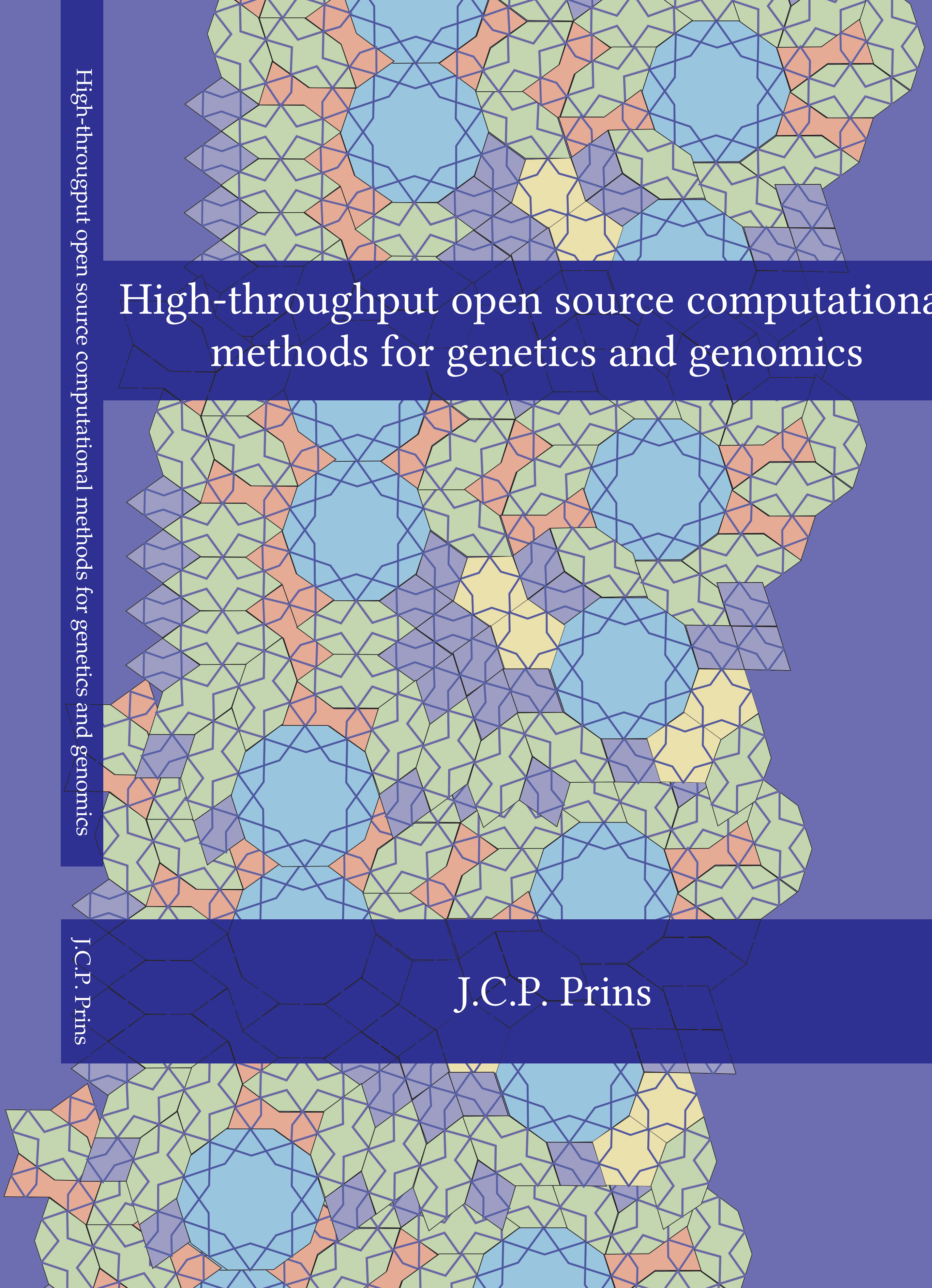
High-throughput open source computational methods for genetics and genomics

J.C.P. Prins



High-throughput open source computational methods for genetics and genomics

J.C.P. Prins

Propositions

1. Fully parallelized data processing as exemplified by Sambamba will be essential to analyze big data.
   (this thesis)

2. Opinions are equally important as results in accomplishing scientific breakthroughs.
   (this thesis)

3. *Homo sapiens* is now a model for other species.

4. Scientists are a factor never accounted for in calculating the false discovery rate.

5. Those who fashionably claim to study complex biology basically should be heed to Einstein's 'if you can't explain it simply, you don't understand it well enough'.

6. If you think humans are the summit of evolution, take a second look at the domesticated cat.

Propositions belonging to the thesis entitled:
'High-throughput open source computational methods for genetics and genomics'

J.C.P. Prins (Pjotr)
Wageningen, 5th of October 2015

High-throughput open source computational methods for genetics and genomics

J.C.P. Prins

Thesis committee

Promotors
Prof. Dr J. Bakker
Professor of Nematology
Wageningen University

Prof. Dr R.C. Jansen
Professor of Bioinformatics
University of Groningen

Co-promotor
Dr G. Smant
Assistant Professor, Laboratory of Nematology
Wageningen University

Other members
Prof. Dr D. de Ridder, Wageningen University
Prof. Dr M.E. Schranz, Wageningen University
Dr V. Guryev, European Research Institute for the Biology of Aging, Groningen
Dr M.A. Swertz, University Medical Centre Groningen, Groningen

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

High-throughput open source computational methods for genetics and genomics

J.C.P. Prins

Thesis
submitted in fulfilment 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 5 October 2015
at 4 p.m. in the Aula.

J.C.P. Prins
High-throughput open source computational methods for genetics and genomics,
137 pages.

PhD thesis, Wageningen University, Wageningen, NL (2015)
With references, with summary in English

ISBN 978-94-6257-459-5

Table of contents

1 Introduction **1**

2 A cross-species genome-wide scan for nematode gene-families subject to diversifying selection **7**

3 GenEST, a powerful bidirectional link between cDNA sequence data and gene expression profiles generated by cDNA-AFLP
Nucleic Acids Research (2001) **19**

4 Identification of imprinted genes subject to parent-of-origin specific expression in *Arabidopsis thaliana* seeds
BMC Plant Biology (2011) **31**

5 R/qtl: high throughput Multiple QTL Mapping
Bioinformatics (2010) **53**

6 Genetical Genomics for Evolutionary Studies
Methods in Molecular Biology (2012) **57**

7 BioRuby: Bioinformatics software for the Ruby programming language
Bioinformatics (2010) **71**

8 BioGem: an effective tool based approach for scaling up open source software development in bioinformatics
Bioinformatics (2012) **77**

9 Sambamba: fast processing of NGS alignment formats
Bioinformatics (2015) **83**

10 Big data, but are we ready?
Nature Reviews Genetics (2011) **87**

11 Towards effective software solutions for big biology
Nature Biotechnology (2015) **91**

12 General Discussion **95**

Summary **109**

Acknowledgements **113**

Bibliography **115**

List of publications **135**

1

Introduction

1.1 Elucidate biological processes from sequenced DNA and RNA

DNA encodes the information for all life on earth. In the living cell, sections of DNA get transcribed to messenger RNA which in turn gets translated into proteins. Proteins are the building blocks and functional molecules which take care of most molecular processes inside an organism. This flow from DNA to protein is known as the central dogma of molecular biology[1].

Since the invention of Sanger sequencing in 1977[2], technology allows for ever faster decoding of both DNA and RNA into the purine bases, adenine 'A' and guanine 'G' and the pyrimidines cytosine 'C' and thymine 'T' (RNA uses uracil in place of thymine). These four nucleotide bases are the building blocks of DNA. The sequencing effort leads to DNA/RNA sequences comprising the genetic code for every individual and, stringing out the genetic code into a combination of A, G, C and T's. The full string of DNA forms the genome.

Genes are the coding part of the genome. In human the genes represent approximately 2% of our DNA. These genes ultimately code for proteins. Intriguingly, as a result of sequencing, we are finding that large parts of, so called, non-coding DNA actually gets transcribed into 'non-coding' RNA and has function in gene-expression regulation and post-transcription modification, i.e., non-coding DNA has function too. The overall landscape of DNA transcription and regulation is complicated: the encyclopedia of DNA Elements (ENCODE) has systematically mapped regions of transcription, transcription factor association, chromatin structure and histone modification with the result that approximately 80% of the human genome appears to have some form of biochemical function[3].

For agriculture, biology and biomedical research an important challenge is to link phenotype, i.e., the observable characteristics or traits, with that of genotype, i.e., the DNA. Examples of phenotype linked to genotype are the number of spots on a leopard, colour blindness, and susceptibility to certain disease. In rare cases a single phenotype can be linked to a single gene. For example, one gene is causal for the human sickle-cell disease[4] and mutations in the BRCA1 gene are responsible for approximately 40% of inherited breast cancers, even though these mutations account for only 2-3% of all breast cancers[5]. It was even found that mutations in one gene define the ambling gait characteristic of the Icelandic horse[6]. With most phenotypes of interest, however, many genes are involved and a wide range of variation in regulation, transcription, translation and post-translational modification complicate the disentanglement of the relation between genotype and phenotype. The discipline of genetics, as it is used here, concerns the study of such quantitative or complex traits. Today, we live in the era of genomics, where the science of genetics is combined with high-troughput technologies, such as DNA sequencing. Computational biology, in conjuction with bioinformatics, makes it possible to help sequence, assemble, and analyse the function and structure of genomes and, ultimately, improve the understanding of how phenotype relates to genotype.

This thesis focuses on the elucidation of biological processes from sequenced DNA and RNA using high-throughput technologies. Software solutions were written and explored for analysing sequenced data, starting with the DNA of the smallest multi-cellular organism on earth, the nematode or roundworm. Nematodes are studied, not only because they are partial to human, animal, and plant disease, with huge associated cost, but also because they are useful model organisms for studying particular biological phenomena which provide insight into the workings of higher organisms. Nematodes belong to the most successful species on earth, living in diverse habitats. There may exist over one million species. So far, about 28,000 species have been described. The genomes of about twenty parasitic and non-parasitic nematodes has been sequenced, including that of the free-living nematode *Caenorhabditis elegans* which was famously the first multi-cellular species to have its genome published[7].

Broadly, bioinformatics can be defined as the application of information technology to the field of molecular biology. In the context of this thesis the term bioinformatics is used for the creation and advancement of computational solutions for genetics and genomics, where genomics concerns the effort of sequencing and analysis of the function and structure of genomes. Bioinformatics is interdisciplinary because it serves both biology and informatics. In this thesis a number of research questions were formulated that were directed more towards the 'bio', and others that were directed more towards the 'informatics' in bioinformatics. For example, bio-type research questions asked were 'How can we identify genes involved in pathogenicity or plant defence from DNA and RNA sequences?' and 'How can we identify genes that are expressed differentially and relate them to a phenotype'. Informatics-type research questions asked were 'How can we improve tools for genetic analysis in the era of high-throughput sequencing?' and 'How can we scale up computations and be prepared for the genomic data deluge?'.

The chapters in this thesis pursue to answer such questions with the aid of software solutions that were developed and published. The usefulness and application of bioinformatics solutions quickly extends beyond the phyla of Nematoda. Software written for nematodes is generally useful for research on other eukaryotes, such as plants, insects and mammals.

Through the internet, free and open source software (FOSS) is published 'early and often' and made available to the wider international biological and biomedical research community who can immediately access and run the tools.

1.2 Outline of this thesis

Identifying nematode genes involved in plant pathogenicity

In Chapter 2 'A cross-species genome-wide scan for nematode gene-families subject to diversifying selection' a strategy was defined for locating genes involved in plant pathogenicity on a genome-wide scale. Plant parasitic nematodes deliver a battery of secreted effectors into the apoplast and cytoplasm of plant cells to enable the invasion of the host and to alter the structure and function of host cells. Some of these nematode effectors interact at a molecular level with the plant innate immune system, which triggers a potent defence response in the plant. The nematode effectors capable of inducing plant immune responses appear to be the products of gene families harbouring stretches of hypervariable coding sequences. This finding led us to investigate whether such footprints of positive selection in genomes of parasitic nematodes can also be used to identify novel nematode effectors by executing a genome-wide scan for evidence of positive selection in multiple species. This resulted in a multi-species database of DNA sequences possibly involved in pathogenicity, including those for nematodes *Meloidogyne incognita* and *Meloidogyne hapla*, providing a resource which grows in value with every genome that is added.

In Chapter 3, 'GenEST, a powerful bidirectional link between cDNA sequence data and gene expression profiles generated by cDNA-AFLP' cDNA-AFLP was combined with available nematode sequence material, this time by efficiently integrating information captured in the form of expressed sequence tags (ESTs) to find novel factors involved in *Globodera rostochiensis* pathogenicity through gene expression profiles.

Identifying maternally controlled plant genes

In Chapter 4 'Identification of imprinted genes subject to parent-of-origin specific expression in *Arabidopsis thaliana* seeds' the challenge was to identify genes that are expressed differentially in alternatively spliced DNA. Essentially we applied the same technique developed for ESTs and nematodes (Chapter 3) to predict splice variants of genes with differentially expressed transcript-derived fragments (TDFs) from the DNA sequence of the plant and model organism *A. thaliana*. The resulting software solution GenFrag predicted genes involved in epigenetic parental imprinting in seed and identified 52 candidate maternally expressed genes in seed from the genome sequence of *A. thaliana*.

Improving tools for genetics in the age of high-throughput sequencing

To find correlation between phenotype and genotype, statistical tools can be applied. Quantitative trait loci (QTL) mapping is a time proven statistical method for linking phenotype to genotype. For example, high throughput microarrays and RNA sequencing (RNA-seq) measure gene expression levels which are quantitative traits and can be analysed in the same way as classical traits. This strategy, when applied to gene expression levels on a genome-wide scale, is known as 'genetical genomics'[8].

In Chapter 5 'R/qtl: high throughput Multiple QTL mapping' we added multiple QTL mapping (MQM), a sensitive approach for QTL mapping, to the R/qtl QTL mapping software-suite for genetical genomics. Not only is this implementation of MQM part of the statistical R/qtl environment, which is both for use on the desktop and for scripting pipe-lined setups, it is also scalable through parallelisation and allows for high-throughput QTL analysis. For determining significance in large data sets we included permutation strategies for determining thresholds of significance relevant for QTL and QTL hot spots. This work has resulted in MQM for R/qtl being a high-performance comprehensive QTL mapping toolbox for the analysis of experimental populations which is increasingly used for research in model organisms, including *Mus musculus* (mouse), *C. elegans*, *A. thaliana*, and *S. lycopersicum* (tomato).

Chapter 6 'Genetical Genomics for Evolutionary Studies' is a review-style book chapter that builds on the previous chapters and suggests a strategy for identifying genes involved in plant-resistance, combining QTL analysis with genome-wide scans for positive selection.

Scaling software development in biology

With the increasing amount of data being churned out by sequencers, the bioinformatics part is increasingly a bottleneck in creating software solutions and executing data analysis. Scalability problems come in multiple forms. Scaling software development for high-throughput DNA and RNA analysis, for example, can be achieved by pooling software engineering resources.

In Chapter 7 'BioRuby: Bioinformatics software for the Ruby programming language' and Chapter 8 'BioGem: an effective tool based approach for scaling up open source software development in bioinformatics', we demonstrate that improved collaborative software development efforts can be applied to scale up software development in bioinformatics.

Scaling software solutions for the genomic data deluge

A different type of scalability problem concerns the non-linear growth of data produced in biology which puts a strain on not only people and software development resources, but also on infrastructure. Data analysis takes significant computational resources. Chapter 9 ‘Sambamba: fast processing of NGS alignment formats’ introduces a software solution that makes effective use of multi-core processing to speed up processing of next generation sequencing (NGS) data.

Currently the genomes of tens of thousands of people are being sequenced. Soon, millions world-wide will be sequenced. In addition RNA sequencing is on the increase. RNA sequencing will drive data growth as it is applied over multiple experimental conditions, tissues, time series, and may include the meta-transcriptome of all bacterial gut data which is much larger than the host genome, e.g., [9]. The same is happening for animal and plant sequencing, with implications for both medicine and agriculture. Sequencing centers are already handling petabytes of sequence data within species. Analysing such large sized data is non-trivial, a point we make in Chapter 10 ‘Big Data, but are we ready?’. In Chapter 11 ‘Towards effective software solutions for big biology’ we point out that significant investments in bioinformatics software development are required to realise the potential of big biology.

A perspective on High-throughput open source computational methods for genetics and genomics

One of the exciting developments in biology and biomedical research is the emergence of NGS. NGS is dramatically faster than older sequencing techniques, but raises its own challenges, in particular when it comes to data size and fidelity. The final Chapter 12 ‘General Discussion’, brings up future perspectives of FOSS development for genetics and genomics, a prelude to the development of new software solutions for NGS and QTL mapping.

2

A cross-species genome-wide scan for nematode gene-families subject to diversifying selection

Persistent infections by parasitic nematodes are recognised as a major health concern in animals, humans, and plants. Host organisms and obligate parasites are embroiled in an evolutionary arms race with reciprocal adaptations driving sequence diversification in duplicated genes. To identify genes subject to diversifying selection in the plant-parasitic nematode *Meloidogyne incognita*, evidence of positive selection was analysed on a genome-wide scale in multiple parasitic and non-parasitic nematode species. Clusters of highly similar duplicated sequences within nematode genomes were investigated for evidence of positive selection using the ratio of non-synonymous to synonymous nucleotide substitution rates (dN/dS or ω). By comparing the presence of positively selected clusters across genomes of nematodes with entirely different life histories, we identified gene families uniquely associated with plant parasitism.

Manuscript in preparation 'A cross-species genome-wide scan for nematode gene-families subject to diversifying selection' by Pjotr Prins, Anna Tomczak, Jaap Bakker, Maria Anisimova, Aska Goverse and Geert Smant

2.1 Introduction

Nematodes are the most ubiquitous and abundant animals on earth, inhabiting all terrestrial and marine habitats. They have a simple build that includes a complete alimentary tract to acquire essential nutrients, a neural system with sensory organs to perceive environmental cues, muscles to move about, and multiple glandular systems to maintain the internal homeostasis and to interact with other organisms in their immediate environment.

Even though nematodes are barely visible with the naked eye, they are successful models for studying organismal development and complex genetic diseases of humans and other higher animals. The first species in the animal kingdom that had its genome of approximately 100 million nucleotide base pairs (Mb) fully sequenced and published in 1998 for this particular purpose was the soil-dwelling nematode *Caenorhabditis elegans*, a free-living species that thrives on bacteria and other microbes[7].

A subset of nematode species has evolved into highly advanced parasites of animals and plants. Over one-quarter of the global human population carries parasitic nematodes, often without clinical symptoms, e.g., [10]. Animal parasitic nematodes included in this study are *Brugia malayi*, *Trichinella spiralis* and *Strongyloides ratti*, all of which employ fundamentally different parasitic strategies to exploit their host. *B. malayi* is one of the three causative agents of lymphatic filariasis (elephantiasis) in humans and was the first parasitic nematode which had its genome sequence published in 2004[11]; *T. spiralis* thrives in muscular tissue of essentially all mammals; and *S. ratti* is a common gastro-intestinal parasite of rats. Not all parasitic nematodes have a negative impact on health in natural ecosystems and many parasitic species are essential parts of food webs[12]. Plants can co-exist in intimate and prolonged relationships with parasitic nematodes. To identify genes involved in plant-parasitism of *M. incognita*, we included the following harmful plant parasitic nematodes in this study: *M. incognita*, the main object of this study, *Meloidogyne hapla*, *Globodera pallida* and *Bursaphelenchus xylophilus*. *M. incognita* is a biotrophic root-knot plant parasitic nematode with a worldwide distribution and numerous hosts, including tomato, cotton, coffee, and grape. *M. hapla* is another ubiquitous nematode that invades many known hosts, including vegetables. *G. pallida* is an important pest in potato cultivation and *B. xylophilus* is causal for Pine wilt disease. The typical persistence of nematode infections gives evidence of extensive mutual adaptations in nematodes and plants that they acquired during a prolonged common evolutionary history. The purpose of the study in hand was to develop a genome-wide scan for gene families subject to positive, diversifying selection and to identify positively selected genes uniquely associated with parasitism of the root-knot nematode *M. incognita*.

The genome sequence of the root-knot nematode *M. incognita* was published in 2008[13]. Most of the assembled sequence of this asexually reproducing nematode, totaling approximately 86 Mb, exists in pairs of homologous but divergent segments. The sequencing effort suggested that ancient allelic regions in *M. incog-*

nita are evolving toward effective haploidy and permitting new mechanisms of adaptation. The number and diversity of plant cell wall degrading enzymes in *M. incognita*, for instance, was unprecedented in any animal for which a genome sequence was available and may be derived from multiple horizontal gene transfers from bacterial sources[13].

To invade their host, all plant-parasitic nematodes have developed mechanisms to alter the structure and function of host cells. Plant-parasitic nematodes deliver a battery of secretory proteins into the apoplast and cytoplasm of host cells. Some of these nematode effectors are known to interact at a molecular level with the plant innate immune system, which may trigger a potent defence response in the plant[14]. Further investigations into the coding sequences of nematode effectors that are recognised by plant immune receptors have revealed evidence of positive diversifying evolutionary selection in nematode genomes. For example, the SPRY Domain-Containing Protein SPRYSEC, from the plant-parasitic nematode *G. rostochiensis*, interacts with a CC-NB-LRR protein from susceptible tomato host *Solanum lycopersicum*)[15]. Many nematode effectors associated with the induction or suppression of plant immune responses appear to be the products of gene families harbouring stretches of hypervariable coding sequences. This finding led us to investigate whether such footprints of positive selection in genomes of parasitic nematodes can also be used to identify novel nematode effectors.

By using sequences from more than two related species and by making comparisons between evolutionary models within a likelihood framework, it is possible to identify both lineage-specific trends and to quantify the relative strengths of positive selection, negative selection, and neutral evolution[16]. Natural selection is inferred by estimation of dN/dS (or ω), i.e., the ratio of non-synonymous (dN, amino acid changing) to synonymous (dS, amino acid retaining) substitution rates. With $\omega < 0$, $\omega = 1$ and $\omega > 1$ representing purifying, neutral and adaptive evolution respectively. Identification of genes with $\omega > 1$ is persuasive evidence for adaptive evolution at a particular locus[17]. A review of the validity of this approach can be found in [18]. The software tool CODEML, part of the phylogenetic analysis by maximum likelihood package (PAML), can identify codon sites which show evidence of positive selection by testing evolutionary models that allow for positive selection against models that do not allow for positive selection and applying a likelihood ratio test (LRT). If the model that allows positive selection fits the data significantly better, as judged by the LRT, positive selection is inferred[16]. To calculate ω, e.g., [19, 20], we investigated clusters of paralogous predicted coding DNA sequences (CDS) for evidence of positive selection. The statistical significance of $\omega > 1$ per site was assessed under PAML's evolutionary models M7 and M8[17](see methods).

Because CDS are not available for all species and predicted CDS sets do not fully represent real gene sets, especially for less-studied species[21], we also tested for evidence of positive selection using clusters of orthologous open reading frames (ORF, see methods).

For the large majority of orthologous gene sequences in different but related species ω is small, confirming the general assumption that non-synonymous mutations are selected against and that purifying selection is the dominant force in evolution[20]. In contrast, paralogous genes in nematodes exhibit a relaxation of selective constraints and may be subject to positive selection. We focused our studies on the adaptive evolution of the families of duplicated genes in nematode genomes[22]. As a 'surrogate outgroup' we analysed the genome of *Phytophthora infestans*, a plant-pathogenic oomycete with a large genome (approximately 240Mb) that causes potato late blight[23].

To identify genes encoding proteins that are likely involved in the host-parasite interactome (HPI), in addition to above inter-species comparison, evidence was collected and added to a database on conserved properties of sequences from NCBI Refseq[24]. Also, information was compiled that predict whether a protein is secreted because most known nematode effectors delivered into the plant contain a signal peptide (SP) at the N-terminus of the protein[15, 25].

Previous analysis of the genome of parasitic nematodes identified families of orthologous genes shared between these species and gene families which are specific to the lineage, e.g., [13, 20]. To our knowledge, this is the first study that includes such a large-scale cross-species comparison of adaptive evolution at the codon level through both predicted CDS and ORFs which were combined with known functional DNA characteristics.

2.2 Results

Genome-wide identification of positive selection

Gene families in plant-parasitic nematodes can be subject to positive, diversifying selection. To identify such gene families, predicted CDS from the genome of a species were clustered at a conservative 70% amino acid sequence identity with the BLASTCLUST tool, part of the NCBI-BLAST software suite (see 2.4). For the plant parasitic nematode, *M. incognita* out of a total of 20358 predicted CDS, 260 clusters were identified containing four or more sequences. Of these clusters, 43 (17%) clusters contain hypervariable codon locations showing significant evidence of positive selection ($p < 0.05$), according to the CODEML algorithm of PAML[16]. For five other nematodes species with predicted CDS available (i.e., *C. elegans*, *G. pallida*, *M. hapla*, *P. pacificus*, and *T. spiralis*) we used the same approach to identify CDS-based positively selected clusters (PSC) (see supplementary figure online Fig. S1 at http://biobeat.org/GWP/).

To expand the genome search space, ten nematode species and one oomycete (i.e., *B. malayi*, *B. xylophilus*, *C. briggsae*, *C. elegans*, *G. pallida*, *M. incognita*, *M. hapla*, *P. pacificus*, *S. ratti*, *T. spiralis*, and *P. infestans*) were included to identify sequence families from ORFs. ORFs were generated by splitting the full genome sequence into fragments on stop codons in six reading frames. Only sequences were included in the cluster analysis that contained more than 60 codons and were less than 10% masked (see methods). For *M. incognita* 1873 ORF clusters were identified containing four or more sequences having a sequence identity of more than 70%, out of a total of 299,643 ORFs. Of these clusters, 325 (17%) clusters contain hypervariable codon locations showing significant evidence of positive selection ($p < 0.05$) (Fig. S1).

To estimate the overlap between ORF PSC and CDS PSC, a within-species nucleotide MegaBLAST search was executed ($p < 10^{-5}$). MegaBLAST is optimised for aligning nucleotide sequences that differ slightly[26]. All matches were filtered for alignment-length $> 60bps$, resulting in 252,089 significant ORF hits against 19,632 (or 96%) of CDS for *M. incognita*, i.e., almost all CDS matched one or more ORFs. Using these results, we identified all individual ORF sequences that make up ORF PSC and matched them against CDS sequences that make up CDS PSC. 59 out of 325 (18%) ORF-based PSC contained one or more sequence matches and thus overlap with CDS-based PSC. The other way, 17 out of 43 (40%) CDS-based PSC contained one or more matches and thus overlap with ORF-based PSC. At the PSC level, ten CDS PSC were matched by more than one ORF PSC, possibly because of shared domains/motifs. For example, the large CDS cluster0001 matches the GATA transcription factor gene families with multiple ORF PSC hits (supplementary tables online ORF CDS matches). Altogether, the combined *M. incognita* CDS-based and ORF-based positive selected clusters resulted in a total of $325 - 59 + 43 = 309$ unique paralogous sequence families subject to positive, diversifying selection.

Identification of PSC that are conserved in annotated species

Not all gene families under positive, diversifying selection, are involved in plant-parasitism. To prune PSC that have conserved functional properties in non-parasitic organisms, PSC were identified that have similarity to members of the NCBI curated non-redundant Refseq database which contains predicted proteins from over 30,000 organisms[24]. For all 11 genomes of the species in this study, a PSC was considered conserved when its member sequences shows significant BLAST similarity ($p < 10^{-5}$) to organisms annotated in Refseq.

For *M. incognita*, 27 (63%) of CDS-based PSC and only 36 (11%) of ORF-based PSC showed significant similarity to accessions from other organisms in protein databases After subtracting the PSC harbouring sequence conservation in multiple other organisms we identified a total of $325 - 36 + 43 - 27 = 305$ unique PSC that may be associated with plant-parasitism of *M. incognita* (see Fig. S1 and supplementary tables online Refseq1 and Refseq2).

Identification of conserved evidence of positive selection across nematode genomes

To identify PSC that display 'conserved' footprints of adaptive evolution (i.e. PSC that show homology to PSC in other species) all sequences contained in PSCs from the 11 genomes were stored in a searchable database. The members of this database were used as queries in a BLAST search ($p < 10^{-5}$) on the database itself to identify PSCs from different genomes sharing significant sequence similarity. The number of PSC of *M. incognita* with significant sequence similarity with PSC from other species in the full comparison set was 29 (67%) and 109 (34%), for CDS PSC and ORF PSC, respectively. Of those, respectively 9 (21%) and 75 (23%) showed similarity to predicted proteins from plant parasitic nematodes only (Fig. S1). Most of sequence similarity was found between ORF-based positively selected clusters in two closely related root-knot nematode species *M. incognita* and *M. hapla*. Some of the ORF-based PSC show sequence similarity with the more distantly related plant-parasite nematode *G. pallida*. No sequence similarity was found between PSCs from *M. incognita,* and animal-parasitic nematodes or *P. infestans*.

Eliminating the PSC of *M. incognita* that have sequence similarity either with the predicted protein sequences from other organisms in RefSeq or with PSC from other genomes in the searchable database resulted in 84 PSC uniquely associated with plant parasitism by root-knot nematodes in this study (Fig. S1).

Identification of PSC encoding nematode secreted proteins

Nematode effectors secreted into the apoplast and cytoplasm of plant cells typically carry a classical eukaryotic amino-terminal signal peptide for secretion and lack a transmembrane domain in the mature protein sequence. To identify PSCs harbouring putatively secreted proteins, we used SignalP[27] and PHOBIUS[28] algorithms to predict whether sequence members of PSC in our database harboured an N-terminal signal peptide and possible transmembrane domain(s) (see methods).

For *M. incognita* both SignalP and PHOBIUS predicted that 20 (47%) CDS-based PSCs contained members with a signal peptide for secretion. Out of those, 13 CDS PSC were predicted to lack TMM activity. Meanwhile, 43 (13%) ORF PSC contain SP according to SignalP and 50 (15%) according to PHOBIUS. Out of those, 35 ORF-based PSC were predicted for TMM activity (supplementary tables online S1 and S2).

Functional ORF PSC consist of exons and contain sequences that are truncated from the 5' and/or 3' end. To identify the full-length genes that correspond to genes from *M. incognita* ORF PSC we reused above MegaBLAST results (for *M. incognita* 252,089 significant ORF hits against 19,632 (96%) of CDS). The full-length CDS sequences which match the short ORF PSC sequences out resulted in identification of an additional 19 SP-protein containing ORF PSC, whereof 11 lack a TMM.

The three sets of SP-proteins (20 from CDS data, 50 from ORF data and 19 from ORF expanded MegaBLAST) were checked for overlap and resulted in a total of 77 PSC that may be secreted by the plant-parasite *M. incognita* (supplementary tables online S3).

A database of positively selected gene families from nematode

To make the analysis reproducible and the data available for further analysis we made the software available for download under a free and open source software license together with a queryable database containing the full body of data analysed in this study. The database includes computed PSC, annotated sequence homology from both Refseq and the special BLAST database of species created for this study, as well as the annotation from SP and TMM topology predictions. This database comes in the form of a linked data resource description framework (RDF) graph containing 14,054,846 data points (triples) which can be loaded, for example, in a 4store triple-store[29] and allows further searches in data and relationships[30].

2.3 Discussion

This study adds weight to the findings that the DNA sequence of plant-parasitic nematode *M. incognita* contains evidence of adaptive evolution driven by DNA sequence diversification through duplicated genes which was suggested by the first published sequence initiative[13]. Abad *et al.* stated that the genome of *M. incognita* harbours ancient allelic regions that are evolving toward effective haploidy and permitting new mechanisms of adaptation. Because evidence of positive selection in highly conserved paralogous sequences is not likely to represent random DNA motifs, we hypothesise that many plant-parasitic nematode PSC may represent a functional role in either the innate immune system or host-parasite interactions.

In this study we applied the CODEML algorithm in PAML to test for evidence of positive selection on a genome-wide scale. We have not quantified or formally controlled the false positive (FP) rate of detecting sequences under positive selection, but we have aimed to minimise false positives by clustering sequences on a conservative 70% sequence identity threshold and by gathering additional evidence from multiple approaches (i.e., inter-species homology comparisons, functional annotation and evidence of DNA coding for signal peptides) to narrow down on possible gene candidates. In the highly-variable sequence regions, inferior alignments are known to be a source of FPs. When comparing alignment algorithms Villanueva-Cañas *et al.*[31] found the estimated fraction of positively selected genes with PRANK alignments was consistently lower than with MAFFT alignments, also in agreement with previous results[32–34]. Based on this information we selected the codon-based PRANK aligner even though PRANK is computationally slower than the other aligners. Villanueva-Cañas *et al.* focused on genome-wide scans with PRANK and PAML assessing protein isoforms of similar length returned less false positives which can be reduced by using protein isoforms of similar length. They state that further improvements in methods for the automated analyses of gene families are highly desirable[31] and, according to Markova and Petrov, one should account for a false positive rate of 50%[33]. We tried automated alignment cleansing procedures as provided, for example, by Gblocks which eliminates poorly aligned positions and divergent regions of a DNA or protein alignment[35]. Gblocks is aimed at making alignments more suitable for phylogenetic analysis. We even wrote our own alignment correction routines as part of Biogems (Chapter 8). In the end we decided that these methods were too stringent and reasoned that the strategy of combining evidence from multiple approaches would render a limited number of results that could be tested and validated in the laboratory.

Two studies relevant for multi-species genome-wide searches for evidence of positive selection were recently published. Roux *et al.* studied positive selection in seven ant species and compared them with ten bee species and twelve fly species, resulting in 24 functional categories of genes which were enriched for positively selected genes in the ant lineage. Roux *et al.* also combined evidence from other sources (mostly using GO categories) to zoom in on gene families[36]. Moretti *et al.* published an update on the running Selectome 'protein evolution' database project which hosts searchable precomputed estimates of positive selection from the CODEML branch-site test. The Selectome database applies more stringent filtering criteria and contains 6810 gene (family) trees that display evidence of positive selection from 81 species[37].

In our search for sequences that make up the host-parasitic 'interactome' we included ORFs because gene predictors are known to be less effective for poorly-studied species[21] (see also Chapter 12). Another reason to include ORFs as source data is that the general view of the genome has become increasingly plastic over time and even non-coding sequences may have functional properties (e.g., [38]) and may form a reservoir for novel effectors in pathogenicity, as was shown in bacteria[39]. Furthermore, non-coding sequences may be precursors for smaller mRNAs[40]. While 96% of CDS match known ORF sequences, only 17 out of 43 (40%) CDS-based PSC contained one or more sequence matches with ORF-based PSC. The limited overlap of sequence families showing evidence of positive selection may (partly) be explained by the fact that the sequences in the ORF-based PSC are much shorter than the sequences in the CDS-based PSC (e.g., [41]) while at the same time they are selected for further analysis based on their hypervariable regions (showing evidence of positive selection). Even if the search for evidence of positive selection is inherently difficult and will underestimate the real number of PSC, adding ORF-based PSCs significantly expanded the final number of sequence families potentially involved in host-pathogen interaction from 20 CDS-based PSCs to 77 PSCs in total.

Future work may include relaxing the 70% identity constraint for original clustering of sequences. This will result in larger sets of gene families to study including those proteins with a smaller conserved scaffold and larger (hyper)variable regions. Another improvement would be to break down highly diversified gene families into smaller sequence clusters, along likely phylogenetic branches, and test them all separately for positive selection. In the current version with large sequence clusters, the number of sequences included in the alignment was limited so that both PRANK and PAML could finish within reasonable computation time.

Other potential improvements would be to include additional branch-site models of evolution, such as provided by PAML[42].

The relatively large genome of plant-parasite *M. incognita* and the PSC discovered in this study suggests that *M. incognita* harbours conserved coding and non-coding sequences under current or recent diversifying selection. Some of these will be part of the host-parasite interactome and may thus help explain the success of *M. incognita* in attacking a large range of hosts. Altogether, our approach has resulted in the identification of 77 PSC unique associated with plant-parasitism in the genome of *M. incognita*. Further functional characterization of these PSC in the laboratory is needed to pinpoint the subset of PSCs with significant contributions to virulence of root-knot nematodes in plants. Our approach was validated by PSCs identified in this study which included previously discovered effectors in *M. incognita*[43]. We conclude that, although our search for positive selection may have returned FPs and although we may have lacked power of detecting positive selection (FNs), the overall approach of clustering paralogues sequences and using PAML for detecting evidence of positive selection, followed by drilling down on functional characteristics of sequences and a between-species comparison, has rendered a novel and useful resource for plant-parasite research that can easily be explored further in the laboratory.

2.4 Materials and Methods

Databases and software

G. pallida CDS and contigs (v1.0) were fetched from Sanger FTP. The other Nematode genome sequences were fetched from Wormbase (release W236)[22]. Predicted CDS transcripts for gene families were fetched for *C. elegans*, *M. hapla*, *P. pacificus*, and *T. spiralis*; as well as the full genomic soft masked whole genome for *C. elegans*, *M. hapla*, *P. pacificus*, *T. spiralis B. malayi*, *B. xylophilus*, *C. briggsae*, *C. elegans*, *M. incognita*, *M. hapla*, *P. pacificus*, *S. ratti*, and *T. spiralis*. The *P. infestans* CDS and genome sequence was fetched from ENSEMBL Genomes FTP (release 20). The RefSeq non-redundant BLAST database release 59 was downloaded from the NCBI FTP site[44] and indexed using a local BLAST installation[45]. BLAST 2.2.26[45], Paml 4.7[16], Prank-msa v130410[46], signalP-4.1[27], Phobius-1.01[28] and EMBOSS-6.6.0[47] were installed and run on Red Hat Linux 4.4.7-3 using gcc version 4.4.7 20120313.

Clustering of sequence families with BLASTCLUST

CDS sequences were prepared and translated for a within-species amino acid BLASTCLUST '-L .7 -b T -S 70', i.e., a 70% percentage identity and length coverage threshold. For ORF sequences EMBOSS getorf was used to create a FASTA file of nucleotide sequences between STOP codons. ORF sequences shorter than nucleotide 180 bps or with more than 10% masked nucleotides were discarded.

After BLASTCLUST all clusters with more than four sequences were collected and rewritten as FASTA nucleotide multi-sequences alignment (MSA) files containing the clustered sequences. To prevent Prank and PAML from seizing up and have computations complete within 24 hours, MSA's were truncated to 19 sequences.

Testing for positive selection with CODEML

For each sequence cluster the MSA was aligned using the codon aligner PRANK. PRANK also provided the phylogenetic tree, used as input for CODEML. Next, the CODEML programme of the PAML software was run on each cluster to predict for positive selection using M7-8 models. Relevant settings were CodonFreq=F3X4, model=0, fix_kappa=0, kappa=4, fix_omega=0, omega=5, ncatG=10. The results of the CODEML runs were turned into a digest of positive selected clusters (PSC) and added to the RDF database (see section 2.4).

Testing for similarity and conservation with BLAST

To compare PSC between species all amino acid sequences contained in PSC were compiled into a BLAST database. Next, every individual sequence in each PSC was 'BLASTed' against this database ($E-value < 10^{-5}$) and results were added to the RDF database. Also every sequence in PSCs was BLASTed against RefSeq[44] ($E - value < 10^{-5}$) and the result was added to the RDF database. A nucleotide MegaBLAST search[26] was executed of all clustered *M. incognita* ORF sequences against all sequences that make up CDS. MegaBLAST default settings were used and post-filtered for alignment-length > 60 and $p < 10^{-5}$. All BLAST results were transformed to RDF using the bioruby-blastxmlparser tool with the blast2rdf-minimal.erb template (see also Chapter 8).

Identification of secreted proteins with Signal-P and PHOBIUS

N-terminal signal peptide and possible transmembrane domain(s) in PSC sequences were identified using SignalP[27] and PHOBIUS[28] prediction algorithms using default settings. The result was added to the RDF database.

RDF database

All results were stored in an RDF graph, available for download from http://biobeat.org/GWP/ and for this study stored in a 4-store v1.1.4 container for SPARQL queries[29]. The SPARQL results were compiled for figures and tables using scripts.

Pipeline and scripts

In addition to the GWP scripts (http://github.com/pjotrp/) we wrote and added PAML-parsing support to BioRuby 1.4.3 (Chapter 7) and alignment support to the

Biogems bio-alignment gem (Chapter 8) as well as tools for parsing blastclust, BLAST XML and PAML output. Gems installed are bioruby-blastxmlparser 1.1.1, bioruby-bigbio 0.1.5 and bioruby-table 0.8.0. For all the tools default settings were used, unless mentioned differently.

All software and scripts for running the invidual steps are available from the git repositories http://github.com/pjotrp/ and supplementary data at http://biobeat.org/GWP/. For information on how to use the scripts in the pipeline, see the protocol document.

3

GenEST, a powerful bidirectional link between cDNA sequence data and gene expression profiles generated by cDNA-AFLP

The release of vast quantities of DNA sequence data by large-scale genome and expressed sequence tag (EST) projects underlines the necessity for the development of efficient and inexpensive ways to link sequence databases with temporal and spatial expression profiles. Here, we demonstrate the power of linking cDNA sequence data (including EST sequences) with transcript profiles revealed by cDNA-AFLP, a highly reproducible differential display method based on restriction enzyme digests and selective amplification under high stringency conditions. We have developed a computer program (GenEST) that predicts the sizes of virtual transcript-derived fragments (TDFs) of in silico digested cDNA sequences retrieved from databases. The vast majority of the resulting virtual TDFs could be traced back among the thousands of TDFs displayed on cDNA-AFLP gels. Sequencing of the corresponding bands excised from cDNA-AFLP gels revealed no inconsistencies. As a consequence, cDNA sequence databases can be screened very efficiently to identify genes with relevant expression profiles. The other way round, it is possible to switch from cDNA-AFLP gels to sequences in the databases. Using the restriction enzyme recognition sites, the primer extensions and the estimated TDF size as identifiers, the DNA sequence(s) corresponding to a TDF with an interesting expression pattern can be identified. In this paper we show examples in both directions by analyzing the plant parasitic nematode *G. rostochiensis*. Various novel pathogenicity factors were identified by combining ESTs from the infective stage juveniles with expression profiles of ∼4000 genes in five developmental stages produced by cDNA-AFLP.

Published as 'GenEST, a powerful bidirectional link between cDNA sequence data and gene expression profiles generated by cDNA-AFLP' by Ling Qin, Pjotr Prins, John T. Jones, Herman Popeijus, Geert Smant, Jaap Bakker and Johannes Helder

Nucleic Acids Res (2001)
29(7):1616-22. doi:10.1093/nar/29.7.1616

3.1 Introduction

With the advent of high throughput techniques for DNA sequencing, whole genome sequences from several organisms have become available[7, 48] and many others will be available in the near future. At the same time, millions of expressed sequence tags (ESTs), single pass sequences of cDNA clones selected randomly from a library, have been generated and deposited in public and private databases. Searching for homologous sequences in databases is usually the first step towards understanding the functions of newly identified genes. Homology information is useful for orthologous genes, but in the case of paralogs the value of this information may be more limited. Furthermore, it is often found that a significant proportion (40-60%) of newly identified DNA sequences lack homology with genes for which the functions are known [7, 49]. Additional tools are therefore needed to allow functional analysis of newly identified genes.

Biological responses and developmental processes are precisely controlled at the level of gene expression. Information on the temporal and spatial regulation of gene expression often sheds light on the potential function of a particular gene. Hence, an essential aspect of functional genomics is the transcriptome, i.e., the analysis of expression patterns of genes on a large scale. There are currently three high throughput techniques for large-scale monitoring of gene expression: serial analysis of gene expression (SAGE)[50], hybridization-based methods[51, 52], gel-based RNA fingerprinting techniques such as differential display[53] and cDNA-AFLP[54]. In principle, SAGE can provide quantitative data concerning gene expression. However, it is expensive and labor intensive when multiple sample points are to be compared. Microarray technology is very powerful in generating a broad view of gene expression. Unlike cDNA arrays, oligonucleotide arrays are able to distinguish between highly homologous sequences. However, the design of oligonucleotide arrays requires comprehensive sequence knowledge at present only available for a small number of organisms. cDNA-AFLP is an inexpensive gel-based method for analysis of gene expression patterns and can be performed in any laboratory.

In the cDNA-AFLP procedure cDNAs synthesized from mRNAs isolated from various sample points are digested by two restriction enzymes. Oligonucleotide adapters are then ligated to the resulting restriction fragments to generate template DNA for PCR. PCR primers complementary to the adapter sequences with additional selective nucleotides at the 3'-ends allow specific amplification of a limited number of cDNA fragments. Unlike differential display methods that make use of small random primers[53], relatively high annealing temperatures can be used and, hence, cDNA-AFLP is more stringent and reproducible. In contrast to most hybridization-based techniques, cDNA-AFLP will distinguish between highly homologous genes from gene families while (contrary to oligonucleotide arrays) no sequence foreknowledge is needed.

Since sequence information is accumulating at an unprecedented rate for a wide variety of organisms, there is an urgent need for efficient and inexpensive ways to screen these databases on genes with interesting expression profiles. Here, we report on the advantages of combining ESTs with cDNA-AFLP data. The potential benefits of this combination in gene discovery and functional analysis prompted us to develop a computer program that creates restriction patterns of cDNAs in silico in accordance with the enzyme combinations used in cDNA-AFLP. The resulting virtual cDNA fragments are ordered according to the extensions of the amplifying primers and their sizes. These virtual fragments can then be traced back on cDNA-AFLP gels to identify the corresponding bands, with primer extensions and fragment sizes as a unique identifier. The program can also be used in the opposite direction by using the size and primer extensions of a potentially interesting band identified on a cDNA-AFLP gel as criteria to search the corresponding cDNA. This simplifies the procedure of cloning full-length genes with interesting temporal and spatial expression patterns.

In this paper we demonstrate the utility of the program by linking EST sequence data and expression profiles of ~4000 genes from the potato cyst nematode *Globodera rostochiensis*, which causes extensive damage to solanaceous crops. Genes potentially related to the nematode's ability to parasitize plants were identified within a pool of hundreds of ESTs. We show that this program could be useful in any system where stage- or tissue-specific genes are to be selected from pools of (uncharacterized) cDNAs.

3.2 Results

ESTs and cDNA-AFLP-based expression profiles

A cDNA library from second stage juveniles in the H stage of the potato cyst nematode *G. rostochiensis* was used to sequence 985 cDNA clones. Starting from the 5'-end, the average read was ~600 bp[55]. In parallel, cDNA-AFLP-based gene expression profiles were generated from five distinct developmental stages, D, S, H, U and P, of this nematode species. The expression profiles were highly reproducible and no significant differences were observed between independent replicates. An average of 32 bands per lane were displayed using EcoRI and TaqI primers with two selective nucleotides (E+NN and T+NN, respectively) extending beyond the adapters into the cDNA. Approximately 8200 TDFs were displayed using the whole set of 256 ($16*16$) primer combinations. In a previous study[56] it was shown that genes involved in plant parasitism are usually up-regulated in developmental stages S and H or in stage H only. Bands showing such expression patterns were excised from gels, cloned and sequenced. Sequencing of > 100 TDFs revealed that the marker-based size estimations corresponded well to the actual sizes of these TDFs (with an accuracy of ±1 nt for bands < 300 nt and ±3 nt for bands > 300 bp).

Generation of virtual TDFs from ESTs using GenEST

We used GenEST to generate virtual TDFs from 985 ESTs. EcoRI and TaqI recognition sites were used as begin and end tags with a length modifier of 22 nt to account for the additional adapter sequences.

A total of 228 virtual TDFs derived from 159 ESTs were predicted by GenEST (Table 3.1). Of these 159 ESTs, 100 were predicted to produce a single virtual TDF, 51 were predicted to give rise to two virtual TDFs each (thereby generating 102 TDFs), six ESTs were predicted to result in three TDFs each (generating 18 TDFs) and two ESTs were predicted to generate four TDFs each (eight TDFs in total).

Table 3.1: Virtual TDFs generated after in silico restriction with EcoRI and TaqI of 985 ESTs randomly picked from a cDNA library from infective juveniles of the potato cyst nematode *G. rostochiensis* using GenEST. E+AN/CN/GN/TN are the extensions of the EcoRI primer (E, core primer). Each EcoRI primer was combined with all TaqI primers (T+NN). Note: E+GA will constitute both a TaqI and an EcoRI recognition sequence (GAATTCGA). In this case TDFs will not be amplified and cannot be traced back on a cDNA-AFLP gel. Therefore, these TDFs were not included in the total count[a].

	E+AN	E+CN	E+GN	E+TN	Total
N = A	20	17	32[a]	2	
C	6	11	15	11	
G	12	14	19	14	
T	14	33	22	18	
Total	52	75	56[a]	45	228[a]

To estimate how many genes are represented by the 8200 TDFs displayed in our study we have randomly extracted 1000 full-length cDNAs of *Caenorhabditis elegans* from GenBank (both the size and average GC content of the *G. rostochiensis* genome are similar to *C. elegans*; [56]. These sequences were processed by GenEST and 336 cDNAs (∼34%, the remaining cDNAs not containing both restriction sites) generated 693 virtual EcoRI/TaqI TDFs. The percentage of genes which produced TDFs is ∼48% (336/693 = 48%) of the total TDF number. Assuming that the average mRNA size and the number of genes of the potato cyst nematode do not differ substantially from *C. elegans*, the 8200 TDFs displayed on cDNA-AFLP gels in this study would represent ∼4000 expressed genes.

From ESTs to the corresponding TDFs on cDNA-AFLP gels

The vast majority of the virtual TDFs predicted could be located at the expected position on the cDNA-AFLP gel. The cases where no matches were found between virtual and real TDFs could usually be explained by the system used. Here, we describe detailed analyses of 52 virtual TDFs that were generated in silico using the primers E+AN in combination with all TaqI primers (T+NN) (Table 3.1). Multiple TDFs that originated from a single EST sequence were all checked. Matching bands could be found on gels for 41 TDFs. Eight virtual TDFs were smaller than the exclusion limit of 50 nt used in this study. As expected, these TDFs were not displayed. Lowering the exclusion limit would allow the display of bands down to 10 nt. Among these eight ESTs, six would produce a second virtual TDF. All these TDFs were identified on gels. Within the size range analyzed only three virtual TDFs could not be traced back on the cDNA-AFLP gels.

The TDF computed from EST GE1867 could not be detected. This EST aligned almost completely with the cloned GR-eng-2 gene from *G. rostochiensis*[57]. Careful examination of the sequence suggested that a 10 bp fragment at the 5'-end of the EST, in which a TaqI recognition site was located, may have originated from another gene. We therefore assumed that a rare recombination event occurred during construction of the cDNA library. A second band predicted for GE1867, 399 nt in length with extensions E+TT/T+TG, was readily identified on the gel.

For one particular EST, GE1782, a TaqI recognition sequence (bold) was found to be partially nested inside the EcoRI recognition site (underlined) (GAATTCGA). Contrary to the E+GA group mentioned in Table 3.1, the TCGA sequence was located at the outside of the TDF. Following the cDNA-AFLP protocol the cDNA was first digested with TaqI and, as a consequence, the EcoRI site was lost. Hence, in this particular case the predicted TDF was not amplified.

ESTs GE1349 and GE1483 were predicted to produce four TDFs. All four TDFs of GE1349 were located on gels at the predicted size. For GE1483 one band was found, the other three being smaller than the cut-off size of the gel.

In summary, from a total of 52 TDFs predicted to be produced by E+AN just one, from EST GE1133, could not be located at the predicted size and primer extensions. This minor discrepancy between the GenEST prediction and the bands displayed on the gel may be caused by a PCR or sequencing error. It is concluded that predicted TDFs from ESTs can always be traced back on cDNA-AFLP gels, when taking PCR and sequencing errors into account.

Validation of virtual TDFs by sequencing the matching bands

As has been described above, sequencing of > 100 bands excised from cDNA-AFLP gels showed that the marker-based size estimation was highly accurate. This accuracy was further confirmed by sequencing 24 bands that matched virtual TDFs. Sequencing of these matching bands revealed no inconsistencies with the computed TDFs. It is therefore concluded that three identifiers, the restriction enzyme recognition sites, the primer extensions and the size of the band, are sufficient to find the corresponding real TDFs on cDNA-AFLP gels.

Expression patterns of virtual TDFs derived from ESTs with putative housekeeping functions

We chose several ESTs with putative housekeeping functions and investigated whether we could find TDFs from these genes on cDNA-AFLP gels at the appropriate positions and with the expected constitutive expression pattern.

EST sequences GE1373, GE1659 and GE1699 share high homology (BLASTX E value $< 10^{-30}$) with elongation factor 1-β from various organisms, GE99 shares high homology (E value 10^{-35}) with 40S ribosomal protein S20 and GE1409 is likely to be a ribosomal protein L20 homolog (E value 10^{-41}). These proteins are essential components in protein synthesis and are constitutively expressed in most eukaryotic organisms. For all individual ESTs, GenEST predicted the generation of at least one TDF. Examination of cDNA-AFLP gels showed discrete bands at the right positions and virtually equal band intensities were observed in the five developmental stages. From one of the developmental stages the amplification products were cloned and sequenced. The resulting sequences were found to perfectly match the corresponding EST sequences.

These results show that the expression profiles were in accord with the predicted functions of the ESTs and that it is feasible to discard or select ESTs by analyzing the expression patterns of the predicted TDFs.

EST to cDNA-AFLP: discarding ESTs on the basis of expression profiles

For many ESTs no function could be inferred from homology searches. About 40% of the ESTs obtained from *G. rostochiensis* were categorized as unknown and many of these genes seemed to be nematode-specific. Proteins encoded by these nematode-specific genes are presumably important in nematode physiology and a few among them may be related to parasitism of host plants[55]. Examination of the expression profiles of the virtual TDFs provides valuable information on whether such ESTs deserve further investigation or not. This is exemplified by ESTs GE54 and GE1084. GE54 was predicted to produce a TDF with extensions E+TC/T+AG and a size of 138 nt and GE1084 with extensions E+AC/T+AC and a size of 85 nt. Their corresponding TDFs were readily identified on cDNA-AFLP gels. Both bands displayed a constitutive expression pattern throughout the five developmental stages. This argues against a direct function of the proteins encoded by these two genes in plant parasitism.

EST to cDNA-AFLP: selection of ESTs on the basis of expression profiles

GE1156 was predicted to produce three TDFs (E+CA/T+TT/65 nt, E+CT/T+GG/73 nt and E+GC/T+AT/82 nt). A band could be found at each of the predicted positions. The bands in the hatched J2 stage showed the highest intensity. Sequence alignment revealed that GE1156 was similar to the dorsal gland-specific gene GR-dgl-2 from the potato cyst nematode[58]. GR-dgl-2 was previously shown to be specifically expressed in PRD-hatched J2. *In situ* hybridization revealed specific expression of GR-dgl-2 in the dorsal gland of the nematode. The proteins produced by this gland may be involved in induction of a feeding site, a so-called syncytium, in the host plant[59]. The protein conceptually translated from the cDNA was predicted to be preceded by a signal peptide for secretion, indicating that this protein might be secreted by the nematode during the infection process.

GE1867 appeared to be identical to GR-eng-2, one of the β-1,4-endoglucanases that is secreted by cyst nematodes. *In situ* hybridization showed that GR-eng-2 was specifically expressed in the subventral gland[57]. Unlike GE1156, the GE1867-derived TDF (E+TT/T+TG/399 nt) showed high expression not only in the H but also in the (earlier) S stage. This points to an earlier transcription activation of subventral gland-specific genes. The proteins encoded by these genes may be important in the early infection process, namely penetration of and migration in the plant root.

cDNA-AFLP to EST: finding (near) full-length cDNAs corresponding to TDFs with relevant expression patterns

The extensions of the EcoRI primer, the extensions of the TaqI primer and the size of a band on a cDNA-AFLP gel constitute a unique identifier for a TDF. These parameters can be used to search in the EST database to find an EST that can produce such a TDF. In this way TDFs with S/H or H stage-specific expression (i.e., gene expression just prior to invasion of the plant) were used to search the list of virtual TDFs generated from the EST database.

One TDF specifically expressed at the H stage, with extensions E+CC/T+CT/ and 137 nt in length, perfectly matched the parameters of the predicted TDF from EST GE2075. This band was subsequently cloned and sequenced. The sequence showed 99% match to EST GE2075. With the help of GenEST the gene sequence representing this H stage-specific band was extended from 137 to 477 bp (Table 3.2). By sequencing the original plasmid of EST GE2075 from the 3'-end, the cDNA sequence was extended to 685 bp.

Another band displaying high expression in the S and H stages, with extensions E+AA and T+TA and a size of 251 nt, perfectly matched the predicted TDF from EST GE1816. Analysis of this EST sequence revealed that the longest open reading frame (ORF) contained 107 amino acids. This gene had no significant homology with existing genes in public databases. Use of the SignalP program[60] predicted that the protein had a cleavable signal sequence at its N-terminus that presumably targets the mature peptide for secretion. Hence, the combination of cDNA-AFLP and EST analysis has allowed us to identify this gene as worthy of further study for its potential role in nematode parasitism of plants.

These two examples illustrate another benefit of combining cDNA-AFLP and EST, which is to facilitate cloning of full-length cDNA sequences from which inter-

Table 3.2: Genes selected on the basis of a combinatorial use of EST sequences and cDNA-AFLP data from the potato cyst nematode *G. rostochiensis*. In each direction the bidirectional program GenEST allowed for selection of putative pathogenicity-related genes out of hundreds of EST sequences and expression profiles of thousands of genes.

Starting point	Corresponds to	Expression pattern on gel	Homology
E+AA/T+TA/251	GE1816	↑ in S and H	Unknown, predicted to have a signal peptide for secretion
E+CC/T+CT/137	GE2075	↑ in H	Nematode dorsal gland-specific gene GR-dgl-2

GE1156	E+CA/T+TT/65;
	E+CT/T+GG/73;
	E+GC/T+AT/82
		↑ in H (three TDFs, same pattern observed)	Nematode dorsal gland-specific gene GR-dgl-2

GE1867	E+TT/T+TG/399	↑ in S and H stages	GR-eng-2 from potato cyst nematode

esting TDFs are derived. The corresponding gene can be readily identified from the EST database even without cloning and sequencing of the TDF displayed on a gel. Once the corresponding EST is identified, obtaining a (nearly) full-length sequence is relatively simple by sequencing the entire cDNA insert from which the EST was originally derived.

In Table 3.2 four examples of putatively interesting ESTs and their corresponding TDFs are given. In online Figure 3.1 an overview of the bidirectional link between ESTs and cDNA-AFLP expression profiles established by GenEST is given.

3.3 Discussion

In this paper we present an efficient and bidirectional link between (partial) cDNA sequences and gene expression profiles as generated by cDNA-AFLP. A program called GenEST establishes this link. The added value of combining cDNA sequence information and cDNA-AFLP profiles is illustrated for one particular case, namely the search for putative pathogenicity factors from a plant parasitic nematode, *G. rostochiensis*. On the one hand, GenEST enabled us to find the expression profile of a given EST among the profiles of thousands of genes. The other way round, it allowed quick extension of TDFs by searching for the corresponding EST(s). As we have shown, the restriction enzyme recognition sites, the primer extensions and the size of the band displayed on a cDNA-AFLP gel constitute a unique set of identifiers for a TDF, the corresponding (nearly) full-length cDNA can be identified even without cloning and sequencing of the TDF of interest. In this way, the bottleneck of identifying the (near) full-length cDNAs in high throughput functional genomics studies using gel-based gene expression monitoring systems can be overcome. Since database similarity searches are more robust when using longer sequence fragments, the possibility of moving directly from a short TDF to a much longer EST may be very useful in further characterizing the putative function of a gene.

Use of GenEST for the selection of putative pathogenicity factors

Selection on the basis of expression profiles of the 228 virtual TDFs that were produced by in silico restriction of 985 ESTs with EcoRI and TaqI revealed four putative pathogenicity-related genes. One was a known gene encoding a cellulase [57]. GE2075 and GE1156 displayed strong homology with a nematode secretory gland-specific gene GR-dgl-2, indicating a possible role in the parasitism of host plants. GE1816 is a novel gene. Its function will be studied further to reveal its role in the nematode infection process. It should be noted that this is the result of a small-scale pilot experiment only. Even on this scale, the value of GenEST, which combines two high throughput technologies, is evident: four putative pathogenicity-related genes were selected out of hundreds of EST sequences and expression profiles. The applicability of this freely available tool is broad as long as the expression of genes of interest is strictly limited, either spatially or temporarily.

Further applications of GenEST

Contrary to EST approaches, the cDNA-AFLP technique is not biased towards abundant transcripts and does not involve selection on insert size. Moreover, there is no unwanted selection due to intolerance of *Escherichia coli* to a subset of the inserts. To estimate the fraction of genes not tagged by ESTs, Penn *et al.*[61] have spotted 10,000 predicted ORFs from the human genome on a cDNA array and monitored expression of these ORFs under various conditions. They concluded that potentially up to 30% of the genes in the human genome will not be discovered by an EST approach. A similar experiment could be performed by linking cDNA-AFLP and EST data with GenEST. Failure to find a good match for a TDF shown on a cDNA-AFLP gel in a large-scale EST database is informative. The corresponding gene is presumably a novel gene expressed at a low level, a small gene or a gene refractory to cloning in *E. coli*. An advantage of our approach is that ESTs and cDNA-AFLP are not linked physically, as is the case for cDNA arrays. This avoids the amplification and spotting of thousands of EST clones, saving huge logistical efforts.

Besides generating restriction patterns of sequences, GenEST can also be used to find other sequence motifs in a large data set, a process which is often too time consuming to be done manually. To illustrate this application, GenEST was used to predict the occurrence of trans-spliced leader sequences from a database composed of ∼1000 ESTs of the root knot nematode *Meloidogyne incognita*. In many nematode species up to 70% of the mature mRNAs are trans-spliced with a 22 nt leader sequence on the 5'-end of the mRNAs[62]. When the EcoRI recognition sequence in the command file is replaced by the trans-spliced leader sequence all the ESTs containing this sequence can be quickly identified using GenEST. This information can be used to estimate the fraction of full-length cDNAs present in a library and to check whether the encoded ORFs start with a peptide signal for secretion. This latter process could be further streamlined by establishing a link between GenEST and search algorithms such as SignalP.

AFLP techniques have been used extensively in genetic mapping in various organisms and a large number of AFLP markers associated with genes of interest have been identified[63, 64]. Such markers combined with a fully sequenced genome (e.g. *Arabidopsis thaliana*; [65]) could facilitate efficient cloning of target genes. To this end, GenEST can be adapted to assist in the identification of the physical locus of an interesting gene by using the identifiers of appropriate AFLP markers.

Further improvement of the EST coverage

Only 16% ($159/985 * 100\%$) of the 985 ESTs were digested in silico by EcoRI and TaqI. To increase the percentage of ESTs from which virtual TDFs are obtained a set of alternative rare cutters, including NcoI, KasI and AseI, are currently being used in combination with TaqI. With three additional primer combinations, more than half of the EST sequences [$1 - (1 - 0.16)^4 = 50.2\%$] will produce at least one virtual TDF, which could be identified on cDNA-AFLP gels. To further increase the coverage of the EST population, cDNA-AFLP can be performed with two frequent cutters. Alternatively, cDNAs could be digested with a frequent cutter only and ligated to the corresponding adapter. Subsequently, 3'-anchored cDNA-AFLP could be performed using an oligo(dT) primer in combination with the rare cutter adapter primer. This approach may be especially useful with organisms for which the entire genome has been sequenced or for which large-scale 3'-end EST sequencing has been performed. Moreover, by increasing the fraction of full-length cDNA sequences, the chance of finding at least one corresponding TDF on a gel would improve significantly.

As shown in this study, the ability to switch between sequence data and expression profiles revealed by cDNA-AFLP and vice versa is a very powerful approach to select genes for further research. This novel link provided by GenEST will be useful for functional genomics studies and is applicable to any organism where differentially expressed genes are of interest. The source code of the GenEST program is freely available.

3.4 Materials and methods

The nucleotide sequences of the ESTs described in this study are available in the GenBank EST division (dbEST) under accession nos BE607308 (GE1867), AW506364 (GE1782), AW506154 (GE1483), AW506045 (GE1349), AW505895 (GE1133), AW506065 (GE1373), AW506280 (GE1659), AW506299 (GE1699), AW505736 (GE99), AW506094 (GE1409), AW505716 (GE54), AW505855 (GE1084), AW506406 (GE1816), BE607310 (GE2075) and BE607309 (GE1156).

Generation of ESTs

The ESTs described by Popeijus *et al.*[55] were used in this study. Briefly, total RNA was extracted from infective second stage juveniles (J2) of the potato cyst nematode *G. rostochiensis* pathotype Ro1 Mierenbos freshly hatched in potato root diffusate (PRD). cDNA primed with an oligo(dT) primer was directionally cloned in the pcDNA II vector (Invitrogen, Leek, The Netherlands). The resulting library contained at least $2.5*106$ recombinant plasmids. ESTs were obtained by random sequencing of the library inserts from the 5'-end.

cDNA-AFLP profile

cDNA-AFLP profiles were generated as described by Qin *et al.*[58]. Briefly, total RNA was extracted from five developmental stages of *G. rostochiensis*: D, dehydrated unhatched J2 in cysts (in diapause); S, rehydrated unhatched J2 in 1-year-old cysts after exposure to sterile tap water for 2 days; H, pre-parasitic J2 (dry cysts incubated in sterile tap water for 1 week, then PRD for a second week); U, developing nematodes (mostly J1) in gravid females 2 months post-inoculation; P, developing nematodes (J2) in gravid females 3 months post-inoculation. cDNA was synthesized with oligo(dT)12-18 as primer. The resulting cDNAs were then digested with the restriction enzymes EcoRI and TaqI and ligated to corresponding adapters. The ligated cDNA fragments were subsequently amplified by EcoRI and TaqI primers that annealed to the adapters in PCR reactions and displayed on polyacrylamide gels.

GenEST program

A command file can be created, with a text editor, which contains restriction enzyme recognition sites to be used as the begin and end tags and the marker length modifier. Multiple combinations can be defined in a command file. GenEST uses the begin tag to search for the tag sequence in the cDNA data, which are contained in the input files in FASTA format. If such a tag is found, it will continue its search for a matching end tag. This search action is executed in both directions for all begin/end tag combinations. The marker length modifier is designed to compensate for the additional adapter sequences present in the transcript-derived fragments (TDFs) as they appear on a cDNA-AFLP gel.

Furthermore, the identifiers of a band on a gel (restriction enzyme recognition sequences, primer extensions and band size) can be used as a search query to quickly identify the corresponding EST(s) in an automated procedure.

4

Identification of imprinted genes subject to parent-of-origin specific expression in *Arabidopsis thaliana* seeds

Epigenetic regulation of gene dosage by genomic imprinting of some autosomal genes facilitates normal reproductive development in both mammals and flowering plants. While many imprinted genes have been identified and intensively studied in mammals, smaller numbers have been characterized in flowering plants, mostly in *Arabidopsis thaliana*. Identification of additional imprinted loci in flowering plants by genome-wide screening for parent-of-origin specific uniparental expression in seed tissues will facilitate our understanding of the origins and functions of imprinted genes in flowering plants.

cDNA-AFLP can detect allele-specific expression that is parent-of-origin dependent for expressed genes in which restriction site polymorphisms exist in the transcripts derived from each allele. Using a genome-wide cDNA-AFLP screen surveying allele-specific expression of 4500 transcript-derived fragments, we report the identification of 52 maternally expressed genes (MEGs) displaying parent-of-origin dependent expression patterns in *A. thaliana* siliques containing F1 hybrid seeds (3, 4 and 5 days after pollination). We identified these MEGs by developing a bioinformatics tool (GenFrag) which can directly determine the identities of transcript-derived fragments from (i) their size and (ii) which selective nucleotides were added to the primers used to generate them. Hence, GenFrag facilitates increased throughput for genome-wide cDNA-AFLP fragment analyses. The 52 MEGs we iden-

Published as ‘Identification of imprinted genes subject to parent-of-origin specific expression in *Arabidopsis thaliana* seeds’ by Peter C McKeown†, Sylvia Laouielle-Duprat†, Pjotr Prins†, Philip Wolff, Marc W Schmid, Mark Donoghue, Antoine Fort, Dorota Duszynska, Aurélie Comte, Nga Lao, Trevor Wennblom, Geert Smant, Claudia Köhler, Ueli Grossniklaus and Charles Spillane

BMC Plant Biol (2011)
Aug 12;11:113. doi:10.1186/1471-2229-11-113
†These authors contributed equally to this manuscript

tified were further filtered for high expression levels in the endosperm relative to the seed coat to identify the candidate genes most likely representing novel imprinted genes expressed in the endosperm of *A. thaliana*. Expression in seed tissues of the three top-ranked candidate genes, ATCDC48, PDE120 and MS5-like, was confirmed by Laser-Capture Microdissection and qRT-PCR analysis. Maternal-specific expression of these genes in *A. thaliana* F1 seeds was confirmed via allele-specific transcript analysis across a range of different accessions. Differentially methylated regions were identified adjacent to ATCDC48 and PDE120, which may represent candidate imprinting control regions. Finally, we demonstrate that expression levels of these three genes in vegetative tissues are MET1-dependent, while their uniparental maternal expression in the seed is not dependent on MET1.

Using a cDNA-AFLP transcriptome profiling approach, we have identified three genes, ATCDC48, PDE120 and MS5-like which represent novel maternally expressed imprinted genes in the *A. thaliana* seed. The extent of overlap between our cDNA-AFLP screen for maternally expressed imprinted genes, and other screens for imprinted and endosperm-expressed genes is discussed.

4.1 Background

Flowering plant (angiosperm) seeds are chimeric structures which contain tissues whose cells have unequal genomic contributions from the maternal and paternal parents[66–68]. Within *A. thaliana* seeds the diploid embryo is comprised of cells containing nuclear genomes inherited equally from the maternal and paternal parents. In contrast, the triploid endosperm contains two maternally inherited nuclear genomes and one paternal genome. In addition, these two fertilisation products are surrounded by a maternally derived diploid seed coat[69]. The triploid endosperm is a terminally differentiated structure which nourishes the developing embryo, while the diploid maternal seed coat plays key roles in supporting the development of the seed and the embryo it harbours[70]. The interactions between these different tissues and genomes during seed development in plants remain poorly understood[71], despite the fundamental economic importance of angiosperm seeds. For any given gene, the relative and absolute contribution of each seed tissue to overall transcript levels in the seed can be difficult to determine.

An important consequence of the unequal contributions of male and female genomes to the chimeric seed is that seed development can be affected by genome dosage and parent-of-origin effects[72]. Such maternal effects include sporophytic maternal effects from the maternally derived seed coat and gametophytic maternal effects derived from the female gametes. Gametophytic maternal effects on seed development can be due (a) to general dosage effects in the endosperm; (b) to deposition of maternal transcripts expressed prior to fertilization in the egg and central cell that give rise to the embryo and endosperm, respectively; or (c) to epigenetic regulation of genes via genomic imprinting, whereby autosomal genes are uniparentally expressed post-fertilisation in a parent-of-origin-specific manner[73, 74].

Genomic imprinting has been predominantly described in mammals and flowering plants where it occurs in nutritive tissues (endosperm, placenta) and the developing embryo, although the latter is rare in plants[75]. While there are many theories regarding the evolution of genomic imprinting in mammals and plants, some focus on imprinting arising due to a 'parental conflict' over resource allocation[76] or due to a necessity to limit gene dosage of key genes during early development[77, 78].

Many imprinted genes (i.e., hundreds, typically arranged in gene clusters along chromosomes) have been identified and intensively studied in mammalian species[79]. Until recently (2010), only 18 imprinted genes had been reported across all flowering plant species, 11 of them in *A. thaliana* (Table S1). Imprinted genes have been identified using a range of different strategies, including: mutant screens for maternally-controlled seed abortion (*A. thaliana* MEA and FIS2[80]); screens for genes regulated by the FIS Polycomb group complex (*A. thaliana* PHE1[81]); microarray analyses searching for genes showing similar responses to known imprinted genes (*A. thaliana* MPC[82]); endosperm mRNA profiling (maize nrp1[83]), and via a combination of microarray profiling and allele-specific expression analysis on endosperm from reciprocally crossed inbred lines (eight maize genes[84]). Using cdka;1 fertilized seeds which lack a paternal genome contribution to the (unfertilised) central cell, Shirzadi *et al.* (2011) used microarray profiling to identify AGL36 as a maternally expressed imprinted gene amongst the 600 genes differentially regulated in the absence of a paternal genome[85]. The advent of next generation sequencing based transcriptomics has facilitated the recent identification of additional imprinted gene candidates in *A. thaliana* seeds[86]. Hsieh et al (2011)[87] identified 43 confirmed imprinted genes (9 paternally expressed, 34 maternally expressed) in F1 hybrid seeds (7-8 days after pollination) from L*er*-0×Col-0 reciprocal crosses. Again using next generation sequencing approaches, Wolff et al (2011)[86] have identified 65 candidate imprinted genes in F1 hybrid seeds (4 days after pollination) from Bur-0×Col-0 reciprocal crosses of which 19 were confirmed in both cross directions (8 paternally expressed, and 11 maternally expressed). Hence, next generation sequencing studies are now being employed to identify putative imprinted genes[86, 87].

An indirect approach for the identification of novel imprinted genes has been conducted based on identification of differentially methylated regions (DMRs) as candidate imprinting control regions (ICRs)[88]. Genes acting as modifiers of genomic imprinting have also been identified in plants and include MET1[89], DDM1[80] and DME[90]. For example, the 5-methylcytosine DNA glycosylase gene DME is preferentially expressed in the central cell of the female gametophyte and can regulate the expression of some imprinted genes in the endosperm through demethylation of their ICRs[90]. In mutant DME endosperm ICRs remain methylated and as a result some imprinted genes are misregulated, which facilitates their detection[90].

While there are a number of genome-wide profiling approaches that can be used to identify allele-specific expression, there are several significant challenges for the definition of novel imprinted genes[91]. To distinguish between allele-specific expression effects that are either parent-of-origin dependent (e.g. imprinting) or independent, it is necessary to demonstrate the parent-of-origin dependency of uniparental expression at imprinted loci by analysis of reciprocal F1 hybrid offspring. Furthermore, where maternal-specific expression is detected in a plant seed, it is necessary to distinguish between seed coat versus endosperm (and/or embryo) expression, and also to distinguish between transcripts maternally deposited in the egg and/or central cell versus transcripts generated post-fertilisation in the developing endosperm and/or embryo[75]. While imprinted genes displaying clear mutant phenotypes (e.g. medea) on seed development can facilitate interpretation of such loci as imprinted[74], many of the imprinted genes identified to date do not display any obvious mutant phenotype in seeds[92]. In some instances, promoter:reporter constructs have been used to identify cis-regulatory regions that are required for im-printing[82, 93], while only one study has demonstrated post-fertilisation nascent uniparental *de novo* transcription of an imprinted gene in the endosperm[80].

The choice of transcript profiling platform is an important consideration for identification of novel imprinted genes. Microarrays are dependent on genes being expressed at a level sufficient to be detectable via hybridization and complementary strategies are necessary to also detect imprinted genes that may be lowly expressed. Hence, in this study we chose cDNA-AFLP[54] for genome-wide screening for novel imprinted genes. Although an early generation transcript profiling technology, as a PCR-based technology, cDNA-AFLP allows the amplification of even lowly expressed transcripts and can identify uniparentally expressed transcripts for all cases where there is a restriction site polymorphism between the parental alleles. To facilitate genome-wide cDNA-AFLP expression profiling, we have developed a gene-identifying bioinformatic software program, GenFrag, which can determine the identity of genes displaying parent-of-origin specific cDNA-AFLP expression profiles.

Our analysis of allele-specific expression of 4500 transcript-derived fragments (TDFs) in an experimental design based on the generation of reciprocal F1 hybrids seeds allowed the identification of 52 genes displaying maternal-specific expression (MEGs). The maternal specific expression of some of these MEGs may be due to genomic imprinting. Within these 52 maternally expressed genes, 18 represent genes that display higher relative and absolute expression levels in the endosperm relative to the maternal seed coat. Hence, the detection of maternal-specific expression of such genes in F1 hybrid seeds 4 days after pollination (DAP) is consistent with such genes being subject to genomic imprinting in the developing endosperm. Four of these 18 MEGs have proximal differentially methylated regions (DMRs) in seed endosperm from wild-type and DME mutant backgrounds that may represent candidate imprinting control elements (ICRs). For the three top ranked candidates (ATCDC48, PDE120 and MS5-like) we confirm maternal-specific expression in F1 hybrid seeds 4DAP and characterise the control of their allele-specific expression at different developmental stages, and in different genetic and mutant backgrounds. Overall, we have identified a range of novel MEGs in *A. thaliana* seeds, from which we further demonstrate that three are novel maternally expressed imprinted genes in *A. thaliana* seeds.

4.2 Results

cDNA-AFLP expression profiling of *A. thaliana* siliques containing F1 hybrid seeds detects 93 uniparentally-expressed TDFs

To identify genes which are uniparentally expressed in F1 hybrid seeds within siliques of *A. thaliana* we employed a genome-wide cDNA-AFLP transcriptome profiling approach. At 3, 4 and 5DAP RNA samples were generated from siliques containing F1 hybrid seeds generated via reciprocal crosses between the accessions Col-0 and L*er*-0Ṫhese three stages correspond to developmental stages from the late globular 3DAP to early and late heart stages 4 and 5DAP of embryo development within the seed. These stages of embryo development were chosen to mitigate against the possibility of detection of maternally deposited long-lived RNAs in the egg cell and/or central cell, and also because zygotic expression from both parental alleles is evident at these developmental stages[94]. In these samples, maternally expressed genes may be detected from either the silique or F1 seed tissues, and within the F1 seeds from either the maternal seed coat or the fertilisation products (i.e., the embryo and/or endosperm).

cDNA-AFLP was performed on cDNA derived from RNA samples following restriction digestion with a frequently cutting enzyme (BstYI) and a rare cutting enzyme (MseI) (Fig. S1 — supplementary data can be found online at http://www.biomedcentral.com/1471-2229/11/113/additional).

Fragments were ligated with adapters complementary to the restriction sites of the enzymes. To reduce the complexity of the mixture of fragments, a series of PCR amplifications were performed to generate subsets of fragments using selective primers. These selective primers share a common sequence, which corresponds to the adapters and a section of the restriction sites but are differentiated by one or two additional nucleotides at the 3'end, called selective nucleotides (Methods; Fig. S1).

The cDNA-AFLP generated transcript derived fragments (TDFs) were run on an ABI-3130xl capillary analyser and visualized with fluorescently labelled probes to accurately estimate their size (see Methods). A total of 10,200 TDFs were detected across the three time points 3, 4, 5DAPṪhe TDFs ranged in size from 50 to 500 nucleotide base pairs (bp) and an average of 80bp was visualized per sample. Of the 10,200 TDFs screened, 4500 showed a polymorphism between cDNA derived from the reciprocal crosses between the two different accessions (genetic backgrounds) with sizes ranging from 100bp to 500bp. Maternally expressed alleles were found in approximately equal numbers when each of the two accessions were used as the maternal parent in a reciprocal cross (Table S2). For example, at the 4DAP time-point, 366 maternally expressed Col-0 alleles were detected in the Col-0×L*er*-0 cross, while 306 maternally expressed L*er*-0 alleles were detected in the reciprocal L*er*-0×Col-0 cross. The numbers of maternally expressed TDFs detected were similar across the three developmental stages indicating consistency of maternal-specific transcription during early silique development. For each polymorphic allele, i.e., Col-0 vs L*er*-0 alleles differing in a restriction site, only one fragment is detectable from each restriction digestion event as only those TDFs proximal to the poly-A tail were isolated for analysis. Hence for each of the two accessions there is no redundancy within the number of TDFs detected at each time-point.

To identify uniparentally expressed genes, cDNA-AFLP profiles for these 4500 polymorphic TDFs were compared between those obtained from siliques containing reciprocal F1 hybrid seeds, i.e., F1 progeny of L*er*-0×Col-0 versus Col-0×L*er*-0 crosses, and those obtained from the equivalent cross between plants of the same accession, i.e., Col-0×Col-0, L*er*-0×L*er*-0. The samples at 3, 4 and 5DAP were used to filter for TDFs which displayed uniparental expression for at least two of the stages sampled. This strategy allowed the identification of 93 uniparentally expressed TDFs. All 93 of the uniparentally expressed TDFs displayed a maternal-specific expression pattern (Table S3).

Direct identification of genes based on TDF size and the selective nucleotides of each primer combination using the GenFrag bioinformatics program

To identify the genes that produced the maternal TDFs detected in *A. thaliana* siliques containing F1 hybrid seeds (Table S3), we developed a bioinformatics program called GenFrag. GenFrag is designed to allow *in silico* identification of sequences of TDFs produced by cDNA-AFLP using publicly available cDNA and EST libraries (which for the well annotated *A. thaliana* genome also includes all curated alternative splice variants[95]). Using these resources, GenFrag is designed to simulate the steps of the cDNA-AFLP *in silico* by scanning existing *A. thaliana* genome information for dual restriction enzyme cutting sites (see Methods and Fig. S1). Given the fragment size (as assessed on the capillary sequencer) and the selective nucleotides added to the primers used to generate the TDF, GenFrag can identify the corresponding sequence of a TDF and thereby the identity of the gene corresponding to the TDF. The GenFrag software is developed as open source software and is freely available for use online (see Chapter 8).

GenFrag-based identification of 52 genes from the set of 93 maternally expressed TDFs

GenFrag was used to identify genes corresponding to the 93 maternal specific TDFs (Table S3). To increase selectivity, we incorporated an option into GenFrag to only return the last matched fragment in a 5'-3' sequence i.e., the fragment closest to the poly-A tail of the mRNA. We combined this adaptation with a stringent range of 1bp deviation between the observed size of the TDF when run on the capillary analyser and the size predicted *in silico* for a candidate sequence. Using these conditions, GenFrag was able to determine unique sequence (i.e., gene ID) matches for 52 of the 93 maternally expressed TDFs identified (i.e., TDFs 1-52 in Table S3). Of the remaining TDFs, 21 matched sequences shared by more than one gene and therefore could not be uniquely distinguished (TDFs 53-73 in Table S3), while the remaining 20 could not be matched to any genes using the GenFrag approach (TDFs 74-93 in Table S3). The lack of identification of these 20 TDFs may be due to aberrant enzyme restriction and/or incomplete coverage of the *A. thaliana* transcriptome. The 52 unique sequence TDFs were matched to genes by BLAST searching the *A. thaliana* genome (TAIR). This allowed us to unambiguously identify 52 maternally expressed genes in *A. thaliana* siliques containing F1 hybrid seeds (Table T1). Gene Ontology enrichment analysis of the 52 maternally expressed genes did not reveal any significant enriched terms (data not shown). Our set of 52 MEGs did not include the known imprinted genes from *A. thaliana*, however, this is not surprising as most of these 52 MEGs have few SNP differences between the alleles from different accessions, and where they do, the SNPs do not disrupt the restriction sites that are scanned by the cDNA-AFLP technique using these restriction enzymes (Table S4). For instance, there are no Col-0/L*er*-0 SNPs in the coding sequence of the maternally expressed imprinted gene MEDEA. The 52 genes we identify represent novel maternally expressed genes (MEGs).

18 candidate imprinted genes in which the observed maternal expression is predominantly derived from higher transcript levels in the endosperm relative to the maternal seed coat

The 52 maternally expressed genes (MEGs) were detected in siliques containing reciprocal F1 hybrid seeds where the maternal-specific expression could be derived from the silique, the maternal seed coat, the endosperm and/or the embryo. Seed expressed genes which are predominantly maternally expressed in the endosperm from 3DAP (late globular stage embryos) are excellent candidates for regulation by genomic imprinting. It was recently shown that embryo development up to the globular stage does not depend on *de novo* transcription while endosperm development requires active transcription following fertilization, suggesting that maternally deposited RNAs do not play a predominant role in the endosperm[96]. Thus, mRNAs detected in the endosperm at ≥3DAP are most likely to be derived from *de novo* transcription post-fertilization. To identify which of the 52 maternally expressed genes are predominantly expressed in the endosperm at high expression levels, we used a publicly available expression dataset (Seed Gene Network - Harada-Goldberg Arabidopsis Laser Capture Microdissection Gene Chip Data Set, [97]) where the relative expression levels of genes in the seed coat and endosperm tissues (peripheral, chalazal and micropylar fractions) of seeds at the globular stage of embryo development 3DAP have been assessed.

From the 52 maternally expressed genes, we could identify 32 genes which had strong signals of expression in the 3DAP seed. Eleven genes were not detected as they did not have probes in the array dataset used or their probes also matched another gene. Nine genes were not expressed in seeds and therefore may be good candidates for silique specific MEGs. Comparing the expression levels between the endosperm and the seed coat, we found three MEGs which were exclusively expressed in the seed coat but no MEGs which were absent from the seed coat but were expressed in the endosperm. 29 MEGs showed expression in both the endosperm and the seed coat. We considered that if maternal-specific expression can be demonstrated in seeds for MEGs where the majority of the expression level signal is from the endosperm, that such a pattern would be strongly indicative of a maternally expressed imprinted gene in the endosperm. Biallelic expression in the endosperm should also be easier to detect in such cases. Hence, for these 29 MEGs, we aimed to identify genes where the majority of the expression detected in the seed is due to the endosperm fraction. We selected the 18 genes out of the 29 that showed higher expression in the endosperm compared to the seed coat and ranked these genes based on the absolute difference of expression levels between the highest expressing endosperm fraction and the seed coat (see online Table T2). We reasoned that genes displaying the highest levels of expression in the endosperm of 3DAP seeds were least likely to be genes where maternal-specific transcripts detected could be due to maternal deposition of transcripts in the central cell [96] or transferred from the maternal seed coat as has recently been proposed [87] i.e., we focussed on genes which are highly expressed in the endosperm relative to the maternal seed coat. As a complementary approach, we also compared these genes on the basis of relative transcription levels (Table S5). For these MEGs with significantly higher expression levels in the endosperm when compared to the seed coat, maternal-specific expression detected in reciprocal F1 hybrid seeds at 4DAP is consistent with regulation via genomic imprinting in the endosperm. Using these approaches, we chose the three top ranked genes as measured by total enrichment of expression in the endosperm, ATCDC48 (At3g09840), PDE120 (At5g16620) and MS5-like (At3g51280) as our strongest imprinted candidates for further investigation. Although PDE120 and MS5-like were less highly expressed in the endosperm in total, they were also the most highly ranked genes as measured by ratio of endosperm to seed coat expression (Table S5) and as noted in online Table T1 have previously been reported as preferentially endosperm-expressed in a microarray study performed by Day et al.[98]. Hence we consider all three of these MEGs to be principally expressed in the F1 endosperm relative to the maternal seed coat.

Laser capture microdissection (LCM) and qRT-PCR confirm expression of ATCDC48, PDE120 and MS5-like in *A. thaliana* seed

To validate the expression patterns of the three top ranked imprinted gene candidates ATCDC48, PDE120 and MS5-like, we used Laser Capture Microdissection (LCM) to microdissect *A. thaliana* seeds 5DAP of accession L*er*-0 into endosperm (ES), seed coat (SC) and embryo (EM) fractions. The three LCM tissues were screened by qualitative end-point RT-PCR to investigate tissue-specific expression of each gene within the seed at 5DAP which confirmed that all three genes are indeed expressed in *A. thaliana* seeds (Fig. S2). Transcripts were detected in both the seed coat and endosperm for all three genes while ATCDC48 and MS5-like were also detected in the embryo. Although this qualitative RT-PCR analysis provided no indication of relative expression levels in each of the three distinct parts of the seed, it served to independently confirm that the three genes are indeed expressed in seed tissues at 5DAP in the tissues predicted by the Seed Gene Network expression database (Table T2).

To determine how the expression levels of these genes in seeds varied over the time-course covered by our cDNA-AFLP experiment, we performed qRT-PCR on seeds at different time-points 3, 4 and 5-6 days after manual pollination. The existing data for whole-seed expression levels in Ws-0 (Seed Gene Network, [97]) predicted that expression of MS5-like and CDC48A would increase through development (across globular, heart and elongated cotyledon stages). In our qRT-PCR analysis, we found that this expression pattern was conserved in both Col-0 and L*er*-0 seeds (see online Fig. F1 A,B) indicating that for these genes there is little effect of accession background on total expression levels. We also found increased expression of PDE120 at the 5-6DAP time-point in both accessions, which differed from the Ws-0 data (Seed Gene Network) (Fig. F1 A,B).

To preclude any differences on expression levels that could be due to a hybrid background, we also measured expression of PDE120 within reciprocal Col-0×L*er*-0 crosses at the 3, 4 and 5-6DAP time-points and again found increased expression through seed development (Fig. F1 C). This suggests that the expression patterns of these three seed-expressed genes, which are similar in both parental accessions, are not significantly altered in their F1 hybrid offspring, although transcript levels of PDE120 might be slightly higher at 3DAP in the Col-0×L*er*-0 cross direction. Because expression increases throughout development, and was, in contrast, lower in pre-fertilized ovules (Fig. F1 D), this suggests that the expression we have detected is due to *de novo* post-fertilisation transcription and not maternal deposition of long-lived RNA transcripts from the central cell and/or egg cell to the post-fertilisation endosperm and/or embryo, respectively. The maternally expressed seed genes ATCDC48, PDE120 and MS5-like are subject to gene-specific imprinting in different genetic backgrounds

Genomic imprinting can be 'gene-specific' (where all alleles of the gene are imprinted in the majority of genetic backgrounds) or 'allele-specific' (where only one or a few alleles are imprinted in specific genetic backgrounds)[91]. To validate the three top-ranked genes as maternally expressed imprinted genes and to test for gene- vs allele-specific imprinting, we identified SNPs in the coding regions of each gene between the Col-0 and C24 accessions, and between the Col-0 and Bur-0 accessions. We sequenced cDNA from reciprocal F1 hybrid seeds 4DAP to detect any evidence of mono-allelic expression patterns consistent with regulation of the genes by genomic imprinting. To confirm the effects in both of the genetic backgrounds used for cDNA-AFLP, we also sequenced SNPs in cDNA from F1 hybrid seeds 4DAP of L*er*-0×Col-0 crosses for PDE120 and MS5-like. In all cases, we found that ATCDC48, PDE120 and MS5-like were maternally expressed in F1 hybrid seeds at 4DAP (see online Fig. F2; S3). While binary imprinted expression (on/off) was observed for ATCDC48 and PDE120, MS5-like displayed preferential expression of the maternally inherited allele (Fig. F2). This indicates that the imprinted status of these three genes, like their expression levels (Fig. F1), is conserved across divergent accessions and that they likely represent cases of gene-specific imprinting.

As a more general validation of the cDNA-AFLP approach to detect maternally expressed seed genes, we chose six further genes predicted to be expressed in seed tissues and sequenced SNPs in cDNA generated from Col-0×C24 and C24×Col-0 F1 hybrid seeds at 4DAPİn all six cases, we validated maternal-specific expression. We have therefore validated $9/52 = 17\%$ of the genes identified as uniparentally expressed by cDNA-AFLP as MEGs (Fig. S4).

For the top ranked imprinted gene ATCDC48, we also quantified the extent of imprinting using Quantification of Allele Specific Expression by Pyrosequencing (QUASEP), a technique based on real-time pyrophosphate (PPi) detection[54, 95, 96], which allows precise relative quantification of SNP frequencies (Fig. F3). QUASEP was performed on the maternally expressed imprinted gene ATCDC48 using cDNA collected from reciprocal Col-0×C24 F1 hybrid seeds 4DAP. The known imprinted genes FWA and PHE1 were used as controls (see online Table T3), which confirmed maternal-specific (binary) and paternal-specific (preferential) expression patterns for these two imprinted genes, respectively[89, 99]. PHE2, the non-imprinted endosperm-expressed homologue of PHE1, was used as a biallelic control (Table T3). We found that in F1 hybrid seeds at 4DAP the relative expression level from the maternally inherited allele of ATCDC48 was 100% for Col-0×C24 and 80.5% for C24×Col-0 indicating that ATCDC48 displays maternal-specific expression (Fig. F2). Although ATCDC48 is subject to expression in the seed coat, it displays high expression levels in the chalazal endosperm (Table T2), which is consistent with post-fertilisation transcription in the endosperm rather than a scenario of deposition of maternal transcripts in the central cell. Thus, the expression pattern of ATCDC48 is consistent with ATCDC48 being a novel maternally expressed imprinted gene in the endosperm of *A. thaliana* seeds.

Both ATCDC48 and MS5-like show high levels of expression in the embryo (Table T2). Biallelic expression at the heart stage of embryo development would be expected for most embryo-expressed genes, following the earlier reactivation of the paternal genome (from the globular embryo stage onwards) in *A. thaliana*[94].

In the case of MS5-like, expression within the seed is largely confined to the embryo and to the peripheral endosperm. It is likely that imprinting of MS5-like occurs exclusively within the 4DAP endosperm whilst expression in the embryo is biallelic, which could explain the partial peak of expression from the paternal allele of this gene (Fig. F2). For ATCDC48 however, the detection of almost exclusively maternal transcripts by sequencing and QUASEP could suggest that ATCDC48 may undergo delayed reactivation of the paternally inherited allele in the 4DAP embryo. Expression of imprinted genes in endosperm of seeds at later developmental stages

In a recent study, Hsieh et al. (2011)[87] screened for novel imprinted genes in 7-8DAP seed from reciprocal crosses between Col-0 and L*er*-0Ṫhe differences between the numbers of uniparental TDFs identified by cDNA-AFLP at 3, 4 and 5DAP (Table S2), with only 92 uniparental TDFs detected at multiple developmental stages, suggests some temporal dynamism in the regulation of imprinting in *A. thaliana* seeds which could potentially explain the lack of overlap between our results and those of Hsieh et al.[87]. To test this, we investigated whether the MEGs we had identified at 4DAP remained monoallelic or became biallelic at later developmental stages. Our results indicate that in cDNA from 7DAP seed, paternal alleles were more highly expressed than at 4DAP for all three of the genes (Fig. F2). In the case of ATCDC48A, this rendered the expression fully biallelic, whilst the maternal allele was still preferentially expressed for MS5-like and PDE120 (Fig. F2). At the 7DAP time-point, while all three genes are expressed from the embryo and endosperm, the relative and absolute contributions of each tissue to total transcript levels in the 7DAP seed are not known. Hence, the increased expression of the paternal allele observed in the 7DAP seed could arise from loss of imprinting and/or a shift in the relative proportion of embryo versus endosperm tissues amounts in the 7DAP seed (compared to the 4DAP seed). In the latter scenario, the MEG could remain imprinted in the endosperm tissue, but be masked by a biallelic expression signal from the more abundant embryo tissue at 7DAPṪhe expression of both alleles would be likely to preclude their identification at the $p < 0.001$ cut-off used for most gene identifications by Hsieh et al.[87]. We also considered the concordance between our dataset and a further next-generation sequencing screen performed by Wolff et al.[86] (Fig. S5) and found no overlap either with our screen or with that of Hsieh et al.[87] (see also Discussion). We also found very little overlap (seven out of 100) between imprinted genes detected by these two studies and differentially methylated regions (DMRs) previously predicted by Gehring et al.[88]. This prompted us to consider the possible existence of unidentified DMRs which could act as imprinting control regions (ICRs) associated with our imprinted genes.

Identification of DMRs at the ATCDC48, PDE120 and MS5-like loci

While the imprinting control regions (ICRs) of imprinted genes in mammals often overlap with differentially methylated regions (DMRs), the genome-wide distribution of DMRs means that only some of these are likely to be ICRs[100–103]. In plant genomes, ICRs that coincide with DMRs have been identified for the imprinted genes FWA[89, 104], PHE1[93], and MPC[82]. As noted above, however, they have not been detected for many other imprinted genes, and induction of imprinting by many putative DMRs[75] remains unconfirmed (Fig. S5). Using available methylation data for wild-type and DME endosperm[105], we searched for DMRs in the genomic vicinity of the maternally expressed imprinted loci ATCDC48, PDE120 and MS5-like.

We identified DMRs that could potentially act as ICRs for PDE120 and ATCDC48 (Fig. F4 A,B) by analysing expression data derived from endosperm of the wild type and endosperm of seeds deficient for a maternal DMEX allele[105]. These were retrieved from ArrayExpress and the percentage of methylation at cytosines situated between the genes immediately upstream and downstream of the gene bodies calculated. A DMR was located 432bp downstream of ATCDC48A containing 26 cytosines, of which 6 are hypermethylated in DME (Fig. F4 A). Four DMRs were located upstream of PDE120 at distances of 8273bp (30 cytosines, 17 hypermethylated in DME), 5377bp (49 cytosines, 6 hypermethylated in DME), 4620bp (46 cytosines, 13 hypermethylated in DME) and 3635bp (115 cytosines, 12 hypermethylated in DME) (Fig. F4 B). No obvious DME-dependent DMRs could be identified in the genomic neighbourhood of the imprinted gene MS5-like (Fig. F4 C). We also analysed our entire portfolio of candidate imprinted genes (Table T2) for potential DMRs in their vicinity. In contrast to our three top ranked imprinted genes, we could only identify DMRs for two additional genes out of the other 49, namely At1g25370 (encoding a protein of unknown function containing a DUF1639) and At2g32000 (encoding a DNA topoisomerase, type 1A) (Fig. S6). Overall, these data suggest that the imprinted MS5-like gene is less likely to be regulated via a methylation-dependent mechanism than the imprinted genes ATCDC48 and PDE120.

Expression levels of imprinted genes ATCDC48 and PDE120 are regulated by methylation pathways

In order to confirm whether DNA methylation changes are associated with altered expression levels of our novel imprinted genes, we performed qRT-PCR on cDNA derived from seedlings of met1-3 plants and found that there is a significant aberrant induction of the imprinted MEGs ATCDC48A and PDE120 in met1-3 mutants (Fig. F5 A). In concordance with the failure to detect a candidate DMR for MS5-like, no such induction occurred for this gene (Fig. F5 A). Interestingly, seeds generated by fertilising wild-type *A. thaliana* with pollen from met1-3 plants did not cause a reactivation of the paternal allele of any of the three genes (Fig. F5 B). The maternal FIS-complex has also been recently shown to regulate imprinting of certain MEGs[99, 106–108]. For the three imprinted loci of focus in this study, however, we found that fertilising fis2 plants with wild-type pollen did not lead to any loss of imprinting either (Fig. F5 B). Overall, this could imply that the proximal DMRs we have identified do not function as ICRs for these imprinted loci. Alternatively, it may suggest the existence of a subset of imprinted MEGs in which imprinted status and expression levels are regulated via a MET1- and DME/FIS-independent pathway. The lack of response of these three genes to these epigenetic modifier pathways offers a further explanation for the failure of Hsieh et al.[87] to detect ATCDC48A, MS5-like and PDE120 as imprinted MEGs, as their filtering approach compared numbers of sequence reads in wild-type crosses with those crossed to such epigenetic modifier backgrounds.

4.3 Discussion

In comparison with current knowledge of genomic imprinting, i.e., regarding number of imprinted genes and regulatory mechanisms, in mammalian genomes, the study of genomic imprinting in plants has been hindered by the low number of imprinted genes that have been reported and studied to date. In this study, we have sought to address this by identifying novel imprinted genes in the model plant *A. thaliana* and considering our results in the light of screens performed by others, and of current theories concerning the regulation of imprinting in plants.

In this study, we have conducted a genome-wide allele-specific expression analysis screen using cDNA-AFLP to identify 93 maternally expressed TDFs from a total of 4500 polymorphic allele-specific TDFs. Some of these may represent candidate maternally expressed genes regulated by imprinting in the model plant *A. thaliana*. To identify the genes represented by each TDF, we developed a novel bioinformatics software program called GenFrag which can directly identify genes in well annotated sequenced genomes, such as Col-0 accession, based only on the size of the TDF and the selective nucleotides of the primers used to generate the TDF. Although cDNA-AFLP is an early generation transcriptomics platform, as a technique it has some distinct advantages over probe hybridisation based approaches such as microarrays. These advantages include: (a) applicability to any species (including species with no genomic information), (b) low cost and reproducibility, (c) small amounts of RNA template needed, (d) detection of lowly expressed genes and (e) high specificity to distinguish closely related genes[109–112]. One of the most time-consuming steps in the cDNA-AFLP technique is the excision of TDFs from gels so that the TDF can be sequenced (typically following amplification and/or subcloning into a plasmid). To increase the throughput of gene identification in cDNA-AFLP experiments involving species with sequenced and well annotated genomes (such as *A. thaliana*), we developed the GenFrag bioinformatics software program.

There have been previous efforts to develop bioinformatic approaches to improve the efficiency of (cDNA-)AFLP techniques. The large amount of DNA sequence data available for several species has been used for *in silico* predictions of virtual transcript profiles. Tailor-made software, such as AFLPinSilico[113] and GenEST (Chapter 3, [114, 115]), allow high-throughput identification of AFLP and cDNA-AFLP TDFs for *A. thaliana* and *Globodera rostochiensis*, respectively. These *in silico* approaches were also developed to enable experiment simulations, decreasing the time needed for AFLP optimisation, and the number of samples which need to be processed[113–115]. The GenFrag program developed in this study is designed to facilitate high throughput direct identification of genes from cDNA-AFLP experiments with fully sequenced well-annotated genomes such as that of *A. thaliana*. We have made the GenFrag program freely available to the research community (see also Chapter 8).

In our study to identify novel imprinted genes in *A. thaliana*, we applied the GenFrag program to the 93 TDFs displaying a maternal-specific expression pattern, and could thereby identify 52 maternally expressed genes (MEGs) in *A. thaliana* (Table T1). By filtering for expression within seeds and enrichment within endosperm tissues, we ranked 18 MEGs on the basis of the absolute difference of their expression levels between the seed coat and the endosperm (Table T2). The identification of MS5-like and PDE120 was also supported by alternative approaches, i.e., comparison with the dataset of Day et al. ([98]; Table T1) and ranking by ratio of Endosperm/Seed Coat expression (Table S5). For any given gene expressed in the developing seed, it is difficult to separate both the absolute and relative contributions of the different seed tissues, especially given their differing ploidies (triploid in the endosperm, diploid maternal in the seed coat, diploid hybrid in the embryo) and the differences in cellular/nuclear abundance for the different tissues (seed coat, endosperm, embryo). As the contributions to total transcription are normalised against units of RNA no direct determination of the absolute contributions from each seed tissue is possible. We can demonstrate, however, that biallelic expression in the seed is detectable at the developmental stage we sample through use of a biallelic endosperm expressed gene (PHE2) as a positive control (Table T3). Our approach does have the advantage of allowing a focus on highly expressed genes, whose transcripts in seeds 4DAP are least likely to have been maternally deposited in the central cell prior to fertilisation. The endosperm is transcriptionally active immediately following fertilization, such that maternally deposited, long-lived RNAs are unlikely to play an important role[96] or be found at high levels in endosperm tissues 4DAPṪhis contrasts with the early development of the embryo, where expression in the embryo is maternally-biased (88% of transcripts at the 2-4 cell stage, for example), with paternal alleles subsequently becoming reactivated at the later globular stages of embryo development[94]. Hence, the top ranked endosperm-enriched genes identified in our study can be considered to be the most likely imprinted genes (Table T2).

A striking finding in our study is that there is little overlap in terms of genes detected between all of the different screens for imprinted genes in *A. thaliana* conducted to date, including our study (Fig. S5). Possible explanations for such lack of overlap can include (a) use of different accessions (genetic backgrounds); (b) use of samples from different developmental stages (where the relative abundance and contribution of embryo versus endosperm tissues will differ); (c) use of different filtering criteria; (d) use of different experimental approaches for isolation of seed, embryo and endosperm tissues and RNA from each tissue; and (e) use of different transcriptome profiling platforms and bioinformatic pipelines. In this study we demonstrate that the imprinted genes we have identified are unlikely to be detected at the later developmental stage used by Hsieh et al.[87], whilst the lack of overlap between the next-generation sequencing approaches of Hsieh et al. (2011) and Wolff et al.[86] is likely contributed to the analysis of different time points 7-8DAP versus 4DAP and different accessions (Col-0×L*er*-0 versus Col-0×Bur-0). There is some overlap (7 genes) between the RNA sequencing approach of[86] (Col-0×Bur-0 crosses) and a screen for genes regulated by DMRs in Col-gl X L*er*-0 crosses[88] suggesting that DMRs may control gene-specific imprinting for a limited number of loci, and/or that their ability to do so may vary according to different genetic backgrounds. Although it seems likely that all these approaches have identified imprinted genes it would seem that detection of imprinted loci (gene-specific or allele-specific) may be dependent upon accessions (genetic backgrounds), developmental stages sampled and experimental methodology. These factors may introduce significant variation between the results of different studies. Given the increasing numbers of allele-specific expression effects being detected in plants, it may be opportune for the imprinting research community to develop some common standards for the definition and validation of imprinted genes in flowering plants (see also[75]).

For the top three ranked genes ATCDC48, PDE120 and MS5-like, using LCM, we could independently detect expression of these genes in 4DAP seed tissues (seed coat, endosperm and embryo) (Fig. S2). For ATCDC48 and PDE120 we also confirmed that expression was low in pre-fertilized ovules but increased during the course of seed development (Fig. F1A, B), which is consistent with these genes being subject to post-fertilisation expression in the developing seed (i.e., not maternally deposited). We confirmed that all three of these endosperm-expressed genes are maternally expressed in 4DAP reciprocal F1 hybrid seeds from different accessions and hence represent novel cases of gene-specific imprinting in *A. thaliana* (Fig. F2 and Fig. F3). While ATCDC48 and PDE120 are subject to binary imprinted expression, MS5-like shows a preferential maternal expression pattern of imprinting[73, 84], as some paternal expression is also detected (Fig. F2). Although the expression levels of MS5-like were similar in Col-0 and L*er*-0 (Fig. F1), and in the pattern determined for Ws-0 (Seed Genes Network), the extent of imprinting did vary, with the C24 and Bur-0 alleles displaying a greater extent of imprinting when paternally inherited.

ICRs of imprinted genes often overlap with DMRs. Hence, we considered that our top-ranked imprinted genes ATCDC48, PDE120 and MS5-like might contain candidate DMRs in their genomic vicinity and that, if so, these could be candidate ICRs. We could identify DMRs upstream of PDE120 and one DMR downstream of ATCDC48 that could potentially act as ICRs (Fig. F4 A and B). The difference in methylation between wild-type and DME endosperm, however, did not reveal any DMR for MS5-like (Fig. F4 C). Expression of DME in the central cell leads to hypomethylation of the maternal genome. The methylation data used[105], however, represent the global methylation status of both the maternal and paternal genomes of the endosperm. This could explain why no DMR could be identified for MS5-like. Control of imprinting at the MS5-like locus may be independent of DNA methylation, or be regulated by a DMR far distal to the gene. Methylation-independent imprinting has been observed for some imprinted loci in mammals[116] and histone methylation by Polycomb group proteins has been shown to regulate several imprinted genes in plants[99, 106, 117]. Our results indicate that lack of MET1 in the male gamete has no effect on imprinting of ATCDC48, PDE120 and MS5-like in the developing seed. In contrast, we find that lack of MET1 leads to overexpression of ATCDC48 and PDE120 in vegetative leaf tissues. No effects of lack of MET1 in vegetative tissues were observed for MS5-like. Taking into consideration the recent findings of[86] and previous reports showing that PcG complexes regulate imprinting[99, 106–108], we also tested for possible effects of the maternal FIS-complex on regulation of the three maternally expressed imprinted genes and found that fertilising fis2 plants with wild-type pollen did not lead to any loss of imprinting. Hence, alternative epigenetic pathways are likely to regulate imprinting of MS5-like. Such regulation can neither be ruled out for ATCDC48 and PDE120. Further characterization of the imprinted ATCDC48, PDE120 and MS5-like loci will provide opportunities for increasing our understanding of the epigenetic mechanisms involved in the regulation of genomic imprinting in angiosperms.

The maternally expressed imprinted gene, ATCDC48A, is a homohexameric AAA(+) ATPase chaperone implicated in cell cycle control and cell proliferation. CDC48/p97 represents a highly conserved protein which plays a role as an initiation factor for DNA replication in many species[118] and has been shown to be essential in a wide range of multicellular and unicellular organisms [119]. In plants, the CDC48A protein has been shown to physically interact with the SOMATIC EMBRYOGENESIS RECEPTOR LIKE KINASE 1 (SERK1) protein[120, 121]. The *A. thaliana* genome contains three CDC48 loci, ATCDC48A (At3g09840), ATCDC48B (At3g53230) and ATCDC48C (At5g03340). ATCDC48A can functionally complement CDC48 mutants of Saccharomyces cerevisiae[118], and loss of the PUX1 negative regulator of ATCDC48 leads to accelerated plant growth due to increased cell division and expansion[122]. Additional studies in *A. thaliana* conducted with T-DNA knockout lines of AtCDC48A have demonstrated that homozygous null seedlings are viable until 5 days old but die shortly thereafter. It was also demonstrated that null Atcdc48a alleles have a drastically reduced transmission efficiency through the male gametophyte (i.e., ATCDC48A is essential for normal pollen germination and tube elongation)[119].

Our results indicate that ATCDC48A is maternally expressed and subject to genomic imprinting in the developing seed (endosperm) (Fig. F1, Fig. F2 and Fig. F3). Although the imprinting status of the maize homolog of ATCDC48A has not yet been determined, it is possible that imprinting of the maize homolog of ATCDC48A (or other cell-cycle genes) could be responsible for the dosage effects on cell-cycle progression observed in endosperm from interploidy crosses of maize[123]. While a clear role for ATCDC48 in the control of DNA replication in plant cells has not yet been established, our findings that ATCDC48 is a maternally expressed imprinted gene in developing endosperm resonates with a role in controlling proliferation as suggested for imprinted genes by the parental conflict theory[124].

Less is known from a functional perspective regarding the other two imprinted genes identified in this study. The MS5-like maternally expressed imprinted gene has sequence similarity to Male Sterile 5 (MS5), a gene that has been shown to be essential for male meiosis in *A. thaliana*[125]. MS5-like also displays sequence similarity with the sulphur deficiency-induced gene AtSDI1[126].

The maternally expressed imprinted gene PDE120 is annotated as a pigment defective embryo (pde) mutant in the SeedGenes database[127, 128]. The nuclear encoded PDE120 locus encodes the TIC40 protein which is a component of the protein import apparatus of the inner envelope of the chloroplast[129]. The identification of a maternally expressed imprinted nuclear gene which encodes a protein product targeted to the maternally-inherited chloroplasts could be suggestive of selection for imprinting at nuclear loci where strong control by maternally-inherited alleles of chloroplast function is essential[130].

4.4 Conclusion

In this study we have identified 52 maternally expressed genes in siliques containing reciprocal F1 hybrid seeds. We have developed and employed a novel bioinformatics tool called GenFrag to facilitate higher-throughput analysis of cDNA-AFLP experiments on organisms with well-annotated sequenced genomes. We ranked the 52 maternally expressed genes according to their relative expression levels in the endosperm versus seed coat tissues at the globular embryo stage and chose the three top-ranked imprinted candidate genes for further investigation. We confirmed expression of the three candidates in 4DAP seeds by LCM RT-PCR and further confirmed maternal-specific expression of the three genes in 4DAP F1 hybrid seeds generated with different *A. thaliana* accessions. Taken together, our results indicate that ATCDC48 is a maternally expressed imprinted gene in the developing *A. thaliana* seed, and is likely imprinted in the endosperm and perhaps the embryo. Confirmation of imprinted maternal expression was also demonstrated for the other two top-ranked genes PDE120 and MS5-like. Where present, DMRs for each of the three imprinted genes and the 18 maternally expressed genes in Table T2 were identified and posited as putative ICRs. Analysis of the imprinted ATCDC48, PDE120 and MS5-like loci with the candidate modifiers met1-3 and fis2 indicates that the regulation of imprinting at these three genes is independent of DNA methylation and the FIS-complex. Overall, our study identifies novel maternally expressed genes in *A. thaliana* seed and validates three genes (ATCDC48, PDE120 and MS5-like) as novel maternally expressed imprinted genes in *A. thaliana* seed. Further analysis of the genes identified here and by others will accelerate efforts to increase our understanding of the epigenetic regulatory mechanisms and evolution of imprinted genes in flowering plants.

4.5 Methods

Plant growth and generation of cDNA

A. thaliana L. of accessions Col-0, L*er*-0 C24 and Bur-0 were grown on 8 parts Westland multipurpose compost (Dungannon, N. Ireland): 1 part perlite: 1 part vermiculite under the following conditions: 200 μmol m-2 s-1 at 21°C/18°C and a 16:8 hr light:dark cycle. F1 hybrid seeds were generated via reciprocal crosses of Col-0 and L*er*-0 Bur-0 and C24 accessions[87, 88]. Plants were manually emasculated before anthesis and reciprocally crossed by hand under a Leica MZ6 dissecting microscope (Leica Microsystems CMS GmbH, Ernst-Leitz-StraÃČÂ§e 17-37, Wetzlar, D-35578, Germany) using Dumostar No. 5 tweezers (Dumont Biology, Switzerland). Siliques and seeds were harvested at the time points described. mRNA was extracted in combination with on-column DNase treatment using an RNase-free DNase kit (Qiagen, USA). 5 μg of total RNA was hybridized to biotinylated oligo dT which binds the streptavidin-coated PCR tube wall (mRNA Capture Kit, Roche) and cDNA synthesis performed (Quantitect Reverse Transcriptase kit, Qiagen). Quality control was performed on the Agilent 2100 Bioanalyzer (Agilent Technologies Schweiz AG, Basel, Switzerland). Samples were stored at -80°C prior to use.

cDNA-AFLP

cDNA from siliques was generated as described, digested with restriction enzymes BstYI and MseI and ligated with adapters complementary to the restriction site of BstYI and MseI. The ligated fragments were selectively amplified a first time using MseI primer and BstYI primers. The amplified fragments were diluted 1:20 and amplified a second time using 128 primer combinations (8 BstYI possible primers×16 MseI possible primers = 128 combinations). Products were run on polyacrylamide gels and visualised with the GelDoc-ItTM Imaging System (Ultra-Violet Products Ltd., Cambridge, UK). Samples were processed using the 16-capillary 3130×l Genetic Analyser (Applied Biosystems Inc.). 0.5 μl reaction products were mixed with 0.4 μl Internal Lane Standard 600 ROXTM size standard (Promega, WI, USA) or GeneScanTM 500 ROXTM size standard (Applied Biosystems, UK), in 9 μl Hi-Di Formamide (Applied Biosystems, UK). Fragments were analysed in a multiplex run and visualised with BstYI+C and BstYI+T primers, respectively labelled with the fluorescent dyes JOE and 6-FAM. Samples were analysed using the GeneMapper v3.7 software, which assigned each TDF an allelic label, or bin, based on its size as determined by comparison to the ILS600-C marker (Promega). Bin assignment permitted a variation of ±0.5bp in the determined size. For cDNA-AFLP samples generated with a given primer combination, the two parental lines, Col-0×Col-0 and L*er*-0×L*er*-0 and the two reciprocal hybrids, Col-0×L*er*-0 and L*er*-0×Col-0 were analysed together within a run to allow identification of polymorphic and differentially expressed TDFs. Fragment-sizing and allele-calling parameters for GeneMapper were normalized to the data using the default Sum-of-Signal method; alleles common between samples were not deleted. This generated electropherograms matching detected peaks with their allele calls, from which genotypes were derived.

Development of GenFrag program & software

We downloaded the two datasets containing the available full-length *A. thaliana* cDNAs from the TIGR v.4.0 (released March 2005) and TAIR databases respectively. *A. thaliana* ESTs were downloaded from the plantgdb.org website and a dataset of alternative splicing variants from the TIGR database (release June 2003).

GenFrag expands on the earlier GenEst package (Chapter 3, [115]) by providing a web interface. In addition, GenFrag provides full named support for all known restriction enzymes as listed in REBASE[131], additional support for primer combinations, their size corrections, and a listing of mismatched fragment sizes. GenFrag also allows a subset of experimental allelic fragments to be selected for analysis on the basis of the potential interest of genes in a candidate sequence list, i.e., rather than sequencing all fragments. The GenFrag software is written in Ruby, and can be run on all platforms supported by Ruby, including Windows, OSX, Linux and the Java virtual machine. The restriction enzyme module is available as part of the Open Bioinformatics Foundation BioRuby toolkit(Chapter 7, [132]) and includes all known restriction enzymes by name. Genomic information can be read in any BioRuby supported format, including FASTA. The web interface is written in Ruby on Rails, and SQLite is used for caching searches. GenFrag can be used in two ways: through a public web interface and as a software module in a computing pipeline.

Expression analysis

Microarray data of gene expression levels and absence calls from Seedgenenetwork (Harada-Goldberg Arabidopsis Laser Capture Microdissection Gene Chip Data Set) were downloaded from Gene Expression Omnibus [133], accession numbers GSM284397 and GSM284398 (seed coat), GSM284390 and GSM284391 (peripheral endosperm), GSM284388 and GSM284389 (micropylar endosperm), GSM284392, GSM284393 and GSM284394 (chalazal endosperm) and GSM284384 and GSM284385 (embryo). The developmental stage sampled by these experiments is the globular stage of embryo development. The mean expression value of all replicates was used. The following genes did not have probes: At1g12420, At1g55320, At2g45315, At3g21465, At4g01000, At4g25315, At5g04895, At5g35737 and At5g40240. Probes for At4g37530 and At1g14880 also matched another gene so were omitted from the analysis due to the possibility of ambiguous results.

Laser capture microdissection (LCM)

Siliques of emasculated and hand-crossed plants of accession L*er*-0 were collected and directly transferred to an ASP200 embedding machine (Leica Microsystems GmbH, Wetzlar, Germany) and dehydrated at room temperature in a graded ethanol series (1 hour at 70%, 3×1 hour at 90%, 3×1 hour at 99.98%) and in xylol (2×1 hour and 1×75 minutes) which was substituted by Paraplast X-tra embedding media (Roth AG, Arlesheim, Switzerland) at 56°C (2 × 1 hour, 1×3 hours), poured into paraffin blocks and stored at 4°C. Paraffin blocks were cut to 10 μm thin sections on an RM2145 microtome (Leica Microsystems GmbH, Wetzlar, Germany) and mounted on nuclease-free membranes held in metal frame slides in methanol, dried overnight at 42°C and deparaffinised in xylol at 56°C (3×10 minutes). Microdissection was performed on thin sections of siliques using the MMI CellCut Plus laser capture microscope (MMI Molecular Machines and Industries AG, Glattburg, Switzerland) to generate *circa* 150 cuts (1500 cells) per sample. Total RNA was extracted from pooled samples using the PicoPure RNA isolation kit (Arcturus Engineering, Mountain View, CA 94043-4019, USA) and single-stranded cDNA generated using the NuGEN WT-Ovation Pico RNA Amplification System (NuGEN Technologies Inc., Brockville, Canada).

RT-PCR Primers for the three top ranked candidate genes were designed using the Universal ProbeLibrary Assay Design Center (Roche, Switzerland) Identical PCR conditions were used for all genes, with Tm of 59°C and 40 amplification cycles. Two replicates were performed (data not shown), one representative result is shown for the three top ranked candidate imprinted genes analysed (Fig. S2). Quantitative RT-PCR was performed on biological triplicate samples using SYBR Green master mix (ABI) and run on a C1000 Thermal CycLer incorporating the CFX Real-Time System. Details of all primers are available on request. DNA sequencing & QUASEP Exonic SNPs between *A. thaliana* accessions were identified at TAIR[134] (PERL0437780 for ATCDC48, PERL0895299 for PDE120, PERL0626585 for MS5-like and Exon 2, 2345566 (C/T) for PHE1). cDNA from seeds of reciprocal Col-0×C24 and Col-0×L*er*-0 crosses was generated as described. Sequences surrounding the SNPs were amplified by PCR performed under standard conditions with GoTaq (Invitrogen) and sequenced by GATC. Quantification of maternally- and paternally-derived SNPs was performed via QUASEP (Quantification of Allele-Specific Expression by Pyrosequencing). RT-PCR was performed with Quantitect RT kits according to manufacturer's instructions. PCR was performed on cDNA using one biotinylated primer per pair using sequences adapted from assays designed by PSQ assay software (sequences available on request). Mean values of parental expression were calculated from at least three replicates. Genomic DNA and the genes FWA, PHE1 and PHE2 were used as controls.

Identification of DMRs

High-throughput bisulfite sequencing data of *A. thaliana* wild-type endosperm and endosperm from seeds deficient for a maternal DME allele[105] were retrieved from ArrayExpress (accession number E-GEOD-15922), corresponding to the TAIR version 8. The percentage of methylation at cytosines situated between the genes immediately upstream and downstream of our candidates was calculated. Regions that showed a difference between DME and wild-type endosperm cytosine methylation percentages were identified as DMRs and potential ICRs.

Online figures and tables

Figures and tables can be found online at
http://www.biomedcentral.com/1471-2229/11/113/additional.

Authors' contributions

PMcK designed assays, performed sequencing and pyrosequencing, and prepared the manuscript. SLD performed the cDNA-AFLP screen, analyzed TDF data, and performed pyrosequencing controls. PP contributed *in silico* cDNA-AFLP and with TJW, developed the GenFrag software. PW generated cDNA sequence traces from Col-0×Bur-0 accessions and crosses to mutant modifiers. MS performed LCM and RT-PCR experiments. MTAD analysed and compared data-sets, determined expression ratios of candidate genes, prepared figure S1 and formatted and edited the manuscript. AF and DD performed qRT-PCR. NTL generated hybrid cDNA and conducted sequencing reactions. AC identified differentially methylated regions and edited the manuscript. TJW and GS supervised the development of GenFrag. CK and UG assisted experimental design, UG supervised and funded the performance of LCM and related expression analyses. CS designed experiments, raised financing for their implementation, oversaw the experiments and development of the project, and prepared the final manuscript. All authors read and confirmed the manuscript.

5

R/qtl: high throughput Multiple QTL Mapping

R/qtl is free and powerful software for mapping and exploring quantitative trait loci (QTL). R/qtl provides a fully comprehensive range of methods for a wide range of experimental cross types. We recently added Multiple QTL Mapping (MQM) to R/qtl. MQM adds higher statistical power to detect and disentangle the effects of multiple linked and unlinked QTL compared to many other methods. MQM for R/qtl adds many new features including improved handling of missing data, analysis of 10,000s of molecular traits, permutation for determining significance thresholds for QTL and QTL hot spots, and visualizations for cis-trans and QTL interaction effects. MQM for R/qtl is the first free and open source implementation of MQM, that is multi-platform, scalable and suitable for automated procedures and large genetical genomics datasets.

R/qtl is free and open source multi-platform software for the statistical language R, and is made available under the GPLv3 license. R/qtl can be installed from http://www.rqtl.org/.

5.1 Introduction

R/qtl is an extensible, interactive environment for the mapping of quantitative trait loci (QTL, see also Chapter 6) in experimental crosses. It is implemented as an add-on package for the freely available and widely used statistical language/-

Published as 'R/qtl: high throughput Multiple QTL Mapping' by Danny Arends†, Pjotr Prins†, Ritsert C. Jansen and Karl W. Broman

Bioinformatics (2010)
26 (23), 2990-2992. 10.1093/bioinformatics/btq565
†These authors contributed equally to this manuscript

software R [135]. Since its introduction, R/qtl[136] has become a reference implementation with an extensive guide on QTL mapping [137]. R/qtl development is continuous, with input from multiple collaborators and users. We have introduced a full testing environment with regression testing, updated the license to the GPL version 3, and hosted the source code repository on Github, which gives R/qtl software development high visibility and transparency. The development of R/qtl reflects trends in quantitative genetics, in particular the use of larger datasets, larger calculations and requirements for controlling the false discovery rate (FDR). These developments are partly driven by high-throughput genetical genomics—the name coined for the study of gene expression QTL (*e*QTL)[8], metabolite QTL (*m*QTL), protein QTL (*p*QTL).

Multiple QTL Mapping (MQM) belongs to a family of QTL mapping methods, that include Haley-Knott regression[140] and composite interval mapping CIM[141]. MQM combines the strengths of generalized linear model regression with those of interval mapping[142, 143] . Recent developments in QTL mapping include Bayesian modelling of multiple QTL, e.g., R/qtlbim package[144, 145]. Bayesian modelling, however, is computationally expensive, and arguably has little additional power when applied to high density maps, and (nearly) complete genotype data[146]. Still, we intend to combine the strengths of the different methods in future versions of R/qtl.

MQM provides a practical, relevant and sensitive approach for mapping QTL in experimental populations. The theoretical framework of MQM was introduced and explored by one of us[147] and explained in the 'Handbook of Statistical Genetics'[146]. MQM has one known commercial implementation[148], which has been used effectively in practical research, resulting in hundreds of papers, e.g., in mouse, plant, and fish, respectively[149–151]. Now, with MQM for R/qtl, we present the first free and open source implementation of MQM, that is multiplatform, scalable and suitable for automated procedures and large datasets.

5.2 Features

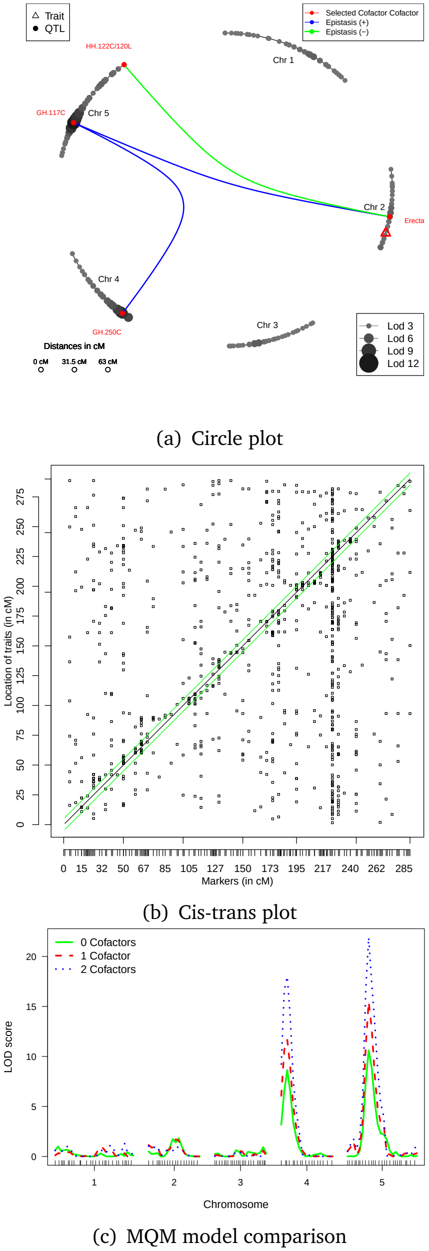
MQM for R/qtl is an **automated** three-stage procedure in which, in the first stage, missing genotype data is 'augmented'. In other words, rather than guessing one likely genotype, multiple genotypes are modelled with their estimated probabilities. In the second stage, important marker cofactors are selected by multiple regression and backward elimination. In the third stage, a QTL is moved along the chromosomes using these pre-selected markers as cofactors. QTL are interval-mapped using the most informative model through maximum likelihood. A refined and automated procedure for cases with large numbers of marker cofactors is included. The method lets users test different QTL models by elimination of non-significant cofactors. MQM for R/qtl brings the following advantages to QTL mapping: (1) Higher power, as long as the QTL explain a reasonable amount of variation; (2) Protection against over-fitting, because MQM fixes the residual variance from the full model, which allows the use of more cofactors than may be used in, for example, composite interval mapping (CIM)[141]; (3) Prevention of ghost QTL detection (between two QTL in coupling phase); and (4) Detection of negating QTL (QTL in repulsion phase).

MQM for R/qtl brings additional advantages to genetical genomics data sets with hundreds to millions of traits: (5) A pragmatic permutation strategy for controlling the FDR and prevention of locating false QTL hot spots, as discussed in Breitling et al. (2008)[152]. Marker data is permuted, while keeping the correlation structure in the trait data; (6) High-performance computing by scaling on multi-CPU computers, as well as clustered computers, by calculating phenotypes in parallel, through the Message Passing Interface (MPI) of the SNOW package for R[153]; (7) Visualizations for exploring interactions in a genomic circle plot (Fig. 5.1a) and cis- and trans-regulation (Fig. 5.1b).

A 40-page tutorial for MQM explores, both the automated procedure, and the manual procedure of adding and removing cofactors, in an *Arabidopsis thaliana* recombinant inbred line (RIL) metabolite (mQTL) dataset with 24 metabolites as phenotypes[138]. In addition, the tutorial visually explains the effects of data augmentation, cofactor selection, model selection, and tweaking of input parameters, such as cofactor significance (Fig. 5.1c). Genetic interactions (epistasis) are explored through effect plots, and an example is given of parallel computation. The tutorial is part of the software distribution of R/qtl and is available online.

5.3 Conclusion

MQM for R/qtl is a significant addition to the QTL mapper's toolbox. R/qtl provides the user with the most frequently used statistical analysis methods: single-marker analysis, interval mapping, Haley-Knott regression[140], CIM[141] and MQM[147]. MQM has improved handling of missing data and allows more powerful and precise detection of QTL, compared to many other methods. Not only is this new implementation of MQM available in the statistical R environment, which allows scripting for pipe-lined setups, it is also highly scalable through parallelisation and paves the way for high-throughput QTL analysis. With MQM, R/qtl is a free and high-performance comprehensive QTL mapping toolbox for the analysis of experimental populations. R/qtl now includes permutation strategies for determining thresholds of significance relevant for QTL and QTL hot spots; the first step towards causal inference and network analysis.



(a) Circle plot

(b) Cis-trans plot

(c) MQM model comparison

Figure 5.1: Three examples of MQM plots included in R/qtl. (a) Circular genome mQTL interaction plot of the *A. thaliana* glucosinolate pathway, metabolite data 2,000 mass peaks in 162 RILs of Arabidopsis generated from a cross between the distant accessions Landsberg erecta (L*er*-0) and Cape Verde Islands (Cvi). These individuals have been genotyped at 117 markers which are nearly evenly distributed along the genome[138]. LOD scores shown at marker positions are scaled (grey circles), with selected cofactors (red circles) and epistasis between multiple cofactors (green and blue splines). (b) Cis-trans plot of significant eQTL (squares) showing cis acting QTL (diagonal) and a trans-band (vertical, chromosome 5) in *Caenorhabditis elegans* gene expression data of 80 N2 and CB4856 RILs, hybridized on 40 two-color microarrays with 23,232 phenotypes (probes)[139]. (c) Comparison of genome-wide mQTL detection in *A. thaliana* when adding 0, 1 and 2 cofactors manually to the model, with dataset as in (a)[138]. LOD score increases when cofactors are added manually to the model. Here, adding more than two cofactors does not improve the model any further (as discussed in the online MQM tutorial).

6

Genetical Genomics for Evolutionary Studies

Genetical genomics combines high-throughput genomic data with genetic analysis. In this chapter, we review and discuss application of genetical genomics for evolutionary studies, where new high-throughput molecular technologies are combined with mapping quantitative trait loci (QTL) on the genome in segregating populations.

The recent explosion of high-throughput data — measuring thousands of proteins and metabolites, deep sequencing, chromatin, and methyl-DNA immunoprecipitation — allows the study of the genetic variation underlying quantitative phenotypes, together termed xQTL. At the same time, mining information is not getting easier. To deal with the sheer amount of information, powerful statistical tools are needed to analyze multidimensional relationships. In the context of evolutionary computational biology, a well designed experiment may help dissect a complex evolutionary trait using proven statistical methods for associating phenotypical variation with genomic locations.

Evolutionary expression QTL (eQTL) studies of the last years focus on gene expression adaptations, mapping the gene expression landscape, and, tentatively, eQTL networks. Here, we discuss the possibility of introducing an evolutionary *prior*, in the form of gene families displaying evidence of positive selection, and using that in the context of an eQTL experiment for elucidating host-pathogen protein-protein interactions. Through the example of an experimental design, we discuss the choice of xQTL platform, analysis methods, and scope of results. The resulting eQTL can be matched, resulting in putative interacting genes and their regulators. In addition, a prior may help distinguish QTL causality from reactivity, or independence of traits, by creating QTL networks.

Published as ‘Genetical Genomics for Evolutionary Studies’ by Pjotr Prins, Geert M. Smant, Ritsert C. Jansen

Methods Mol Biol (2012)
856:469-85. 10.1007/978-1-61779-585-5_19

6.1 Introduction

Genetics, as it is used here, concerns the study of quantitative, or complex, traits. A quantitative trait is influenced by multiple factors, including gene interactions and environmental factors, and typically does not lead to discrete phenotypes. Many traits of interest, such as milk production in cattle, response to fertilizer in crops and most human, animal, and plant diseases, are complex traits. Associating, or linking, complex traits with certain positions on the genome are achieved through the mapping of the so called quantitative trait loci (QTL, see also Chapter 5).

Mapping QTL in experimental populations is possible when linkage and/or association information is available. When we have a population of individuals with known genotypes, it may be possible to link a phenotype with a certain genotype. To genotype individuals, first marker maps are created. A marker is a known genomic location, where the genotype of an individual can be determined. In the early days, the genotype was determined with visible chromosome features, later with Restriction Fragment Length Polymorphism (RFLP), and Amplified Fragment Length Polymorphism (AFLP, see also Chapter 3 and Chapter 4, [154]), and, these days increasingly, with SNP/haplotype data(see also Chapter 12, [155]). Say, all individuals with genotype A, at a marker location somewhere on the genome, are susceptible to a disease and all other individuals with genotype B are not, there is linkage/association, or a QTL. If it is clear cut, it may even be a single gene effect. When it is not a single gene effect, significance statistics are required to link phenotype with genotype.

It is also possible to map QTL in natural populations through linkage disequilibrium (LD). Linkage disequilibrium occurs when certain stretches of the genome (haplotypes) show nonrandom behavior, based on allele frequencies and recombination. Associating haplotype frequencies with phenotypes potentially renders QTL. Kim *et al.* describe the genome-wide pattern of LD in a sample of 19 *Arabidopsis thaliana* accessions using SNP microarrays[156]. LD is tested, for example by Dixon *et al.*, to globally map the effect of polymorphism on gene expression in 400 children from families recruited through a proband with asthma[157].

The use of terms 'association' and 'linkage' can be confusing, even in literature. Here, we use association with haplotypes in natural populations of unrelated individuals, and linkage with markers in experimental populations. Note that in some association studies, such as Dixon *et al.*, individuals are related, i.e., some within-family linkage information is available for 400 children from 206 families.

Statistical power can be increased by using experimental crosses instead of natural populations. For example, recombinant inbred lines (RILs) are homozygous at every genomic location, simplifying genetics and increasing statistical power at the same time. For model organisms, such as *A. thaliana*, *Caenorhabditis elegans*, *Drosophila melanogaster* and *Mus musculus*, genotyped experimental crosses are available; i.e., for these species it is not always necessary to generate a new cross. Compared with natural populations, experimental crosses may introduce some bias, for example with recessive lethal alleles. Also, these individuals are rarely

100% homozygous. Finally, populations that have been maintained for some time will likely contain genotyping 'errors', mutations over generations and there is evidence that 4% line swaps can be expected due to human error[158]. Data analysis should account for such sources of bias.

Genetical genomics combines genetics with high-throughput molecular technologies. In 2001, Jansen and Nap coined the term Genetical Genomics[8] for mapping QTL in segregating populations with gene expression as a phenotype. Combining gene expression, as measured by microarray probes, with linkage leads to gene expression QTL (*e*QTL). Such *e*QTL studies elucidate how genotypic variation underlies, for example morphological phenotypes, by using gene expression levels as intermediate molecular phenotypes. In other words, the expression level as measured by a microarray probe, or probe set, is treated as a phenotype, i.e., a gene expression trait. This phenotype is associated with the genome in the form of one or more *e*QTL. With microarrays, the probe represents a known gene, and therefore genomic location. Therefore, expression phenotype and probe connect two types of genomic information: *e*QTL location(s) and gene location. It is usually assumed that *e*QTL loci represent *cis*- or *trans*-transcription regulators of the target gene[159]. If the *e*QTL is located close to the gene on the genome, the *e*QTL may point to a *cis*-regulator. If the *e*QTL is located far from the gene on the genome, the *e*QTL may point to a *trans*-regulator of a single gene or even *trans*-bands for multiple regulated genes (Chapter 5, [139]).

In a similar fashion, abundance of thousands of proteins and metabolites can be measured to map protein QTL (*p*QTL) and metabolite QTL (*m*QTL). Deep sequencing, chromatin, and methyl-DNA immunoprecipitation are just a few of the latest technologies that add to the arsenal of tools available for the study of the genetic variation underlying quantitative phenotypes. Together, *e*QTL, *m*QTL, and *p*QTL are referred to as *x*QTL. Different *x*QTL appear to confirm each other, for example, with the *A. thaliana* glucosinolate pathway[160]. Such causal inference can lead to dissecting pathways and gene networks which is an active field of research.

Evolutionary *x*QTL studies

From the perspective of evolutionary biology, genetical genomics has been applied to elucidate evolutionary adaptations of transcript regulation. For example, Fraser *et al.* introduced a test for lineage-specific selection and analyzed the directionality of microarray *e*QTL for 112 haploid segregants of a genetic cross between two strains of the budding yeast *Saccharomyces cerevisiae*; reanalyising the two-color cDNA microarray data of Brem and Kruglyak[161]. They found that hundreds of gene expression levels have been subject to lineage-specific selection. Comparing these findings with independent population genetic evidence of selective sweeps suggests that this lineage-specific selection has resulted in recent sweeps at over a hundred genes, most of which led to increased transcript levels. Fraser *et al.* suggest that adaptive evolution of gene expression is common in yeast, that regulatory adaptation can occur at the level of entire pathways, and that similar genome-wide scans may be possible in other species, including human[162].

In another *S. cerevisiae* study, Zou *et al.*, by reanalyising the same two-color cDNA microarray data, uncovered genetic regulatory network divergence between duplicate genes. They found evidence that the regulation of the ancestral gene diverged since gene duplication[163].

Li *et al.* studied plasticity of gene expression in *C. elegans*, using a set of 80 RILs generated from a cross of N2 (Bristol) and CB4856 (Hawaii), representing two genetic and ecological extremes of *C. elegans*. Differential expression induced in a RIL population by temperatures of 16 °C and 24 °C has a strong genetic component. With a group of trans-genes there was prominent evidence for a common master regulator: a trans-band of 66 coregulated genes appeared at 24 °C. The results suggest widespread genetic variation of differential expression responses to environmental impacts and demonstrate the potential of genetical genomics for mapping the molecular determinants of phenotypic plasticity[139], leading to a more generalized genetical genomics, where value is added from environmental perturbation[164].

Kliebenstein *et al.* detected significant gene network variation in 148 RILs originating from a cross between two *A. thaliana* accessions, Bay-0 and Shahdara. They were able to identify *e*QTL controlling network responses for 18 out of 20 *a priori*-defined gene networks, representing 239 genes[165].

According to Gilad, *e*QTL studies show that (i) variation in gene expression levels is both widespread and highly heritable; (ii) gene expression levels are highly amenable to genetic mapping; and (iii) most strong *e*QTL are found near the target gene, suggesting that variation in cis-regulatory elements underlies much of the observed variation in gene expression levels[166]. Meanwhile, Alberts *et al.* suggest that sequence polymorphisms may cause many false cis *e*QTL, which should be accounted for[167].

Adding a prior

QTL link complex traits with one or more locations on the genome (Fig. 6.1). Such a location is a wide measure, because a QTL is a statistical estimate, and rarely a precise indicator. On the genome, a single QTL may represent tens, hundreds, or even thousands of real genes. Combining the QTL with high-throughput technologies, such as microarrays, can add information. To zoom in on the genes underlying QTL, information from other sources can be utilized. Such *a priori* knowledge could consist of results from traditional linkage studies or association studies of, for example, human disease. That way one can assign a specific regulatory role to polymorphic sites in a genomic region known to be associated with disease[166]. Other useful priors can be existing information on gene ontology terms, metabolic pathways, and protein-protein interactions, which can be used to identify genes and pathways[168], provided these databases are sufficiently informative.

Zou *et al.*, for example, used gene ontology as a *prior* and concluded that *trans*-acting *e*QTL divergence between duplicate pairs is related to fitness defect under treatment conditions, but not with fitness under normal condition[163].

Chen *et al.* identified strong candidate genes for resistance to leaf rust in barley and on the general pathogen response pathway using a custom barley microarray on 144 doubled haploid lines of the St/Mx population[169]. 15,685 *e*QTL were mapped from 9,557 genes. Correlation analysis identified 128 genes that were correlated with resistance, of which 89 had *e*QTL colocating with the phenotypic QTL (phQTL), or classic QTL. Transcript abundance in the parents and conservation of synteny with rice prioritized six genes as candidates for Rphq11, the phQTL of largest effect[169].

Evidence of positive selection as the prior

In this Chapter 6 we discuss the steps needed to design an *x*QTL experiment to make use of genetical genomics in evolutionary studies more concrete. As the *prior* we add information on plant host genes showing evidence of positive selection.

6.2 Designing an evolutionary *x*QTL experiment

An experimental design based on genetical genomics can highlight sections of the genome showing correlation with an evolutionary trait. One such evolutionary trait of interest is plant resistance against pathogens. Plants have developed mechanisms to defend themselves against pests. When a pathogen, such as potato blight *Phytophthora infestans*, or a nematode, such as *Meloidogyne hapla*, infects a plant, it uses a battery of, so-called, effectors to help invade the plant (see also Chapter 2). Some of these effector molecules act to dissolve cellulose[170]. Intriguingly, other molecules are involved in actively reprogramming plant cells. Such plant pathogen effectors have been shown to mimic plant transcription factors[171] and switch on genes that help the pathogen[172]. A susceptible plant allows the pathogen to suppress defense mechanisms and to change cell configuration. For example, the nematodes *M. hapla* and *Globodera rostochiensis* transform plant cells, so they become elaborate feeding structures. The genetics of this plant-pathogen interaction is potentially even relevant for human medicine, as an increased understanding of host-pathogen relationships may help understand the workings of the innate immune system and helminth immunomodulation, e.g., [173, 174]. The innate immune system, through plant resistance genes (R-genes, see Box 6.2), influences susceptibility to infections in all multicellular organisms and is a much older evolutionary mechanism than the advanced adaptive immune system of higher organisms.

In this chapter we do not limit ourselves to (known) R-genes. Plants have evolved a complex array of chemical and enzymatic defenses, both constitutive and inducible, that are not involved in pathogen detection but whose effectiveness influences pathogenesis and disease resistance. The genes underlying these defenses comprise a substantial portion of the host genome. Based on genomic sequencing, it is estimated that some 14% of the 21,000 genes in *A. thaliana* are directly related to defense [175]. Most of these genes are not involved in pathogen detection, but possibly their products do molecularly interact directly with pathogen proteins or protein products. Among these proteins, for example, are chitinases and endoglucanases that attack and degrade the cell walls of pathogens, and which pathogens counterattack with inhibitors. Such systems of antagonistically interacting proteins provide the opportunity for molecular coevolution of individual systems of attack and resistance[176].

In this chapter we suggest ther design of an experiment to look for all gene families showing evidence of positive selection (see also Chapter 2). This information is the prior for *e*QTL analysis: combining known genomic locations of gene families with *e*QTL locations derived from gene expression variation in a host-pathogen interaction experiment, which hopefully results in zooming in on gene families involved in plant resistance. The prior adds statistical power in locating putative gene families involved in host-pathogen coevolution (Fig. 6.1). Note that, in this chapter, the term ‘interaction’ is used in two ways. The first is QTL interaction, where two QTL on the genome interact statistically. The second is host-pathogen gene-for-gene interaction, where gene products from different species interact physically.

Box 6.2

Plant resistance genes (R-genes) are a homologous family of genes, formed by gene duplication events and hypothesized to be involved in an evolutionary arms race with pathogen effectors. R-genes are involved in recognizing specific pathogens with cognate avirulence genes and initiating defense signaling that results in disease resistance[14]. R-genes are characterized by a molecular gene-for-gene interaction[177] in which a specific allele of a disease resistance gene recognizes an avirulence protein or pathogen allele. This specificity is often encoded, at least in part, in a relatively fast-evolving Leucine-Rich-Repeat (LRR) region[178], which consists of a varying number of LRR modules. Activation of at least some of these proteins are regulated in trans, as has been shown for RPM1 and RPS2[179].

A single *A. thaliana* plant has about 150 R-genes, representing a subset of R-genes in the overall population. The protein products of R-genes are involved in molecular interactions. They generally have a recognition site which can dock against, i.e. recognise, another one or more specific molecule(s). The proteins encoded by the largest class of R-genes carry a nucleotide-binding site LRR domain (NB-LRR, also referred to as NB-ARC-LRR and NBS-LRR). NB-LRR R-genes can be further subdivided based on their N-terminal structural features into TIR-NB-LRR, which have homology to the Drosophila Toll and mammalian interleukin-1 receptors and CC-NB-LRR, which contain a putative coiled-coil motif[180]. The LRR domain appears to mediate specificity in pathogen recognition, while the N-terminal TIR, or coiled-coil motif, is likely to play a role in downstream signaling[178]. When a molecule is docked, the R-protein is able to activate pathways in the cell, resulting in, for example, a hypersensitive response causing apoptosis and preventing spread of infection.

Meanwhile, one single R-protein only recognizes one type of invading molecules. Therefore, through its R-genes, one individual plant only recognizes a limited number of strains of invading pathogens, as the individual pathogens have variation in effectors too. When a pathogen evolves to use nonrecognized effectors, the plant becomes susceptible. The success of plant defense is determined by both evolution and the variation of specificity in a population. Unlike the evolved mammal immune system, which can change in a living organism and learn about invasions ‘on the fly’[181], plant R-genes depend on the variation inside a gene pool to provide the resistance against a pathogen; see for example Holub *et al.*[182]. Even so, many genes involved in pathogen recognition undergo rapid adaptive evolution[176], and studies have found that *A. thaliana* R-genes show evidence of positive selection, e.g., [183–185].

Create a prior with PAML

To create the prior we use Ziheng Yang's Codeml implementation of phylogenetic analysis by maximum likelihood (PAML)[17]. PAML can find amino acid sites which show evidence of positive selection using dN/dS ratios, which is the ratio of non-synonymous over synonymous substitution (ω, see Chapter 2). The calculation of maximum likelihood for multiple evolutionary models is computationally expensive, and executing PAML over an alignment of a hundred sequences may take hours, sometimes days, on a PC.

The software for generating the prior is prepackaged on BioNode, including BLAST[45], Muscle[186], pal2nal[187], PAML[17], and BioRuby[132].

It is possible to find nonoverlapping large gene families by using BLASTCLUST, a tool that is part of the BLAST tool set[45]. After fetching the *A. thaliana* cDNA sequences from the Arabidopsis Information Resource (TAIR)[188], convert the sequences to a protein BLAST database format. Based on a homology criterium, the identity score, genes are clustered into putative gene families by running BLASTCLUST with 70% amino acid sequence identity. Note that the percentage identity may not render all families, and will leave out a number of genes. It is used here for demonstration purposes only. For *A. thaliana* such a genome-wide search finds at least 60 gene families, including some R-gene families.

After aligning all family sequences, use PAML's Codeml to find evidence of positive selection in the gene families. Muscle is used to align the amino acid sequences, and create a phylogenetic tree. Next, pal2nal creates CODON alignments, which can be used by PAML. Finally run PAML's Codeml M0-M3 tests and M7-M8 tests in a computing cluster environment using, for example, BioNode and the 'rq' job scheduler.

An M0-M3 χ^2 test finds that 43 gene families (out of 60) show significant evidence of positive selection. M7-M8, meanwhile, finds 35 gene families. Therefore, based on the described procedure, approximately half the families show significant evidence of positive selection and can therefore be considered candidate gene families involved in host-pathogen interactions. Note that this figure contains false positives because the evolutionary model may be too simplistic; see also[189]. Nevertheless, these candidate gene families can be used as an effective filter for further research.

When a gene family displays evidence of positive selection, the genome locations can be used as a prior for genetical genomics (Fig. 6.1). With the full genome sequence of *A. thaliana* available, the location of gene families showing evidence of positive selection is known. For example, in the Columbia-0 (Col-0) ecotype, the majority of the 149 R-genes are combined in clusters spreading 2 to 9 loci; the remaining 40 are isolated. Clusters are organized in, so-called, superclusters[65, 180]. Phylogenetic analysis shows that such clusters are the result of both old segmental duplications and recent chromosome rearrangements[180, 190].

Select a suitable experimental population

To select a suitable experimental population, the choice of parents is key. Here, we want a descriptive evolutionary prior based on gene families with known genome locations. This means that one of the parents has to have a sequenced genome. The choice of parents for QTL analysis is normally based on large (classical) phenotypic differences. For testing pathogen resistance, the choice would ideally be one susceptible parent and one resistant (nonsusceptible) parent. For *e*QTL, the phylogenetical distance can be used, when there is no obvious phenotype. In general, it is a good idea to use common library strains based on, for example, Colombia (Col), Landsberg *erecta* (L*er*-0), Wassilewskijai (Ws), or Kashmir (Kas), as one of the parents because experimental resources and online information will be available. In addition a reference genetic background is provided in this way, which allows the comparison of the effects of QTL and mutant alleles[191]. A number of RIL populations can be found through TAIR, a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials, and community[188].

Which *x*QTL technology?

A large part of published *x*QTL studies is based on gene expression *e*QTL partly because gene expression probe provides a direct genomic link. When it comes to selecting single-color or two-color arrays one consideration may be that two-color arrays have higher efficiency when using a distant pair design[192].

Deep sequencing technology (RNA-Seq, [193]) will soon be affordable for *e*QTL studies. The main advantage over microarrays is improved signal-to-noise ratios, and possibly improved coverage depending on the reference genome (see Chapter 12). Microarrays are noisy partly due to cross-hybridization, e.g.,[194], and have limited signal on low expressors; both facts are detrimental to significance. Deep sequencing is no *panacea*, however, since it accentuates the high expressors. High expressors are expressed thousands of times higher than low expressors. Low expressors may lack significance for differential expression. Worse because deep sequencing is stochastic, many low expressors may even be absent. Another point to consider is that currently at least 1 in 1,000 nucleotide base pairs is misread, which makes it harder to disentangle error from genetical variation. Only when a sequence polymorphism is measured many times (say 20), it is confirmed to be genetical variation.

Also a choice of *e*QTL technology may take into account that, when looking at differential gene expression analysis, different microarray platforms agree with each other, but overlap between microarray and deep sequencing is much lower, suggesting a technical bias[195].

For an example of a metabolite *m*QTL study, see Keurentjes *et al.*[196] and Fu *et al.*[138]. For a study integrating *e*QTL, *p*QTL, *m*QTL and classical phenotypic QTL, see Fu *et al.*[197], and Jansen *et al.*[160].

Sizing the experimental population

The size of the experimental population should be large enough to give informative results. For classical QTL analysis, the sizing may be assisted using estimates of total environmental variance and the total genetic variance derived from the accessions, selected as parents. Roughly, population sizes of 200 RILs, without replications, will allow detection of large-effect QTL with an explained variance of 10% in confidence intervals of 10-20 cM. Detection of small-effect QTL, or mapping accuracy below 5% requires increasing the population size to at least 300 RILs[191]. It is important to see that QTL mapping accuracy is a function of both marker density and number of individuals tested. The promise of extreme dense marker maps, such as delivered by SNPs, does not automatically translate to higher accuracy. It is the number of recombination events in the population for a particular QTL that limits QTL interval size. In fact, current marker maps, in the order of thousands of (evenly spread) markers per genome, suite population sizes of a few hundred RILs. It is a fallacy, for example, to expect higher mapping power combining an ultradense SNP map with just 20 individuals.

For high-throughput *x*QTL, the experimental population should be sized against an acceptable false discovery rate (FDR). This can be achieved using a permutation strategy to assess statistical significance, maintaining the correlation of the expression traits while destroying any genetic linkages, or associations in natural populations: marker data is permuted while keeping the correlation structure in the trait data, such as presented by Breitling *et al.*[152]. Unfortunately, this information differs for every experiment and is only available afterward! Analyzing a similar experiment, using the same tissue and data acquisition technology, may give an indication[197], but when no such material is available a crude estimate may be had by taking the thresholds of a (classic) single-trait QTL experiment, and adjust that for multiple testing by the Bonferonni correction. Note that this results in a very conservative estimate.

Analyzing the *x*QTL experiment with R/qtl

R/qtl is extensible, interactive free software for the mapping of *x*QTL in experimental crosses. It is implemented as an add-on package for the widely used statistical language/software R (Chapter 5). Since its introduction, R/qtl has become a reference implementation with an extensive guide on QTL mapping[137].

R/qtl includes Multiple QTL Mapping (MQM), as described in Chapter 5 and [198], an automated procedure, which combines the strengths of generalized linear model regression with those of interval mapping. MQM can handle missing data by analyzing probable genotypes. MQM selects important marker cofactors by multiple regression and backward elimination. QTL are moved along the chromosomes using these preselected markers as cofactors. QTL are interval mapped using the most informative model through maximum likelihood. MQM for R/qtl brings the following advantages to QTL mapping: (i) higher power, as long as the

QTL explain a reasonable amount of variation; (ii) protection against overfitting, because MQM fixes the residual variance from the full model; (iii) prevention of ghost QTL detection (between two QTL in coupling phase); and (iv) detection of negating QTL (QTL in repulsion phase)[198].

MQM for R/qtl brings additional advantages to genetical genomics data sets with hundreds to millions of traits: (v) a pragmatic permutation strategy for control of the FDR and prevention of locating false QTL hot spots, as discussed above; (vi) High-performance computing by scaling on multi-CPU computers, as well as clustered computers, by calculating phenotypes in parallel, through the Message Passing Interface (MPI) of the SNOW package for R[153]; (vii) visualizations for exploring interactions in a genomic circle plot and cis- and trans-regulation (see Chapter 5 Fig. 5.1). A 40-page tutorial for MQM is part of the software distribution of R/qtl and is available online[199].

Matching the prior

After detecting *e*QTL, we have a map of gene regulation in the form of a cis-trans map. When taking *a priori* information into account, i.e., genomic locations derived through other methods, we can potentially match the genomic locations of genes and gene families with the *e*QTL cis-trans map. Until now, there has been no combined QTL and evolutionary study, involving PAML, for host-pathogen relationships in plants, though they have been conducted separately.

Combining xQTL results: causality, network inference

In addition to identifying *e*QTL or *x*QTL, it is possible to think in terms of grouping related traits by correlations. Molecular and phenotypic traits can be informative for inferring underlying molecular networks. When two traits share multiple QTL, something that is not likely to happen at random, inference of a functional relationship is possible (Fig. 6.1). Thus, distinguishing trait causality, reactivity, or independence can be based upon logic involving underlying QTL. This was the basic idea in Jansen & Nap 2001[8]. Later, people started to use the biological variation as extra source for reasoning because biological variation in trait A is propagated to B and not vice versa if A affects B. This assumes there is no hidden trait C affecting both A and B; see also Li *et al.*[200].

Mapping phenotypes for thousands of traits is the first step in attempting to reconstruct gene networks. Not only can network reconstruction be used within a particular layer, say within *e*QTL analysis, i.e., transcript data only, but also across layers. Such interlevel (system) analysis integrates transcript *e*QTL, protein *p*QTL, metabolite *m*QTL, and classical QTL[160].

The examination of pairwise correlation between traits can lead to the hypothesis of a functional relationship when that correlation is high. Beyond the detected QTL, the correlation between residuals among traits, after accounting for QTL effects, or correlations between traits conditional on other traits is further evidence for a network connection. To infer directional effects, it is necessary to analyze the correlations among pairs of traits in detail. If trait A maps to a subset of the QTL of trait B, then the common QTL can be taken as evidence for their network connection while the distinct QTL can be used to infer the direction (Fig. 6.1), unless all the common QTL have widespread pleiotropic effects, which is when a single gene influences multiple traits. If traits A and B have common QTL, without QTL that are distinct, then the inference is more complicated and further analysis is needed to discriminate pleiotropy from any of the possible orderings among traits[160, 200].

Li *et al.*[200] point out that, despite the exciting possibilities of correlation analysis, extreme caution is advised, especially in intralevel analyses, owing to the potential impact of correlated measurement error (leading to false-positive connections). By introducing a prior, however, causal inference becomes feasible for realistic population sizes[200] . The outcome of a causal inference on two traits sharing a common QTL may be either that one is causal for the other or that they are independent. In the first case, QTL-induced variation is propagated from one trait to the other while in the latter case the two traits may even be regulated by different genes or polymorphisms within the QTL region and their apparent relationship (correlation) is explained by linkage disequilibrium and not by a shared biological pathway[200].

6.3 Discussion

A QTL is a statistical property connecting genotype with phenotype. In this chapter, we reviewed studies which, with various degrees of success, combine some type of prior information with *x*QTL. We propose that a search for genome-wide evidence of positive selection can produce a valid and interesting prior for *x*QTL analysis. This is achieved by tying genomic locations of putative gene families, possibly involved in plant-pathogen interactions, with QTL locations derived from a genetical genomics experiment. Both the *e*QTL example and the search for genome-wide evidence of positive selection pressure are essentially exploratory and result in a list of putative genes, or gene families, with known genomic locations. The combined information yields candidate genes and pathways that are under positive selection pressure and, potentially, involved in host-pathogen interactions. We explain that it is possible to design an *e*QTL experiment using existing experimental populations, e.g. using an *A. thaliana* RIL population, and analyze results with existing free and open source software, such as the R/qtl tool set.

Genetical genomics bridges the study of quantitative traits with molecular biology and gives new impetus to QTL population studies. Genetic variation at multiple loci in combination with environmental factors can induce molecular or phenotypic variation. Variation may manifest itself as linear patterns among traits at different levels that can be deconstructed. Correlations can be attributed to detectable QTL and a logical framework based on common and distinct QTL and propagation of biological variation, which can be used to infer network causality, reactivity, or independence[200]. Unexplained biological variation can be used to infer direction between traits that share a common QTL and have no distinct QTL, though it may be difficult to separate biological from technical variation. Prior knowledge and complementary experiments, such as deletion mapping followed by independent gene expression studies between parental lines, may validate or disprove implicated network connections[201].

Evolutionary genetical genomics can help dissect the underlying genetics of pathogen susceptibility in plants. Where 'Evolutionary Genetics' describes how evolutionary forces shape biodiversity, as observed in nature, 'Evolutionary Genetical Genomics' describes how phenotype variation in a population is formed by genotype variation between, for example, host and pathogen involved in an evolutionary arms race.

If you want to know more about *e*QTL, we suggest the review by Gilad *et al.*[166], which also discusses *e*QTL in genome-wide association studies (GWAS), useful in situations where experimental crosses are not available (such as with many pathogens and humans). For further reading on R-gene evolution, we recommend Bakker *et al.*[178]. For R/qtl analysis, we recommend the R/qtl guide[137] and our MQM tutorial online[199]. For integrating different *x*QTL methods and causal inference, we recommend Li *et al.*[200] and Jansen *et al.*[160].

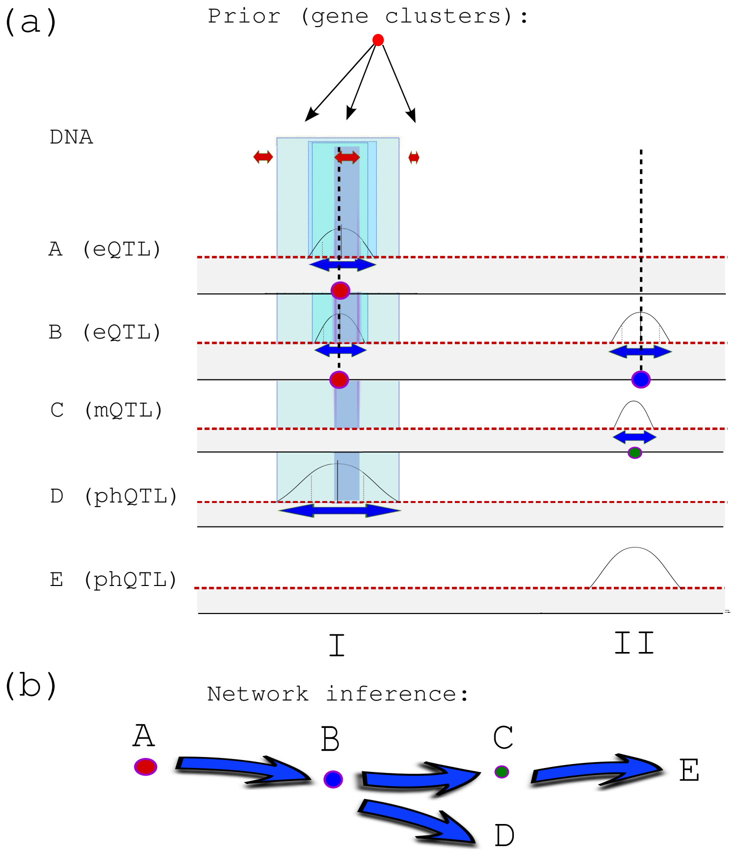


Figure 6.1: In this hypothetical and schematic example, related to mapped locations on a chromosome, prior information is combined with multiple phenotype-genotype QTL mappings to zoom in on genomic areas and to reason about causal relations between different layers of information. **(a)** The prior (red area on the chromosome) points out that certain sections are of interest; these sections consist of related genes with high homology showing evidence of positive selection, as discussed in the main text. The blue double arrow points out the confidence interval for each QTL, above the significance threshold (red dotted line). The accumulated evidence (light blue areas) leads to a narrowed down section on the genome, where in this case the prior information is the most specific. In addition, A and B point to exact gene locations (dotted line, based on exact probe information). **(b)** To infer causal relationships network inference is possible. On the left (vertical I), traits A, B, and D map to one hot spot, where A may be a regulator of B, as one QTL is shared. B causes metabolite C, again a shared QTL. Phenotype D matches A and B, and phenotype E matches A, B, and C. These causal relationships are drawn by arrows. The figure suggests that, while individual QTL are not very informative, accumulated evidence, including a prior starts to paint a picture.

7

BioRuby: Bioinformatics software for the Ruby programming language

The BioRuby software toolkit contains a comprehensive set of free development tools and libraries for bioinformatics and molecular biology, written in the Ruby programming language. BioRuby has components for sequence analysis, pathway analysis, protein modelling and phylogenetic analysis; it supports many widely used data formats and provides easy access to databases, external programs and public web services, including BLAST, KEGG, GenBank, MEDLINE and GO. BioRuby comes with a tutorial, documentation and an interactive environment, which can be used in the shell, and in the web browser.

BioRuby is free and open source software, made available under the free and open source Ruby license. BioRuby runs on all platforms that support Ruby, including Linux, Mac OS X and Windows. And, with JRuby, BioRuby runs on the Java Virtual Machine.

The source code is available from http://www.bioruby.org/

7.1 INTRODUCTION

Research in molecular biology depends critically on access to databases and web services. The BioRuby project was conceived in 2000 to provide easy access to bioinformatics resources through free and open source tools and libraries for Ruby,

Published as 'BioRuby: Bioinformatics software for the Ruby programming language' by Naohisa Goto†, Pjotr Prins†, Mitsuteru Nakao, Raoul Bonnal, Jan Aerts and Toshiaki Katayama
Bioinformatics (2010)
26(20):2617–2619. doi:10.1093/bioinformatics/btq475
†These authors contributed equally to this manuscript

```
% gem install bio
% bioruby

bioruby> puts keggapi.bfind("module citrate cycle")
md:M00009 Citrate cycle (TCA cycle, Krebs cycle)
md:M00010 Citrate cycle, first carbon oxidation
md:M00011 Citrate cycle, second carbon oxidation

bioruby> entry = keggapi.bget("md:M00010")
bioruby> md10 = flatparse(entry)

bioruby> entry = keggapi.bget("md:M00011")
bioruby> md11 = flatparse(entry)

bioruby> md10.pathways
  ==> {"ko00020"=>"Citrate cycle (TCA cycle)"}
bioruby> md11.pathways
  ==> {"ko00020"=>"Citrate cycle (TCA cycle)"}

bioruby> md10ko = md10.orthologs_as_array
bioruby> md11ko = md11.orthologs_as_array

bioruby> ko_list = md10ko + md11ko
bioruby> bg_list = ["#000099"] * md10ko.size + ["#660033"] * md11ko.size
bioruby> fg_list = ["#cccccc"] * ko_list.size

bioruby> url = keggapi.color_pathway_by_objects("path:map00020", ko_list, fg_list, bg_list)

bioruby> savefile "map00020.png", open(url).read
```

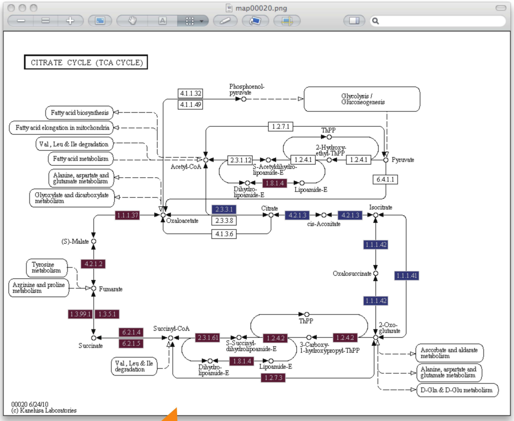


Figure 7.1: BioRuby shell example of fetching a KEGG graph using BioRuby's KEGG API[202]. After installing BioRuby the 'bioruby' command starts the interactive shell. With the *bfind* command the KEGG module database is queried for entries involved in the metabolic 'cytrate cycle', or tricarboxylic acid cycle (TCA cycle). The purple and blue colours, in input and output, reflect two modules in the carbon oxidation pathway. The user loads and confirms entries by using *flatparse* and *pathways* commands. Next, KEGG ORTHOLOGY database IDs are fetched and the colours are assigned to enzymes in each module. Finally KEGG generates the coloured image of the 'cytrate cycle' pathway and the image is saved locally.

a dynamic open source programming language with a focus on simplicity and productivity (see www.ruby-lang.org).

The BioRuby software components cover a wide range of functionality that is comparable to that offered by other Bio* projects, each targeting a different computer programming language[203], such as BioPerl[204], Biopython[205] and BioJava[206]. BioRuby software components are written in standard Ruby, so they run on all operating systems that support Ruby itself, including Linux, OS X, FreeBSD, Solaris and Windows. With JRuby, BioRuby also can run inside a Java Virtual Machine (JVM), allowing interaction with Java applications and libraries, like Cytoscape for visualisation[207].

Both BioRuby and Ruby are used in bioinformatics for scripting[208], scripting against applications[209], modelling[210, 211], analysis[212], visualisation and service inte-gration[213].

The web development framework ‘Ruby on Rails’ is used to create web applications and web services[214, 215]. BioRuby provides connection functionality for major web services, such as the Kyoto Encyclopedia of Genes and Genomes (KEGG; see example in Fig. 7.1)[202], and the TogoWS service, which provides a uniform web service front-end for the major bioinformatics databases[216].

The BioRuby source tree contains over 580 documented classes, 2800 public methods and 20,000 unit test assertions. Source code is kept under Git version control, which allows anyone to clone the source tree and start submitting. We have found that Git substantially lowers the barrier for new people to start contributing to the project. In the last two years the source tree has gained 100 people tracking changes and 32 people cloned the repository.

The BioRuby project is part of the Open Bioinformatics Foundation (OBF), which hosts the project website and mailing list, and organises the annual Bioinformatics Open Source Conference (BOSC) together with the other Bio* projects. A bignum of BioRuby features support Bio* cross-project standards, such as the BioSQL relational model for interoperable storage of certain data objects, or their implementation is coordinated across the Bio* projects, including support for the FASTQ[217] and phyloXML[218] data exchange formats.

7.2 FEATURES

BioRuby covers a wide range of functional areas which have been logically divided into separate *modules* (Table 7.1).

Table 7.1: BioRuby Modules

Category	Module list
Object	Sequence, pathway, tree, bibliography reference
Sequence	Manipulation, translation, alignment, location, mapping, feature table, molecular weight, design siRNA, restriction enzyme
Format	GenBank, EMBL, UniProt, KEGG, PDB, MEDLINE, REBASE, FASTA, FASTQ, GFF, MSF, ABIF, SCF, GCG, Lasergene, GEO SOFT, Gene Ontology
Tool	BLAST, FASTA, EMBOSS, HMMER, InterProScan, GenScan, BLAT, Sim4, Spidey, MEME, ClustalW, MUSCLE, MAFFT, T-Coffee, ProbCons
Phylogeny	PHYLIP, PAML, phyloXML, NEXUS, Newick
Web service	NCBI, EBI, DDBJ, KEGG, TogoWS, PSORT, TargetP, PTS1, SOSUI, TMHMM
ODBA	BioSQL, BioFetch, indexed flat files

BioRuby allows accessing a comprehensive range of public bioinformatics resources. For example, BioRuby supports the Open Biological Database Access (ODBA) as a generic and standardised way of accessing *biological data sources*. In addition BioRuby can directly process local database files in a variety of different flat file formats, including FASTA, FASTQ (Fig. 7.2), GenBank and PDB. BioRuby also allows querying and accessing remote online resources through their interfaces for programmatic access, such as those provided by KEGG, the DNA Databank of Japan (DDBJ), the National Center for Biotechnology Information (NCBI), and the European Bioinformatics Institute (EBI).

BioRuby has online documentation, tutorials and code examples. It is straightforward to get started with BioRuby and use it to replace, or glue together, legacy shell scripts, or to mix Ruby on Rails into an existing web application.

BioRuby comes with an interactive environment, both for the command-line shell and in the browser. Ideas can be quickly prototyped in the interactive environment, and can be saved as ‘scripts’ for later use. Such an interactive environment has shown to be especially useful for bioinformatics training and teaching (Fig. 7.1).

New features, and refinements of existing ones, are constantly being added to the BioRuby code base. Current development activity focuses on adding support for the semantic web, and on designing a plugin system that allows adding entirely new components in a loosely coupled manner, such that experimental new code can be developed without having an impact on BioRuby’s core stability and portability.

```
require ’bio’
quality_threshold = 60
Bio::FlatFile.open(’sample.fastq’).each do |entry|
    hq_seq = entry.mask(quality_threshold)
    puts hqseq.output_fasta(entry.entry_id)
end
```

Figure 7.2: BioRuby example of masking sequences from next generation sequencing data in FASTQ format using a defined quality_threshold, and writing the results in FASTA format.

7.3 CONCLUSION

The BioRuby software toolkit provides a broad range of functionality for molecular biology and easy access to bioinformatics resources. BioRuby is written in Ruby, a dynamic programming language with a focus on simplicity and productivity, which targets all popular operating systems and the JVM. The BioRuby project is an international and vibrant collaborative software initiative that delivers life-science programming resources for those researchers who want to benefit from the productivity features of the Ruby language, as well as from the larger Ruby ecosystem of reusable open source components.

8

BioGem: an effective tool based approach for scaling up open source software development in bioinformatics

BioGem provides a software development environment for the Ruby programming language, which encourages community-based software development for bioinformatics while lowering the barrier to entry and encouraging best practices. BioGem, with its targeted modular and decentralized approach, software generator, tools, and tight web integration, is an improved general model for scaling up collaborative open source software development in bioinformatics.

BioGem and modules are free and open source software. BioGem runs on all systems that support recent versions of Ruby, including Linux, Mac OS X and Windows. Further information at www.biogems.info. A tutorial is available at www.biogems.info/howto.html

8.1 Introduction

In biomedical science, new technologies, data formats, and methods emerge continuously. Scientists want to take advantage of these developments as soon as possible, which requires bioinformatics software to keep up with new require-

Published as ‘BioGem: an effective tool based approach for scaling up open source software development in bioinformatics’ by Raoul J.P. Bonnal†Jan Aerts, George Githinji, Naohisa Goto, Dan MacLean, Chase A Miller, Hiroyuki Mishima, Massimiliano Pagani, Ricardo Ramirez-Gonzalez, Geert Smant, Francesco Strozzi, Rob Syme, Rutger Vos, Trevor J Wennblom, Ben J Woodcroft, Toshiaki Katayama†, and Pjotr Prins†

Bioinformatics (2012)
28(7):1035–1037, doi:2012.10.1093/bioinformatics/bts080
†These authors contributed equally to this manuscript

ments. We support the notion of the Open Bioinformatics Foundation (OBF) that development of collaborative open source software (OSS) is essential for bioinformatics. The OBF represents a number of important projects, such as BioPerl[204], Biopython[205], BioRuby[132], and BioJava[219]. These Bio-star (Bio*) projects effectively function as community centres and share a centralised approach in software development with large source code repositories. Bio* projects, generally, aim for consolidated tools, a stable application programming interface (API), and backwards compatibility.

Within the BioRuby project we experienced the drive for stability easily overwhelmed and discouraged developers. Not only because of the complexity of the existing code base, but also because coding standards are enforced, and extensive tests and documentation are required. Furthermore, newly contributed code may be subject to community scrutiny, and in many cases further demands for improving the code follow. The full process introduces a significant delay between initial idea and final acceptance of the code in the main project. Months, even years, may pass between stable releases of main Bio* projects. It may take a long time before a new feature is publicly released.

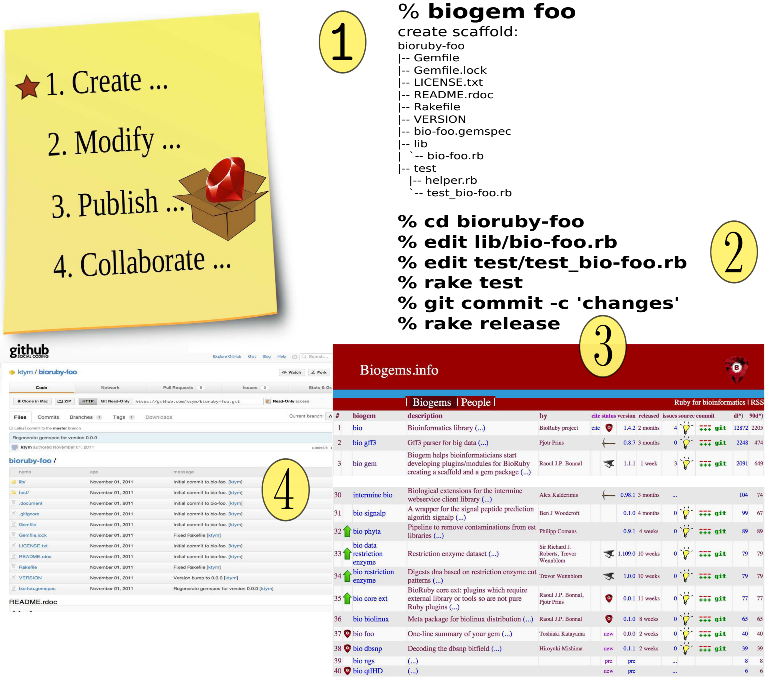
To scale up collaborative software development in BioRuby, we recognised existing and new developers need to be encouraged to contribute more code. To achieve this, we created BioGem a Ruby application framework for rapid creation of decentralised, internet published software modules written to lower the barrier to entry. BioGem was initially inspired by the R/Bioconductor packaging system[220], which encourages software developers to publish software modules independently using simple rules; and Ruby on Rails (RoR) plugins[221], which provides a software generator and modular software plugin system.

8.2 Features

For BioGem we created specific tools to support the creation of bioinformatics software functionalities and to support development 'best practises', i.e., infrastructure for software specification, documentation and tests. We also provide tight web integration based on public websites and services. These websites publish and distribute software modules and give web based access to source code, complete with revision history (see Fig. 8.1). BioGem exposes Ruby bioinformatics modules, and makes developer productivity and module popularity visible.

The primary tool of the BioGem framework is a software generator consisting of templates for bioinformatics scripts, source code, software specification, documentation, and tests. With the generator, required directories and files are automatically created from templates for a new software module. Templates are included for commonly encountered tasks, such as command line parameter handling, error handling, make files etc.

Another BioGem tool publishes the versioned module with its dependencies on the internet. The published module is immediately available for download and



```
% biogem foo
create scaffold:
bioruby-foo
|-- Gemfile
|-- Gemfile.lock
|-- LICENSE.txt
|-- README.rdoc
|-- Rakefile
|-- VERSION
|-- bio-foo.gemspec
|-- lib
|   `-- bio-foo.rb
|-- test
    |-- helper.rb
    `-- test_bio-foo.rb

% cd bioruby-foo
% edit lib/bio-foo.rb
% edit test/test_bio-foo.rb
% rake test
% git commit -c 'changes'
% rake release
```

https://github.com/ktym/bioruby-foo.git

Latest commit to the **master** branch

Regenerate gemspec for version 0.0.0

ktym authored November 01, 2011

bioruby-foo /

name	age	message
lib/	November 01, 2011	Initial commit to bio-foo. [ktym]
test/	November 01, 2011	Initial commit to bio-foo. [ktym]
.document	November 01, 2011	Initial commit to bio-foo. [ktym]
.gitignore	November 01, 2011	Initial commit to bio-foo. [ktym]
Gemfile	November 01, 2011	Initial commit to bio-foo. [ktym]
Gemfile.lock	November 01, 2011	Fixed Rakefile [ktym]
LICENSE.txt	November 01, 2011	Initial commit to bio-foo. [ktym]
README.rdoc	November 01, 2011	Initial commit to bio-foo. [ktym]
Rakefile	November 01, 2011	Fixed Rakefile [ktym]
VERSION	November 01, 2011	Version bump to 0.0.0 [ktym]
bio-foo.gemspec	November 01, 2011	Regenerate gemspec for version 0.0.0 [ktym]

README.rdoc

#	biogem	description	by	cite	status	version	released	issues	source	commit	dl*)	90d*)
1	bio	Bioinformatics library (...)	BioRuby project	cite		1.4.2	2 months	4		git	12872	2205
2	bio gff3	Gff3 parser for big data (...)	Pjotr Prins			0.8.7	3 months	0		git	2248	474
3	bio gem	Biogem helps bioinformaticians start developing plugins/modules for BioRuby creating a scaffold and a gem package (...)	Raoul J.P. Bonnal			1.1.1	1 week	3		git	2091	649
30	intermine bio	Biological extensions for the intermine webservice client library (...)	Alex Kalderimis			0.98.1	3 months	...			104	74
31	bio signalp	A wrapper for the signal peptide prediction algorith signalp (...)	Ben J Woodcroft			0.1.0	4 months	0		git	99	67
32	bio phyta	Pipeline to remove contaminations from est libraries (...)	Philipp Comans			0.9.1	4 weeks	0		git	89	89
33	bio data restriction enzyme	Restriction enzyme dataset (...)	Sir Richard J. Roberts, Trevor Wennblom			1.109.0	10 weeks	0		git	79	79
34	bio restriction enzyme	Digests dna based on restriction enzyme cut patterns (...)	Trevor Wennblom			1.0.0	10 weeks	0		git	79	79
35	bio core ext	BioRuby core ext: plugins which require external library or tools so are not pure Ruby plugins (...)	Raoul J.P. Bonnal, Pjotr Prins			0.0.1	11 weeks	0		git	77	77
36	bio biolinux	Meta package for biolinux distribution (...)	Raoul J.P. Bonnal			0.1.0	8 weeks	0		git	65	65
37	bio foo	One-line summary of your gem (...)	Toshiaki Katayama		new	0.0.0	2 weeks	0		git	40	40
38	bio dbsnp	Decoding the dbsnp bitfield (...)	Hiroyuki Mishima		new	0.1.1	2 weeks	0		git	39	39
39	bio ngs	(...)			pre	pre		...			8	8
40	bio qtlHD	(...)			new	pre		...			6	6

Figure 8.1: Biogem eases publication of new bioinformatics Ruby software modules on the Internet, in a few steps. (1) The software generator creates the directory layout and files for a new software module named 'foo'. (2) The developer writes or modifies source code, and (3) quickly and easily publishes the source code and module online, for others to read, install and use. Collaboration (4) is facilitated by publishing source code and changes to navigationable websites. Then the workflow continues again at (2). The http://biogems.info website tracks published modules. Popularity of each published module is tracked, as well as source code changes, updates, bugs, and issues. Unlike with the practise of publishing scientific papers, collaboration on software often comes *post factum*, i.e. after original publishing of a software module. Therefore it pays to publish software modules early and often. This is reflected in the Biogem workflow.

installation to bioinformatics users in the form of a Ruby gem (i.e., an archive of modular Ruby code with all the supporting files and information needed for installation by 'package manager' software). We refer to a BioGem module as a 'BioRuby plugin' if the module extends the BioRuby project. Published software modules are easily repackaged by software distributions, e.g., Debian Bio Med[222] and BioLinux[223].

The BioGem website (see abstract) makes it easy to find and install software modules. The website also allows people to track releases, software dependencies, development activity, outstanding issues, integration test results, documentation and popularity of published modules. A map shows the location of Biogem developers to help foster a sense of international community.

BioGem encourages software development best practices by providing templates for documentation and multiple test driven development strategies; such as unit tests, behaviour driven development, and a natural language parser for software specification[224]. A notable difference to the traditional code contribution procedures of the Bio* projects is that best practices are encouraged, rather than enforced.

Templates are also included for certain types of functionality, e.g., to generate portable SQL database handlers, and to build a dynamic web site. With BioGem it is possible to create a functional web application, or service, in just a few steps. Generating the different features is handled through work flows (Fig. 8.1).

We added tutorials for BioGem, which explain the software generators, templates and software publishing. These tutorials are part of the software distribution and available online.

We created 'collections' that bundle important modules together as specific releases. For example, 'bio-core' contains stable modules, and 'bio-core-ext' contains stable modules with bindings to C libraries. Special purpose collections exist such as 'bio-biolinux', which is distributed by the Cloud Biolinux project and merged with the Galaxy CloudMan project [225].

In the first eight months of the BioGem functionality becoming available, over twenty new modules have been published through BioGem, showing a wide variety of subjects. These modules, for example, target big data handling, next generation sequencing, and parsing of bioinformatics data formats (Table 1).

8.3 Conclusion

BioGem provides an environment for rapid bioinformatics software development with a low barrier to entry. BioGem frees potential contributors from code maturity expectations that can be deterring, and encourages Ruby developers to contribute experimental source code early to the BioRuby community. Through BioGem software is published in a modular way, and best practises are encouraged through infrastructure for software specification and testing. All this results in better utilisation of existing and new software development manpower, thereby scaling up open source software development in bioinformatics.

We suggest BioGem can serve as a generic model; not by replacing existing Bio* projects, but by supplementing them with a decentralised and evolutionary model for collaborative software development.

Table 8.1: The introduction of BioGem has led to a broad range of new BioRuby plugins. An up-to-date list can be found at http://biogems.info

Name	Description
bio assembly	read and write assembly data
bio blastxmlparser	fast, low memory, big data BLAST parser
bio bwa	Burrows Wheeler aligner
bio cnls scraper	nuclear localisation signal prediction
bio six frame	sequence translation
bio genomic interval	detect intervals
bio gff3	fast, low memory, big data GFF3 parser
bio isoelectric point	calculate protein isoelectric point
bio kb illumina	Illumina annotations
bio lazyblastxml	another BLAST XML parser
bio logger	sane error handling
bio nexml	NeXML support, for phylogenetic data
bio ngs	NGS workflows and display, incl. support for bio bwa, Bowtie, TopHat, and Cufflinks
bio octopus	transmembrane domain predictor interface
bio restriction enzyme	DNA cutting operations with REBASE
bio samtools	samtools API
bio signalp	signal peptide prediction interface
bio sge	split huge files for cluster computing
bio tm hmm	transmembrane predictor interface
bio ucsc api	UCSC Genome Database binding

9 Sambamba: fast processing of NGS alignment formats

Sambamba is a high performance robust tool and library for working with SAM, BAM and CRAM sequence alignment files; the most common file formats for aligned next generation sequencing (NGS) data. Sambamba is a faster alternative to samtools that exploits multi-core processing and dramatically reduces processing time. Sambamba is being adopted at sequencing centers, not only because of its speed, but also because of additional functionality, including coverage analysis and powerful filtering capability.

9.1 Introduction

Processing speed matters, not only for diagnostics, but also for analysis and sharing of computational resources. NGS is increasingly used as a genetic screening tool in diagnostics[226] and reducing time from sample intake to test result/diagnosis potentially saves lives. Introducing multi-core processing can accelerate steps in a pipeline when the CPU is the bottleneck[227].

Since its introduction by the 1000 Genomes Project[228], the sequence alignment/map format (SAM) and its compressed binary counterpart (BAM) have become the *de facto* file formats used for storing and distributing NGS data. Samtools is the original tool for SAM/BAM files processing, including data extraction and filtering[229]. Recently samtools added the CRAM format as a compressed alternative to SAM/BAM[230]. While samtools exploits the speed of the low-level

Published as 'Sambamba: fast processing of NGS alignment formats' by Artem Tarasov, Albert J. Vilella, Edwin Cuppen, Isaac J. Nijman and Pjotr Prins
Bioinformatics (2015)
31(12):2032–2034. doi:10.1093/bioinformatics/btv098

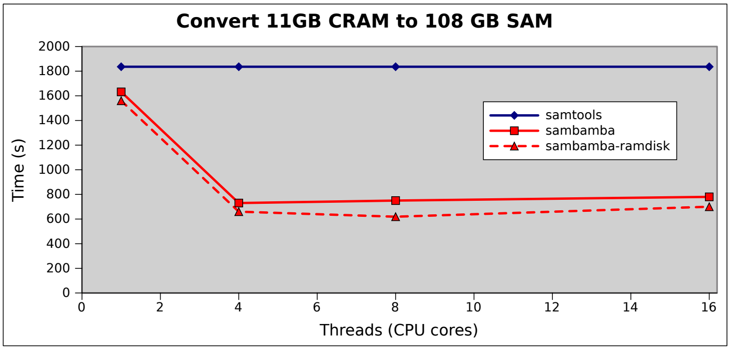


Figure 9.1: Processing speed comparison of samtools and sambamba. Wall-clock time (s) versus number of threads to convert a 11 GB CRAM (1000genomes HG00110) to 108 GB SAM. With Samtools, VIEW is bound to a single thread at CPU 90%. With Sambamba, IO gets saturated at approx. CPU 250%. When using a faster RAM-disk, IO gets saturated at approx. CPU 350%. For samtools a RAM-disk makes no difference. When adding more threads, performance reproducibly degrades because of CPU cache contention. All timings were performed on a server-class machine with 512 GB of RAM and 48 CPU cores (4×12-core AMD Opteron(tm) Processor 6174 @2.2Ghz with 6Mb L2 cache) Samtools version v1.0-15 using htslib v1.0-1 and sambamba v0.5.0 compiled with the LLVM D-compiler v0.14.0.

C programming language and uses streamed data for efficiency, it has limited support for parallel processing (Fig. 9.1). Samtools has inspired a number of other BAM processors, notably Picard[231], samblaster[232], biobambam[233], and Scramble[234], each of which is either slower than samtools, or offers a subset of its functionality.

To accelerate analysis pipelines we created sambamba, a new incarnation of samtools that fully utilises parallel processing. Sambamba (which means ‘parallel’ in Swahili) is written in the D programming language, a modern programming language with run-time performance similar to that of C[235]. D has powerful abstractions for parallel computing which make it possible to scale computations with the number of cores (Fig. 9.1). When running a Human cancer exome SNV calling pipeline on the results of a single Illumina HiSeq 2500 flowcell in fast mode (2000 genes, 300 million reads, 100bp read length and average read depth of 100 for 6 samples) following standard best practice guidelines, the bioinformatics processing time was reduced from 2 hours to 30 minutes by replacing Picard MARKDUP and samtools INDEX, FLAGSTAT, MERGE and VIEW.

9.2 Features

Sambamba introduces full parallelised data processing of SAM, BAM and CRAM files. Sambamba primarily uses D’s parallel processing capabilities. For CRAM

Table 9.1: Examples of processing of 31GB BAM and matching 11GB CRAM of HG00110 with sambamba and samtools. Wall-clock time (t in seconds) reflects improved analysis time. CPU (×100%) reflects effective multi-core utilisation. See caption Fig. 9.1 for description of hardware, software and measurements.

	samtools		sambamba		speedup
	t(s)	CPU%	t(s)	CPU%	
BAM view	1506		429	785%	3.5×
filter[a]	195		35	471%	3.5×
sort	12288	396%	1265	945%	10×
index	577		137	562%	4×
markdup[b]	5220		2296	278%	2×
merge	3090	571%	2247	1015%	1.5×
mpileup[c]	7750		584	4409%	13×
BAM to CRAM	4354		640	796%	7×
CRAM to SAM	1850		729	347%	2.5×
CRAM index	9		9		=

[a] Filter on $q > 30$ and $Chr1$

[b] For markdup samtools v0.19 was used

[c] mpileup to VCF on 2GB BAM of $Chr1$ only

support the htslib C-library was linked against[234]. And for mpileup support the original samtools program is called in map-reduce fashion. This resulted in improved processing speed on multi-core computers (Table 9.1). Sambamba is most effective on machines where CPU utilisation is the constraining factor (Fig. 9.1). The gain may be therefore be limited on cluster setups where shared storage is a bottleneck, e.g., [227].

Compatibility: Sambamba is a robust replacement for the commonly used samtools commands: INDEX, SORT, VIEW, MPILEUP, MARKDUP, MERGE and FLAGSTAT. The output of sambamba compares to that of samtools, except for markdup, where the Picard 'sum of base qualities' method was chosen . Sambamba's RAM utilisation compares to that of samtools; only with SORT sambamba uses significantly less RAM.

New functionality: Sambamba adds new functionality compared to existing tools. To be able to calculate coverage statistics, read DEPTH analysis was added. To speed up splitting BAM files, SLICE was added which copies large regions without decompression. And when a BED file is supplied to VIEW, the index is used to decompress only those regions that are actually visited.

To further shorten processing time, index files are created on the fly by SORT, VIEW, MARKDUP and MERGE. And to combine multiple steps into one, powerful filtering with logic operators and regular expressions was added. For example, to filter on mapping quality and CIGAR

Finally, to make it easier to process results, sambamba VIEW can generate output in the standard Javascript object notation (JSON) format.

Source code: Sambamba abides by the rules of the 'Small tools MANIFESTO for Bioinformatics'[236]. The sambamba source code is extensible and maintainable. For SAM parsing we opted for Ragel, a finite-state machine compiler, which generates a fast look-ahead parser with input validation, making the code base even more compact[237]. Sambamba uses a unit testing framework with continuous integration testing, so that existing functionality is validated every time the code base is changed.

9.3 Conclusion

Sambamba is a software engineering example that shows how to make effective use of the D programming language and multi-core computers to reduce the time needed to get from sample to result. Whole genome sequencing and growing sample numbers make such performance improvements increasingly relevant.

10

Big data, but are we ready?

We welcome the timely Review by Schadt *et al.* (Computational solutions to large-scale data management and analysis. *Nature Rev. Genet.* 11, 647-657 (2010))[238], which presents cloud and heterogeneous computing as solutions for tackling large-scale and high-dimensional data-sets. These technologies have been around for years, which raises the question: why are they not used more often in bioinformatics? The answer is that, apart from introducing complexity, they quickly break down when much data is communicated between computing nodes.

In their review, Schadt and colleagues state that computational analysis in biology is high-dimensional, and predict that petabytes, even exabytes, of data will be soon stored and analyzed. We agree with this predicted scenario and performed a simple calculation to illustrate how suitable current computational technologies really are at dealing with such large volumes of data.

As shown in Fig. 10.1, processing 500 GB on each of 1000 cloud nodes takes minimally 9 hours, and currently costs $3000 (500 GB - 500 TB of data). The bottleneck in this process is the input/output (IO) hardware that links data storage to the calculation node (Fig. 10.1). All nodes are idle for long periods, waiting for data to arrive from storage; and shipping the data on a hard disk to the data storage does not resolve the bottleneck. We calculate that 1000 cloud nodes each processing one petabyte (1 petabyte to 1 exabyte of total data) currently takes two years, and costs $6,000,000.

A less expensive option would be to use heterogeneous computing, in which

Published as 'Big data, but are we ready?' by Oswaldo Trelles†, Pjotr Prins†, Marc Snir and Ritsert C. Jansen

Nature Reviews Genetics (2011)
12(3):224. doi:10.1038/nrg2857-c1

†These authors contributed equally to this manuscript

graphics processing units (GPUs) are used to boost speed. A similar calculation shows, however, that GPUs are idle 98% of the time when processing 500 GB of data. GPU performance rapidly degrades when large volumes of data are communicated, even with state-of-the-art disk arrays. Furthermore, GPUs are vector processors that are suitable for a subset of computational problems only.

What is the best way forward? Computer systems that provide 'fast' access to peta-bytes of data will be essential. As high-dimensional, large datasets exacerbate IO issues, the future lies in developing highly parallelized IO using the shortest possible path between storage and CPUs. Examples of this trend are Oracle Exadata[239] and IBM Netezza[240], which offer parallelized exabyte analysis by providing CPUs on the storage itself. Another trend for improving speed is the integration of photonics and electronics [241],[242].

To fully exploit the parallelization of computation, bioinformaticians will also have to adopt new programming languages, tools and practices, because writing correct software for concurrent processing that is efficient and scalable is difficult[243], [244]. The popular R programming language, for example, has only limited support for writing parallelized software (e.g. [153], [245]), whereas other languages, [246], [245] make parallel programming easier, e.g. through abstracting threads[247] and shared memory[244].

So, not only do cloud and heterogeneous computing suffer from severe hardware bottlenecks, they also introduce (unwanted) software complexity. It is our opinion that large multi-CPU computers are the preferred choice for handling big data. Future machines will integrate CPUs, vector processors, and RAM, with parallel high speed interconnections to optimize raw processor performance. Our calculations show that for peta-byte sized high-dimensional data, bioinformatics will require unprecedented fast storage and IO to perform calculations within an acceptable time frame.

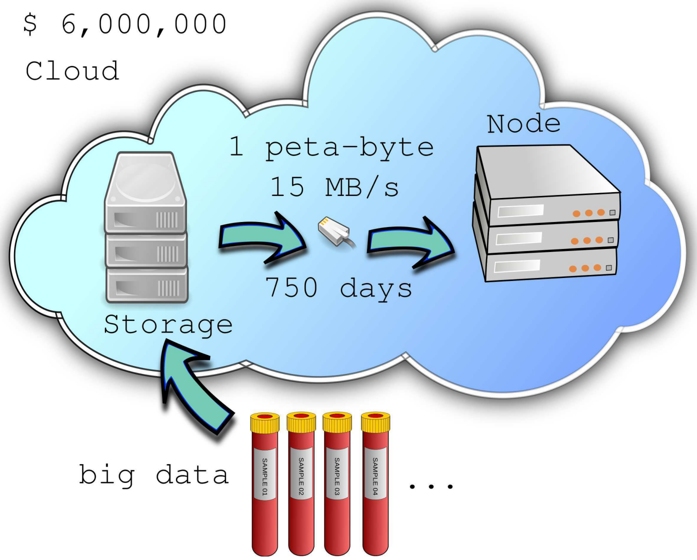


Figure 10.1: Input/output (IO) bottleneck between data storage and calculation node. In our calculations, 1000 computational nodes each processing a 500 GB dataset would take 500GB/15MB/s, or 9 hr, using large nodes at \$0.34/hr. The total cost for a single analysis run would be $1000 \times 9 \times 0.34 = \$3,060$. In reality, throughput will be lower because of competition for access to data storage caused by parallel processing. There are significant throughput instability and abnormal delay variations, even when the network is lightly utilized[248]. In the illustrated example, 1000 cloud nodes each processing a peta-byte dataset take 1PB/15MB/s, or 750 days, and cost $1000 \times 750 \times 24 \times 0.34 = \$6,120,000$.

11

Towards effective software solutions for big biology

Leading scientists tell us that the problem of large data and data integration, referred to as Big Data, is acute and hurting research. Recently, Snijder *et al.* (Toward effective sharing of high-dimensional immunology data), *Nature Biotechnology* **32**, 755-759 (2014)[249]) suggest a culture change with scientists to share high-dimensional data between laboratories. The elephant in the room is bioinformatics and bioinformatics software development in particular - which, despite being crucially important, mostly fails to address the requirements of 'big data'.

Whereas Internet companies such as Google, Facebook and Skype have built infrastructure and developed innovative software solutions to cope with vast amounts of data, the bioscience community seems to be struggling to realize big data software projects. This has led to problems in sharing, annotation, computation and reproducibility of data[227, 250, 251].

Before we can devise software solutions for big data, there are more basic pressing concerns with bioinformatics software development that need to be resolved. Biologists are not formally trained for software engineering, so much of the bioinformatics software available today has been developed by PhD biologists in relative isolation on the back of funded experimental research programs. This model of software development tied to wet-lab research can work well but has resulted in a culture of 'one-offs'. The aim of most research projects is to obtain

Published as 'Towards effective software solutions for big biology' by Pjotr Prins†, Joep de Ligt†, Artem Tarasov, Ritsert C. Jansen, Edwin Cuppen and Philip E. Bourne

Nature Biotechnology (2015)
33, 686-687 doi:10.1038/nbt.3240
†These authors contributed equally to this manuscript

results in the shortest possible time, and this is often achieved by writing prototype software rather than developing well-engineered and scalable solutions. Even when funding is obtained to develop software, there are usually no long-term resources allocated to software maintenance, which results in problems with bug fixing, continuity and reproducibility.

Instead of working alone to develop software, researchers can join or start collaborative free and open-source software (FOSS) projects, thereby improving their coding skills through the scrutiny of their peers. True FOSS projects have licenses that allow continuation of projects that were abandoned by the original developers, thereby enabling modular development. We published a bioinformatics manifesto[236] as a practical guide for FOSS-style development that aims to provide process and architecture guidelines for early-career bioinformaticians and their supervisors. Bioinformatics already has vibrant collaborative FOSS projects, such as Galaxy, Cytoscape, BioPerl and Biopython, but these projects are often worked on parttime owing to lack of or inadequate funding and will not service the requirements of big biology without major additional investment. For example, after initial funding from the US National Institutes of Health (NIH) and the National Science Foundation (NSF), the Galaxy project is now seeking new funding to continue its work, and no funds at all have been granted by scientific agencies to work on Biopython. The amount of dedicated funding for bioinformatics software development remains small. For example, the NIH has a budget of $30 billion, of which an estimated 2-4% is allocated to computation and bioinformatics grants. We estimate that only a small fraction of this funding is used for big data software development. By comparison, the nonprofit Mozilla Foundation turns over $300 million annually for software development and FOSS promotion, and Google invests an estimated $6.7 billion annually in R&D. Private donors could, in principle, establish a foundation to support software development for integrative web-based services on large computer clusters. If investments in sharing data resources for biomedical research, such as the NIH Big Data to Knowledge (BD2K) initiative, with an annual budget of $24 million, and the European Bioinformatics Institute's smaller BioSamples project, were matched by serious investments in software development, maintenance and reproducibility, these projects would render better returns.

One way to solve the challenge is to wait for companies, such as 23andMe, that have made multimillion-dollar deals with pharma to realize large-scale investments and create big data solutions. However, such solutions would need to be purchased and, owing to their proprietary nature, would be difficult to adapt or benchmark. Another solution would be for biology funding agencies to establish initiatives for centralized software development. A different solution, and the one that we favor, is to use FOSS as a distributed development effort and develop collaborative software projects, such as those developed by the Linux, Mozilla and Apache foundations, which include private sector participation. For example, the goal of the Linux Foundation (which includes members such as IBM and Intel) is to fund Linux development.

Most of the bioinformatics software in use today does not scale for terabytes of data. R software programs typically load all data in RAM and suffer from its memory and runtime inefficiencies, and they are not designed for simultaneous use of multiple CPUs to speed up computations[227]. Where programming languages such as R, Python, Perl and Ruby are great for prototyping and quick analysis, they fail to deliver when it comes to big data processing. Solving the scalability problem will require embracing programming languages that are more efficient and have abstractions for multi-CPU computations[227], even if switching languages proves hard for most bioinformatician programmers.

Attribution for bioinformatics software development is also problematic. In a post titled 'You're not allowed bioinformatics anymore' on his blog Opiniomics[252] Mick Watson eloquently explains that bioinformatics is a scientific discipline in its own right and that bioinformaticians need career development. Ironically, in many of the most-cited biology research publications, there is a substantial bioinformatics contribution (usually the analytic method), often delivered as novel software solutions and data. However, it is rare for bioinformaticians to feature either as first or last authors on publications in high impact journals. Authorship of community software projects can be troublesome as well, because the original authors tend to receive credit for the lifetime of the project, even when later code amendments and added functionality are equally or more important than the initial software. Lack of scientific attribution for software development hurts career development and can force bioinformaticians to opt for careers in traditional biology.

To solve the issue of attribution and related career development, we propose that the software contribution itself counts toward scientific track record. Every versioned software release and accompanying source code can be assigned a digital object identifier (DOI) with clear attribution for all contributors. The relative contribution of authors could be checked by visiting the software version control, such as that delivered by web services such as GitHub. This would make published software accountable, reproducible and citable. DOI citations could count as conventional citations, because they express the impact of a piece of software by its use.

In conclusion, our view is that to tackle the challenge of big biology software development, leading scientists need to acknowledge that software development is an integral part of research and not just an underpinning method. Projects need to promote bioinformatics collaborations and create scientific rewards. Universities need to increase their efforts to promote interdisciplinary research, to ensure that informatics is embedded in the life sciences curriculum and encourage talented software developers and biologists to get involved in big data by tailoring individual career-development plans.

Funding agencies can add institutional focus; emphasize collaborative FOSS approaches; build on existing grassroots initiatives[253]; create split funding streams for software and hardware; support maintenance of projects; encourage collaboration with experts in high-performance computing and software engineering; and fund larger projects dedicated to big biology software solutions.

12

General Discussion

The work described in this thesis provides computational methods and solutions for genetics in the era of high-throughput sequencing, and a road map for the development of software in big data genomics. The approach was to develop computational methods to answer the following research questions:

1. 'How can we identify genes involved in pathogenicity or plant defence from DNA and RNA sequences?'

2. 'How can we identify genes that are expressed differentially and relate them to a phenotype'

3. 'How can we improve tools for genetic analysis in the era of high-throughput sequencing?' and

4. 'How can we scale up computations and be prepared for the genomic data deluge?'.

Based on individual research cases, every chapter presents generic and tangible bioinformatics software solutions that come in the form of free and open source software (FOSS) tools and libraries that can be used by the wider research community.

This discussion chapter builds up on the software solutions presented in this thesis, painting a picture of further challenges in bioinformatics computational solutions. The chapter starts with a discussion on the merits and shortcomings of each individual software solution presented in this thesis (section 12.1), followed by a perspective and recommendations on software solutions for next generation sequencing, data integration and future research (section 12.2).

12.1 Merits and shortcomings of software solutions presented in this thesis

GWP: A cross-species genome-wide scan for nematode gene-families subject to diversifying selection

To identify genes subject to diversifying selection in the plant-parasitic nematode *Meloidogyne incognita,* evidence of positive selection was analysed on a genome-wide scale in multiple parasitic and non-parasitic nematode species. Special software was written for executing the pipeline on a compute cluster (Chapter 2), including a PAML parser (part of Chapter 7), a BLAST XML parser and Semantic

Web resource description framework (RDF) generators (also part of Chapter 8). Clusters of highly similar duplicated sequences within nematode genomes were investigated for evidence of positive selection by calculating dN/dS (ω), the ratio of non-synonymous to synonymous nucleotide substitution rates.

The merit of the bioinformatics approach presented in Chapter 2 concerns the comprehensiveness of including all available DNA material, including ORFs, and comparing the presence of positively selected clusters across genomes of nematodes with entirely different life histories. Gene families were compared for sequence homology and classified as putatively associated with plant parasitism. The results were compiled into a linked data resource (see also 12.2), thereby creating an automated annotation pipeline that can grow in value when more species are added.

The relatively large genome of plant-parasite *M. incognita* and the positive selection clusters (PSC) discovered in this study suggested that *M. incognita* harbours conserved coding and non-coding sequences under current or recent diversifying selection which makes up the host-pathogen interactome and possibly may help explain the success of *M. incognita* in attacking a large range of hosts.

Even though 77 putative sequence families under positive selection were identified, this approach should be viewed as an exploratory method. The approach of clustering sequences, aligning them and testing for positive selection involves parameter choices at every step. Future work may include, for example, relaxing the 70% identity constraint for clustering PSC. This may result in a larger set of gene families to study, including those proteins with a smaller conserved scaffold and larger hypervariable regions. With large sequence clusters, the number of sequences included in the alignment was reduced to a maximum of 19, so that PRANK and PAML would finish within reasonable computation time (24hr). Rather than taking the first 19 sequences, future work could break down highly diversified gene families into smaller sequence clusters, preferably along phylogenetic branches, and test them all separately for positive selection. Other potential improvements would be to include additional branch site models of evolution, such as provided by PAML[42].

GenEST and Genfrag: software for the identification of expressed genes from cDNA-AFLP

cDNA-AFLP is a form of high-throughput PCR-based transcript selection where identified targets can be sequenced (Chapter 3 and Chapter 4).

The GenEST software presented in Chapter 3 was the first in-silico cDNA-AFLP tool forming a bidirectional link between virtual transcript derived fragments (TDFs) derived from predictions on DNA sequences in an EST database and TDFs as resolved in the cDNA-AFLP lab protocol. The power of GenEST was demonstrated by the identification and validation of novel effectors from the nematode *Globodera rostochiensis* and linking hundreds of EST sequences to cDNA-AFLP expression profiles, and *vice versa*.

The Genfrag software presented in Chapter 4 evolved from GenEST to allow for more flexible gene identification when full genome data is available. Genfrag can handle larger data sizes than GenEST, comes with a choice of standard restriction enzymes and adapters and can run as an interactive web server with database attached. With Genfrag we successfully matched predicted splice variants of genes with differentially expressed TDFs of the plant, and model organism, *A. thaliana*. The use of GenFrag resulted in evidence of epigenetic parental imprinting in seed and identified 52 candidate maternally expressed genes in seed from the genome sequence of *A. thaliana*.

In Chapter 4 we concluded that cDNA-AFLP can be particularly useful when high specificity in distinguishing the expression of closely related genes is needed. A possible shortcoming of cDNA-AFLP is that the protocol is laborious. Even so, cDNA-AFLP in conjuction with Sanger sequencing is stringent and reproducible and, in contrast to microarray techniques and RNA sequenced data (RNA-seq, see also 12.2), it may still be more successful in distinguishing lowly expressed genes, gene heterozygosity and gene expression in highly similar paralogues.

MQM for R/qtl: software for genetics in the era of high-throughput sequencing

The Multiple QTL Mapping (MQM) method provides a sensitive approach for mapping quantitative trait loci (QTL) in experimental populations. In Chapter 5 we described a FOSS implementation of MQM. The main merit of MQM for R/qtl is that it is a robust and scalable implementation of the original MQM method which combines the strengths of linear model regression with those of interval mapping [142, 143] .

Parallelisation of calculations paves the way for high-throughput QTL analysis. To determine significance in large data sets we added permutation strategies for determining thresholds of significance relevant for QTL and QTL hot spots (Chapter 5). This way, MQM for R/qtl has become a parallelised comprehensive QTL mapping toolbox for the analysis of experimental populations and is increasingly used for research in, for example, *Mus musculus* [254], *A. thaliana* [255], and *Solanum lycopersicum* [256].

With MQM for R/qtl there are, however, also some shortcomings. To support the trend of the rapidly increasing number of phenotypes and genotypes in studies the software still needs major work, mostly because the containing R environment has severe limitations when it comes to fine-grained multi-threading and memory use. A new project named ‘qtlHD’ has been started in the D programming language to make use of fine-grained multi-core processing (see 12.2).

BioRuby and Biogems: software solutions for the genomic data deluge

In biomedical science, new technologies, data formats, and methods emerge continuously. Scientists want to take advantage of these developments as soon as possible, which requires bioinformatics software to keep up with new requirements. Bio-star (Bio*) projects, such as BioPerl[204] and Biopython[205], effectively function as (virtual) community centres and share a centralised approach in software development with large source code repositories. Bio* projects, generally, aim for consolidated tools, a stable application programming interface (API), and backwards compatibility. Where before data formats were a major challenge, today it is dealing with the data deluge caused by sequencing and the accompanying problem of data integration (see also 12.2).

The BioRuby project, published in 'BioRuby: Bioinformatics software for the Ruby programming language' (Chapter 7), is an international and vibrant collaborative software development initiative that delivers life-science programming resources for the Ruby programming language. BioRuby has components for sequence analysis, pathway analysis, protein modelling and phylogenetic analysis; it supports widely used data formats and provides access to databases, external programs and public web services.

We also created 'Biogem: an effective tool based approach for scaling up open source software development in bioinformatics' (Chapter 8), a tool based approach for rapid creation of decentralised internet published software modules to facilitate the FOSS publication of bioinformatics software modules written in Ruby.

All Ruby software created in the context of this thesis was contributed as FOSS to initially the main BioRuby project, e.g. the PAML parser of Chapter 2, and later as Biogems, e.g. the bio-blastxmlparser, bio-alignment, bigbio and bio-rdf biogems for Chapter 2, and three Genfrag related biogems for Chapter 4. Over 16 modules were contributed by the author as Ruby FOSS projects and are listed on the biogems.info website.

Because of the open nature of the BioRuby project, both BioRuby and Biogem software modules are increasingly used in biomedical research, not only in genomics, e.g., [257], but also in phylogenetics and prediction of protein structural complexes[258] and data integration[216]. The success of the Biogem approach can be measured by the increase and variety of publications and software written for biology and by the increasing number of contributors to Ruby bioinformatics. In 2011 the there were five active developers contributing to the BioRuby project. Two years after the introduction of Biogems by late 2012, there are over 120 new software modules contributed by over 30 software developers world-wide and there are over one hundred publications citing these two papers, according to Google Scholar (June 2015).

Sambamba: fast processing of NGS alignment formats

The ‘sambamba software’ is a good example of successfully scaling computations through the use of multiple cores on a computer (Chapter 9). Sambamba is a replacement for the popular samtools tool[259], a commonly used software tool for working with aligned output from sequencers. Sambamba makes use of multicore processing and is written in the D programming language[260, 261]. Not only does sambamba outperform samtools, but it already comes with an improved deduplication routine and other facilities, such as easy filtering of data.

The main shortcoming of sambamba is that it does not address the IO issue discussed in Chapter 10. Fig. 9.1 shows that from 8 CPU cores onwards performance does not improve. Having a more efficient alignment format may help reduce the IO bottleneck. The other shortcoming is that, for one commonly used functionality, sambamba uses the samtools mpileup routines instead of having its own routines. For future versions it would be an improvement to rewrite the (complex) underlying algorithms so sambamba can be deployed without samtools.

12.2 Perspective

In the following section we identify critical areas of work for bioinformatics in the coming years in relation to topics treated in this thesis.

Identification of genes involved in pathogenicity or plant defence

In Chapter 2 and Chapter 3 we developed methods for identifying genes involved in pathogenicity or plant defence. The success of these methods depends largely on the quality of the reference genome and prediction of genes and alternative splicing variants. The exact mechanism involved in transcribing RNA from DNA and the way RNA is spliced is complex and poorly understood[21]. In eukaryotes there are common patterns, such as a TATA box, which is usually required to initiate transcription, and there is an upstream coding region promoting transcription. It is also known that different factors are necessary for binding polymerase to a eukaryote promotor, and that the transcription is influenced by other factors, such as DNA folding, histone location and methylation. The problem is that these factors interplay differently between species. For this reason, the best software that predicts genes and splice variants from DNA is based on machine learning algorithms, which needs prior information in the form of a learning set[21]. When the genome is close enough to that of a well studied species, such as that of humans and mice, the prediction software can do a reasonable job, certainly in combination with homology searches and RNA-seq/EST/cDNA-based annotation[262], but for many genomes the contents are *terra incognita*[263]. With nematodes, for example, a gene predictor trained on *C. elegans* turned out to be inappropriate for plant-pathogens *M. hapla* and *M. incognita*. Therefore, in Chapter 2, we had to to resort to using ORFs, rather than relying on predicted genes alone.

Sequencing costs have fallen dramatically so that a single laboratory can afford to sequence large genomes (see also section 12.2). Although sequencing has become easier, genome analysis and annotation has not become less challenging. Several factors are responsible for this. In addition to mentioned problems with gene prediction, the shorter read lengths of second generation sequencing platforms mean that current genome assemblies rarely attain the contiguity of the 'classic' shotgun assemblies. Another challenge is posed by the need to update and merge annotation data sets. RNA-seq provides an obvious means for updating older annotation data sets; doing so, however, is non-trivial[249]. Furthermore, it is not unusual today for multiple groups to annotate the same genome using different annotation procedures. Merging these to produce a consensus annotation data set is a complex task[249, 263], also discussed in section 12.2.

In Chapter 2 we assumed effector proteins involved in plant-pathogen interactions to contain hypervariable regions in the DNA sequence encoding the protein. Our method, therefore, only identifies effectors which are represented by such regions and gene families. Effectors produced through gene conversion or, perhaps, forms of post-transcriptional processing will be missed. Also, our method discards a wide range of candidate gene families by only selecting candidates that are represented in multiple species.

Identification of genes that are expressed differentially and relate them to phenotype

In Chapter 4 differential gene expression was linked with maternal imprinting. In Chapter 5 and Chapter 6 genetical genomics was introduced where gene expression is used as a phenotype that gets linked to genotype. This leads directly to the next question of improving genetics in the era of high-throughput sequencing:

Improve genetics in the era of high-throughput sequencing

In Chapter 5 we provided a sensitive approach for QTL mapping that makes optimal use of phenotype and genotype information for calculating QTL. QTL mapping offers statistical analysis of high-throughput data[264]. It is important to realise, however, that with RNA-seq high-throughput sequencing the number of phenotypes increases rapidly. The power of QTL mapping in model species depends on the amount of individuals used and the (detected) underlying DNA recombination in the experimental population (Chapter 6). Also, power can be increased by multi-trait (*e*QTL) analysis and intepretation[8].

As mentioned in section 12.1, the current crop of QTL mapping tools needs to improve further to scale for large datasets. To facilitate new requirements in software it is sometimes a good idea to start from scratch and design on a new architecture. Therefore we started the 'qtlHD' project (recently renamed to 'R/qtl2') as a clean follow-up on MQM for R/qtl (Chapter 5). The QTL mapping functionality is being moved out of the R container. The new code may be written in the high-performing ‘D’ programming language. D is binary compatible with C and can be bound to R. This means researchers can still work in the R environment, if required.

In addition to solving the scalability requirements, next generation QTL mapping tools have to address demands from the genetics community, especially support for new crosses (e.g. [265]), data visualisation suitable for the web[266], and access to shared data sources. Other bioinformatics demands from the genetics community are (in no particular order): drilling down on ‘lower’ traits, dealing with causality and inference, working with epigenetic & structural variation, model within cell and between cell interactions, dissecting the environment component at the cellular level, support microbiome and between-species interactions and provide access to human relevant biomedical data. These functionalities do not necessarily have to be part of a qtlHD solution, but reflect current developments in genetics that require bioinformatics support.

Scale up software solutions and be prepared for the genomic data deluge

In Chapter 10 we discussed some of the problems around big data analysis. The sequencing effort is causing a deluge of data, often (inaccurately) referred to as the ‘big data’ problem in biology. Big data is defined as ‘a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications’[267]. At this stage data sets are becoming large, but the complexity of such large datasets in biology is still fairly low, even though the underlying biology and data analysis itself can be complex.

DNA sequencers are churning out terabytes (TB) per day. Even the final processed data that is stored for further analysis is bulky. For example, for the Genome of the Netherlands (GoNL)[268], the genomes of 750 people have been sequenced representing 60TB data and, as of March 2013, the 1000 genomes project contains approx. 2700 individuals and counts 464TB[269]. These numbers are growing fast and world-wide computational capacity is not keeping up with data growth. And, as predicted in Chapter 10, the IO bottleneck is increasingly problematic. An example of a solution we published for speeding up processing is ‘probabilistic fast file fingerprinting’ (Pfff)[270] which speeds up the mundane task of file comparisons. Comparing and transferring files is computationally expensive and ties down shared resources in data centers. Pfff exploits the intrinsic variation present in biological data and computes file fingerprints by sampling randomly from the file instead of reading it in full. This way, file comparison has a flat performance characteristic and is not correlated with file size[270].

Data analysis takes significant computational resources. With GoNL, locating DNA variation between individuals took 145,000 CPU hours using the GATK tool[271]. And, even though GATK SNP calling is approximately computationally linear, data growth is non-linear. Currently the genomes of tens of thousands of people are being sequenced, and soon millions world-wide will be sequenced. Especially RNA sequencing will drive data future growth as it is applied over multiple experimental conditions, tissues, time series, and may include the meta-transcriptome of all bacterial gut data which is much larger than the host genome, e.g., [9].

To solve large data processing, as suggested in Chapter 10, computer systems that provide 'fast' access to peta-bytes of data will be essential. As high-dimensional, large datasets exacerbate IO issues, the future lies in developing highly parallelized IO using the shortest possible path between storage and CPUs. Examples of this trend are Oracle Exadata[239] and IBM Netezza[240], which offer parallelized exabyte analysis by providing CPUs on the storage itself. Another trend for improving speed is the integration of photonics and electronics [241],[242].

To fully exploit the parallelization of computation, bioinformaticians will also have to adopt new programming languages, tools and practices, because writing correct software for concurrent processing that is efficient and scalable is difficult[243], [244]. The popular R programming language, for example, has only limited support for writing parallelized software (e.g. [153], [245]), whereas other languages, [246], [245] make parallel programming easier, e.g. through abstracting threads[247] and shared memory[244].

NGS challenges

A thesis built on sequencing should discuss the emergence of next generation sequencing (NGS). NGS is dramatically faster than older sequencing techniques, but raises its own challenges. Both the technology and biology strive to confound the hunt for tangible results and clear markers. In general, technology related problems are a result of short-reads, bias in short-read selection, stochastic effects and misreads which increase towards the end of a read, e.g., [272]. Biology related problems are a result of repetitive information in the genome, few or flawed reference genomes, issues around ploidy and the difficulty of functional gene prediction and the functional impact of DNA nucleotide variants. Also, to get enough starter material multiple cells in different states are sequenced together, such as somatic variations in DNA variant calling, and expression variations in RNA-seq calling. Together these factors result in a hunt for variation that is often close to the noise level, whether it concerns *de-novo* genome sequencing or sequencing for SNP/mutation scoring, e.g., [273].

In human, the problems with NGS sequencing are less pronounced than with less studied species because after billions of dollars worth of research, much is known about the human genome and its genes. The '$1,000 genome and the $100,000 analysis'[274] is therefore probably better restated as the '$1,000 genome and the $1,000,000,000 analysis' when taking the human genome project as a reference for work on other species.

Most tools in use today are based on using a single reference genome. The reasons are that a single reference gives an anchor for (i) speedily mapping of reads, (ii) easy variant calling (iii) useful visualisation of the genome and variants. Unfortunately, using a single reference genome introduces challenges and pitfalls of its own. Current software development activity gives a perspective on such challenges in bioinformatics. This activity includes improvements on variant calling by local realignment against the reference genome, and imputed reference genomes (see below).

Improving SNP calls and indels is an active area of software development. For example, the non-FOSS GATK haplotype caller[271] has made significant strides in 2013 improving sensitivity and specificity by realigning reads locally and local indel variant quality score recalibration, a technique also picked up, for example, by FreeBayes, a FOSS variant caller[275]. Local realignment mostly corrects for small indels and is necessary because of of faulty positioning calls by the mapper on top of a single reference genome. Mapping of reads (not to be confused with QTL mapping) is the process of locating sequenced reads on a reference genome. When the reads do not map, i.e., when the individual does not compare against the reference genome, the reads are simply discarded (often in the order of 30% of total reads). When reads do map, they can map against the wrong position, especially in highly similar or repetitive regions. Local realignment can not fix all these issues, and, in fact, can introduce artifacts of its own.

To address the problem of mapping reads wrongly against a single reference genome, one immediate solution is to calculate an imputed 'reference' genome for every individual from the sequenced parent/population genomes using all available haplotype information, and map reads against the imputed genome, similar to [276] and [277]. For example, in *M. musculus*, imputed genomes are now calculated for homozygous mice and other model organisms, which improves RNA-seq results greatly[278]. Also with human heterozygous genomes it is possible to improve the scoring of SNPs and other variants by piecing together a closest 'reference' genome by finding the closest matching haplotypes. In fact, the latest reference human genome (HG20) is no longer a single reference genome, but consists of sections of multiple alternate or ALT loci, especially for sections considered problematic or variable, such as the major histocompatibility complex (MHC). Selecting a closer match reduces the number of differences between that of the individual and the calculated reference genome, thereby (hopefully) increasing the fidelity of mapping. The latest version of BWA, a software package for mapping low-divergent sequences against a large reference genome, recently added support for ALT sequences[200]. And the latest versions of GATK have some form of haplotype support. There is currently, however, no mapper that can take full account of population haplotype information.

In the near future, it may be possible to let go of the concept of a reference genome altogether. The main advantage of using a reference genome is that it allows for fast read-mapping software, such as BWA, and it also provides a scaffold for comparison. The downside is that individuals, or part of their genomes, are closer or further removed from the reference genome. Finding variants, therefore, can more or less powerful between individuals, depending on the sequence difference or distance. Missing data on the reference genome is especially problematic.

With the right software it may be feasible to assemble the genome of every individual from scratch using all available haplotype information at the time of assembly, including that of the reference genome with the growing number of ALTS and effectively use these and population haplotypes as 'hints'. These hints take this type of assembly some steps up from a pure *de-novo* assembly. Also genomic evidence derived from other technologies, such as longer nanopore reads or even optical mapping techniques, e.g., [262], can be incorporated right at the assembly step to produce the genome for an individual. Different types of assemblers can also be used to generate new hints for the main assembler. Current *de-novo* assemblers in use are based on using greedy and overlap-layout-consensus (OLC) [279] or de Bruin graphs[280]. Slower statistics based assembers that use the full read length have also produced good results[281]. Combining the strengths of the different assembly methods in a haplotype setting may improve results further.

Individually assembled genomes, when they closely represent the actual genome, are especially likely to improve RNA-seq calling, SNP and MNP calling; this was shown with ABRA, which reassembles a genome to account for INDELs[282]. Also for (larger) structural variant (SV) calling in DNA, typically operating at the noise level, it is likely that calls will be improved and SNP callers, for example, can then correct for underlying SV.

Recent large scale DNA sequencing studies in human, nematode and plant populations have provided evidence that structural DNA variation is omnipresent between individuals. This suggests that phenotype effects of SV may be underestimated and that tools for variant detection may benefit from improved detection methods that account for variation in populations. The GoNL project recently showed that structural variation in DNA between humans is larger than thought before, even between parent and child (communication Victor Guryev). This implies that we should correct for SV when calling SNP and SNV variants.

To benefit from the power of detecting variants in populations new tools are required. Current variant calling tools lack the ability to take full population information into account and only support a single reference genome. The GATK tool has shown that haplotype calling improves SNV detection even by posterior comparison of variants. Even so, GATK can not handle larger populations in the early variant calling stages and has no facility for SV correction. A new crop of SV tools require significant software engineering because accounting for multiple individuals is essentially an $O(n^2)$ (quadratic) problem and requires smart solutions for RAM utilisation, multicore programming and, possibly, interaction between running processes on a compute cluster.

Creating individually assembled genomes for species that have larger genomes, using combined hints from multiple assemblers, multiple reference genomes and population haplotype information, requires a full redesign of the current variant calling software stack, starting from assembly all the way to variant calling. Incorporating all available evidence at time of assembly will be computationally intensive, will require significant computational power from compute clusters, and will require clever data storage systems. Intriguingly, by having an individual assembled genome, it may be possible to remove the mapping step altogether, or use the mapping step solely to check the fidelity of the assembled genome. Finally, there will also be the challenge of designing a user interface for presenting variants based on different individual genomes to the research community. For visualisation it may be advisable to fall back on a single reference genome again. Here, the single genome acts as a scaffold for visualisation with the added benefit of reusing existing visualisation tools.

To solve the problems around the reference genome we will have to produce solutions that are able to assemble large genomes using 'hints', even if it proves computationally expensive. Well engineered software that allows for creating individually assembled genomes and calling variants thereon will improve variant detection in well-studied species. This will especially be a boon for species that are less studied and have incomplete reference genomes or highly variable genomes such as found in pathogenic species, as has been shown by work on bacteria[283].

The data integration challenge

Next to the rate of data generation, one of the most important bioinformatics challenges is the problem of data integration. Not only for data generated by genomics, but also for data generated by other technologies from -omics, such as metabolomics and proteomics. Other types of data that need integration are the phenotypes given by diagnostics and research programmes, including tissue comparisons, time-series and even visualisations (images). Effective integrating of such multi-level data and mining that data is one of the major challenges in biomedical research (Fig. 12.1).

Decennia of work have led to an accumulation of databases world-wide, including important resources NCBI GEO[284], KEGG[285], ENCODE[3] and UNIPROT[286]. Lately, new data acquisition technologies, especially next generation sequencing (NGS), are rapidly increasing the amount of information available online, from data published with papers all the way to large scale collaborations, such as Wormbase for nematodes (used in Chapter 2) and The Arabidopsis Information Resource (TAIR, used in Chapter 4)[287] and 1,001 genomes for *A. thaliana*[288] offering information on sequenced genomes, gene expression, gene onthologies, pathways etc. To reach information every service offers a different approach and there is no unified way of accessing and querying this data with automated tools. Collaborative projects, such as BioPerl[204] BioRuby (Chapter 7) and Biogem (Chapter 8), put major efforts in writing specialised software for accessing these resources.

The solution to data integration, the Semantic Web, originated before 2001[289]. The Semantic Web refers to information that is explicitly encoded in a standardised machine-readable syntax with relationships between entities[289]. The Semantic Web is useful for biomedical data because it links data without enforcing rigid two-dimensional data structures, which is the current standard way of representing data in biology, including the tabular structure, the database table and the spreadsheet. Not only is modelling data up-front required for tabular data, also such two-dimensional structures are highly constrained. For this reason, almost all successful biological data standards introduce columns containing attributes or key-value pairs, adding dimensions to a two-dimensional table. Examples of such successful formats that add key-value pairs in a two-dimensional format are GFF3 and GVF tag-value pairs[290], SAM/BAM tags[259], and VCF key-value pairs[291]. The flexibility of such key-value pairs is required because data use cases evolve over time and the format needs to support them. Without such flexibility the 'standard' format would become obsolete quickly. The downside is that these formats evolve quickly, become less standardized, become hard to test for correctness, incompatibilities arise between tools and data interchange and integration becomes hard.

The Semantic Web takes the key-value idea a step further. Not only is the Semantic Web flexible because it allows linking data in any way (as it represents a graph rather than a table or tree) but also because it formalizes direction and multi-dimensionality, which is a natural way of modelling biological and biomedical data. Such 'linked data' can also be integrated site-to-site across multiple independent providers via queries that span multiple data-endpoints, because Semantic Web standards cater for that. Linked data technologies and conventions, therefore, facilitate data exploration and evaluation by removing the need to design an integrative schema, download, homogenise, and finally warehouse data subsets in order to ask common domain-spanning questions[292]. In short, the Semantic Web represents a wide range of standards and software solutions that are very useful for biological research and should be adopted when data integration is a concern.

The Semantic Webification of data and structures is an on-going and expanding exercise. It is indicative that, for example, EBI, KEGG and UNIPROT increasingly make data available through Semantic Web tools and that NCBI resources are being made available[293, 294]. In Chapter 2 we used Semantic Web technologies to annotate the genomes of pathogenic nematodes by pinpointing regions which may be involved with pathogenicity and published the results in RDF. Therefore, this database can easily be accessed and mined over the internet using standard Semantic Web technologies, in the same way as described in the Bio2RDF paper which also transforms and publishes public bioinformatics databases such as PDB, MGI, HGNC and several of NCBI's databases[295].

To solve the data intergration problem, RDF and linked data are going to play an increasingly important role in the biomedical sciences, next to tabular file formats, SQL databases and the more tree-like storage systems provided by 'NoSQL' systems, such as CouchDB and MongoDB. Most biomedical data is naturally stored in a graph-database and today's SPARQL query engines provide sufficient power for federated data integration purposes. Each of the different database systems

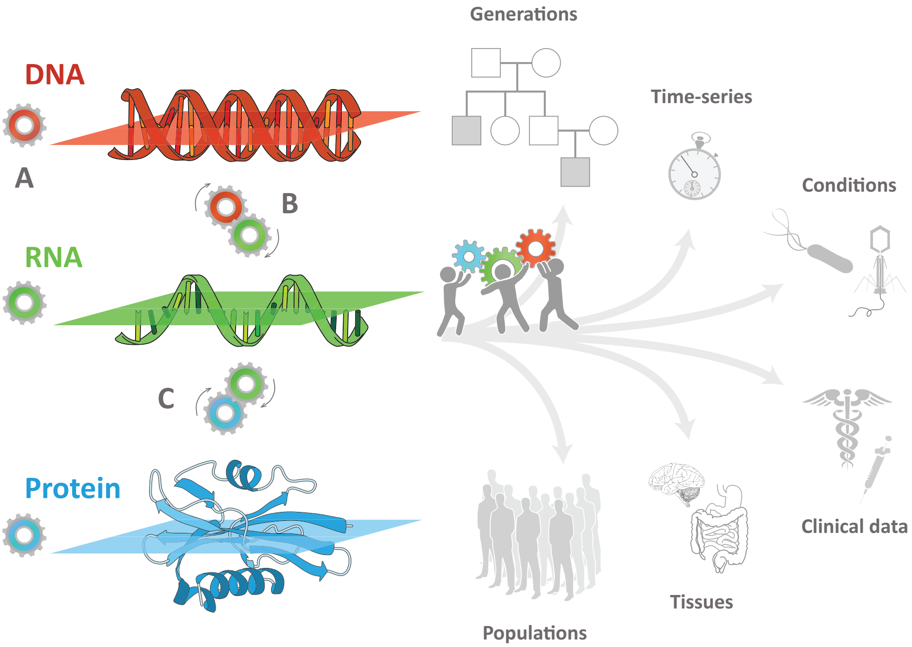


Figure 12.1: Biology is increasingly data-driven and getting more complex and harder to process using existing bioinformatics software solutions. Solutions have to integrate available data sources (e.g. A-B-C) and scale up in a reproducible way for hundreds conditions, tissues and time-series and 100,000s of individuals in multiple generations. Furthermore, clinical data and imaging have to be linked with such molecular evidence to unravel diseases, such as cancer. We should be aiming at effective exploration of biomedical relevant big biology, beyond current systems biology approaches (figure compiled by Joep de Ligt and Pjotr Prins).

should be used for their different strengths, so in the biomedical sciences we are likely to end up with a mixture of systems.

Summary

Biology is increasingly data driven by virtue of the development of high-throughput technologies, such as DNA and RNA sequencing. Computational biology and bioinformatics are scientific disciplines that cross-over between the disciplines of biology, informatics and statistics; which is clearly reflected in this thesis. Bioinformaticians often contribute crucial insights and novelty to scientific research because they are central to data analysis and contribute concrete algorithms and software solutions. In addition, bioinformaticians have an important role to play when it comes to organising data and software and making it accessible to others. In this thesis, in addition to contributing to biological questions, I discuss issues around accessing and sharing data, with the challenges of handling large data, input/output (IO) bottlenecks and effective use of multi-core computations.

By creating software solutions together with molecular biologists, I contributed and published insights in biological processes in nematodes and plants. I published software solutions that made it easier for others to analyse data, which impacts the wider research community. I created solutions that made it easier for others to publish software solutions by themselves. The introduction of computing and the internet makes it possible to share ideas and computational methods. I am convinced it is a good idea to publish software solutions as 'free and open source' software (FOSS) in the public domain so that we can continue to build on the work of others.

Chapter 2 presents a computational method for identifying gene families in a sequenced genome that may be involved in pathogenicity, i.e., those genes that code for proteins that interact with molecules of an infected host. Such nematode proteins are known to contain highly variable DNA sections that code for the biochemical properties of an interaction site. By applying phylogenetic analysis through maximum likelihood (PAML) and comparison of homologues sequences in other organisms with comparable and different life styles, we discovered 77 unique candidate sequence families in the plant pathogen *M. incognita* that deserve further investigation in the laboratory.

Chapter 3 presents GenEST, a computational method for predicting which fragments captured by the cDNA-AFLP high-throughput technology matched known expressed sequence tags (ESTs). The cDNA-AFLP biochemical process was calculated *in silico* and fragments matching the fragment lengths as given by cDNA-

AFLP were matched. Through this technique novel effectors from the nematode *Globodera rostochiensis*, putatively involved in pathogenicity, were identified and partly confirmed in the laboratory.

Chapter 4 presents GenFrag, a computational method that expands on GenEST for predicting which fragments captured by cDNA-AFLP matched fragments of a fully sequenced genome with its known spliced gene variants. Through this *in silico* technique genes were identified in the plant *Arabidopsis thaliana* putatively involved in maternal genomic imprinting and partly confirmed in the laboratory.

Chapter 5 presents multiple QTL mapping (MQM), a high-throughput computational method for predicting what sections of a genome correlate with, for example, gene expression. The study of finding such eQTL is challenging, not least because many of them are potentially false positives. The MQM parallelized algorithm is embedded in the R/qtl software package which makes it widely available to researchers. The impact thereof means that it is widely cited by studies on model organisms, such as mouse, rat, the nematode *Caenorhabditis elegans* and the plant *A. thaliana*.

Chapter 6 presents a theoretical framework in the form of a review for identifying plant-resistance genes (R-genes) that combines the lessons learnt in the previous chapters. Plants lack an adaptive immune system and therefore, next to having physical defences, use R-genes to code for proteins that recognise molecules and proteins from invading pathogens, with an example on *A. thaliana*. These R-genes can be viewed as the counterparts of effectors identified in Chapter 3 and Chapter 4. By introducing the concept of a prior the chapter discusses *e*QTL or broader *x*QTL techniques as presented in the Chapter 5 to narrow down on gene candidates involved in plant defence.

Chapter 7 and Chapter 8 present FOSS bioinformatics tools, and modules that make use the Ruby programming language. BioRuby (Chapter 7) has components for sequence analysis, pathway analysis, protein modelling and phylogenetic analysis; it supports widely used data formats and provides access to databases, external programs and public web services. All Ruby software created in the context of this thesis was contributed initially to the main BioRuby project, e.g. the PAML parser of Chapter 2, and later as individual Biogems (Chapter 8), e.g. the bio-blastxmlparser, bio-alignment, bigbio and bio-rdf biogems for Chapter 2, and three Genfrag related biogems for Chapter 4. Over 16 modules were contributed by the author as Ruby FOSS projects and are listed on the http://biogems.info/ website. Because of the open nature of the BioRuby project, both BioRuby and BioGem software modules are increasingly used and cited in biomedical research, not only in genomics, but also in phylogenetics and prediction of protein structural complexes and data integration.

Chapter 9 presents sambamba, a software tool for scaling up next generation sequencing (NGS) alignment processing through the use of multiple cores on a computer. Sambamba is a replacement for samtools, a commonly used software tool for working with aligned output from sequencers. Sambamba makes use of multi-core processing and is written in the D programming language. Not only does sambamba outperform samtools, but it already comes with an improved deduplication routine and other facilities, such as easy filtering of data. The Sambamba software is now used in the large sequencing centres around the world.

Chapter 10 'Big Data, but are we ready?' gives a response to a publication on using cloud computing for large data processing. The chapter discusses computational bottlenecks and proves prescient because the number of citations of this paper increases every year.

Chapter 11 'Towards effective software solutions for big biology' discusses the need for a change of strategy with regard to bioinformatics software development in the biomedical sciences to realise big biology software projects. This includes improved scientific career tracks for bioinformaticians and dedicated funding for big data software development.

Chapter 12 discusses the computational methods and software solutions presented in this thesis, painting a picture of further challenges in bioinformatics computational solutions for the elucidation of biological processes. The chapter starts with a discussion on the merits and shortcomings of each individual software solution presented in this thesis, followed by a perspective on next generation sequencing, data integration and future research in software solutions.

Acknowledgements

Scientific work is rarely produced alone and in isolation. There were a great many people involved in the work leading up to this thesis. If I were to thank everyone it would run longer than some chapters in this thesis. For that reason, I limit myself to the few people who I am extremely grateful to and who contributed significantly to my development as a scientist.

First, I wish to thank the promotors Jaap Bakker, Ritsert Jansen and Geert Smant for support and inspiration and never giving up and moulding a self proclaimed hacker into a scientist.

Special mention for those who were particularly instrumental in writing this thesis: Charles Spillane, Karl Broman, Maria Anisimova, Oswaldo Trelles, Artem Tarasov and Danny Arends.

Also, Joep de Ligt, Harm Nijveen, Edwin Cuppen, David Brown and Jeroen Saeij gave ideas, inspiration and direction. I feel lucky to find such great people and fruitful research collaborations.

Fitting for a thesis on free and open source software (FOSS), I wish to thank the bioinformatics FOSS community, especially Toshiaki Katayama, Raoul Bonnal, Francesco Strozzi, Naohisa Goto, Steffen Möller, Chris Fields, Peter Cock, Brad Chapman, John Woods and Hilmar Lapp for everything we achieved and all the fun we have.

Current and former colleagues, especially from the Wageningen Laboratory of Nematology, the Groningen Bioinformatics Centre, the Department of Human Genetics at the University Medical Centre Utrecht and recently the department of Anatomy and Neurobiology at the University of Tennessee: I thank you all for making working in science such a wonderful experience.

Finally, I thank my parents Willem and Jos, both graduates of Wageningen University themselves. And I am forever grateful of Eva, my support throughout the years and who tolerates my geeky friends, hobbies and shares my love for feline creatures and exotic locations.

I also wish to mention two special people no longer around anymore. Ling Qin drew me into biology and science and Jack Leunissen always had an open door and shared the love of programming languages. I miss you both.

Bibliography

1. F. Crick. Central dogma of molecular biology. *Nature*, 227(5258):561–563, 1970.

2. F. Sanger, S. Nicklen, and A. R. Coulson. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A*, 74(12):5463–5467, 1977.

3. B. E. Bernstein, E. Birney, I. Dunham, E. D. Green, C. Gunter, and M. Snyder. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414): 57–74, 2012.

4. Y. W. Kan and A. M. Dozy. Polymorphism of DNA sequence adjacent to human beta-globin structural gene: relationship to sickle mutation. *Proc Natl Acad Sci U S A*, 75 (11):5631–5635, 1978.

5. Eliot M. Rosen, Saijun Fan, Richard G. Pestell, and Itzhak D. Goldberg. Brca1 gene in breast cancer. *Journal of Cellular Physiology*, 196(1):19–41, 2003. ISSN 1097-4652.

6. L. S. Andersson, M. Larhammar, F. Memic, H. Wootz, D. Schwochow, C. J. Rubin, K. Patra, T. Arnason, L. Wellbring, G. Hjalm, F. Imsland, J. L. Petersen, M. E. McCue, J. R. Mickelson, G. Cothran, N. Ahituv, L. Roepstorff, S. Mikko, A. Vallstedt, G. Lindgren, L. Andersson, and K. Kullander. Mutations in DMRT3 affect locomotion in horses and spinal circuit function in mice. *Nature*, 488(7413):642–646, 2012.

7. Genome sequence of the nematode C. elegans: a platform for investigating biology. *Science*, 282(5396):2012–2018, 1998.

8. R. C. Jansen and J. P. Nap. Genetical genomics: the added value from segregation. *Trends Genet*, 17(7):388–391, 2001.

9. R. Knight, J. Jansson, D. Field, N. Fierer, N. Desai, J. A. Fuhrman, P. Hugenholtz, D. van der Lelie, F. Meyer, R. Stevens, M. J. Bailey, J. I. Gordon, G. A. Kowalchuk, and J. A. Gilbert. Unlocking the potential of metagenomics through replicated experimental design. *Nat Biotechnol*, 30(6):513–520, 2012.

10. S. Tawill, L. Le Goff, F. Ali, M. Blaxter, and J. E. Allen. Both free-living and parasitic nematodes induce a characteristic Th2 response that is dependent on the presence of intact glycans. *Infect Immun*, 72(1):398–407, 2004.

11. C. Whitton, J. Daub, M. Quail, N. Hall, J. Foster, J. Ware, M. Ganatra, B. Slatko, B. Barrell, and M. Blaxter. A genome sequence survey of the filarial nematode Brugia malayi: repeats, gene discovery, and comparative genomics. *Mol Biochem Parasitol*, 137(2):215–227, 2004.

12. A. Dobson, K. D. Lafferty, A. M. Kuris, R. F. Hechinger, and W. Jetz. Colloquium paper:

homage to Linnaeus: how many parasites? How many hosts? *Proc Natl Acad Sci U S A*, 105 Suppl 1:11482–11489, 2008.

13. P. Abad, J. Gouzy, et al. Genome sequence of the metazoan plant-parasitic nematode Meloidogyne incognita. *Nat Biotechnol*, 26(8):909–915, 2008.

14. J. L. Dangl and J. D. Jones. Plant pathogens and integrated defence responses to infection. *Nature*, 411(6839):826–833, 2001.

15. S. Rehman, W. Postma, T. Tytgat, P. Prins, L. Qin, H. Overmars, J. Vossen, L. N. Spiridon, A. J. Petrescu, A. Goverse, J. Bakker, and G. Smant. A secreted SPRY domain-containing protein (SPRYSEC) from the plant-parasitic nematode Globodera rostochiensis interacts with a CC-NB-LRR protein from a susceptible tomato. *Mol Plant Microbe Interact*, 22(3):330–340, 2009.

16. Z. Yang. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol*, 24 (8):1586–1591, 2007.

17. Z. Yang. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci*, 13(5):555–556, 1997.

18. J. Zhang, R. Nielsen, and Z. Yang. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Mol Biol Evol*, 22(12): 2472–2479, 2005.

19. S. Moretti, R. Murri, S. Maffioletti, A. Kuzniar, B. Castella, N. Salamin, M. Robinson-Rechavi, and H. Stockinger. gcodeml: a Grid-enabled tool for detecting positive selection in biological evolution. *Stud Health Technol Inform*, 175:59–68, 2012.

20. R. D. Emes and Z. Yang. Duplicated paralogous genes subject to positive selection in the genome of Trypanosoma brucei. *PLoS One*, 3(5):e2295, 2008.

21. N. Goel, S. Singh, and T. C. Aseri. A comparative analysis of soft computing techniques for gene prediction. *Anal Biochem*, 438(1):14–21, 2013.

22. E. M. Schwarz, I. Antoshechkin, C. Bastiani, T. Bieri, D. Blasiar, P. Canaran, J. Chan, N. Chen, W. J. Chen, P. Davis, T. J. Fiedler, L. Girard, T. W. Harris, E. E. Kenny, R. Kishore, D. Lawson, R. Lee, H. M. Muller, C. Nakamura, P. Ozersky, A. Petcherski, A. Rogers, W. Spooner, M. A. Tuli, K. Van Auken, D. Wang, R. Durbin, J. Spieth, L. D. Stein, and P. W. Sternberg. WormBase: better software, richer content. *Nucleic Acids Res*, 34(Database issue):D475–D478, 2006.

23. B. J. Haas, S. Kamoun, et al. Genome sequence and analysis of the Irish potato famine pathogen Phytophthora infestans. *Nature*, 461(7262):393–398, 2009.

24. K. D. Pruitt, T. Tatusova, G. R. Brown, and D. R. Maglott. NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Res*, 40(Database issue):D130–D135, 2012.

25. S. Rivas and S. Genin. A plethora of virulence strategies hidden behind nuclear targeting of microbial effectors. *Front Plant Sci*, 2:104, 2011.

26. A. Morgulis, G. Coulouris, Y. Raytselis, T. L. Madden, R. Agarwala, and A. A. Schaffer. Database indexing for production MegaBLAST searches. *Bioinformatics*, 24(16): 1757–1764, 2008.

27. Thomas Nordahl Petersen, Søren Brunak, Gunnar von Heijne, and Henrik Nielsen.

SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nature methods*, 8(10):785–786, October 2011. ISSN 1548-7105.

28. L. Kall, A. Krogh, and E. L. Sonnhammer. A combined transmembrane topology and signal peptide prediction method. *J Mol Biol*, 338(5):1027–1036, 2004.

29. S. Harris, N. Lamb, and N. Shadbolt. 4store: The design and implementation of a clustered rdf store. In *5th International Workshop on Scalable Semantic Web Knowledge Base Systems (SSWS2009)*, 2009.

30. T. Katayama, K. Arakawa, et al. The DBCLS BioHackathon: standardization and interoperability for bioinformatics web services and workflows. The DBCLS BioHackathon Consortium*. *J Biomed Semantics*, 1(1):8, 2010.

31. J. L. Villanueva-Canas, S. Laurie, and M. M. Alba. Improving genome-wide scans of positive selection by using protein isoforms of similar length. *Genome Biol Evol*, 5(2): 457–467, 2013.

32. W. Fletcher and Z. Yang. The effect of insertions, deletions, and alignment errors on the branch-site test of positive selection. *Mol Biol Evol*, 27(10):2257–2267, 2010.

33. P. Markova-Raina and D. Petrov. High sensitivity to aligner and high rate of false positives in the estimates of positive selection in the 12 Drosophila genomes. *Genome Res*, 21(6):863–874, 2011.

34. G. Jordan and N. Goldman. The effects of alignment error and alignment filtering on the sitewise detection of positive selection. *Mol Biol Evol*, 29(4):1125–1139, 2012.

35. G. Talavera and J. Castresana. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst Biol*, 56(4): 564–577, 2007.

36. J. Roux, E. Privman, S. Moretti, J. T. Daub, M. Robinson-Rechavi, and L. Keller. Patterns of positive selection in seven ant genomes. *Mol Biol Evol*, 31(7):1661–1685, 2014.

37. S. Moretti, B. Laurenczy, W. H. Gharib, B. Castella, A. Kuzniar, H. Schabauer, R. A. Studer, M. Valle, N. Salamin, H. Stockinger, and M. Robinson-Rechavi. Selectome update: quality control and computational improvements to a database of positive selection. *Nucleic Acids Res*, 42(Database issue):D917–D921, 2014.

38. S. van Heesch, M. van Iterson, J. Jacobi, S. Boymans, P. B. Essers, E. de Bruijn, W. Hao, A. W. Macinnes, E. Cuppen, and M. Simonis. Extensive localization of long noncoding RNAs to the cytosol and mono- and polyribosomal complexes. *Genome Biol*, 15(1): R6, 2014.

39. R. Gil and A. Latorre. Factors behind junk DNA in bacteria. *Genes (Basel)*, 3(4): 634–650, 2012.

40. M. E. Dinger, K. C. Pang, T. R. Mercer, and J. S. Mattick. Differentiating protein-coding and noncoding RNA: challenges and ambiguities. *PLoS Comput Biol*, 4(11): e1000176, 2008.

41. Terence A. Brown. *Genomes*. Oxford: Wiley-Liss, 2nd edition, 2002. ISBN 0-471-25046-5.

42. Z. Yang and M. dos Reis. Statistical properties of the branch-site test of positive

selection. *Mol Biol Evol*, 28(3):1217–1228, 2011.

43. G. Huang, B. Gao, T. Maier, R. Allen, E. L. Davis, T. J. Baum, and R. S. Hussey. A profile of putative parasitism genes expressed in the esophageal gland cells of the root-knot nematode Meloidogyne incognita. *Mol Plant Microbe Interact*, 16(5):376–381, 2003.

44. K. D. Pruitt, G. R. Brown, et al. RefSeq: an update on mammalian reference sequences. *Nucleic Acids Res*, 42(Database issue):D756–D763, 2014.

45. S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, 25(17):3389–3402, 1997.

46. A. Loytynoja. Phylogeny-aware alignment with PRANK. *Methods Mol Biol*, 1079: 155–170, 2014.

47. P. Rice, I. Longden, and A. Bleasby. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet*, 16(6):276–277, 2000.

48. A. Goffeau, B. G. Barrell, H. Bussey, R. W. Davis, B. Dujon, H. Feldmann, F. Galibert, J. D. Hoheisel, C. Jacq, M. Johnston, E. J. Louis, H. W. Mewes, Y. Murakami, P. Philippsen, H. Tettelin, and S. G. Oliver. Life with 6000 genes. *Science*, 274(5287): 546, 563–546, 567, 1996.

49. M. Blaxter. Caenorhabditis elegans is a nematode. *Science*, 282(5396):2041–2046, 1998.

50. V. E. Velculescu, L. Zhang, B. Vogelstein, and K. W. Kinzler. Serial analysis of gene expression. *Science*, 270(5235):484–487, 1995.

51. M. Schena. Genome analysis with gene expression microarrays. *Bioessays*, 18(5): 427–431, 1996.

52. D. J. Lockhart, H. Dong, M. C. Byrne, M. T. Follettie, M. V. Gallo, M. S. Chee, M. Mittmann, C. Wang, M. Kobayashi, H. Horton, and E. L. Brown. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat Biotechnol*, 14 (13):1675–1680, 1996.

53. P. Liang and A. B. Pardee. Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction. *Science*, 257(5072):967–971, 1992.

54. C. W. Bachem, R. S. van der Hoeven, S. M. de Bruijn, D. Vreugdenhil, M. Zabeau, and R. G. Visser. Visualization of differential gene expression using a novel method of RNA fingerprinting based on AFLP: analysis of gene expression during potato tuber development. *Plant J*, 9(5):745–753, 1996.

55. H. Popeijus, V.C. Blok, L. Cardle, E. Bakker, M.S. Phillips, J. Helder, G. Smant, and J.T. Jones. Analysis of genes expressed in second stage juveniles of the potato cyst nematodes globodera rostochiensis and g. pallida using the expressed sequence tag approach. *Nematology*, pages 567–574, 2000.

56. J.N.A.M. Rouppe van der Voort, H.J. van Eck, P. van Zandvoort, H. Overmars, J. Helder, and J. Bakker. Linkage analysis by genotyping of sibling populations: a genetic map for the potato cyst nematode constructed using a £pseudo-f2£ mapping strategy. *Mol. Gen. Genet.*, 261:1021–1031, 1999.

57. G. Smant, J. P. Stokkermans, Y. Yan, J. M. de Boer, T. J. Baum, X. Wang, R. S. Hussey,

F. J. Gommers, B. Henrissat, E. L. Davis, J. Helder, A. Schots, and J. Bakker. Endogenous cellulases in animals: isolation of β-1,4-endoglucanase genes from two species of plant-parasitic cyst nematodes. *Proc Natl Acad Sci U S A*, 95(9):4906–4911, 1998.

58. L. Qin, H. Overmars, J. Helder, H. Popeijus, J. R. van der Voort, W. Groenink, P. van Koert, A. Schots, J. Bakker, and G. Smant. An efficient cDNA-AFLP-based strategy for the identification of putative pathogenicity factors from the potato cyst nematode Globodera rostochiensis. *Mol Plant Microbe Interact*, 13(8):830–836, 2000.

59. V. M. Williamson and R. S. Hussey. Nematode pathogenesis and resistance in plants. *Plant Cell*, 8(10):1735–1745, 1996.

60. H. Nielsen, J. Engelbrecht, S. Brunak, and G. von Heijne. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng*, 10 (1):1–6, 1997.

61. S. G. Penn, D. R. Rank, D. K. Hanzel, and D. L. Barker. Mining the human genome using microarrays of open reading frames. *Nat Genet*, 26(3):315–318, 2000.

62. M. Blaxter and L. Liu. Nematode spliced leaders–ubiquity, evolution and utility. *Int J Parasitol*, 26(10):1025–1033, 1996.

63. P. Vos, R. Hogers, M. Bleeker, M. Reijans, T. van de Lee, M. Hornes, A. Frijters, J. Pot, J. Peleman, M. Kuiper, and a. l. .. et. AFLP: a new technique for DNA fingerprinting. *Nucleic Acids Res*, 23(21):4407–4414, 1995.

64. R. Buschges, K. Hollricher, R. Panstruga, G. Simons, M. Wolter, A. Frijters, R. van Daelen, T. van der Lee, P. Diergaarde, J. Groenendijk, S. Topsch, P. Vos, F. Salamini, and P. Schulze-Lefert. The barley Mlo gene: a novel control element of plant pathogen resistance. *Cell*, 88(5):695–705, 1997.

65. Analysis of the genome sequence of the flowering plant arabidopsis thaliana. *Nature*, 408(6814):796–815, 2000.

66. V. Walbot and M. M. Evans. Unique features of the plant life cycle and their consequences. *Nat Rev Genet*, 4(5):369–379, 2003.

67. E. M. Lord and S. D. Russell. The mechanisms of pollination and fertilization in plants. *Annu Rev Cell Dev Biol*, 18:81–105, 2002.

68. T. Dresselhaus. Cell-cell communication during double fertilization. *Curr Opin Plant Biol*, 9(1):41–47, 2006.

69. F. Berger. Endosperm: the crossroad of seed development. *Curr Opin Plant Biol*, 6(1): 42–50, 2003.

70. G. Haughn and A. Chaudhury. Genetic analysis of seed coat development in Arabidopsis. *Trends Plant Sci*, 10(10):472–477, 2005.

71. A. J. Johnston, P. Meier, J. Gheyselinck, S. E. Wuest, M. Federer, E. Schlagenhauf, J. D. Becker, and U. Grossniklaus. Genetic subtraction profiling identifies genes essential for Arabidopsis reproduction and reveals interaction between the female gametophyte and the maternal sporophyte. *Genome Biol*, 8(10):R204, 2007.

72. R. J. Scott, M. Spielman, J. Bailey, and H. G. Dickinson. Parent-of-origin effects on seed development in Arabidopsis thaliana. *Development*, 125(17):3329–3341, 1998.

73. B. P. Dilkes and L. Comai. A differential dosage hypothesis for parental effects in seed

development. *Plant Cell*, 16(12):3174–3180, 2004.

74. U. Grossniklaus, J. P. Vielle-Calzada, M. A. Hoeppner, and W. B. Gagliano. Maternal control of embryogenesis by MEDEA, a polycomb group gene in Arabidopsis. *Science*, 280(5362):446–450, 1998.

75. M. T. Raissig, C. Baroux, and U. Grossniklaus. Regulation and flexibility of genomic imprinting during seed development. *Plant Cell*, 23(1):16–26, 2011.

76. T. Kinoshita, Y. Ikeda, and R. Ishikawa. Genomic imprinting: a balance between antagonistic roles of parental chromosomes. *Semin Cell Dev Biol*, 19(6):574–579, 2008.

77. O. Garnier, S. Laoueille-Duprat, and C. Spillane. Genomic imprinting in plants. *Epigenetics*, 3(1):14–20, 2008.

78. O'Connell, M. J. and Loughran, N. B. and Walsh, T. A. and Donoghue, M. T. and Schmid, K. J. and Spillane, C. A phylogenetic approach to test for evidence of parental conflict or gene duplications associated with protein-encoding imprinted orthologous genes in placental mammals. *Mamm Genome*, 21(9-10):486–498, 2010.

79. I. M. Morison, J. P. Ramsay, and H. G. Spencer. A census of mammalian imprinting. *Trends Genet*, 21(8):457–465, 2005.

80. J. P. Vielle-Calzada, J. Thomas, C. Spillane, A. Coluccio, M. A. Hoeppner, and U. Grossniklaus. Maintenance of genomic imprinting at the Arabidopsis medea locus requires zygotic DDM1 activity. *Genes Dev*, 13(22):2971–2982, 1999.

81. C. Kohler, L. Hennig, C. Spillane, S. Pien, W. Gruissem, and U. Grossniklaus. The Polycomb-group protein MEDEA regulates seed development by controlling expression of the MADS-box gene PHERES1. *Genes Dev*, 17(12):1540–1553, 2003.

82. S. Tiwari, R. Schulz, Y. Ikeda, L. Dytham, J. Bravo, L. Mathers, M. Spielman, P. Guzman, R. J. Oakey, T. Kinoshita, and R. J. Scott. Maternally Expressed PAB C-Terminal, a novel imprinted gene in Arabidopsis, encodes the conserved C-terminal domain of polyadenylate binding proteins. *Plant Cell*, 20(9):2387–2398, 2008.

83. M. Guo, M. A. Rupe, O. N. Danilevskaya, X. Yang, and Z. Hu. Genome-wide mRNA profiling reveals heterochronic allelic variation and a new imprinted gene in hybrid maize endosperm. *Plant J*, 36(1):30–44, 2003.

84. R. M. Stupar, P. J. Hermanson, and N. M. Springer. Nonadditive expression and parent-of-origin effects identified by microarray and allele-specific expression profiling of maize endosperm. *Plant Physiol*, 145(2):411–425, 2007.

85. R. Shirzadi, E. D. Andersen, K. N. Bjerkan, B. M. Gloeckle, M. Heese, A. Ungru, P. Winge, C. Koncz, R. B. Aalen, A. Schnittger, and P. E. Grini. Genome-wide transcript profiling of endosperm without paternal contribution identifies parent-of-origin-dependent regulation of AGAMOUS-LIKE36. *PLoS Genet*, 7(2):e1001303, 2011.

86. P. Wolff, I. Weinhofer, J. Seguin, P. Roszak, C. Beisel, M. T. Donoghue, C. Spillane, M. Nordborg, M. Rehmsmeier, and C. Kohler. High-resolution analysis of parent-of-origin allelic expression in the Arabidopsis Endosperm. *PLoS Genet*, 7(6):e1002126, 2011.

87. T. F. Hsieh, J. Shin, R. Uzawa, P. Silva, S. Cohen, M. J. Bauer, M. Hashimoto, R. C. Kirkbride, J. J. Harada, D. Zilberman, and R. L. Fischer. Regulation of imprinted gene expression in Arabidopsis endosperm. *Proc Natl Acad Sci U S A*, 108(5):1755–1762, 2011.

88. M. Gehring, K. L. Bubb, and S. Henikoff. Extensive demethylation of repetitive elements during seed development underlies gene imprinting. *Science*, 324(5933): 1447–1451, 2009.

89. T. Kinoshita, A. Miura, Y. Choi, Y. Kinoshita, X. Cao, S. E. Jacobsen, R. L. Fischer, and T. Kakutani. One-way control of FWA imprinting in Arabidopsis endosperm by DNA methylation. *Science*, 303(5657):521–523, 2004.

90. Y. Choi, M. Gehring, L. Johnson, M. Hannon, J. J. Harada, R. B. Goldberg, S. E. Jacobsen, and R. L. Fischer. DEMETER, a DNA glycosylase domain protein, is required for endosperm gene imprinting and seed viability in arabidopsis. *Cell*, 110(1):33–42, 2002.

91. C. Baroux, C. Spillane, and U. Grossniklaus. Genomic imprinting during seed development. *Adv Genet*, 46:165–214, 2002.

92. P. E. Jullien and F. Berger. Gamete-specific epigenetic mechanisms shape genomic imprinting. *Curr Opin Plant Biol*, 12(5):637–642, 2009.

93. C. B. Villar, A. Erilova, G. Makarevich, R. Trosch, and C. Kohler. Control of PHERES1 imprinting in Arabidopsis by direct tandem repeats. *Mol Plant*, 2(4):654–660, 2009.

94. D. Autran, C. Baroux, M. T. Raissig, T. Lenormand, M. Wittig, S. Grob, A. Steimer, M. Barann, U. C. Klostermeier, O. Leblanc, J. P. Vielle-Calzada, P. Rosenstiel, D. Grimanelli, and U. Grossniklaus. Maternal epigenetic pathways control parental contributions to Arabidopsis early embryogenesis. *Cell*, 145(5):707–719, 2011.

95. B. J. Haas, A. L. Delcher, S. M. Mount, J. R. Wortman, R. K. Jr Smith, L. I. Hannick, R. Maiti, C. M. Ronning, D. B. Rusch, C. D. Town, S. L. Salzberg, and O. White. Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res*, 31(19):5654–5666, 2003.

96. M. Pillot, C. Baroux, M. A. Vazquez, D. Autran, O. Leblanc, J. P. Vielle-Calzada, U. Grossniklaus, and D. Grimanelli. Embryo and endosperm inherit distinct chromatin and transcriptional states from the female gametes in Arabidopsis. *Plant Cell*, 22(2):307–320, 2010.

97. B. H. Le, C. Cheng, A. Q. Bui, J. A. Wagmaister, K. F. Henry, J. Pelletier, L. Kwong, M. Belmonte, R. Kirkbride, S. Horvath, G. N. Drews, R. L. Fischer, J. K. Okamuro, J. J. Harada, and R. B. Goldberg. Global analysis of gene activity during Arabidopsis seed development and identification of seed-specific transcription factors. *Proc Natl Acad Sci U S A*, 107(18):8063–8070, 2010.

98. R. C. Day, R. P. Herridge, B. A. Ambrose, and R. C. Macknight. Transcriptome analysis of proliferating Arabidopsis endosperm reveals biological implications for the control of syncytial division, cytokinin signaling, and gene expression regulation. *Plant Physiol*, 148(4):1964–1984, 2008.

99. C. Kohler, D. R. Page, V. Gagliardini, and U. Grossniklaus. The Arabidopsis thaliana MEDEA Polycomb group protein controls expression of PHERES1 by parental imprint-

ing. *Nat Genet*, 37(1):28–30, 2005.

100. B. Tycko. Allele-specific DNA methylation: beyond imprinting. *Hum Mol Genet*, 19 (R2):R210–R220, 2010.

101. E. L. Meaburn, L. C. Schalkwyk, and J. Mill. Allele-specific methylation in the human genome: implications for genetic studies of complex disease. *Epigenetics*, 5(7):578–582, 2010.

102. R. Shoemaker, J. Deng, W. Wang, and K. Zhang. Allele-specific methylation is prevalent and is contributed by CpG-SNPs in the human genome. *Genome Res*, 20(7): 883–889, 2010.

103. L. C. Schalkwyk, E. L. Meaburn, R. Smith, E. L. Dempster, A. R. Jeffries, M. N. Davies, R. Plomin, and J. Mill. Allelic skewing of DNA methylation is widespread across the genome. *Am J Hum Genet*, 86(2):196–212, 2010.

104. Y. Kinoshita, H. Saze, T. Kinoshita, A. Miura, W. J. Soppe, M. Koornneef, and T. Kakutani. Control of FWA gene silencing in Arabidopsis thaliana by SINE-related direct repeats. *Plant J*, 49(1):38–45, 2007.

105. T. F. Hsieh, C. A. Ibarra, P. Silva, A. Zemach, L. Eshed-Williams, R. L. Fischer, and D. Zilberman. Genome-wide demethylation of Arabidopsis endosperm. *Science*, 324 (5933):1451–1454, 2009.

106. C. Baroux, V. Gagliardini, D. R. Page, and U. Grossniklaus. Dynamic regulatory interactions of Polycomb group genes: MEDEA autoregulation is required for imprinted gene expression in Arabidopsis. *Genes Dev*, 20(9):1081–1086, 2006.

107. M. Gehring, J. H. Huh, T. F. Hsieh, J. Penterman, Y. Choi, J. J. Harada, R. B. Goldberg, and R. L. Fischer. DEMETER DNA glycosylase establishes MEDEA polycomb gene self-imprinting by allele-specific demethylation. *Cell*, 124(3):495–506, 2006.

108. P. E. Jullien, T. Kinoshita, N. Ohad, and F. Berger. Maintenance of DNA methylation during the Arabidopsis life cycle is essential for parental imprinting. *Plant Cell*, 18 (6):1360–1372, 2006.

109. H. Wenz, J. M. Robertson, S. Menchen, F. Oaks, D. M. Demorest, D. Scheibler, B. B. Rosenblum, C. Wike, D. A. Gilbert, and J. W. Efcavitch. High-precision genotyping by denaturing capillary electrophoresis. *Genome Res*, 8(1):69–80, 1998.

110. R. J. Cho, M. Huang, M. J. Campbell, H. Dong, L. Steinmetz, L. Sapinoso, G. Hampton, S. J. Elledge, R. W. Davis, and D. J. Lockhart. Transcriptional regulation and function during the human cell cycle. *Nat Genet*, 27(1):48–54, 2001.

111. R. Fukumura, H. Takahashi, T. Saito, Y. Tsutsumi, A. Fujimori, S. Sato, K. Tatsumi, R. Araki, and M. Abe. A sensitive transcriptome analysis method that can detect unknown transcripts. *Nucleic Acids Res*, 31(16):e94, 2003.

112. M. Reijans, R. Lascaris, A. O. Groeneger, A. Wittenberg, E. Wesselink, J. van Oeveren, E. de Wit, A. Boorsma, B. Voetdijk, H. van der Spek, L. A. Grivell, and G. Simons. Quantitative comparison of cDNA-AFLP, microarrays, and GeneChip expression data in Saccharomyces cerevisiae. *Genomics*, 82(6):606–618, 2003.

113. S. Rombauts, Y. Van De Peer, and P. Rouze. AFLPinSilico, simulating AFLP fingerprints. *Bioinformatics*, 19(6):776–777, 2003.

114. L. Qin, P. Prins, and J. Helder. Linking cDNA-AFLP-based gene expression patterns and ESTs. *Methods Mol Biol*, 317:123–138, 2006.

115. L. Qin, P. Prins, J. T. Jones, H. Popeijus, G. Smant, J. Bakker, and J. Helder. GenEST, a powerful bidirectional link between cDNA sequence data and gene expression profiles generated by cDNA-AFLP. *Nucleic Acids Res*, 29(7):1616–1622, 2001.

116. J. Gribnau, K. Hochedlinger, K. Hata, E. Li, and R. Jaenisch. Asynchronous replication timing of imprinted loci is independent of DNA methylation, but consistent with differential subnuclear localization. *Genes Dev*, 17(6):759–773, 2003.

117. J. N. Fitz Gerald, P. S. Hui, and F. Berger. Polycomb group-dependent imprinting of the actin regulator AtFH5 regulates morphogenesis in Arabidopsis thaliana. *Development*, 136(20):3399–3404, 2009.

118. A. Deichsel, J. Mouysset, and T. Hoppe. The ubiquitin-selective chaperone CDC-48/p97, a new player in DNA replication. *Cell Cycle*, 8(2):185–190, 2009.

119. S. Park, D. M. Rancour, and S. Y. Bednarek. In planta analysis of the cell cycle-dependent localization of AtCDC48A and its critical roles in cell division, expansion, and differentiation. *Plant Physiol*, 148(1):246–258, 2008.

120. J. Aker, J. W. Borst, R. Karlova, and S. de Vries. The Arabidopsis thaliana AAA protein CDC48A interacts in vivo with the somatic embryogenesis receptor-like kinase 1 receptor at the plasma membrane. *J Struct Biol*, 156(1):62–71, 2006.

121. J. Aker, R. Hesselink, R. Engel, R. Karlova, J. W. Borst, A. J. Visser, and S. C. de Vries. In vivo hexamerization and characterization of the Arabidopsis AAA ATPase CDC48A complex using forster resonance energy transfer-fluorescence lifetime imaging microscopy and fluorescence correlation spectroscopy. *Plant Physiol*, 145(2):339–350, 2007.

122. D. M. Rancour, S. Park, S. D. Knight, and S. Y. Bednarek. Plant UBX domain-containing protein 1, PUX1, regulates the oligomeric structure and activity of arabidopsis CDC48. *J Biol Chem*, 279(52):54264–54274, 2004.

123. P. E. Jullien and F. Berger. Parental genome dosage imbalance deregulates imprinting in Arabidopsis. *PLoS Genet*, 6(3):e1000885, 2010.

124. David Haig and Mark Westoby. Genomic imprinting in endosperm: Its effect on seed development in crosses between species, and between different ploidies of the same species, and its implications for the evolution of apomixis. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 333(1266):1–13, 1991. ISSN 0962-8436. doi: 10.1098/rstb.1991.0057.

125. J. Glover, M. Grelon, S. Craig, A. Chaudhury, and E. Dennis. Cloning and characterization of MS5 from Arabidopsis: a gene critical in male meiosis. *Plant J*, 15(3): 345–356, 1998.

126. J. R. Howarth, S. Parmar, P. B. Barraclough, and M. J. Hawkesford. A sulphur deficiency-induced gene, sdi1, involved in the utilization of stored sulphate pools under sulphur-limiting conditions has potential as a diagnostic indicator of sulphur nutritional status. *Plant Biotechnol J*, 7(2):200–209, 2009.

127. I. Tzafrir, A. Dickerman, O. Brazhnik, Q. Nguyen, J. McElver, C. Frye, D. Patton, and D. Meinke. The Arabidopsis SeedGenes Project. *Nucleic Acids Res*, 31(1):90–93, 2003.

128. I. Tzafrir, R. Pena-Muralla, A. Dickerman, M. Berg, R. Rogers, S. Hutchens, T. C. Sweeney, J. McElver, G. Aux, D. Patton, and D. Meinke. Identification of genes required for embryo development in Arabidopsis. *Plant Physiol*, 135(3):1206–1220, 2004.

129. J. Bedard, S. Kubis, S. Bimanadham, and P. Jarvis. Functional similarity between the chloroplast translocon component, Tic40, and the human co-chaperone, Hsp70-interacting protein (Hip). *J Biol Chem*, 282(29):21404–21414, 2007.

130. J. B. Wolf. Cytonuclear interactions can favor the evolution of genomic imprinting. *Evolution*, 63(5):1364–1371, 2009.

131. R. J. Roberts, T. Vincze, J. Posfai, and D. Macelis. REBASE–restriction enzymes and DNA methyltransferases. *Nucleic Acids Res*, 33(Database issue):D230–D232, 2005.

132. N. Goto, P. Prins, M. Nakao, R. Bonnal, J. Aerts, and T. Katayama. BioRuby: bioinformatics software for the Ruby programming language. *Bioinformatics*, 26(20):2617–2619, 2010.

133. T. Barrett, D. B. Troup, S. E. Wilhite, P. Ledoux, D. Rudnev, C. Evangelista, I. F. Kim, A. Soboleva, M. Tomashevsky, K. A. Marshall, K. H. Phillippy, P. M. Sherman, R. N. Muertter, and R. Edgar. NCBI GEO: archive for high-throughput functional genomic data. *Nucleic Acids Res*, 37(Database issue):D885–D890, 2009.

134. D. Swarbreck, C. Wilks, P. Lamesch, T. Z. Berardini, M. Garcia-Hernandez, H. Foerster, D. Li, T. Meyer, R. Muller, L. Ploetz, A. Radenbaugh, S. Singh, V. Swing, C. Tissier, P. Zhang, and E. Huala. The Arabidopsis Information Resource (TAIR): gene structure and function annotation. *Nucleic Acids Res*, 36(Database issue):D1009–D1014, 2008.

135. R Development Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2010.

136. K. W. Broman, H. Wu, S. Sen, and G. A. Churchill. R/qtl: QTL mapping in experimental crosses. *Bioinformatics*, 19(7):889–890, 2003.

137. K.W. Broman and Ś. Sen. *A Guide to QTL Mapping with R/qtl*. Springer Verlag, 2009. ISBN 0387921249.

138. J. Fu, M. A. Swertz, J. J. Keurentjes, and R. C. Jansen. MetaNetwork: a computational protocol for the genetic study of metabolic networks. *Nat Protoc*, 2(3):685–694, 2007.

139. Y. Li, O. A. Alvarez, E. W. Gutteling, M. Tijsterman, J. Fu, J. A. Riksen, E. Hazendonk, P. Prins, R. H. Plasterk, R. C. Jansen, R. Breitling, and J. E. Kammenga. Mapping determinants of gene expression plasticity by genetical genomics in C. elegans. *PLoS Genet*, 2(12):e222, 2006.

140. C. S. Haley and S. A. Knott. A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. *Heredity*, 69(4):315–324, 1992.

141. Z. B. Zeng. Precision mapping of quantitative trait loci. *Genetics*, 136(4):1457–1468, 1994.

142. R. C. Jansen. Interval mapping of multiple quantitative trait loci. *Genetics*, 135(1):205–211, 1993.

143. R. C. Jansen and P. Stam. High resolution of quantitative traits into multiple loci via interval mapping. *Genetics*, 136(4):1447–1455, 1994.

144. B. S. Yandell, T. Mehta, S. Banerjee, D. Shriner, R. Venkataraman, J. Y. Moon, W. W. Neely, H. Wu, R. von Smith, and N. Yi. R/qtlbim: QTL with Bayesian Interval Mapping in experimental crosses. *Bioinformatics*, 23(5):641–643, 2007.

145. S. Banerjee, B. S. Yandell, and N. Yi. Bayesian quantitative trait loci mapping for multiple traits. *Genetics*, 179(4):2275–2289, 2008.

146. R. C. Jansen. Quantitative trait loci in inbred lines. In D. J. Balding, M. Bishop, and C. Cannings, editors, *Handbook of Statistical Genetics*, pages 589-622. John Wiley & Sons, Ltd., 2007.

147. R. C. Jansen. Controlling the type I and type II errors in mapping quantitative trait loci. *Genetics*, 138(3):871–881, 1994.

148. J. W. Van Ooijen, M. P. Boer, R. C. Jansen, and C. Maliepaard. MapQTL 4.0, Software for the Calculation of QTL Position on Genetic Maps, 2002.

149. J. G. de Mooij-van Malsen, H. A. van Lith, H. Oppelaar, B. Olivier, and M. J. Kas. Evidence for epigenetic interactions for loci on mouse chromosome 1 regulating open field activity. *Behav Genet*, 39(2):176–182, 2009.

150. M. J. Jeuken, N. W. Zhang, L. K. McHale, K. Pelgrom, E. den Boer, P. Lindhout, R. W. Michelmore, R. G. Visser, and R. E. Niks. Rin4 causes hybrid necrosis and race-specific resistance in an interspecific lettuce hybrid. *Plant Cell*, 21(10):3368–3378, 2009.

151. J. Kitano, J. A. Ross, S. Mori, M. Kume, F. C. Jones, Y. F. Chan, D. M. Absher, J. Grimwood, J. Schmutz, R. M. Myers, D. M. Kingsley, and C. L. Peichel. A role for a neo-sex chromosome in stickleback speciation. *Nature*, 461(7267):1079–1083, 2009.

152. R. Breitling, Y. Li, B. M. Tesson, J. Fu, C. Wu, T. Wiltshire, A. Gerrits, L. V. Bystrykh, G. de Haan, A. I. Su, and R. C. Jansen. Genetical genomics: spotlight on QTL hotspots. *PLoS Genet*, 4(10):e1000232, 2008.

153. L. Tierney, A.J. Rossini, and N. Li. SNOW: a parallel computing framework for the R system. *International Journal of Parallel Programming*, 37(1):78–90, 2009. ISSN 0885-7458.

154. S. Nandi, P. K. Subudhi, D. Senadhira, N. L. Manigbas, S. Sen-Mandi, and N. Huang. Mapping QTLs for submergence tolerance in rice by AFLP analysis and selective genotyping. *Mol Gen Genet*, 255(1):1–8, 1997.

155. E. Meaburn, L. M. Butcher, L. C. Schalkwyk, and R. Plomin. Genotyping pooled DNA using 100K SNP microarrays: a step towards genomewide association scans. *Nucleic Acids Res*, 34(4):e27, 2006.

156. S. Kim, V. Plagnol, T. T. Hu, C. Toomajian, R. M. Clark, S. Ossowski, J. R. Ecker, D. Weigel, and M. Nordborg. Recombination and linkage disequilibrium in Arabidopsis thaliana. *Nat Genet*, 39(9):1151–1155, 2007.

157. A. L. Dixon, L. Liang, M. F. Moffatt, W. Chen, S. Heath, K. C. Wong, J. Taylor, E. Burnett, I. Gut, M. Farrall, G. M. Lathrop, G. R. Abecasis, and W. O. Cookson. A genome-wide association study of global gene expression. *Nat Genet*, 39(10):1202–1207, 2007.

158. H. J. Westra, R. C. Jansen, R. S. Fehrmann, G. J. te Meerman, D. van Heel, C. Wijmenga, and L. Franke. MixupMapper: correcting sample mix-ups in genome-wide

datasets increases power to detect small genetic effects. *Bioinformatics*, 27(15):2104–2111, 2011.

159. G. Gibson and B. Weir. The quantitative genetics of transcription. *Trends Genet*, 21 (11):616–623, 2005.

160. R. C. Jansen, B. M. Tesson, J. Fu, Y. Yang, and L. M. McIntyre. Defining gene and QTL networks. *Curr Opin Plant Biol*, 12(2):241–246, 2009.

161. R. B. Brem and L. Kruglyak. The landscape of genetic complexity across 5,700 gene expression traits in yeast. *Proc Natl Acad Sci U S A*, 102(5):1572–1577, 2005.

162. H. B. Fraser, A. M. Moses, and E. E. Schadt. Evidence for widespread adaptive evolution of gene expression in budding yeast. *Proc Natl Acad Sci U S A*, 107(7):2977–2982, 2010.

163. Y. Zou, Z. Su, J. Yang, Y. Zeng, and X. Gu. Uncovering genetic regulatory network divergence between duplicate genes using yeast eQTL landscape. *J Exp Zool B Mol Dev Evol*, 312(7):722–733, 2009.

164. Y. Li, R. Breitling, and R. C. Jansen. Generalizing genetical genomics: getting added value from environmental perturbation. *Trends Genet*, 24(10):518–524, 2008.

165. D. J. Kliebenstein, M. A. West, H. van Leeuwen, O. Loudet, R. W. Doerge, and D. A. St Clair. Identification of QTLs controlling gene expression networks defined a priori. *BMC Bioinformatics*, 7:308, 2006.

166. Y. Gilad, S. A. Rifkin, and J. K. Pritchard. Revealing the architecture of gene regulation: the promise of eQTL studies. *Trends Genet*, 24(8):408–415, 2008.

167. R. Alberts, P. Terpstra, Y. Li, R. Breitling, J. P. Nap, and R. C. Jansen. Sequence polymorphisms cause many false cis eQTLs. *PLoS One*, 2(7):e622, 2007.

168. L. Franke, H. van Bakel, L. Fokkens, E. D. de Jong, M. Egmont-Petersen, and C. Wijmenga. Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. *Am J Hum Genet*, 78(6):1011–1025, 2006.

169. X. Chen, C. A. Hackett, R. E. Niks, P. E. Hedley, C. Booth, A. Druka, T. C. Marcel, A. Vels, M. Bayer, I. Milne, J. Morris, L. Ramsay, D. Marshall, L. Cardle, and R. Waugh. An eQTL analysis of partial resistance to Puccinia hordei in barley. *PLoS One*, 5(1): e8598, 2010.

170. L. Qin, U. Kudla, E. H. Roze, A. Goverse, H. Popeijus, J. Nieuwland, H. Overmars, J. T. Jones, A. Schots, G. Smant, J. Bakker, and J. Helder. Plant degradation: a nematode expansin acting on plants. *Nature*, 427(6969):30, 2004.

171. Y. Saijo and P. Schulze-Lefert. Manipulation of the eukaryotic transcriptional machinery by bacterial pathogens. *Cell Host Microbe*, 4(2):96–99, 2008.

172. L. Q. Chen, B. H. Hou, S. Lalonde, H. Takanaga, M. L. Hartung, X. Q. Qu, W. J. Guo, J. G. Kim, W. Underwood, B. Chaudhuri, D. Chermak, G. Antony, F. F. White, S. C. Somerville, M. B. Mudgett, and W. B. Frommer. Sugar transporters for intercellular exchange and nutrition of pathogens. *Nature*, 468(7323):527–532, 2010.

173. J. P. Hewitson, J. R. Grainger, and R. M. Maizels. Helminth immunoregulation: the role of parasite secreted proteins in modulating host immunity. *Mol Biochem Parasitol*, 167(1):1–11, 2009.

174. D. M. Bird and C. H. Opperman. The secret(ion) life of worms. *Genome Biol*, 10(1): 205, 2009.

175. M. Bevan, I. Bancroft, et al. Analysis of 1.9 Mb of contiguous sequence from chromosome 4 of Arabidopsis thaliana. *Nature*, 391(6666):485–488, 1998.

176. J. G. Bishop, A. M. Dean, and T. Mitchell-Olds. Rapid evolution in plant chitinases: molecular targets of selection in plant-pathogen coevolution. *Proc Natl Acad Sci U S A*, 97(10):5322–5327, 2000.

177. H.H. Flor. The Complementary Genic Systems in Flax and Flax Rust*. *Advances in genetics*, 8:29–54, 1956. ISSN 0065-2660.

178. E. G. Bakker, C. Toomajian, M. Kreitman, and J. Bergelson. A genome-wide survey of R gene polymorphisms in Arabidopsis. *Plant Cell*, 18(8):1803–1818, 2006.

179. D. Mackey, Y. Belkhadir, J. M. Alonso, J. R. Ecker, and J. L. Dangl. Arabidopsis RIN4 is a target of the type III virulence effector AvrRpt2 and modulates RPS2-mediated resistance. *Cell*, 112(3):379–389, 2003.

180. E. Richly, J. Kurth, and D. Leister. Mode of amplification and reorganization of resistance genes during recent Arabidopsis thaliana evolution. *Mol Biol Evol*, 19(1): 76–84, 2002.

181. R. Medzhitov and C. A. Jr. Janeway. Innate immunity: impact on the adaptive immune response. *Curr Opin Immunol*, 9(1):4–9, 1997.

182. E. B. Holub. The arms race is ancient history in Arabidopsis, the wildflower. *Nat Rev Genet*, 2(7):516–527, 2001.

183. S. Xiao, B. Emerson, K. Ratanasut, E. Patrick, C. O'Neill, I. Bancroft, and J. G. Turner. Origin and maintenance of a broad-spectrum disease resistance locus in Arabidopsis. *Mol Biol Evol*, 21(9):1661–1672, 2004.

184. M. Mondragon-Palomino, B. C. Meyers, R. W. Michelmore, and B. S. Gaut. Patterns of positive selection in the complete nbs-lrr gene family of Arabidopsis thaliana. *Genome Res*, 12(9):1305–1315, 2002.

185. X. Sun, Y. Cao, and S. Wang. Point mutations with positive selection were a major force during the evolution of a receptor-kinase resistance gene family of rice. *Plant Physiol*, 140(3):998–1008, 2006.

186. R. C. Edgar. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*, 32(5):1792–1797, 2004.

187. M. Suyama, D. Torrents, and P. Bork. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res*, 34 (Web Server issue):W609–W612, 2006.

188. S. Y. Rhee, W. Beavis, T. Z. Berardini, G. Chen, D. Dixon, A. Doyle, M. Garcia-Hernandez, E. Huala, G. Lander, M. Montoya, N. Miller, L. A. Mueller, S. Mundodi, L. Reiser, J. Tacklind, D. C. Weems, Y. Wu, I. Xu, D. Yoo, J. Yoon, and P. Zhang. The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community. *Nucleic Acids Res*, 31(1):224–228, 2003.

189. M. Anisimova, R. Nielsen, and Z. Yang. Effect of recombination on the accuracy of

the likelihood method for detecting positive selection at amino acid sites. *Genetics*, 164(3):1229–1236, 2003.

190. R. W. Michelmore and B. C. Meyers. Clusters of resistance genes in plants evolve by divergent selection and a birth-and-death process. *Genome Res*, 8(11):1113–1130, 1998.

191. J. Salinas and J.J. Sanchez-Serrano. *Arabidopsis protocols*. Humana Pr Inc, 2006. ISBN 1588293955.

192. J. Fu and R. C. Jansen. Optimal design and analysis of genetic studies on gene expression. *Genetics*, 172(3):1993–1999, 2006.

193. A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods*, 5(7):621–628, 2008.

194. A. C. Eklund, L. R. Turner, P. Chen, R. V. Jensen, G. deFeo, A. R. Kopf-Sill, and Z. Szallasi. Replacing cRNA targets with cDNA reduces microarray cross-hybridization. *Nat Biotechnol*, 24(9):1071–1073, 2006.

195. P. A. Hoen, Y. Ariyurek, H. H. Thygesen, E. Vreugdenhil, R. H. Vossen, R. X. de Menezes, J. M. Boer, G. J. van Ommen, and J. T. den Dunnen. Deep sequencing-based expression analysis shows major advances in robustness, resolution and inter-lab portability over five microarray platforms. *Nucleic Acids Res*, 36(21):e141, 2008.

196. J. J. Keurentjes, R. Sulpice, Y. Gibon, M. C. Steinhauser, J. Fu, M. Koornneef, M. Stitt, and D. Vreugdenhil. Integrative analyses of genetic variation in enzyme activities of primary carbohydrate metabolism reveal distinct modes of regulation in Arabidopsis thaliana. *Genome Biol*, 9(8):R129, 2008.

197. J. Fu, J. J. Keurentjes, H. Bouwmeester, T. America, F. W. Verstappen, J. L. Ward, M. H. Beale, R. C. de Vos, M. Dijkstra, R. A. Scheltema, F. Johannes, M. Koornneef, D. Vreugdenhil, R. Breitling, and R. C. Jansen. System-wide molecular evidence for phenotypic buffering in Arabidopsis. *Nat Genet*, 41(2):166–167, 2009.

198. D. Arends, P. Prins, R. C. Jansen, and K. W. Broman. R/qtl: high-throughput multiple QTL mapping. *Bioinformatics*, 26(23):2990–2992, 2010.

199. D. Arends, P. Prins, K. W. Broman, and R. C. Jansen. Tutorial - Multiple-QTL Mapping (MQM) Analysis, 2010.

200. Y. Li, B. M. Tesson, G. A. Churchill, and R. C. Jansen. Critical reasoning on causal inference in genome-wide linkage and association studies. *Trends Genet*, 26(12): 493–498, 2010.

201. M. L. Wayne and L. M. McIntyre. Combining mapping and arraying: An approach to candidate gene identification. *Proc Natl Acad Sci U S A*, 99(23):14903–14906, 2002.

202. M. Kanehisa, M. Araki, S. Goto, M. Hattori, M. Hirakawa, M. Itoh, T. Katayama, S. Kawashima, S. Okuda, T. Tokimatsu, and Y. Yamanishi. KEGG for linking genomes to life and the environment. *Nucleic Acids Res*, 36(Database issue):D480–D484, 2008.

203. J. E. Stajich and H. Lapp. Open source tools and toolkits for bioinformatics: significance, and where are we? *Brief Bioinform*, 7(3):287–296, 2006.

204. J. E. Stajich, D. Block, K. Boulez, S. E. Brenner, S. A. Chervitz, C. Dagdigian, G. Fu-

ellen, J. G. Gilbert, I. Korf, H. Lapp, H. Lehvaslaiho, C. Matsalla, C. J. Mungall, B. I. Osborne, M. R. Pocock, P. Schattner, M. Senger, L. D. Stein, E. Stupka, M. D. Wilkinson, and E. Birney. The Bioperl toolkit: Perl modules for the life sciences. *Genome Res*, 12(10):1611–1618, 2002.

205. P. J. Cock, T. Antao, J. T. Chang, B. A. Chapman, C. J. Cox, A. Dalke, I. Friedberg, T. Hamelryck, F. Kauff, B. Wilczynski, and M. J. de Hoon. Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422–1423, 2009.

206. R. C. Holland, T. A. Down, M. Pocock, A. Prlic, D. Huen, K. James, S. Foisy, A. Drager, A. Yates, M. Heuer, and M. J. Schreiber. BioJava: an open-source framework for bioinformatics. *Bioinformatics*, 24(18):2096–2097, 2008.

207. P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*, 13(11):2498–2504, Nov 2003.

208. J. Aerts and A. Law. An introduction to scripting in Ruby for biologists. *BMC Bioinformatics*, 10:221, 2009.

209. K. Katoh, K. Kuma, H. Toh, and T. Miyata. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res*, 33(2):511–518, 2005.

210. S. Lee and T. L. Blundell. Ulla: a program for calculating environment-specific amino acid substitution tables. *Bioinformatics*, 25(15):1976–1977, 2009.

211. Z. Metlagel, Y. S. Kikkawa, and M. Kikkawa. Ruby-Helix: an implementation of helical image processing based on object-oriented scripting language. *J Struct Biol*, 157(1): 95–105, 2007.

212. J. T. Prince and E. M. Marcotte. mspire: mass spectrometry proteomics in Ruby. *Bioinformatics*, 24(23):2796–2797, 2008.

213. S. Philippi. Light-weight integration of molecular biological databases. *Bioinformatics*, 20(1):51–57, 2004.

214. A. Jacobsen, A. Krogh, S. Kauppinen, and M. Lindow. miRMaid: a unified programming interface for microRNA data resources. *BMC Bioinformatics*, 11:29, 2010.

215. A. Biegert, C. Mayer, M. Remmert, J. Soding, and A. N. Lupas. The MPI Bioinformatics Toolkit for protein sequence analysis. *Nucleic Acids Res*, 34(Web Server issue):W335–W339, 2006.

216. Toshiaki Katayama, Mitsuteru Nakao, and Toshihisa Takagi. TogoWS: integrated SOAP and REST APIs for interoperable bioinformatics Web services. *Nucleic Acids Res.*, 38(Web Server issue):W706–W711, 2010.

217. P. J. Cock, C. J. Fields, N. Goto, M. L. Heuer, and P. M. Rice. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Res*, 38(6):1767–1771, 2010.

218. M. V. Han and C. M. Zmasek. phyloXML: XML for evolutionary biology and comparative genomics. *BMC Bioinformatics*, 10:356, 2009.

219. R. C. Holland, T. A. Down, M. Pocock, A. Prlic, D. Huen, K. James, S. Foisy, A. Drager,

A. Yates, M. Heuer, and M. J. Schreiber. BioJava: an open-source framework for bioinformatics. *Bioinformatics*, 24(18):2096–2097, 2008.

220. Robert C Gentleman, Vincent J. Carey, Douglas M. Bates, and others. Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biology*, 5:R80, 2004.

221. Viswa Viswanathan. Rapid web application development: A ruby on rails tutorial. *IEEE Software*, 25(6):98–106, 2008.

222. S. Möller, H.N. Krabbenhöft, A. Tille, D. Paleino, A Williams, K. Wolstencroft, G. Goble, R. Holland, D. Belhachemi, and C. Plessy. Community-driven computational biology with Debian Linux. *BMC Bioinformatics*, 11 Suppl 12:S5, 2010.

223. D. Field, B. Tiwari, T. Booth, S. Houten, D. Swan, N. Bertrand, and M. Thurston. Open software for biologists: from famine to feast. *Nat Biotechnol*, 24(7):801–803, 2006.

224. David Chelimsky. *The RSpec Book (incomplet ref)*. 2007?

225. E. Afgan, D. Baker, N. Coraor, B. Chapman, A. Nekrutenko, and J. Taylor. Galaxy CloudMan: delivering cloud compute clusters. *BMC Bioinformatics*, 11 Suppl 12:S4, 2010.

226. R. R. Gullapalli, K. V. Desai, L. Santana-Santos, J. A. Kant, and M. J. Becich. Next generation sequencing in clinical medicine: Challenges and lessons for pathology and biomedical informatics. *J Pathol Inform*, 3:40, 2012.

227. O. Trelles, P. Prins, M. Snir, and R. C. Jansen. Big data, but are we ready? *Nat Rev Genet*, 12(3):224, 2011.

228. N. Siva. 1000 Genomes project. *Nat Biotechnol*, 26(3):256, 2008.

229. H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, 2009.

230. G. Cochrane, B. Alako, et al. Facing growth in the European Nucleotide Archive. *Nucleic Acids Res*, 41(Database issue):D30–D35, 2013.

231. Picard. Picard. http://picard.sourceforge.net/, 2009.

232. G. G. Faust and I. M. Hall. SAMBLASTER: fast duplicate marking and structural variant read extraction. *Bioinformatics*, 30(17):2503–2505, 2014.

233. G. Tischler and S. Leonard. biobambam: tools for read pair collation based algorithms on BAM files. *Source Code for Biology and Medicine*, 9(13), 2014.

234. J. K. Bonfield. The Scramble conversion tool. *Bioinformatics*, 30(19):2818–2819, 2014.

235. Andrei Alexandrescu. *The D programming language*. Addison-Wesley, Upper Saddle River, NJ, 2010. ISBN 9780321635365.

236. Pjotr Prins et al. Small tools MANIFESTO for Bioinformatics. doi: 10.5281/zenodo.11321, Aug. 2014.

237. Adrian D. Thurston. Parsing computer languages with an automaton compiled from a single regular expression. In Oscar H. Ibarra and Hsu-Chun Yen, editors, *CIAA*, volume 4094 of *Lecture Notes in Computer Science*, pages 285–286. Springer, 2006.

ISBN 3-540-37213-X.

238. E. E. Schadt, M. D. Linderman, J. Sorenson, L. Lee, and G. P. Nolan. Computational solutions to large-scale data management and analysis. *Nat Rev Genet*, 11(9):647–657, 2010.

239. Grancher E. Oracle and storage ios, explanations and experience at cern. *Journal of Physics: Conference Series*, 219(5):052004, 2010.

240. G. S. Davidson, K. W. Boyack, R. A. Zacharski, S.C. Helmreich, and Cowie. J. R. *Data-Centric Computing with the Netezza Architecture*. Sandia Report, 2006.

241. Y. Vlasov, W. M. J. Green, and F. Xia. High-throughput silicon nanophotonic wavelength-insensitive switch for on-chip optical networks. *Nature Photonics*, (2): 242 – 246, 2008.

242. G. T. Reed. *Silicon Photonics: The State of the Art*. Wiley-Interscience, New York, NY, USA, 2008. ISBN 0470025794, 9780470025796.

243. T. Mattson, B. Sanders, and B. Massingill. *Patterns for parallel programming*. Addison-Wesley Professional, 2004. ISBN 0321228111.

244. T. Harris, A. Cristal, O. S. Unsal, E. Ayguade, F. Gagliardi, B. Smith, and M. Valero. Transactional memory: An overview. *IEEE Micro*, 27:8–29, 2007. ISSN 0272-1732. doi: 10.1109/MM.2007.63.

245. J. M. Kraus and H. A. Kestler. Multi-core parallelization in clojure: a case study. In *Proceedings of the 6th European Lisp Workshop*, ELW '09, pages 8–17, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-539-0.

246. J. Armstrong. *Programming Erlang: Software for a Concurrent World*. Pragmatic Bookshelf, 2007. ISBN 193435600X, 9781934356005.

247. P. Haller and M. Odersky. Scala actors: Unifying thread-based and event-based programming. *Theoretical Computer Science*, 410(2-3):202 – 220, 2009. ISSN 0304-3975.

248. G. Wang and T. E. S. Ng. The impact of virtualization on network performance of amazon ec2 data center. In *Proc. IEEE INFOCOM*, San Diego, CA, March 2010.

249. B. Snijder, R. K. Kandasamy, and G. Superti-Furga. Toward effective sharing of high-dimensional immunology data. *Nat Biotechnol*, 32(8):755–759, 2014.

250. F. S. Collins and L. A. Tabak. Policy: NIH plans to enhance reproducibility. *Nature*, 505(7485):612–613, 2014.

251. V. Marx. Biology: The big challenges of big data. *Nature*, 498(7453):255–260, 2013.

252. M. Watson. You're not allowed bioinformatics anymore. http://bio-mickwatson.word-press.com/2014/07/21/youre-not-allowed-bioinformatics-anymore/, July 2014.

253. S. Moller, E. Afgan, M. Banck, R. J. Bonnal, T. Booth, J. Chilton, P. J. Cock, M. Gumbel, N. Harris, R. Holland, M. Kalas, L. Kajan, E. Kibukawa, D. R. Powel, P. Prins, J. Quinn, O. Sallou, F. Strozzi, T. Seemann, C. Sloggett, S. Soiland-Reyes, W. Spooner, S. Steinbiss, A. Tille, A. J. Travis, R. Guimera, T. Katayama, and B. A. Chapman. Community-driven development for computational biology at Sprints, Hackathons and Codefests. *BMC Bioinformatics*, 15 Suppl 14:S7, 2014.

254. M. A. White, B. Steffy, T. Wiltshire, and B. A. Payseur. Genetic dissection of a key reproductive barrier between nascent species of house mice. *Genetics*, 189(1):289–304, 2011.

255. K. Choi, X. Zhao, K. A. Kelly, O. Venn, J. D. Higgins, N. E. Yelina, T. J. Hardcastle, P. A. Ziolkowski, G. P. Copenhaver, F. C. Franklin, G. McVean, and I. R. Henderson. Arabidopsis meiotic crossover hot spots overlap with H2A.Z nucleosomes at gene promoters. *Nat Genet*, 45(11):1327–1336, 2013.

256. N. Khan, R. H. Kazmi, L. A. Willems, A. W. van Heusden, W. Ligterink, and H. W. Hilhorst. Exploring the natural variation for seedling traits and their link with seed dimensions in tomato. *PLoS One*, 7(8):e43991, 2012.

257. S. Nygaard, G. Zhang, M. Schiott, C. Li, Y. Wurm, H. Hu, J. Zhou, L. Ji, F. Qiu, M. Rasmussen, H. Pan, F. Hauser, A. Krogh, C. J. Grimmelikhuijzen, J. Wang, and J. J. Boomsma. The genome of the leaf-cutting ant Acromyrmex echinatior suggests key adaptations to advanced social life and fungus farming. *Genome Res*, 21(8):1339–1348, 2011.

258. M. Tyagi, K. Hashimoto, B. A. Shoemaker, S. Wuchty, and A. R. Panchenko. Large-scale mapping of human protein interactome using structural complexes. *EMBO Rep*, 13(3):266–271, 2012.

259. P. Li. Exploring virtual environments in a decentralized lab. *ACM SIGITE Newsletter*, 6(1):4–10, 2009.

260. A. Tarasov, A. J. Vilella, E. Cuppen, I. J. Nijman, and P. Prins. Sambamba: fast processing of NGS alignment formats. *Bioinformatics*, 31(12):2032–2034, 2015.

261. Artem Tarasov and Pjotr Prins. Sambamba v0.5.0. doi: 10.5281/zenodo.13200, Dec 2014.

262. Y. Dong, M. Xie, et al. Sequencing and automated whole-genome optical mapping of the genome of a domestic goat (Capra hircus). *Nat Biotechnol*, 31(2):135–141, 2013.

263. M. Yandell and D. Ence. A beginner's guide to eukaryotic genome annotation. *Nat Rev Genet*, 13(5):329–342, 2012.

264. T. J. Aitman, C. Boone, G. A. Churchill, M. O. Hengartner, T. F. Mackay, and D. L. Stemple. The future of model organisms in human disease research. *Nat Rev Genet*, 12(8):575–582, 2011.

265. G. A. Churchill, D. C. Airey, et al. The Collaborative Cross, a community resource for the genetic analysis of complex traits. *Nat Genet*, 36(11):1133–1137, 2004.

266. K. W. Broman. R/qtlcharts: Interactive Graphics for Quantitative Trait Locus Mapping. *Genetics*, 2014.

267. Wikipedia. Big data — wikipedia, the free encyclopedia. http://en.wikipedia.org/w/index.php?title=Big_data&oldid=582004355, 2013.

268. D. I. Boomsma, C. Wijmenga, et al. The Genome of the Netherlands: design, and project goals. *Eur J Hum Genet*, 2013.

269. G. R. Abecasis, A. Auton, L. D. Brooks, M. A. DePristo, R. M. Durbin, R. E. Handsaker, H. M. Kang, G. T. Marth, and G. A. McVean. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65, 2012.

270. K. Tretyakov, S. Laur, G. Smant, J. Vilo, and P. Prins. Fast probabilistic file fingerprinting for big data. *BMC Genomics*, 14 Suppl 2:S8, 2013.

271. A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, and M. A. DePristo. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*, 20(9):1297–1303, 2010.

272. S. Ossowski, K. Schneeberger, R. M. Clark, C. Lanz, N. Warthmann, and D. Weigel. Sequencing of natural strains of Arabidopsis thaliana with short reads. *Genome Res*, 18(12):2024–2033, 2008.

273. M. Harakalova, I. J. Nijman, J. Medic, M. Mokry, I. Renkens, J. D. Blankensteijn, W. Kloosterman, A. F. Baas, and E. Cuppen. Genomic DNA pooling strategy for next-generation sequencing-based rare variant discovery in abdominal aortic aneurysm regions of interest-challenges and limitations. *J Cardiovasc Transl Res*, 4(3):271–280, 2011.

274. E. R. Mardis. The $1,000 genome, the$100,000 analysis? *Genome Med*, 2(11):84, 2010.

275. Erik Garrison and Gabor Marth. Haplotype-based variant detection from short-read sequencing, 2012.

276. J. R. Wang, F. P. de Villena, H. A. Lawson, J. M. Cheverud, G. A. Churchill, and L. McMillan. Imputation of single-nucleotide polymorphisms in inbred mice using local phylogeny. *Genetics*, 190(2):449–458, 2012.

277. P. Zhang, X. Zhan, N. A. Rosenberg, and S. Zollner. Genotype imputation reference panel selection using maximal phylogenetic diversity. *Genetics*, 195(2):319–330, 2013.

278. S. C. Munger, N. Raghupathy, K. Choi, A. K. Simons, D. M. Gatti, D. A. Hinerfeld, K. L. Svenson, M. P. Keller, A. D. Attie, M. A. Hibbs, J. H. Graber, E. J. Chesler, and G. A. Churchill. RNA-Seq alignment to individualized genomes improves transcript abundance estimates in multiparent populations. *Genetics*, 198(1):59–73, 2014.

279. M. Pop. Genome assembly reborn: recent computational challenges. *Brief Bioinform*, 10(4):354–366, 2009.

280. P. A. Pevzner, H. Tang, and M. S. Waterman. An Eulerian path approach to DNA fragment assembly. *Proc Natl Acad Sci U S A*, 98(17):9748–9753, 2001.

281. M. Howison, F. Zapata, and C. W. Dunn. Toward a statistically explicit understanding of de novo sequence assembly. *Bioinformatics*, 29(23):2959–2963, 2013.

282. L. E. Mose, M. D. Wilkerson, D. N. Hayes, C. M. Perou, and J. S. Parker. ABRA: improved coding indel detection via assembly-based realignment. *Bioinformatics*, 30 (19):2813–2815, 2014.

283. A. D. Prjibelski, I. Vasilinetc, A. Bankevich, A. Gurevich, T. Krivosheeva, S. Nurk, S. Pham, A. Korobeynikov, A. Lapidus, and P. A. Pevzner. ExSPAnder: a universal repeat resolver for DNA fragment assembly. *Bioinformatics*, 30(12):i293–i301, 2014.

284. Tanya Barrett, Dennis B Troup, Stephen E Wilhite, Pierre Ledoux, Carlos Evangelista, Irene F Kim, Maxim Tomashevsky, Kimberly A Marshall, Katherine H Phillippy, Patti M Sherman, Rolf N Muertter, Michelle Holko, Oluwabukunmi Ayanbule, Andrey

Yefanov, and Alexandra Soboleva. Ncbi geo: archive for functional genomics data sets–10 years on. *Nucleic Acids Res*, 39(Database issue):D1005–D1010, Jan 2011.

285. S. Okuda, T. Yamada, M. Hamajima, M. Itoh, T. Katayama, P. Bork, S. Goto, and M. Kanehisa. Kegg atlas mapping for global analysis of metabolic pathways. *Nucleic Acids Res*, 36(Web Server issue):W423–W426, 2008.

286. F. Jungo, L. Bougueleret, I. Xenarios, and S. Poux. The UniProtKB/Swiss-Prot Tox-Prot program: A central hub of integrated venom protein data. *Toxicon*, 60(4):551–557, 2012.

287. P. Lamesch, T. Z. Berardini, D. Li, D. Swarbreck, C. Wilks, R. Sasidharan, R. Muller, K. Dreher, D. L. Alexander, M. Garcia-Hernandez, A. S. Karthikeyan, C. H. Lee, W. D. Nelson, L. Ploetz, S. Singh, A. Wensel, and E. Huala. The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res*, 40 (Database issue):D1202–D1210, 2012.

288. D. Weigel and R. Mott. The 1001 genomes project for Arabidopsis thaliana. *Genome Biol*, 10(5):107, 2009.

289. T. Berners-Lee and J. Hendler. Publishing on the semantic web. *Nature*, 410(6832): 1023–1024, 2001.

290. M. G. Reese, B. Moore, C. Batchelor, F. Salas, F. Cunningham, G. T. Marth, L. Stein, P. Flicek, M. Yandell, and K. Eilbeck. A standard variation file format for human genome sequences. *Genome Biol*, 11(8):R88, 2010.

291. P. Danecek, A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, G. Lunter, G. T. Marth, S. T. Sherry, G. McVean, and R. Durbin. The variant call format and VCFtools. *Bioinformatics*, 27(15):2156–2158, 2011.

292. Biohackathon. The 3rd DBCLS BioHackathon: improving life science data integration with Semantic Web technologies. *J Biomed Semantics*, 4(1):6, 2013.

293. A. Anguita, M. Garcia-Remesal, D. de la Iglesia, and V. Maojo. NCBI2RDF: enabling full RDF-based access to NCBI databases. *Biomed Res Int*, 2013:983805, 2013.

294. S. Jupp, J. Malone, J. Bolleman, M. Brandizi, M. Davies, L. Garcia, A. Gaulton, S. Gehant, C. Laibe, N. Redaschi, S. M. Wimalaratne, M. Martin, N. Le Novere, H. Parkinson, E. Birney, and A. M. Jenkinson. The EBI RDF platform: linked open data for the life sciences. *Bioinformatics*, 30(9):1338–1339, 2014.

295. F. Belleau, M. A. Nolin, N. Tourigny, P. Rigault, and J. Morissette. Bio2RDF: towards a mashup to build bioinformatics knowledge systems. *J Biomed Inform*, 41(5):706–716, 2008.

List of publications

1. **P. Prins**†, J. de Ligt, A. Tarasov, R. C. Jansen, E. Cuppen, and P. E. Bourne. Toward effective software solutions for big biology. *Nat Biotechnol*, 33(7):686–687, 2015.
2. A. Tarasov, A. J. Vilella, E. Cuppen, I. J. Nijman, and **P. Prins**†. Sambamba: fast processing of NGS alignment formats. *Bioinformatics*, 31(12):2032–2034, 2015.
3. T. Katayama, M. D. Wilkinson, et al. BioHackathon series in 2011 and 2012: penetration of ontology and linked data in life science domains. *J Biomed Semantics*, 5(1):5, 2014
4. S. Möller, E. Afgan, et al. Community-driven development for computational biology at Sprints, Hackathons and Codefests. *BMC Bioinformatics*, 15 Suppl 14:S7, 2014
5. T. Katayama, M. D. Wilkinson, et al. The 3rd DBCLS BioHackathon: improving life science data integration with Semantic Web technologies. *J Biomed Semantics*, 4(1):6, 2013
6. K. Tretyakov, S. Laur, G. Smant, J. Vilo, and **P. Prins**†. Fast probabilistic file fingerprinting for big data. *BMC Genomics*, 14 Suppl 2:S8, 2013
7. R. J. Bonnal, J. Aerts, G. Githinji, N. Goto, D. MacLean, C. A. Miller, H. Mishima, M. Pagani, R. Ramirez-Gonzalez, G. Smant, F. Strozzi, R. Syme, R. Vos, T. J. Wennblom, B. J. Woodcroft, T. Katayama†, and **P. Prins**†. Biogem: an effective tool-based approach for scaling up open source software development in bioinformatics. *Bioinformatics*, 28 (7):1035–1037, 2012
8. D. Arends, K. J. van der Velde, **P. Prins**, K. W. Broman, S. Möller, R. C. Jansen, and M. A. Swertz. xQTL workbench: a scalable web environment for multi-level QTL analysis. *Bioinformatics*, 28(7):1042–1044, 2012
9. **P. Prins**†, N. Goto, A. Yates, L. Gautier, S. Willis, C. Fields, and T. Katayama. Sharing programming resources between Bio* projects through remote procedure call and native call stack strategies. *Methods Mol Biol*, 856:513–527, 2012
10. **P. Prins**†, G. Smant, and R. C. Jansen. Genetical genomics for evolutionary studies. *Methods Mol Biol*, 856:469–485, 2012

†First or last (shared) authorships

Google Scholar citations of **P. Prins** (as of 5 June 2015) total 718, H-index 12

11. **P. Prins**†, D. Belhachemi, S. Möller, and G. Smant. Scalable computing for evolutionary genomics. *Methods Mol Biol*, 856:529–545, 2012
12. C. Durrant, M. A. Swertz, R. Alberts, D. Arends, S. Moller, R. Mott, **P. Prins**, K. J. van der Velde, R. C. Jansen, and K. Schughart. Bioinformatics tools and database resources for systems genetics analysis in mice–a short review and an evaluation of future needs. *Brief Bioinform*, 13(2):135–142, 2012.
13. O. Trelles†, **P. Prins**†, M. Snir, and R. C. Jansen. Big data, but are we ready? *Nat Rev Genet*, 12(3):224, 2011
14. P. C. McKeown†, S. Laouielle-Duprat†, **P. Prins**†, P. Wolff, M. W. Schmid, M. T. Donoghue, A. Fort, D. Duszynska, A. Comte, N. T. Lao, T. J. Wennblom, G. Smant, C. Kohler, U. Grossniklaus, and C. Spillane. Identification of imprinted genes subject to parent-of-origin specific expression in *Arabidopsis thaliana* seeds. *BMC Plant Biol*, 11: 113, 2011
15. E. Bakker, T. Borm, **P. Prins**, E. van der Vossen, G. Uenk, M. Arens, J. de Boer, H. van Eck, M. Muskens, J. Vossen, G. van der Linden, R. van Ham, R. Klein-Lankhorst, R. Visser, G. Smant, J. Bakker, and A. Goverse. A genome-wide genetic map of NB-LRR disease resistance loci in potato. *Theor Appl Genet*, 123(3):493–508, 2011
16. T. Katayama, K. Arakawa, et al. The DBCLS BioHackathon: standardization and interoperability for bioinformatics web services and workflows. The DBCLS BioHackathon Consortium*. *J Biomed Semantics*, 1(1):8, 2010.
17. N. Goto†, **P. Prins**†, M. Nakao, R. Bonnal, J. Aerts, and T. Katayama. BioRuby: bioinformatics software for the Ruby programming language. *Bioinformatics*, 26(20):2617–2619, 2010
18. D. Arends†, **P. Prins**†, R. C. Jansen, and K. W. Broman. R/qtl: high throughput Multiple QTL mapping. *Bioinformatics*, 26 (23), 2990-2992, 2010
19. S. Rehman, W. Postma, T. Tytgat, **P. Prins**, L. Qin, H. Overmars, J. Vossen, L. N. Spiridon, A. J. Petrescu, A. Goverse, J. Bakker, and G. Smant. A secreted SPRY domain-containing protein (SPRYSEC) from the plant-parasitic nematode *Globodera rostochiensis* interacts with a CC-NB-LRR protein from a susceptible tomato. *Mol Plant Microbe Interact*, 22(3):330–340, 2009
20. Y. Li, O. A. Alvarez, E. W. Gutteling, M. Tijsterman, J. Fu, J. A. Riksen, E. Hazendonk, **P. Prins**, R. H. Plasterk, R. C. Jansen, R. Breitling, and J. E. Kammenga. Mapping determinants of gene expression plasticity by genetical genomics in *C. elegans*. *PLoS Genet*, 2(12):e222, 2006
21. L. Qin, **P. Prins**, and J. Helder. Linking cDNA-AFLP-based gene expression patterns and ESTs. *Methods Mol Biol*, 317:123–138, 2006
22. L. Qin, **P. Prins**, J. T. Jones, H. Popeijus, G. Smant, J. Bakker, and J. Helder. GenEST, a powerful bidirectional link between cDNA sequence data and gene expression profiles generated by cDNA-AFLP. *Nucleic Acids Res*, 29(7):1616–1622, 2001

†First or last (shared) authorships

Education Statement of the Graduate School

Experimental Plant Sciences

Issued to:	**J.C.P. (Pjotr) Prins**
Date:	**5 October 2015**
Group:	**Laboratory of Nematology**
University:	**Wageningen University & Research Centre**

	1) Start-up phase	date
►	**First presentation of your project**	
	Open source methods for genetics and genomics	Apr 01, 2005
►	**Writing or rewriting a project proposal**	
►	**Writing a review or book chapter**	
	Genetical Genomics for Evolutionary Studies, Springer Series, Jan 2011	Dec 20, 2010
►	**Msc courses**	
►	**Laboratory use of isotopes**	
	Subtotal Start-up Phase	7.5 credits*

	2) Scientific Exposure	date
►	**EPS PhD student days**	
	EPS PhD Student Day, Utrecht University	Jun 01, 2010
	EPS PhD Student Day, University of Amsterdam	Nov 20, 2012
►	**EPS theme symposia**	
	EPS Theme 4 Symposium 'Genome Plasticity', Wageningen University	Dec 10, 2010
	EPS Theme 4 Symposium 'Genome Biology', Radboud University Nijmegen	Dec 07, 2012
►	**NWO Lunteren days and other National Platforms**	
	NBIC GALAXY Community conference	May 25-26, 2011
►	**Seminars (series), workshops and symposia**	
	SYSGENET Groningen Workshop	Sep 07-09, 2010
	Benelux Bioinformatics Conference	Apr 14-15, 2005
►	**Seminar plus**	
►	**International symposia and congresses**	
	CTC Groningen	Jun 26-29, 2005
	MGED8 Norway	Sep 11-13, 2005
	MPMI Conference, Cancun	Dec 14-18, 2005
	Biohackathon Japan (grant from DBCLS)	Feb 10-16, 2008
	ISMB/BOSC	Jun 27-Jul 02, 2009
	Biohackathon Japan (grant from DBCLS)	Feb 08-12, 2010
	COST Braunschweig SYSGENET	Apr 07-09, 2010
►	**Presentations**	
	Cfruby / Cfengine Wageningen UR (Talk)	Dec 14, 2004
	CTC Groningen (shared w. Olga Alvaraz) (Talk)	Jun 26, 2005
	Ruby on Rails, NLUUG (Talk)	Nov 17, 2005
	GBIC Microrrays (Talk)	Jun 12, 2006
	BIOEXPLOIT (Talk)	Mar 31, 2007
	Virtualization Wageningen UR (Talk)	Jun 19, 2007
	Biohackathon (Talk)	Feb 15, 2008
	University of Malaga (Talk)	May 15, 2008
	BOSC (Talk)	Jul 19, 2008
	BOSC (Talk)	Jun 27, 2009
	Indian Institute of Science (Talk)	Feb 15, 2010
	IBAB (Talk)	Feb 16, 2010
	University of Madras (Talk)	Feb 18, 2010
	Univerity of Kerala (Talk)	Mar 04, 2010
	Braunschweig 2010 (Talk)	Apr 08, 2010
	EPS Theme 4 Symposium 2010 on MQM (Talk)	Dec 10, 2010
►	**IAB interview**	
	Meeting with a member of the International Advisory Board of EPS	Nov 2012
►	**Excursions**	
	Subtotal Scientific Exposure	29.3 credits*

	3) In-Depth Studies	date
►	**EPS courses or other PhD courses**	
	BIT1 (accelerated 1 week Ph.D. course by Peter Schaap)	Sep 2005
	Jackson Lab 2009 System Genetics QTL	Oct 19-25, 2009
►	**Journal club**	
	Member of a literature discussion group at Nematology	2005-2010
►	**Individual research training**	
	Groningen Bioinformatics (ongoing days at GBIC lab w. Ritsert Jansen, Danny Arends)	2008-2012
	Subtotal In-Depth Studies	8.5 credits*

	4) Personal development	date
►	**Skill training courses**	
	Writing course Mike Grossman (EPS)	Jun 29- Jul 02, 2010
	Google Summer of Code Programme, which develops many skills, incl. writing, communicating,	Mar-Aug 2009
►	**Organisation of PhD students day, course or conference**	
	WUR minisymposium	Jun 19, 2007
	EU-codefest, Lodi, 2012	Jul 19-20, 2012
►	**Membership of Board, Committee or PhD council**	
	WURLUG Board - president of WUR Linux User Group http://wurlug.org/	Oct 2003 - present
	Subtotal Personal Development	6.7 credits*

TOTAL NUMBER OF CREDIT POINTS*	**52.0**

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

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