T., Mohorianu, I., Walts, B., Chamala, S., Brad Barbazuk, W., Li, J., Guo, H., Lee, T.-H., Wang, Y., Zhang, D., Paterson, A.H., Wang, X., Tang, H., Barone, A., Luisa Chiusano, M., Raffaella Ercolano, M., D'Agostino, N., Di Filippo, M., Traini, A., Sanseverino, W., Frusciante, L., Seymour, G.B., Elharam, M., Fu, Y., Hua, A., Kenton, S., Lewis, J., Lin, S., Najar, F., Lai, H., Qin, B., Qu, C., Shi, R., White, D., White, J., Xing, Y., Yang, K., Yi, J., Yao, Z., Zhou, L., Roe, B.A., Vezzi, A., D'Angelo, M., Zimbello, R., Schiavon, R., Caniato, E., Rigobello, C., Campagna, D., Vitulo, N., Valle, G., Nelson, D.R., De Paoli, E., Szinay, D., de Jong, H.H., Bai, Y., Visser, R.G.F., Klein Lankhorst, R.M., Beasley, H., McLaren, K., Nicholson, C., Riddle, C., Gianese, G., Sato, S., Tabata, S., Mueller, L.A., Huang, S., Du, Y., Li, C., Cheng, Z., Zuo, J., Han, B., Wang, Y., Ling, H., Xue, Y., Ware, D., Richard McCombie, W., Lippman, Z.B., Stack, S.M., Tanksley, S.D., Van de Peer, Y., Mayer, K., Bishop, G.J., Butcher, S., Kumar Singh, N., Schiex, T., Bouzayen, M., Granell, A., Carrari, F., De Bellis, G., Giuliano, G., Bryan, G., van Eijk, M.J.T., Fukuoka, H., Chattopadhyay, D., van Ham, R.C.H.J., Choi, D., Rogers, J., Fei, Z., Giovannoni, J.J., Wing, R., Schoof, H., Meyers, B.C., Khurana, J.P., Tyagi, A.K., Dalmay, T., Paterson, A.H., Wang, X., Frusciante, L., Seymour, G.B., Roe, B.A., Valle, G., de Jong, H.H. and Klein Lankhorst, R.M. (2012) The tomato genome sequence provides insights into fleshy fruit evolution. *Nature*, **485**, 635-641.

Scally, A., Dutheil, J.Y., Hillier, L.W., Jordan, G.E., Goodhead, I., Herrero, J., Hobolth, A., Lappalainen, T., Mailund, T., Marques-Bonet, T., McCarthy, S., Montgomery, S.H., Schwalie, P.C., Tang, Y.A., Ward, M.C., Xue, Y., Yngvadottir, B., Alkan, C., Andersen, L.N., Ayub, Q., Ball, E.V., Beal, K., Bradley, B.J., Chen, Y., Clee, C.M., Fitzgerald, S., Graves, T.A., Gu, Y., Heath, P., Heger, A., Karakoc, E., Kolb-Kokocinski, A., Laird, G.K., Lunter, G., Meader, S., Mort, M., Mullikin, J.C., Munch, K., O'Connor, T.D., Phillips, A.D., Prado-Martinez, J., Rogers, A.S., Sajjadian, S., Schmidt, D., Shaw, K., Simpson, J.T., Stenson, P.D., Turner, D.J., Vigilant, L., Vilella, A.J., Whitener, W., Zhu, B., Cooper, D.N., de Jong, P., Dermitzakis, E.T., Eichler, E.E., Flicek, P., Goldman, N., Mundy, N.I., Ning, Z., Odom, D.T., Ponting, C.P., Quail, M.A., Ryder, O.A., Searle, S.M., Warren, W.C., Wilson, R.K., Schierup, M.H., Rogers, J., Tyler-Smith, C. and Durbin, R. (2012) Insights into hominid evolution from the gorilla genome sequence. *Nature*, **483**, 169-175.

Schauer, N., Semel, Y., Roessner, U., Gur, A., Balbo, I., Carrari, F., Pleban, T., Perez-Melis, A., Bruedigam, C., Kopka, J., Willmitzer, L., Zamir, D. and Fernie, A.R. (2006) Comprehensive metabolic profiling and phenotyping of interspecific introgression lines for tomato improvement. *Nat Biotechnol*, **24**, 447-454.

Schloissnig, S., Arumugam, M., Sunagawa, S., Mitreva, M., Tap, J., Zhu, A., Waller, A., Mende, D.R., Kultima, J.R., Martin, J., Kota, K., Sunyaev, S.R., Weinstock, G.M. and Bork, P. (2013) Genomic variation landscape of the human gut microbiome. *Nature*, **493**, 45-50.

Schmitz, R.J., Schultz, M.D., Urich, M.A., Nery, J.R., Pelizzola, M., Libiger, O., Alix, A., McCosh, R.B., Chen, H., Schork, N.J., Ecker, J.R. (2013) Patterns of population epigenomic diversity. *Nature,* **495**, 193-198.

Schmutz, J., Cannon, S.B., Schlueter, J., Ma, J., Mitros, T., Nelson, W., Hyten, D.L., Song, Q., Thelen, J.J., Cheng, J., Xu, D., Hellsten, U., May, G.D., Yu, Y., Sakurai, T., Umezawa, T., Bhattacharyya, M.K., Sandhu, D., Valliyodan, B., Lindquist, E., Peto, M., Grant, D., Shu, S., Goodstein, D., Barry, K., Futrell-Griggs, M., Abernathy, B., Du, J., Tian, Z., Zhu, L., Gill, N., Joshi, T., Libault, M., Sethuraman, A., Zhang, X.C., Shinozaki, K., Nguyen, H.T., Wing, R.A., Cregan, P., Specht, J., Grimwood, J., Rokhsar, D., Stacey, G., Shoemaker, R.C. and Jackson, S.A. (2010) Genome sequence of the palaeopolyploid soybean. *Nature*, **463**, 178-183.

Schmutzer, T., Ma, L., Pousarebani, N., Bull, F., Stein, N., Houben, A. and Scholz, U. (2014) Kmasker - A Tool for in silico Prediction of Single-Copy FISH Probes for the Large-Genome Species Hordeum vulgare. *Cytogenetic and Genome Research*, **142**, 66-78.

Schnable, P.S., Ware, D., Fulton, R.S., Stein, J.C., Wei, F., Pasternak, S., Liang, C., Zhang, J., Fulton, L., Graves, T.A., Minx, P., Reily, A.D., Courtney, L., Kruchowski, S.S., Tomlinson, C., Strong, C., Delehaunty, K., Fronick, C., Courtney, B., Rock, S.M., Belter, E., Du, F., Kim, K., Abbott, R.M., Cotton, M., Levy, A., Marchetto, P., Ochoa, K., Jackson, S.M., Gillam, B., Chen, W., Yan, L., Higginbotham, J., Cardenas, M., Waligorski, J., Applebaum, E., Phelps, L., Falcone, J., Kanchi, K., Thane, T., Scimone, A., Thane, N., Henke, J., Wang, T., Ruppert, J., Shah, N., Rotter, K., Hodges, J., Ingenthron, E., Cordes, M., Kohlberg, S., Sgro, J., Delgado, B., Mead, K., Chinwalla, A., Leonard, S., Crouse, K., Collura, K., Kudrna, D., Currie, J., He, R., Angelova, A., Rajasekar, S., Mueller, T., Lomeli, R., Scara, G., Ko, A., Delaney, K., Wissotski, M., Lopez, G., Campos, D., Braidotti, M., Ashley, E., Golser, W., Kim, H., Lee, S., Lin, J., Dujmic, Z., Kim, W., Talag, J., Zuccolo, A., Fan, C., Sebastian, A., Kramer, M., Spiegel, L., Nascimento, L., Zutavern, T., Miller, B., Ambroise, C., Muller, S., Spooner, W., Narechania, A., Ren, L., Wei, S., Kumari, S., Faga, B., Levy, M.J., McMahan, L., Van Buren, P., Vaughn, M.W., Ying, K., Yeh, C.T., Emrich, S.J., Jia, Y., Kalyanaraman, A., Hsia, A.P., Barbazuk, W.B., Baucom, R.S., Brutnell, T.P., Carpita, N.C., Chaparro, C., Chia, J.M., Deragon, J.M., Estill, J.C., Fu, Y., Jeddeloh, J.A., Han, Y., Lee, H., Li, P., Lisch, D.R., Liu, S., Liu, Z., Nagel, D.H., McCann, M.C., SanMiguel, P., Myers, A.M., Nettleton, D., Nguyen, J., Penning, B.W., Ponnala, L., Schneider, K.L., Schwartz, D.C., Sharma, A., Soderlund, C., Springer, N.M., Sun, Q., Wang, H., Waterman, M., Westerman, R., Wolfgruber, T.K., Yang, L., Yu, Y., Zhang, L., Zhou, S., Zhu, Q., Bennetzen, J.L., Dawe, R.K., Jiang, J., Jiang, N., Presting, G.G., Wessler, S.R., Aluru, S., Martienssen, R.A., Clifton, S.W., McCombie, W.R., Wing, R.A. and Wilson, R.K. (2009) The B73 maize genome: complexity, diversity, and dynamics. *Science (New York, N.Y.)*, **326**, 1112-1115.

Schneeberger, K., Ossowski, S., Lanz, C., Juul, T., Petersen, A.H., Nielsen, K.L., Jorgensen, J.E., Weigel, D. and Andersen, S.U. (2009) SHOREMAP: simultaneous mapping and mutation identification by deep sequencing. *Nature Methods*, **6**, 550-551.

Schwarzacher, T., Anamthawat-Jónsson, K., Harrison, G.E., Islam, A.K.M.R., Jia, J.Z., King, I.P., Leitch, A.R., Miller, T.E., Reader, S.M., Rogers, W.J., Shi, M. and Heslop-Harrison, J.S. (1992) Genomics *in situ* hybridization to identify alien chromosomes and chromosome segments in wheat. *Theor. Appl. Genet.* **84**, 778-786.

Shapiro, M.D., Kronenberg, Z., Li, C., Domyan, E.T., Pan, H., Campbell, M., Tan, H., Huff, C.D., Hu, H., Vickrey, A.I., Nielsen, S.C., Stringham, S.A., Hu, H., Willerslev, E., Gilbert, M.T., Yandell, M., Zhang, G. and Wang, J. (2013) Genomic diversity and evolution of the head crest in the rock pigeon. *Science (New York, N.Y.)*, **339**, 1063-1067.

Shearer, L.A., Anderson, L.K., de Jong, H., Smit, S., Goicoechea, J.L., Roe, B.A., Hua, A., Giovannoni, J.J. and Stack, S.M. (2014) Fluorescence *In situ* Hybridization and Optical Mapping to Correct Scaffold Arrangement in the Tomato Genome. *G3: Genes|Genomes|Genetics*, **4**, 1395-1405.

Sherman, J.D. and Stack, S.M. (1995) Two-dimensional spreads of synaptonemal complexes from solanaceous plants. VI. High-resolution recombination nodule map for tomato (Lycopersicon esculentum). *Genetics*, **141**, 683-708.

Sierro, N., Battey, J.N., Ouadi, S., Bakaher, N., Bovet, L., Willig, A., Goepfert, S., Peitsch, M.C. and Ivanov, N.V. (2014) The tobacco genome sequence and its comparison with those of tomato and potato. *Nature communications*, **5**, 3833.

Simpson, J.T., Wong, K., Jackman, S.D., Schein, J.E., Jones, S.J. and Birol, I. (2009) ABySS: a parallel assembler for short read sequence data. *Genome research*, **19**, 1117-1123.

Sims, G.E., Jun, S.R., Wu, G.A. and Kim, S.H. (2009) Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. *Proceedings of the National Academy of Sciences of the United States of America*, **106**, 2677-2682.

Singh, R.J. (2007) Genetic Resources, Chromosome Engineering, and Crop Improvement: Vegetable Crops Volume 3. Boca Raton, USA: CRC Press.

Skinner, M.E., Uzilov, A.V., Stein, L.D., Mungall, C.J. and Holmes, I.H. (2009) JBrowse: a next-generation genome browser. *Genome Res.* **19**, 1630-1638.

Smit, A.F.A., R. Hubley, and P. Green. (2013) RepeatMasker Open-4.0. 2013–2015.

Smits, S.L., Bodewes, R., Ruiz-Gonzalez, A., Baumgartner, W., Koopmans, M.P., Osterhaus, A.D. and Schurch, A.C. (2014) Assembly of viral genomes from metagenomes. *Frontiers in microbiology*, **5**, 714.

Smyth, D. (2012) Faculty of 1000 evaluation for The tomato genome sequence provides insights into fleshy fruit evolution. In *F1000 - Post-publication peer review of the biomedical literature*: Faculty of 1000, Ltd.

Souciet, J.L., Dujon, B., Gaillardin, C., Johnston, M., Baret, P.V., Cliften, P., Sherman, D.J., Weissenbach, J., Westhof, E., Wincker, P., Jubin, C., Poulain, J., Barbe, V., Segurens, B., Artiguenave, F., Anthouard, V., Vacherie, B., Val, M.E., Fulton, R.S., Minx, P., Wilson, R., Durrens, P., Jean, G., Marck, C., Martin, T., Nikolski, M., Rolland, T., Seret, M.L., Casaregola, S., Despons, L., Fairhead, C., Fischer, G., Lafontaine, I., Leh, V., Lemaire, M., de Montigny, J., Neuveglise, C., Thierry, A., Blanc-Lenfle, I., Bleykasten, C., Diffels, J., Fritsch, E., Frangeul, L., Goeffon, A., Jauniaux, N., Kachouri-Lafond, R., Payen, C., Potier, S., Pribylova, L., Ozanne, C., Richard, G.F., Sacerdot, C., Straub, M.L. and Talla, E. (2009) Comparative genomics of protoploid Saccharomycetaceae. *Genome research*, **19**, 1696-1709.

Spooner, D.M., Peralta, I.E. and Knapp, S. (2005) Camparison of AFLPs with other markers for phylogenetic inference in wild tomatoes [*Solanum* L. section *Lycopersicon* (Mill.) Wettst.]. *Taxon*, 54, 43-61.

St Pierre, S.E., Ponting, L., Stefancsik, R., McQuilton, P. and FlyBase, C. (2014) FlyBase 102--advanced approaches to interrogating FlyBase. *Nucleic acids research*, **42**, D780-788.

Stack, S.M., Clarke, C.R., Cary, W.E. and Muffly, J.T. (1974) Different kinds of heterochromatin in higher plant chromosomes. *Journal of cell science*, **14**, 499-504.

Stack, S.M., Royer, S.M., Shearer, L.A., Chang, S.B., Giovannoni, J.J., Westfall, D.H., White, R.A. and Anderson, L.K. (2009) Role of Fluorescence *in situ* Hybridization in Sequencing the Tomato Genome. *Cytogenetic and Genome Research*, **124**, 339-350.

Staden, R. (1979) A strategy of DNA sequencing employing computer programs. *Nucl Acids Res*, **6**, 2601-2610.

Städler, T., Roselius K. and Stephan, W. (2005) Genealogical footprints of speciation processes in wild tomatoes: demography and evidence for historical gene flow. *Evolution*, 59, 1268-1279.

Stein, L.D., Knoppers, B.M., Campbell, P., Getz, G. and Korbel, J.O. (2015) Data analysis: Create a cloud commons. *Nature*, **523**, 149-151.

Stein, L.D., Mungall, C., Shu, S., Caudy, M., Mangone, M., Day, A., Nickerson, E., Stajich, J.E., Harris, T.W., Arva, A. and Lewis, S. (2002) The generic genome browser: a building block for a model organism system database. *Genome Res*. **12**, 1599-1610.

Strimmer, K., Forslund, K., Holland, B. and Moulton, V. (2003) A novel exploratory method for visual recombination detection. *Genome Biol*. **4**, R33.

Szinay, D., Wijnker, E., Van den Berg, R., Visser, R.G.F., De Jong, H., and Bai, Y. (2012) Chromosome evolution in *Solanum* traced by cross-species BAC-FISH. *New Phyt*. 195, 688-698.

Szinay, D.r., Chang, S.-B., Khrustaleva, L., Peters, S., Schijlen, E., Bai, Y., Stiekema, W.J., van Ham, R.C.H.J., de Jong, H. and Klein Lankhorst, R.M. (2008) High-resolution chromosome mapping of BACs using multi-colour FISH and pooled-BAC FISH as a backbone for sequencing tomato chromosome af6. *The Plant Journal*, **56**, 627-637.

Tabata, S., Kaneko, T., Nakamura, Y., Kotani, H., Kato, T., Asamizu, E., Miyajima, N., Sasamoto, S., Kimura, T., Hosouchi, T., Kawashima, K., Kohara, M., Matsumoto, M., Matsuno, A., Muraki, A., Nakayama, S., Nakazaki, N., Naruo, K., Okumura, S., Shinpo, S., Takeuchi, C., Wada, T., Watanabe, A., Yamada, M., Yasuda, M., Sato, S., de la Bastide, M., Huang, E., Spiegel, L., Gnoj, L., O'Shaughnessy, A., Preston, R., Habermann, K., Murray, J., Johnson, D., Rohlfing, T., Nelson, J., Stoneking, T., Pepin, K., Spieth, J., Sekhon, M., Armstrong, J., Becker, M., Belter, E., Cordum, H., Cordes, M., Courtney, L., Courtney, W., Dante, M., Du, H., Edwards, J., Fryman, J., Haakensen, B., Lamar, E., Latreille, P., Leonard, S., Meyer, R., Mulvaney, E., Ozersky, P., Riley, A., Strowmatt, C., Wagner-McPherson, C., Wollam, A., Yoakum, M., Bell, M., Dedhia, N., Parnell, L., Shah, R., Rodriguez, M., See, L.H., Vil, D., Baker, J., Kirchoff, K., Toth, K., King, L., Bahret, A., Miller, B., Marra, M., Martienssen, R., McCombie, W.R., Wilson, R.K., Murphy, G., Bancroft, I., Volckaert, G., Wambutt, R., Dusterhoft, A., Stiekema, W., Pohl, T., Entian, K.D., Terryn, N., Hartley, N., Bent, E., Johnson, S., Langham, S.A., McCullagh, B., Robben, J., Grymonprez, B., Zimmermann, W.,

Ramsperger, U., Wedler, H., Balke, K., Wedler, E., Peters, S., van Staveren, M., Dirkse, W., Mooijman, P., Lankhorst, R.K., Weitzenegger, T., Bothe, G., Rose, M., Hauf, J., Berneiser, S., Hempel, S., Feldpausch, M., Lamberth, S., Villarroel, R., Gielen, J., Ardiles, W., Bents, O., Lemcke, K., Kolesov, G., Mayer, K., Rudd, S., Schoof, H., Schueller, C., Zaccaria, P., Mewes, H.W., Bevan, M., Fransz, P., Kazusa, D.N.A.R.I., Cold Spring, H., Washington University in St Louis Sequencing, C. and European Union Arabidopsis Genome Sequencing, C. (2000) Sequence and analysis of chromosome 5 of the plant *Arabidopsis thaliana*. *Nature*, **408**, 823-826.

Tabata, S., Kaneko, T., Nakamura, Y., Kotani, H., Kato, T., Asamizu, E., Miyajima, N., Sasamoto, S., Kimura, T., Hosouchi, T., Kawashima, K., Kohara, M., Matsumoto, M., Matsuno, A., Muraki, A., Nakayama, S., Nakazaki, N., Naruo, K., Okumura, S., Shinpo, S., Takeuchi, C., Wada, T., Watanabe, A., Yamada, M., Yasuda, M., Sato, S., de la Bastide, M., Huang, E., Spiegel, L., Gnoj, L., O'Shaughnessy, A., Preston, R., Habermann, K., Murray, J., Johnson, D., Rohlfing, T., Nelson, J., Stoneking, T., Pepin, K., Spieth, J., Sekhon, M., Armstrong, J., Becker, M., Belter, E., Cordum, H., Cordes, M., Courtney, L., Courtney, W., Dante, M., Du, H., Edwards, J., Fryman, J., Haakensen, B., Lamar, E., Latreille, P., Leonard, S., Meyer, R., Mulvaney, E., Ozersky, P., Riley, A., Strowmatt, C., Wagner-McPherson, C., Wollam, A., Yoakum, M., Bell, M., Dedhia, N., Parnell, L., Shah, R., Rodriguez, M., See, L.H., Vil, D., Baker, J., Kirchoff, K., Toth, K., King, L., Bahret, A., Miller, B., Marra, M., Martienssen, R., McCombie, W.R., Wilson, R.K., Murphy, G., Bancroft, I., Volckaert, G., Wambutt, R., Dusterhoft, A., Stiekema, W., Pohl, T., Entian, K.D., Terryn, N., Hartley, N., Bent, E., Johnson, S., Langham, S.A., McCullagh, B., Robben, J., Grymonprez, B., Zimmermann, W., Ramsperger, U., Wedler, H., Balke, K., Wedler, E., Peters, S., van Staveren, M., Dirkse, W., Mooijman, P., Lankhorst, R.K., Weitzenegger, T., Bothe, G., Rose, M., Hauf, J., Berneiser, S., Hempel, S., Feldpausch, M., Lamberth, S., Villarroel, R., Gielen, J., Ardiles, W., Bents, O., Lemcke, K., Kolesov, G., Mayer, K., Rudd, S., Schoof, H., Schueller, C., Zaccaria, P., Mewes, H.W., Bevan, M. and Fransz, P. (2000) Sequence and analysis of chromosome 5 of the plant *Arabidopsis thaliana*. *Nature*, **408**, 823-826.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M. and Kumar, S. (2011) MEGA5: Molecular Evolutionary Genetics Analysis using Maximum Likelyhood, Evolutionary Distance, and Maximum Parimony Methods. *Mol. Biol. Evol.* 28, 2731-2739.

Tang, X., Szinay, D., Lang, C., Ramanna, M.S., van der Vossen, E.A., Datema, E., Klein Lankhorst, R.K., de Boer, J., Peters, S.A., Bachem, C., Stiekema, W., Visser, R.G. and de Jong, H., Bai, Y. (2008) Cross-species bacterial artificial chromosome-fluorescence *in situ* hybridization painting of the tomato and potato chromosome 6 reveals undescribed chromosomal rearrangements. *Genetics*, **180**, 1319-1328.

Tanksley, S.D., Ganal, M.W., Prince, J.P., de Vicente, M.C., Bonierbale, M.W., Broun, P., Fulton, T.M., Giovannoni, J.J., Grandillo, S., Martin, G.B. and *et al.***,** (1992) High density molecular linkage maps of the tomato and potato genomes. *Genetics*, **132**, 1141-1160.

Tanksley, S.D., Ganal, M.W., Prince, J.P., De Vicente, M.C., Bonierbale, M.W., Broun, P., Fulton, T.M., Giovannoni, J.J., Grandillo, S., Martin, G.B., Messeguer, R., Miller, J.C., Miller, L., Paterson, A.H., Pineda, O., Röder, M.S., Wing, R.A., Wu, W. and Young, N.D. (1992) High density molecular linkage maps of the tomato and potato genomes. *Genetics*, 132, 1141-1160.

The 1000 genomes project consortium. (2010) A map of human genome variation from population-scale sequencing. *Nature*, 467, 1061-1073

The Tomato Genome Consortium. (2012) The tomato genome sequence provides insights into fleshy fruit evolution. *Nature,* **485**, 635-641.

The Tomato sequencing Consortium. (2012) The tomato sequence provides insights into fleshy fruit tomato. *Nature* 485, 635-641.

Thomas, H.M., Morgan, W.G., Meredith, M.R., Humphreys, M.W., Thomas, H. and Leggett, J.M. (1994) Identification of parental and RECOMBINED chromosomes in hybrid derivatives of *Lolium multiforum* x *Festuca pratensis* by genomic *in situ* hybridization. *Theor. Appl. Genet.* **88**, 909-913.

Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*. 22, 4673-4680.

Todesco, S., Campagna, D., Levorin, F., D'Angelo, M., Schiavon, R., Valle, G. and Vezzi, A. (2008) PABS: An online platform to assist BAC-by-BAC sequencing projects. *BioTechniques*, **44**, 60-64.

Tomato Genome, C. (2012) The tomato genome sequence provides insights into fleshy fruit evolution. *Nature*, **485**, 635-641.

Touchman, J. (2010) Comparative Genomics. *Nature Education Knowledge*, **3**, 13.

Tran, N.H. and Chen, X. (2014) Comparison of next-generation sequencing samples using compression-based distances and its application to phylogenetic reconstruction. *BMC research notes*, **7**, 320.

Tuskan, G.A., Difazio, S., Jansson, S., Bohlmann, J., Grigoriev, I., Hellsten, U., Putnam, N., Ralph, S., Rombauts, S., Salamov, A., Schein, J., Sterck, L., Aerts, A., Bhalerao, R.R., Bhalerao, R.P., Blaudez, D., Boerjan, W., Brun, A., Brunner, A., Busov, V., Campbell, M., Carlson, J., Chalot, M., Chapman, J., Chen, G.L., Cooper, D., Coutinho, P.M., Couturier, J., Covert, S., Cronk, Q., Cunningham, R., Davis, J., Degroeve, S., Dejardin, A., Depamphilis, C., Detter, J., Dirks, B., Dubchak, I., Duplessis, S., Ehlting, J., Ellis, B., Gendler, K., Goodstein, D., Gribskov, M., Grimwood, J., Groover, A., Gunter, L., Hamberger, B., Heinze, B., Helariutta, Y., Henrissat, B., Holligan, D., Holt, R., Huang, W., Islam-Faridi, N., Jones, S., Jones-Rhoades, M., Jorgensen, R., Joshi, C., Kangasjarvi, J., Karlsson, J., Kelleher, C., Kirkpatrick, R., Kirst, M., Kohler, A., Kalluri, U., Larimer, F., Leebens-Mack, J., Leple, J.C., Locascio, P., Lou, Y., Lucas, S., Martin, F., Montanini, B., Napoli, C., Nelson, D.R., Nelson, C., Nieminen, K., Nilsson, O., Pereda, V., Peter, G., Philippe, R., Pilate, G., Poliakov, A., Razumovskaya, J., Richardson, P., Rinaldi, C., Ritland, K., Rouze, P., Ryaboy, D., Schmutz, J., Schrader, J., Segerman, B., Shin, H., Siddiqui, A., Sterky, F., Terry, A., Tsai, C.J., Uberbacher, E., Unneberg, P., Vahala, J., Wall, K., Wessler, S., Yang, G., Yin, T., Douglas, C., Marra, M., Sandberg, G., Van de Peer, Y. and Rokhsar, D. (2006) The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). *Science (New York, N.Y.)*, **313**, 1596-1604.

van de Wouw, M., Kik, C., van Hintum, T., van Treuren, R. and Visser, B. (2009) Genetic erosion in crops: concept, research results and challenges. *Plant Genetic Resources*, **8**, 1.

Van der Hoeven, R., Ronning, C., Giovannoni, J., Martin, G. and Tanksley, S. (2002) Deductions about the number, organization, and evolution of genes in the tomato genome based on analysis of a large expressed sequence tag collection and selective genomic sequencing. *Plant Cell* 14, 1441-56.

Van Gent, M., Bart, M.J., Van der Heide, H.G.J., Heuvelman, K.J., Kallonen, T., He, Q., Mertsola, J., Advani, A., Hallander, H.O., Janssens, K., Hermans, P.W. and Mooi, F.R. (2011) SNP-based typing: a useful tool to study *Bordetella pertussis* populations. *PLoS ONE,* 6, e20340.

Vaughan, D.A., Balazs, E. and Heslop-Harrison, J.S. (2007) From Crop Domestication to Super-domestication. *Annals of Botany*, **100**, 893-901.

Velasco, R., Zharkikh, A., Affourtit, J., Dhingra, A., Cestaro, A., Kalyanaraman, A., Fontana, P., Bhatnagar, S.K., Troggio, M., Pruss, D., Salvi, S., Pindo, M., Baldi, P., Castelletti, S., Cavaiuolo, M., Coppola, G., Costa, F., Cova, V., Dal Ri, A., Goremykin, V., Komjanc, M., Longhi, S., Magnago, P., Malacarne, G., Malnoy, M., Micheletti, D., Moretto, M., Perazzolli, M., Si-Ammour, A., Vezzulli, S., Zini, E., Eldredge, G., Fitzgerald, L.M., Gutin, N., Lanchbury, J., Macalma, T., Mitchell, J.T., Reid, J., Wardell, B., Kodira, C., Chen, Z., Desany, B., Niazi, F., Palmer, M., Koepke, T., Jiwan, D., Schaeffer, S., Krishnan, V., Wu, C., Chu, V.T., King, S.T., Vick, J., Tao, Q., Mraz, A., Stormo, A., Stormo, K., Bogden, R., Ederle, D., Stella, A., Vecchietti, A., Kater, M.M., Masiero, S., Lasserre, P., Lespinasse, Y., Allan, A.C., Bus, V., Chagne, D., Crowhurst, R.N., Gleave, A.P., Lavezzo, E., Fawcett, J.A., Proost, S., Rouze, P., Sterck, L., Toppo, S., Lazzari, B., Hellens, R.P., Durel, C.E., Gutin, A., Bumgarner, R.E., Gardiner, S.E., Skolnick, M., Egholm, M., Van de Peer, Y., Salamini, F. and Viola, R. (2010) The genome of the domesticated apple (Malus x domestica Borkh.). *Nature genetics*, **42**, 833-839.

Venter, J.C. (2001) The Sequence of the Human Genome. *Science*, **291**, 1304-1351.

Viquez-Zamora, M., Vosman, B., van de Geest, H., Bovy, A., Visser, R.G., Finkers, R. and van Heusden, A.W. (2013) Tomato breeding in the genomics era: insights from a SNP array. *BMC Genomics*, **14**, 354.

Vos, P., Hogers, R., Bleeker, M., Reijans, M., Lee, T.v.d., Hornes, M., Friters, A., Pot, J., Paleman, J., Kuiper, M. and Zabeau, M. (1995) AFLP: a new technique for DNA fingerprinting. *Nucl Acids Res*, **23**, 4407-4414.

Wade, C.M., Giulotto, E., Sigurdsson, S., Zoli, M., Gnerre, S., Imsland, F., Lear, T.L., Adelson, D.L., Bailey, E., Bellone, R.R., Blocker, H., Distl, O., Edgar, R.C., Garber, M., Leeb, T., Mauceli, E., MacLeod, J.N., Penedo, M.C., Raison, J.M., Sharpe, T., Vogel, J., Andersson, L., Antczak, D.F., Biagi, T., Binns, M.M., Chowdhary, B.P., Coleman, S.J., Della Valle, G., Fryc, S., Guerin, G., Hasegawa, T., Hill, E.W., Jurka, J., Kiialainen, A., Lindgren, G., Liu, J., Magnani, E., Mickelson, J.R., Murray, J., Nergadze, S.G., Onofrio, R., Pedroni, S., Piras, M.F., Raudsepp, T., Rocchi, M., Roed, K.H., Ryder, O.A., Searle, S., Skow, L., Swinburne, J.E., Syvanen, A.C., Tozaki, T., Valberg, S.J., Vaudin, M., White, J.R., Zody, M.C., Lander, E.S. and Lindblad-Toh, K. (2009) Genome sequence, comparative analysis, and population genetics of the domestic horse. *Science (New York, N.Y.)*, **326**, 865-867.

Wang, J. (2002) RePS: A Sequence Assembler That Masks Exact Repeats Identified from the Shotgun Data. *Genome Research*, **12**, 824-831.

Wang, J., Wurm, Y., Nipitwattanaphon, M., Riba-Grognuz, O., Huang, Y.C., Shoemaker, D. and Keller, L. (2013a) A Y-like social chromosome causes alternative colony organization in fire ants. *Nature*, **493**, 664-668.

Wang, X., Weigel, D. and Smith, L.M. (2013b) Transposon variants and their effects on gene expression in *Arabidopsis*. *PLoS Genetics*, **9**, e1003255.

Wang, Y. (2005) Euchromatin and Pericentromeric Heterochromatin: Comparative Composition in the Tomato Genome. *Genetics*, **172**, 2529-2540.

Wang, Y., van der Hoeven, R.S., Nielsen, R., Mueller, L.A. and Tanksley, S.D. (2005) Characteristics of the tomato nuclear genome as determined by sequencing undermethylated EcoRI digested fragments. *Theor Appl Genet*, **112**, 72-84.

Warren, R.L., Sutton, G.G., Jones, S.J.M. and Holt, R.A. (2006) Assembling millions of short DNA sequences using SSAKE. *Bioinformatics*, **23**, 500-501.

Weigel, D. and Mott, R. (2009) The 1001 genomes project for *Arabidopsis thaliana*. *Genome Biol*. **10**, 107.

Werren, J.H., Richards, S., Desjardins, C.A., Niehuis, O., Gadau, J., Colbourne, J.K., Werren, J.H., Richards, S., Desjardins, C.A., Niehuis, O., Gadau, J., Colbourne, J.K., Beukeboom, L.W., Desplan, C., Elsik, C.G., Grimmelikhuijzen, C.J., Kitts, P., Lynch, J.A., Murphy, T., Oliveira, D.C., Smith, C.D., van de Zande, L., Worley, K.C., Zdobnov, E.M., Aerts, M., Albert, S., Anaya, V.H., Anzola, J.M., Barchuk, A.R., Behura, S.K., Bera, A.N., Berenbaum, M.R., Bertossa, R.C., Bitondi, M.M., Bordenstein, S.R., Bork, P., Bornberg-Bauer, E., Brunain, M., Cazzamali, G., Chaboub, L., Chacko, J., Chavez, D., Childers, C.P., Choi, J.H., Clark, M.E., Claudianos, C., Clinton, R.A., Cree, A.G., Cristino, A.S., Dang, P.M., Darby, A.C., de Graaf, D.C., Devreese, B., Dinh, H.H., Edwards, R., Elango, N., Elhaik, E., Ermolaeva, O., Evans, J.D., Foret, S., Fowler, G.R., Gerlach, D., Gibson, J.D., Gilbert, D.G., Graur, D., Grunder, S., Hagen, D.E., Han, Y., Hauser, F., Hultmark, D., Hunter, H.C.t., Hurst, G.D., Jhangian, S.N., Jiang, H., Johnson, R.M., Jones, A.K., Junier, T., Kadowaki, T., Kamping, A., Kapustin, Y., Kechavarzi, B., Kim, J., Kim, J., Kiryutin, B., Koevoets, T., Kovar, C.L., Kriventseva, E.V., Kucharski, R., Lee, H., Lee, S.L., Lees, K., Lewis, L.R., Loehlin, D.W., Logsdon, J.M., Jr., Lopez, J.A., Lozado, R.J., Maglott, D., Maleszka, R., Mayampurath, A., Mazur, D.J., McClure, M.A., Moore, A.D., Morgan, M.B., Muller, J., Munoz-Torres, M.C., Muzny, D.M., Nazareth, L.V., Neupert, S., Nguyen, N.B., Nunes, F.M., Oakeshott, J.G., Okwuonu, G.O., Pannebakker, B.A., Pejaver, V.R., Peng, Z., Pratt, S.C., Predel, R., Pu, L.L., Ranson, H., Raychoudhury, R., Rechtsteiner, A., Reese, J.T., Reid, J.G., Riddle, M., Robertson, H.M., Romero-Severson, J., Rosenberg, M., Sackton, T.B., Sattelle, D.B., Schluns, H., Schmitt, T., Schneider, M., Schuler, A., Schurko, A.M., Shuker, D.M., Simoes, Z.L., Sinha, S., Smith, Z., Solovyev, V., Souvorov, A., Springauf, A., Stafflinger, E., Stage, D.E., Stanke, M., Tanaka, Y., Telschow, A., Trent, C., Vattathil, S., Verhulst, E.C., Viljakainen, L., Wanner, K.W., Waterhouse, R.M., Whitfield, J.B., Wilkes, T.E., Williamson, M., Willis, J.H., Wolschin, F., Wyder, S., Yamada, T., Yi, S.V., Zecher, C.N., Zhang, L. and Gibbs, R.A. (2010) Functional and evolutionary insights from the genomes of three parasitoid Nasonia species. *Science (New York, N.Y.)*, **327**, 343-348.

Westesson, O., Skinner, M. and Holmes, I. (2013) Visualizing next-generation sequencing data with JBrowse. *Brief. Bioinf.* **14**, 172–177.

Wetterstrand, K.A. (2013) DNA sequencing costs: data from the NHGRI Genome Sequencing Program (GSP). In *National Human Genome Research Institute*.

Wijnker, E., Velikkakam James G., Ding, J., Becker, F., Klasen, J.R., Rawat, V., Rowan, B.A., de Jong, D.F., de Snoo, C.B., Zapata, L., Huettel, B., de Jong, H., Ossowski, S., Weigel, D., Koornneef, M., Keurentjes, J.J. and Schneeberger, K. (2013) The genomic landscape of meiotic crossovers and gene conversions in *Arabidopsis thaliana*. *eLife*, e01426.

Williams, J.G.K., Kubelik, A.R., Livak, K.J., Rafalski, J.A. and Tingey, S.V. (1990) DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. *Nucl Acids Res*, **18**, 6531-6535.

Wood, D.E. and Salzberg, S.L. (2014) Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome biology*, **15**, R46.

Wood, V., Gwilliam, R., Rajandream, M.A., Lyne, M., Lyne, R., Stewart, A., Sgouros, J., Peat, N., Hayles, J., Baker, S., Basham, D., Bowman, S., Brooks, K., Brown, D., Brown, S., Chillingworth, T., Churcher, C., Collins, M., Connor, R., Cronin, A., Davis, P., Feltwell, T., Fraser, A., Gentles, S., Goble, A., Hamlin, N., Harris, D., Hidalgo, J., Hodgson, G., Holroyd, S., Hornsby, T., Howarth, S., Huckle, E.J., Hunt, S., Jagels, K., James, K., Jones, L., Jones, M., Leather, S., McDonald, S., McLean, J., Mooney, P., Moule, S., Mungall, K., Murphy, L., Niblett, D., Odell, C., Oliver, K., O'Neil, S., Pearson, D., Quail, M.A., Rabbinowitsch, E., Rutherford, K., Rutter, S., Saunders, D., Seeger, K., Sharp, S., Skelton, J., Simmonds, M., Squares, R., Squares, S., Stevens, K., Taylor, K., Taylor, R.G., Tivey, A., Walsh, S., Warren, T., Whitehead, S., Woodward, J., Volckaert, G., Aert, R., Robben, J., Grymonprez, B., Weltjens, I., Vanstreels, E., Rieger, M., Schafer, M., Muller-Auer, S., Gabel, C., Fuchs, M., Dusterhoft, A., Fritzc, C., Holzer, E., Moestl, D., Hilbert, H., Borzym, K., Langer, I., Beck, A., Lehrach, H., Reinhardt, R., Pohl, T.M., Eger, P., Zimmermann, W., Wedler, H., Wambutt, R., Purnelle, B., Goffeau, A., Cadieu, E., Dreano, S., Gloux, S., Lelaure, V., Mottier, S., Galibert, F., Aves, S.J., Xiang, Z., Hunt, C., Moore, K., Hurst, S.M., Lucas, M., Rochet, M., Gaillardin, C., Tallada, V.A., Garzon, A., Thode, G., Daga, R.R., Cruzado, L., Jimenez, J., Sanchez, M., del Rey, F., Benito, J., Dominguez, A., Revuelta, J.L., Moreno, S., Armstrong, J., Forsburg, S.L., Cerutti, L., Lowe, T., McCombie, W.R., Paulsen, I., Potashkin, J., Shpakovski, G.V., Ussery, D., Barrell, B.G. and Nurse, P. (2002) The genome sequence of Schizosaccharomyces pombe. *Nature*, **415**, 871-880.

Xie W., Chen Y., Zhou G., Wang L., Zhang C., Zhang J., Xiao J., Zhu T., Zhang Q. (2009) Single feature polymorphisms between two rice cultivars detected using a median polish method. *Theor Appl Genet*. 2009 Jun;119(1):151-64

Xu, J. and Earle, E.D. (1996) High resolution physical mapping of 45S (5.8S, 18S and 25S) rDNA gene loci in the tomato genome using a combination of karyotyping and FISH of pachytene chromosomes. *Chromosoma*, **104**, 545-550.

Xu, Q., Chen, L.L., Ruan, X., Chen, D., Zhu, A., Chen, C., Bertrand, D., Jiao, W.B., Hao, B.H., Lyon, M.P., Chen, J., Gao, S., Xing, F., Lan, H., Chang, J.W., Ge, X., Lei, Y., Hu, Q., Miao, Y., Wang, L., Xiao, S., Biswas, M.K., Zeng, W., Guo, F., Cao, H., Yang, X., Xu, X.W., Cheng, Y.J., Xu, J., Liu, J.H., Luo, O.J., Tang, Z., Guo, W.W., Kuang, H., Zhang, H.Y., Roose, M.L., Nagarajan, N., Deng, X.X. and Ruan, Y. (2013) The draft genome of sweet orange (Citrus sinensis). *Nature genetics*, **45**, 59-66.

Xu, X., Pan, S., Cheng, S., Zhang, B., Mu, D., Ni, P., Zhang, G., Yang, S., Li, R., Wang, J., Orjeda, G., Guzman, F., Torres, M., Lozano, R., Ponce, O., Martinez, D., De la Cruz, G., Chakrabarti, S.K., Patil, V.U., Skryabin, K.G., Kuznetsov, B.B., Ravin, N.V., Kolganova, T.V., Beletsky, A.V., Mardanov, A.V., Di Genova, A., Bolser, D.M., Martin, D.M., Li, G., Yang, Y., Kuang, H., Hu, Q., Xiong, X., Bishop, G.J., Sagredo, B., Mejia, N., Zagorski, W., Gromadka, R., Gawor, J., Szczesny, P., Huang, S., Zhang, Z., Liang, C., He, J., Li, Y., He, Y., Xu, J., Zhang, Y., Xie, B., Du, Y., Qu, D., Bonierbale, M., Ghislain, M., Herrera Mdel, R., Giuliano, G., Pietrella, M., Perrotta, G., Facella, P., O'Brien, K., Feingold, S.E., Barreiro, L.E., Massa, G.A., Diambra, L., Whitty, B.R., Vaillancourt, B., Lin, H., Massa, A.N., Geoffroy, M., Lundback, S., DellaPenna, D., Buell, C.R., Sharma, S.K., Marshall, D.F., Waugh, R., Bryan, G.J., Destefanis, M., Nagy, I., Milbourne, D., Thomson, S.J., Fiers, M., Jacobs, J.M., Nielsen, K.L., Sonderkaer, M., Iovene, M., Torres, G.A., Jiang, J., Veilleux, R.E., Bachem, C.W., de Boer, J., Borm, T., Kloosterman, B., van Eck, H., Datema, E., Hekkert, B., Goverse, A., van Ham, R.C. and Visser, R.G. (2011) Genome sequence and analysis of the tuber crop potato. *Nature*, **475**, 189-195.

Yamamoto, T., Nagasaki, H., Yonemaru, J., Ebana, K., Nakajima, M., Shibaya, T. and Yano, M. (2010) Fine definition of the pedigree haplotypes of closely related rice cultivars by means of genome-wide discovery of single-nucleotide polymorphisms. *BMC genomics*, **11**, 267.

Yang, B., Peng, Y., Leung, H.C., Yiu, S.M., Chen, J.C. and Chin, F.Y. (2010) Unsupervised binning of environmental genomic fragments based on an error robust selection of l-mers. *BMC bioinformatics*, **11 Suppl 2**, 6.

Ye, J., Coulouris, G., Zaretskaya, I., Cutcutache, I., Rozen, S. and Madden, T.L. (2012) Primer-BLAST: A tool to design target-specific primers for polymerase chain reaction. *BMC Bioinformatics*, **13**, 134.

Yi, H. and Jin, L. (2013) Co-Phylog: an assembly-free phylogenomic approach for closely related organisms. *Nucleic acids research*, **41**, e75.

Zamir, D. (2001) Improving plant breeding with exotic genetic libraries. *Nature Rev. Gen.* **2**, 983-989.

Zhan, X., Pan, S., Wang, J., Dixon, A., He, J., Muller, M.G., Ni, P., Hu, L., Liu, Y., Hou, H., Chen, Y., Xia, J., Luo, Q., Xu, P., Chen, Y., Liao, S., Cao, C., Gao, S., Wang, Z., Yue, Z., Li, G., Yin, Y., Fox, N.C., Wang, J. and Bruford, M.W. (2013) Peregrine and saker falcon genome sequences provide insights into evolution of a predatory lifestyle. *Nature genetics*, **45**, 563-566.

Zhang, Y., Wiggins, B.E., Lawrence, C., Petrick, J., Ivashuta, S. and Heck, G. (2012) Analysis of plant-derived miRNAs in animal small RNA datasets. *BMC genomics*, **13**, 381.

Zhong, X.-B., de Jong, J.H. and Zabel, P. (1996) Preparation of tomato meiotic pachytene and mitotic metaphase chromosomes suitable for fluorescence*in situ* hybridization (FISH). *Chromosome Research*, **4**, 24-28.

Zhong, X.b., Lizardi, P.M., Huang, X.h., Bray-Ward, P.L. and Ward, D.C. (2001) Visualization of oligonucleotide probes and point mutations in interphase nuclei and DNA fibers using rolling circle DNA amplification. *Proceedings of the National Academy of Sciences*, **98**, 3940-3945.

Zimin, A.V., Delcher, A.L., Florea, L., Kelley, D.R., Schatz, M.C., Puiu, D., Hanrahan, F., Pertea, G., Van Tassell, C.P., Sonstegard, T.S., Marcais, G., Roberts, M., Subramanian, P., Yorke, J.A. and Salzberg, S.L. (2009) A whole-genome assembly of the domestic cow, Bos taurus. *Genome biology*, **10**, R42.

Zygier, S., Chaim, A.B., Efrati, A., Kaluzky, G., Borovsky, Y. and Paran, I. (2005) QTLs mapping for fruit size and shape in chromosomes 2 and 4 in pepper and a comparison of the pepper QTL map with that of tomato. *Theor Appl Genet*, **111**, 437-445.

Summaries

English

Knowledge of natural variation of crops and wild varieties is of pivotal importance for plant genetics and breeding. Due to the high costs associated with genomics, this data could, until recently, only be obtained in small numbers on a case-by-case basis. With the advent of NGS and the rapid decrease in sequencing costs in recent years, it is now possible to sample a large number of varieties and have their genomes sequenced. The emerging sequencing technologies and increased data production power poses both new genome bioinformatics possibilities and challenges. Due to the need of high sequencing coverage in NGS experiments, the volume of data created can be overwhelming to systems not designed for big data analysis. Also, some methodologies applied for the analysis of a few samples are not able to scale up with the increased volume of data generated by NGS.

Tomato is a commercially important crop and it is a model species for fleshy fruit plants and the important nightshade clade. With the availability of the tomato and potato genomes, breeders became aware of the need to expand their skills, bioinformatics tools, and understanding of the genetic variation in their collection of cultivars, heirlooms, landraces and subspecies. In my comparative genomics study (Chapter 2) I described the results of such an inventory by mapping 81 shallow sequenced tomato genomes and assembling *de novo* the genomes of 3 deep sequenced wild tomato species, namely, *Solanum arcanum*, *S. habrochaites* and *S. pennellii*.

After sequencing and mapping the 84 Solanum genomes to the golden reference *Solanum lycopersicum* cv Heinz 1706, we needed a tool to analyse such dataset for introgressions. Although there are tools available that can visualise SNPs or find introgressions, there was no software available to efficiently analyse such a large number of samples simultaneously. To that end I developed the Introgression Browser, a tool that allows for the first time to analyse a large number of samples in parallel for signs of introgression. With this bioinformatics tool I was able to identify and visualize an introgression from *S. pimpinellifolium* into three crop tomatoes as well as SNP poor regions in *Arabidopsis thaliana* that is indicative of a known inversion in chromosome 4. This tool enables insight in the consequences of introgressive hybridization breeding using NGS data, without being dependent on time consuming and less accurate marker dependent analysis methods. As such, the Introgression Browser shows how bioinformatics can assist in making breeding more targeted.

During the analysis of the data created in Chapter 2, it became clear that some samples were mislabelled. This prompted me to create a program, called Cnidaria (Chapter 4), which is able to identify species based on raw NGS data. Although this tool was only fully developed by the end of my thesis it could have been implemented in the large scale comparative genomics study previously addressed. For now I demonstrated how it is possible to correct the label of several samples described in the comparative genomics

study as well as classifying 164 samples ranging from fungi and plants to humans as a proof of concept.

Despite all the information made available by genome mapping, *de novo* assembly and introgression analysis, one of the arguably more important question about chromosome structure and assembly correctness is frequently overseen in current genomics projects. Originally, genome sequencing projects would create Bacterial/Yeast Artificial Chromosomes in order to link sequenced data to their chromosome molecules. This practice is prohibitively expensive and new methods such as Optical Mapping, Genome Mapping and PCR Fluorescence *in situ* Hybridization (FISH) have been developed to replace B/YAC libraries. For PCR-FISH, it is imperative to generate sets of unique primer pairs which produce amplicons not containing repetitive sequences. The third application of bioinformatics in my thesis (Chapter 5) involves the improvement of the existing Mathematically Defined Repeats (MDR) method to aid the chromosomal visualization of selected DNA sequences from the assembled tomato genome. The reason for this approach is the need of positioning contigs of fragmented genomes that could not be properly assembled even after expensive and low throughput post-sequencing manual correction and scaffolding. MDR have been used in the past for this purpose and we expanded this method with extra verifications of uniqueness of the amplicons in order to create short, unique probes which amplify directly from genomic DNA without the need for B/YAC libraries.

In this thesis I have created a large dataset of 82 resequenced varieties from 14 Solanum species, 3 new *de novo* Solanum genomes and 3 new programs for assisting in new genomic initiatives. From introgression analysis, species identification and assembly closure, the body of work presented in this thesis should serve the genomics research community, and especially those involved in plant breeding, to improve their workflow and reduce analysis time.

Keywords: comparative genomics; genomics; NGS; Whole Genome Sequencing; FISH; tomato

Dutch

Kennis ten aan zien van de (genetische) diversiteit van voedsel gewassen en wilde verwante soorten op basis van genoom sequentie informatie is van belang voor de plantengenetica en plantenveredeling. Mede door hoge kosten was de productie van sequentie data tot voor kort beperkt. Door de ontwikkeling van 'Next Generation Sequencing' technologie is daarin verandering gekomen. We zijn steeds beter in staat om de genoom sequentie van soorten te ontsluiten. De enorme berg aan sequentie informatie opent nieuwe mogelijkheden voor de genoom bioinformatica, maar is tegelijkertijd een uitdaging als het gaat om de verwerking, interpretatie en visualisatie ervan. Om aan de groeiende berg informatie het hoofd te kunnen blijven bieden moeten er nieuwe methoden worden ontwikkeld.

In dit proefschrift beschrijf ik het onderzoek aan tomaat dat samen met aardappel, economisch gezien, de belangrijkste vertegenwoordiger is van de nachtschade familie. Het beschikbaar komen van de tomaat en aardappel genoom sequentie was een belangrijke mijlpaal voor de vergelijkede genomica en luidde in veel opzichten de start in voor de ontwikkeling van nieuwe tools en inzichten. De veredeling en de genoom bioinformatica zag zich niet alleen geconfronteerd met een enorme berg aan data, maar was ook genoodzaakt om nieuwe gereedschappen te ontwikkelen om de genetische diversiteit in kaart te kunnen brengen. In hoofdstuk 2 beschrijf ik de reconstructie van 3 referentie genomen, te weten *Solanum arcanum*, *S. habrochaites* en *S. pennellii* en een vergelijkende genoom studie van 81 verschillende tomaat accessies. Na het sequeneren en het karteren van 84 Solanum accessies tegen het referentie genoom van *S. lycopersicum* cv Heinz 1706, was het doel om introgressies in de accessies te analyseren. Hoewel er software beschikbaar was om introgressies op basis van SNPs te detecteren, waren deze tools niet geschikt om een groot aantal genomen tegelijkertijd te screenen. Deze functionaliteit is met de constructie van IBrowser nu beschikbaar gekomen. De IBrowser applicatie maakt het mogelijk om op basis van SNP patronen introgressies te detecteren en die te visualiseren met behulp van een reguliere web browser functionaliteit. Introgressies afkomstig van *S. pimpinellifolium* in drie verschillende tomaat accessies, alsook een inversie in de korte arm van chromosoom 4 in Arabidopsis, konden met behulp van IBrowser worden gedetecteerd. Met IBrowser is de detectie van introgressies accurater en minder afhankelijk geworden van tijdrovende genetische merker analyse methoden. IBrowser is daarmee een waardevol toepassing voor de intogressieve hybridisatie veredeling geworden.

Tijdens de data analyse werd de foutieve identiteit van een aantal accessies in het 150 tomaat genoom project aangetoond. Dit was de start voor de constructie van een nieuwe tool, Cnidaria, dat op basis van ruwe NGS sequenties soorten alsook accessies kan identificeren, hetgeen ik beschrijf is hoofdstuk 4 van dit proefschrift. In dit hoofdstuk beschrijf ik de functionaliteit van Cnidaria en demonstreer ik de classificatie van

164 soorten variërend van schimmels, planten, tot aan mens toe, alsook de gecorrigeerde identiteit van een aantal foutief gemerkte samples.

Ondanks alle kennis die de de novo assemblage, genoom kartering en introgressie analyse heeft opgeleverd, wordt er slechts relatief weinig aandacht besteed aan de verificatie van genoom assemblages en chromosoom structuur. Oorspronkelijk werd in genoom projecten gebruik gemaakt van het sequeneren van BAC of YAC kloons die werden verankerd aan een genetische kaart. Na het karteren van de BAC kloons werd op basis van overlap en insertie lengte criteria de meest optimale set aan BAC/YAC klonen geselecteerd, die vervolgens werden gesequeneerd om de genoom sequentie te kunnen reconstrueren. Deze methode is relatief duur en tegenwoordig zijn nieuwe methoden beschikbaar om sequenties te karteren, zoals Optisch Karteren (Optical Mapping) en PCR FISH. Bij PCR FISH is het noodzakelijk om van unieke primer paren gebruik te kunnen maken. Daarmee kunnen gelabelde amplicons worden geproduceerd die vrij zijn van repetitieve sequenties. Met behulp van FISH leveren deze amplicons een uniek signaal en een unieke positie op het chromosoom op dat gebruikt kan worden om de juiste volgorde van genoom sequenties te ontrafelen. De derde applicatie beschrijf ik in hoofdstuk 5 van mijn proefschrift en betreft een verbeterde methode om repetitieve sequenties te detecteren. De verbeterde MDR (Mathematically Defined Repeats) methode is erop gericht repetitieve sequenties te detecteren in een geassembleerde genoom, zoals bijvoorbeeld het tomaat referentie genoom. Voorheen was het verankeren van contigs die repetitieve sequenties bevatten, zelfs na opnieuw sequeneren en handmatige filtering van complexe sequenties, problematisch gebleken. Met behulp van de verbeterde MDR methode kan met grotere zekerheid de uniciteit van amplicon sequenties worden voorspeld uit een geassembleerde genoom sequentie. Uiteindelijk leverde dit verbeterde FISH merkers op, die direct van het genomisch DNA geamplificeerd kunnen worden zonder BAC/YAC klonen.

In dit proefschrift beschrijf ik een sequentie data set van 82 tomaat accessies verdeeld over 14 Solanum soorten, waarvan er 3 de novo zijn gesequeneerd alsmede 3 nieuwe applicaties ter ondersteuning van genoom assemblage, introgressie analyse en soort identificatie. Dit resultaat komt ten goede aan een verbeterde verwerking, interpretatie en visualisatie van genoom sequenties en de karakteristieke eigenschappen ervan en dient zowel de genomica alsook het veredeling onderzoek.

Acknowledgements

First, I would like to acknowledge the importance of the first people I met -- my nuclear family, Valnê, Déa, Héber, and Sílvia. The certainty of your support gives me the drive to try anything. My extended family: uncles, aunts, and cousins, particularly Elson, Neisa, Samir, Thiago, and Vandilson -- besides being family, you are great friends. With Aline Pimentel and Roseli Paraguassu, we have a connection that defies the difficulties of distance, and every time we meet it is as if we talked yesterday.

Now, officially finishing my studying years, I would like to thank my middle and high school friends: Cleidelene Azevedo, Anderson Araújo, Andréa Negrão, Ivia Natália, and Jaqueline Almeida. You made the school experience a pleasure because you were there.

After deciding to pursue an academic career in biology, I'm so glad I've made friends for life in my bachelor at UFBA: Alexandre Ramos, Bruno Cosme, Cris Brito, Igor Cruz, Jilmária Oliveira, Lia Meyer, Maria Betânia Figueiredo Silva, Rogério Lima, and Wendell Vilas Boas. You are not only friends but also advisors and confidants. To Thales Francisco, Juliana Munduruca, and their beautiful twins (Rodrigo and Nina), you are family to me.

I also made fantastic friends during my bachelor's internship at UCSAL: André Luiz Fagundes, Cimille Antunes, Claudineéia Pelacani, Fábio & Graça Ferreira, Juan Carlos Rossi, Rafael Simões, Sidnei Cerqueira Santos, and Wilson Matos. Our time together is unforgettable; you all have a place in my heart. In the same internship I've had great supervisors and co-supervisors: Luzimar Gonzaga Fernandez and Renato Delmondez. Nobody outside my family had such an impact in the unfolding of my life, and I'm grateful for having you in my life. Marta Bruno Loureiro, you joined the group after I left, and I've met you in Wageningen in a work mission with Renata Mann; it was a short stay, but I consider you both great friends.

Before knowing I was accepted to Wageningen, I started a master in Fortaleza, Brazil, under the supervision of Benildo Cavada, a great scientist, where I made friends who welcomed and helped me with endless generosity: Andreia Nikokavouras, Gustavo Arruda, Raquel Benevides, Taianá Maia, Victor Carneiro, and Kelly Coelho. You received me without reservation when I had nobody to lean on and became like a family to me.

When first arriving in Wageningen, thanks to the personal help of Monique Montenarie, the introduction week gave me pleasure and the honour of meeting some amazing people whom I'm glad to call my brothers and sisters: Ana Maria Blindu, Dilek Sagram, Ferdie José, Isabel Verhulst, Milkha Margaretta, and Reiko Kiwamoto & Tjibbe Wubbels. In a difficult time adapting to a new country, language, and temperatures, you gave me warmth, friendship, and weekly dinners where we shared experiences and laughs. Che-Yang Liao, the second generation, my partner in archery and beer, the gentleman.

In a new land with new experiences and new people on every corner, the feeling of being at home was always certain when meeting my fellow Brazilians. To the families

Aans (Wiebe, Anabele, Mateus, Eduardo, and Hidde), Santos (Cinara, Larissa and Jânio), and Viana (Vanja, Thaynã, and Leonor): I can't put into words the important role you play in my life. Knowing you are here for me has given me the confidence to want to stay, knowing I'm not alone in this land. Felipe & Magdalena van der Struijk, Isabela Dutra, Michael Daamen & Itziar Sevilla & Iara, Tamara Borges and Ulisses Rocha, my partners in crime -- too many stories to tell which are better left untold because "what happens in Wageningen, stays in Wageningen". And the many, many, many Brazilians who came here for either short or long periods of time and left both great memories and great holes when they left but, most of all, left a piece of themselves in my life: Beatriz Oliveira, Carol Mosca, Charles Moreira, Deborah & Joachim & Olivier Bargsten, Dorys dos Santos, Flávia Saia, Haissa Cardarelli, Isabela Nougalli, Julio Maia, Larissa Barreto, Luciana Soler, Marcelo Maia, Melina Macatelli, Raquel Mendonça, Régis Corrêa & Lucia Yanez, Rita & Celso & Pedro von Randow, Sandra Crispim, Tatiana Nicz and several more. The world is round and small, so we will meet again soon; but due to something called the internet, we can, every once in a while, catch up. Carol & Dennis & Bento Souza da Silva-Bijl, Guilherme & Gosia & Sebastian van der Struijk, Lara Vita, Maria Cecília Dias da Costa, and Mauricio Dimitrov & Fernanda Paganelli, I'm glad you are still here with me. With Claire Kamei & Edouard Severing, it took a bit of work of my part to make you two take the first step towards each other, but I'm glad for my meddling.

During my master's work in Wageningen in 2006, I got a great reception from my supervisors and colleagues from Plant Physiology: Henk Hilhorst, Ronny Joosen, and Wilco Ligterink. It was great working with you, and I'm proud to have worked with you. My colleagues in the bioinformatics master, which still meet every Thursday for drinks -- Benoit Carrères, Erik Ittman, Erik Roijen, Harm-Jan Westra, Heleen de Weerd, Paul Boekschoten, and Pierre-Yves Chibon -- it is great to be surrounded by like-minded and like-nerded people.

Believe it or not, nerds also have a social life; and for my friends from salsa and the international club, cheers: Jose Lozano, Katharina Zellmer, Marjorie van Strien, Pierce Cicilia, Jr., Wouter Lee, Abdul Drame, Florene Slotboom, Harmen Boerboom, Kitty Cruden, Nelleke van Schoonhoven, and Waeil Haider. Marie del Marmol, Laurence Duvivier & Rinaldo Kembel, my dear friends, I hope I will find more time to visit you, I miss you dearly and I cherish all of our many memories.

In the two years I've spent in Utrecht working at CBS I have also met great colleagues: Anna Kolecka, Bart Theelen, Ferry Hagen, Kantarawee Khayhan, Karolina Dukik, Neriman & Cobus Visage, Rolf Boesten, Teun Boekhout, and Ulrike Damm. It was a great time; I cherish our time together. Utrecht University fortunately sponsored weekly drinks for foreign students. Fortunately because, due to that, I met wonderful people, directly and indirectly, which I can't imagine that city without: Afrouz & Dewi le Bars, Ana Chies, Arjen & Ester Den Admirant, Arnaud Bataille, Daniel Schiavini & Karolina Vos, Diana Papazova, Edwin Lima, Eimear Murphy, Elínborg Kristjánsdóttir, Elke Roovers,

Emanuelle Santos, Erick Plauschinn, Guðrún Stefánsdóttir & Vinicius do Ó, Igor Khavkine, Joana Caldas, Louise Kindt, Margarida Avo, Maria Augusta Sartori, Marie-Claire Genin, Priscila Rosetto, Rosja Mastop, Scott McDonald, Tony Booth, Vânia Vicente (and family), and many others.

A PhD is a collective work. In this spirit I would like to recognize the important help from my colleagues: Colleagues from genetics -- Dóra Szinay, José van de Belt, and Paola Gaiero. Colleagues from PRI -- Aalt-Jan van Dijk, Bas te Lintel Hekkert, Elio Schijlen, Erwin Datema, Felipe Leal Valentin, Henri van de Geest, Jan Peter Nap, Jan van Haarst, Jose Muino, Linda Baker, Paul Mooijman, Pieter Lukasse, Sevgin Demirci, and Sven Warris. Colleagues from the chair group bioinformatics -- Harm Nijveen, Ke Lin, Luca Santuari, and Sandra Smit. Also, colleagues from diverse groups and management -- Ana Marcela Viquez, Andries Koops, Eric Schanz, Hana Nobels, Hendrik-Jan Megens, Ingrid van der Meer, Monique Montenaire, Paul Franz, Richard Finkers, and Robert Hall. Colleagues from NIPGR in India -- Debasis Chattopadhyay, Kamlesh & Asha Sahu, and Sabiha Parween. Special thanks to the cluster leader from applied bioinformatics who believed in me and hired me, Roeland van Ham, and the cluster leader who replaced him and did all the work of managing the group and my project, Gabino Sanchez-Perez. Thank you both.

Thank you very much, Sander Peters and Hans de Jong. When I was left without supervisors because of the passing away of Prof Jack Leunissen, the two of you took me under your guidance and made this PhD possible, always reminding me that coding is great but the scientific question must come first. And Dick de Ridder, who took over the chair group bioinformatics, inheriting me and my project in the last stretch, which turned out to be the bulk of the work -- thank you for your patience and guidance.

From all the friends I've made, I have to thank my best friend, my partner, my wife, Suzanne Hoogstrate for supporting me throughout this PhD. Some say that if you can stand your partner during his PhD, you will have a merry life. I think we are set for life then. I would also like to thank the Hoogstrate family: Ad, Koos, Johnny, Helma & Rob & Lucas van Woerkom. Thank you for having me as part of your family and for laughing of my unfunny, sometimes tactless, jokes. Thank you to "the girls": Marijke Jansen, Marjon Hakkeling, Nanda Spanjer and Rina de Zwaan. I know how important Suzanne is for you, and I thank you for lending her to me.

I would also like to mention Jan & Femmy Peelen -- thank you for making my childhood dream of coming to Europe a reality.

Curriculum Vitae

Saulo Alves Aflitos was born on the 20th of May 1982 in Porto Trombetas, Pará, Brazil. In 1999 he obtained his high school degree from the Brazilian Baptist College in Salvador, Bahia, Brazil. In the year 2000 he started his bachelor in Biology with emphasis on ecology at the Federal University of Bahia in Salvador, Bahia, Brazil. He was awarded his degree in January 2006 with a thesis entitled "Computational approach for the identification of lectin domain coding genes in the genome of *Streptococcus mutans*." During his bachelor he conducted a four-year internship in biostatistics and biochemistry at the Catholic University of Salvador under the supervision of Prof. Dr Luzimar Gonzaga Fernandez and Prof. Dr Renato Delmondez, working with heavy metal contamination in estuarine regions and the genome sequencing of the phytopathogen *Moniliophthora perniciosa* (the causal agent of the "Witches' Broom Disease" of the cocoa tree, *Theobroma cacao*). For one semester he held the position of substitute teacher in biochemistry at the Federal University of Bahia. He left this position in March 2006 to start his, incomplete, master in biochemistry at the Federal University of Ceará, Fortaleza, Brazil, under supervision of Prof. Dr. Benildo Cavada. In September 2006, he started a master program in bioinformatics at Wageningen University, Wageningen, The Netherlands, defending in 2008 his thesis entitled "Integrated promoter extraction by sequence information content from several software tools: a case study with *Arabidopsis thaliana* seed dormancy and germination gene sets" in the laboratory of plant physiology under the supervision of Prof. Dr Henk Hilhorst. Between 2008 and 2010 he worked as a bioinformatics technician at the Centraal Bureau voor Schimmelcultures of the Royal Netherlands Academy of Arts and Sciences, Utrecht, The Netherlands, where he developed projects both in the genome assembly of the human pathogen *Cryptococcus neoformans* and the design of diagnosis probes for the human pathogen *Candida albicans* under the guidance of Dr Teun Boekhout. In 2011 he started his PhD in Wageningen University, placed at the cluster bioinformatics of Plant Research International under the supervision of Prof. Dr Jack Leunissen and Prof. Dr Roeland van Ham, and later under the supervision of Prof. Dr Hans de Jong, Prof. Dr Dick de Ridder, Dr Sander Peters, and Dr Gabino Sanchez-Perez. Originally working on the analysis of an Inbred Line (IL) population on the "mechanisms of introgression breeding in tomato", his project was later expanded to include the analysis of 150 tomato genomes in a large scale sequencing project, cultimating in the defense of his thesis entitled "High-throughput comparative genomics for plant breeding and its application in the tomato clade" in December 2015.

List of Publications

Aflitos, Saulo, Elio Schijlen, Hans Jong, Dick Ridder, Sandra Smit, Richard Finkers, Jun Wang *et al.*, "**Exploring genetic variation in the tomato (*Solanum* section *Lycopersicon*) clade by whole-genome sequencing**." The Plant Journal 80, no. 1 (2014): 136-148. DOI: 10.1111/tpj.12616

Anthony Bolger, Federico Scossa, Marie E Bolger, Christa Lanz, Florian Maumus, Takayuki Tohge, Hadi Quesneville, Saleh Alseekh, Iben Sørensen, Gabriel Lichtenstein, Eric A Fich, Mariana Conte, Heike Keller, Korbinian Schneeberger, Rainer Schwacke, Itai Ofner, Julia Vrebalov, Yimin Xu, Sonia Osorio, **Saulo Alves Aflitos**, Elio Schijlen, José M Jiménez-Goméz, Malgorzata Ryngajllo, Seisuke Kimura, Ravi Kumar, Daniel Koenig, Lauren R Headland, Julin N Maloof, Neelima Sinha, Roeland CHJ van Ham, René Klein Lankhorst, Linyong Mao, Alexander Vogel, Borjana Arsova, Ralph Panstruga, Zhangjun Fei, Jocelyn KC Rose, Dani Zamir, Fernando Carrari, James J Giovannoni, Detlef Weigel, Björn Usadel, Alisdair R Fernie. "**The genome of the stress-tolerant wild tomato species *Solanum pennellii*.**" Nature genetics (2014). DOI: 10.1038/ng.3046

Aflitos, Saulo Alves, Gabino Sanchez-Perez, Dick Ridder, Paul Fransz, Michael E. Schranz, Hans Jong, Sander A. Peters. "**Introgression browser: high-throughput whole-genome SNP visualization**." The Plant Journal 82, no. 1 (2015): 174-182. DOI: 10.1111/tpj.12800

Aflitos, Saulo Alves, Edouard Severing, Gabino Sanchez-Perez, Sander Peters, Hans de Jong, Dick de Ridder. "**Cnidaria fast, reference-free clustering of raw and assembled genome and transcriptome NGS data**", accepted for publication by BMC Bioinformatics journal

Education Statement of the Graduate School

Experimental Plant Sciences

The Graduate School EXPERIMENTAL PLANT SCIENCES

Issued to: Saulo Alves Aflitos
Date: 16 December 2015
Group: PRI - Bioscience / Bioinformatics
University: Wageningen University & Research Centre

1) Start-up phase	date
► **First presentation of your project**	
Introduction 'Sequence QC'	Apr 10, 2011
► **Writing or rewriting a project proposal**	
"Whole genome sequence analysis in tomato: alien DNA footprints by introgression breeding"	Sep 10, 2011
► **Writing a review or book chapter**	
► **MSc courses**	
C++ For Bioinformatics	Oct 03-28, 2011
GEN-30306 Genetic Analysis, Tools and Concepts (GATC)	Sep 2012
► **Laboratory use of isotopes**	
Subtotal Start-up Phase	*13.5 credits**

2) Scientific Exposure	date
► **EPS PhD student days**	
EPS PhD Student Day 2011, Wageningen University	May 20, 2011
EPS PhD Student Day 2013, Leiden University	Nov 29, 2013
► **EPS theme symposia**	
EPS Theme 4 Symposium 'Genome Biology', Wageningen University	Dec 09, 2011
EPS Theme 4 Symposium 'Genome Biology', Radboud University	Dec 07, 2012
EPS Theme 4 Symposium 'Genome Biology', Wageningen University	Dec 13, 2013
EPS Theme 4 Symposium 'Genome Biology', Wageningen University	Dec 03, 2014
► **NWO Lunteren days and other National Platforms**	
ALW meeting experimental plant science Lunteren	Apr 02-03, 2012
ALW meeting experimental plant science Lunteren	Apr 22-23, 2013
► **Seminars (series), workshops and symposia**	
Bioscience Day	Nov 03, 2011
EPS Expectation Day	Nov 18, 2011
EPS Mini-Symposium 'Plant Breeding in the Genomics Era'	Nov 25, 2011
Life Science Momentum Rotterdam	Nov 22, 2011
WUR Sequencing Seminar	Dec 07, 2011
CBSG Summit	Mar 29-Apr 01, 2012
CBSG Workshop - IPR in a breeding context	Apr 07, 2012
NBIC workshop Software Licensing and Valorization	May 25, 2012
CBSG Summit	Mar 11-13, 2013
Next Generation Sequencing (NGS) methods for identification of mutations and large structural variants	Mar 11-12, 2014
NBIC Conference	Apr 08-09, 2014
EPS Symposium 'Omics Advances for Academia and Industry - Towards True Molecular Plant Breeding'	Dec 11, 2014
► **Seminar plus**	
► **International symposia and congresses**	
International Symposium on Integrative Bioinformatics 2011	Mar 21-23, 2011
SOL2012 9th Solanaceae Conference, From the Bench to Innovative Applications	Aug 26-30, 2012
Plant Genomics Congress - London	May 12-13, 2014
► **Presentations**	
EPS Theme 4 Symposium 'Genome Biology' - Talk	Dec 09, 2011
CBSG Summit - Poster	Mar 01, 2012
EPS Theme 4 Symposium 'Genome Biology' - Talk	Dec 07, 2012
CBSG Summit - Poster	Feb 11, 2013
NBIC Conference - Poster	Apr 08-09, 2014
NBIC Conference - Talk	Apr 08-09, 2014
► **IAB interview**	
Meeting with a member of the International Advisory Board of EPS	Sep 29, 2014
► **Excursions**	
Subtotal Scientific Exposure	*17.8 credits**

3) In-Depth Studies	date
► **EPS courses or other PhD courses**	
WIAS course Statistics for the Life Sciences	May 20-27, 2011
NBIC Comparative Genomics - from evolution to function	Jun 27-Jul 01, /2011
MGC Promovendi Course 'Next Generation Sequencing'	Sep 05-07, 2011
Identity by Descent (IBD)	Jun 03-06, 2012
Natural Variation in Plants	Aug 21-24, 2012
► **Journal club**	
Member of literature discussion group	2011-2013
► **Individual research training**	
Subtotal In-Depth Studies	*9.8 credits**

4) Personal development	date
► **Skill training courses**	
Competence Assessment	Nov 23, 2011
Dutch for employees	Sep 08, 2013
Scientific writting 1	Mar 28, 2014
► **Organisation of PhD students day, course or conference**	
► **Membership of Board, Committee or PhD council**	
Subtotal Personal Development	*3.6 credits**

TOTAL NUMBER OF CREDIT POINTS*	**44.7**

Herewith the Graduate School declares that the PhD candidate has complied with the educational requirements set by the Educational Committee of EPS which comprises of a minimum total of 30 ECTS credits

** A credit represents a normative study load of 28 hours of study.*

The research described in this thesis was financially supported by the Centre for BioSystems Genomics (CBSG) under the grant number TO09.

Financial support from Wageningen University for printing this thesis is gratefully acknowledged.