Proteomic and genomic data mining with applications in plant science

Hong Luo

Propositions

1. It is far easier to develop a biological database than it is to maintain it.
   (this thesis)

2. The sorghum *Dry* gene is a domestication gene.
   (this thesis)

3. AI cannot replace human curation of scientific knowledgebases.

4. The algorithms developed by social media companies should be protected.

5. Elon Musk is a scientist.

6. Teleworking cannot be the long-term normal after the Covid-19 pandemic.

Propositions belonging to the thesis, entitled

Proteomic and genomic data mining with applications in plant science

Hong Luo
Wageningen, 21 March 2023

Proteomic and genomic data mining with applications in plant science

Hong Luo

Thesis committee

Promotor
Prof. Dr D. de Ridder
Professor of Bioinformatics
Wageningen University & Research

Co-promoter
Dr H. Nijveen
Researcher, Bioinformatics Group
Wageningen University & Research

Other members
Prof. Dr A.B. Bonnema, Wageningen University & Research
Dr W. Ligterink, Crop Innovation Unit, KeyGene, Wageningen
Dr S. Warris, Wageningen University & Research
Dr L. Fokkens, Wageningen University & Research

Proteomic and genomic data mining with applications in plant science

Hong Luo

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 Tuesday 21 March 2023
at 11 a.m. in the Omnia Auditorium.

Hong Luo
Proteomic and genomic data mining with applications in plant science,
138 pages.

PhD thesis, Wageningen University, Wageningen, the Netherlands (2023)
With references, with summary in English

ISBN 978-94-6447-472-5
DOI https://doi.org/10.18174/579767

Contents

CHAPTER 1

1

Introduction

Bioinformatics is an interdisciplinary field of life science. It encompasses the integration of mathematics, statistics along with computer science to interpret and understand biological data. The scope of bioinformatics research and analysis ranges from the analysis of protein sequence and genomics data, to the development of workflow and data management systems, from genome annotation and protein function prediction, to the mining of biological literature and images. Recently, the astronomical accumulation of omics data, boosted by advanced sequencing and molecular technology, has become the main driving force for the development of bioinformatics and the training of bioinformaticians. The main task is to develop reliable and feasible procedures through mathematical/statistical modeling, facilitating the interpretation of biological phenomena.

Plants are of immense significance for life on earth. As the primary staple food source, crop plants are crucial for humans to survive and thrive. The important foci of crop research are domesticating varieties superior in quality and quantity and enhancing their adaptation capacity to biotic and abiotic/environmental stress factors. Nowadays， crop plants are investigated at various levels, including genomics, proteomics, molecular genetics and breeding. However, knowledge is still limited in areas like genetic architecture, population structure and domestication model. This shows the need for fundamental and applied bioinformatics research, developing and applying novel computational methods and tools for the analysis and integration of crop omics data. The cross-talk of high-throughput data and bioinformatics calls for advanced approaches in functional genomics and population genetics, to solve the scientific questions underlying modern agriculture. Moreover, it is necessary to develop relevant databases, software, tools, and web resources to ease access for the research community.

This PhD thesis focuses on applying bioinformatics methodology to two topics: the discovery of repeats in protein sequences and the analysis of crop plant omics data. In this introductory chapter, I first give a brief overview of the history of bioinformatics, followed by a more detailed discussion on sequence analysis methodologies, pattern recognition algorithms and database management principles for biological data. Then I review the methodology of applying next-generation sequencing (NGS) technology and bioinformatics approaches to crop omics research, combined with the formulation of scientific questions in crop functional genomics and population genetics. Finally, I present the outline and contribution of this thesis.

1.1 Bioinformatics

During recent decades, bioinformatics has become recognized as a pervasive technology, made possible by the marriage of biology and computer science. It is widely accepted that the term 'bioinformatics' was first used in the early 1970s by Paulien Hogeweg and Ben Hesper. They defined it as 'the study of informatic processes in biotic systems' (1). From this perspective, bioinformatics deals with computational and mathematical approaches for understanding how biological systems process information, following the central dogma of molecular biology as observed in the sequences of the genome, transcriptome and proteome, as well as the cellular phenotype. More recently, bioinformatics has become a branch of science aiming to extract knowledge from biological data, which includes collecting, manipulating and modelling data for analysis, visualization or prediction through the development of algorithms, software and pipelines.

With the development of bioinformatics, the profession of bioinformatician emerged. Margaret Dayhoff, who developed a number of algorithms and tools for protein sequence analysis (2,3), is commonly believed to have been the first bioinformatician (4). Currently, a bioinformatician is regarded as a person with core competencies including using current techniques, skills, and tools necessary to process, analyze and interpret biological measurement data, and applying statistical methods in the contexts of molecular biology, genomics, medicine, and/or population genetics (5). Besides these competencies, both academic and industrial communities expect more professional capacities from bioinformaticians, such as in-depth knowledge in one or more areas of biology, proficiency in programming, script development and database construction, and a thorough understanding of algorithms for data integration, text mining and machine learning applied to biological big data. In short, a full-fledged bioinformatician should be a "missing link" to bridge the multiple disciplines that need to come together to solve pressing problems in biology.

Although the information flow of molecular biology starts from the DNA sequence, bioinformatics started with protein sequence analysis in the early 1960s, as the first sequencing method was for proteins (6). In this period, pioneering bioinformaticians developed tools and algorithms for determining protein primary structure (3) and making pairwise protein sequence alignments (7). With the development and improvement of the Sanger chain termination method (8) for DNA sequencing, the paradigm of bioinformatics gradually shifted to the integrated analysis of protein and DNA sequences, which commonly includes comparisons between sequences from different organisms, inferences of the phylogenetic relationships via orthologous sequences, and discovery of patterns in sequences. At the same time, the demands for comprehensive sequence analysis tools boosted the development of classical bioinformatics software and packages, such as Staden (9), GCG (10), and MUMmer (11), some of which are still in widespread use today. Moreover, centralized bioinformatics databases such as the European Nucleotide Archive (ENA), GenBank and DNA Data Bank of Japan (DDBJ) were established and integrated to standardize data formats defining minimal information for reporting nucleotide sequences and facilitating data sharing (12).

With the initiation of the Human Genome Project (HGP) (13) and the growth of the Internet, the 1990s saw the beginning of a new era of bioinformatics. It was not only the dawn of genomics, spurred by multiple genome sequencing projects, but also of proteomics, promoted by advances in protein sequence pattern recognition algorithms and computational 3D structure prediction methods (14). At the same time, the World Wide Web (WWW) allowed the release of online bioinformatics resources and packages, such as UniProtKB/Swiss-Prot (15), PubMed (16), and wEMBOSS (17,18), most of which, following the philosophy of free software, were released in a free and open-source manner. With this, building web services with easy-to-use graphical interfaces for bioinformatics tools has steadily become routine for bioinformaticians.

Since the early 2000s, the popularization of NGS technologies (19), which can sequence millions of fragments of DNA molecules in a single machine run, increased the need for bioinformatics to offer more sophisticated algorithms, exploiting the higher available computational hardware capacity. It also led to a larger number of professional bioinformaticians that graduated from newly started education programmes. At the same time, the impressive decrease of sequencing costs accelerated the generation of biological “big data”, currently beyond the exabyte (20) level. The statistics of GenBank reveal a staggering increase from the first release in December 1982 of 606 sequences to the latest release in February 2022 of more than 236 million sequences (https://www.ncbi.nlm.nih.gov/genbank/statistics/).

Recently, third-generation sequencing (TGS) technologies (21) emerged, which allow real-time single molecule sequencing by omitting the conventional PCR amplification step indispensable in the NGS protocol. One of the most significant advantages of TGS is the ultra-long reads they generate (22), which are crucial to gain high-resolution contiguous reference genomes by spanning repeats that hamper genome assembly based on short reads (23). Although the early TGS protocols had a high sequencing error rate, optimized strategies such as the circular consensus sequencing (CCS) on the PacBio (24) platform and new base-calling algorithms (25) on the nanopore platform promise better sequencing accuracy comparable to NGS protocols. Therefore, bioinformatics work has shifted to long-read mapping and assembly algorithms (26).

Today, bioinformatics faces multiple chances together with challenges, such as interpreting big biological data with advanced machine learning and deep learning methods, ensuring the reproducibility of results by coordinating collaborations and formulating uniform standards among bioinformatics communities, and proper integration of computer science and biology courses and training into academic bioinformatics curricula. There is a growing consensus that biology and bioinformatics are so intertwined that eventually, it may become unnecessary to distinguish one from the other. Integrated systems biology could make the next leap of modelling the living cell, organs or whole organisms. Such models should include complete genomes, transcriptomes, metabolomes, phenomes and environments and take all interactions between these into account simultaneously (27,28).

1.2 Biological sequence analysis and management

1.2.1 Biological sequence alignment

Biological sequence analysis is a fundamental task in bioinformatics to make sense of the vast accumulated biological data. It is a challenging task as biological sequences encode for most of the complexity of molecular biology, although they can be represented as simple strings of bases or amino acids. These strings do not convey the immense richness of biological signals, as they have been shaped by multiple evolutionary forces, such as natural selection and genetic drift.

One of the most fundamental tasks in molecular biology is to establish the relatedness of genes or proteins. Similarity of two biological molecules at the sequence level suggests that they are homologous, i.e. share a common ancestor. To evaluate the similarity of biological sequence pairs, identifying a plausible alignment between them is intuitive and effective. Dynamic programming is the methodology for finding an optimal alignment given a specific score scheme describing the probability of an amino acid substitution, such as point accepted mutation (PAM) matrices (2) and blocks substitution matrices (BLOSUM) (29). The introduction of probabilistic matrices considers the features that constrain primary sequence evolution to grant the biologically most likely alignment the highest score. For sequences with high similarity and of roughly equal size, global alignment is more practical. For finding regions containing similar motifs in divergent sequences, local alignment is more appropriate. Dynamic programming can be applied to produce global alignments via the Needleman-Wunsch algorithm (30) and local alignments via the Smith-Waterman algorithm (31).

The classical methodologies developed for pairwise sequence alignment can be extended to multiple sequences alignment (MSA). In MSA, sequences are aligned by bringing similar characters into the same column of the alignment, which could reflect the evolutionary history of the sequences. Progressive alignment approaches use dynamic programming to build an MSA, starting with the two most similar sequences and then progressively adding less similar sequences to the initial alignment (32). This approach has the inherent limitation that it is sensitive to initial alignment error. Iterative approaches were therefore developed to enhance the alignment quality by obtaining information from repeated alignment procedures (33). The applications of MSA for biological sequence analysis are more extensive than pairwise sequence alignment. For example, phylogenetic prediction algorithms often begin with producing an optimal MSA as the first step toward making a phylogenetic tree.

1.2.2 Pattern recognition in biological sequences

A pattern can be defined "as the opposite of a chaos" (34). Patterns in biological sequences could be genes, sequence motifs or protein domains with functional implications. Pattern recognition in bioinformatics is concerned with developing and applying systems that learn to solve a given problem using a set of biological data, each represented by some features. The development of statistical pattern recognition algorithms is mainly concerned with theory and methods including clustering, dimensionality reduction and classification (35). In DNA and protein sequence analysis, pattern matching and detection algorithms have been widely used to explore and exploit specified and novel patterns. Due to the fivefold higher variety of sequence characters in proteins, it is much easier to detect patterns of sequence similarity between protein sequences than between DNA sequences (36). Once biologists started to read protein sequences and genomes, they learned that large parts of gene and genome sequences consist of periodic patterns (37). This led to the question what possible biological role they might have. Repeat patterns in DNA sequences have been widely investigated, leading to a plethora of detection algorithms and the discovery of roles for DNA repeat elements in diverse biological processes (38,39). Nevertheless, repeat patterns in protein sequences have different implications. Repeated amino acids could participate in the formation of secondary and three-dimensional structures of proteins to create and alter protein function.

In principle, identifying repeats from protein sequences is achieved with pattern recognition (35). It should start from the explicit definition and classification of various amino acid repeats by considering repeat unit features regarding their sequence pattern and potential biological significance. Based on the repeated amino acid unit's complexity, similarity, and distance within a protein, a protein with a repeat embedded amino acid sequence could be classified as complex or simple, tandem or sequentially interspersed, perfect or imperfect repeat-containing protein (RCP). Various sequence pattern recognition strategies can detect different repeat patterns, such as string suffix trees (40), complexity measurements (41), discrete Fourier and stationary wavelet transforms (42), hidden Markov model (HMM) based self-comparisons (43), and trained neural networks (44). Nevertheless, multiple repeat patterns are commonly intertwined within one RCP so that no single algorithm can uncover all different cryptic repeat patterns. A rational strategy is to identify repeat fragments by multiple algorithms on the same RCP, then merge or distinguish these based on their positions and repeat unit patterns.

1.2.3 Searching and managing biological databases

Biological databases play a central role in bioinformatics. They offer scientists access to a wide variety of biologically relevant data, including genomic sequences, population variations, information on gene structure and protein family classification of an increasingly broad range of organisms. They provide fertile ground for biologists to better design and interpret their experiments in the laboratory, fulfilling the promise of bioinformatics of advancing and accelerating biological discovery (45).

Database search is remarkably useful for finding the function of genes or proteins whose sequences have been determined in the laboratory. In addition, the biological function of particular sequences in model organisms could help predict the function of similar sequences in other organisms. Thus, an important application of database searching is to identify similar sequences. Such searches have become commonplace and are greatly facilitated by programs such as basic local alignment search tool (BLAST) (46). BLAST was designed to accelerate database search using a heuristic method. In contrast to the original version of the Smith-Waterman algorithm, BLAST searches are not guaranteed to find the optimal alignments. They limit the search space by scanning a database for possible short sequence matches before performing more rigorous alignments, which is essential to save the search time. BLAST was the first program to apply rigorous statistics to obtain scores for local sequence alignments. The NCBI BLAST server (http://www.ncbi.nlm.nih.gov/BLAST) is probably the most widely used sequence analysis facility in the world and provides similarity searching to all currently available sequences.

The centralized and primary databases generally store data generated directly from sequencing and experimental platforms, annotated by automated pipelines. Secondary databases on the other hand contain information that was mined from primary databases. The users of secondary databases commonly focus on specific species or research topics. They have more refined requirements for database content, interface and analysis tools. Constructing such databases should follow the principles of friendly-to-use, fast-to-query, intuitive-to-analysis and informative-to-learn (47). Secondary databases also take advantage of data mining and database management methods to integrate multiple cross-references extracted from other data resources. Moreover, regular maintenance and timely updates of these biological databases are essential to the scientific community. For example, all databases published in the NAR Database Issue (48) are expected to be maintained under the same URL for at least five years after the publication date. Graduation or retirement of the database developers is not a valid reason for the termination of the database.

1.3 Studying integrative methodology on plant omics

1.3.1 Omics technologies and bioinformatics methodologies

Omics technologies are used to measure the entire complement of a given level of biological molecules and information. Primary applications are the detection of genes (genomics), mRNA (transcriptomics), proteins (proteomics), metabolites (metabolomics), and phenotypes (phenomics) in a specific biologic sample in a non-targeted and non-biased manner. It encompasses a variety of new high-throughput technologies to analyze very large numbers of biological data in a combination of procedures, by which to enable a system-level understanding of correlations and dependencies between molecular components (49-51).

It is believed that the initiation of high-throughput omics was triggered by the revolution in NGS technologies. In the early 2000s, strategies of massively parallel sequencing of clonally amplified DNA molecules that are spatially separated in a flow cell were developed (52). Although different academic and industrial platforms were developed in the early stages, the sequencing by synthesis system of Illumina/Solexa (http://www.illumina.com) has become the dominant standard, as it has tailored sample preparation and data generation protocols which find the balance between performance and cost. NGS methodologies were next successfully applied to map and quantify transcriptomes by altering the library preparation protocols for mRNAs, non-coding RNAs and small RNAs, an approach commonly called RNA-seq. RNA-seq came with all the advantages of NGS including high-throughput, high accuracy for quantifying expression levels, and high levels of reproducibility (53). Next to DNA and RNA sequencing, a number of high-throughput methodologies aimed at other aspects of omics, such as mass spectrometry (54) for proteomics, ChIP-seq (55) for interactomics, and DNA methylation sequencing (56) for epigenomics.

Essentially, bioinformatics methodologies for omics constitute a crossover between bottom-up hypothesis-driven and top-down data-driven approaches, which is indispensable to integrate the acquisition, analysis and management of multiple omics data with genome-scale statistical and mathematical modelling and simulation (57). During recent decades, bioinformatics developed a score of methods for sequence analysis, in terms of sequence assembly, genome annotation, comparative genomics, genetics and population genomics, computational evolutionary biology; for expression analysis, in terms of gene expression, protein expression, metabolite profiling; for structural bioinformatics, in terms of genome 3D chromatin modelling, RNA secondary structure prediction, protein structure prediction, homology modelling; for network and system biology, in terms of biological network analysis, gene co-expression analysis, molecular interaction networks; and for data management, in terms of database and web service development, data curation, data visualization, workflow management, and so on.

1.3.2 Perspective of crop omics

A global food crisis can soon develop, given the rapid growth of the world population and the lagging pace of yield increase for major crops, including rice, maize, wheat, and sorghum. Domestication, design, and development of genetically improved, stress-resilient, and environmental adapted crops have become the research priority. The crop omics perspective necessitates the convergence of low-cost genome sequencing with improved computational power and high-throughput molecular phenotyping technologies to accelerate the identification of genes and/or loci underlying important agronomic traits relevant to food production and quality (58).

As the basis of crop omics, the first essential step is assembly and decoding of the reference genomes. This was a challenge for NGS based methodologies as the size and dynamic nature of the plant genome are complicated and diversified (59,60). Plants tend to have more multigene families and a higher frequency of polyploidy than other forms of life (61). This commonly resulted in an expanded and repeat-rich genome that cannot be fully determined by short reads. Paralogy is a substantial issue in plant genomics, so a series of compensation methods were required to obtain the high quality reference genome (62). The development of long read sequencing technologies (21) significantly improved the situation, together with advanced chromosome physical mapping protocols such as BioNano optical mapping (63) and chromosome conformation capture (Hi-C) (64). This yielded higher assembly quality for complex plant genomes with lower computational prices (65) and prompted scientific communities to construct many complete plant genomes (66,67).

Based on fully assembled and well-annotated crop reference genomes, crop improvement will depend on comparisons of individual plant genomes. Some of the best opportunities may lie in using combinations of new genetic mapping strategies and evolutionary analyses to direct and optimize the discovery and use of genetic variation. Besides, conventional molecular population genetics using a limited numbers of DNA-based markers has evolved to population genomics, adapting to the increased availability of genome-wide DNA variation data of many individuals in natural crop populations (68). However, the high levels of nucleotide diversity in crop genomes poses challenges. For example, the maize and human genomes are similar in size, but an average pair of maize individuals differ at ten times more sites than any two humans do (69). This is due to the higher incidence of interspecific and introgressive hybridization between crop subpopulations, regarded as an important mechanism for their adaptive evolution (70). Next to SNPs and short insertions/deletions (InDels), larger genome structure variations induced by transposon translocation or polyploidization events could impact crop genotypes associated with crucial agronomic traits for their domestication and improvement.

Currently, hundreds of plant reference genomes have been assembled and annotated along with the generation of large amount of omics data. Accordingly, several integrated plant hub databases were developed such as Phytozome (71), Ensembl Plants (72) and Gramene (73). These are crucial for researchers from broad plant science communities to access the data and make comparative genomics analyses available. For some widely used plants such as major crops, the need for specific secondary databases is growing. Example databases include MaizeGDB (74) for maize, WheatGenome.info for wheat (75), and RAP-DB (76) for rice. These contain general genomic datasets such as genome sequence, gene models, functional annotation, and polymorphic loci of these crops. They also integrate information on the variome and phenome such as breeding status, population structure and kinship, linkage disequilibrium (LD) mapping, and phenotype data for multiple agronomic traits. The experiences of constructing these databases could be applied to studying other crops, such as sorghum, the fifth cereal crop in the world.

1.4 Applying multiple bioinformatics approaches to sorghum populations and functional genomics

1.4.1 Background of sorghum genomics

Sorghum is a grass species which diverged from rice ~50 million years and from maize ~12 million years ago (77,78). It uses the C4 carbon fixation photosynthetic process. Its relatively small genome (diploid, ten chromosomes, ~730M, ~34,000 genes) (79,80) makes sorghum an appealing model organism for C4 grass species with a more complex genome, such as wheat and sugarcane (77). As sorghum improvement has relied on public research more heavily than that of genetically modified crops such as corn or soybean, many genetic resources of sorghums serve a dual purpose for academic and commercial pursuits (81). Since the first release of the sorghum reference genome (*BTx623*) sequenced by the Sanger method in 2009 (79) and the improved assembly and annotation later by supplementary NGS protocols (80), two genomes of sorghum varieties have recently been *de novo* assembled using TGS technologies (82,83). Although comparisons with maize genomes showed a higher degree of collinearity and structural conservation in sorghum genome (84), it has been argued that the expected genomic diversity of sorghum populations should be higher than observed (81). The underestimation is due to the limited number of accessions being sequenced, that do not represent the full spectrum of diversity in the sorghum germplasm. A pan-genome panel represented as the nonredundant collection of genes and/or DNA sequences (85) in sorghum is required. This will allow to better study genetic mechanisms of variation, which should incorporate more divergent genotypes and wild relatives along with closely related but phenotypically divergent germplasms. Nevertheless, the cost of constructing a pan-genome panel consisting of dozens of *de novo* assembled sorghum varieties is still high. Using whole-genome resequencing, variome data has been collected in population genomics studies involving thousands of genotypes representing diversified gene pools of sorghum.

1.4.2 Domestication and breeding history of sorghum population

Recent decades have a strong interest in the origins of crop domestication (70,86), which are vital to understand crop evolutionary mechanisms and enhance crop agricultural traits. The current abundance in population genomics data generated by advanced omics technology provides unprecedented opportunities to study crop domestication and breeding. Sorghum is the fifth major cereal crop originating from Africa, which was initially domesticated as early as 4,000-6,000 years ago (87,88). Differing from rice, maize and wheat that are primarily cultivated as food sources, sorghum has multiple end uses as food, feed, fodder, fuel, fibre, broom and beverage. Four major breeding subpopulations with diversified agronomic traits are in cultivation worldwide, including grain sorghum with high seed yield and quality, sweet sorghum with juicy and sugary stem, forage sorghum with good tillers and biomass production, and broom sorghum with long fibres of panicle (89,90). Besides, sorghum was subjected to multiple domestication processes, in which selections of wild relatives happened in different regions at various time points (86,87).

Sweet sorghum is a unique cultivated subgroup, distinguished by its juicy and sugar-rich stem at maturity. The juicy stem allows transporting and storing mass and minimizes postharvest loss of fermentable sugars. Furthermore, sorghum has higher energy utilization efficiency and drought tolerance capacity than maize and sugarcane, which is fundamental to exploiting sweet sorghum's potential as a second generation biofuel crop (91). Promoted by a large scale program for the development of sweet sorghum cultivars that started in the 1970s (92), studies on essential genes altering stem juicy and sugar content of sweet sorghum were performed. The *Dry* locus, which controls the pithy/juicy stem trait, was discovered over a century ago (93). Moreover, previous studies identified a major quantitative trait locus (QTL) for midrib colour, sugar yield, juice volume, and moisture at ~51.8 Mb on Chromosome 6 (94). Nevertheless, few studies investigated the origin and evolution of *Dry* locus molecular basis, and its impact on the domestication and breeding of sweet sorghum.

1.4.3 Studying sorghum population and functional genomics via bioinformatics approaches

Recent advances in omics have substantially increased research opportunities for sorghum. As cost is no longer an obstacle, acquiring high-density markers of whole-genome variations across hundreds of varieties is now possible. Analyzing such data, classical genetic and population biology theory needs to be supported by novel bioinformatics and statistical approaches to help better explain the biological processes.

In sorghum molecular breeding research, it is essential to identify candidate genes corresponding to specific functions affecting crucial agronomic traits. Map-based cloning is the traditional approach, using segregation populations derived from the cross of two parental lines. However, as recombination within the biparental populations is typically limited, the mapped regions are commonly too large to quickly identify the underlying genes. Further fine mapping is formidable task that requires tedious work in the field and laboratory. To tackle the problem, a strategy is to gather a large number of natural sorghum varieties and obtain sufficient natural variation covering the candidate locus by whole-genome resequencing. A genome-wide association study (GWAS) can then be performed to fine map the candidate genes.

Genome-wide variations data provide a wealth of research material for sorghum population genomics to address new scientific questions. Population structure and phylogenetic relationships are fundamental aspects in sorghum population genomics. These result from the individual survival, dispersal, and reproduction histories, which reflect sorghum’s evolution and breeding background. The demographic history and geographic differentiation of sorghum population is important because it shapes patterns and levels of extant genetic diversity (95). Previous demographic analyses rested on assumptions based on archaeological evidence. Confirmation of these assumptions is necessary to infer the time of the possible domestication bottlenecks, which was crucial in reducing genetic diversity and eroding sorghum fitness. Besides, it is useful to infer the potential split and mixture events and conceivable migration events within and between sorghum populations using statistical models (96) on genome-wide allele frequency data. To learn how natural and artificial selection have shaped genes and their regulation network, thereby influence the diversity and fitness of the sorghum population, we can perform whole-genome screens with measures of population genetics/genomics parameters, such as nucleotide diversity, population divergence, and tests of selection via statistical models (68,97).

1.5 Contribution of this thesis

In this thesis, I present a number of contributions in managing and analysing large amounts of genome, proteome and variome data. In ***Chapter 2*** and ***Chapter 3***, I focus on proteomics, reviewing and integrating protein repeat detection algorithms and constructing a repository for collecting, cataloguing and managing protein repeats. ***Chapter 4*** reviews NGS technologies and their application in plant genomics and breeding. In ***Chapter 5***, I describe the construction of a repository for sorghum variome data and show how it can be used to study evolution and support breeding. In ***Chapter 6***, we demonstrate the importance of variomes in a study on an essential trait of sorghum. Finally, in ***Chapter 7***, I discuss the contribution and limitations of the thesis and present an outlook on future developments.

References

1. Hogeweg, P. (2011) The Roots of Bioinformatics in Theoretical Biology. *PLoS Computational Biology*, **7**, e1002021.
2. Dayhoff, M.O., Schwartz, R.M. and Orcutt, B.C. (1978) A Model of Evolutionary Change in Proteins. *Atlas of Protein Sequence and Structure*, **5**, 345-352.
3. Dayhoff, M.O. and Ledley, R.S. (1962), *Proceedings of the December 4-6, 1962, fall joint computer conference*. Association for Computing Machinery, Philadelphia, Pennsylvania, pp. 262–274.
4. Gauthier, J., Vincent, A.T., Charette, S.J. and Derome, N. (2018) A brief history of bioinformatics. *Briefings in Bioinformatics*, **20**, 1981-1996.
5. Welch, L., Lewitter, F., Schwartz, R., Brooksbank, C., Radivojac, P., Gaeta, B. and Schneider, M.V. (2014) Bioinformatics Curriculum Guidelines: Toward a Definition of Core Competencies. *PLoS Computational Biology*, **10**, e1003496.
6. Edman, P. (1949) A method for the determination of amino acid sequence in peptides. *Archives of Biochemistry*, **22**, 475.
7. Needleman, S.B. and Wunsch, C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, **48**, 443-453.
8. Sanger, F., Nicklen, S. and Coulson, A.R. (1977) DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, **74**, 5463-5467.
9. Staden, R. (1979) A strategy of DNA sequencing employing computer programs. *Nucleic Acids Research*, **6**, 2601-2610.
10. Devereux, J., Haeberli, P. and Smithies, O. (1984) A comprehensive set of sequence analysis programs for the VAX. *Nucleic Acids Research*, **12**, 387-395.
11. Marçais, G., Delcher, A.L., Phillippy, A.M., Coston, R., Salzberg, S.L. and Zimin, A. (2018) MUMmer4: A fast and versatile genome alignment system. *PLoS Computational Biology*, **14**, e1005944.
12. Karsch-Mizrachi, I., Takagi, T., Cochrane, G. and Collaboration, o.b.o.t.I.N.S.D. (2017) The international nucleotide sequence database collaboration. *Nucleic Acids Research*, **46**, D48-D51.
13. Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860-921.

14. Wooley, J.C. and Ye, Y. (2007) In Xu, Y., Xu, D. and Liang, J. (eds.), *Computational Methods for Protein Structure Prediction and Modeling: Volume 1: Basic Characterization*. Springer New York, New York, NY, pp. 1-43.

15. Consortium, T.U. (2018) UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Research*, **47**, D506-D515.

16. Sayers, E.W., Beck, J., Brister, J.R., Bolton, E.E., Canese, K., Comeau, D.C., Funk, K., Ketter, A., Kim, S., Kimchi, A. *et al.* (2019) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, **48**, D9-D16.

17. Rice, P., Longden, I. and Bleasby, A. (2000) EMBOSS: The European Molecular Biology Open Software Suite. *Trends in Genetics*, **16**, 276-277.

18. Sarachu, M. and Colet, M. (2004) wEMBOSS: a web interface for EMBOSS. *Bioinformatics*, **21**, 540-541.

19. Goodwin, S., McPherson, J.D. and McCombie, W.R. (2016) Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, **17**, 333-351.

20. Li, Y. and Chen, L. (2014) Big Biological Data: Challenges and Opportunities. *Genomics, Proteomics & Bioinformatics*, **12**, 187-189.

21. van Dijk, E.L., Jaszczyszyn, Y., Naquin, D. and Thermes, C. (2018) The Third Revolution in Sequencing Technology. *Trends in Genetics*, **34**, 666-681.

22. Jain, M., Koren, S., Miga, K.H., Quick, J., Rand, A.C., Sasani, T.A., Tyson, J.R., Beggs, A.D., Dilthey, A.T., Fiddes, I.T. *et al.* (2018) Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nature Biotechnology*, **36**, 338-345.

23. Chaisson, M.J.P., Huddleston, J., Dennis, M.Y., Sudmant, P.H., Malig, M., Hormozdiari, F., Antonacci, F., Surti, U., Sandstrom, R., Boitano, M. *et al.* (2015) Resolving the complexity of the human genome using single-molecule sequencing. *Nature*, **517**, 608-611.

24. Travers, K.J., Chin, C.-S., Rank, D.R., Eid, J.S. and Turner, S.W. (2010) A flexible and efficient template format for circular consensus sequencing and SNP detection. *Nucleic Acids Research*, **38**, e159-e159.

25. Wang, Y., Zhao, Y., Bollas, A., Wang, Y. and Au, K.F. (2021) Nanopore sequencing technology, bioinformatics and applications. *Nature Biotechnology*, **39**, 1348-1365.

26. Sedlazeck, F.J., Lee, H., Darby, C.A. and Schatz, M.C. (2018) Piercing the dark matter: bioinformatics of long-range sequencing and mapping. *Nature Reviews Genetics*, **19**, 329-346.

27. Goldberg, A.P., Szigeti, B., Chew, Y.H., Sekar, J.A.P., Roth, Y.D. and Karr, J.R. (2018) Emerging whole-cell modeling principles and methods. *Current Opinion in Biotechnology*, **51**, 97-102.

28. Karr, Jonathan R., Sanghvi, Jayodita C., Macklin, Derek N., Gutschow, Miriam V., Jacobs, Jared M., Bolival, B., Assad-Garcia, N., Glass, John I. and Covert, Markus W. (2012) A Whole-Cell Computational Model Predicts Phenotype from Genotype. *Cell*, **150**, 389-401.

29. Henikoff, S. and Henikoff, J.G. (1992) Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences of the United States of America*, **89**, 10915-10919.

30. Needleman, S.B. and Wunsch, C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, **48**, 443-453.

31. Smith, T.F., Waterman, M.S. and Fitch, W.M. (1981) Comparative biosequence metrics. *Journal of Molecular Biology*, **18**, 38-46.

32. Feng, D.F. and Doolittle, R.F. (1987) Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *Journal of Molecular Biology*, **25**, 351-360.

33. Bawono, P., Dijkstra, M., Pirovano, W., Feenstra, A., Abeln, S. and Heringa, J. (2017) In Keith, J. M. (ed.), *Bioinformatics: Volume I: Data, Sequence Analysis, and Evolution*. Springer New York, New York, NY, pp. 167-189.

34. Watanabe, S. (1985) *Pattern recognition: human and mechanical*. John Wiley & Sons, Inc.

35. de Ridder, D., de Ridder, J. and Reinders, M.J.T. (2013) Pattern recognition in bioinformatics. *Briefings in Bioinformatics*, **14**, 633-647.

36. Mount, D.W. (2001) *Bioinformatics: sequence and genome analysis*. Cold spring harbor laboratory press Cold Spring Harbor, NY.

37. Coward, E. and Drablos, F. (1998) Detecting periodic patterns in biological sequences. *Bioinformatics*, **14**, 498-507.

38. Verstrepen, K.J., Jansen, A., Lewitter, F. and Fink, G.R. (2005) Intragenic tandem repeats generate functional variability. *Nature Genetics*, **37**, 986-990.

39. Ellegren, H. (2004) Microsatellites: simple sequences with complex evolution. *Nature Reviews Genetics*, **5**, 435-445.

40. Gusfield, D. and Stoye, J. (2004) Linear time algorithms for finding and representing all the tandem repeats in a string. *Journal of Computer and System Sciences*, **69**, 525-546.

41. Albà, M.M., Laskowski, R.A. and Hancock, J.M. (2002) Detecting cryptically simple protein sequences using the SIMPLE algorithm. *Bioinformatics*, **18**, 672-678.

42. Marsella, L., Sirocco, F., Trovato, A., Seno, F. and Tosatto, S.C.E. (2009) REPETITA: detection and discrimination of the periodicity of protein solenoid repeats by discrete Fourier transform. *Bioinformatics*, **25**, i289-i295.

43. Edgar, R.C. and Sjölander, K. (2004) COACH: profile–profile alignment of protein families using hidden Markov models. *Bioinformatics*, **20**, 1309-1318.

44. Palidwor, G.A., Shcherbinin, S., Huska, M.R., Rasko, T., Stelzl, U., Arumughan, A., Foulle, R., Porras, P., Sanchez-Pulido, L., Wanker, E.E. *et al.* (2009) Detection of Alpha-Rod Protein Repeats Using a Neural Network and Application to Huntingtin. *PLoS Computational Biology*, **5**, e1000304.

45. Baxevanis, A.D. and Bateman, A. (2015) The Importance of Biological Databases in Biological Discovery. *Current Protocols in Bioinformatics*, **50**, 1.1.1-1.1.8.

46. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool. *Journal of Molecular Biology*, **215**, 403-410.

47. Helmy, M., Crits-Christoph, A. and Bader, G.D. (2016) Ten Simple Rules for Developing Public Biological Databases. *PLoS Computational Biology*, **12**, e1005128.

48. Rigden, D.J. and Fernández, X.M. (2021) The 2022 Nucleic Acids Research database issue and the online molecular biology database collection. *Nucleic Acids Research*, **50**, D1-D10.

49. Horgan, R.P. and Kenny, L.C. (2011) 'Omic' technologies: genomics, transcriptomics, proteomics and metabolomics. *The Obstetrician & Gynaecologist*, **13**, 189-195.

50. Vailati-Riboni, M., Palombo, V. and Loor, J.J. (2017) In Ametaj, B. N. (ed.), *Periparturient Diseases of Dairy Cows: A Systems Biology Approach*. Springer International Publishing, Cham, pp. 1-7.

51. Mayer, B. (2011) *Bioinformatics for omics data: methods and protocols*. Springer.

52. Voelkerding, K.V., Dames, S.A. and Durtschi, J.D. (2009) Next-Generation Sequencing: From Basic Research to Diagnostics. *Clinical Chemistry*, **55**, 641-658.

53. Wang, Z., Gerstein, M. and Snyder, M. (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nature Review Genetics*, **10**, 57-63.

54. Aebersold, R. and Mann, M. (2003) Mass spectrometry-based proteomics. *Nature*, **422**, 198-207.

55. Park, P.J. (2009) ChIP-seq: advantages and challenges of a maturing technology. *Nature Reviews Genetics*, **10**, 669-680.

56. Hardcastle, T.J. (2013) High-throughput sequencing of cytosine methylation in plant DNA. *Plant Methods*, **9**, 16.

57. Goh, H.-H. (2018) In Aizat, W. M., Goh, H.-H. and Baharum, S. N. (eds.), *Omics Applications for Systems Biology*. Springer International Publishing, Cham, pp. 69-80.

58. Steinwand, M.A. and Ronald, P.C. (2020) Crop biotechnology and the future of food. *Nature Food*, **1**, 273-283.

59. Morrell, P.L., Buckler, E.S. and Ross-Ibarra, J. (2011) Crop genomics: advances and applications. *Nature Review Genetics*, **13**, 85-96.

60. Kersey, P.J. (2019) Plant genome sequences: past, present, future. *Current Opinion in Plant Biology*, **48**, 1-8.

61. Lockton, S. and Gaut, B.S. (2005) Plant conserved non-coding sequences and paralogue evolution. *Trends in Genetics*, **21**, 60-65.

62. Van Bel, M., Bucchini, F. and Vandepoele, K. (2019) Gene space completeness in complex plant genomes. *Current Opinion in Plant Biology*, **48**, 9-17.

63. Lam, E.T., Hastie, A., Lin, C., Ehrlich, D., Das, S.K., Austin, M.D., Deshpande, P., Cao, H., Nagarajan, N., Xiao, M. *et al.* (2012) Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly. *Nature Biotechnology*, **30**, 771-776.

64. Doğan, E.S. and Liu, C. (2018) Three-dimensional chromatin packing and positioning of plant genomes. *Nature Plants*, **4**, 521-529.

65. Belser, C., Istace, B., Denis, E., Dubarry, M., Baurens, F.-C., Falentin, C., Genete, M., Berrabah, W., Chèvre, A.-M., Delourme, R. *et al.* (2018) Chromosome-scale assemblies of plant genomes using nanopore long reads and optical maps. *Nature Plants*, **4**, 879-887.

66. Michael, T.P. and VanBuren, R. (2020) Building near-complete plant genomes. *Current Opinion in Plant Biology*, **54**, 26-33.

67. Twyford, A.D. (2018) The road to 10,000 plant genomes. *Nature Plants*, **4**, 312-313.

68. Casillas, S. and Barbadilla, A. (2017) Molecular Population Genetics. *Genetics*, **205**, 1003-1035.

69. Rafalski, A. and Morgante, M. (2004) Corn and humans: recombination and linkage disequilibrium in two genomes of similar size. *Trends in Genetics*, **20**, 103-111.

70. Purugganan, M.D. (2019) Evolutionary insights into the nature of plant domestication. *Current Biology*, **29**, R705-R714.

71. Goodstein, D.M., Shu, S., Howson, R., Neupane, R., Hayes, R.D., Fazo, J., Mitros, T., Dirks, W., Hellsten, U., Putnam, N. *et al.* (2011) Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Research*, **40**, D1178-D1186.

72. Howe, K.L., Contreras-Moreira, B., De Silva, N., Maslen, G., Akanni, W., Allen, J., Alvarez-Jarreta, J., Barba, M., Bolser, D.M., Cambell, L. *et al.* (2019) Ensembl Genomes 2020 - enabling non-vertebrate genomic research. *Nucleic Acids Research*, **48**, D689-D695.

73. Tello-Ruiz, M.K., Naithani, S., Stein, J.C., Gupta, P., Campbell, M., Olson, A., Wei, S., Preece, J., Geniza, M.J., Jiao, Y. *et al.* (2017) Gramene 2018: unifying comparative genomics and pathway resources for plant research. *Nucleic Acids Research*, **46**, D1181-D1189.

74. Portwood, J.L., II, Woodhouse, M.R., Cannon, E.K., Gardiner, J.M., Harper, L.C., Schaeffer, M.L., Walsh, J.R., Sen, T.Z., Cho, K.T., Schott, D.A. *et al.* (2018) MaizeGDB 2018: the maize multi-genome genetics and genomics database. *Nucleic Acids Research*, **47**, D1146-D1154.

75. Lai, K., Berkman, P.J., Lorenc, M.T., Duran, C., Smits, L., Manoli, S., Stiller, J. and Edwards, D. (2011) WheatGenome.info: An Integrated Database and Portal for Wheat Genome Information. *Plant and Cell Physiology*, **53**, e2-e2.

76. Sakai, H., Lee, S.S., Tanaka, T., Numa, H., Kim, J., Kawahara, Y., Wakimoto, H., Yang, C.-c., Iwamoto, M., Abe, T. *et al.* (2013) Rice Annotation Project Database (RAP-DB): An Integrative and Interactive Database for Rice Genomics. *Plant and Cell Physiology*, **54**, e6-e6.

77. Paterson, A.H., Bowers, J.E., Feltus, F.A., Tang, H., Lin, L. and Wang, X. (2009) Comparative Genomics of Grasses Promises a Bountiful Harvest. *Plant Physiology*, **149**, 125-131.

78. Swigoňová, Z., Lai, J., Ma, J., Ramakrishna, W., Llaca, V., Bennetzen, J.L. and Messing, J. (2004) Close Split of Sorghum and Maize Genome Progenitors. *Genome Research*, **14**, 1916-1923.

79. Paterson, A.H., Bowers, J.E., Bruggmann, R., Dubchak, I., Grimwood, J., Gundlach, H., Haberer, G., Hellsten, U., Mitros, T., Poliakov, A. *et al.* (2009) The Sorghum bicolor genome and the diversification of grasses. *Nature*, **457**, 551-556.

80. McCormick, R.F., Truong, S.K., Sreedasyam, A., Jenkins, J., Shu, S., Sims, D., Kennedy, M., Amirebrahimi, M., Weers, B.D., McKinley, B. *et al.* (2018) The *Sorghum bicolor* reference genome: improved assembly, gene annotations, a transcriptome atlas, and signatures of genome organization. *The Plant Journal*, **93**, 338-354.

81. Boyles, R.E., Brenton, Z.W. and Kresovich, S. (2019) Genetic and genomic resources of sorghum to connect genotype with phenotype in contrasting environments. *The Plant Journal*, **97**, 19-39.

82. Cooper, E.A., Brenton, Z.W., Flinn, B.S., Jenkins, J., Shu, S., Flowers, D., Luo, F., Wang, Y., Xia, P., Barry, K. *et al.* (2019) A new reference genome for Sorghum bicolor reveals high levels of sequence similarity between sweet and grain genotypes: implications for the genetics of sugar metabolism. *BMC Genomics*, **20**, 420.

83. Deschamps, S., Zhang, Y., Llaca, V., Ye, L., Sanyal, A., King, M., May, G. and Lin, H. (2018) A chromosome-scale assembly of the sorghum genome using nanopore sequencing and optical mapping. *Nature Communications*, **9**, 4844.

84. Hirsch, C.N., Foerster, J.M., Johnson, J.M., Sekhon, R.S., Muttoni, G., Vaillancourt, B., Peñagaricano, F., Lindquist, E., Pedraza, M.A., Barry, K. *et al.* (2014) Insights into the Maize Pan-Genome and Pan-Transcriptome. *The Plant Cell*, **26**, 121-135.

85. Lei, L., Goltsman, E., Goodstein, D., Wu, G.A., Rokhsar, D.S. and Vogel, J.P. (2021) Plant Pan-Genomics Comes of Age. *Annual Review of Plant Biology*, **72**, 411-435.

86. Meyer, R.S. and Purugganan, M.D. (2013) Evolution of crop species: genetics of domestication and diversification. *Nature Review Genetics*, **14**, 840-852.

87. Fuller, D.Q. and Stevens, C.J. (2018) In Mercuri, A. M., D'Andrea, A. C., Fornaciari, R. and Höhn, A. (eds.), *Plants and People in the African Past: Progress in African Archaeobotany*. Springer International Publishing, Cham, pp. 427-452.

88. Winchell, F., Stevens, C.J., Murphy, C., Champion, L. and Fuller, D. (2017) Evidence for Sorghum Domestication in Fourth Millennium BC Eastern Sudan: Spikelet Morphology from Ceramic Impressions of the Butana Group. *Current Anthropology*, **58**, 673-683.

89. Dahlberg, J., Berenji, J., Sikora, V. and Latković, D. (2012) Assessing sorghum [*Sorghum bicolor* (L) Moench] germplasm for new traits: food, fuels & unique uses. *Maydica*, **56**.

90. Aruna, C., Visarada, K., Bhat, B.V. and Tonapi, V.A. (2018) *Breeding sorghum for diverse end uses*. Woodhead Publishing.

91. Anami, S.E., Zhang, L.M., Xia, Y., Zhang, Y.M., Liu, Z.Q. and Jing, H.C. (2015) Sweet sorghum ideotypes: genetic improvement of the biofuel syndrome. *Food Energy Security*, **4**, 159-177.

92. Umakanth, A.V., Kumar, A.A., Vermerris, W. and Tonapi, V.A. (2019) In Aruna, C., Visarada, K. B. R. S., Bhat, B. V. and Tonapi, V. A. (eds.), *Breeding Sorghum for Diverse End Uses*. Woodhead Publishing, pp. 255-270.

93. Hilson, G. (1916) On the inheritance of certain stem characters in sorghum. *Agriculture Journal India*, **11**, 150-155.

94. Burks, P.S., Kaiser, C.M., Hawkins, E.M. and Brown, P.J. (2015) Genomewide Association for Sugar Yield in Sweet Sorghum. *Crop Science*, **55**, 2138-2148.

95. Gaut, B.S., Seymour, D.K., Liu, Q. and Zhou, Y. (2018) Demography and its effects on genomic variation in crop domestication. *Nature Plants*, **4**, 512-520.

96. Pickrell, J.K. and Pritchard, J.K. (2012) Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genetics*, **8**, e1002967.

97. Chen, H., Patterson, N. and Reich, D. (2010) Population differentiation as a test for selective sweeps. *Genome Research*, **20**, 393-402.

CHAPTER 2

2

Understanding and identifying amino acid repeats

Hong Luo and Harm Nijveen

Based on: Briefings in Bioinformatics, 2013;15;582-591

Abstract

Amino acid repeats (AARs) are abundant in protein sequences. They have particular roles in protein function and evolution. Simple repeat patterns generated by DNA slippage tend to introduce length variations and point mutations in repeat regions. Loss of ordinary and gain of abnormal function owing to their variable length are potential risks leading to diseases. Repeats with complex patterns mainly refer to the functional domain repeats, such as the well-known leucine-rich repeat and WD repeat, which are frequently involved in protein-protein interaction. They are primarily derived from internal gene duplication events and stabilized by 'gate-keeper' residues, which play crucial roles in preventing inter-domain aggregation. AARs are widely distributed in different proteomes across various taxonomic ranges and are especially abundant in eukaryotic proteins. However, their specific evolutionary and functional scenarios are still poorly understood. Identifying AARs in protein sequences is the first step for the further investigation of their biological function and evolutionary mechanism. In principle, this is an NP-hard problem, as most of the repeat fragments are shaped by a series of sophisticated evolutionary events and become latent periodical patterns. It is not possible to define a uniform criterion for detecting and verifying various repeat patterns. Instead, people have developed different algorithms based on different strategies to cope with varying patterns of repeat. In this review, we attempt to describe the amino acid repeat-detection algorithms currently available and compare their strategies based on an in-depth analysis of the biological significance of protein repeats.

2.1 Introduction

Amino acid repeats (AARs) are abundant in protein sequences either as periodic elements in structural proteins such as collagens, keratins, silk and cell wall proteins, or as structural modules in functional proteins such as transcription factors, receptors, ion channels, histones, ubiquitins and calcium storage proteins. Table 2.1 shows some well-known examples of human repeat-containing proteins (RCPs) gathered in the UniProt/Swiss-Prot Knowledgebase (http://www.uniprot.org/). For example, the major prion protein (PRIO_HUMAN) contains an N-terminal repeat region with several octamers (PHGGGWGQ); the extra-embryonic spermatogenesis homeobox 1 protein (ESX1_HUMAN) has a sequence motif PPxxPxPPx repeated nine times and the alpha-1 type I collagen protein contains a repeat of various lengths of the periodic tri-amino acid GPP. The giant muscle protein Titin composed of 34 350 amino acid residues (TITIN_HUMAN) contains several types of repeating domains. Single amino acid repeats (SAARs) are also common, such as the polyQ repeats in the Forkhead box protein P2 (FOXP2_HUMAN), the androgen receptor (ANDR_HUMAN) and the Huntington's disease (HD) protein (HD_HUMAN). Other SAARs including polyL, polyA and polyH can also be found in many other proteins. RCPs are distributed in all life kingdoms, and especially abundant in eukaryotes (1).

It is known that some AARs such as the leucine rich repeats (LRRs) form the structural framework for protein–protein interaction, and the repeat fragment in zinc finger transcription factors binds to *cis*-elements of DNA promoters. AARs can also cause problems such as the mis-folding of prion proteins (2). Furthermore, modification of repeat length may introduce abnormal function. A typical case is the expansion of polyQ, resulting in several neurological disorders such as mental retardation, HD, inherited ataxias and muscular dystrophy.

Table 2.1 **Some examples of amino acid repeats**

UniProt ID	Description	AA	Repeat Pattern
SECR_HUMAN	Secretin	121	polyL
PRIO_HUMAN	Major prion protein	253	$(PHGGGWGQ)_4$
ANKR1_HUMAN	Ankyrin repeat domain-containing protein 1	319	Ankyrin repeat
CASQ2_HUMAN	Calsequestrin-2	399	D/E-Rich
ESX1_HUMAN	Homeobox protein ESX1	406	$(PPxxPxPPx)_9$
WDR1_HUMAN	WD repeat-containing protein 1	606	WD repeat
UBC_HUMAN	Polyubiquitin-C	685	Ubiquitin
FOXP2_HUMAN	Forkhead box protein P2	715	polyQ
LRRN1_HUMAN	Leucine-rich repeat neuronal protein 1	716	Leucine Rich Repeat
ANDR_HUMAN	Androgen receptor	919	polyQ, polyG, polyP
SRBP2_HUMAN	Sterol regulatory element-binding protein 2	1141	polyS, $(PQ)_{4,}$ $(SGSS)_2$
BRD4_HUMAN	Bromodomain-containing protein 4	1362	polyP, polyH, polyQ K-Rich S-Rich
CO1A1_HUMAN	Collagen alpha-1(I) chain	1464	$(GPP)_n$
CAC1A_HUMAN	Brain calcium channel I	2505	polyQ, polyH, polyG
HD_HUMAN	Huntington disease protein	3142	polyQ, polyP, polyT, polyE, HEAT domain
MLL2_HUMAN	Histone-lysine N-methyltransferase MLL2	5537	$(S/P-P-P-E/P-E/A)_{15}$
TITIN_HUMAN	Titin	34350	Several types of repeating domains: TPR WD RCC1 PEVK Kelch Z Ig repeats

2.2 Classification of amino acid repeat patterns at sequence level

Mathematical and statistical methodologies can be applied to study the particular functional and evolutionary background of an AAR. Several approaches have been proposed to classify AARs into different categories depending on the characteristics of repeat units, including the sequence similarity among repeat units, the distance between adjacent repeat units and the complexity of the sequence pattern of the repeat units.

The first approach is to classify AARs according to the similarity among the repeat units. Based on this approach, AARs can be classified into two main groups: perfect repeats and imperfect repeats. The repeat units in perfect repeat fragments are identical, e.g. AAAAAAA and PQPQPQPQ, whereas the repeat units in imperfect repeat fragments are not exactly the same, e.g. AAWAAAA and QQQMLQQQFL. Imperfect repeats with highly variable, but still recognizable, repeat units are also called divergent repeats.

The second approach for repeat classification is based on the distance between adjacent units. AARs can be classified as tandem repeats (TRs) or non-tandem repeats (NTRs). The units in TRs are continuously distributed in the repeat sequence, whereas the units in NTRs are sequentially interspersed.

The third approach takes the complexity of the sequence pattern of the repeat units into consideration. Based on this approach, AARs can be roughly classified as simple repeats or complex repeats. Simple repeats generally refer to the continuous or interrupted runs of single amino acid residues or short peptides. The regions in a protein sequence containing simple repeats are often called simple sequences (SSs) or low complexity regions (LCRs). On the other hand, most of the complex repeats usually have sophisticated patterns of repeat units with variable lengths ranging from 10 to >100 residues, and these complex repeats patterns are frequently recognized as repeated protein domains (3).

In practice, it is rather difficult to strictly distinguish the different classes owing to the complicated patterns of AARs. For example, some domain repeats also contain SSs, such as the abundant leucine residues found in an LRR domain. And in the case of point mutations or insertions/deletions (INDELs), the original perfectly repeated units in proteins could gradually evolve into non-perfect tandem repeats (NPTRs).

The above approaches used to classify AARs are all based on the protein sequence. However, they are insufficient to reveal the biological significance of AARs, as proteins play their functional roles by folding into particular secondary and tertiary structures, which are difficult to deduce through amino acid patterns at sequence level. Data from several experiments show that proteins with similar tertiary structures may share low sequence identity (4,5). And similar functional domains of proteins do not necessarily correspond to recognizable sequence repeat patterns (3,6-8). Therefore, in-depth study of protein repeats requires better understanding of the correspondence of repeat sequences with their structures and functions. In addition, the acquisition of such biological knowledge is more sophisticated than simply classifying sequential repeat data.

2.3 Biological significance of different patterns of AARs

Biologically, different amino acid repeat patterns imply different functional and evolutionary backgrounds. Repeats with simple patterns, such as single AARs, mainly exist in intrinsically unstructured regions (IURs) of proteins (9,10). Such protein regions that do not fold into a 3D structure commonly have functions related to molecular recognition and molecular assembly (11,12). Single amino acid or trinucleotide repeats like polyQ are involved in neurodegenerative diseases such as HD

(13), where their length variations often result in either loss of normal or gain of abnormal function (14,15).

Most SAARs are presumed to be originally derived from replicative DNA slippage (16) in the coding region. Expansion of some SAARs might also result from unequal chromosomal crossover, such as the polyA in the human HOX13 gene (17). In general, perfect amino acid runs are inherently mutable and are frequently interrupted by point mutations (18) to become simple sequences (19).

In addition to SAARs, sequential tandem repeats (PTRs and NPTRs) with highly similar units are prevalent in protein sequences. We have found that ~13% of all proteins deposited in the public protein databases contain at least one tandem repeat fragment. And >40% of the tandem repeats are PTRs, while ~60% PTRs are single amino acid runs (1). Errors in sequencing and automatic annotation procedures might have introduced some false-positive PTRs into the public protein knowledgebase. However, this cannot undermine the biological significance of frequently occurring PTRs in protein sequences, especially considering the fact that functional PTRs are being continuously experimentally identified, and most of them are conserved among orthologous proteins (20-22).

Consistent with this scenario, conservation of amino acid tandem repeats is a strong indication for biological relevance. The phylogenetically conserved repeat fragments among orthologous proteins should have a conserved function, such as the conserved polyQ regions in primate FOXP2 proteins (23). In contrast, however, variable repeat unit length in corresponding regions of orthologous proteins indicates a different scenario. These repeats are probably going through a rapid change driven by selection (24). More interestingly, tandem repeats have been shown to play an important role in micro-evolution by catalysing the rapid production of genetic and phenotypic variation among organisms (25-28).

Repeats with complex patterns have comparatively stable structures and conserved functions, which are generally called domain repeats. Domain repeats are among the most common protein motifs in the Pfam database (29), such as LRRs, Zinc finger repeats, Ankyrin repeats and Tetratricopeptide repeats (TPRs) (30). These domain repeats are mostly involved in transcription regulation, cell-cycle control and signal transduction (31-34) and widely spread in the proteomes of different species across different life kingdoms (35). Many genes containing these domain repeats in the coding region are significant in certain diseases (36), as sequence identity increases the chance of protein aggregation (37) and mis-folding. Domain repeats are thought to have evolved through internal gene duplications arising from recombination events (3,38), such as unequal crossing over (39) and exon shuffling (40). The duplications may involve several domains at a time (3,41) In addition, a number of specific sequence-based signals such as the ‘gate-keeper’ residues (41) play a crucial role in preventing inter-domain aggregation. Therefore, these repeat patterns are generally obscure at sequence level, and a sophisticated search is required to detect them.

2.4 Repeat Detection Strategies

During the past decade, several strategies for the identification of AARs from protein sequences have been reported. Among these approaches, the three major ones are self-comparison, pattern recognition and complexity measurement. Table 2.1 shows the algorithms and publicly available tools including online resources that can be used to detect AARs of various types. In the following section, we will give a brief introduction to the amino acid repeat-detection strategies focusing on the general principles behind these strategies.

Table 2.2 **Repeat detection algorithms**

Method	Repeat type[a]	Ref	Availability
Self-comparison			
REP	Domain	(42)	http://www.embl.de/~andrade/papers/rep/search.html
COACH	Domain	(43)	http://www.drive5.com/lobster/
TPRpred	Domain	(44)	http://tprpred.tuebingen.mpg.de/
REPRO	Domain	(45)	http://www.ibi.vu.nl/programs/reprowww/
TRUST	Divergent	(46)	http://www.ibi.vu.nl/programs/trustwww/
Internal Repeat Finder	Divergent	(47)	http://nihserver.mbi.ucla.edu/Repeats/
HHrep	Divergent	(48)	http://hhrep.tuebingen.mpg.de/hhrep/
RADAR	Divergent	(49)	http://www.ebi.ac.uk/Tools/Radar/
HHrepID	Divergent	(50)	http://toolkit.tuebingen.mpg.de/hhrepid/
Pattern recognition			
REPETITA	Solenoid	(51)	http://protein.bio.unipd.it/repetita/
LSTM	Domain	(52)	http://www.bioinf.jku.at/software/LSTM_protein/
ARD	Alpha-Rod	(53)	http://www.ogic.ca/projects/ard/
Complexity measurement			
SIMPLE	Simple	(19)	http://www.biochem.ucl.ac.uk/bsm/SIMPLE/
GBA	Simple	(54)	xli@cise.ufl.edu
Others			
XSTREAM	ATR	(55)	http://jimcooperlab.mcdb.ucsb.edu/xstream/
Apriod	PPP	(56)	hwan@mindgen.org
LocRepeat	PPP	(57)	http://www.cs.cityu.edu.hk/~lwang/software/LocRepeat/
REPfind	ATR	(58)	adebiyi@informatik.uni-tuebingen.de
Reptile	Perfect	(59)	http://reptile.unibe.ch/
SUFFIX	Perfect	(60)	http://www.cs.ucdavis.edu/~gusfield/strmat.html

[a] NPTR=non-perfect tandem repeat; PPP=pseudo-periodic partitions.

2.4.1 The self-comparison strategy

One of the most intuitive strategies to detect repeat patterns in protein sequences is the self-comparison method. The idea of this approach is rather simple, i.e. comparing a protein sequence to itself. Sequence comparison is a fundamental bioinformatics method that has been extensively used to search similar regions among biological sequences. The global sequence alignment method was first proposed in the 1970s (61) and focuses on finding the optimal alignment of two entire biological sequences using dynamic programming. Soon after, the Smith-Waterman local alignment algorithm (62) was developed to recognize the better aligned sub-regions between two sequences in order to show meaningful biological relevance.

On aligning a sequence with itself for the purpose of identifying repeat patterns, the sub-optimal alignments become obscured by the best (and most obvious) alignment. This optimal alignment should be excluded from the initial search. The reliability of identifying sub-optimal alignments of protein sequences using the dynamic programming method has been evaluated (62). A very distinguishing feature of this method is the use of a scoring system that gives scores to paired amino acids and penalties to unmatched gaps. Substitution matrices such as PAM (63) and BLOSUM (64) are the basis of the scoring system and represent the specific evolutionary relevance among different amino acids. More specifically tuned scoring matrices have also been proposed. These matrices take special features of amino acids such as polarity, electrostatic charge, structure, molecular volume and codon bias (65) into account. One of the greatest advantages of using a scoring system for identifying sub-optimal alignments is that statistical models can be applied to define reliable criteria (66,67).

In principle, the self-alignment repeat-detection methods are the extension of an alignment-based homology-detection approach. Thus, they have inherited characteristics that are more suitable for detecting divergent internal repeats in protein sequences. The units of these repeats generally have low identities and ambiguous boundaries, but share evolutionarily conserved sites or motifs, which are presumed to have crucial functions. As such, the accurate definition of repeat length and repeat number according to substantial biological significance is a sophisticated problem. And this is especially true for detecting repeat patterns without prior knowledge, also called 'de novo' repeat detection. On the other hand, the algorithms depending on prior knowledge, such as REP, COACH and TPRpred (42-44), generally search repeat patterns from sequence databases by profiles constructed with known repeat families using hidden Markov models (HMMs) (68). Therefore, the repeat patterns identified by these programs are usually well-known, and some of them are experimentally studied functional protein domain repeats.

It is generally believed that detecting repeat patterns with a self-alignment-based method is a feasible strategy. However, it also has some flaws and limitations. First, the computational complexity of performing self-alignment is high, as the general complexity for a sequence with n amino acids is O (n^2) for both time and space. Fortunately, this problem is not too serious for protein sequences, as their average length is around 320 AA (69). And the computational capacity of current computer hardware is powerful enough to handle this problem within acceptable time and space.

In addition, several optimization strategies have been recently applied to sequence alignments, such as the implementation of the Smith–Waterman algorithm with the new technology of graphics processing units (GPUs) (70), and the parallel computing version of the REPRO (71) algorithm (72) can handle much longer sequences within a reasonable time.

One of the main purposes for detecting AARs is to find novel repeat patterns and infer their functional and evolutionary roles. As the majority of repeat patterns in protein sequences have not been well studied, de novo repeat-detection algorithms are more widely used, such as PEPRO, Internal Repeat Finder, RADAR, TRUST, HHrep and HHrepID (45-50,56,57). All of them identify repeats using the self-comparison strategy, but differ in some aspects. For example, Internal Repeat Finder assumes that the statistically significant sub-optimal alignment scores should have a Poisson distribution (47). TRUST uses the particular strategy on sub-optimal alignments, which could increase the chance and reliability to identify divergent repeats (46). HHrep (48) and its optimized version HHrepID (50) compares a sequence with itself by the HMM–HMM (73) strategy, which looks for the sub-optimal alignments using a profile HMM constructed by iterations of PSI-BLAST (74).

2.4.2 The pattern recognition strategy

The second strategy to detect AARs from protein sequences uses the conventional method of pattern recognition. The two main algorithms of this strategy are the discrete Fourier transform (DFT) and neural networks.

DFT has been widely applied in the research area of signal processing. Generally, it can decompose signals into constituent frequencies, so that the cryptic patterns hidden in the signals could be analysed intuitively. Early studies showed that DFT can be used to detect periodic patterns in collagen protein (75), but also has some fundamental difficulties which limit its usage (45). The accuracy of DFT-based methods is easily biased by the length variation of the repeat units caused by mutations or INDELs, as this will weaken the periodical pattern of the transformed Fourier spectral amplitudes.

Some recent algorithms make efforts to provide better discrimination on Fourier spectral amplitudes using newly developed methods. For example, REPETITA yields better accuracy than self-alignment methods on detecting protein solenoid repeats (51) by introducing several optimized strategies of the DFT-based method (51). In addition, the stationary wavelet packet transform has been widely used in bioinformatics and computational biology in recent years (76). As a state of the art optimization DFT algorithm (77), it has been shown to have good quality on detecting protein repeat patterns (78).

The neural network-based method is another well-studied pattern-recognition strategy, which is also capable of identifying similar patterns in protein sequences (79). A well-established neural network is able to associate homologous patterns in the protein sequence with the input patterns and can be trained to adapt the patterns. Several neural network algorithms show good accuracy and time efficiency on protein homologue detection. LSTM is able to combine amino acid properties with patterns and does not rely on pre-defined scoring matrices for similarity measurements (52). The ARD neural network is designed to identify specific alpha-rod repeat patterns and has been applied to the analysis of Huntingtin protein sequences (53).

2.4.3 The complexity measurement strategy

The third approach of identifying AARs takes complexity measurement into consideration. LCRs are widely distributed in protein sequences. LCRs commonly contain particular repeat patterns that have continuous repetitions of very short units, such as the SAARs and cryptically simple sequences (19). Apparently, these repeats have special functional and evolutionary properties that differ from the repeats with more complex patterns and longer units. Their typical short unit length makes both the self-comparison and the pattern recognition-based strategies less well suited to identify LCR repeats efficiently.

Fortunately, several algorithms have been introduced to detect repeats involved in LCRs, most of them using a strategy to measure the complexity of sequences within a sliding window. As for complexity measuring, SIMPLE (19) awards simplicity score to the central amino acid of each window, and is most suitable for detecting short unit cryptic repeats. SEG (80), DSR (81), and CARD (82) are based on Shannon entropy (83), which displays several limitations when decoding complex protein sequences (43).

The main drawback of sliding windows-based algorithms is that they all require a pre-specified window size, and repeats that are longer or shorter than the window is not detectable. On the other hand, non-sliding window algorithms show more flexibility on detecting repeats in LCRs. GBA (54) constructs a graph for each protein sequence, and finds short subsequences as LCR candidates through traversing. Coronado et al. (84) introduce the composition-modified scoring matrices to identify LCRs within cell wall proteins of fungi. These algorithms are an important complement to the sliding window-based algorithms.

2.4.4 Other strategies

As described above, the self-comparison strategy and the pattern recognition strategy are mostly suitable for detecting divergent repeats, whereas the complexity measurement strategy is mostly suitable for detecting simple unit repeats. In addition, exclusive and optimized strategies for sequential tandem repeats are also particularly useful. Sequential tandem repeats implicated in the amino acid fragments with tandem repeat patterns are comparatively more explicit than divergent repeats. They are widely spread in many proteomes across wide taxonomic ranges, but are still insufficiently studied.

Hamming distance (85) and edit distance, also called Levenshtein distance (86) are widely used for measuring the similarity of sequential tandem repeats (87-90). Differing from hamming distance, which only accounts for point mutations, edit distance-measuring algorithms also consider insertions and deletions. In addition, Apriod (56) and LocRepeat (57) focus on finding the ‘pseudo-periodic partitions’, which are gradually evolved patterns among repeat units. Given that NPTRs are originally evolved from PTRs, Xstream (55) and REPfind (58) detect NPTRs based on the extension of exact repeats seeds, which could decrease the computational complexity of both time and space.

Most of the repeat-detection algorithms can identify PTRs together with other repeat patterns incidentally. But as some of the PTRs are nested in larger NPTR fragments, which can hardly be distinguished by the common strategies, an exclusive algorithm for detecting PTRs is also necessary. For example, the suffix tree-based strategy is supportive to identify all PTRs in a protein sequence with linear time complexity (60). Reptile uses a ‘brute-force’ strategy to detect PTRs from the proteins of parasite antigens (59). Following the definition of statistically significant repeat runs in protein sequences (91), the cut-off sizes of five, four, three and two of the repeat unit repetitions are common criteria for identifying mono-amino, di-amino, tri-amino and all other repeats, respectively.

2.5 Summary and perspective

Identifying repeat patterns in proteins is the first step towards the understanding of their physiological function and evolutionary mechanism. During evolution, these patterns become so intricate that no single algorithm is adequate to identify all of them. There is no doubt that an in-depth investigation of their biological background is required to choose proper algorithms for the identification of specific patterns. In general, self-comparison algorithms are suitable to detect de novo repeats with complex patterns. Pattern recognition-based algorithms are suitable to detect repeats with low sequence identities but high intrinsic biological similarities. Complexity measurement-based algorithms can be applied to detect repeats with simple patterns involved in LCRs. For the tandem repeats that have more sequentially repetitive patterns, one should consider the strategies that measure the similarity of repeat units by edit or hamming distance.

The biological significance of protein repeats has been discussed for years. Internal duplication in genomes is one of the most important evolutionary mechanisms for species to adapt the environment (92-94). As a result, repetitive patterns at the DNA level such as interspersed microsatellites and tandem tri-nucleotide repeats are prevalent. Intragenic repeats are presumed to have potential roles on generating functional variability (95,96). Moreover, repeats in coding regions corresponding to AARs are more likely to go through adaptive competition (24,97,98). Therefore, protein repeats are increasingly considered interesting, with potentially large functional consequences (99). At the same time, their variable characters and involvement in disorder and diseases have been scientific puzzles for a long time. Frequently asked questions are: are the characteristics of similar repeat patterns coherent in different proteomes across life kingdoms? Could the functional and evolutionary roles of certain repeats correspond to their particular characters, such as position bias, GC content constraints and codon usage? How could the conserved functions of particular repeats have been evolved by selection? And what structure and sequence-based strategies does the cell employ to prevent repeats from aggregation?

The insufficient understanding of protein repeats is not only due to the difficulty of identification, but also because of the lack of an integrated repository for large-scale investigation and comparison of repeats among a variety of proteomes across different kingdoms. To that end, we developed ProRepeat (http://prorepeat.bioinformatics.nl), which integrates non-redundant tandem repeats detected by several algorithms from the UniProt (69) and RefSeq (100) protein databases and offers powerful analysis tools for finding biologically interesting properties of query results. In addition, we also integrated ProRepeat with ProGMap - a tool we developed for the integration of annotation resources for protein orthology (101). With this set-up, we will be making large-scale orthologous comparisons on protein repeats over a broad taxonomy range especially eukaryotes in the near future.

Acknowledgements

This article is dedicated to the memory of Jack Leunissen, a former Professor of Bioinformatics at Wageningen University and Research Center in the Netherlands, and a former Editorial Board member of Briefings in Bioinformatics, who passed away on 14 May 2012. He conceived and supervised the Protein Repeat project, which was supported partly by the BioRange project of the Netherlands Bioinformatics Centre (NBIC). Thanks to Dr Lisa Mullan for critically reading and language editing of this manuscript. Thanks to Professor Lettie Lubsen for critically reading and informative comments of the manuscript.

References

1. Luo, H., Lin, K., David, A., Nijveen, H. and Leunissen, J.A.M. (2011) ProRepeat: an integrated repository for studying amino acid tandem repeats in proteins. *Nucleic Acids Research*, **40**, D394-D399.

2. Cordeiro, Y., Kraineva, J., Gomes, M.P.B., Lopes, M.H., Martins, V.R., Lima, L.M.T.R., Foguel, D., Winter, R. and Silva, J.L. (2005) The Amino-Terminal PrP Domain Is Crucial to Modulate Prion Misfolding and Aggregation. *Biophysical Journal*, **89**, 2667-2676.

3. Björklund, A.K., Ekman, D. and Elofsson, A. (2006) Expansion of protein domain repeats. *PLoS Computational Biology*, **2**, e114.

4. Sussman, J.L., Lin, D., Jiang, J., Manning, N.O., Prilusky, J., Ritter, O. and Abola, E.E. (1998) Protein Data Bank (PDB): Database of Three-Dimensional Structural Information of Biological Macromolecules. *Acta Crystallographica Section D*, **54**, 1078-1084.

5. Chikenji, G., Fujitsuka, Y. and Takada, S. (2006) Shaping up the protein folding funnel by local interaction: lesson from a structure prediction study. *Proceedings of the National Academy of Sciences of the United States of America*, **103**, 3141-3146.

6. Ferreiro, D.U., Walczak, A.M., Komives, E.A. and Wolynes, P.G. (2008) The energy landscapes of repeat-containing proteins: topology, cooperativity, and the folding funnels of one-dimensional architectures. *PLoS Computational Biology*, 10.1371/journal.pcbi.1000070.

7. Main, E.R.G., Lowe, A.R., Mochrie, S.G.J., Jackson, S.E. and Regan, L. (2005) A recurring theme in protein engineering: the design, stability and folding of repeat proteins. *Current Opinion in Structural Biology*, **15**, 464-471.

8. Ferreiro, D.U. and Komives, E.A. (2007) The plastic landscape of repeat proteins. *Proceedings of the National Academy of Sciences of the United States of America*, **104**, 7735-7736.

9. Simon, M. and Hancock, J.M. (2009) Tandem and cryptic amino acid repeats accumulate in disordered regions of proteins. *Genome Biology*, **10**, R59.

10. Dunker, A.K., Brown, C.J., Lawson, J.D., Iakoucheva, L.M. and Obradović, Z. (2002) Intrinsic disorder and protein function. *Biochemistry*, **41**, 6573-6582.

11. Dunker, A.K., Silman, I., Uversky, V.N. and Sussman, J.L. (2008) Function and structure of inherently disordered proteins. *Current Opinion in Structural Biology*, **18**, 756-764.

12. Dunker, A.K. and Uversky, V.N. (2010) Drugs for 'protein clouds': targeting intrinsically disordered transcription factors. *Current opinion in Pharmacology*, **10**, 782-788.

13. Orr, H.T. and Zoghbi, H.Y. (2007) Trinucleotide Repeat Disorders. *Annual Review of Neuroscience*, **30**, 575-621.

14. Buchanan, G., Yang, M., Cheong, A., Harris, J.M., Irvine, R.A., Lambert, P.F., Moore, N.L., Raynor, M., Neufing, P.J., Coetzee, G.A. *et al.* (2004) Structural and functional consequences of glutamine tract variation in the androgen receptor. *Human Molecular Genetics*, **13**, 1677-1692.

15. Brown, L., Paraso, M., Arkell, R. and Brown, S. (2005) In vitro analysis of partial loss-of-function ZIC2 mutations in holoprosencephaly: alanine tract expansion modulates DNA binding and transactivation. *Human Molecular Genetics*, **14**, 411-420.

16. Levinson, G. and Gutman, G.A. (1987) Slipped-strand mispairing: a major mechanism for DNA sequence evolution. *Molecular Biology and Evolution*, **4**, 203-221.

17. Warren, S.T., Muragaki, Y., Mundlos, S., Upton, J. and Olsen, B.R. (1997) Polyalanine Expansion in Synpolydactyly Might Result from Unequal Crossing-Over of HOXD13. *Science*, **275**, 408-409.

18. Hancock, J.M. and Vogler, A.P. (2000) How slippage-derived sequences are incorporated into rRNA variable-region secondary structure: implications for phylogeny reconstruction. *Molecular Phylogenetics and Evolution*, **14**, 366-374.

19. Albà, M.M., Laskowski, R.A. and Hancock, J.M. (2002) Detecting cryptically simple protein sequences using the SIMPLE algorithm. *Bioinformatics*, **18**, 672-678.

20. Salichs, E., Ledda, A., Mularoni, L., Albà, M.M. and de la Luna, S. (2009) Genome-Wide Analysis of Histidine Repeats Reveals Their Role in the Localization of Human Proteins to the Nuclear Speckles Compartment. *PLoS Genetics*, **5**, e1000397.

21. Anan, K., Yoshida, N., Kataoka, Y., Sato, M., Ichise, H., Nasu, M. and Ueda, S. (2006) Morphological Change Caused by Loss of the Taxon-Specific Polyalanine Tract in Hoxd-13. *Molecular Biology and Evolution*, **24**, 281-287.

22. Wu, H.-T., Su, Y.-N., Hung, C.-C., Hsieh, W.-S. and Wu, K.-J. (2009) Interaction between PHOX2B and CREBBP mediates synergistic activation: mechanistic implications of PHOX2B mutants. *Human Mutation*, **30**, 655-660.

23. Enard, W., Przeworski, M., Fisher, S.E., Lai, C.S.L., Wiebe, V., Kitano, T., Monaco, A.P. and Pääbo, S. (2002) Molecular evolution of FOXP2, a gene involved in speech and language. *Nature*, **418**, 869-872.

24. Mularoni, L., Ledda, A., Toll-Riera, M. and Albà, M.M. (2010) Natural selection drives the accumulation of amino acid tandem repeats in human proteins. *Genome Research*, **20**, 745-754.

25. Caburet, S., Vaiman, D. and Veitia, R.A. (2004) A Genomic Basis for the Evolution of Vertebrate Transcription Factors Containing Amino Acid Runs. *Genetics*, **167**, 1813-1820.

26. Fondon, J.W. and Garner, H.R. (2004) Molecular origins of rapid and continuous morphological evolution. *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 18058-18063.

27. Caburet, S., Cocquet, J., Vaiman, D. and Veitia, R.A. (2005) Coding repeats and evolutionary "agility". *BioEssays*, **27**, 581-587.

28. Huntley, M.A. and Clark, A.G. (2007) Evolutionary Analysis of Amino Acid Repeats across the Genomes of 12 Drosophila Species. *Molecular Biology and Evolution*, **24**, 2598-2609.

29. El-Gebali, S., Mistry, J., Bateman, A., Eddy, S.R., Luciani, A., Potter, S.C., Qureshi, M., Richardson, L.J., Salazar, G.A., Smart, A. *et al.* (2018) The Pfam protein families database in 2019. *Nucleic Acids Research*, **47**, D427-D432.

30. Kajander, T., Cortajarena, A.L., Main, E.R.G., Mochrie, S.G.J. and Regan, L. (2005) A New Folding Paradigm for Repeat Proteins. *Journal of the American Chemical Society*, **127**, 10188-10190.

31. Kobe, B. and Kajava, A.V. (2001) The leucine-rich repeat as a protein recognition motif. *Current Opinion in Structural Biology*, **11**, 725-732.

32. D'Andrea, L.D. and Regan, L. (2003) TPR proteins: the versatile helix. *Trends in Biochemical Sciences*, **28**, 655-662.

33. Gamsjaeger, R., Liew, C.K., Loughlin, F.E., Crossley, M. and Mackay, J.P. (2007) Sticky fingers: zinc-fingers as protein-recognition motifs. *Trends in Biochemical Sciences*, **32**, 63-70.

34. Li, J., Mahajan, A. and Tsai, M.-D. (2006) Ankyrin Repeat: A Unique Motif Mediating Protein−Protein Interactions. *Biochemistry*, **45**, 15168-15178.

35. Apic, G., Gough, J. and Teichmann, S.A. (2001) Domain combinations in archaeal, eubacterial and eukaryotic proteomes. *Journal of Molecular Biology*, **310**, 311-325.

36. Ponting, C.P., Mott, R., Bork, P. and Copley, R.R. (2001) Novel Protein Domains and Repeats in Drosophila melanogaster: Insights into Structure, Function, and Evolution. *Genome Research*, **11**, 1996-2008.

37. Wright, C.F., Teichmann, S.A., Clarke, J. and Dobson, C.M. (2005) The importance of sequence diversity in the aggregation and evolution of proteins. *Nature*, **438**, 878-881.

38. Andrade, M.A., Perez-Iratxeta, C. and Ponting, C.P. (2001) Protein Repeats: Structures, Functions, and Evolution. *Journal of Structural Biology*, **134**, 117-131.

39. Weatherall, D.J. and Clegg, J.B. (1979) Recent developments in the molecular genetics of human hemoglobin. *Cell*, **16**, 467-479.

40. Patthy, L. (1999) Genome evolution and the evolution of exon-shuffling - a review. *Gene*, **238**, 103-114.

41. Han, J.-H., Batey, S., Nickson, A.A., Teichmann, S.A. and Clarke, J. (2007) The folding and evolution of multidomain proteins. *Nature Reviews Molecular Cell Biology*, **8**, 319.

42. Andrade, M.A., Ponting, C.P., Gibson, T.J. and Bork, P. (2000) Homology-based method for identification of protein repeats using statistical significance estimates11Edited by J. Thornton. *Journal of Molecular Biology*, **298**, 521-537.

43. Edgar, R.C. and Sjölander, K. (2004) COACH: profile–profile alignment of protein families using hidden Markov models. *Bioinformatics*, **20**, 1309-1318.

44. Karpenahalli, M.R., Lupas, A.N. and Söding, J. (2007) TPRpred: a tool for prediction of TPR-, PPR- and SEL1-like repeats from protein sequences. *BMC Bioinformatics*, **8**, 2.

45. Heringa, J. and Argos, P. (1993) A method to recognize distant repeats in protein sequences. *Proteins: Structure, Function, and Bioinformatics*, **17**, 391-411.

46. Szklarczyk, R. and Heringa, J. (2004) Tracking repeats using significance and transitivity. *Bioinformatics*, **20**, i311-i317.

47. Pellegrini, M., Marcotte, E.M. and Yeates, T.O. (1999) A fast algorithm for genome-wide analysis of proteins with repeated sequences. *Proteins*, **35**, 440-446.

48. Söding, J., Remmert, M. and Biegert, A. (2006) HHrep: de novo protein repeat detection and the origin of TIM barrels. *Nucleic Acids Research*, **34**, W137-W142.

49. Heger, A. and Holm, L. (2000) Rapid automatic detection and alignment of repeats in protein sequences. *Proteins: Structure, Function, and Bioinformatics*, **41**, 224-237.

50. Biegert, A. and Söding, J. (2008) De novo identification of highly diverged protein repeats by probabilistic consistency. *Bioinformatics*, **24**, 807-814.

51. Marsella, L., Sirocco, F., Trovato, A., Seno, F. and Tosatto, S.C.E. (2009) REPETITA: detection and discrimination of the periodicity of protein solenoid repeats by discrete Fourier transform. *Bioinformatics*, **25**, i289-i295.

52. Hochreiter, S., Heusel, M. and Obermayer, K. (2007) Fast model-based protein homology detection without alignment. *Bioinformatics*, **23**, 1728-1736.

53. Palidwor, G.A., Shcherbinin, S., Huska, M.R., Rasko, T., Stelzl, U., Arumughan, A., Foulle, R., Porras, P., Sanchez-Pulido, L., Wanker, E.E. *et al.* (2009) Detection of Alpha-Rod Protein Repeats Using a Neural Network and Application to Huntingtin. *PLoS Computational Biology*, **5**, e1000304.

54. Li, X. and Kahveci, T. (2006) A Novel algorithm for identifying low-complexity regions in a protein sequence. *Bioinformatics*, **22**, 2980-2987.

55. Newman, A.M. and Cooper, J.B. (2007) XSTREAM: A practical algorithm for identification and architecture modeling of tandem repeats in protein sequences. *BMC Bioinformatics*, **8**, 382.

56. Li, L., Jin, R., Kok, P.-L. and Wan, H. (2004) Pseudo-periodic partitions of biological sequences. *Bioinformatics*, **20**, 295-306.

57. Liu, X. and Wang, L. (2006) Finding the region of pseudo-periodic tandem repeats in biological sequences. *Algorithms for Molecular Biology*, **1**, 2.

58. Adebiyi, E.F., Jiang, T. and Kaufmann, M. (2001) An efficient algorithm for finding short approximate non-tandem repeats. *Bioinformatics*, **17**, S5-S12.

59. Fankhauser, N., Nguyen-Ha, T.-M., Adler, J. and Mäser, P. (2007) Surface antigens and potential virulence factors from parasites detected by comparative genomics of perfect amino acid repeats. *Proteome Science*, **5**, 20.

60. Gusfield, D. and Stoye, J. (2004) Linear time algorithms for finding and representing all the tandem repeats in a string. *Journal of Computer and System Sciences*, **69**, 525-546.

61. Needleman, S.B. and Wunsch, C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, **48**, 443-453.

62. Smith, T.F. and Waterman, M.S. (1981) Identification of common molecular subsequences. *Journal of Molecular Biology*, **147**, 195-197.

63. Dayhoff, M. (1978) A model of evolutionary change in proteins. *Atlas of Protein Sequence & Structure*, **5**, 345-352.

64. Henikoff, S. and Henikoff, J.G. (1992) Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences of the United States of America*, **89**, 10915-10919.

65. Atchley, W.R., Zhao, J., Fernandes, A.D. and Drüke, T. (2005) Solving the protein sequence metric problem. *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 6395-6400.

66. Karlin, S. and Altschul, S.F. (1990) Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proceedings of the National Academy of Sciences of the United States of America*, **87**, 2264-2268.

67. Karlin, S. and Altschul, S.F. (1993) Applications and statistics for multiple high-scoring segments in molecular sequences. *Proceedings of the National Academy of Sciences of the United States of America*, **90**, 5873-5877.

68. Rabiner, L.R. (1990) In Waibel, A. and Lee, K.-F. (eds.), *Readings in Speech Recognition*. Morgan Kaufmann, San Francisco, pp. 267-296.

69. Consortium, T.U. (2018) UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Research*, **47**, D506-D515.

70. Alexandrov, V.N., Van Albada, G.D., Sloot, P.M. and Dongarra, J.J. (2006) *Computational Science-ICCS 2006: 6th International Conference, Reading, UK, May 28-31, 2006, Proceedings*. Springer Science & Business Media.

71. George, R.A. and Heringa, J. (2000) The REPRO server: finding protein internal sequence repeats through the Web. *Trends in Biochemical Sciences*, **25**, 515-517.

72. Romein, J.W., Heringa, J. and Bal, H.E. (2003), *Proceedings of the 2003 ACM/IEEE conference on Supercomputing*. Association for Computing Machinery, Phoenix, AZ, USA, pp. 20.

73. Söding, J. (2004) Protein homology detection by HMM–HMM comparison. *Bioinformatics*, **21**, 951-960.

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

75. McLachlan, A.D. (1977) Analysis of periodic patterns in amino acid sequences: Collagen. *Biopolymers*, **16**, 1271-1297.

76. Liò, P. (2003) Wavelets in bioinformatics and computational biology: state of art and perspectives. *Bioinformatics*, **19**, 2-9.

77. Sweldens, W. (1996) Wavelets: What next? *Proceedings of the IEEE*, **84**, 680-685.

78. Vo, A., Nguyen, N. and Huang, H. (2010) Solenoid and non-solenoid protein recognition using stationary wavelet packet transform. *Bioinformatics*, **26**, i467-i473.

79. Bishop, C.M. (1995) *Neural networks for pattern recognition*. Oxford university press.

80. Wootton, J.C. and Federhen, S. (1996), *Methods in Enzymology*. Academic Press, Vol. 266, pp. 554-571.

81. Wan, H., Li, L., Federhen, S. and Wootton, J.C. (2003) Discovering simple regions in biological sequences associated with scoring schemes. *Journal of Computational Biology*, **10**, 171-185.

82. Shin, S.W. and Kim, S.M. (2004) A new algorithm for detecting low-complexity regions in protein sequences. *Bioinformatics*, **21**, 160-170.

83. Shannon, C.E. and Weaver, W. (1962) *The Mathematical Theory of Communication*. University of Illinois Press, Urbana.

84. Coronado, J.E., Attie, O., Epstein, S.L., Qiu, W.-G. and Lipke, P.N. (2006) Composition-Modified Matrices Improve Identification of Homologs of *Saccharomyces cerevisiae* Low-Complexity Glycoproteins. *Eukaryotic Cell*, **5**, 628-637.

85. Hamming, R.W. (1950) Error Detecting and Error Correcting Codes. *Bell System Technical Journal*, **29**, 147-160.

86. Levenshtein, V.I. (1966), *Soviet Physics Doklady*, Vol. 10, pp. 707-710.

87. Katti, M.V., Sami-Subbu, R., Ranjekar, P.K. and Gupta, V.S. (2000) Amino acid repeat patterns in protein sequences: Their diversity and structural-functional implications. *Protein Science*, **9**, 1203-1209.

88. Groult, R., Léonard, M. and Mouchard, L. (2004) Speeding up the detection of evolutive tandem repeats. *Theoretical Computer Science*, **310**, 309-328.

89. Hammock, E.A.D. and Young, L.J. (2005) Microsatellite Instability Generates Diversity in Brain and Sociobehavioral Traits. *Science*, **308**, 1630-1634.

90. Bannen, R.M., Bingman, C.A. and Phillips, G.N. (2007) Effect of low-complexity regions on protein structure determination. *Journal of Structural and Functional Genomics*, **8**, 217-226.

91. Karlin, S. (1995) Statistical significance of sequence patterns in proteins. *Current Opinion in Structural Biology*, **5**, 360-371.

92. Long, M., Betrán, E., Thornton, K. and Wang, W. (2003) The origin of new genes: glimpses from the young and old. *Nature Reviews Genetics*, **4**, 865-875.

93. Ohno, S. (2013) *Evolution by Gene Duplication*. Springer Science & Business Media.

94. Crow, K.D. and Wagner, G.P. (2005) What Is the Role of Genome Duplication in the Evolution of Complexity and Diversity? *Molecular Biology and Evolution*, **23**, 887-892.

95. Verstrepen, K.J., Jansen, A., Lewitter, F. and Fink, G.R. (2005) Intragenic tandem repeats generate functional variability. *Nature Genetics*, **37**, 986-990.

96. Gibbons, J.G. and Rokas, A. (2008) Comparative and Functional Characterization of Intragenic Tandem Repeats in 10 Aspergillus Genomes. *Molecular Biology and Evolution*, **26**, 591-602.

97. Rorick, M.M. and Wagner, G.P. (2010) The Origin of Conserved Protein Domains and Amino Acid Repeats Via Adaptive Competition for Control Over Amino Acid Residues. *Journal of Molecular Evolution*, **70**, 29-43.

98. Haerty, W. and Golding, G.B. (2010) Genome-wide evidence for selection acting on single amino acid repeats. *Genome Research*, **20**, 755-760.

99. Haerty, W. and Golding, G.B. (2010) Low-complexity sequences and single amino acid repeats: not just "junk" peptide sequences. *Genome*, **53**, 753-762.

100. O'Leary, N.A., Wright, M.W., Brister, J.R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D. *et al.* (2015) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, **44**, D733-D745.

101. Kuzniar, A., Lin, K., He, Y., Nijveen, H., Pongor, S. and Leunissen, J.A.M. (2009) ProGMap: an integrated annotation resource for protein orthology. *Nucleic Acids Research*, **37**, W428-W434.

CHAPTER 3

3

ProRepeat: an integrated repository for studying amino acid tandem repeats in proteins

Hong Luo, Ke Lin, Audrey David, Harm Nijveen,
Jack A. M. Leunissen

Based on: Nucleic Acids Research, 2012;40;D394-399.

Abstract

ProRepeat (http://prorepeat.bioinformatics.nl/) is an integrated curated repository and analysis platform for in-depth research on the biological characteristics of amino acid tandem repeats. ProRepeat collects repeats from all proteins included in the UniProt knowledgebase, together with 85 completely sequenced eukaryotic proteomes contained within the RefSeq collection. It contains non-redundant perfect tandem repeats, approximate tandem repeats and simple, low-complexity sequences, covering the majority of the amino acid tandem repeat patterns found in proteins. The ProRepeat web interface allows querying the repeat database using repeat characteristics like repeat unit and length, number of repetitions of the repeat unit and position of the repeat in the protein. Users can also search for repeats by the characteristics of repeat containing proteins, such as entry ID, protein description, sequence length, gene name and taxon. ProRepeat offers powerful analysis tools for finding biological interesting properties of repeats, such as the strong position bias of leucine repeats in the N-terminus of eukaryotic protein sequences, the differences of repeat abundance among proteomes, the functional classification of repeat containing proteins and GC content constrains of repeats' corresponding codons.

3.1 Introduction

Amino acid tandem repeats, as one of the most prevalent patterns in protein sequences, have inspired the interests of researchers for many years in terms of their pathological, functional and evolutionary roles. According to the patterns of units, repeats in proteins can be generally classified into several categories.

Single amino acid repeats (SAARs), also known as homo peptides, have the simplest repeat unit. Some of the SAARs have been extensively studied as they are involved in numbers of human neurodegenerative diseases, such as those with variable polyglutamines (polyQ) and polyalanines (polyA) (1). Since they are important modulation factors on protein–protein interactions (2,3), the insertions, deletions, substitutions, as well as growing or shrinking of the repeats result in either loss-of-function or gain of abnormal function (4,5) by altering the conformation of protein tertiary structures. As for other types of SAARs, leucine runs are mainly located in the N-terminus of eukaryotic proteins, which are presumed to be involved in the signal peptide (6). Higher frequency of proline repeats in mammalian proteomes is considered to significantly contribute to network evolution (7). In addition, histidine repeats play a crucial role in the localization of human proteins to the nuclear speckle compartment (8).

Amino acid tandem repeats with complex unit patterns have also been studied frequently. Different from SAARs, most of them are comparatively conserved in their structure. Well-known patterns include the leucine rich repeats (LRRs) that commonly act as the structural framework for the formation of protein-protein interactions (9), the ankyrin repeats that contain the binding site for the huge titin proteins that are involved in muscle ultrastructure and elasticity (10,11), and the polyubiquitins that are synthesized as repetitive polyproteins (12).

Although the biological significance of particular amino acid tandem repeats has been demonstrated continually during the past years in several model organisms, no convincing conclusions can be drawn until now. Questions are posed on several aspects: Is the role of similar repeat patterns coherent in different proteomes across the kingdoms of life? Could the functional and evolutionary roles of certain repeats correspond to their particular characteristics, such as position bias, GC content constraints and codon usage? How could the conserved functions of particular repeats have evolved by natural selection? Why are repeats so common in protein sequences even under the scenario that their instable characteristics often cause disorder and disease (5,13,14)? And what are structural and sequence-based strategies to prevent repeats from possible aggregation (15,16)?

The dilemma of contradicting explanations of the role of repeats is partly because of the lack of repositories for large-scale investigation and comparison of repeats among the variety of proteomes across different kingdoms. Several databases of amino acid repeats were constructed during the recent decade. Unfortunately, some of these databases are no longer accessible or functional, such as COPASAAR (17), RepSeq (18) and ProtRepeatDB (19). As for the remaining ones, TRIPS gathered repeats generated from a very old version of SwissProt (year 1999) (20), RCPdb offers the codon usage bias data of homopeptides (SAARs) of 13 completely sequenced eukaryotic species (21), and the PolyQ database collects the sequences of all human proteins containing runs of seven or more glutamine residues (22).

To change the incompatible situation between the rapid increase of protein sequence data and the lack of a large scale, well-annotated protein repeat repository, we have constructed an online database of protein repeat sequences (ProRepeat, http://prorepeat.bioinformatics.nl/). ProRepeat recruits both perfect and approximate tandem repeats from all taxa of UniProtKB (23) and supplied by 85 complete sequenced and well annotated eukaryotic proteomes. ProRepeat also gathers the corresponding nucleotide sequences of the repeat fragments for the purpose of codon usage analysis. The latest update of ProRepeat is based on the datasets of UniProtKB release 2011_05 and RefSeq (24) release 40. An easy-to-use web interface was designed for users to query the database, and to perform statistical analyses on the query results. We believe that ProRepeat provides the user community with a useful resource for the exploration of function and evolution of protein repeats.

3.2 Repeat detection and dataset generation

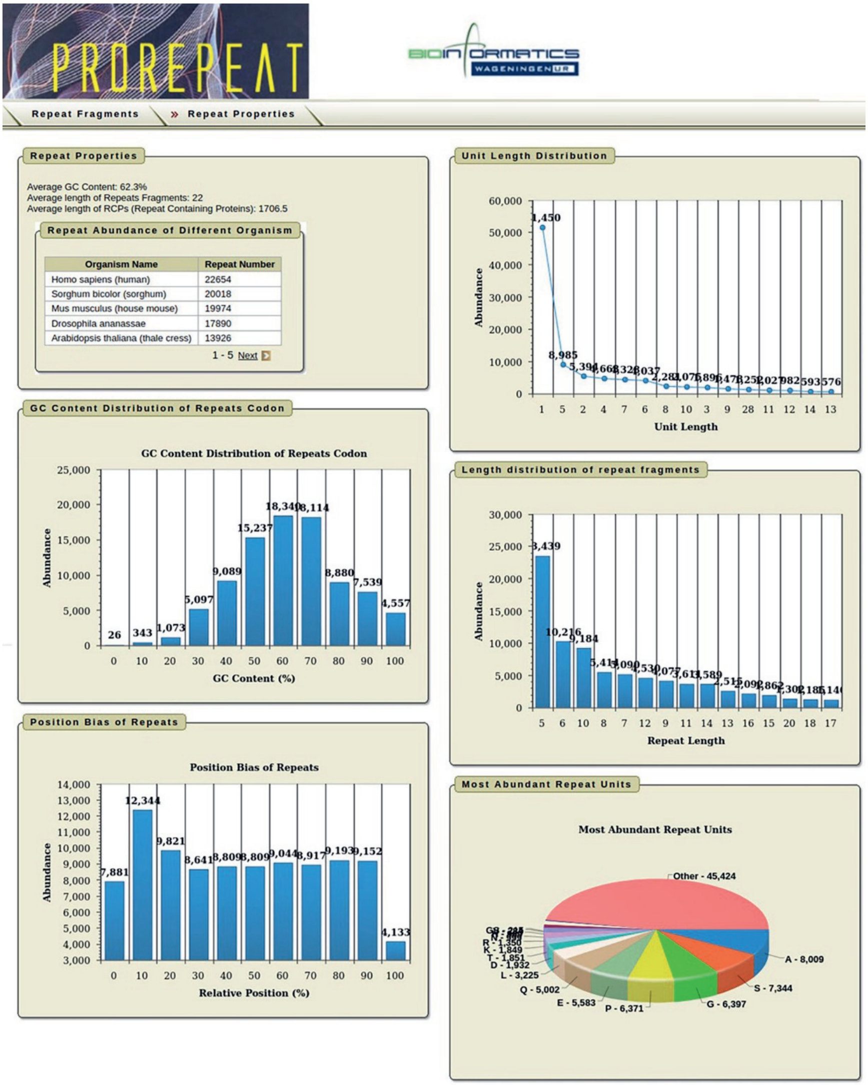
We collect three types of repeat patterns including perfect tandem repeats (PTRs), approximate tandem repeats (ATRs) and simple sequences (SSs) in proteins. The PTRs were detected using an in house developed C/C++ procedure we implemented based on the suffix tree algorithm which identifies all perfect tandem repeats in a protein sequence (25), the ATRs were detected by XSTREAM (26) and the SSs were detected by SIMPLE (27). Following the definition of statistically significant repeat runs in protein sequences (28), we used cutoffs of five, four, three and two of the repeat unit repetitions to identify mono-amino, di-amino, tri-amino and all other repeats, respectively.

It is possible that different algorithms identify repeats with the same unit and overlapped position in the same protein. To remove this redundancy, we developed a PL/SQL procedure which distinguishes between unique and overlapping repeats. The repeats datasets were merged followed by a sorting step based on the identifier of repeat containing proteins (RCPs), repeat unit and the position of the repeat. The repeats with the same unit and overlapping position within the same protein were merged into a single fragment. If the begin and end positions of these repeats were also the same, only one of them was retained as they were actually the same repeat identified by different algorithms. We classified perfect and approximate repeats separately and marked them as such in the database, so that the user can search them individually.

The repeat datasets were generated based on the protein entries collected in UniProtKB release 2011_05. For the convenience of comparative analysis, we also generated the repeat datasets of the completely sequenced eukaryotic proteomes based on RefSeq release 40. For the selection of completely sequenced eukaryotic proteomes, we obtained the list of the complete published eukaryotic organisms from the genomes online database (GOLD) (29). For each organism, we compared the number of ORFs given by GOLD with the number of proteins collected by RefSeq. If the two numbers were approximately consistent, i.e., the difference was <5%, we considered the proteome collected by RefSeq as complete and retrieved repeats from it. Thus, ProRepeat contains repeats from 85 complete sequenced eukaryotic proteomes including 14 vertebrates, 8 plants, 22 fungi, 12 insects and 29 other organisms. The gene ontology cross references of RCPs were generated based on GOA (30), and RefSeq annotations for gene ontology. The corresponding nucleotide sequences of the repeat fragments were obtained via EMBL and RefSeq cross-references within each UniProtKB and RefSeq protein entry, respectively.

3.3 The web interface

The ProRepeat database can be accessed using an intuitive web interface. An Introduction page provides information about the types of repeats that the database contains, the tools that were used to create the database, and the background of functional studies of repeats. The Statistics page lists several characteristics of the database, like the abundance of different repeat types across the different life kingdoms. The Help page offers practical examples to help users find interesting repeats and perform online analyses. On the Query page, users can search for repeats in one or more species, using annotations of RCPs including entry ID, protein description and gene name. Users can also specify the repeat unit, unit length, repeat sequence length, number of units and position of repeat in the protein sequence. For all the repeat units,



Repeat Properties

Average GC Content: 62.3%
Average length of Repeats Fragments: 22
Average length of RCPs (Repeat Containing Proteins): 1706.5

Repeat Abundance of Different Organism

Organism Name	Repeat Number
Homo sapiens (human)	22654
Sorghum bicolor (sorghum)	20018
Mus musculus (house mouse)	19974
Drosophila ananassae	17890
Arabidopsis thaliana (thale cress)	13926

1 - 5 Next

Figure 3.1 **The statistical analysis result of repeat properties for PTRs of all taxa in UniProt with the 'Isomorphic Search' option on, and the default evidence at protein level.**

ProRepeat offers two additional options. For example, the repeat unit of two repeat fragments DEDEDEDE and EDEDEDED could be identified as DE and ED, and defined as isomorphic repeats. By switching on the 'Isomorphic Unit Search' option, users can obtain all cyclic permutations of this repeat pattern. With the 'ProSite SyntaxSearch' option, users can specify a regular expression as search pattern used by the ProSite database. In addition, ProRepeat classifies the repeats as PTRs and ATRs defined by the similarities of repeat units. Users can choose either 'Perfect Tandem Repeats', 'Approximate Tandem Repeats' or 'All Repeats' (both PTRs and ATRs).

Users can query individual species of 85 eukaryotic complete proteomes obtained from RefSeq through their taxonomic names, or from broader taxonomic ranges collected from the UniProt Knowledgebase. The query results are displayed as interactive web pages in tabular format. Columns contain information about the position and length of the repeat, as well as the corresponding protein. Clicking on a repeat brings up a page with the corresponding DNA sequence and the codon usage pattern. Users can save the query results in Microsoft Excel format, or perform a secondary search using the query results. ProRepeat provides users with an online tool to perform statistical analyses on the query results. For example, the query results of PTRs from all species in UniProt at protein level can be analyzed to show various properties including the repeat abundance in different organisms, the gene ontology annotation of the RCPs, the position bias of repeats, the distributions of GC content of repeat codon, the unit length and repeat fragment length (Figure 3.1).

3.4 Preliminary Analysis of Protein Repeats

Based on UniProtKB datasets, ProRepeat gathers ~3.75 million repeat fragments contained in 2 million RCPs from 0.1 million organisms. The distribution of repeats over eukaryota, bacteria, archaea and viruses are shown in Table 3.1. The relative repeat abundance normalized by the number of proteins of the different kingdoms indicates that eukaryotic proteins are four times more likely to have tandem repeats than prokaryotic proteins, and the possibility of having tandem repeats in viruses and prokaryotes is similar. This supports the idea that large numbers of protein repeats arose after the divergence of prokaryota and eukaryota (31).

Table 3.1 **Repeat abundance in four kingdoms**

Kingdom	Repeat Number		Repeat abundance[a]	Protein abundance[b]	Relative
	PTR	ATR	(%)	(%)	abundance[c]
Eukaryota	1,163,368	1,195,655	63.10	27.2	2.32
Bacteria	498,071	705,575	32.20	63.9	0.50
Archaea	12,584	18,631	0.85	1.8	0.47
Viruses	75,109	68,821	3.85	6.9	0.56

[a]Percentage of repeats in four kingdoms, [b]Percentage of proteins in four kingdoms based on UniProtKB (0.2% unclassified entries are not listed), [c]Percentage of protein abundance divided by percentage of repeat abundance.

There is a long-standing debate about the roles of repeats in proteins. Some early viewpoints ascribe large amounts of SSs to ‘junk protein’ (32) as few of them have identified stable tertiary structures (33) and are thought to be non-functional. However, more and more evidences show that they are not just ‘junk’ peptide sequences (34) and might have particular function and structure (35). Important subsets of SSs, in particular cryptic and identical SAARs, have been reported to be actively evolving (36-38). As a result, the evolutionary footprint and functional implication of repeats which are being modulated by selection could be inferred from their properties. For example, in Drosophila and Arabidopsis, the RCPs are mostly involved in gene regulation, signaling and developmental processes, but significantly under-represented in the process of DNA recombination and DNA replication. In addition, the positional distribution of repeats in proteins of Drosophila and Arabidopsis is also non-random (39,40).

Using ProRepeat, we made a comparison of repeat properties including repeat length, RCPs length, repeat position and repeat codon usage in model organisms across different kingdoms (Table 3.2). In general, glutamic acid (E), serine (S), glutamine (Q), proline (P), alanine (A) and leucine (L) are widely used by SAARs in all taxa, but the pattern varies between different taxa. For example, polyL and polyP are preferred by prokaryotes and eukaryotes, respectively; *Drosophila melanogaster* uses polyQ more frequently than most of the other organisms; polyE is extremely abundant in Hepatitis delta virus, and for human immunodeficiency virus, although polyE has the highest frequency, arginine (R) is actually the most commonly used amino acid (near 80%) when approximate SAARs are combined. When looking at the N-terminal perfect SAARs, polyL is the most popular especially in eukaryotes and bacteria, in which they play functional roles, for instance, in signal peptides (6).

Table 3.2 **Repeat properties in representative species**

Species	Most abundant SAARs(%)	N/C SAARs	GC1	GC2	L1	L2
HIV	E(45.3), A(27.0), N(8.6)	SA/INP	42.0	41.9	462	10.7
HDV	E(99.6), P(0.4)	Na/Na	Na	41.5	113	5.2
E. coli	L(32.0), A(29.5), G(9.4)	LAT/GAV	50.0	58.0	765	18.7
B. subtilis	A(23.8), L(19.8), S(19.8)	LKA/KSG	43.5	48.5	481	15.0
A. fulgidus	E(22.0), V,(18.0), L(18.0)	ER/KTL	48.6	51.9	389	10.1
M. jannaschii	E(25.9), K(22.2), L(11.1)	ILE/KGR	31.0	31.7	412	10.8
S. cerevisiae	S(24.0), Q(18.7), N(11.7)	SQN/KDQ	38.1	44.3	759	18.5
A. thaliana	S(27.2), G(12.3), P(11.5)	SLE/GES	36.0	50.9	812	16.0
C. elegans	S(14.9), T(13.8), Q(13.6)	SLQ/QGS	35.0	51.7	1103	25.0
D. melanogaster	Q(31.9), A(15.2), S(11.3)	QAS/QAS	41.0	61.3	1338	15.6
D. rerio	S(21.4), E(17.6), P(13.1)	LAG/ESK	37.6	54.4	1286	37.9
G. gallus	E(17.7), P(15.1), S(13.4)	LAG/ESK	50.0	62.5	1099	20.8
M. musculus	E(19.2), P(14.6), A(11.6)	LAG/EPA	41.7	60.9	1304	26.6
H. sapiens	E(16.0), P(16.0), A(14.3)	LAG/ESP	40.9	63.0	1390	31.2

N/C SAARs, most abundant N- and C-terminal SAARs corresponding to 5% and 95% of RCP length, respectively; the middle point of the repeat fragments is defined as the position of repeats; GC1, genomic GC content; GC2, average GC content of repeat codon; L1, average length of RCPs; L2, average length of repeat fragments. Na: not available.

The positive correlation between GC content and genes rich in coding repeats has also been noticed in recent years (41-43), which suggests that the formation and evolution of coding repeats is constrained by sequence composition at the genome level. Other studies also indicate that the length of the coding sequence is directly proportional to higher GC content (44) as the stop codon has a bias toward A and T, thus the shorter the sequence the higher the AT bias (45). To investigate this, we used ProRepeat to compute the average GC content of repeat codons across taxa. The result shows that the average GC content of repeat codons is much higher than the genome GC content in nearly all species (Table 3.2). This is especially true in eukaryotes which have higher repeat abundance than prokaryotes and viruses. On the other hand, although the average length of RCPs is much greater than the average length of proteins in different kingdoms, i.e. 361 AA in Eukaryotes, 267 AA in Bacteria and 247 AA in Archaea, respectively (46), the relationship between GC content and the length of RCPs is not very strong.

3.5 Future Directions

To cope with the fast development of genome sequencing and annotating, we have been keeping ProRepeat updated to the latest version of the protein databases UniProtKB and RefSeq protein. Furthermore, as our repeat integrating strategy merges different datasets generated by different algorithms, we will integrate more repeat patterns into ProRepeat detected by more algorithms in the future.

Comparing specific repeat fragments among orthologous RCPs is a widely used strategy to discover their potential evolutionary and functional roles. The former analysis across Drosophila, rodents and primates (13,37,39,42) shows its reliability. As ProRepeat contains data over a broad taxonomy range, it may serve as an excellent platform to perform orthologous analysis on repeats. To meet such requirements, we are currently integrating ProRepeat with ProGMap - the integrated annotation resource for protein orthology (47) we developed earlier. With this setup users can compare repeats among orthologous RCPs in ProRepeat.

Acknowledgements

The authors like to thank Wilfried de Jong for starting the project together with J.A.M.L., Maarten van den Bosch for providing the prototype implementation of the database, Guido Kappé for programming the suffix tree program, Arnold Kuzniar for fruitful discussions and Jingchu Luo for critically reading the manuscript.

References

1. Orr, H.T. and Zoghbi, H.Y. (2007) Trinucleotide Repeat Disorders. *Annual Review of Neuroscience*, **30**, 575-621.

2. Buchanan, G., Yang, M., Cheong, A., Harris, J.M., Irvine, R.A., Lambert, P.F., Moore, N.L., Raynor, M., Neufing, P.J., Coetzee, G.A. *et al.* (2004) Structural and functional consequences of glutamine tract variation in the androgen receptor. *Human molecular genetics*, **13**, 1677-1692.

3. Brown, L., Paraso, M., Arkell, R. and Brown, S. (2005) In vitro analysis of partial loss-of-function ZIC2 mutations in holoprosencephaly: alanine tract expansion modulates DNA binding and transactivation. *Human molecular genetics*, **14**, 411-420.

4. Brown, L.Y. and Brown, S.A. (2004) Alanine tracts: the expanding story of human illness and trinucleotide repeats. *Trends in Genetics*, **20**, 51-58.

5. Gatchel, J.R. and Zoghbi, H.Y. (2005) Diseases of Unstable Repeat Expansion: Mechanisms and Common Principles. *Nature Reviews Genetics*, **6**, 743-755.

6. Łabaj, P.P., Leparc, G.G., Bardet, A.F., Kreil, G. and Kreil, D.P. (2010) Single amino acid repeats in signal peptides. *The FEBS Journal*, **277**, 3147-3157.

7. Hancock, J.M. and Simon, M. (2005) Simple sequence repeats in proteins and their significance for network evolution. *Gene*, **345**, 113-118.

8. Salichs, E., Ledda, A., Mularoni, L., Albà, M.M. and de la Luna, S. (2009) Genome-Wide Analysis of Histidine Repeats Reveals Their Role in the Localization of Human Proteins to the Nuclear Speckles Compartment. *PLoS Genetics*, **5**, e1000397.

9. Kobe, B. and Kajava, A.V. (2001) The leucine-rich repeat as a protein recognition motif. *Current Opinion in Structural Biology*, **11**, 725-732.

10. Miller, M.K., Bang, M.-L., Witt, C.C., Labeit, D., Trombitas, C., Watanabe, K., Granzier, H., McElhinny, A.S., Gregorio, C.C. and Labeit, S. (2003) The Muscle Ankyrin Repeat Proteins: CARP, ankrd2/Arpp and DARP as a Family of Titin Filament-based Stress Response Molecules. *Journal of Molecular Biology*, **333**, 951-964.

11. Labeit, S. and Kolmerer, B. (1995) Titins: Giant Proteins in Charge of Muscle Ultrastructure and Elasticity. *Science*, **270**, 293-296.

12. Callis, J., Carpenter, T., Sun, C.W. and Vierstra, R.D. (1995) Structure and evolution of genes encoding polyubiquitin and ubiquitin-like proteins in Arabidopsis thaliana ecotype Columbia. *Genetics*, **139**, 921-939.

13. Simon, M. and Hancock, J.M. (2009) Tandem and cryptic amino acid repeats accumulate in disordered regions of proteins. *Genome Biology*, **10**, R59.

14. Karlin, S., Brocchieri, L., Bergman, A., Mrázek, J. and Gentles, A.J. (2002) Amino acid runs in eukaryotic proteomes and disease associations. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 333-338.

15. Monsellier, E. and Chiti, F. (2007) Prevention of amyloid-like aggregation as a driving force of protein evolution. *EMBO reports*, **8**, 737-742.

16. Han, J.-H., Batey, S., Nickson, A.A., Teichmann, S.A. and Clarke, J. (2007) The folding and evolution of multidomain proteins. *Nature Reviews Molecular Cell Biology*, **8**, 319-330.

17. Depledge, D.P. and Dalby, A.R. (2005) COPASAAR – A database for proteomic analysis of single amino acid repeats. *BMC Bioinformatics*, **6**, 196.

18. Depledge, D.P., Lower, R.P.J. and Smith, D.F. (2007) RepSeq – A database of amino acid repeats present in lower eukaryotic pathogens. *BMC Bioinformatics*, **8**, 122.

19. Kalita, M.K., Ramasamy, G., Duraisamy, S., Chauhan, V.S. and Gupta, D. (2006) ProtRepeatsDB: a database of amino acid repeats in genomes. *BMC Bioinformatics*, **7**, 336.

20. Katti, M.V., Sami-Subbu, R., Ranjekar, P.K. and Gupta, V.S. (2000) Amino acid repeat patterns in protein sequences: Their diversity and structural-functional implications. *Protein Science*, **9**, 1203-1209.

21. Faux, N.G., Huttley, G.A., Mahmood, K., Webb, G.I., Garcia de la Banda, M. and Whisstock, J.C. (2007) RCPdb: An evolutionary classification and codon usage database for repeat-containing proteins. *Genome Research*, **17**, 1118-1127.

22. Robertson, A.L., Bate, M.A., Androulakis, S.G., Bottomley, S.P. and Buckle, A.M. (2010) PolyQ: a database describing the sequence and domain context of polyglutamine repeats in proteins. *Nucleic Acids Research*, **39**, D272-D276.

23. Consortium, T.U. (2018) UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Research*, **47**, D506-D515.

24. O'Leary, N.A., Wright, M.W., Brister, J.R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D. *et al.* (2015) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, **44**, D733-D745.

25. Gusfield, D. and Stoye, J. (2004) Linear time algorithms for finding and representing all the tandem repeats in a string. *Journal of Computer and System Sciences*, **69**, 525-546.

26. Newman, A.M. and Cooper, J.B. (2007) XSTREAM: A practical algorithm for identification and architecture modeling of tandem repeats in protein sequences. *BMC Bioinformatics*, **8**, 382.

27. Albà, M.M., Laskowski, R.A. and Hancock, J.M. (2002) Detecting cryptically simple protein sequences using the SIMPLE algorithm. *Bioinformatics*, **18**, 672-678.

28. Karlin, S. (1995) Statistical significance of sequence patterns in proteins. *Current Opinion in Structural Biology*, **5**, 360-371.

29. Mukherjee, S., Stamatis, D., Bertsch, J., Ovchinnikova, G., Katta, H.Y., Mojica, A., Chen, I.-M.A., Kyrpides, N.C. and Reddy, T. (2018) Genomes OnLine database (GOLD) v.7: updates and new features. *Nucleic Acids Research*, **47**, D649-D659.

30. Huntley, R.P., Sawford, T., Mutowo-Meullenet, P., Shypitsyna, A., Bonilla, C., Martin, M.J. and O'Donovan, C. (2014) The GOA database: Gene Ontology annotation updates for 2015. *Nucleic Acids Research*, **43**, D1057-D1063.

31. Marcotte, E.M., Pellegrini, M., Yeates, T.O. and Eisenberg, D. (1999) A census of protein repeats11Edited by J. M. Thornton. *Journal of Molecular Biology*, **293**, 151-160.

32. Lovell, S.C. (2003) Are non-functional, unfolded proteins ('junk proteins') common in the genome? *FEBS Letters*, **554**, 237-239.

33. Huntley, M.A. and Golding, G.B. (2002) Simple sequences are rare in the Protein Data Bank. *Proteins: Structure, Function, and Bioinformatics*, **48**, 134-140.

34. Haerty, W. and Golding, G.B. (2010) Low-complexity sequences and single amino acid repeats: not just "junk" peptide sequences. *Genome*, **53**, 753-762.

35. Dunker, A.K., Silman, I., Uversky, V.N. and Sussman, J.L. (2008) Function and structure of inherently disordered proteins. *Current Opinion in Structural Biology*, **18**, 756-764.

36. Haerty, W. and Golding, G.B. (2010) Genome-wide evidence for selection acting on single amino acid repeats. *Genome Research*, **20**, 755-760.

37. Mularoni, L., Ledda, A., Toll-Riera, M. and Albà, M.M. (2010) Natural selection drives the accumulation of amino acid tandem repeats in human proteins. *Genome Research*, **20**, 745-754.

38. Huntley, M.A. and Golding, G.B. (2006) Selection and Slippage Creating Serine Homopolymers. *Molecular Biology and Evolution*, **23**, 2017-2025.

39. Huntley, M.A. and Clark, A.G. (2007) Evolutionary Analysis of Amino Acid Repeats across the Genomes of 12 Drosophila Species. *Molecular Biology and Evolution*, **24**, 2598-2609.

40. Zhang, L., Yu, S., Cao, Y., Wang, J., Zuo, K., Qin, J. and Tang, K. (2006) Distributional gradient of amino acid repeats in plant proteins. *Genome*, **49**, 900-905.

41. Nakachi, Y., Hayakawa, T., Oota, H., Sumiyama, K., Wang, L. and Ueda, S. (1997) Nucleotide compositional constraints on genomes generate alanine-, glycine-, and proline-rich structures in transcription factors. *Molecular Biology and Evolution*, **14**, 1042-1049.

42. Albà, M.M. and Guigó, R. (2004) Comparative Analysis of Amino Acid Repeats in Rodents and Humans. *Genome Research*, **14**, 549-554.

43. Cocquet, J., De Baere, E., Caburet, S. and Veitia, R.A. (2003) Compositional Biases and Polyalanine Runs in Humans. *Genetics*, **165**, 1613-1617.

44. Pozzoli, U., Menozzi, G., Fumagalli, M., Cereda, M., Comi, G.P., Cagliani, R., Bresolin, N. and Sironi, M. (2008) Both selective and neutral processes drive GC content evolution in the human genome. *BMC Evolutionary Biology*, **8**, 99.

45. Wuitschick, J.D. and Karrer, K.M. (1999) Analysis of genomic G + C content, codon usage, initiator codon context and translation termination sites in Tetrahymena thermophila. *Journa of Eukaryptic Microbiology*, **46**, 239-247.

46. Brocchieri, L. and Karlin, S. (2005) Protein length in eukaryotic and prokaryotic proteomes. *Nucleic Acids Research*, **33**, 3390-3400.

47. Kuzniar, A., Lin, K., He, Y., Nijveen, H., Pongor, S. and Leunissen, J.A.M. (2009) ProGMap: an integrated annotation resource for protein orthology. *Nucleic Acids Research*, **37**, W428-W434.

CHAPTER 4

4

Next-generation sequencing technology for genetics and genomics of sorghum

Hong Luo, Anne Mocoeur, Hai-Chun Jing

Based on: Genetics, Genomics and Breeding of Sorghum. CRC Press (2014), pp. 226-250.

Abstract

The invention and application of Next-Generation Sequencing (NGS) technologies have revolutionized the study of genetics and genomics. Much research that would not even be considered a decade before is now being executed in many laboratories as routine. In this chapter, we first introduce the development and consequences of NGS technology, including a comparison of the available NGS platforms and a summary of the rapidly accumulated genome sequencing data. Then we provide a perspective on how NGS-based methodologies have been applied to examine the genomes and proteomes of sorghum populations. Finally, we list NGS-based applications for further genetic improvement and breeding of sorghum.

4.1 Introduction

Decoding the genome of a species has become one of the most challenging tasks for biologists and bioinformaticians over the recent decades. It is essential to know an organism's genome sequence to understand all biological implications following the central dogma. The early genome sequencing projects mainly targeted viruses and bacteria (1,2) with small and compact genomes, as the experimental and computational sequence assembly procedures in those days were costly and time-consuming. With the development and optimization of the Sanger sequencing method (3) at the beginning of the 1990s, more and more eukaryotic organisms with larger and more complex genomes have been successfully sequenced. The first batch of sequenced organisms, *Saccharomyces cerevisiae*, *Caenorhabditis elegans*, *Drosophila melanogaster* and *Arabidopsis thaliana*, are regarded as model organisms. Knowing their genome was particularly relevant because these organisms are usually amenable to experimental manipulations such as cultivation, transformation, and inbreeding.

During this decade, exciting discoveries have been published based on the study of genomes. Sequencing the genomes of more essential and particular species, such as human beings, became possible. With the efforts of numerous scientists, the Human Genome Project (HGP), one of the most significant international collaborative scientific projects in human history, spent at least 30 years and three billion US$. The success of the HGP started a new era of genome sequencing and computational biology. It determined the three billion nucleic acid bases and the ~19,000 genes (4) in the human genome and drastically improved bioinformatics tools and algorithms for data analysis.

Furthermore, the competition and cooperation through the HGP project also stimulated the rapid progress of industrial sequencing technologies. As a result, several massively parallel DNA sequencing platforms were marketed soon after the HGP finished, employing a similar sequencing-by-synthesis strategy with different single-molecule amplification protocols (5). The new platforms, which are commonly called NGS machines, included 454, Illumina, and SOLID. They significantly increased the sequencing throughput with a dramatic decrease in instrumental and labor expenses. With the wide use of NGS technologies, the average cost of sequencing declined from approximately $1 per base to less than $1 per million bases.

Consequently, more and more laboratories choose to employ NGS-based methods for research, in addition to the traditional low throughput protocols. At the same time, the large data volumes generated by NGS platforms brought many new problems in terms of quality control, manipulation, storage, analysis and sharing of data. Many bioinformatics algorithms and tools have been developed to process the NGS data. The combination of these opportunities and challenges led to the development of the interdisciplinary research field.

4.2 The Rapid increase of genome sequencing projects

The recent years have witnessed an explosive expansion of genome sequencing projects promoted by the rapid development of sequencing technologies. According to published statistics from Genome Online Database (GOLD) (6), a steep increase in the number of genome sequencing projects appeared around 2011 when the NGS technologies matured (Figure 4.1). However, the growth slowed down in the wake of the peak due to the inherited read length limitations of the NGS platforms. As genomes with highly repetitive content cannot be assembled completely from short reads, many of them remained in a draft state with numerous unfilled gaps and thousands of unplaced scaffolds (contigs). The turning point came after 2015, owing to the development of the third-generation sequencing (TGS) technologies and novel scaffolding approaches, such as the chromosome conformation capture (Hi-C) (7) and BioNano optical maps (8). These new technologies triggered a new wave of genome sequencing projects.

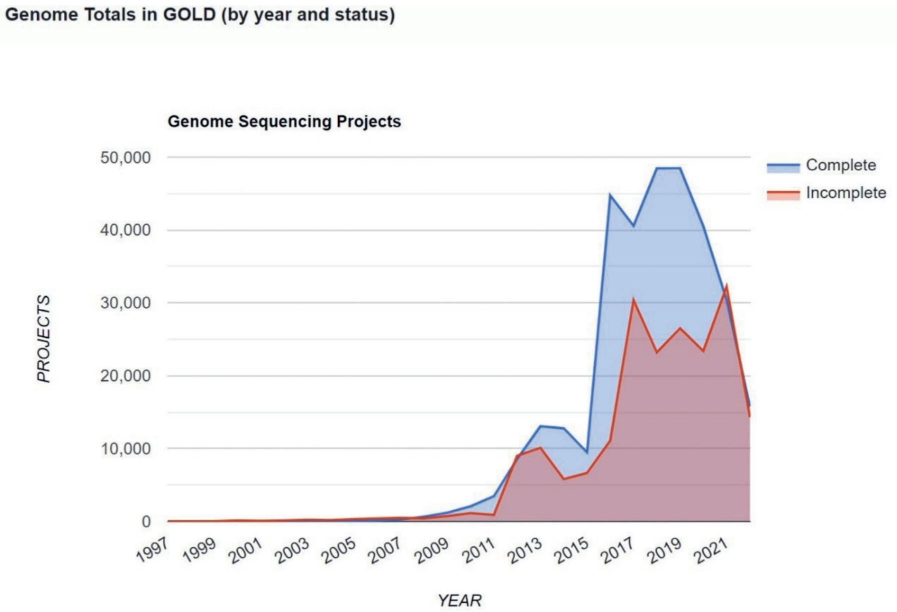


Figure 4.1 **The rapid increase of genome sequencing projects**

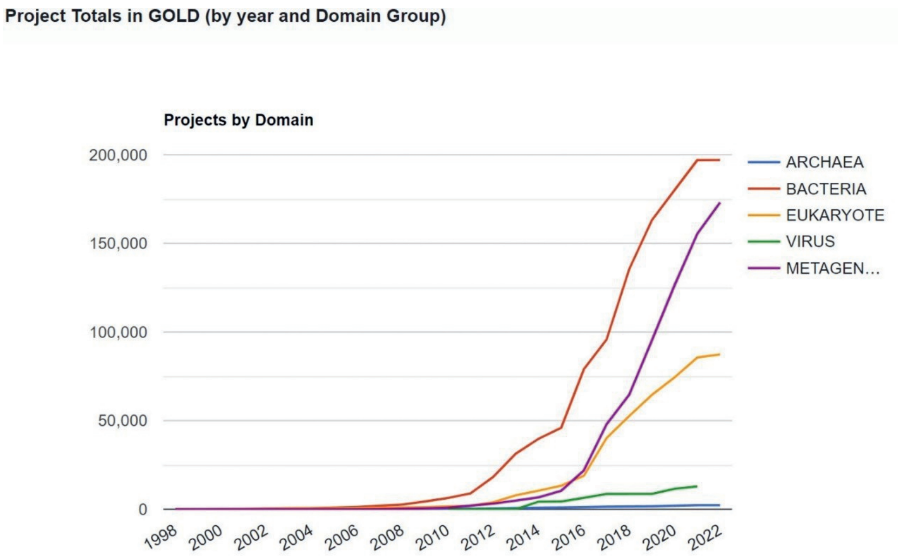


Figure 4.2 **Genome Projects according to Phylogenetic Groups**

The statistics (June 2022; https://gold.jgi.doe.gov/statistics) show a total of 472,698 sequencing projects. More than 30,000 new sequencing projects have been started during 2020 despite the onset of the Covid-19 pandemic at the beginning of the year. Along with the bacterial genome projects, the increasing number of more complex eukaryotic genomes and metagenomes is noticeable (Figure 4.2)

4.3 Next generation sequencing methods

4.3.1 From first-generation sequencing to the second-generation sequencing

First-generation sequencing protocols such as the Sanger method were widely used in many important sequencing projects in the early days. Over the past decades, Sanger sequencing has been improved to achieve read lengths up to ~1,000 bp, per-base accuracy as high as 99.999%, and cost as low as $ 0.50 per kilobase (5). Nevertheless, the Sanger protocol cannot fulfill the requirements of current genome sequencing applications. Promoted by the rapid progress of instrument and algorithm development, several second (also termed as next-generation) generation (NGS) sequencing platforms have been developed and are now available.

4.3.2 NGS platforms

The key steps in the next-generation sequencing protocol generally include DNA sample preparation, DNA sample fragmentation, DNA fragment amplification, sequencing by synthesis (SBS) of the DNA fragments - commonly called reads - and assembly of the sequenced reads by bioinformatics tools. Several frequently used sequencing platforms are reviewed below.

Table 4.1 **Comparison of frequently used sequencing platforms**

Platform	**3730XL**	**HiSeq 2000**	**Ion Torrent**
Amplification	Clonal plasmid	Bridge PCR	emPCR on bead
Chemistry	DCT[a]	Pyro-sequencing	H^+ detection
Instrument cost	$376k	$690k	$67.5k
Yields per run	1.9-84 kb	600 Gb	1 Gb
Read length	800-900 bp	100 bp	200 bp
Reagent cost	$96	$23,610	$925
Cost per Mb	$1600-$2400	$0.04-$0.07	$1
Error rate	0.1-1%	> 0.1%	~ 1%
Advantage	Low cost of small study	Most output at lowest cost	Fast run, low cost, and trajectory to longer read
Disadvantage	High cost for large study	High capital cost and computation need	Unreliable for long homopolymer region

The data are obtained from (9,10) and official documentations of Illumina and Life Technology.

[a]Dideoxy chain termination

The Illumina/Solexa amplifies DNA fragments by a bridge amplification approach (9) and can yield large amounts of data per run with a very low error rate. Illumina is commercially successful in the sequencing market as it launched different products to suit different scales of sequencing applications, such as the HiSeq and Miseq series platforms. As a result, Illumina has become the most widely used platform. The Ion Torrent from Life Technologies is another platform that uses semiconductors instead of optical devices in the SBS step (11). It does not detect the light signal from fluorescent dyes but measures the pH change resulting from the release of the H+ ion. The sequencing running cycle of Ion Torrent is the shortest among all the NGS platforms, so it has been used with some urgent sequencing applications for small genomes like pathogenic viruses or bacteria (12).

4.3.3 Third-generation sequencing

The main limitations of the current NGS platforms are the DNA fragmentation and amplification steps. The random fragmentation and PCR process inevitably introduces bias due to the specific characters of certain genomes, such as GC content and repeat content. To solve these problems, some post-modern generation (also called third generation) sequencing strategies have been developed. During the most recent decade, third-generation sequencing technologies matured to generate long reads spanning over 10 kb directly reading from DNA and RNA molecules. Two technologies currently dominate the long-read sequencing space: Pacific Biosciences (PacBio) single-molecule real-time (SMRT) sequencing (13) and Oxford Nanopore Technologies (ONT) nanopore sequencing (14). The ONT platform can generate longer reads, reaching 60 kilobytes (kb) in average length with 95 to 98% modal accuracy. PacBio high-fidelity (HiFi) reads, although shorter (~15 kb), are >99% accurate (Table 4.2). Long-read sequencing offers several advantages over short-read sequencing to improve *de novo*

Table 4.2 **Comparison of frequently used long-read sequencing platforms**

Platform	Pacific Bioscience Sequel II	Oxford Nanopore PromethION
Sequencing Principle	Detecting fluorescence events that correspond to the addition of one specific nucleotide by a polymerase	Measuring ionic current fluctuations when single-stranded nucleic acids pass through biological nanopores
Detection	Fluorescent	Electrical Conductivity
Average Read length (bases)	~30kb (HiFi reads ~15kb)	~60kb (Ultra-long reads >800kb)
Output/flow cell	600Gb	250Gb
Equipment cost	~$500,000	~$300,000
Max Accuracy	HiFi reads > 99%	R10 modal >98%
Advantages	Higher sequencing accuracy in HiFi model	Fast-sequencing; allows direct detection of DNA modifications
Disadvantages	Expensive and larger footprint of the sequencing equipment	Historically higher error rate (continues to improve)

The data are obtained from (17) and official documentation of Pacific Biosciences and Oxford Nanopore technologies.

assembly, mapping certainty, transcript isoform identification, and detection of structural variants (15,16).

4.3.4 NGS applications and analysis protocols

A complete sequencing project includes not only the generation of the DNA fragment reads, but also the assembly of these reads and the interpretation of the resulting sequence information. According to different biological questions, experimental designs and analysis protocols, NGS applications can be roughly divided into *de novo* sequencing and assembly, genome re-sequencing, and RNA-sequencing.

De novo sequencing assembles large numbers of reads without prior knowledge of the genome. It is a process similar to constructing a jigsaw puzzle without knowing the whole picture. Early Sanger sequencing applications usually followed large-scale shotgun assembly strategies to deal with the problem, which generally generates relatively low (~10x) coverage of reads and employs the Overlap-Layout-Consensus (OLC) approach to make the assembly. OLC is an intuitive algorithm, which includes three general steps: find Overlapping reads (O), Layout overlapping reads (L) and infer the Consensus sequence (C) (18). The combination of Sanger sequencing and the OLC protocol may give a good result but at a very high computational cost, which becomes one of the main limiting factors of its application range. The wide use of the NGS platforms significantly pushed forward *de novo* sequencing projects, as the sequencing cost dropped dramatically. Although the reads generated by commonly used NGS platforms are still not as long as those found by the Sanger method, new sample preparation and assembly strategies have been developed to solve the problem. The NGS short reads' most widely used assembly algorithm is based on the De Bruijn Graph (DBG). DBG assembly uses graph theory to infer the genome sequence from the De Bruijn graph formed by all the k-mers contained in all reads. Although the strategy of using NGS short reads for genome assembly was debated (19,20), more and more *de novo* assemblies of large-scale genomes have been carried out in this way (21,22).

With the rapid increase of available assembled genomes and the concomitant decrease in sequencing cost, the range of NGS applications is no longer limited to *de novo* sequencing. Instead, it has extended to genome re-sequencing. The variations between individual genomes within a species, represented by single nucleotide polymorphisms (SNPs), have been described earlier (23,24). However, early research on SNP identification was mainly based on Expressed Sequence Tag (EST) sequences and array data (25,26), making it very costly for large scale genome-wide mapping. As the sequence quality of these data is usually not high, its use is also limited by the high false-negative rates. NGS technologies have brought a big change to this application, as it allows fast and large-scale parallel identification of mutations between individuals or lines with a reasonable cost. It has been used to perform genome re-sequencing on natural strains of model organisms (27) and to identify SNPs and insertions and deletions (indels) with bioinformatics methods (28). With the development of new analysis tools (21,29) and an increase in computational capacity, mapping more specific genome-wide patterns of genetic variation gradually became the standard re-sequencing procedure, including not only SNPs but also indels, Copy Number Variations (CNVs) and Structural Variations (SVs).

The study of the transcriptome is another application extensively promoted by the spread of NGS technologies. Compared to the DNA microarray, which was the only choice for large-scale gene expression analysis in the early days, the NGS based RNA-sequencing (RNA-seq) protocol has advantages in many aspects. First, it can overcome the limitations of microarray-based protocols regarding background signal noise interference and the inconsistency of the expression level among different experimental replicates. The results of RNA-Seq also show high levels of reproducibility, for both technical and biological replicates (30). Furthermore, computational methods designed for RNA-seq data can discover new transcripts and alternative-spliced forms (31,32), which could extend the application scope of DNA microarray. Currently, RNA-seq applications aim not only at the study of mRNAs, but also at noncoding RNAs and small RNAs; not only to the transcriptional structure of genes, but also to their post-transcriptional modifications.

Besides the above applications, other newly invented NGS-based methods are also becoming popular, focusing on different biological questions. For example, directly sequencing the cytosine methylome (methylC-seq) is developed to map the epigenetic regulations. In plants, relevant work has been done in Arabidopsis (33,34), which sheds new light on how plants respond to the environment. Other NGS applications such as chromatin immune-precipitation followed by sequencing (ChIP-seq) (35) and whole exome sequencing of target-enriched genomic DNA (exome-seq) (36) are also widely used by more and more researchers.

4.4 NGS application in sorghum crop improvement

Sorghum (*Sorghum bicolor*) ranks the fifth most-grown cereal among the principal crops worldwide (USDA 2020). It feeds over 500 million people in nearly 100 countries (37), with an estimated 60 million tons produced yearly on 44 million hectares, making it a remarkable target for genetic improvement. Although this African-originated cereal (38) is grown as a food crop mainly in arid and semi-arid regions, due to its efficiency to produce high biomass under harsh conditions with low inputs (39), it has recently been explored for biofuel production fields, especially sweet sorghum, which accumulates fermentable sugars in the stem. This biofuel-associated trait makes sorghum a versatile crop providing food, feed, fiber and fuel. This C4 plant, belonging to the Poaceae subfamily (grass), which also includes maize, millet, sugarcane, miscanthus and switchgrass, is becoming a model system for biofuel crops and C4 carbon-fixing plants, justifying its early sequencing as the third plant in 2007 (40,41).

4.4.1 Sorghum genome

Rice has been the first fully-sequenced cereal plant (International Rice Genome Sequencing Project 2005). Given its importance as a crop and model plant, its genome is representative of C3 carbon-fixing grasses. As a C4 crop, one of the critical reasons for the sorghum genome sequencing by Patterson et al. (40) is its relatively small genome of 730 Mb compared to more complex and duplicated genomes such as that of the ancient allotetraploid maize (2.3 Gb) (42), believed to have diverged from sorghum 11.9 Mya (43). The BTx623 genotype was used as a reference and sequenced by Sanger Whole Genome Shotgun (WGS), a method also used for earlier genome sequencing projects including *Arabidopsis thaliana* (Arabidopsis Genome Initiative 2000) and *Oryza sativa* (International Rice Genome Sequencing Project 2005). The homozygous and diploid sorghum genome contains 34,496 genes over ten chromosomes, with about 28,000 *bona fide* annotated genes which have orthologues in rice, Arabidopsis and poplar (40). The increased larger sorghum genome size, as compared to the rice genome (roughly 400 Mb), is attributed to the significant proliferation of retrotransposon elements leading to a high repeat content (62%). The number and duplication level of genes and intron and exon sizes are similar to those described in rice or even Arabidopsis (Table 4.3). Despite the high number of genes conserved between sorghum and other sequenced plants, 24% of genes families revealed by the sequencing study are grasses-specific, while 7% are found to be species-specific (40). Ten years later, an improved version of the genome was published. Compared with the original reference genome, the new reference genome size added 29.6 Mb, the number of annotated genes increased by 24% to 34,211, and the sequencing error rate was reduced to 1 per 100 kb (44).

The availability of the sorghum genome sequence has provided new perspectives for sorghum improvement. Enhancing the understanding and knowledge of the genome is an essential step towards analyzing agronomic traits and the physiological progress involved in expressing these desirable traits. The sequenced genome combined with

Table 4.3 **Features of the sorghum genome**

Type	Average length
Gene	2856 bp
Transcript	1426 bp
Exon	267 bp
Intron	419 bp
Protein	409 aa

Data from http://genome.jgi-psf.org/Sorbi1/Sorbi1.info.html

NGS technologies has accelerated the pace of scientific discovery, leading to more routine utilization of NGS to analyze sorghum transcript expression, genome variations and grasses evolution (40,43,45,46).

Thanks to the development of TGS technologies, more chromosome-level *de novo* assembly were generated from other sorghum elite lines with ONT (14) and PacBio (13) platforms. In addition to the high-quality genome sequences from individual sorghum lines, we recently constructed the first sorghum pan-genome, consisting of the genome sequences assembled *de novo* from 16 diverse sorghum lines. This pan-genome sequence has a size of 955 Mb and contains 44,079 gene families. The percentage of dispensable genes in the sorghum pan-genome is much higher than that reported in *Oryza sativa* (54%), *Glycine max* (49%) and *Brachypodium distachyon* (45%), indicating that sorghum may have greater genetic diversity (47)

4.4.2 Genotyping-by-sequencing in sorghum

Advances in NGS sequencing platforms have accelerated the pace of genetic variation discovery and genotyping, which was previously performed through microarray platforms. Genetic variants are defined by sequence variant types, including SNPs, indels, microsatellites (Simple Sequence Repeats, SSRs) and transposable elements. Structural variations are also considered genetic variations and are designated as large-scale insertions/deletions, i.e., Presence/Absence Variations (PAVs) and Copy Number Variants (CNVs). Genetic variation is a helpful resource in plant breeding. It is responsible for observed phenotypic variation (45,48,49) and can be used to develop molecular markers towards molecular breeding programs, functional and evolutionary studies, association mapping (50,51) or a wide range of other applications. SNPs are now the most commonly used molecular markers in plant breeding and genetic studies but are limited to major crops due to the time and cost associated with their discovery, validation and utilization when processed by traditional molecular methods. SNPs are revealed by comparing two DNA sequences from two or more different accessions. NGS, combined with the accessibility of entire genome sequences, has directly characterized genetic variants at a reduced cost. Whole-genome SNP discovery and genotyping can now be easily accomplished for small genomes via NGS, termed genotyping-by-sequencing (GBS) (52-55).

Application of genotyping-by-sequencing in sorghum was first reported by Nelson et al. (56), where they sequenced three DNA libraries generated from eight grain sorghum accessions and aligned the resulting 247 M reads to the BTx623 reference genome. They found a higher SNPs distribution at the chromosome ends due to a high content of repetitive sequences near the centromeres. The SNP frequencies detected in sorghum are similar to those described in other plants. Indel discovery revealed higher proportions of in-frame indels than frame-shifting indels and a nonsynonymous-to-synonymous ratio of about 0.8, significantly lower than those calculated for soybean or rice but comparable to Arabidopsis (57-59).

Thus far, sequencing studies in sorghum for genetic variation discovery have been limited to grain sorghum. As sweet sorghum cultivars are gaining interest for biofuel production thanks to the high content of fermentable sugars in their tall stems, understanding the genetic patterns involved in these biofuel-associated traits is the first step towards breeding programs targeting these traits. Zheng et al. (60) conducted a genome-wide genetic variation study using GBS to identify genome regions and metabolic pathways potentially implicated in the sweet sorghum phenotype, found in several local races of bicolor species (61) but so far not genetically distinguished by diversity studies (60,62,63). The cited study informs on the re-sequencing of two sweet sorghum lines, Keller and E-Tian, and a Chinese kaoliang grain sorghum, Ji2731, to compare the sequence polymorphism and structural variants by aligning to the reference sorghum genome. The whole-genome shotgun strategy and Illumina sequencing used in this study yielded 620,72 million 44 bp paired-ends reads, detecting an overall of 1,057,018 SNPs, 99,948 indels, 16,487 PAVs and 17,111 CNVs among the four sorghum cultivars (Fig. 4.3). Fourteen gene families were enriched with large effect SNPs, which are predicted to have a potentially disabling effect on gene function, comprising families involved in biotic and abiotic stress responses. Indel proportions were 9.7%, 75.7% and 14.6% in coding, intronic and UTR regions, respectively, with lengths between 1–10 bp. Larger deletions were less common, with a higher abundance of 3 bp indels detected over the three lines. Frame-shifting indels were rarely encountered in coding regions. Indels and PAVs were observed within the same gene families where large-effect SNPs were located. The average length of PAVs was 2,394 bp, with 1,416 PAVs found in coding regions. CNVs ranged between 2 kb to 48 Mb in length, with 13,427 gains compared to 3,684 losses. A total of 1,442 genes discriminating sweet and grain sorghum participate in particular pathways, including starch and sucrose metabolism, lignin and coumarin biosynthesis associated phenylpropanoid biosynthesis.

As described in previous sections, extensive studies have been conducted on sorghum genetic variations, relationships with non-cultivated sorghum species, evolution, and genome-wide association studies (GWAS). Although data was released on sorghum diversity in the last decades, a limited number of molecular markers and the number and provenance of the sorghum accessions tested have restricted the use and integration

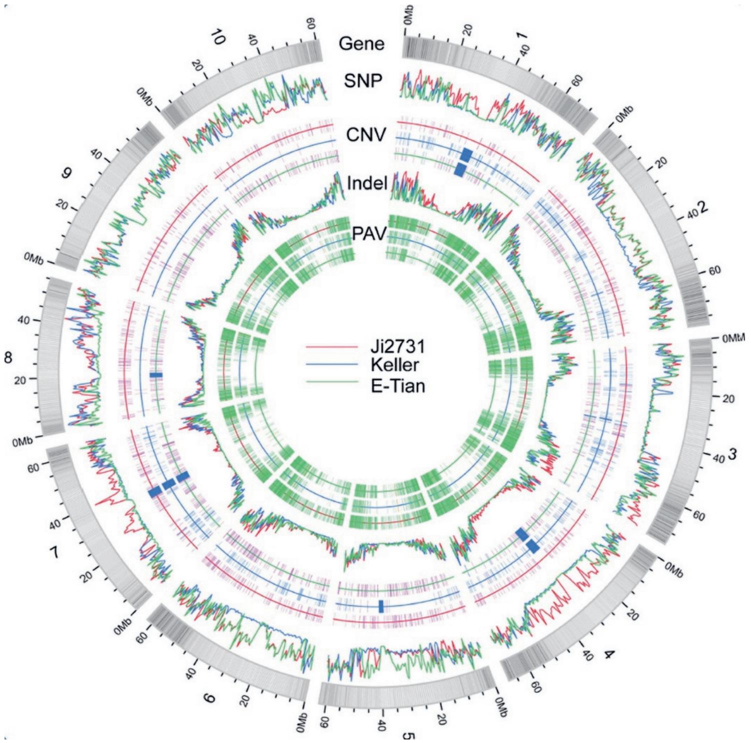


Figure 4.3. **Genome-wide variation between three sweet and grain sorghum lines** (60). Gene density of chromosomes is visualized by line intensity; the more genes on a chromosome region, the darker the color. The purple and blue colors in the CNV ring represent copy number gain and loss, respectively. For PAVs, the green color stands for the absence of variation, pink for the presence of variation.

of these data for breeding applications due to a lack of sensitivity and resolution to detect many associations.

As discussed in work conducted by Zheng et al. (60) and Nelson et al. (56), new sequencing platforms are bringing new approaches for analyzing genetic variation and SNP discovery in sorghum germplasm. Morris et al. (64) re-sequenced 971 sorghum accessions, combining landraces, breeding material and wide relatives using the Illumina Genome Analyser IIx/HiSeq platform, generating a total of 6.13 million unique 64 bp reads, 95% of which were aligned to the sorghum genome. They identified a total of 265,487 SNPs, equivalent to one SNP every 2.7 kb of the genome; 72% of these were within genes, 99% were located within 10 kb from 27,412 *bona fide* sorghum genes. High sorghum genome coverage with sufficient SNP markers dramatically increases the resolution for GWAS application, with an estimated minimum of 100,000 SNPs needed to achieve sufficient resolution in sorghum (65). As an example, application in GWAS, strong associations were found in this study for plant height components and inflorescence architecture. The sorghum population structure according to morphological characters and geographic regions proposed before was also validated through re-sequencing; the evolution of the sorghum races during domestication and crop spreading among several regions displaying different agroclimatic environments was also validated. As a result of this diversity study, it was shown that agroclimatic conditions and stresses had played a significant role, as crucial as geographic isolation, in shaping the diffusion process. Moreover, several domestication-related candidate genes were detected, such as the transcription factor opaque 2 and a region on chromosome 2 harboring a dwarf gene (66). This study resulted in a high number of SNP markers within genes, providing potential tools for Marker-Assisted Selection (MAS) or GWAS application to be integrated into breeding programs as well as for in situ sorghum germplasm conservation. Further increasing the number of sorghum accessions re-sequenced will provide a deeper understanding of genetic patterns between sorghum accessions.

4.4.3 Analysis of the sorghum transcriptome with NGS technologies

NGS applications in sorghum rapidly started after the completion of the genome sequence of BTx623. Transcriptome analysis allows the capture of gene expression variation between tissues and genotypes that control important traits. Transcriptome profiles can be regarded as heritable quantitative traits segregating in a population. Before the sorghum genome sequence release and NGS, transcriptome analyses in sorghum were conducted by cDNA microarrays, measuring in responses to several biotic and abiotic stresses, including the effects of PEG-induced osmotic stress, exogenous abscisic acid (ABA), salt, jasmonic acid and wounding by insects (67-69). As in other organisms, microarray technology has been a high-throughput method of choice in the last decade for transcriptome analysis of sorghum but is now replaced by RNA-seq, providing a deeper insight into plant gene expression and RNA sequences in general, permitting novel genome annotation and non-coding RNA discovery.

4.4.3.1 Transcriptome analysis for abiotic stress response studies

Drought tolerance is a significant trait of interest in sorghum; after millet, sorghum is the most drought-tolerant crop. However, little work has been done on the physiological processes and gene networks involved in this trait in sorghum, in contrast to extensive studies carried out for drought tolerance in rice (70-72), maize (73-75) or the model plant Arabidopsis (76-81). Buchanan et al. (67) studied gene expression response to ABA and PEG-induced osmotic stresses using microarray and revealed 12,982 involved genes. Dugas et al. (82) conducted a whole-genome transcriptome analysis on sorghum using the Illumina GAIIx sequencer. They produced 689.5 million reads of 50 bp, of which 87% were mapped to the sorghum genome: 72% in exons, 3% in introns, 10% in intergenic regions and 15% in splice junctions. Of the 34,144 sorghum genes detected, 84% showed a transcriptional activity. Of these, 92% corresponded to *bona fide* high confidence protein-coding genes. They also found that between 2,300 and 1,650 transcripts were upregulated, and 2,600 and 700 were downregulated by ABA and osmotic stress, respectively. Further study of the affected transcripts revealed that 29 of the 60 genes studied were considered putative or unannotated, and others were possibly involved in abscisic acid and 13-lipoxygenase, salicylic acid, jasmonic acid and plant defense pathways.

4.4.3.2 Transcriptome analysis for biotic stress response studies

Another recent deployment of NGS was reported by Mizuno et al. (83) to identify and characterize key genes responsible for resistance to a fungal pathogen. Target leaf spot, caused by the necrotrophic fungus *Bipolaris sorghicola* (84), is a main foliar disease in sorghum when grown under humid conditions, damaging plant biomass yield. When infected by the fungus, sorghum produces a unique class of phytoalexins, named 3-deoxyanthocyanidins (85). Whole transcriptome sequencing was carried out on BTx623 infected by *Bipolaris sorghicola*, generating 34 million reads containing 7,674 unannotated transcripts at 6,063 different loci. Differentially expressed transcripts encode genes responsible for biosynthesis of molecules such as 3-deoxyanthocyanidin, or for enzymes catalyzing reactions producing suberin, an important component on the lipophilic cell wall barrier. In response to the pathogen infection, they found that the TCA cycle changed its function from energy production to cell wall components biosynthesis. Also, amino-acid metabolizing enzyme encoding genes were upregulated, activating the phytoalexin synthesis and the sulfur-dependent detoxification pathway.

4.4.3.3 Transcriptome analysis of miRNA components in sorghum

Small RNAs, defined as miRNAs between 18 to 25 nucleotides, are valuable elements of the transcriptome. Once transcribed from DNA, these small RNA fragments methylate to form the complex miRNA:miRNA hybrid (perfect or imperfect) and are loaded onto the RNA induced silencing complex. They act as the regulator by repressing targeted mRNAs, leading to their degradation inside the cytoplasm (86-90). They have been well described in several major crops, with 213 miRNA families reported for Arabidopsis and another 462 miRNAs described and characterized in rice (91,92). Before 2011, only conserved miRNAs had been predicted in sorghum, but none of the studies thus far had revealed novel miRNAs (40,93). Calviño et al. (41) were the first to publish on miRNA characterization in sorghum after a transcriptome study. They exploited the SOLID 3 platform to sequence small RNA fragments extracted from stem tissues at the flowering time of two cultivars (BTx623, grain sorghum and Rio, a sweet sorghum cultivar) and an F2 population segregating for the flowering time and stem sugar content. The sequencing output provided 38,336,769 sequence reads, 60% mapping to the BTx623 genome. Sorghum miRNAs were found with a higher abundance of 25 and 24 nt length classes, and a second peak at 22 nt, higher than those for the 20 and 21 nt classes, was also reported. The sorghum genome sequencing predicted 149 miRNAs belonging to 27 miRNA families (40) while Calviño et al. (41) detected miRNA members from 25 families, with a higher concentration from miR172 family accounting for 6% of the total reads and 15% in the BTx623 library. They also discovered nine new miRNAs.

Shortly after, studies (94,95) discovered 13 novel miRNA families, including seven also conserved in related monocots, and measured their temporal expression and mRNA targets. They sequenced small RNAs of 18 to 26 nt from a 3-week-old sweet sorghum plant (M81E). Among the 619,010 sequence reads generated, they noticed a 24 nt miRNA abundance peak and a 21 nt peak. They obtained a total of 113 conserved miRNA homologs belonging to 31 miRNA families and, unlike the previous study, the most abundantly expressed miRNA family was miR166. The miR169 and miR444 families are represented with 14 and 12 members, respectively. About 100 genes targeted by the miRNAs were predicted, the majority of which were transcription factor encoding genes and genes likely involved in growth and development processes, nutrient translocation, assimilation pathways as well as responses to biotic and abiotic stresses. Of the 25 predicted target genes, two were validated and annotated in the sorghum genome, underlining the use of NGS for the discovery of novel genes through miRNA sequencing. These studies provide a better understanding of the spatiotemporal regulation of target genes towards the improvement for biomass accumulation, biotic and abiotic stresses and biofuel-associated traits.

The mRNA-sequencing technologies are becoming vastly popular for sorghum transcriptome studies due to several advantages. RNA transcript sequencing provides information on all transcribed genes, sequences, and expressions without genomic sequencing information. Furthermore, these transcript sequences are useful resources towards genome-wide and comparative studies on ortholog genes, species evolution, species-specific novel transcripts and gene discovery. As such technologies also reveal distinct expression levels between duplicated genes, it is a method of choice for studying duplicated genomes such as sorghum.

4.5 Perspectives of NGS in sorghum breeding

NGS has revolutionized the field of plant breeding by enabling the sequencing of crop genomes, discovery of polymorphisms, gene expression studies, genotyping breeding populations, diversity studies and GWAS. Moreover, breeding population development and typing have indirectly evolved as a result of these newer and faster genetic techniques, allowing exploration of genetic variation at a larger scale and lower cost. NGS technologies provide large volumes of genetic data, improving read length and data accuracy, which opens doors toward a broad range of applications in plant sciences, including more complex genomes such as wheat (96).

4.5.1 Next-generation populations

As discussed in the previous sections, essential molecular biology tools previously employed in plant breeding were rapidly replaced by NGS tools. When basic plant breeding populations were limited to simple crosses involving two contrasting parents, this restricted the number of polymorphisms screened and rare alleles detected. New populations, so-called next-generation populations, have recently been designed. Nested Association Mapping (NAM) populations are an excellent example of these. Developed through crossing diverse accessions with a reference parent, NAM populations have been successful as new breeding material in maize(97,98). Another example is the multi-parent advanced generation inter-cross (MAGIC) populations, which result from intercrossing multiple parents to form better populations (99).

Although next-generation populations are replacing traditional mapping populations, the approach will need to be adapted to each crop as they differ in their reproductive systems. For instance, strictly selfing crops display substantial barriers while intercrossed. Thus, a particular population needs to be designed, taking into account the special physiological properties of the studied crop.

4.5.2 Molecular marker development

As sequenced genomes are becoming available for most major crops, the development of massive numbers of molecular markers is now becoming feasible with DNA/RNA sequencing approaches, especially for SNP and SSR marker systems, which are commonly used for diversity and Quantitative Trait Loci (QTL) mapping studies. With increasing read depth of NGS, SNP detection accuracy is proportionally improving. Using NGS for SSR discovery will be less time consuming than traditional methods. Genome resequencing with NGS such as Roche 454 or Illumina GA has already been described as highly successful and effective for generating thousands of molecular markers for crops by alignment to reference genome sequences. Extensive collections of molecular markers will be highly desired for genetic map construction, GWAS, and diversity studies. Moreover, GBS technology will soon replace laborious laboratory work for discovering and selecting polymorphism markers over parental breeding lines, calculating recombination frequencies and ordering markers on physical maps. GBS is increasing the likelihood of finding and tagging causal polymorphisms on physical maps by generating tens of thousands of usable markers.

Resequencing platforms, together with sequenced genomes, are valuable tools and resources to speed up the discovery of thousands of novel genetic variants among genomes. These variations can be located in agronomic trait-controlled genome regions, and they can be applied in breeding programs to track desirable characters. Using NGS in breeding programs will identify genome-wide diversity even if this was not phenotypically recorded. It could be useful in breeding genetic materials for diverse environments.

4.5.3 QTL mapping and association mapping

SNP and SSR markers discovered through NGS can be used for QTL mapping. With a higher number of mapped markers, the probability of obtaining markers linked or close to genes controlling a QTL increase, which also increases QTL mapping power and accuracy for the tight association. However, since bi-parental QTL mapping is based on segregation between two parents, association mapping (GWAS) has been used for many crops due to NGS technologies. GWAS allows the discovery of genetic differences in extensive germplasm collections and can be directly introduced into breeding programs to map complex traits when many molecular markers are available.

In conclusion, the emergence and maturation of NGS technologies allow the sorghum breeding community to explore genetic and genomic diversity at an unprecedented pace and scale. Many research activities in sorghum research and breeding unimaginable ten years ago can now be performed routinely. It dramatically enhances the genetic improvement and breeding of ideotypes for this important crop.

References

1. Sanger, F., Air, G.M., Barrell, B.G., Brown, N.L., Coulson, A.R., Fiddes, J.C., Hutchison, C.A., Slocombe, P.M. and Smith, M. (1977) Nucleotide sequence of bacteriophage φX174 DNA. *Nature*, **265**, 687-695.
2. Fleischmann, R., Adams, M., White, O., Clayton, R., Kirkness, E., Kerlavage, A., Bult, C., Tomb, J., Dougherty, B., Merrick, J. *et al.* (1995) Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science*, **269**, 496-512.
3. Hunkapiller, T., Kaiser, R., Koop, B. and Hood, L. (1991) Large-scale and automated DNA sequence determination. *Science*, **254**, 59-67.
4. Ezkurdia, I., Juan, D., Rodriguez, J.M., Frankish, A., Diekhans, M., Harrow, J., Vazquez, J., Valencia, A. and Tress, M.L. (2014) Multiple evidence strands suggest that there may be as few as 19,000 human protein-coding genes. *Human Molecular Genetics*, **23**, 5866-5878.
5. Shendure, J. and Ji, H. (2008) Next-generation DNA sequencing. *Nature Biotechnology*, **26**, 1135-1145.
6. Mukherjee, S., Stamatis, D., Bertsch, J., Ovchinnikova, G., Katta, H.Y., Mojica, A., Chen, I.-M.A., Kyrpides, N.C. and Reddy, T. (2018) Genomes OnLine database (GOLD) v.7: updates and new features. *Nucleic Acids Research*, **47**, D649-D659.
7. Kalhor, R., Tjong, H., Jayathilaka, N., Alber, F. and Chen, L. (2012) Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. *Nature Biotechnology*, **30**, 90-98.
8. Lam, E.T., Hastie, A., Lin, C., Ehrlich, D., Das, S.K., Austin, M.D., Deshpande, P., Cao, H., Nagarajan, N., Xiao, M. *et al.* (2012) Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly. *Nature Biotechnology*, **30**, 771-776.
9. Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., Lin, D., Lu, L. and Law, M. (2012) Comparison of Next-Generation Sequencing Systems. *Journal of Biomedicine and Biotechnology*, **2012**, 251364.
10. Metzker, M.L. (2010) Sequencing technologies — the next generation. *Nature Reviews Genetics*, **11**, 31-46.
11. Rothberg, J.M., Hinz, W., Rearick, T.M., Schultz, J., Mileski, W., Davey, M., Leamon, J.H., Johnson, K., Milgrew, M.J., Edwards, M. *et al.* (2011) An integrated semiconductor device enabling non-optical genome sequencing. *Nature*, **475**, 348-352.
12. Rasko, D.A., Webster, D.R., Sahl, J.W., Bashir, A., Boisen, N., Scheutz, F., Paxinos, E.E., Sebra, R., Chin, C.-S., Iliopoulos, D. *et al.* (2011) Origins of the E. coli Strain Causing an Outbreak of Hemolytic–Uremic Syndrome in Germany. *New England Journal of Medicine*, **365**, 709-717.
13. Cooper, E.A., Brenton, Z.W., Flinn, B.S., Jenkins, J., Shu, S., Flowers, D., Luo, F., Wang, Y., Xia, P., Barry, K. *et al.* (2019) A new reference genome for Sorghum bicolor reveals high levels of sequence similarity between sweet and grain genotypes: implications for the genetics of sugar metabolism. *BMC Genomics*, **20**, 420.
14. Deschamps, S., Zhang, Y., Llaca, V., Ye, L., Sanyal, A., King, M., May, G. and Lin, H. (2018) A chromosome-scale assembly of the sorghum genome using nanopore sequencing and optical mapping. *Nature Communications*, **9**, 4844.
15. Burgess, D.J. (2018) Next regeneration sequencing for reference genomes. *Nature Reviews Genetics*, **19**, 125-125.
16. Pollard, M.O., Gurdasani, D., Mentzer, A.J., Porter, T. and Sandhu, M.S. (2018) Long reads: their purpose and place. *Human Molecular Genetics*, **27**, R234-R241.
17. Hu, T., Chitnis, N., Monos, D. and Dinh, A. (2021) Next-generation sequencing technologies: An overview. *Human Immunology*.

18. Li, Z., Chen, Y., Mu, D., Yuan, J., Shi, Y., Zhang, H., Gan, J., Li, N., Hu, X., Liu, B. *et al.* (2011) Comparison of the two major classes of assembly algorithms: overlap–layout–consensus and de-bruijn-graph. *Briefings in Functional Genomics*, **11**, 25-37.

19. Alkan, C., Sajjadian, S. and Eichler, E.E. (2011) Limitations of next-generation genome sequence assembly. *Nature Methods*, **8**, 61-65.

20. Birney, E. (2011) Assemblies: the good, the bad, the ugly. *Nature Methods*, **8**, 59-60.

21. Li, R., Zhu, H., Ruan, J., Qian, W., Fang, X., Shi, Z., Li, Y., Li, S., Shan, G., Kristiansen, K. *et al.* (2010) De novo assembly of human genomes with massively parallel short read sequencing. *Genome Research*, **20**, 265-272.

22. Gnerre, S., MacCallum, I., Przybylski, D., Ribeiro, F.J., Burton, J.N., Walker, B.J., Sharpe, T., Hall, G., Shea, T.P., Sykes, S. *et al.* (2011) High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proceedings of the National Academy of Sciences*, **108**, 1513-1518.

23. Altshuler, D., Pollara, V.J., Cowles, C.R., Van Etten, W.J., Baldwin, J., Linton, L. and Lander, E.S. (2000) An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature*, **407**, 513-516.

24. Mullikin, J.C., Hunt, S.E., Cole, C.G., Mortimore, B.J., Rice, C.M., Burton, J., Matthews, L.H., Pavitt, R., Plumb, R.W., Sims, S.K. *et al.* (2000) An SNP map of human chromosome 22. *Nature*, **407**, 516-520.

25. Batley, J., Barker, G., O'Sullivan, H., Edwards, K.J. and Edwards, D. (2003) Mining for Single Nucleotide Polymorphisms and Insertions/Deletions in Maize Expressed Sequence Tag Data. *Plant Physiology*, **132**, 84-91.

26. Schmid, K.J., Sörensen, T.R., Stracke, R., Törjék, O., Altmann, T., Mitchell-Olds, T. and Weisshaar, B. (2003) Large-Scale Identification and Analysis of Genome-Wide Single-Nucleotide Polymorphisms for Mapping in Arabidopsis thaliana. *Genome Research*, **13**, 1250-1257.

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

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

29. Li, R., Li, Y., Fang, X., Yang, H., Wang, J., Kristiansen, K. and Wang, J. (2009) SNP detection for massively parallel whole-genome resequencing. *Genome Research*, **19**, 1124-1132.

30. Wang, Z., Gerstein, M. and Snyder, M. (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nature Review Genetics*, **10**, 57-63.

31. Nicolae, M., Mangul, S., Măndoiu, I.I. and Zelikovsky, A. (2011) Estimation of alternative splicing isoform frequencies from RNA-Seq data. *Algorithms for Molecular Biology*, **6**, 9.

32. Roberts, A., Pimentel, H., Trapnell, C. and Pachter, L. (2011) Identification of novel transcripts in annotated genomes using RNA-Seq. *Bioinformatics*, **27**, 2325-2329.

33. Lister, R., O'Malley, R.C., Tonti-Filippini, J., Gregory, B.D., Berry, C.C., Millar, A.H. and Ecker, J.R. (2008) Highly Integrated Single-Base Resolution Maps of the Epigenome in Arabidopsis. *Cell*, **133**, 523-536.

34. Qian, W., Miki, D., Zhang, H., Liu, Y., Zhang, X., Tang, K., Kan, Y., La, H., Li, X., Li, S. *et al.* (2012) A Histone Acetyltransferase Regulates Active DNA Demethylation in *Arabidopsis*. *Science*, **336**, 1445-1448.

35. Park, P.J. (2009) ChIP-seq: advantages and challenges of a maturing technology. *Nature Reviews Genetics*, **10**, 669-680.

36. Clark, M.J., Chen, R., Lam, H.Y.K., Karczewski, K.J., Chen, R., Euskirchen, G., Butte, A.J. and Snyder, M. (2011) Performance comparison of exome DNA sequencing technologies. *Nature Biotechnology*, **29**, 908-914.

37. Pennisi, E. (2009) How Sorghum Withstands Heat and Drought. *Science*, **323**, 573-573.

38. Doggett, H. (1967) Yield Increase from Sorghum Hybrids. *Nature*, **216**, 798-799.

39. Sasaki, T. and Antonio, B.A. (2009) Sorghum in sequence. *Nature*, **457**, 547-548.

40. Paterson, A.H., Bowers, J.E., Bruggmann, R., Dubchak, I., Grimwood, J., Gundlach, H., Haberer, G., Hellsten, U., Mitros, T., Poliakov, A. *et al.* (2009) The Sorghum bicolor genome and the diversification of grasses. *Nature*, **457**, 551-556.

41. Calviño, M. and Messing, J. (2012) Sweet sorghum as a model system for bioenergy crops. *Current Opinion in Biotechnology*, **23**, 323-329.

42. Schnable, P.S., Ware, D., Fulton, R.S., Stein, J.C., Wei, F., Pasternak, S., Liang, C., Zhang, J., Fulton, L., Graves, T.A. *et al.* (2009) The B73 Maize Genome: Complexity, Diversity, and Dynamics. *Science*, **326**, 1112-1115.

43. Swigonova, Z., Lai, J., Ma, J., Ramakrishna, W., Llaca, V., Bennetzen, J.L. and Messing, J. (2004) On the Tetraploid Origin of the Maize Genome. *Comparative and Functional Genomics*, **5**, 670102.

44. McCormick, R.F., Truong, S.K., Sreedasyam, A., Jenkins, J., Shu, S., Sims, D., Kennedy, M., Amirebrahimi, M., Weers, B.D., McKinley, B. *et al.* (2018) The *Sorghum bicolor* reference genome: improved assembly, gene annotations, a transcriptome atlas, and signatures of genome organization. *The Plant Journal*, **93**, 338-354.

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

46. Swaminathan, K., Alabady, M.S., Varala, K., De Paoli, E., Ho, I., Rokhsar, D.S., Arumuganathan, A.K., Ming, R., Green, P.J., Meyers, B.C. *et al.* (2010) Genomic and small RNA sequencing of *Miscanthus × giganteusshows* the utility of sorghum as a reference genome sequence for Andropogoneae grasses. *Genome Biology*, **11**, R12.

47. Tao, Y., Luo, H., Xu, J., Cruickshank, A., Zhao, X., Teng, F., Hathorn, A., Wu, X., Liu, Y., Shatte, T. *et al.* (2021) Extensive variation within the pan-genome of cultivated and wild sorghum. *Nature Plants*.

48. Ganal, M.W., Altmann, T. and Röder, M.S. (2009) SNP identification in crop plants. *Current Opinion in Plant Biology*, **12**, 211-217.

49. Langridge, P. and Fleury, D. (2011) Making the most of 'omics' for crop breeding. *Trends in Biotechnology*, **29**, 33-40.

50. Huang, X., Wei, X., Sang, T., Zhao, Q., Feng, Q., Zhao, Y., Li, C., Zhu, C., Lu, T., Zhang, Z. *et al.* (2010) Genome-wide association studies of 14 agronomic traits in rice landraces. *Nature Genetics*, **42**, 961-967.

51. Kump, K.L., Bradbury, P.J., Wisser, R.J., Buckler, E.S., Belcher, A.R., Oropeza-Rosas, M.A., Zwonitzer, J.C., Kresovich, S., McMullen, M.D., Ware, D. *et al.* (2011) Genome-wide association study of quantitative resistance to southern leaf blight in the maize nested association mapping population. *Nature Genetics*, **43**, 163-168.

52. Sánchez, C.C., Smith, T.P.L., Wiedmann, R.T., Vallejo, R.L., Salem, M., Yao, J. and Rexroad, C.E. (2009) Single nucleotide polymorphism discovery in rainbow trout by deep sequencing of a reduced representation library. *BMC Genomics*, **10**, 559.

53. Wiedmann, R.T., Smith, T.P.L. and Nonneman, D.J. (2008) SNP discovery in swine by reduced representation and high throughput pyrosequencing. *BMC Genetics*, **9**, 81.

54. Van Tassell, C.P., Smith, T.P.L., Matukumalli, L.K., Taylor, J.F., Schnabel, R.D., Lawley, C.T., Haudenschild, C.D., Moore, S.S., Warren, W.C. and Sonstegard, T.S. (2008) SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. *Nature Methods*, **5**, 247-252.

55. Kerstens, H.H.D., Crooijmans, R.P.M.A., Veenendaal, A., Dibbits, B.W., Chin-A-Woeng, T.F.C., den Dunnen, J.T. and Groenen, M.A.M. (2009) Large scale single nucleotide polymorphism discovery in unsequenced genomes using second generation high throughput sequencing technology: applied to turkey. *BMC Genomics*, **10**, 479.

56. Nelson, J.C., Wang, S., Wu, Y., Li, X., Antony, G., White, F.F. and Yu, J. (2011) Single-nucleotide polymorphism discovery by high-throughput sequencing in sorghum. *BMC Genomics*, **12**, 352.

57. McNally, K.L., Childs, K.L., Bohnert, R., Davidson, R.M., Zhao, K., Ulat, V.J., Zeller, G., Clark, R.M., Hoen, D.R., Bureau, T.E. *et al.* (2009) Genomewide SNP variation reveals relationships among landraces and modern varieties of rice. *Proceedings of the National Academy of Sciences of the United States of America*, **106**, 12273-12278.

58. Lam, H.-M., Xu, X., Liu, X., Chen, W., Yang, G., Wong, F.-L., Li, M.-W., He, W., Qin, N., Wang, B. *et al.* (2010) Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. *Nature Genetics*, **42**, 1053-1059.

59. Arai-Kichise, Y., Shiwa, Y., Nagasaki, H., Ebana, K., Yoshikawa, H., Yano, M. and Wakasa, K. (2011) Discovery of Genome-Wide DNA Polymorphisms in a Landrace Cultivar of Japonica Rice by Whole-Genome Sequencing. *Plant and Cell Physiology*, **52**, 274-282.

60. Zheng, L.-Y., Guo, X.-S., He, B., Sun, L.-J., Peng, Y., Dong, S.-S., Liu, T.-F., Jiang, S., Ramachandran, S., Liu, C.-M. *et al.* (2011) Genome-wide patterns of genetic variation in sweet and grain sorghum (Sorghum bicolor). *Genome Biology*, **12**, R114.

61. Ritter, K.B., McIntyre, C.L., Godwin, I.D., Jordan, D.R. and Chapman, S.C. (2007) An assessment of the genetic relationship between sweet and grain sorghums, within *Sorghum bicolor* ssp. *bicolor* (L.) Moench, using AFLP markers. *Euphytica*, **157**, 161-176.

62. Draye, X., Lin, Y.-R., Qian, X.-y., Bowers, J.E., Burow, G.B., Morrell, P.L., Peterson, D.G., Presting, G.G., Ren, S.-x., Wing, R.A. *et al.* (2001) Toward Integration of Comparative Genetic, Physical, Diversity, and Cytomolecular Maps for Grasses and Grains, Using the Sorghum Genome as a Foundation. *Plant Physiology*, **125**, 1325-1341.

63. Paterson, A.H. (2008) Genomics of sorghum. *International Journal of Plant Genomics*, **2008**, 362451.

64. Morris, G.P., Ramu, P., Deshpande, S.P., Hash, C.T., Shah, T., Upadhyaya, H.D., Riera-Lizarazu, O., Brown, P.J., Acharya, C.B., Mitchell, S.E. *et al.* (2013) Population genomic and genome-wide association studies of agroclimatic traits in sorghum. *Proceedings of the National Academy of Sciences of the United States of America*, **110**, 453-458.

65. Bouchet, S., Pot, D., Deu, M., Rami, J.-F., Billot, C., Perrier, X., Rivallan, R., Gardes, L., Xia, L., Wenzl, P. *et al.* (2012) Genetic Structure, Linkage Disequilibrium and Signature of Selection in Sorghum: Lessons from Physically Anchored DArT Markers. *PLoS One*, **7**, e33470.

66. Brown, P.J., Upadyayula, N., Mahone, G.S., Tian, F., Bradbury, P.J., Myles, S., Holland, J.B., Flint-Garcia, S., McMullen, M.D., Buckler, E.S. *et al.* (2011) Distinct Genetic Architectures for Male and Female Inflorescence Traits of Maize. *PLoS Genetics*, **7**, e1002383.

67. Buchanan, C.D., Lim, S., Salzman, R.A., Kagiampakis, I., Morishige, D.T., Weers, B.D., Klein, R.R., Pratt, L.H., Cordonnier-Pratt, M.-M., Klein, P.E. *et al.* (2005) Sorghum bicolor's Transcriptome Response to Dehydration, High Salinity and ABA. *Plant Molecular Biology*, **58**, 699-720.

68. Salzman, R.A., Brady, J.A., Finlayson, S.A., Buchanan, C.D., Summer, E.J., Sun, F., Klein, P.E., Klein, R.R., Pratt, L.H., Cordonnier-Pratt, M.-M. *et al.* (2005) Transcriptional Profiling of Sorghum Induced by Methyl Jasmonate, Salicylic Acid, and Aminocyclopropane Carboxylic Acid Reveals Cooperative Regulation and Novel Gene Responses. *Plant Physiology*, **138**, 352-368.

69. Park, S.-J., Huang, Y. and Ayoubi, P. (2006) Identification of expression profiles of sorghum genes in response to greenbug phloem-feeding using cDNA subtraction and microarray analysis. *Planta*, **223**, 932-947.

70. Rabbani, M.A., Maruyama, K., Abe, H., Khan, M.A., Katsura, K., Ito, Y., Yoshiwara, K., Seki, M., Shinozaki, K. and Yamaguchi-Shinozaki, K. (2003) Monitoring Expression Profiles of Rice Genes under Cold, Drought, and High-Salinity Stresses and Abscisic Acid Application Using cDNA Microarray and RNA Gel-Blot Analyses. *Plant Physiology*, **133**, 1755-1767.

71. Zhao, B., Liang, R., Ge, L., Li, W., Xiao, H., Lin, H., Ruan, K. and Jin, Y. (2007) Identification of drought-induced microRNAs in rice. *Biochemical and Biophysical Research Communications*, **354**, 585-590.

72. Degenkolbe, T., Do, P.T., Zuther, E., Repsilber, D., Walther, D., Hincha, D.K. and Köhl, K.I. (2009) Expression profiling of rice cultivars differing in their tolerance to long-term drought stress. *Plant Molecular Biology*, **69**, 133-153.

73. Luo, M., Liu, J., Lee, R.D., Scully, B.T. and Guo, B. (2010) Monitoring the expression of maize genes in developing kernels under drought stress using oligo-microarray. *Journal of Integrative Plant Biology*, **52**, 1059-1074.

74. Hayano-Kanashiro, C., Calderón-Vázquez, C., Ibarra-Laclette, E., Herrera-Estrella, L. and Simpson, J. (2009) Analysis of Gene Expression and Physiological Responses in Three Mexican Maize Landraces under Drought Stress and Recovery Irrigation. *PLoS One*, **4**, e7531.

75. Zheng, J., Fu, J., Gou, M., Huai, J., Liu, Y., Jian, M., Huang, Q., Guo, X., Dong, Z., Wang, H. *et al.* (2010) Genome-wide transcriptome analysis of two maize inbred lines under drought stress. *Plant Molecular Biology*, **72**, 407-421.

76. Seki, M., Narusaka, M., Abe, H., Kasuga, M., Yamaguchi-Shinozaki, K., Carninci, P., Hayashizaki, Y. and Shinozaki, K. (2001) Monitoring the Expression Pattern of 1300 Arabidopsis Genes under Drought and Cold Stresses by Using a Full-Length cDNA Microarray. *The Plant Cell*, **13**, 61-72.

77. Hoth, S., Morgante, M., Sanchez, J.-P., Hanafey, M.K., Tingey, S.V. and Chua, N.-H. (2002) Genome-wide gene expression profiling in *Arabidopsis thaliana* reveals new targets of abscisic acid and largely impaired gene regulation in the *abi1-1* mutant. *Journal of Cell Science*, **115**, 4891-4900.

78. Kreps, J.A., Wu, Y., Chang, H.-S., Zhu, T., Wang, X. and Harper, J.F. (2002) Transcriptome Changes for Arabidopsis in Response to Salt, Osmotic, and Cold Stress. *Plant Physiology*, **130**, 2129-2141.

79. Oono, Y., Seki, M., Nanjo, T., Narusaka, M., Fujita, M., Satoh, R., Satou, M., Sakurai, T., Ishida, J., Akiyama, K. *et al.* (2003) Monitoring expression profiles of Arabidopsis gene expression during rehydration process after dehydration using ca. 7000 full-length cDNA microarray. *The Plant Journal*, **34**, 868-887.

80. Kilian, J., Whitehead, D., Horak, J., Wanke, D., Weinl, S., Batistic, O., D'Angelo, C., Bornberg-Bauer, E., Kudla, J. and Harter, K. (2007) The AtGenExpress global stress expression data set: protocols, evaluation and model data analysis of UV-B light, drought and cold stress responses. *The Plant Journal*, **50**, 347-363.

81. Matsui, A., Ishida, J., Morosawa, T., Okamoto, M., Kim, J.-M., Kurihara, Y., Kawashima, M., Tanaka, M., To, T.K., Nakaminami, K. *et al.* (2010) In Sunkar, R. (ed.), *Plant Stress Tolerance: Methods and Protocols*. Humana Press, Totowa, NJ, pp. 141-155.

82. Dugas, D.V., Monaco, M.K., Olson, A., Klein, R.R., Kumari, S., Ware, D. and Klein, P.E. (2011) Functional annotation of the transcriptome of Sorghum bicolor in response to osmotic stress and abscisic acid. *BMC Genomics*, **12**, 514.

83. Mizuno, H., Kawahigashi, H., Kawahara, Y., Kanamori, H., Ogata, J., Minami, H., Itoh, T. and Matsumoto, T. (2012) Global transcriptome analysis reveals distinct expression among duplicated genes during sorghum-Bipolaris sorghicolainteraction. *BMC Plant Biology*, **12**, 121.

84. Kawahigashi, H., Kasuga, S., Ando, T., Kanamori, H., Wu, J., Yonemaru, J.-i., Sazuka, T. and Matsumoto, T. (2011) Positional cloning of ds1, the target leaf spot resistance gene against *Bipolaris sorghicola* in sorghum. *Theoretical and Applied Genetics*, **123**, 131-142.

85. Snyder, B.A. and Nicholson, R.L. (1990) Synthesis of Phytoalexins in Sorghum as a Site-Specific Response to Fungal Ingress. *Science*, **248**, 1637-1639.

86. Song, X., Li, Y., Cao, X. and Qi, Y. (2019) MicroRNAs and Their Regulatory Roles in Plant–Environment Interactions. *Annual Review of Plant Biology*, **70**, 489-525.

87. Mallory, A.C. and Vaucheret, H. (2006) Functions of microRNAs and related small RNAs in plants. *Nature Genetics*, **38**, S31-S36.

88. Brodersen, P. and Voinnet, O. (2009) Revisiting the principles of microRNA target recognition and mode of action. *Nature Reviews Molecular Cell Biology*, **10**, 141-148.

89. Cui, J., You, C. and Chen, X. (2017) The evolution of microRNAs in plants. *Current Opinion in Plant Biology*, **35**, 61-67.

90. Jones-Rhoades, M.W., Bartel, D.P. and Bartel, B. (2006) MicroRNAs and Their Regulatory Roles IN Plants. *Annual Review of Plant Biology*, **57**, 19-53.

91. Sunkar, R. and Jagadeeswaran, G. (2008) *In silico*identification of conserved microRNAs in large number of diverse plant species. *BMC Plant Biology*, **8**, 37.

92. Sunkar, R. and Zhu, J.K. (2004) Novel and stress-regulated microRNAs and other small RNAs from Arabidopsis. *Plant Cell*, **16**, 2001-2019.

93. Bedell, J.A., Budiman, M.A., Nunberg, A., Citek, R.W., Robbins, D., Jones, J., Flick, E., Rholfing, T., Fries, J., Bradford, K. *et al.* (2005) Sorghum genome sequencing by methylation filtration. *PLoS Biology*, **3**, e13.

94. Zhang, L., Zheng, Y., Jagadeeswaran, G., Li, Y., Gowdu, K. and Sunkar, R. (2011) Identification and temporal expression analysis of conserved and novel microRNAs in Sorghum. *Genomics*, **98**, 460-468.

95. Calviño, M., Bruggmann, R. and Messing, J. (2011) Characterization of the small RNA component of the transcriptome from grain and sweet sorghum stems. *BMC genomics*, 10.1186/1471-2164-12-356.

96. Appels, R., Eversole, K., Feuillet, C., Keller, B., Rogers, J., Stein, N., Pozniak, C.J., Stein, N., Choulet, F., Distelfeld, A. *et al.* (2018) Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science*, **361**.

97. Yu, J., Holland, J.B., McMullen, M.D. and Buckler, E.S. (2008) Genetic Design and Statistical Power of Nested Association Mapping in Maize. *Genetics*, **178**, 539-551.

98. Buckler, E.S., Holland, J.B., Bradbury, P.J., Acharya, C.B., Brown, P.J., Browne, C., Ersoz, E., Flint-Garcia, S., Garcia, A., Glaubitz, J.C. *et al.* (2009) The Genetic Architecture of Maize Flowering Time. *Science*, **325**, 714-718.

99. Cavanagh, C., Morell, M., Mackay, I. and Powell, W. (2008) From mutations to MAGIC: resources for gene discovery, validation and delivery in crop plants. *Current Opinion in Plant Biology*, **11**, 215-221.

CHAPTER 5

SorGSD: a sorghum genome SNP database

Hong Luo, Wenming Zhao, Yanqing Wang, Yan Xia, Xiaoyuan Wu, Limin Zhang, Bixia Tang, Junwei Zhu, Lu Fang, Zhenglin Du, Wubishet A. Bekele, Shuaishuai Tai, David R. Jordan, Ian D. Godwin, Rod J. Snowdon, Emma S. Mace, Jingchu Luo, Hai-Chun Jing

Based on: Biotechnology for Biofuels, (2016) 9: 6.

Abstract

Sorghum (*Sorghum bicolor*) is one of the most important cereal crops globally and a potential energy plant for biofuel production. In order to explore genetic gain for a range of important quantitative traits, such as drought and heat tolerance, grain yield, stem sugar accumulation, and biomass production, via the use of molecular breeding and genomic selection strategies, knowledge of the available genetic variation and the underlying sequence polymorphisms is required.

Based on the assembled and annotated genome sequences of *Sorghum bicolor* (v2.1) and the recently published sorghum re-sequencing data, ~62.9M SNPs were identified among 48 sorghum accessions and included in a newly developed sorghum genome SNP database, SorGSD (http://sorgsd.big.ac.cn). The diverse panel of 48 sorghum lines can be classified into four groups, improved varieties, landraces, wild and weedy sorghums, and a wild relative *Sorghum propinquum*. SorGSD has a web-based query interface to search or browse SNPs from individual accessions, or to compare SNPs among several lines. The query results can be visualized in tables, or rendered as graphics in a genome browser. Users can also look into the SNPs annotations such as synonymous or non-synonymous, start, stop of splice variants, chromosome locations, and links to the annotation on Phytozome (www.phytozome.net) genome database. In addition, general information related to sorghum research such as online sorghum resources and literature references can be found on the website. All the SNP data and annotations can be freely downloaded from the website.

SorGSD is a comprehensive web-portal providing a database of large-scale genome variation across all racial types of cultivated sorghum and wild relatives. It can serve as a bioinformatics platform for a range of genomics and molecular breeding activities for sorghum and for other C4 grasses.

5.1 Background

Sorghum (Sorghum bicolor) originated from Africa and became an important cereal crop after a long period of domestication and selective breeding (1). Nowadays, it feeds over 500 million people in 98 countries (2), with an estimation of 42 million hectares of cultivated area and 62 million tons of yield per year (FAOSTAT data 2013, http://faostat3.fao.org). In contrast to C_3 crops such as rice and wheat, sorghum has the C_4 photosynthetic pathway, which leads to higher photosynthetic efficiency under circumstances of intense light, high temperature and low water supply (2-4). As such, sorghum has remarkable drought and heat tolerance, and can produce high yield and biomass in areas of harsh conditions with low inputs. Sorghum is not only used for food, but also cultivated with other important economic end-uses for forage, sugars and biomass. Furthermore, in recent years sorghum has been regarded as a promising bioenergy feedstock (5), which is comparable to other important biofuel grasses such as maize, sugarcane, Miscanthus and switch grass (6,7). Moreover, the compact genome and high degree of genetic synteny to other C_4 grasses make sorghum a potential genetic model for molecular design breeding of C_4 crops (8,9).

Sorghum's genome is relatively small (~730 M) and simple (10 chromosomes, diploid) compared to other C4 crops in the *Poaceae* subfamily, such as maize and sugarcane. The recent completion and availability of a whole genome reference sequence, based on the elite line BTx623, has accelerated the pace of genetic and genomic research in sorghum (10). The genetic basis of a range of important agronomic traits in sorghum has been elucidated, including drought tolerance and maturity (2). Nevertheless, to better understand the genetic basis for the considerable phenotypic variation observed in many more agronomic and bioenergy traits of different sorghum accessions, it is necessary to have insight into genomic variation including single nucleotide polymorphisms (SNPs), insertions/deletions (indels) and structure variation (SV).

Recently, various high throughput strategies have been developed for genome re-sequencing (11-13), resulting in a large amount of SNP data being generated for sorghum (14-18). These SNP data, representing high density biomarkers, are a valuable resource for researchers to perform genetic and breeding studies, such as genotyping by sequencing (GBS) (19-21), bulked segregant analysis (BSA) (22), and genome-wide association studies (GWAS) (18,23,24). These studies will not only lead to the highly efficient discovery of key quantitative trait locus (QTLs) or genes relevant to important traits, but also contribute to the understanding of the evolutionary relationship of cultivated and wild Sorghum species and subspecies.

To enhance the utility of sorghum SNP data, we developed a web-based large-scale genome variation database (SorGSD, http://sorgsd.big.ac.cn). SorGSD contains ~62.9 million SNPs from a diverse panel of 48 sorghum accessions divided into four groups, including improved inbreds, landraces, wild/weedy sorghums, and accessions of the wild relative S*orghum propinquum*. These SNP data have been annotated and an easy-to-use web interface has been designed for users to browse, search and analyze the SNPs efficiently. SorGSD allows users to query the SNP information and their relevant annotations for individual samples. The search results can be visualized graphically in a genome browser or displayed in formatted tables. Users can also compare SNP data between two or more sorghum accessions. The output of query results can be downloaded for further investigation, or users can bulk download the entire SNP dataset of 48 accessions. SorGSD also manages additional sorghum related information, such as general descriptions of sorghum and its genome, sorghum research institutions around the world, and lists of sorghum literature references.

5.2 Result and discussion

5.2.1 Database content

SorGSD contains ~62.9 million SNPs identified from the re-sequencing data of 48 sorghum lines mapped to the reference genome BTx623. These sorghum lines represent major cultivated races grouped into landraces or improved varieties, and weedy or wild

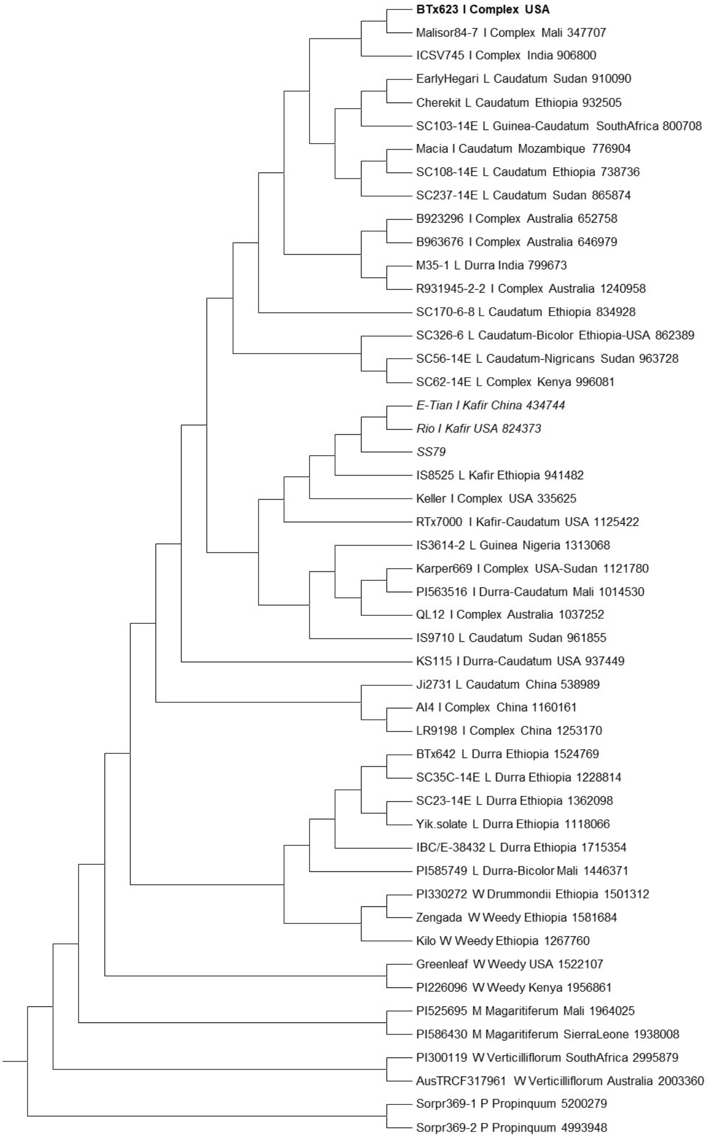

BTx623 I Complex USA
Malisor84-7 I Complex Mali 347707
ICSV745 I Complex India 906800
EarlyHegari L Caudatum Sudan 910090
Cherekit L Caudatum Ethiopia 932505
SC103-14E L Guinea-Caudatum SouthAfrica 800708
Macia I Caudatum Mozambique 776904
SC108-14E L Caudatum Ethiopia 738736
SC237-14E L Caudatum Sudan 865874
B923296 I Complex Australia 652758
B963676 I Complex Australia 646979
M35-1 L Durra India 799673
R931945-2-2 I Complex Australia 1240958
SC170-6-8 L Caudatum Ethiopia 834928
SC326-6 L Caudatum-Bicolor Ethiopia-USA 862389
SC56-14E L Caudatum-Nigricans Sudan 963728
SC62-14E L Complex Kenya 996081
E-Tian I Kafir China 434744
Rio I Kafir USA 824373
SS79
IS8525 L Kafir Ethiopia 941482
Keller I Complex USA 335625
RTx7000 I Kafir-Caudatum USA 1125422
IS3614-2 L Guinea Nigeria 1313068
Karper669 I Complex USA-Sudan 1121780
PI563516 I Durra-Caudatum Mali 1014530
QL12 I Complex Australia 1037252
IS9710 L Caudatum Sudan 961855
KS115 I Durra-Caudatum USA 937449
Ji2731 L Caudatum China 538989
AI4 I Complex China 1160161
LR9198 I Complex China 1253170
BTx642 L Durra Ethiopia 1524769
SC35C-14E L Durra Ethiopia 1228814
SC23-14E L Durra Ethiopia 1362098
Yik.solate L Durra Ethiopia 1118066
IBC/E-38432 L Durra Ethiopia 1715354
PI585749 L Durra-Bicolor Mali 1446371
PI330272 W Drummondii Ethiopia 1501312
Zengada W Weedy Ethiopia 1581684
Kilo W Weedy Ethiopia 1267760
Greenleaf W Weedy USA 1522107
PI226096 W Weedy Kenya 1956861
PI525695 M Magaritiferum Mali 1964025
PI586430 M Magaritiferum SierraLeone 1938008
PI300119 W Verticilliflorum SouthAfrica 2995879
AusTRCF317961 W Verticilliflorum Australia 2003360
Sorpr369-1 P Propinquum 5200279
Sorpr369-2 P Propinquum 4993948

Figure 5.1 **A dendrogram showing the phylogenetic relationships among the diverse set of sorghum lines.** Each sample is labelled as follows; the genotype name, sample type (coded, as detailed below), racial type, geographic origin, and total number of SNPs identified. Sample type codes: I improved variety, L landrace, W weedy or wild, M *margaritiferum*, P *Sorghum propinquum* (allopatric Asian species being as the outgroup). The sorghum reference genome BTx623 is shown in bold, sweet sorghums are in italic. (Adapted from Mace et al. (16) and redrawn using the tool "Display Newick Trees" under MEGA 6.0, SS79 was added based on the output results of the SNPhylo program (25) using the SNP data).

subspecies. Figure 5.1 shows the phylogenetic relationship among these sorghum lines (16), with the genotype name and group indicated. Racial type and geographic origin

Table 5.1 Distribution of SNPs in different genomic regions in 48 sorghum accessions

Genotype	Type*	Racial type	Geographic Origin	Total SNP numbers						
				All	Intergenic	5' UTR	Intronic	Non-Syn	Syn	3' UTR
BTx623	I	Complex	USA	0	0	0	0	0	0	0
Malisor84-7	I	Complex	Mali	347707	284944	2079	36036	7834	7175	8261
ICSV745	I	Complex	India	906800	762772	6166	81888	17476	15943	19300
EarlyHegari	L	Caudatum	Sudan	910090	748915	6893	90114	20375	19267	20719
Cherekit	L	Caudatum	Ethiopia	932505	763491	7179	96419	19765	18989	22799
SC103-14E	L	Guinea-Caudatum	South Africa	800708	657087	5589	82356	17216	15899	19275
Macia	I	Caudatum	Mozambique	776904	632772	5698	84057	16103	15199	19937
SC108-14E	L	Caudatum	Ethiopia	738736	600093	5969	78205	16647	15926	18758
SC237-14E	L	Caudatum	Sudan	865874	708878	7299	87805	18947	18174	21154
B923296	I	Complex	Australia	652758	537078	4312	66567	13219	12591	16395
B963676	I	Complex	Australia	646947	521677	5095	71404	14277	14007	17587
M35-1	L	Durra	India	799673	659631	5629	81167	15727	15239	19067
R931945-2-2	I	Complex	Australia	1240958	1045243	10365	109678	22904	21989	26306
SC170-6-8	L	Caudatum	Ethiopia	834928	698227	6107	77492	16224	15499	18180
SC326-6	L	Caudatum-Bicolor	Ethiopia-USA	862389	702263	6869	91952	18410	17529	21739
SC56-14E	L	Caudatum-Nigricans	Sudan	963728	788668	7831	98457	21451	19783	23541
SC62-14E	L	Complex	Kenya	996081	803234	8304	108872	23186	21516	26629
E-Tian	I	Kafir	China	434744	323422	8666	45605	19683	20023	14334
Rio	I	Kafir	USA	824373	660153	7410	92526	19751	18890	21916
SS79	I	Kafir	Ethiopia	1291250	1048752	16348	122623	34079	32586	30973
IS8525	L	Kafir	Ethiopia	941482	777365	7926	92487	19527	18210	22188
Keller	I	Complex	USA	335625	238622	4096	50617	13143	12560	14148
RTx7000	I	Kafir-Caudatum	USA	1125422	943142	9873	102492	21075	19846	24795
IS3614-2	L	Guinea	Nigeria	1313068	1102724	8066	123657	22931	21749	29188
Karper669	I	Complex	USA-Sudan	1121780	935393	7839	106347	22061	20738	25193
PI563516	I	Durra-Caudatum	Mali	1014530	835382	8632	100645	20999	20101	24679
QL12	I	Complex	Australia	1037252	860948	7376	101401	20297	18933	24245
IS9710	L	Caudatum	Sudan	961866	783299	6937	102930	20584	20071	24013
KS115	I	Durra-Caudatum	USA	937449	767552	4773	102892	17830	16454	24350
Ji2731	L	Caudatum	China	538989	395020	10246	60269	24847	25652	19250
AI4	I	Complex	China	1160161	963494	7757	112978	22722	22193	26553
LR9198	I	Complex	China	1253170	1039361	9778	121486	24483	23565	29609
BTx642	L	Durra	Ethiopia	1524769	1287876	12862	132322	27541	26021	32749
SC35C-14E	L	Durra	Ethiopia	1228814	1028072	7766	115689	23108	22143	27565
SC23-14E	L	Durra	Ethiopia	1362098	1146377	9130	123135	24986	23680	29949
Yik.solate	L	Durra	Ethiopia	1118066	933012	5181	116030	17059	15380	27540
IBC/E-38432	L	Durra	Ethiopia	1715354	1430193	11247	167795	30353	29061	40485
PI585749	L	Durra-Bicolor	Mali	1446371	1210097	11321	133531	27449	25917	32590
PI330272	W	Drummondii	Ethiopia	1501312	1242394	10899	147465	30448	29048	35194
Zengada	W	Weedy	Ethiopia	1581684	1315478	10824	155882	28624	27247	37720
Kilo	W	Weedy	Ethiopia	1267760	1047467	5627	137344	21473	19909	31449
Greenleaf	W	Weedy	USA	1522107	1268287	10468	145993	29247	28130	34204
PI226096	W	Weedy	Kenya	1956801	1641444	16255	179268	35250	33730	43508
PI525695	M	Margaritiferum	Mali	1964025	1628455	12730	197292	36202	35569	46286
PI586430	M	Margaritiferum	Sierra Leone	1938008	1594348	13766	198477	38894	38431	46271
PI300119	W	Verticilliflorum	South Africa	2995879	2482294	26648	290919	56213	56617	71315
AusTRCF317961	W	Verticilliflorum	Australia	2003360	1625419	12596	226288	38953	39283	52566
Sorpr369-1	P	Propinquum	-	5200279	3971685	58105	713492	124517	141591	163430
Sorpr369-2	P	Propinquum	-	4993948	3794524	53315	704812	118631	135432	160696

*I - Improved variety; L - Landrace; W - Wild/Weedy; M - Margaritiferum; P - *Sorghum propinquum*

are also included. Additionally, the total number of SNPs identified per sample is indicated. The two *margaritiferum* cultivars (PI525695 M *Margaritiferum Mali 1964025* and PI586430 M *Margaritiferum Sierra Leone 1938008*) are separated into a distinct group since they are highly divergent from other *S. bicolor* races (Figure 5.1). Two samples of the allopatric Asian species *Sorghum. propinquum* are clustered within a distant group as the outgroup. The SNP numbers of each sample give an overview of the genomic difference between the reference genome BTx623 and individual genomes. Detailed information about distribution of SNPs in different genomic regions, including genic, intergenic, and intronic regions is provided (Table 5.1). For genic regions, SNPs

Table 5.2 Distribution of major effect SNPs in different genic sites and regions in 48 sorghum accessions

Genotype	Type	Racial type	Geographic	Start Codon			Stop Codon			Splice sites		
			Origin	Gain	Lost	Variant	Gain	Lost	Retain	Donor	Acceptor	Region
BTx623	I	Complex	USA	0	0	0	0	0	0	0	0	0
Malisor84-7	I	Complex	Mali	380	16	1	130	39	10	28	32	742
ICSV745	I	Complex	India	1084	45	9	246	64	21	69	74	1643
EarlyHegari	L	Caudatum	Sudan	1281	35	12	282	81	18	69	69	1960
Cherekit	L	Caudatum	Ethiopia	1283	42	12	291	68	18	66	72	2011
SC103-14E	L	Guinea-Caudatum	South Africa	1014	39	8	244	72	20	51	75	1763
Macia	I	Caudatum	Mozambique	996	38	5	239	62	26	51	60	1661
SC108-14E	L	Caudatum	Ethiopia	1054	42	9	230	60	20	51	68	1604
SC237-14E	L	Caudatum	Sudan	1242	50	6	282	79	22	57	72	1807
B923296	I	Complex	Australia	770	26	9	199	53	19	53	72	1395
B963676	I	Complex	Australia	962	30	5	192	62	18	57	64	1510
M35-1	L	Durra	India	1048	39	5	246	68	21	66	74	1646
R931945-2-2	I	Complex	Australia	1703	74	15	331	81	29	88	110	2042
SC170-6-8	L	Caudatum	Ethiopia	1065	44	5	245	69	27	55	58	1631
SC326-6	L	Caudatum-Bicolor	Ethiopia-USA	1220	34	10	266	82	26	62	96	1831
SC56-14E	L	Caudatum-Nigricans	Sudan	1433	40	10	328	78	24	58	80	1946
SC62-14E	L	Complex	Kenya	1455	38	10	295	99	28	71	94	2250
E-Tian	I	Kafir	China	1430	41	14	228	65	25	57	66	1085
Rio	I	Kafir	USA	1273	47	16	259	72	21	65	75	1899
SS79	I	Kafir	Ethiopia	2602	78	21	419	116	37	102	130	2384
IS8525	L	Kafir	Ethiopia	1353	46	17	261	69	24	64	87	1858
Keller	I	Complex	USA	750	27	7	212	48	12	36	45	1302
RTx7000	I	Kafir-Caudatum	USA	1605	47	15	284	93	32	78	97	1948
IS3614-2	L	Guinea	Nigeria	1421	48	13	358	95	37	83	112	2586
Karper669	I	Complex	USA-Sudan	1362	52	11	301	93	24	63	94	2209
PI563516	I	Durra-Caudatum	Mali	1427	53	13	298	86	27	62	94	2032
QL12	I	Complex	Australia	1321	45	10	313	79	25	64	98	2097
IS9710	L	Caudatum	Sudan	1265	38	10	301	85	21	73	78	2161
KS115	I	Durra-Caudatum	USA	900	35	12	270	77	33	71	89	2111
Ji2731	L	Caudatum	China	1666	52	13	265	76	23	74	62	1474
AI4	I	Complex	China	1416	45	12	291	78	24	90	99	2409
LR9198	I	Complex	China	1735	47	10	331	95	27	103	107	2433
BTx642	L	Durra	Ethiopia	2114	75	23	363	107	38	93	99	2486
SC35C-14E	L	Durra	Ethiopia	1402	48	16	317	89	32	84	94	2389
SC23-14E	L	Durra	Ethiopia	1587	55	14	384	107	31	87	108	2468
Yik.solate	L	Durra	Ethiopia	990	25	8	249	67	28	79	90	2328
IBC/E-38432	L	Durra	Ethiopia	1965	70	14	442	113	45	108	121	3342
PI585749	L	Durra-Bicolor	Mali	1930	65	17	388	109	43	95	130	2689
PI330272	W	Drummondii	Ethiopia	1865	56	11	458	123	49	100	148	3054
Zengada	W	Weedy	Ethiopia	1864	59	13	413	111	45	95	147	3162
Kilo	W	Weedy	Ethiopia	1058	35	5	294	73	32	85	106	2803
Greenleaf	W	Weedy	USA	1871	60	16	411	122	34	110	116	3038
PI226096	W	Weedy	Kenya	2767	76	16	495	145	48	148	148	3503
PI525695	M	Margaritiferum	Mali	2318	73	15	524	135	46	136	162	4082
PI586430	M	Margaritiferum	Sierra Leone	2525	82	15	562	144	47	138	175	4133
PI300119	W	Verticilliflorum	South Africa	4441	132	29	786	204	90	211	224	5756
AusTRCF317961	W	Verticilliflorum	Australia	2278	80	16	521	145	53	163	185	4814
Sorpr369-1	P	Propinquum	-	9859	249	42	1519	378	236	407	481	14288
Sorpr369-2	P	Propinquum	-	9169	241	41	1437	359	240	405	465	14181

*I - Improved variety; L - Landrace; W - Wild/Weedy; M - *Margaritiferum*; P - *Sorghum propinquum*

found in specific positions such as start and stop codons, splice donator and acceptor sites are listed (Table 5.2).

All the SNP data shown in the two tables can be easily accessed either as statistical information through the Help page of the database, or through the user interface. The original data of sequencing short reads, the assembled sequence and the SNP data of each accession can be downloaded.

5.2.2 User interface

SorGSD offers three main functions (search, compare and browse), for users to search, display and retrieve the SNPs and their annotations.

The search function provides a user-friendly web interface to query SNP information. Users can search SNPs by specifying chromosomal co-ordinates or the locus ID. Users can also query SNPs based on their genotypes, and predicted variant effects. In addition, users can compare the SNPs between two and more sorghum lines. The query results can be shown as a formatted table which contains the information of ID, chromosome position, genomic location and predicted coding effects, 5' and 3' flanking sequences, reference and derived alleles, respectively. SNPs from the stringent set identified by both pipelines (see description in "Methods" and Figure 5.2 for details) are highlighted with a green background in the result page. The output of the query results can be downloaded as flat text or formatted tables for further investigation.

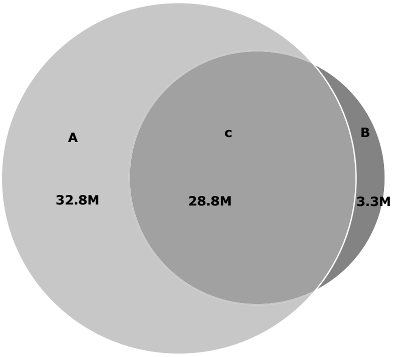


Figure 5.2 **Venn diagram of SNPs identified by two pipelines.** A. SNPs called by the GATK-based pipeline. B. SNPs called by the SOAPsnp- and realSFS-based pipeline. C. The set of highly reliable SNPs as identified by both pipelines.

SorGSD also provides several data browsing functionalities under the "Browse" pull-down menu. The "Total SNPs" tab lists the SNP numbers on 10 chromosomes of all 48 accessions. Users can select a group, e.g., Landraces, to display the SNP numbers of these accessions within this group. Mouse-clicking these SNP numbers will bring up the list of SNPs of a specific accession. Given that the different location in genes such as coding regions, as well as the non-synonymous information are often of great interest for further study, the "Genic SNP" tab lists several submenus including "Coding SNP", "Synonymous SNP", and "Non-synonymous SNP" so that information can be tailored to user requirements.

The "Browse on Chromosome" tab leads to an interactive graphic window to visualize SNPs in a genome browser. Users can customize the visualization interface by selecting different data types, including SNPs, genes, transcripts, allele frequencies, and the SNP density information. Users can obtain a pie chart showing the allele frequency, SNP density in 300 kb windows size, related gene and transcript information.

5.2.3 Help information

SorGSD provides a help resource for users to better access the SNP data, as well as proving links to additional sorghum research related resources.

The help menu provides a "How to" page, which gives a number of examples for users to learn how to search and compare target SNPs. For example, a step-by-step user-guide shows how to obtain non-synonymous SNPs in chromosome 1 of sweet sorghum E-Tian, and how to compare SNPs between sweet sorghum E-Tian and two grain sorghum Ji2731 and Keller. A FAQ page provides answers to a range of frequently asked questions, not only about the content and usage of SorGSD but more broadly about sorghum genomics. Detailed information including software tools, parameters and data sources is presented in the "Pipeline" page. The "Statistics" page shows the SNP numbers distributed in different genomic regions (Table 5.1) and specific genic sites (Table 5.2). The "Data source" page shows the general information of 48 sorghum lines, including their geographic origins, and links to the US Germplasm Resources Information Network (http://www.ars-grin.gov).

The "About" tab contains several pages related to sorghum research. The Sorghum Genome page provides a brief introduction to the reference genome BTx623, including genome size and gene number. The Resource page provides links to online databases, research institutions, sorghum producers and handbooks. The reference page lists selected recently published papers in the fields of sorghum genomics, genetics, QTLs, etc., with links to full lists in PubMed.

5.3 Conclusions and future directions

High coverage resequencing data from two previous sorghum studies (15,16) were used to identify SNPs among 48 sorghum genotypes by combining three SNP calling tools and updating the SNPs datasets using the sorghum reference annotation (Version 2.1). In addition, we annotated the effect of SNP variants on genes of each sorghum accession. SorGSD has already received over two thousand visits from more than 30 countries around the world since it went online a few months ago. During the review process of this manuscript, we were happy to learn that a new website Sorghum Genomics (https://www.purdue.edu/sorghumgenomics) developed at Purdue University became available as a functional gene discovery platform.

We will improve the SNP calling pipeline and the annotation procedure to obtain more accurate SNP data and upload them into the database. Furthermore, we will include additional types of genome variation data detected by newly developed pipelines, including indels and copy number variants (CNVs). At the same time, we will improve the web interface especially in the search function and give more examples in the user guide to help novice users to access the database easily. We will add more analytical functionalities so that users can perform more analyses such as Blast search, sequence alignment and phylogenetic analysis.

SorGSD can serve as a bioinformatics platform to inform wet-lab experiments including biomarker development, allele mining and gene function assessment. In addition to the collaboration among research groups involving in this work, we will collaborate with other domestic and international laboratories in the sorghum research community to sequence and annotate more sorghum accessions in the future.

We will update the database regularly and add SNP datasets with newly available re-sequenced sorghum accessions. We hope that the high density of these SNP data at genomic level collected from the major races of cultivated sorghum as well as other subspecies is a rich repository for a broader research community working in biomarker identification, genetic analysis and molecular breeding, especially for energy plant sweet sorghum cultivation.

5.4 Methods

The construction of SorGSD was a multi-step process. Firstly, the sorghum re-sequencing paired-end raw reads reported in the previously published works were downloaded (15,16). In addition, the paired-end raw reads generated in-house for a sweet sorghum line SS79 were included [unpublished data]. Secondly, the raw reads were mapped to the reference sorghum genome (BTx623) (10) using the BWA program (26). SNPs were identified using the software GATK (27,28), realSFS (http://popgen.dk/angsd/index.php/RealSFS) and SOAPsnp (29) and annotated using SnpEff (30). With the SNP matrix finalized, a web interface was designed for users to browse and search the SNPs and related annotations. Details for the database construction are described below and are also available on the designated website.

5.4.1 Data source

The raw sequencing reads originated from three original datasets. The largest dataset (16) contains 44 sorghum accessions representing the major races of cultivated sorghum as well as their wild relatives. The second dataset (31) contains three accessions of cultivated sorghums. The raw reads of these two datasets can be downloaded from the NCBI sequence read archive (SRA) (accessions SRS378430-SRS378473, and accessions SRX100115-SRX100138). The third dataset contains the paired-end reads of sorghum line SS79, a cultivated sweet sorghum inbred. These data were recently generated in our laboratory using an Illumina HiSeq 2000 platform with insert size of 500 bp (accessible from ftp://download.big.ac.cn/SorghumVB/sra). The average sequencing depth of all sorghum accessions is about 20×, ranging from 12× to 54×.

5.4.2 SNP calling pipeline

After trimming adapters, the clean reads were mapped to version 2.1 of the (http://phytozome.jgi.doe.gov/pz/portal.html#!info?alias=Org_Sbicolor) reference genome using the BWA program (26), allowing a maximum of five mismatches and disabling long gaps in the mapping procedure. The average mapping rate, unique mapping rate and mapping coverage were 95.7%, 68.1% and 88.1% respectively, excluding the two *Sorghum propinquum* accessions. The SAM tools package (32) was used to convert mapping results to BAM format, and then the Picard program (http://picard.sourceforge.net) was applied to eliminate duplicated reads generated during the process of library construction.

Subsequently, the GATK tools (27,28) were used to recalibrate the base quality score to obtain more accurate quality scores for each base and realign reads around known indels. The refined data from all individuals were jointly used to call a raw SNPs set by GATK HaplotypeCaller. Finally, a set of SNPs were identified, using the variant quality score to recalibrate the procedure in GATK. In total, we identified 62,888,582 SNPs across all 48 sorghum lines, corresponding to 15,357,261 sites in the reference genome. The GATK based SNP calling pipeline is similar to that reported in a recent publication (33). SNPs were additionally identified using the pipeline described previously using realSFS (34) and SOAPsnp (29) described by Mace et al. (16). Approximately 28 million highly stringent SNPs were in common between the two SNP identification pipelines (Figure 5.2) with the GATK-based pipeline identifying more SNPs than the SOAPsnp-based pipeline. The total number of SNPs called by the GATK based pipeline was found to be comparable to the study by Evans et al. (35), which employed the CLC Workbench software (CLC Bio-Qiagen, Aarhus, Denmark). All the SNPs identified by the GATK pipeline were stored in SorGSD, with the subset of 28 million highly stringent SNPs highlighted in the results page. Finally, the effect of variants on all the v2.1 predicted gene models for each sorghum accession were predicted and annotated using the SnpEff program (version 4.0e) (30).

5.4.3 Database implementation

The SNP data and their related annotations were formatted into tables and stored in SorGSD using the MySQL database management system (version 5). The web interface of SorGSD was designed by JAVA/JSP (JDK 1.6) under the Apache/Tomcat web server (version 2.0) running under a Linux operation system (CentOS 6). We installed the generic genome browser GBrowse (36) as a chromosome-based visualization tool to display these genomic SNPs and annotations.

References

1. Doggett, H. (1967) Yield Increase from Sorghum Hybrids. *Nature*, **216**, 798-799.
2. Pennisi, E. (2009) How Sorghum Withstands Heat and Drought. *Science*, **323**, 573-573.
3. Osborne, C.P. and Beerling, D.J. (2006) Nature's green revolution: the remarkable evolutionary rise of C_4 plants. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **361**, 173-194.
4. Sasaki, T. and Antonio, B.A. (2009) Sorghum in sequence. *Nature*, **457**, 547-548.
5. Rooney, W.L., Blumenthal, J., Bean, B. and Mullet, J.E. (2007) Designing sorghum as a dedicated bioenergy feedstock. *Biofuels, Bioproducts and Biorefining*, **1**, 147-157.
6. Carpita, N.C. and McCann, M.C. (2008) Maize and sorghum: genetic resources for bioenergy grasses. *Trends in Plant Science*, **13**, 415-420.
7. Vermerris, W. (2011) Survey of Genomics Approaches to Improve Bioenergy Traits in Maize, Sorghum and Sugarcane. *Journal of Integrative Plant Biology*, **53**, 105-119.

8. Calviño, M. and Messing, J. (2012) Sweet sorghum as a model system for bioenergy crops. *Current Opinion in Biotechnology*, **23**, 323-329.

9. Mullet, J., Morishige, D., McCormick, R., Truong, S., Hilley, J., McKinley, B., Anderson, R., Olson, S.N. and Rooney, W. (2014) Energy Sorghum—a genetic model for the design of C4 grass bioenergy crops. *Journal of Experimental Botany*, **65**, 3479-3489.

10. Paterson, A.H., Bowers, J.E., Bruggmann, R., Dubchak, I., Grimwood, J., Gundlach, H., Haberer, G., Hellsten, U., Mitros, T., Poliakov, A. *et al.* (2009) The Sorghum bicolor genome and the diversification of grasses. *Nature*, **457**, 551-556.

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

12. Davey, J.W., Hohenlohe, P.A., Etter, P.D., Boone, J.Q., Catchen, J.M. and Blaxter, M.L. (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics*, **12**, 499-510.

13. Wang, S., Meyer, E., McKay, J.K. and Matz, M.V. (2012) 2b-RAD: a simple and flexible method for genome-wide genotyping. *Nature Methods*, **9**, 808-810.

14. Nelson, J.C., Wang, S., Wu, Y., Li, X., Antony, G., White, F.F. and Yu, J. (2011) Single-nucleotide polymorphism discovery by high-throughput sequencing in sorghum. *BMC Genomics*, **12**, 352.

15. Zheng, L.-Y., Guo, X.-S., He, B., Sun, L.-J., Peng, Y., Dong, S.-S., Liu, T.-F., Jiang, S., Ramachandran, S., Liu, C.-M. *et al.* (2011) Genome-wide patterns of genetic variation in sweet and grain sorghum (Sorghum bicolor). *Genome Biology*, **12**, R114.

16. Mace, E.S., Tai, S., Gilding, E.K., Li, Y., Prentis, P.J., Bian, L., Campbell, B.C., Hu, W., Innes, D.J., Han, X. *et al.* (2013) Whole-genome sequencing reveals untapped genetic potential in Africa's indigenous cereal crop sorghum. *Nature Communication*, **4**, 2320.

17. Bekele, W.A., Wieckhorst, S., Friedt, W. and Snowdon, R.J. (2013) High-throughput genomics in sorghum: from whole-genome resequencing to a SNP screening array. *Plant Biotechnology Journal*, **11**, 1112-1125.

18. Morris, G.P., Ramu, P., Deshpande, S.P., Hash, C.T., Shah, T., Upadhyaya, H.D., Riera-Lizarazu, O., Brown, P.J., Acharya, C.B., Mitchell, S.E. *et al.* (2013) Population genomic and genome-wide association studies of agroclimatic traits in sorghum. *Proceedings of the National Academy of Sciences of the United States of America*, **110**, 453-458.

19. Nielsen, R., Paul, J.S., Albrechtsen, A. and Song, Y.S. (2011) Genotype and SNP calling from next-generation sequencing data. *Nature Reviews Genetics*, **12**, 443-451.

20. Spindel, J., Wright, M., Chen, C., Cobb, J., Gage, J., Harrington, S., Lorieux, M., Ahmadi, N. and McCouch, S. (2013) Bridging the genotyping gap: using genotyping by sequencing (GBS) to add high-density SNP markers and new value to traditional bi-parental mapping and breeding populations. *Theoretical and Applied Genetics*, **126**, 2699-2716.

21. Morishige, D.T., Klein, P.E., Hilley, J.L., Sahraeian, S.M.E., Sharma, A. and Mullet, J.E. (2013) Digital genotyping of sorghum – a diverse plant species with a large repeat-rich genome. *BMC Genomics*, **14**, 448.

22. Han, Y., Lv, P., Hou, S., Li, S., Ji, G., Ma, X., Du, R. and Liu, G. (2015) Combining Next Generation Sequencing with Bulked Segregant Analysis to Fine Map a Stem Moisture Locus in Sorghum (Sorghum bicolor L. Moench). *PLoS One*, **10**, e0127065.

23. Rhodes, D.H., Hoffmann, L., Rooney, W.L., Ramu, P., Morris, G.P. and Kresovich, S. (2014) Genome-Wide Association Study of Grain Polyphenol Concentrations in Global Sorghum [Sorghum bicolor (L.) Moench] Germplasm. *Journal of Agricultural and Food Chemistry*, **62**, 10916-10927.

24. Adeyanju, A., Little, C., Yu, J. and Tesso, T. (2015) Genome-Wide Association Study on Resistance to Stalk Rot Diseases in Grain Sorghum. *G3: Genes|Genomes|Genetics*, **5**, 1165-1175.

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

26. Li, H. and Durbin, R. (2010) Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics*, **26**, 589-595.

27. DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., del Angel, G., Rivas, M.A., Hanna, M. *et al.* (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, **43**, 491-498.

28. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M. *et al.* (2010) The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, **20**, 1297-1303.

29. Li, R., Li, Y., Fang, X., Yang, H., Wang, J., Kristiansen, K. and Wang, J. (2009) SNP detection for massively parallel whole-genome resequencing. *Genome Research*, **19**, 1124-1132.

30. Cingolani, P., Platts, A., Wang, L.L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X. and Ruden, D.M. (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly*, **6**, 80-92.

31. Zheng, L.Y., Guo, X.S., He, B., Sun, L.J., Peng, Y., Dong, S.S., Liu, T.F., Jiang, S., Ramachandran, S., Liu, C.M. *et al.* (2011) Genome-wide patterns of genetic variation in sweet and grain sorghum (Sorghum bicolor). *Genome Biology*, **12**, R114.

32. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. and Subgroup, G.P.D.P. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078-2079.

33. McCormick, R.F., Truong, S.K. and Mullet, J.E. (2015) RIG: Recalibration and Interrelation of Genomic Sequence Data with the GATK. *G3: Genes|Genomes|Genetics*, **5**, 655-665.

34. Korneliussen, T.S., Albrechtsen, A. and Nielsen, R. (2014) ANGSD: Analysis of Next Generation Sequencing Data. *BMC Bioinformatics*, **15**, 356.

35. Evans, J., McCormick, R.F., Morishige, D., Olson, S.N., Weers, B., Hilley, J., Klein, P., Rooney, W. and Mullet, J. (2013) Extensive variation in the density and distribution of DNA polymorphism in sorghum genomes. *PLoS One*, **8**, e79192.

36. Stein, L.D., Mungall, C., Shu, S., Caudy, M., Mangone, M., Day, A., Nickerson, E., Stajich, J.E., Harris, T.W., Arva, A. *et al.* (2002) The Generic Genome Browser: A Building Block for a Model Organism System Database. *Genome Research*, **12**, 1599-1610.

CHAPTER 6

6

Sweet sorghum originated through selection of *Dry*, a plant-specific NAC transcription factor gene

Li-Min Zhang[1], Chuan-Yuan Leng[1], Hong Luo[1], Xiao-Yuan Wu, Zhi-Quan Liu, Yu-Miao Zhang, Hong Zhang, Yan Xia, Li Shang, Chun-Ming Liu, Dong-Yun Hao, Yi-Hua Zhou, Cheng-Cai Chu, Hong-Wei Cai, Hai-Chun Jing

[1] First author

Based on: The Plant Cell, Oct 2018, 30 (10) 2286-2307

Abstract

Sorghum (*Sorghum bicolor*) is the fifth most popular crop worldwide and a C4 model plant. Domesticated sorghum comes in many forms, including sweet cultivars with juicy stems and grain sorghum with dry, pithy stems at maturity. The *Dry* locus, which controls the pithy/juicy stem trait, was discovered over a century ago. Here, we found that *Dry* gene encodes a plant-specific NAC transcription factor. *Dry* was either deleted or acquired loss-of-function mutations in sweet sorghum, resulting in cell collapse and altered secondary cell wall composition in the stem. Twenty-three *Dry* ancestral haplotypes, all with dry, pithy stems, were found among wild sorghum and wild sorghum relatives. Two of the haplotypes were detected in domesticated landraces, with four additional dry haplotypes with juicy stems detected in improved lines. These results imply that selection for *Dry* gene mutations was a major step leading to the origin of sweet sorghum. The *Dry* gene is conserved in major cereals; fine-tuning its regulatory network could provide a molecular tool to control crop stem texture.

6.1 Introduction

Sorghum (*Sorghum bicolor*) is the fifth most important cereal crop after wheat (*Triticum aestivum*), rice (*Oryza sativa*), maize (*Zea mays*), and barley (*Hordeum vulgare*). This crop has numerous advantages, including remarkable stress tolerance and high photosynthetic efficiency, and is widely grown in arid or semi-arid areas. Similar to common grain sorghum, sweet sorghum is widely found in many geographical regions, with various landraces, and is considered to be an ideal biofuel crop for first- and second-generation bioethanol production. Sorghum can produce high biomass on marginal lands that are not suitable for food or feed production (1).

In vascular plants, stems/shoots have evolved as an important link between roots and reproductive organs by providing strong support and efficient water and nutrient transport. Since the driving force for survival and plant breeding is to maximize reproductive success and yield, most crops, especially cereals, often have dry, pithy, and sometimes hollow stems/shoots at maturity. One remarkable feature of sweet sorghum is that its stems accumulate high amounts of juice and directly fermentable sugars at maturity, in contrast to grain sorghum, which often has dry, pithy stems (2). Hence, sweet sorghum provides an interesting model to examine the roles of stem/shoots in whole plant water transport and carbon partitioning in cereals. The sweet sorghum model is also useful for comparative genomics for plants with stems as the primary harvest targets, such as sugarcane (*Saccharum. officinarum*; (3)). Since sweet sorghum exhibits large variations in stem water content, this crop could be a useful model to study stem water transport by examining within-species genomic variation and hence may provide a new angle to complement similar studies in *Arabidopsis thaliana*, rice, and the moss *Physcomitrella patens* (4,5). Identifying the genes controlling stem juiciness and the molecular process underlying how sweet sorghum evolved is a first key step in this analysis.

Classical genetic studies have shown that sorghum stem juiciness is governed by a single locus, named *Dry* or *D*, and that dry stem trait is dominant over the juice-rich stem trait (6-8). Previous quantitative trait locus analyses have localized the *Dry* locus between RFLP markers umc34 and txs1030 on Chromosome 6 (9); this locus is also tightly linked to marker loci txp97 (10) and txp145 (11). Further fine mapping efforts have narrowed down the locus to the interval between SSR markers sm06068 and Sb6-2, which harbors only six candidate genes (12), which is consistent with the findings obtained by bulked segregant analysis and deep sequencing analyses (13). Such classical quantitative trait locus mapping results have been further confirmed by genome-wide association study (GWAS) analysis, which identified a major quantitative trait locus for midrib color, sugar yield, juice volume, and moisture at ~51.8 Mb on Chromosome 6 (14). This raises the interesting question of whether the *Dry* locus is an important locus for the origin of sweet sorghum. So far, the molecular nature of the *Dry* locus remains elusive; therefore, it is not possible to examine its selection history during domestication and breeding. This is partially hindered by the continuous variation in the amount of extractible juice among juicy genotypes and among the offspring of crosses between juicy and pithy or juicy and juicy lines (15). Sweet sorghum can be found among different landraces. Previous work combining morphological observations and molecular markers demonstrates that sweet sorghum is of polyphyletic origin (more than one ancestor) and cannot be distinctively separated from grain sorghum lines (16-18). Together, such complications require a new approach to dissect the molecular basis of the origin of sweet sorghum.

We hypothesize that the *Dry* locus is an important selection target that led to the development of sweet sorghum and initiated work on the cloning and characterization of the *Dry* gene using a population genomics approach. We show here that the *Dry* gene encodes a plant-specific NAC transcription factor and that stem juiciness results from loss of function of the *Dry* gene. We provide evidence suggesting that the *Dry* gene is an important first-layer master switch for secondary cell wall biosynthesis. Our findings suggest that sweet sorghum originated as a consequence of intensive breeding selection of the *Dry* locus.

6.2 Results

6.2.1 Variation in the stem juiciness trait in a diverse panel of 241 sorghum lines

In our previous work comparing the genomic and phenotypic variation between sweet and grain sorghum, we noticed remarkable variation in the stem juice content among four different sorghum lines (19) (Figure 6.1A). This prompted us to further examine this trait in detail using a diverse panel of 241 sorghum lines assembled over the years, which includes one accession of *Sorghum propinquum*, a close relative of sorghum, 42 wild sorghums, and 198 domesticated/cultivated sorghums (Supplemental Data Set 1). At maturity, the water content for wild sorghum and the dry, pithy sorghum lines ranged from 55 to 63% and 46 to 64%, respectively, in contrast to that of the juicy sorghums, which ranged from 66 to 83% (Figure 6.1B).

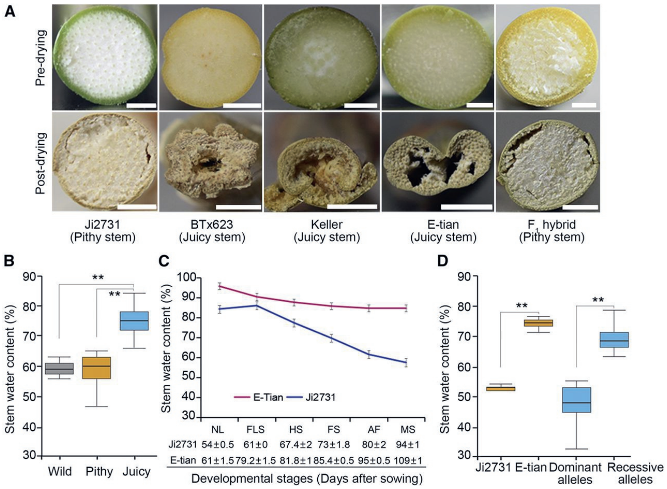


Figure 6.1. **Characterization of pithy/juicy sorghum stems.** (A) Cross sections of the middle internodes of sorghum stems at maturity before (top) and after drying (bottom) from different genetic materials (Chinese kaoliang Ji2731, BTx623, sweet sorghum cultivars Keller and E-tian, and an F1 hybrid generated from a cross between Ji2731 and E-tian). Bars = 0.5 cm. (B) Box plot of stem water content for the 241 sorghum lines examined in this study at maturity, including one *Sorghum propinquum*, 42 wild, 86 dry, pithy sorghums, and 112 juice-rich sorghum lines. The value of stem water content for each sorghum line is the average of more than three individual plants (3-5 plants for 56 sorghum lines, 6-10 plants for 123 sorghum lines, more than 10 plants for 62 sorghum lines). (C) Changes in stem water content during development for Ji2731 and E-tian. The observations were performed in Beijing in 2015 and monitored for developmental stages ranging from NL (emergence of the ninth leaves) to FLS (flag leaf stage), HS (heading stage), FS (flowering stage), AF (1 week after flowering), and MS (milk stage). Values are the means ± se of three individual plants per sorghum line at each developmental stage. For developmental stages, days after sowing were scored for five individual plants and are presented as means ± se. (D) Box plot of stem water content for the two parental lines (each 10 plants) and their F3 RIL lines (174 dominant alleles; 176 recessive alleles) at maturity. Asterisks in (B) and (D) represent significant differences determined by two-tailed Student’s t test at P < 0.01 (Supplemental File 7).

To examine the development and genetics of the trait, we took advantage of a unique feature of Chinese kaoliang sorghum (e.g., Ji2731), which normally has a complete dry, pithy stem at maturity, unlike sweet and juice-rich sorghums (e.g., E-tian, Keller, and BTx623) (Figure 6.1A). In *kaoliang* sorghum, the water loss process normally began at the jointing stage (NL, the emergence of the ninth leaf) and accelerated after flowering, while the water content in sweet sorghum stems remained at relatively high levels throughout development (Figure 6.1C). We constructed four F2 populations by crossing Chinese *kaoliang* sorghum Ji2731 with juicy stem sorghum lines E-tian, Keller, and BTx623, respectively. The F1 plants showed the same complete dry, pithy stems as kaoliang Ji2731 (Figure 6.1A), and the F2 individuals showed a 3:1 segregation ratio for the dry/pithy:juicy/sweet stem trait. In the recombinant inbred lines (RILs) that were subsequently developed from the Ji2731/E-tian crosses, the stem water content at maturity in RILs with recessive alleles was 63 to 78%, while that of RILs with dominant alleles was 33 to 55% (Figure 6.1D; Supplemental Table 1). Hence, our results suggest that a single genetic locus with a completely dominant effect of dry/pithy over juicy/sweet explains this trait.

6.2.2 Genome resequencing and population genomics

To facilitate the identification of the gene underlying the stem juiciness trait, we performed genome resequencing of the 241 sorghum accessions. Whole-genome resequencing generated a total of ~7.39 G paired-end reads 150 bp in length (~1.11 Tb in total), obtaining an average sequencing depth of ~5.67× and an average genome coverage of ~89.6% (Supplemental Data Sets 2 and 3). After mapping to the sorghum reference genome BTx623 (the whole genome was sequenced as a sorghum reference) (20) and single-nucleotide polymorphism (SNP) calling, we obtained 31,946,640 high-quality SNPs and 4,266,768 indels (insertions-deletions) from the 241 sorghum accessions (Supplemental Data Sets 2 and 3). Users can download all of the variation data from the sorghum genome variation database (SorGSD) (21).

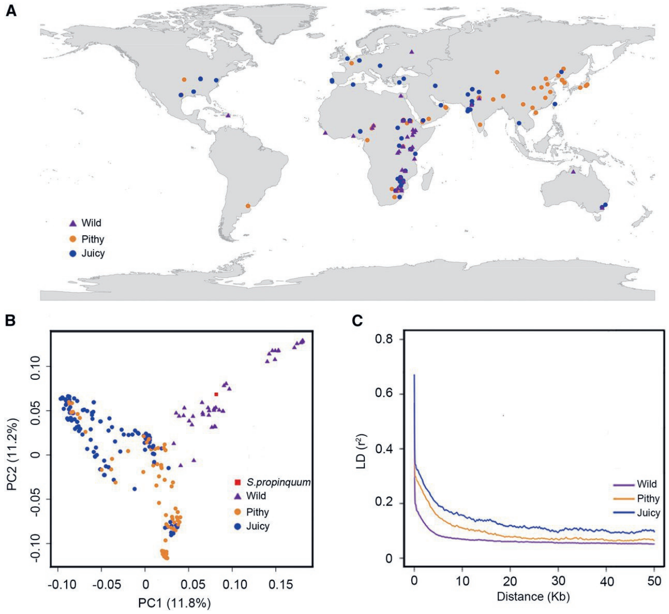


Figure 6.2. **Geographical distribution and genetic analysis of 241 sorghum lines.** (A) Map showing the geographic origin of the accessions used in this study. Different symbols indicate different subgroups. The map contains 210 accessions with known location information obtained from Germplasm Resources Information Network (GRIN; https://www.ars-grin.gov/) and Chinese Crop Germplasm Resources Information System (CGRIS; http://www.cgris.net/); the 31 accessions without location information are not shown. For accessions with only information about the country of collection, the collection sites are depicted in the capitals of the countries; for accessions in which the collection province could be obtained, the collection sites in the provincial capitals are depicted. See also Supplemental Data Set 1. (B) PCA plot of the 241 sorghum lines using 499,130 SNPs filtered with MAF 0.05, max missing rate 0.5, and LD (r^2) 0.2. The wild subgroup could clearly be separated from the pithy and juicy subgroups, while the pithy and juicy subgroups are interconnected. Numbers in brackets denote the variance explained by the first and second principal component. (C) LD decay plots for the three subgroups of sorghum. LD is lower in the wild subgroup compared with pithy sorghums but higher in juicy compared with pithy sorghums. See also Supplemental Data Sets 2 and 3.

Figure 6.2A shows the geographic distribution of the panel, which indicates that the panel has captured a fairly broad representation of the genetic diversity of sorghum. Principal component analysis (PCA) based on SNP data showed that *Sorghum propinquum* and all wild sorghums were clustered together and distinctly separated from the domesticated/cultivated sorghums (Figure 6.2B); among cultivated sorghums, those with dry, pithy stems could largely be separated from those with juicy stems, although there were a few exceptions in both groups. The PCA results, together with the observation that a large proportion of the juicy sorghums are inbred lines and the pithy sorghums are primarily landraces in this collection (Supplemental Data Set 1), imply that the juiciness trait has multiple independent origins among cultivated sorghums with different backgrounds. This notion is supported by the results of linkage disequilibrium (LD) decay analysis (Figure 6.2C), which also confirmed that wild, cultivated juicy, and cultivated dry, pithy sorghums had obviously different decay rates and fell into distinct subgroups.

6.2.3 Identification of the Dry gene by GWAS and map-based cloning

Since our genetic analysis demonstrated that only a single genetic locus controls the pithy/juicy stem trait, we scored the dry pithy stem trait as 1 and the juicy stem trait as 0 and used the SNPs with a minor allele frequency (MAF) ≥ 0.05 for our GWAS. Under the mixed linear model (MLM) with the PCA and familial kinship (K), we detected 79 genetic variants exceeding the significance threshold [$-\log_{10}(P) = 8.06$] (Figure 6.3A). All 79 genetic variants reside on Chromosome 6 within a 290-kb region ranging from 50,613,891 to 50,904,567 bp, and the most significant locus is at 50,893,225 bp [$-\log_{10}(P) = 12.6$]. Further analysis showed that except for seven variants, the majority of the genetic variants were narrowed down to a 15-kb region from 50,890,065 to 50,904,567 bp on Chromosome 6, harboring only one predicted gene (*Sobic.006G147400*) (Supplemental Figure 1; all the supplemental figures and tables can be found on https://academic.oup.com/plcell/article/30/10/2286/6099342)

In parallel, we performed classic map-based cloning using segregation populations derived from crosses between Chinese *kaoliang* sorghum Ji2731 and juicy stem sorghum E-tian, which narrowed down the locus to a 45-kb region from 50,880,559 to 50,925,294 bp on Chromosome 6. This region harbors four candidate genes, coincident with the GWAS mapped region (Figure 6.3C). This region was previously reported to contain a genetic locus named *D* or *Dry*, with a well-matched model of genetic action that the pithy stem trait was dominant over the juicy stem trait and that the pithy/juicy phenotype was controlled by a single locus (6). Hence, both our GWAS and map-based cloning results confirmed that we rediscovered the important *Dry* locus controlling the pithy/juicy trait.

In subsequent work to clone the *Dry* gene, we combined the results from both the GWAS significant region and the map-based cloning interval, and, to be on the safe side, included the four candidate genes in the vicinity of the *Dry* locus for detailed sequence comparisons and expression analysis. The four genes include three hypothetical genes (*Sobic.006G147350*, *Sobic.006G147450*, and *Sobic.006G147500*) and a gene predicted to encode a NAC transcription factor (*Sobic.006G147400*) (Figure 6.3B). A comparison of the genomic sequences between the candidate regions of the parents Ji2731 and E-tian revealed no differences in the coding regions of these three hypothetical genes. For *Sobic.006G147400*, we successfully amplified the full-length NAC gene in Ji2731, while a 14-kb region containing 3.6 kb of the *Dry* genomic sequence (including the first two exons) could not be amplified in E-tian or any recessive allele tested (i.e., RILs carrying the *dry* allele in a mixed Ji2731 and E-tian genetic background) (Figure 6.3B; Supplemental Figure 2A). Furthermore, only the NAC transcription factor gene displayed the expected differential expression patterns between Ji2731/the dominant allele and E-tian/the recessive allele (Supplemental Figure 2B). Similarly, the 14-kb region could not be amplified in Keller and 62 additional juicy sorghum lines (see below for more information). Therefore, we regarded the NAC gene as the functional candidate of the *Dry* locus.

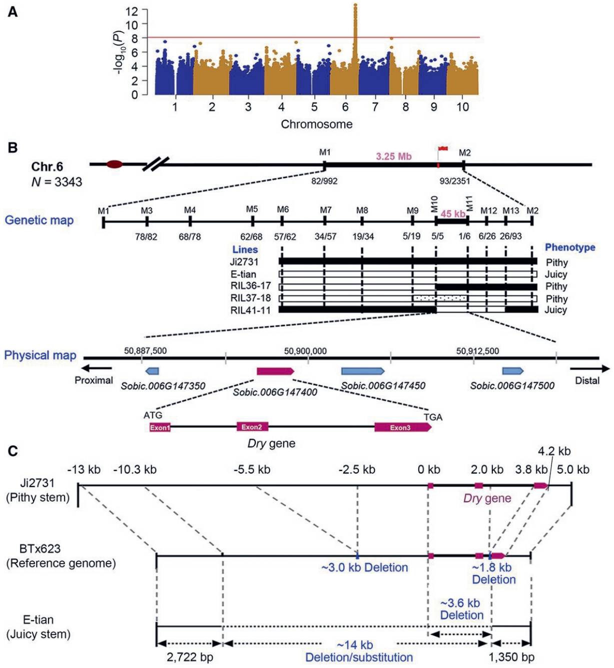


Figure 6.3. **Cloning of the *Dry* gene in sorghum.** (A) Manhattan plot of GWAS of the stem juicy/pithy trait using a panel of 241 sorghum accessions. The red horizontal line depicts the Bonferroni-adjusted significance threshold [$-\log_{10}$ (P) = 8.06]. See also Supplemental Figure 1. (B) Diagram illustrating the fine mapping of the *Dry* gene. In the genetic map, numbers below the lines indicate the number of recombinants versus the total number of individuals examined from both ends. The white bars represent the homozygous regions of E-tian; the dotted bar represents the heterozygous region of Ji2731 and E-tian; and the black bars represent the homozygous regions of Ji2731. In the physical map, the pink arrows/rectangles indicate the *Dry* gene and its exon regions, and the three blue arrows indicate the three remaining genes within the candidate region. See also Supplemental Table 1. (C) Diagram showing the genomic variation in the vicinity of the *Dry* locus among Ji2731, E-tian, and the reference genome BTx623. See also Supplemental Figures 2 and 3.

Compared with Ji2731, BTx623 had two large deletions (a ~1.8-kb deletion in the second intron region and another ~3-kb deletion in the upstream region) in the *Dry* gene (Figure 6.3C; Supplemental Figure 2). We also verified erroneous annotation of the *Dry* gene in the reference genome using rapid amplification of cDNA ends (RACE). According to the annotation (Sorghum bicolor v3.1.1) in the Phytozome database, the *Dry* locus has four exons and three introns. However, we found that the first intron (39-bp length) of the *Dry* gene was transcribed in BTx623, as revealed by our RACE analysis. Meanwhile, a C (Ji2731) to T (BTx623) transition introduces a premature termination codon within the wrongly annotated intron, resulting in the loss of function of the *Dry* gene in BTx623, which is consistent with the finding that BTx623 has a juicy stem (Supplemental Figure 3).

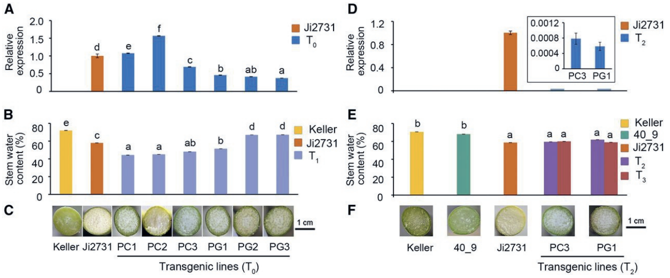


Figure 6.4. **Complementation test of the *Dry* Gene.** (A) and (D) RT-qPCR analysis of the expression of the transgenes in the stem pith in the T0 and T2 generations. The inset in (D) is an enlarged view of the gene expression patterns in PC3 and PG1. The expression level in Ji2731 was set to 1.0. Results are means ± se of three technical replicates (replicates within an experiment). (B) and (E) Water content in the T1, T2, and T3 generations. Values are means ± se of at least three positive plants for each transgenic line. (C) and (F) The stem juicy/pithy phenotype in the T0 and T2 generations. See also Supplemental Figure 4. Keller is the recipient plant for sorghum transformation, 40_9 is null sibling of the transgenic lines in the T2 generation, and Ji2731 was used as a positive control. Significantly different values ($P < 0.05$) are indicated by different letters, as determined by one-way ANOVA (Supplemental File 7).

We introduced both the *Dry* open reading frame and the genomic sequences of Ji2731 driven by its own promoter and the constitutive ubiquitin promoter, respectively, into sweet sorghum line Keller (with a deletion in the *Dry* gene similar to that of E-tian) by Agrobacterium tumefaciens-mediated transformation (Supplemental Figure 4A). Twenty-two independent transgenic plants (T0 plants) were generated, which displayed various degrees of dryness and pithy stems like that of Ji2731 at similar developmental stages. We examined the progeny of transgenic T0 plants with elevated *Dry* gene expression and found that the water content in T1 transgenic plants was reduced compared with Keller (Figures 6.4A to 6.4C). We also monitored the expression levels of the transgenes and stem water contents in the T2 and T3 generations for transgenic lines PC3 (line harboring *pDry:Dry_CDS*) and PG1 (line harboring *pUbi:Dry_genomic_sequence*). A significant difference in stem water content was still observed in the T2 and T3 generations, although the expression levels of the transgene were substantially reduced (Figures 6.4D to 6.4F). Further observations of field-grown

T2 transgenic lines showed that the dry, pithy stem phenotype was retained, and no obvious differences occurred in any other agronomic traits examined (Supplemental Figures 4B and 4C). Hence, we confirm that the NAC transcription factor encoded by the *Dry* locus controls the pithy/juicy stem phenotype.

6.2.4 Molecular and cellular characterization of the Dry gene

The *Dry* gene is predicted to encode a protein belonging to the NAC1 transcription factor subfamily (22), with a typical nuclear localization (Figure 6.5A). RNA *in situ* hybridization showed that *Dry* transcripts were predominantly found in the vascular bundle and parenchyma tissues in Ji2731, while, as expected, in E-tian, no signals above the background level were detected (Figure 6.5B). Our sequence comparison and yeast two-hybrid assay revealed the presence of the conserved NAC domain at the N terminus and the activation domain at the C terminus, with its self-activating region located from 227 to 259 amino acids (Figure 6.5C; Supplemental Figure 5). Gene expression analysis showed that the *Dry* gene is predominantly expressed in well-developed pithy stems (including the stem pith and the epidermal sclerenchyma layer of the stem) (Figure 6.5D). Furthermore, an examination of *Dry* gene expression along the stem internodes revealed that its expression level is generally associated with the degree of stem pithiness, suggesting that the expression level of the *Dry* gene directly controls the degree of stem pithiness (Figure 6.5E).

6.2.5 Dry regulates cell wall biosynthesis

The results of expression pattern analysis, together with findings for other species, indicate that *Dry* gene subfamily NAC transcription factors often regulate the cell wall network (23,24). This finding prompted us to examine cell morphology and cell wall composition in sorghum stems. Normal rounded parenchymal cells were found in Ji2731 (containing the functional *Dry* allele; Figure 6.6A), whereas irregularly shaped cells and striking cell collapse were ubiquitous in the parenchyma of E-tian (containing the null *dry* allele; Figure 6.6B). Transverse sections observed by transmission electron microscopy further showed the presence of irregularly shaped vessel elements in the juicy variety, in contrast to the fully developed vessel elements in the pithy variety (Figures 6.6C and 6.6D). All of the abnormalities detected in E-tian were also observed in Keller, another sweet sorghum (Figures 6.6G and 6.6I) and were partially restored in transgenic lines expressing the functional Dry allele (Figures 6.6H and 6.6J). The cell wall thickness of fiber cells was substantially reduced in xylem tissues of E-tian and Keller, but the fiber cells of the transgenic lines were much thicker and appeared normal, like those of Ji2731 (Figures 6.6E, 6.6F, 6.6K, 6.6L, and 6.6M). Thus, a range of defects in cell morphology occurred in the juicy, sweet sorghum lines due to the loss of function of *Dry*.

To better understand the cellular and molecular basis of cell collapse associated with xylem tissues, we analyzed the cell wall compositions and found that juicy sorghum had significantly elevated cellulose content but reduced hemicellulose content in the stem pith (Figures 6.6N and 6.6O; Supplemental Figure 6). Furthermore, the cellulose

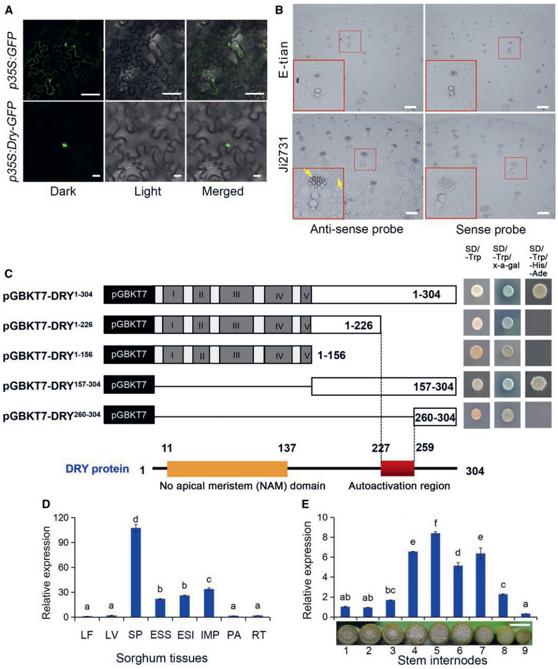


Figure 6.5. **Cellular and molecular characterization of the *Dry* gene and DRY protein.** (A) Subcellular localization of the DRY:GFP fusion protein in the epidermal cells of *N. benthamiana* leaves. Left, GFP (green) fluorescence images; middle, bright-field images; and right, merged images. p35S:GFP was used as a control. Bars = 40 µm in p35S:GFP and 10 µm in p35S:Dry-GFP. (B) Representative micrographs showing the RNA in situ hybridization with the antisense and sense probes (negative control) to probe cross sections of 1-month-old Ji2731 and E-tian sorghum seedlings showing that higher than background transcript levels were found only in the Ji2731 antisense section. The yellow arrows indicate vascular tissue and parenchyma cells, respectively. The insets within the red squares are 6.25-fold enlarged sections showing the obviously higher levels of the *Dry* transcripts in vascular tissues and parenchyma cells. Bars = 100 µm. (C) Identification of the transcriptional activation region in the DRY protein via a yeast two-hybrid assay. The deduced functionally important NAM domain is shown in orange. The red box between amino acids 227 and 259 indicates the autoactivation region of DRY. See also Supplemental Figure 5. (D) Relative expression of the *Dry* gene in different tissues of Ji2731 at the mature stage. LF, leaf; LV, leaf vein; SP, stem pith; ESS, epidermal sclerenchyma layer of the stem; ESI, epidermal sclerenchyma layer of the intercalary meristem; IMP, pith of the intercalary meristem; PA, panicle; RT, root. The expression level in the leaf was set to 1.0. (E) Relative expression of the *Dry* gene in different internodes of Ji2731 at the heading stage. Top, relative expression of the *Dry* gene; bottom, pithy phenotypes of different stem internodes. The first internode is near the ground. All data shown in (D) and (E) are means ± se of three technical repeats, and for each data point, four different plants (biological replicates) were tested. The experiments were repeated twice. Statistically significant differences were tested at $P < 0.05$ by one-way ANOVA, and bars with different letters are significantly different (Supplemental File 7).

and hemi-cellulose content of the transgenic lines reverted to levels similar to those of dry, pithy sorghum line Ji2731 (Figures 6.6N and 6.6O; Supplemental Figure 6). Therefore, we postulate that the higher water content retained in juicy sorghum stems could be attributed to changes in cell morphology and cell wall compositions in the stem.

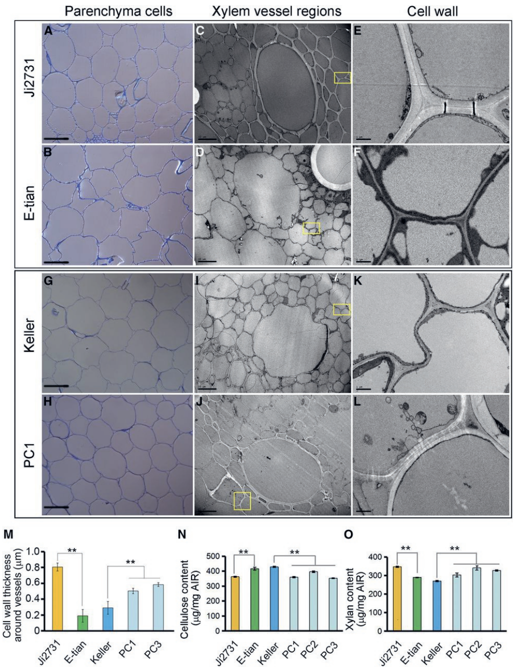


Figure 6.6. **Observation of cell morphology and measurements of cell wall composition in Ji2731, E-tian, Keller, and representative transgenic lines.** (A), (B), (G), and (H) Micrographs are semithin sections of parenchyma cells in the pithy stem line Ji2731 (A), juicy stem lines E-tian and Keller ([B] and [G]), and transgenic line PC1, which partially complemented the juicy stem phenotype (H). In E-tian (B) and Keller (G), irregular parenchyma cells were observed, but in Ji2731 (A) and PC1 (H), parenchyma cells have a normal, rounded shape. Bars = 50 µm. (C), (D), (I), and (J) Micrographs are transverse sections of cells in the xylem vessel regions of Ji2731 (C), E-tian (D), Keller (I), and PC1 (J), respectively, observed by transmission electron microscopy. Bars = 10 µm. (E), (F), (K) and (L) Micrographs are close-up view of the cell walls of parenchyma cells surrounding xylem vessels. (E), (F), (K), and (L) correspond to the yellow boxes in (C), (D), (I), and (J), respectively. In PC1 (L), the secondary cell wall thickness in parenchyma cells was significantly increased compared with Keller (K). Bars = 1 µm. (M) Measurements of secondary cell wall thickness in parenchyma cells around vessels in Ji2731, E-tian, Keller, and transgenic lines PC1 and PC3. Results are means ± se (n = 10 micrographs). (N) and (O) Cell wall composition in the stem piths of Ji2731, E-tian, Keller, and transgenic lines PC1, PC2, and PC3, which partially complemented the juicy stem phenotype. Pithy stem line Ji2731 was included as a positive control. AIR, alcohol-insoluble residues. Results are means ± se of three to five technical replicates, and statistical significance was evaluated by two-tailed Student's t test: $*P < 0.05$ and $**P < 0.01$ (Supplemental Figure 6 and Supplemental File 7).

The family of SECONDARY WALL-ASSOCIATED NAC (SWN) proteins includes three subfamilies, VND, SMB, and NST. Phylogenetic analysis revealed high similarity of the DRY protein to the VND subfamily of maize, rice, and Arabidopsis, suggesting that the *Dry* gene might be a key master regulator of secondary cell wall biosynthesis (Figure 6.7A; Supplemental Figure 7 and Supplemental Files 1 to 3). SWNs activate a battery of downstream MYB transcription factor genes (25). . We identified the homologs of these MYB transcription factors in sorghum (Figures 6.7B; Supplemental

Figure 6.8 and Supplemental Files 1, 4, and 5). To further explore the possible involvement of *Dry* in the cell wall biosynthesis network, we performed RNA-seq analysis of Ji2731, E-tian, transgenic sorghum line 56_11, and negative control line 40_9. Transgenic sorghum line 56_11 was the offspring of PC3 (T0 generation), and the negative control line 40_9 was the null sibling of the transgenic lines. Comparison of the transcriptional profiles between 56_11 versus 40_9 and Ji2731 versus E-tian showed that the expression of the sorghum MYB52/54 homolog was inhibited in the loss-of-function dry gene background, and the homologous genes for cellulose and

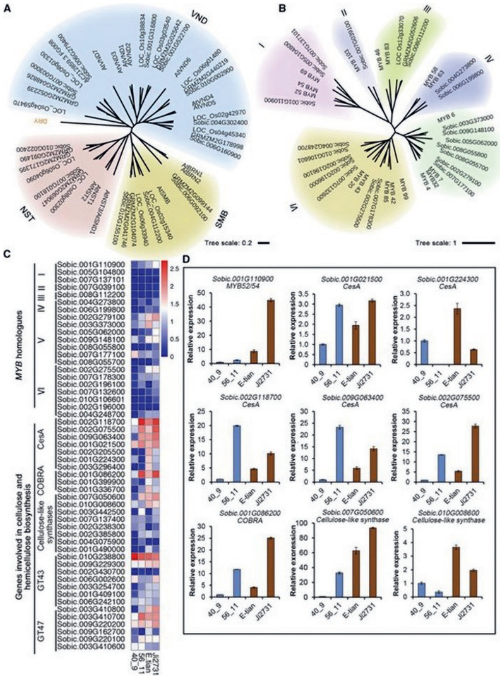


Figure 6.7. **The *Dry* gene regulates cell wall biosynthesis.** (A) Phylogenetic tree of SWN proteins in sorghum, maize, rice and Arabidopsis. The DRY protein is neighboring to the VND proteins. Different colors represent different subfamilies of SWN proteins. See also Supplemental Figure 7. The scale bar corresponds to 0.2 estimated amino acid substitutions per site. The bootstrap values and the alignments are shown in Supplemental Files 1 and 2, respectively. (B) Phylogenetic tree of the MYB proteins associated with cell wall synthesis in sorghum, maize, rice, and Arabidopsis. Different colors represent different clusters of MYB proteins. See also Supplemental Figure 8. The scale bar corresponds to 1 estimated amino acid substitutions per site. The bootstrap values and the alignments are shown in Supplemental Files 1 and 4, respectively. Sequences of SWN proteins (A) and MYB proteins (B) in sorghum, maize, rice, and Arabidopsis were obtained from PlantTFDB according to a previous study (23). Protein sequences were aligned by ClustalW (26), and MEGA (27) was used to construct the unrooted phylogenetic tree by the deduced neighbor-joining method. (C) Expression profiles of *MYB* genes and genes involved in cellulose and hemicellulose synthesis in Ji2731, E-tian, transgenic line 56_11, and negative control 40_9, as determined by RNA-seq. The heat map was plotted on a $\log_{10}$ scale of FPKM (the values of $\log_{10}(\text{Value}_{\text{FPKM}})$ <0 was turned as 0). See also Supplemental Data Sets 4 and 5. (D) RT-qPCR verification of the expression levels of nine randomly selected genes involved in cell wall biosynthesis. The results are consistent with the RNA-seq data. Results are means ± se of three technical replicates.

xylan biosynthesis showed the expected altered expression patterns (Figure 6.7C; Supplemental Data Sets 4 and 5) like those of Arabidopsis (28,29). Other relevant *MYBs* such as *MYB20*, *MYB42*, *MYB43*, and *MYB85* (23), as well as their close homologs in sorghum, showed similar expression patterns to those of *MYB52/54*, likely also acting as negative regulators of cellulose biosynthesis as in other plants (Figure 6.7C; Supplemental Data Set 4). Interestingly, the expression level of the sorghum *MYB46/83* homolog was very low in Ji2731 and E-tian and was not detected in transgenic line 56_11 or the negative control line 40_9 (Figure 6.7C; Supplemental Data Set 4). These results suggest that the sorghum *MYB46/83* homolog may have a different mode of action from those of Arabidopsis. Finally, we compared the expression levels of nine genes involved in cell wall biosynthesis in the four genotypes and found strong associations between the expression levels of these genes and the presence of *Dry/dry* gene alleles (Figure 6.7D). Together, we suggest that the *Dry* gene most likely acts as a regulator of the cell wall biosynthesis network.

6.2.6 The Dry gene in juicy sorghum is under positive selection

The finding that sorghum lines with dry, pithy stems carry the functional *Dry* alleles, while lines with juicy stems carry nonfunctional alleles, implies that different selection signals could be expected on the *Dry* locus between the dry, pithy and juicy sorghum subgroups. To test this hypothesis, we performed a XP-CLR (cross-population composite likelihood ratio) (30) scan around the *Dry* locus region within the pithy cultivated subgroups compared with the wild lines and the juicy compared with the cultivated dry, pithy subgroups. Indeed, in the *Dry* gene flanking regions, no significant signals were detected in the comparison between the pithy cultivated and wild subgroups (Figure 6.8A), whereas conspicuous positive selection signals were detected when comparing the juicy to the dry, pithy subgroups of cultivated sorghums (Figure 6.8B). Furthermore, a scan of nucleotide diversity following the same grouping showed similar nucleotide diversity (π) values around the *Dry* locus between the wild and the dry, pithy cultivated subgroups but distinctly different values between the pithy and juicy subgroups (Figures 6.8C and 6.8D). The nucleotide diversity around the *Dry* locus

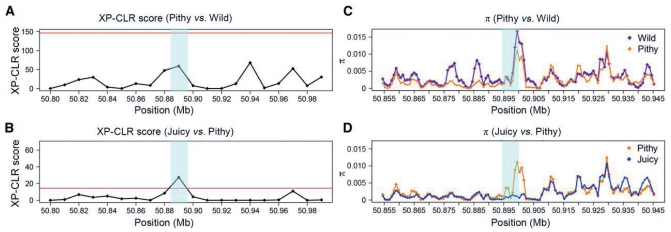


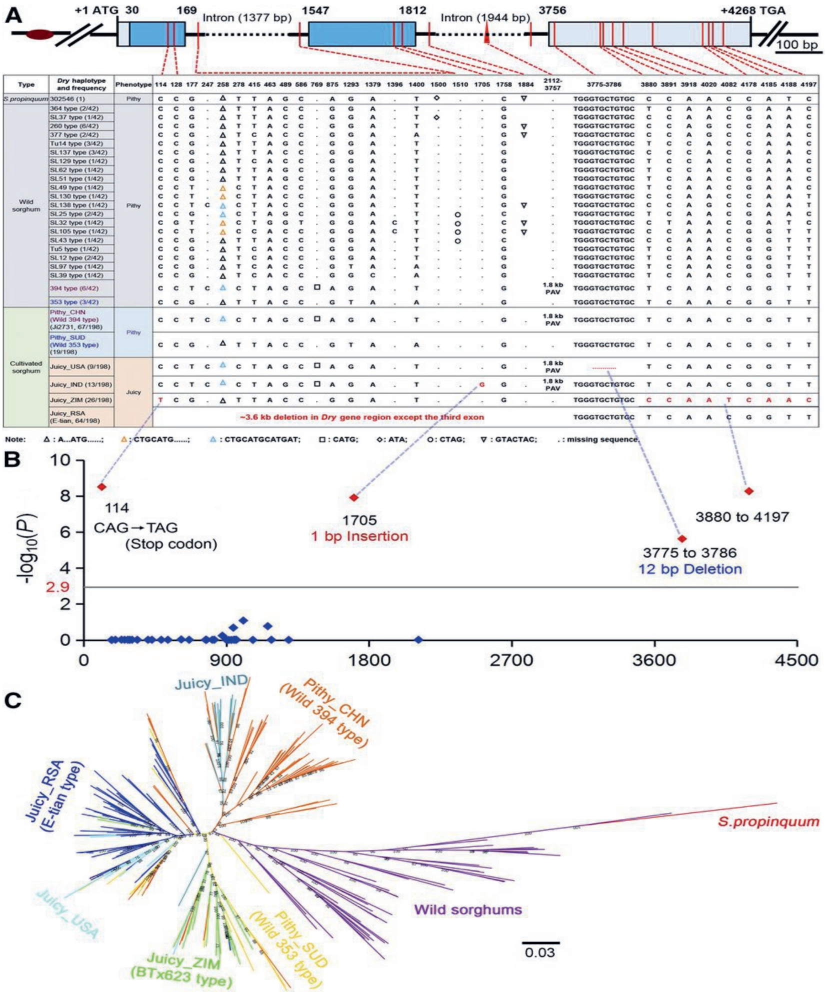
Figure 6.8. **Comparison of the XP-CLR and nucleotide diversity (π) in the vicinity of the *Dry* gene within the pithy cultivated subgroups versus the wild accessions and the juicy versus cultivated dry, pithy subgroups.** (A) and (B) XP-CLR analysis for two comparisons (pithy versus wild, and juicy versus pithy). The red lines indicate the threshold (5% upper-quantile XP-CLR scores across the whole genome) for each comparison. (C) and (D) show nucleotide diversity (π) analysis for two comparisons (pithy versus wild and juicy versus pithy). The regions shaded in cyan indicate the *Dry* gene region.

within the juicy subgroup was markedly lower than that of the pithy subgroup, which also indicates strong selection signals. Hence, our analysis provides evidence that intensive positive selection occurred on the *Dry* gene in juicy sorghums but not in dry, pithy sorghums.

6.2.7 Dry gene haplotypes among sorghum populations

We sequenced the full-length *Dry* genes across 60 sorghum accessions using Sanger sequencing to validate the resequencing data, which included most of the wild sorghums and representative cultivated sorghums (Supplemental Data Set 6). Sanger sequencing verified the variation data, which were then used to classify the *Dry* gene haplotypes across the 241-line sorghum population. In total, 67 variants were identified, including 30 SNPs, 36 indels, and a 1.8-kb PAV (presence and absence variation) in *Sorghum propinquum* and wild sorghums, which classified the *Dry* gene into 23 haplotypes (Figure 6.9A; Supplemental Data Set 7). The *Sorghum propinquum* line represents an independent haplotype with the most variations, and 22 different haplotypes were found for 42 wild sorghums, revealing a high degree of polymorphism (Figure 6.9A). Despite the existence of the diverse haplotypes in the wild sorghum population, these *Dry* gene variants did not lead to frameshifts or changes to conserved sites within the functional domains. Hence, the diverse haplotypes likely maintain the full functions of the encoded DRY proteins, which is consistent with the dry, pithy stem phenotype found in the wild sorghum and wild sorghum relatives (Supplemental Figure 9).

Within the 86 cultivated dry, pithy sorghums, we identified two haplotypes, Pithy_CHN and Pithy_SUD (named according to their most likely geographical origins) (Supplemental Figure 10), which mainly differed in the 1.8-kb PAV of the second intron and were exactly the same as the two wild sorghum haplotypes, 394_type and 353_type, respectively (Figure 6.9A). Due to the higher estimated Tajima's *D* value for the *Dry* gene within the cultivated dry, pithy sorghum population than the wild sorghum population (pithy versus wild = 1.36 versus 0.83), we speculate that this dramatic decrease in haplotypes in the cultivated pithy sorghum population is due to a recent bottleneck. Such a bottleneck would correspond to a preferential loss of low-frequency variants (31). On the other hand, four juicy stem haplotypes were defined by four new variations, all causing the loss of function of the *Dry* gene by either the 12-bp deletion (Juicy_USA, mostly representing the accessions with no original collecting site information, which were likely bred in the US) within the conserved functional NAC domain, frameshifts (Juicy_IND), premature stop mutations (Juicy_ZIM), or deletion of the gene (Juicy_RSA) (Figure 6.9A; Supplemental Figure 10). An association test of the dry, pithy, and juicy phenotypes of stems across 134 cultivated sorghums (without the Juicy_RSA haplotype, as its *Dry* gene was deleted) showed that the three mutations were significantly associated with the loss of function of the *Dry* gene (P value = 2.37×10^{-6} to 3.07×10^{-9}). Another important mutation at the 3' end was linked to mutation site 114 in Juicy_ZIM (Figure 6.9B).



Type	*Dry* haplotype and frequency	Phenotype	114	128	177	247	258	278	415	463	489	586	769	875	1293	1379	1396	1400	1500	1510	1705	1758	1884	2112-3757	3775-3786	3880	3891	3918	4020	4082	4178	4185	4188	4197
S.propinquum	302546 (1)	Pithy	C	C	G	.	Δ	T	T	A	G	C	.	A	G	A	.	T	◇	.	.	C	▽	.	TGGGTGCTGTGC	C	C	A	A	C	C	A	T	C
Wild sorghum	364 type (2/42)	Pithy	C	C	G	.	Δ	T	T	A	C	C	.	G	G	A	.	T	.	.	.	G		.	TGGGTGCTGTGC	T	C	A	A	C	G	A	A	C
	SL37 type (1/42)		C	C	G	.	Δ	T	T	A	C	C	.	G	G	A	.	T	◇	.	.	G		.	TGGGTGCTGTGC	C	C	A	A	C	G	A	A	C
	260 type (6/42)		C	C	G	.	Δ	T	T	A	C	C	.	G	G	A	.	T	.	.	.	G	▽	.	TGGGTGCTGTGC	C	C	A	G	C	C	A	A	C
	377 type (2/42)		C	C	G	.	Δ	T	C	A	C	C	.	G	G	A	.	A	.	.	.	G	▽	.	TGGGTGCTGTGC	C	C	A	G	C	C	A	A	C
	Tu14 type (3/42)		C	C	G	.	Δ	T	T	A	G	C	.	G	G	A	.	T	.	.	.	G		.	TGGGTGCTGTGC	T	C	C	A	C	C	A	A	C
	SL137 type (3/42)		C	C	G	.	Δ	T	T	A	C	C	.	G	G	A	.	T	.	.	.	G		.	TGGGTGCTGTGC	C	C	C	A	C	G	A	A	C
	SL129 type (1/42)		C	C	G	.	Δ	T	C	A	C	C	.	G	G	A	.	T	.	.	.	G		.	TGGGTGCTGTGC	T	C	C	A	C	C	A	A	C
	SL62 type (1/42)		C	C	G	.	Δ	T	T	A	G	C	.	G	G	A	.	T	.	.	.	G		.	TGGGTGCTGTGC	T	C	C	A	C	G	A	A	C
	SL51 type (1/42)		C	C	G	.	Δ	T	T	A	G	C	.	G	G	A	.	T	.	.	.	G		.	TGGGTGCTGTGC	T	C	A	A	C	G	A	A	C
	SL49 type (1/42)		C	C	T	.	Δ	C	T	A	C	C	.	G	G	A	.	T	.	.	.	G		.	TGGGTGCTGTGC	T	C	A	A	C	G	A	A	C
	SL130 type (1/42)		C	C	T	.	Δ	C	T	A	C	C	.	G	G	A	.	T	.	.	.	G		.	TGGGTGCTGTGC	C	C	A	A	C	C	A	A	T
	SL138 type (1/42)		C	C	T	C	Δ	C	T	A	C	C	.	G	G	A	.	T	.	.	.	G	▽	.	TGGGTGCTGTGC	C	C	A	G	C	C	A	A	T
	SL25 type (2/42)		C	C	G	.	Δ	C	T	A	G	C	.	G	G	A	.	T	.	○	.	C		.	TGGGTGCTGTGC	T	C	A	A	C	G	A	A	C
	SL32 type (1/42)		C	G	T	.	Δ	C	T	G	G	T	.	G	G	A	C	T	.	○	.	C	▽	.	TGGGTGCTGTGC	C	T	A	A	C	G	A	T	C
	SL105 type (1/42)		C	C	T	.	Δ	C	C	A	C	C	.	G	G	A	C	T	.	○	.	C	▽	.	TGGGTGCTGTGC	C	C	A	A	C	G	G	T	T
	SL43 type (1/42)		C	C	G	.	Δ	T	T	A	C	C	.	G	G	A	.	T	.	○	.	C		.	TGGGTGCTGTGC	C	C	A	A	C	G	G	T	T
	Tu5 type (1/42)		C	C	G	.	Δ	T	C	A	C	C	.	G	G	A	.	T	.	.	.	G		.	TGGGTGCTGTGC	T	C	A	A	C	G	G	T	T
	SL12 type (2/42)		C	C	G	.	Δ	T	C	A	C	C	.	G	G	A	.	T	.	.	.	G		.	TGGGTGCTGTGC	T	C	A	A	C	G	G	T	T
	SL97 type (1/42)		C	C	G	.	Δ	T	C	A	C	C	.	G	T	A	.	A	.	.	.	G		.	TGGGTGCTGTGC	T	C	A	A	C	G	G	T	T
	SL39 type (1/42)		C	C	G	.	Δ	T	C	A	C	C	.	G	G	C	.	A	.	.	.	G		.	TGGGTGCTGTGC	T	C	A	A	C	G	G	T	T
	394 type (6/42)		C	C	T	C	Δ	C	T	A	G	C	□	A	G	A	.	T	.	.	.	G		1.8 kb PAV	TGGGTGCTGTGC	T	C	A	A	C	G	G	T	T
	353 type (3/42)		C	C	G	.	Δ	T	T	A	C	C	.	G	T	A	.	A	.	.	.	G		.	TGGGTGCTGTGC	T	C	A	A	C	G	G	T	T
Cultivated sorghum	Pithy_CHN (Wild 394 type) (Ji2731, 67/198)	Pithy	C	C	T	C	Δ	C	T	A	G	C	□	A	G	A	.	T	.	.	.	G	.	1.8 kb PAV	TGGGTGCTGTGC	T	C	A	A	C	G	G	T	T
	Pithy_SUD (Wild 353 type) (19/198)		C	C	G	.	Δ	T	T	A	C	C	.	G	T	A	.	A	.	.	.	G	.	.	TGGGTGCTGTGC	T	C	A	A	C	G	G	T	T
	Juicy_USA (9/198)	Juicy	C	C	T	C	Δ	C	T	A	G	C	□	A	G	A	.	T	.	.	.	G	.	1.8 kb PAV	....	T	C	A	A	C	G	G	T	T
	Juicy_IND (13/198)		C	C	T	C	Δ	C	T	A	G	C	□	A	G	A	.	T	.	.	G	G	.	1.8 kb PAV	TGGGTGCTGTGC	T	C	A	A	C	G	G	T	T
	Juicy_ZIM (26/198)		T	C	G	.	Δ	T	T	A	C	C	.	G	G	A	.	T	.	.	.	G	.	.	TGGGTGCTGTGC	C	C	A	A	T	C	A	A	C
	Juicy_RSA (E-tian, 64/198)		~3.6 kb deletion in *Dry* gene region except the third exon																						TGGGTGCTGTGC	T	C	A	A	C	G	G	T	T

Note: Δ : A...ATG......; Δ : CTGCATG......; Δ : CTGCATGCATGAT; □ : CATG; ◇: ATA; ○: CTAG; ▽ : GTACTAC; . : missing sequence.

Figure 6.9. **Variant alleles and association test of the *Dry* gene.** (A) Analysis of variant alleles of the *Dry* gene for different haplotypes, including one *Sorghum propinquum* line, 42 wild sorghum lines, and 198 natural varieties. Twenty-three ancestral Dry haplotypes were found among one *Sorghum propinquum* and 42 wild sorghum lines, two of which were retained in the dry, pithy landraces (86), and four *dry* haplotypes with juicy stems were found among the improved lines (112). The positions of the variants are shown using the consensus sequence of the *Dry* gene as a reference, with the start codon designated as position 1. The four most important variants are highlighted in red. Boxes indicate exons, and black bars between the boxes indicate introns. Deduced functionally important NAM domains are shown in blue. ATG, start codon; TGA, stop codon. Different symbols indicate different InDel variants, as described in the footnote. See also Supplemental Figure 9 and Supplemental Data Set 7. (B) Association test of the dry, pithy, and juicy phenotypes of stems across 134 cultivated sorghums; the Juicy_RSA haplotype was not included, as its *Dry* gene was destroyed. Forty-three variations occur in the *Dry* gene region in cultivated sorghum. Line, $-\log_{10}(P) = 2.9$. See also Supplemental Figure 1. (C) Phylogenetic tree of 241 sorghum accessions. The phylogenetic tree was constructed using genome-wide SNPs. Different colors indicate different haplotypes of the *Dry* gene. The scale bar corresponds to 0.03 estimated nucleotide substitutions per site. Numbers in the nodes indicate the bootstrap values from 1000 trials (see also Supplemental Files 1). The alignments are shown in Supplemental Files 6.

Thus, our results suggest a strong association between the juicy/dry, pithy stem phenotype and the *Dry* haplotypes in cultivated sorghum.

Phylogenetic and population structural analysis further showed that the Pithy_CHN haplotype in cultivated sorghum is likely most closely related to wild sorghum (Figure 6.9C; Supplemental Figure 10 and Supplemental Files 1 and 6). Another obvious link was that the Juicy_IND haplotype might have been derived from the Pithy_CHN haplotype by deliberate breeding selection for the *dry* alleles (Figure 6.9C). The Juicy_USA haplotype has roughly the same geographical distribution as the Juicy_RSA haplotype (Supplemental Data Set 1 and Supplemental Figure 10; primarily collected from in the GRIN database). The Pithy_SUD haplotype appeared to represent an independent subgroup and was closely related to the Juicy_ZIM haplotype. The most prominent juicy haplotype, Juicy_RSA (representing 57% of the juicy stem sorghum lines), with the deletion of the *Dry* gene, had a mixed background and might have been derived from the Pithy_CHN haplotype via breeding selection.

6.2.8 Conservation of the Dry gene in cereals

Cereal crops such as rice, maize, foxtail millet, wheat, and barley often have dry, pithy and sometimes hollow stems/shoots at the mature stage, which prompted us to examine the possible orthologs of *Dry* genes in cereals. We found that the genomic region surrounding the sorghum *Dry* locus shows strong micro-synteny with those of maize, foxtail millet, rice, and *Brachypodium distachyon* (Figure 6.10A). We examined the features of ~120-kb syntenic blocks in these genes, including 15 genes in sorghum, and found that most of the 15 genes had various numbers of orthologs and similar arrangements in four of these additional grass species (Figure 6.10B). Furthermore, we found that the *Dry* gene and its orthologs, including *GRMZM2G081930* in maize, *Seita.7G166500* in foxtail millet, *LOC_Os04g43560* in rice, and *Bradi5g15587* in *Brachypodium distachyon*, possess similar gene structures, with three exons and two introns (Figure 6.10C). These results imply that the *Dry* genes are likely functionally conserved in the five species and that the genomic regions containing the *Dry* locus in cereals may be derived from a common ancestor. Further work is needed to exploit the molecular functions of these genes.

6.3 Discussion

In this study, we characterized the *Dry* gene, an important sorghum gene controlling the stem pithy/juicy trait. Using a range of mapping resources, gene expression studies, and transgenic manipulations, we showed that *Dry* encodes a plant-specific NAC transcription factor. Our population genomics analyses also provided evidence that selection on the *Dry* gene has implications for the history of sweet sorghum domestication.

The control of stem juiciness in sorghum has been a long-standing question in the sorghum academic and breeding communities. This trait was first shown to be controlled by a single locus, *Dry*, over a century ago (6). Yet, the simple assay scoring

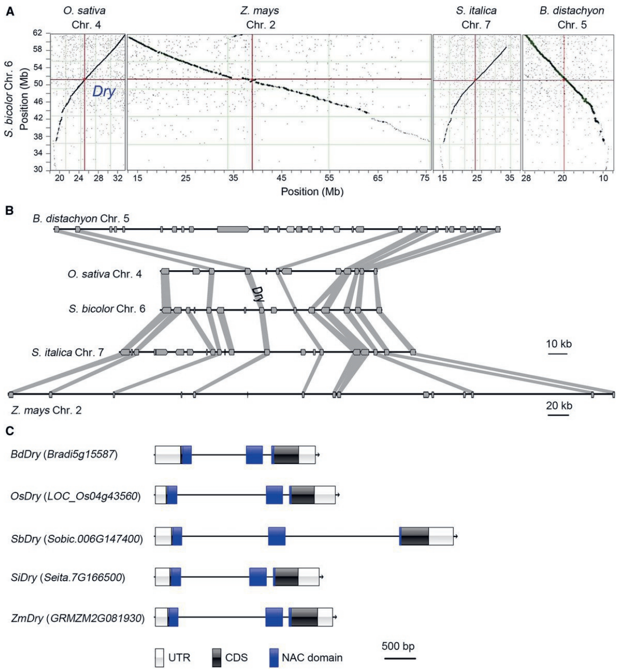


Figure 6.10. **Collinearity of the *Dry* locus in cereals.** (A) High collinearity in genomic regions close to the *Dry* loci in sorghum, maize, rice, *Setaria italica*, and *Brachypodium distachyon*. The genomic map was plotted based on BLASTP results of pairwise genome analyses from CoGe (32); each putative homologous gene-pair is drawn as a gray dot on the dot plot with its x and y position corresponding to the genomic position of each gene in their respective genomes. The dot plot alignments indicate the collinearity of genomic regions. (B) Collinear genomic regions around the *Dry* loci in S. bicolor (JGI v3.1.1), rice (JGI MSU v7.0), maize (JGI B73 v3), *Setaria italica* (JGI, v2.0.39), and *Brachypodium distachyon* (JGI v3.1). Genes are indicated by gray arrowed boxes, and shaded areas connect conserved genes. The 10-kb bar indicates the sizes of regions in sorghum, rice, foxtail millet, and *Brachypodium distachyon*, and the 20-kb bar indicates the sizes of regions in maize. (C) Structural comparison of *Dry* genes among the five species. The five genes show similar structures, with three exons and two introns. Gray boxes indicate untranslated region (UTR) sequences, black boxes the coding sequence (CDS), and blue boxes the NAM domain coding sequences.

the stem pithy/juicy phenotype described in this study was somehow neglected and has not been followed. Early studies also indicated that white/green midrib color co-segregated with the pithy/juicy trait and was tightly associated with the *Dry* locus (6-8). Such complications have made it difficult to pin down the juiciness gene(s) over the years, although the *Dry* locus has been consistently identified as a major-effect quantitative trait locus using various genetic populations and a number of phenotyping parameters, including extractable juice contents, green midribs, sugar yield, and moisture (9,10,12-14)

In this study, we explored the completely dry, pithy stem feature of the Chinese kaoliang sorghum. This feature appears to be more readily distinguishable than the juicy feature and remains quite stable across various growth conditions. This allowed us to score the pithy/juicy trait as the qualitative trait, as was done over 100 years ago, and to effectively build high-resolution fine maps in the linkage populations or within-gene associations in the GWAS populations. Much more effort is required to identify the genes associated with water content and their relationships with the *Dry* gene before we can obtain a holistic picture of the regulation of the juiciness trait in sweet sorghum.

We also noticed a close link between the green/white midribs and *dry/Dry* alleles. In sorghum natural population, we observed at least three leaf midrib colors, green, yellow, and white, and we did not detect any separation of the linkage between midrib color and the *Dry* alleles in the advanced RIL populations. This tight association is most likely a consequence of the highly effective visual selection of the leaf midrib phenotype. Hence, at this stage, it does not seem possible to draw any conclusion.

The origin of sweet sorghum remains debatable. Previous work showed that sweet sorghum could be found across all the major sorghum landraces (16-18). Indeed, our results provide evidence to support the argument that sweet sorghum originated following intensive breeding selection of the *Dry* gene performed simultaneously on various landraces. The observation that no juicy haplotypes were found in wild sorghum or early-domesticated landraces suggests that sweet sorghum with juicy stems occurred later due to breeding selection of the loss-of-function *dry* gene (Figure 6.11). The purpose for the early selection of juicy sweet sorghum was primarily for high-efficiency syrup production (33), while breeding selection for sweet, juicy sorghum has gradually been intensifying as this crop has come to be recognized as an important biofuel feedstock over the past several decades (1,2,34). The expansion of the *Dry* gene haplotypes from dry, pithy sorghum to juicy sweet sorghum also points to this tendency (Figure 6.11). Furthermore, the existence of two ancient *Dry* gene haplotypes in cultivated sorghums (Figure 6.11; Supplemental Figure 10) begs for further analysis of the history of sorghum domestication.

In this study, we provided evidence that the *Dry* gene might be an important first-layer master switch for the regulation of secondary cell wall biosynthesis in sorghum. Yet, more work is still required, including examining the molecular processes underlying how the *Dry* gene regulates cell wall biosynthesis and the subsequent accumulation of juice and sugars in the stem and how DRY functions within protein-protein interaction networks. Several sweet sorghum lines possess juice- and sugar-rich stems and have relatively high grain yields (1,2,34,35)， implying that it is possible to breed sorghum varieties with dual purposes as a biofuel and food crop. The finding that the *Dry* gene has conserved orthologs in other major cereals suggests that this gene is an attractive molecular target for developing crop varieties that convert dry, pithy cereal stems into juicy stems for silage production without compromising grain production. Sweet sorghum is especially appropriate for this purpose, as it is regarded as a model bioenergy crop (35) with potential evolutionary relationships with major cereal crops (36), as well as genomic collinearity with other biofuel crops with more complex genomes, such as switchgrass and sugarcane (3).

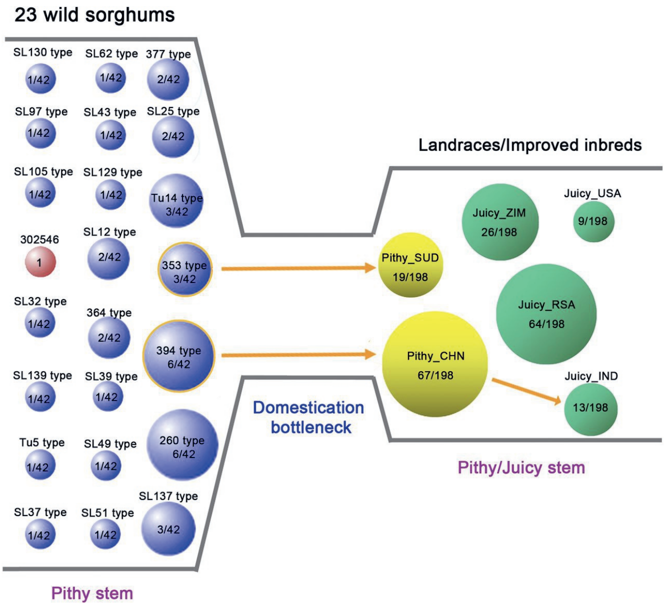


Figure 6.11. **Proposed model illustrating the origin of sweet sorghum via intensive breeding selection of the Dry gene.** Twenty-three ancestral Dry haplotypes were found in *Sorghum propinquum* (red circle) and wild sorghums (blue circles), two of which (blue circles outlined in orange) were maintained in the dry, pithy landraces (yellow circles), and four dry haplotypes (green circles) with the juicy stems were found in landraces/improved inbreds. There is a clear bottleneck effect on the *Dry* gene during sorghum domestication. The Pithy_CHN haplotype and Pithy_SUD haplotype in cultivated sorghum are likely two independent subgroups. Another obvious link is that the Juicy_IND haplotype was likely derived from the Pithy_CHN haplotype by breeding selection. Each circle represents one haplotype of the *Dry* gene. The number in the circle corresponds to the distribution frequency of sorghum lines in different haplotypes: the larger the circle, the more sorghum lines. See also Supplemental Data Set 7.

6.4 Methods

(Selected sections relevant to bioinformatics, see http://www.plantcell.org/content/30/10/2286.long for the full Methods text)

Resequencing and SNP Detection

All 241 sorghum lines were sequenced on the Illumina HiSeq X Ten platform using paired-end sequencing. Sequencing libraries were constructed using a TruseqNano DNA HT sample preparation kit (Illumina; catalog no. FC-121-4003) following the manufacturer's recommendations and index codes were added to tag each sample. The insertion length and reads length were 350 and 150 bp, respectively. The average sequencing depth of all sorghum accessions was ~5.67×. After trimming adapters and filtering low-quality reads (reads with ≥10% unidentified nucleotides [N]; >10 nucleotides aligned to the adaptor, allowing ≤10% mismatches; >50% bases having Phred quality score < 5), the clean reads were mapped to the reference genome BTx623 (v3.1) with Burrows-Wheeler Alignment software (version 0.7.8) using the mem command (37). While mapping, minimum seed length was set to 32 (-k 32), and marking shorter split hits as secondary (-M) was allowed. The SAMtools (version 1.3) package was used to convert the mapping results to BAM format and to eliminate the duplicated reads and multi-aligned reads (38).

After filtering, the variations in each sorghum accession, including SNPs and InDels, were called separately using the Genome Analysis Toolkit (GATK, version 3.1, HaplotypeCaller) (39) and the SAMtools (version 1.3) (38) package. Co-current variations detected by both tools were extracted after strict filtering. GATK was used to perform BQSR (recalibrating the base quality score) to obtain more accurate quality scores for each base using the co-current variations set as known sites (39). The recalibrated data after BQSR were used for variation detection by HaplotypeCaller in GATK. The co-current variation set (the same as used in the BQSR step) was then used as the "training and truth" set to perform variant quality score recalibration. At the same time, a strictly filtered variation set after GATK calling (the first step calling) was used as the "known" set. In total, 31,946,640 SNPs and 4,266,768 InDels were identified across 241 sorghum lines.

LD Decay and PCA Analyses

The LD levels in all three sorghum subgroups (Wild, Pithy, and Juicy) were measured based on the pairwise correlation coefficient (r^2) of alleles. The r^2 for each pairwise comparison of SNPs was calculated with Haploview (40) using previously described parameters (41), and the distance between each pair of SNPs was recorded. The average r^2 value for each distance was calculated, and an LD decay plot for the three sorghum subgroups was created using an in-house R script, in which the LD decay simulation curves were estimated using the smooth.spline function. Based on the LD decay values, the SNPs were filtered with MAF 0.05, maximum missing rate 0.5 and LD (r^2) 0.2, then the filtered SNPs were used to perform PCA with GCTA software (42), and the first two eigenvectors were plotted in 2D.

GWAS

GWAS for the stem pithy/juicy trait was performed using GAPIT (v2017.08.18) (43) under the MLM (PCA+K) model. SNPs for GWAS were filtered with MAF 0.05, maximum missing rate of 0.5 using VCFtools. The PCA of the filtered SNPs was determined using GCTA software (42). The first five PCs explained 38.95% of the variation and were adopted in the MLM model. The kinship matrix (K) was constructed according to the GAPIT manual.

Haplotype Analysis of the Dry Gene

The genomic sequences of *Dry* gene were amplified from 60 sorghum accessions by PCR using KOD-Plus-Neo (Toyobo; catalog no. KOD-401) and sequenced by Sanger sequencing. The sequences were assembled and aligned to identify the variations. An additional 181 sorghum inbred lines were then selected as the diversity panel to examine the *Dry* haplotypes.

Phylogenetic Analysis

The sorghum NAC and MYB protein sequences were downloaded from plantTFDB (44). The NAC and MYB proteins involved in the secondary cell wall synthesis in sorghum (*Sorghum bicolor*), maize (*Zea mays*), rice (*Oryza sativa*), and *Arabidopsis thaliana* were chosen as previously reported (23). The protein sequences were aligned with ClustalW, and a neighbor-joining phylogenetic tree was constructed using MEGA 6.0 software (version 6.0.6). Bootstrapping was performed with 1000 replications. The online tool iTOL (45) was used to visualize the phylogenetic tree. The scale bar corresponds to estimated amino acids per site.

Transcriptome Analysis of Sorghum

Sorghum plants for the RNA-seq experiments were grown in the experimental fields at the Institute of Botany, Chinese Academy of Sciences (Beijing, China) in 2016. At the heading stage, the stem pith in the middle internodes, including the 6th in Ji2731, the 8th in E-tian, and the 7th in transgenic line 56_11 and the negative control 40_9 in the Keller background, were collected for RNA extraction. Three different plants (biological replicates) were processed for Ji2731 and E-tian, respectively. Only one biological replicate was processed for 56_11 and 40_9. Total RNA was extracted using RNAiso Plus reagent (TaKaRa; catalog no. 9108) and purified using DNase I (TaKaRa; catalog no. 2270A). RNA-seq libraries were constructed using a NEBNext Ultra RNA Library Prep Kit for Illumina (NEB; catalog no. E7530L), and high-throughput sequencing was performed using Illumina HiSeq 2500. Reads generated in the paired-end sequencing were 150 bp long. After removing the adaptor sequences and low-quality reads, the clean reads were mapped to the reference genome BTx623 (v3.1) using HISAT (version 0.1.6) (46). FPKM (fragment per kilobases per million fragments) (47) values were used to measure gene expression levels with HTSeq (version 0.6.1) software (48).

Selection Signal Scanning

XPCLR software (version 1.0) (30) was used to calculate the XP-CLR score between the subgroups with a nonoverlapping window size of 10 kb. SNPs for XP-CLR analysis were filtered with a maximum missing rate 0.5, minimum depth 2, minimum genotype quality 5, and only biallelic sites with at least one of the reference alleles were retained. To calculate the XP-CLR score, the average genetic distances for Chromosomes 1 to 10 were estimated as 2.5, 2.6, 2.32, 2.22, 2, 2.6, 2, 2.64, 2.02, and 2.8 came/Mb, respectively, using information from previous studies (11,49). VCFtools software was used to calculate the nucleotide diversity (π) for each subgroup with a window size of 1500 bp and a step of 750 bp (50). Tajima's *D* (51) statistic of the *Dry* gene was calculated using VCFtools (–TajimaD) with a window size across the gene region (Chromosome 6: 50,896,169 to 50,898,604 bp). The number of SNPs used to calculate the Tajima's *D* in the wild and pithy population was 37 and 30, respectively.

Population Genetics Analysis

SNPhylo (version 20140701) (52) was used to conduct the phylogenetic analysis. SNPs of all accessions were filtered with minimum depth 2, MAF 0.1, and maximum missing rate 0.5 and LD (r^2) 0.2. Sequences were generated from the filtered SNPs and used to perform multiple alignments by MUSCLE (53). The maximum likelihood tree was then constructed by running DNAML programs in the PHYLIP (54). Bootstrapping analysis for the tree was performed using the "phangorn" package (55), and bootstrap values were estimated from 1000 trials. *Sorghum propinquum* 302546 was used as the outgroup. All of the above procedures were integrated in SNPhylo with the parameters -p 5 -c 2 -l 0.2 -M 0.5 -o 302546 -A -b -B 1000. The online tool iTOL was used to visualize the phylogenetic tree. The scale bar corresponds to estimated nucleotide substitutions per site. The population structure of the panel was determined using fastSTRUCTURE (56). SNPs for structure determination were filtered with MAF 0.05, maximum missing rate 0.5, and LD (r^2) 0.2.

Comparative Genomics

For genomic collinearity plotting, BLASTP of SynMap on CoGe was used to perform pairwise genome comparisons (32). Genome sequence data sources were selected for the following cereals: *Sorghum bicolor* (BTx623, id331), rice (Nipponbare, id3), maize (B73, id333), *Setaria italic* (Yugu1, id32546), and *Brachypodium distachyon* (Bd21, id33982). Syntenic gene pairs were identified by DAGChainer (57) and colored based on their synonymous substitution rate as calculated using CodeML from the PAML package (58). The region from 30 to 62 Mb on Chromosome 6 of sorghum and its collinear regions in four other species are shown in the dot plot.

For collinearity plotting of the orthologs, orthologs of genes in the syntenic regions were identified using the PhytoMine tool in the Phytozome database (https://phytozome.jgi.doe.gov/phytomine/begin.do). Sequences of syntenic regions around the *Dry* locus and gene annotations in five species, including *Sorghum bicolor* (JGI v3.1.1), rice (JGI MSU v7.0), maize (JGI B73 v3), *Setaria italica* (JGI, v2.0.39), and *Brachypodium distachyon* (JGI v3.1), were obtained from the Phytozome database. Genes in collinear regions were marked on the sequences with IBS software (59) according to the gene annotations.

For gene structure plotting, structure information was extracted from the annotations in the Phytozome database. Structures of genes were drawn with IBS software (60).

Statistical Test

The two-tailed Student's *t*-test and one-way ANOVA were performed using SPSS 17.0. For values represented as means ± s.

References

1. Anami, S.E., Zhang, L.-M., Xia, Y., Zhang, Y.-M., Liu, Z.-Q. and Jing, H.-C. (2015) Sweet sorghum ideotypes: genetic improvement of the biofuel syndrome. *Food and Energy Security*, **4**, 159-177.
2. Mathur, S., Umakanth, A.V., Tonapi, V.A., Sharma, R. and Sharma, M.K. (2017) Sweet sorghum as biofuel feedstock: recent advances and available resources. *Biotechnol Biofuels*, **10**, 146.
3. Wang, J., Roe, B., Macmil, S., Yu, Q., Murray, J.E., Tang, H., Chen, C., Najar, F., Wiley, G., Bowers, J. *et al.* (2010) Microcollinearity between autopolyploid sugarcane and diploid sorghum genomes. *BMC Genomics*, **11**, 261.
4. Xu, B., Ohtani, M., Yamaguchi, M., Toyooka, K., Wakazaki, M., Sato, M., Kubo, M., Nakano, Y., Sano, R., Hiwatashi, Y. *et al.* (2014) Contribution of NAC Transcription Factors to Plant Adaptation to Land. *Science*, **343**, 1505-1508.
5. Plackett, A.R., Di Stilio, V.S. and Langdale, J.A. (2015) Ferns: the missing link in shoot evolution and development. *Front in Plant Scicence*, **6**, 972.
6. Hilson, G. (1916) On the inheritance of certain stem characters in sorghum. *Agriculture Journal India*, **11**, 150-155.
7. Swanson, A.F. and Parker, J.H. (1931) Inheritance of smut resistance and juiciness of stalk: In the sorghum cross, Red Amber X Feterita. *Journal of Heredity*, **22**, 51-56.
8. Rangaswami, G., Ayyangar, N., Ayyar, M.S. and Rao, V.P. (1937), *Proceedings of the Indian Academy of Sciences-Section B*. Springer India, Vol. 5, pp. 1-3.
9. Xu, W., Subudhi, P.K., Crasta, O.R., Rosenow, D.T., Mullet, J.E. and Nguyen, H.T. (2000) Molecular mapping of QTLs conferring stay-green in grain sorghum (Sorghum bicolor L. Moench). *Genome*, **43**, 461-469.
10. Hart, G.E., Schertz, K.F., Peng, Y. and Syed, N.H. (2001) Genetic mapping of Sorghum bicolor (L.) Moench QTLs that control variation in tillering and other morphological characters. *Theoretical and Applied Genetics*, **103**, 1232-1242.
11. Mace, E.S. and Jordan, D.R. (2010) Location of major effect genes in sorghum (Sorghum bicolor (L.) Moench). *Theoretical and Applied Genetics*, **121**, 1339-1356.
12. Zhai, G.W., Zou, G.H., Yan, S., Wang, H., Shao, J.F. and Tao, Y.Z. (2014) Identification and fine mapping of the gene associated with moisture content of stem in sorghum[Sorghum bicolor(L.) Moench]. *Acta Agriculturae Zhejiangensis*, **26**.
13. Han, Y., Lv, P., Hou, S., Li, S., Ji, G., Ma, X., Du, R. and Liu, G. (2015) Combining Next Generation Sequencing with Bulked Segregant Analysis to Fine Map a Stem Moisture Locus in Sorghum (Sorghum bicolor L. Moench). *PLoS One*, **10**, e0127065.
14. Burks, P.S., Kaiser, C.M., Hawkins, E.M. and Brown, P.J. (2015) Genomewide Association for Sugar Yield in Sweet Sorghum. *Crop Science*, **55**, 2138-2148.
15. Lekgari, A.L. (2010) Genetic mapping of quantitative trait loci associated with bioenergy traits, and the assessment of genetic variability in sweet sorghum (Sorghum bicolor (L.). Moench) *PhD dissertation (University of Nebraska-Lincoln)*
16. Ritter, K.B., McIntyre, C.L., Godwin, I.D., Jordan, D.R. and Chapman, S.C. (2007) An assessment of the genetic relationship between sweet and grain sorghums, within Sorghum bicolor ssp. bicolor (L.) Moench, using AFLP markers. *Euphytica*, **157**, 161-176.
17. Ali, M., Rajewski, J., Baenziger, P., Gill, K., Eskridge, K. and Dweikat, I. (2008) Assessment of genetic diversity and relationship among a collection of US sweet sorghum germplasm by SSR markers. *Molecular Breeding*, **21**, 497-509.
18. Wang, M.L., Zhu, C., Barkley, N.A., Chen, Z., Erpelding, J.E., Murray, S.C., Tuinstra, M.R., Tesso, T., Pederson, G.A. and Yu, J. (2009) Genetic diversity and population structure analysis of accessions in the US historic sweet sorghum collection. *Theoretical and Applied Genetics*, **120**, 13-23.
19. Zheng, L.-Y., Guo, X.-S., He, B., Sun, L.-J., Peng, Y., Dong, S.-S., Liu, T.-F., Jiang, S., Ramachandran, S., Liu, C.-M. *et al.* (2011) Genome-wide patterns of genetic variation in sweet and grain sorghum (Sorghum bicolor). *Genome Biology*, **12**, R114.
20. Paterson, A.H., Bowers, J.E., Bruggmann, R., Dubchak, I., Grimwood, J., Gundlach, H., Haberer, G., Hellsten, U., Mitros, T., Poliakov, A. *et al.* (2009) The Sorghum bicolor genome and the diversification of grasses. *Nature*, **457**, 551-556.

21. Luo, H., Zhao, W., Wang, Y., Xia, Y., Wu, X., Zhang, L., Tang, B., Zhu, J., Fang, L., Du, Z. *et al.* (2016) SorGSD: a sorghum genome SNP database. *Biotechnology for Biofuels*, **9**, 6.

22. Zhu, T., Nevo, E., Sun, D. and Peng, J. (2012) Phylogenetic Analysis Unravel the Evolutonary History of NAC Protein in Plants. *Evolution*, **66**, 1833-1848.

23. Nakano, Y., Yamaguchi, M., Endo, H., Rejab, N.A. and Ohtani, M. (2015) NAC-MYB-based transcriptional regulation of secondary cell wall biosynthesis in land plants. *Frontiers in Plant Science*, **6**.

24. Zhong, R. and Ye, Z.-H. (2014) Secondary Cell Walls: Biosynthesis, Patterned Deposition and Transcriptional Regulation. *Plant and Cell Physiology*, **56**, 195-214.

25. Zhong, R., Lee, C., Zhou, J., McCarthy, R.L. and Ye, Z.-H. (2008) A Battery of Transcription Factors Involved in the Regulation of Secondary Cell Wall Biosynthesis in *Arabidopsis*. *The Plant Cell*, **20**, 2763-2782.

26. Thompson, J.D., Gibson, T.J. and Higgins, D.G. (2003) Multiple Sequence Alignment Using ClustalW and ClustalX. *Current Protocols in Bioinformatics*, **00**, 2.3.1-2.3.22.

27. Kumar, S., Stecher, G., Li, M., Knyaz, C. and Tamura, K. (2018) MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Molecular Biology and Evolution*, **35**, 1547-1549.

28. McKinley, B., Rooney, W., Wilkerson, C. and Mullet, J. (2016) Dynamics of biomass partitioning, stem gene expression, cell wall biosynthesis, and sucrose accumulation during development of Sorghum bicolor. *Plant Journal*, **88**, 662-680.

29. Ko, J.-H., Jeon, H.-W., Kim, W.-C., Kim, J.-Y. and Han, K.-H. (2014) The MYB46/MYB83-mediated transcriptional regulatory programme is a gatekeeper of secondary wall biosynthesis. *Annals of Botany*, **114**, 1099-1107.

30. Chen, H., Patterson, N. and Reich, D. (2010) Population differentiation as a test for selective sweeps. *Genome Research*, **20**, 393-402.

31. Zhu, Q., Zheng, X., Luo, J., Gaut, B.S. and Ge, S. (2007) Multilocus Analysis of Nucleotide Variation of Oryza sativa and Its Wild Relatives: Severe Bottleneck during Domestication of Rice. *Molecular Biology and Evolution*, **24**, 875-888.

32. Lyons, E., Pedersen, B., Kane, J. and Freeling, M. (2008) The Value of Nonmodel Genomes and an Example Using SynMap Within CoGe to Dissect the Hexaploidy that Predates the Rosids. *Tropical Plant Biology*, **1**, 181-190.

33. Hunter, E. and Anderson, I. (1997) Sweet sorghum. *Horticultural reviews*, **21**, 73-104.

34. Mullet, J., Morishige, D., McCormick, R., Truong, S., Hilley, J., McKinley, B., Anderson, R., Olson, S.N. and Rooney, W. (2014) Energy Sorghum—a genetic model for the design of C4 grass bioenergy crops. *Journal of Experimental Botany*, **65**, 3479-3489.

35. Calviño, M. and Messing, J. (2012) Sweet sorghum as a model system for bioenergy crops. *Current Opinion in Biotechnology*, **23**, 323-329.

36. Paterson, A.H., Bowers, J.E., Feltus, F.A., Tang, H., Lin, L. and Wang, X. (2009) Comparative Genomics of Grasses Promises a Bountiful Harvest. *Plant Physiology*, **149**, 125-131.

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

38. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. and Subgroup, G.P.D.P. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078-2079.

39. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M. *et al.* (2010) The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, **20**, 1297-1303.

40. Barrett, J.C., Fry, B., Maller, J. and Daly, M.J. (2004) Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics*, **21**, 263-265.

41. Mace, E.S., Tai, S., Gilding, E.K., Li, Y., Prentis, P.J., Bian, L., Campbell, B.C., Hu, W., Innes, D.J., Han, X. *et al.* (2013) Whole-genome sequencing reveals untapped genetic potential in Africa's indigenous cereal crop sorghum. *Nature Communication*, **4**, 2320.

42. Yang, J., Lee, S.H., Goddard, M.E. and Visscher, P.M. (2011) GCTA: A Tool for Genome-wide Complex Trait Analysis. *The American Journal of Human Genetics*, **88**, 76-82.

43. Lipka, A.E., Tian, F., Wang, Q., Peiffer, J., Li, M., Bradbury, P.J., Gore, M.A., Buckler, E.S. and Zhang, Z. (2012) GAPIT: genome association and prediction integrated tool. *Bioinformatics*, **28**, 2397-2399.

44. Jin, J., Tian, F., Yang, D.-C., Meng, Y.-Q., Kong, L., Luo, J. and Gao, G. (2016) PlantTFDB 4.0: toward a central hub for transcription factors and regulatory interactions in plants. *Nucleic Acids Research*, **45**, D1040-D1045.

45. Letunic, I. and Bork, P. (2016) Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Research*, **44**, W242-W245.

46. Kim, D., Langmead, B. and Salzberg, S.L. (2015) HISAT: a fast spliced aligner with low memory requirements. *Nature Methods*, **12**, 357-360.

47. Roberts, A., Trapnell, C., Donaghey, J., Rinn, J.L. and Pachter, L. (2011) Improving RNA-Seq expression estimates by correcting for fragment bias. *Genome Biology*, **12**, R22.

48. Anders, S., Pyl, P.T. and Huber, W. (2014) HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics*, **31**, 166-169.

49. Shen, X., Liu, Z.-Q., Mocoeur, A., Xia, Y. and Jing, H.-C. (2015) PAV markers in Sorghum bicolour: genome pattern, affected genes and pathways, and genetic linkage map construction. *Theoretical and Applied Genetics*, **128**, 623-637.

50. Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T. *et al.* (2011) The variant call format and VCFtools. *Bioinformatics*, **27**, 2156-2158.

51. Hartl, D.L., Clark, A.G. and Clark, A.G. (1997) *Principles of population genetics*. Sinauer associates Sunderland, MA.

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

53. Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, **32**, 1792-1797.

54. Felsenstein, J. (1993) *PHYLIP (phylogeny inference package), version 3.5 c*. Joseph Felsenstein.

55. Schliep, K.P. (2010) phangorn: phylogenetic analysis in R. *Bioinformatics*, **27**, 592-593.

56. Raj, A., Stephens, M. and Pritchard, J.K. (2014) fastSTRUCTURE: variational inference of population structure in large SNP data sets. *Genetics*, **197**, 573-589.

57. Haas, B.J., Delcher, A.L., Wortman, J.R. and Salzberg, S.L. (2004) DAGchainer: a tool for mining segmental genome duplications and synteny. *Bioinformatics*, **20**, 3643-3646.

58. Yang, Z. (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. *Bioinformatics*, **13**, 555-556.

59. Chin, C.-S., Peluso, P., Sedlazeck, F.J., Nattestad, M., Concepcion, G.T., Clum, A., Dunn, C., O'Malley, R., Figueroa-Balderas, R., Morales-Cruz, A. *et al.* (2016) Phased diploid genome assembly with single-molecule real-time sequencing. *Nature Methods*, **13**, 1050-1054.

60. Liu, W., Xie, Y., Ma, J., Luo, X., Nie, P., Zuo, Z., Lahrmann, U., Zhao, Q., Zheng, Y., Zhao, Y. *et al.* (2015) IBS: an illustrator for the presentation and visualization of biological sequences. *Bioinformatics*, **31**, 3359-3361.

CHAPTER 7

Discussion

7.1 Contributions and limitations of the thesis

7.1.1 Investigating amino acid repeats across the universal protein sequences

This thesis starts from a systematic review (***Chapter 2***) of the biological significance of repeat patterns in protein sequences. As multiple repeat patterns are commonly intertwined within one repeat-containing protein (RCP), no uniform algorithm can uncover all cryptic repeat patterns. We explicitly defined and classified amino acid repeats based on repeating unit features, considering their sequence patterns and potential biological significances. We reviewed multiple state-of-the-art amino acid repeat detection algorithms and tools and developed a novel pipeline to recognize and integrate repeat patterns in protein sequences and construct the ProRepeat database (***Chapter 3***).

Unlike most previous databases containing specific repeat patterns, ProRepeat contains non-redundant perfect tandem repeats, approximate tandem repeats and simple, low-complexity sequences, covering the majority of the amino acid tandem repeat patterns found in proteins. We aim to identify as many repeat fragments as possible by applying multiple algorithms on the same protein, merging or distinguishing them based on repeat positions and unit patterns. ProRepeat found ~3.75 million repeat fragments in two million proteins from 0.1 million organisms (1). It is one of the few curated databases (1,2), allowing researchers to make comparative genomics analysis on multiple repeat properties such as repeat length, RCP length, repeat position and repeat codon usage in model organisms across different kingdoms.

Most of the early studies on RCP focus on specific repeat patterns in limited species (3-6), merely offering fragmentary pieces of the whole puzzle. ProRepeat can help researchers uncover the complete picture of protein repeats' functional and evolutionary roles in the perspective of universal proteomes across all kingdoms of life. For example, the universal comparison among Eukaryota, Bacteria, Archaea and Viruses performed with ProRepeat supports the idea that large numbers of protein repeats arose after the divergence of Prokaryota and Eukaryota. Previous studies showed a higher percentage of proteins containing repeats within the *Drosophila* genus than in most of other eukaryotes (7). With ProRepeat, we could further identify that *Drosophila melanogaster* uses polyglutamine more frequently than other organisms.

ProRepeat also collected the corresponding coding DNA sequences of the repeat fragments, which is essential to infer their evolutionary backgrounds. Taking the example of the simplest repeat pattern, polyglutamine, which could play different roles in different proteins encoded by different genes or even in orthologous genes in different organisms (3,6-8). Previous studies identified long and conserved tracts of polyglutamine within the protein coded by human FOXP2, the first gene presumed to be involved in the development of linguistic capacity (9). The mixture of CAG and CAA codons of the polyglutamine tracts is one of the critical pieces of evidence to support their functional role (10), as they have a good chance to escape form the DNA slippage mechanism (11), keeping the conserved functions of the polyglutamine tract in normal individuals. Using ProRepeat, it is convenient to study such repeats and genes.

7.1.2 From comparative genomics to comparative population genomics

Since the accomplishment of sequencing projects for human and other model organisms, methods based on comparative genomics became dominant and successful in locating novel functional elements, illuminating the evolutionary history of known functional elements, and estimating the percentage of functional sites in a genome (12-14). Comparative genomics seeks patterns of evolutionary conservation in aligned sequences between species to identify functional elements (15). It assumes that the mutations in functional elements should be removed by purifying selection, leaving a signature of sequence conservation between species. While traditional species-to-species comparative genomics methods serve to detect functional elements maintained over a long evolutionary history, they are not very sensitive to recently arisen functionality (16). For example, we found that the conservation of some repeat patterns among species is low, although some of these are involved in functional motifs of domains in model organisms. This makes it challenging to deduce their biological roles by traditional comparative genomics. Tandem repeats in proteins have been shown to play essential roles in micro-evolution by catalyzing the rapid production of genetic and phenotypic variation among organisms (7,17-19). Thus, the polymorphism of repeats within a species could offer a snapshot of the evolutionary history.

With the development of NGS technology, whole-genome resequencing generates a large amount of population variation data at a low cost. This boosted comparative population genomics, which integrates and extends methods and paradigms from traditional comparative genomics, population genetics, and evolutionary biology. It represents an exciting new prospect for detecting functional genomic elements, describing their functional relevance, and imputing their evolutionary history (16). Early successful studies in animal (20) and plant (21) demonstrated the utility of comparative population genomics methodology. Recent studies highlighted its power to answer old questions with new accumulated multispecies population variation datasets (22). Research on amino acid repeats also benefited from this: scientists have recently demonstrated that the length of the polyglutamine repeat in the *ELF3* gene correlates with its thermal responsiveness in *Arabidopsis* (23), and that protein-coding repeat polymorphisms strongly shaped diverse human phenotypes (24).

7.1.3 Practice on sorghum comparative population genomics

Sorghum (*Sorghum bicolor*) is the fifth most important cereal crop globally and is among the first ten published plant genomes (25). It has a medium-size diploid genome (~730 Mb) among plants (the median size of sequenced plant genome is 575.5 Mb (26)). Compared to the highly repetitive maize genome (27) and large hexaploid wheat genome (28), the complexity of the sorghum genome is substantially lower. Nevertheless, research on sorghum functional genomics has lagged behind other major crops for a long time. A noted example is the single *Dry* locus controlling sorghum stem juiciness (29): it took over a century to identify the gene responsible for this agriculturally and commercially important trait.

In ***Chapter 6***, we successfully decoded the *Dry* gene by comprehensive comparative population genomics approaches. We first resequenced 241 sorghum accessions with a broad geographical distribution, including wild and cultivated lines, which generated one of the most extensive sorghum variome datasets (30-34). We assessed and integrated state-of-the-art pipelines (35-37) and tuned the parameters for sorghum. We then performed GWAS analysis on the natural population of 241 accessions and traditional map-based cloning analysis on the segregation populations derived from crosses between pithy and juicy sorghums. The results demonstrated that GWAS works better than map-based cloning, as it can narrow down *Dry* locus directly to a single gene. It also demonstrated the advantage of the population genomics approach based on the high-density SNPs generated by NGS technology, compared to the traditional labor-consuming approach based on low-density restriction fragment length polymorphisms (RFLP) and simple sequence repeats (SSR) markers.

We next attempted to trace the domestication history of the *Dry* gene in sweet sorghum, which is crucial for sorghum biomass and yield breeding applications. We employed multiple classical comparative population genomics approaches and tools. The *Dry* gene's haplotype analysis highlighted its high degree of polymorphism in the cultivated sorghum population, indicating the polyphyletic origin of sweet sorghum. However, the complete lack of juicy haplotypes in wild sorghum strongly supports its occurrence due to relatively recent and intensive breeding selection, primarily for syrup production. We found strong evidence for this by identifying a strong selective sweep signal between wild and cultivated sorghums within the *Dry* gene region.

Originating in Africa and subsequently spreading to different continents, sorghum has experienced multiple attempts at domestication and intensive breeding selection for various end uses. However, how these processes have shaped sorghum genomes is not fully understood. Our findings pave the way for developing multipurpose sorghum varieties with juicy stems for biofuel production and other applications concomitant with high grain yield. In addition, the conservation of the *Dry* gene in cereals shows its potential use in other crops. In addition to characterizing the *Dry* gene, our work provides tantalizing historical insights into the domestication of sorghum.

In a follow-up study published recently (38), we extend the variome data panel containing 445 sorghum accessions, covering more wild sorghum and four end-use subpopulations with diverse agronomic traits. The population admixture analysis shows the frequent genetic exchanges and gene flows among major sorghum subpopulations. The whole-genome selective signatures during sorghum domestication and improvement demonstrate that the *Dry* gene, which regulates stem juiciness, was unintentionally selected during the improvement of grain sorghum. Moreover, we extend the comparative population genomics analysis to more domestication and breeding genes in sorghum and other cereal crops. Via the haplotype analysis of two genes, *Sh1* and *SbTB1*, which control crucial domestication syndromes in sorghum, we propose the domestication model of sorghum. Taken together, these findings provide new genomic insights into sorghum domestication and breeding selection and will facilitate further dissection of the domestication and molecular breeding of sorghum.

7.1.4 Principles to construct longevous biological databases

This PhD thesis presented two biological databases, ProRepeat and SorGSD (***Chapter 3 and Chapter 5***). ProRepeat was first published in 2011 (2) and classified in the "databases of individual protein families" category by *Nucleic Acids Research Database Issue* and the online Molecular Biology Database Collection (39). All the databases collected by the issue are expected to be maintained under the same URL for at least five years after the publication. Graduation or retirement of the database developers is not a valid reason for the termination of the database (https://academic.oup.com/nar/pages/Ms_Prep_Database).

ProRepeat was built on an ORACLE database using an academic license, which is essentially a commercial system with embedded schema, packages and functions. Although ORACLE is a powerful system offering many data mining, optimization and tuning tools, it was not an open-source software with sufficient flexibility to migrating, upgrading and maintaining. For example, the server running ProRepeat suffered a hard disk failure which damaged the database system in the fifth years since it published. The recovery from a backup was hindered by technical problems as our academic license of ORACLE has expired. Therefore, we only recovered part of the original functions of ProRepeat. The lessons we learn from this are not uncommon for biological databases. All amino acid repeat databases we reviewed in ***Chapter 3*** are not available anymore. Currently, two newly constructed protein tandem repeat databases (PRDB (1) and RepeatsDB (40)) are running well.

SorGSD is a comprehensive web portal providing access to a database of large-scale genome variation across all racial types of cultivated sorghum and wild relatives. It contains the variome data from 48 sorghum accessions published in two independent studies (31,32). Taking lessons from ProRepeat, we employed open-source software (MySQL, Apache/Tomcat web server) and development tools (JAVA/JSP), which are flexible to deploy and come with support from the development community. SorGSD follows the uniform database constructing framework of biodiversity resources maintained by the National Genomics Data Center (NGDC), China National Center for Bioinformation (CNCB) (41) to support its longevity and renewability. Recently, an updated version of SorGSD was published (42). The variome data was expanded to 289 sorghum lines with SNPs and small insertions/deletions (INDELs), aligned to the newly assembled and annotated sorghum genome BTx623 (v3.1). In addition, we added phenotypic data, implemented new tools including ID conversion, homologue search and genome browser, and updated the general information related to sorghum research.

SorGSD provides a valuable resource for sorghum researchers to find variations they are interested in and download high-throughput datasets for further analysis.

7.2 Challenges and opportunities in plants genomics

7.2.1 Development of plant genome sequencing

After the first *Arabidopsis thaliana* genome was released in 2000 (43), only a few model plant and crop genomes were published during the first decade (2000-2010). Sorghum (44) was among these, which was considered a significant breakthrough because it provided an essential foundation and gave new insights for subsequent research. Early plant genome projects utilized bacterial artificial chromosomes (BACs) sequenced with the Sanger technology, both labor-intensive and time-consuming. NGS technology significantly increased sequencing capacity and lowered cost, leading to expanded growth in the number of sequenced genomes. Nowadays, more than 1,000 genomes of nearly 800 plant species have been sequenced and published (26). The availability of these genomes, especially the high-quality ones, has facilitated studies of plant population genetics and functional genomics. In addition to a substantial increase in quantity, the quality of reference genomes has also improved. For sorghum, the improved reference genome identified 24% new genes and ~29.6 Mb more sequences with a tenfold reduction of genome error rate (45).

NGS technology has advanced the study of plant genomes and has provided insights into diversity and evolution. However, plant genomes more than animal genomes are often characterized by a high polyploidy, heterozygosity, and repetitiveness (46). This leads to erroneous or incomplete assemblies of genomes generated by NGS. Repetitiveness is one of the significant barriers to reconstructing complete chromosomes. For example, nested long terminal repeats (LTRs) found in a genome can span 20-200 kb, which is beyond NGS reads' resolving capacity. High heterozygosity, with numerous SNPs and SV between homologous chromosomal regions in a diploid or polyploid species, can drastically impede genome assembly algorithms (47). The centromeres, telomeres, and ribosomal DNA repeats remained unassembled in the early assemblies of the *Arabidopsis* genome (43), although it is small and not very repetitive compared to many other plants. Hundreds of plant genomes are still in a draft state, containing numerous unfilled gaps and thousands of unoriented and unplaced contigs (scaffolds).

7.2.2 Obtaining high-quality plant genomes via technology revolution

Third-generation sequencing (TGS) technologies (48) overcome the major limiting factor of NGS, namely short read-length. Two TGS platforms, PacBio and ONT, are currently available, enabling routine generation of read N50 lengths in the 20-50 kb range, that can span large repetitive regions where short reads tend to fail. Once again, technologies are revolutionizing plant genome sequencing and assembly. Nowadays, it is routine to adopt hybrid sequencing strategies that combine short reads, long reads with different insertion sizes, and other scaffolding approaches that provide further genomic information, such as 10X (49), Hi-C (50), BioNano (51), and Strand-Seq (52). Although the long reads generated by the early versions of both platforms have a higher error rate than previous NGS sequencing platforms, they have facilitated genome assemblies with significantly increased genome contiguity, or completeness, compared to previous technologies (53). For example, a sorghum genome assembled by ONT reads reaches a scaffold N50 of 33.28 Mbps and covers 90 percent of the expected genome length (54). Other studies demonstrate high-quality assembly for three even larger plant genomes using ONT and BioNano methods (55). Moreover, many efforts are taken to alleviate the consequences of high error-rate long reads by either experimental or software strategies. For example, PacBio updated its circular consensus sequencing (CCS) approach to generate long high fidelity (HiFi) 15 kb reads with accuracy upwards of 99.8% at a 5x cost increase trade-off (56). Several plant genomes (57-59) have already been sequenced with HiFi reads and exhibited high quality and robustness.

With the improvement of TGS and scaffolding approaches, many draft plant genomes are expected to become finished genomes soon. An update to the *Arabidopsis* genome using ONT spanned chromosome arms (telomere to centromere) and resolved previously identified gaps (60). Soon after, another study corrected previous assembly errors caused by BAC sequencing due to a long repeat structure in the Arabidopsis genome for relatively little cost (61). The latest assembly of *Arabidopsis* accession Columbia (Col-CEN) resolves all five centromeres to derive insights into the chromatin and recombination landscapes within the *Arabidopsis* centromeres and how these regions evolve (62). The next target for plant genome sequencing projects is to build near-complete genomes (63). Except for *Arabidopsis*, gapless assemblies of more complex crop genomes were published, including rice (64,65), maize (66) and *Ginkgo biloba* (67). Meanwhile, the scale of plant genome sizes is remarkable, with the currently largest know genome (*Paris japonica*) having a size of 150 Gb (68), more than 1000-fold the size of the *Arabidopsis* genome. However, the biggest plant genome sequenced thus far is that of sugar pine at 31 Gb with a contig N50 of only 4.25 Kb (69). The advanced sequencing technologies also allow us to access these more complex genomes. PacBio and Hi-C enabled assembly and phasing of the octoploid sugarcane (70) and the even more complex allotetraploid peanut (71), teff (72), and broomcorn millet genomes (73).

7.2.3 Building plant reference pan-genomes and identifying structure variation

Early resequencing efforts in *Arabidopsis* (74,75) have demonstrated the vast genetic diversity across diverse accessions in plant species. The high degree of genomic variation observed led to the realization that a single reference genome is not adequate to represent the complete genomic repertoire of a species. A more robust and comprehensive approach to studying plant genomes is pan-genomics, which aims to capture all variation in a species. The pan-genome concept was introduced for the genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae* (76), represented as the nonredundant collection of genes and DNA sequences in a species

(77). A pan-genome broadly comprises of the core genome and the dispensable genome. In plants, the core genome is indispensable for survival; the dispensable genome may contain genes responsible for adaptation and survival in different environments (78). Comparison of the core genome and the dispensable genome among wild species and cultivated species can help uncover the effect of domestication (79).

The construction and interpretation of plant pan-genomes is generally more challenging than in species from other kingdoms. This is partly due to the presence of comprehensive structural variation (SV) in the form of copy number variants (CNVs), presence/absence variants (PAVs), and large-scale chromosomal rearrangements among the genomes in many plant populations (80). Recent studies show that SVs influence agronomic phenotypes and suggest that SVs induce gene fusion events (81). SVs can also be used as markers, to resolve complex haplotypes missed by traditional SNP-based GWAS. Identifying functional, evolutionarily relevant SVs is becoming a significant task in plant pan-genomes studies. In early plant genomic studies, SVs were typically identified by resequencing genomes of individuals using short-read sequencing technology, and comparing these to a reference genome sequence. Short-read approaches, however, have been reported to lack sensitivity (only 10% to 70% of SVs detected), exhibit very high false-positive rates (up to 89%), and misinterpret complex or nested SVs (82). To overcome these issues, some efforts were made to combine multiple SV-calling algorithms into a single pipeline to generate a unified SV call set comprising primarily overlapping calls (83-85).

Pan-genomic research in sorghum lags behind other major crops. As the first step to improve this, we recently published a sorghum pan-genome resource (86). We assembled 13 sorghum genomes representing cultivated sorghum and its wild relatives, and integrated these with three other published genomes to generate a pan-genome of 44,079 gene families with 222.6 Mb of the new sequence identified. We also identified SVs under selection during sorghum domestication and improvement, and demonstrated that this variation had important phenotypic outcomes to improve the crop.

7.3 Leveraging bioinformatics for future plant omics

7.3.1 Towards long-read sequencing

Advances in DNA sequencing and high-throughput omics approaches provide researchers with a wealth of population-scale information. Bioinformatics has a fundamental role in exploiting and integrating this fast-accumulating wealth of data. Several large plant genome-sequencing projects, including the 10KP (10,000 Plants) (87) and Earth BioGenome projects (88,89), were launched with the aim to sequence, catalogue, and characterize biodiversity. Population scale long-read-based *de novo* assembly will likely replace short-read-based whole-genome resequencing for pan-genome and population genetics analyses (90). The increased availability and affordability of long-read sequencing data require new genome assembly tools, focusing on computational demand, contiguity, completeness, and correctness (91,92). Moreover, the field of sequencing and building plant pan-genomes is still in its infancy. The golden standards for a plant pan-genome are yet to be established. Questions about efficient data structures, algorithms, and statistical methods to perform bioinformatic analyses of pan-genomes gave rise to the discipline of computational pan-genomics (93).

Recent advances in genomic technologies, particularly long-read sequencing and whole-genome mapping, promise the production of high-quality plant genome and pan-genome assemblies and access to a broad range of SVs to assess their potential role in plant phenotypic variation (94,95). A new strategy of constructing an integrated reference pan-genome using a graph methodology is promising for high-quality SVs identification in population-scale (96). In the graph-based pan-genome, the reference and alternative genetic variants determined through genomic comparison are recorded as nodes and edges of a graph, respectively. The graph-based pan-genome enables fast and accurate computation for reads mapping and variation calling, but it is less readable by traditional approaches. Early tools have been developed to craft genome graphs in human (97,98). However, the tools and pipelines for plants are still not yet as mature as those for linear reference genome analyses. In Wageningen the PanTools software package (99) is developed that offers functionality to construct and annotate pan-genomes as well as sequence and homology search functions.

7.3.2 Facing new big data

Data-driven progress is one of the distinct advantages of bioinformatics. New data generated by new technologies helps answer old but significant biological questions or overturn classical but misleading concepts. For example, one of challenges of the protein repeat research is the limited number of manually reviewed records in the UniProt database. Most protein sequences were predicted by automatic annotation pipelines, possibly based on erroneous genome sequencing and assembly, imposing multi-level challenges for genome and protein databases (100). Upcoming technology promises to sequence single proteins at single-amino acid resolution (101). Furthermore, new deep learning algorithms can make highly accurate protein structure predictions by incorporating physical, biological, and bioinformatics knowledge (102,103). Using these high-quality data, the research on protein repeats on a large population scale should become more reliable (24).

Many classical tools and algorithms are resurgent, promoted by the newly generated high-quality data. For example, the widely used MUMmer software for genome alignment has been updated to a new version after more than a decade of stasis. It can now work with input sequences of any biologically realistic length by altering the 32-bit suffix tree data structure to a 48-bit suffix array (104). The graph extension of the classic Burrows-Wheeler transform (BWT) algorithm is used to store the haplotypes and make the index for accelerating read mapping speed on a graph pan-genome (105,106).

The rapid increase in availability of various omics technologies also leads to the dissemination of poorly curated datasets in the form of raw collections or preliminary draft results. Consequently, this can affect the establishment of stable and reliable resources. Advances in gene annotation have lagged behind improvements in genome assembly, and generating accurate gene predictions is still a major limitation. Improving annotation quality will require new technologies and new algorithms to make better predictions of functional genomic elements. Currently, handling the datasets from hundreds of samples is still a huge task, particularly for species with large genomes, such as wheat. To deal with 10K or more genomes from different plant species will be a challenge in the near future. While the development of computing technology may partially solve this problem, new strategies are needed to store, present, and analyze the new big data.

7.3.3 Integrating and interpreting multi-omics data

Nowadays, the target of plant omics is shifting from harnessing the potential of modern genome resources and characterizing allelic variations, to creating novel diversity and facilitating their rapid and efficient incorporation in crop improvement programs (107). Some pioneering works start to design future crops by *de novo* domestication (108), high-throughput phenotyping (109), speed breeding (110), genomics selection (111) and gene editing (112). These works generate multi-omics data across a hierarchy of biological scales, such as genome sequence, epigenomic marks, three-dimensional chromosome conformation, gene expression data, organismal phenotypes, and field ecosystem (113). It is a major task for bioinformatics to integrating and interpreting the large, noisy, and heterogeneous data set by further exploiting advanced computational approaches.

Machine learning is a promising approach to tackle the complexity of the multi-omics data (114). Using deep learning (115), the massive omics data provides a rich training resource. The main architectures of deep learning include feed-forward, convolutional and recurrent (116), which are applicable for different omics data including biological sequences, semantic texts and phenomic images at various biochemical, cellular and macroscopic levels. In plant omics and crop improvement, one of the key steps to design the deep learning approach is understanding and integrating information flow among different data levels. In addition, researchers must be aware of the need for sufficient high-quality labelled data for deep learning. Last but not least, although deep learning can achieve high accuracy, interpreting results is more challenging than standard statistical models. A good practice is to compare against simpler machine learning models on the same dataset.

References

1. Jorda, J., Baudrand, T. and Kajava, A.V. (2012) PRDB: Protein Repeat DataBase. *Proteomics*, **12**, 1333-1336.
2. Luo, H., Lin, K., David, A., Nijveen, H. and Leunissen, J.A.M. (2011) ProRepeat: an integrated repository for studying amino acid tandem repeats in proteins. *Nucleic Acids Research*, **40**, D394-D399.
3. Björklund, A.K., Ekman, D. and Elofsson, A. (2006) Expansion of protein domain repeats. *PLoS Computational Biology*, **2**, e114.
4. Andrade, M.A., Perez-Iratxeta, C. and Ponting, C.P. (2001) Protein Repeats: Structures, Functions, and Evolution. *Journal of Structural Biology*, **134**, 117-131.
5. Haerty, W. and Golding, G.B. (2010) Genome-wide evidence for selection acting on single amino acid repeats. *Genome Research*, **20**, 755-760.
6. Mularoni, L., Ledda, A., Toll-Riera, M. and Albà, M.M. (2010) Natural selection drives the accumulation of amino acid tandem repeats in human proteins. *Genome Research*, **20**, 745-754.
7. Huntley, M.A. and Clark, A.G. (2007) Evolutionary Analysis of Amino Acid Repeats across the Genomes of 12 Drosophila Species. *Molecular Biology and Evolution*, **24**, 2598-2609.
8. Undurraga, S.F., Press, M.O., Legendre, M., Bujdoso, N., Bale, J., Wang, H., Davis, S.J., Verstrepen, K.J. and Queitsch, C. (2012) Background-dependent effects of polyglutamine variation in the *Arabidopsis thaliana* gene *ELF3*. *Proceedings of the National Academy of Sciences of the United States of America*, **109**, 19363-19367.
9. Lai, C.S.L., Fisher, S.E., Hurst, J.A., Vargha-Khadem, F. and Monaco, A.P. (2001) A forkhead-domain gene is mutated in a severe speech and language disorder. *Nature*, **413**, 519-523.
10. Enard, W., Przeworski, M., Fisher, S.E., Lai, C.S.L., Wiebe, V., Kitano, T., Monaco, A.P. and Pääbo, S. (2002) Molecular evolution of FOXP2, a gene involved in speech and language. *Nature*, **418**, 869-872.
11. Levinson, G. and Gutman, G.A. (1987) Slipped-strand mispairing: a major mechanism for DNA sequence evolution. *Molecular Biology and Evolution*, **4**, 203-221.
12. Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., Richards, S. *et al.* (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research*, **15**, 1034-1050.
13. Lindblad-Toh, K., Garber, M., Zuk, O., Lin, M.F., Parker, B.J., Washietl, S., Kheradpour, P., Ernst, J., Jordan, G., Mauceli, E. *et al.* (2011) A high-resolution map of human evolutionary constraint using 29 mammals. *Nature*, **478**, 476-482.
14. Rubin, G.M., Yandell, M.D., Wortman, J.R., Gabor, G.L., Miklos, Nelson, C.R., Hariharan, I.K., Fortini, M.E., Li, P.W., Apweiler, R. *et al.* (2000) Comparative Genomics of the Eukaryotes. *Science*, **287**, 2204-2215.
15. Hardison, R.C. (2003) Comparative genomics. *PLoS Biololgy*, **1**, E58.
16. Lawrie, D.S. and Petrov, D.A. (2014) Comparative population genomics: power and principles for the inference of functionality. *Trends in Genetics*, **30**, 133-139.
17. Caburet, S., Vaiman, D. and Veitia, R.A. (2004) A Genomic Basis for the Evolution of Vertebrate Transcription Factors Containing Amino Acid Runs. *Genetics*, **167**, 1813-1820.
18. Fondon, J.W. and Garner, H.R. (2004) Molecular origins of rapid and continuous morphological evolution. *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 18058-18063.
19. Caburet, S., Cocquet, J., Vaiman, D. and Veitia, R.A. (2005) Coding repeats and evolutionary "agility". *BioEssays*, **27**, 581-587.
20. Romiguier, J., Gayral, P., Ballenghien, M., Bernard, A., Cahais, V., Chenuil, A., Chiari, Y., Dernat, R., Duret, L., Faivre, N. *et al.* (2014) Comparative population genomics in animals uncovers the determinants of genetic diversity. *Nature*, **515**, 261-263.

21. Hufford, M.B., Xu, X., van Heerwaarden, J., Pyhajarvi, T., Chia, J.M., Cartwright, R.A., Elshire, R.J., Glaubitz, J.C., Guill, K.E., Kaeppler, S.M. *et al.* (2012) Comparative population genomics of maize domestication and improvement. *Nature Genetics*, **44**, 808-811.

22. McGrath, C. (2022) Highlight: Comparative Population Genomics—Answering Old Questions with New Data. *Genome Biology and Evolution*, **14**.

23. Jung, J.-H., Barbosa, A.D., Hutin, S., Kumita, J.R., Gao, M., Derwort, D., Silva, C.S., Lai, X., Pierre, E., Geng, F. *et al.* (2020) A prion-like domain in ELF3 functions as a thermosensor in Arabidopsis. *Nature*.

24. Mukamel, R.E., Handsaker, R.E., Sherman, M.A., Barton, A.R., Zheng, Y., McCarroll, S.A. and Loh, P.-R. (2021) Protein-coding repeat polymorphisms strongly shape diverse human phenotypes. *Science*, **373**, 1499-1505.

25. Michael, T.P. and Jackson, S. (2013) The First 50 Plant Genomes. *The Plant Genome*, **6**, plantgenome2013.2003.0001in.

26. Sun, Y., Shang, L., Zhu, Q.-H., Fan, L. and Guo, L. (2021) Twenty years of plant genome sequencing: achievements and challenges. *Trends in Plant Science*.

27. Schnable, P.S., Ware, D., Fulton, R.S., Stein, J.C., Wei, F., Pasternak, S., Liang, C., Zhang, J., Fulton, L., Graves, T.A. *et al.* (2009) The B73 Maize Genome: Complexity, Diversity, and Dynamics. *Science*, **326**, 1112-1115.

28. Avni, R., Nave, M., Barad, O., Baruch, K., Twardziok, S.O., Gundlach, H., Hale, I., Mascher, M., Spannagl, M., Wiebe, K. *et al.* (2017) Wild emmer genome architecture and diversity elucidate wheat evolution and domestication. *Science*, **357**, 93-97.

29. Hilson, G. (1916) On the inheritance of certain stem characters in sorghum. *Agriculture Journal India*, **11**, 150-155.

30. Nelson, J.C., Wang, S., Wu, Y., Li, X., Antony, G., White, F.F. and Yu, J. (2011) Single-nucleotide polymorphism discovery by high-throughput sequencing in sorghum. *BMC Genomics*, **12**, 352.

31. Zheng, L.-Y., Guo, X.-S., He, B., Sun, L.-J., Peng, Y., Dong, S.-S., Liu, T.-F., Jiang, S., Ramachandran, S., Liu, C.-M. *et al.* (2011) Genome-wide patterns of genetic variation in sweet and grain sorghum (Sorghum bicolor). *Genome Biology*, **12**, R114.

32. Mace, E.S., Tai, S., Gilding, E.K., Li, Y., Prentis, P.J., Bian, L., Campbell, B.C., Hu, W., Innes, D.J., Han, X. *et al.* (2013) Whole-genome sequencing reveals untapped genetic potential in Africa's indigenous cereal crop sorghum. *Nature Communications*, **4**, 2320.

33. Bekele, W.A., Wieckhorst, S., Friedt, W. and Snowdon, R.J. (2013) High-throughput genomics in sorghum: from whole-genome resequencing to a SNP screening array. *Plant Biotechnology Journal*, **11**, 1112-1125.

34. Morris, G.P., Ramu, P., Deshpande, S.P., Hash, C.T., Shah, T., Upadhyaya, H.D., Riera-Lizarazu, O., Brown, P.J., Acharya, C.B., Mitchell, S.E. *et al.* (2013) Population genomic and genome-wide association studies of agroclimatic traits in sorghum. *Proceedings of the National Academy of Sciences of the United States of America*, **110**, 453-458.

35. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M. *et al.* (2010) The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, **20**, 1297-1303.

36. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. and Subgroup, G.P.D.P. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078-2079.

37. McCormick, R.F., Truong, S.K. and Mullet, J.E. (2015) RIG: Recalibration and Interrelation of Genomic Sequence Data with the GATK. *G3: Genes|Genomes|Genetics*, **5**, 655-665.

38. Wu, X., Liu, Y., Luo, H., Shang, L., Leng, C., Liu, Z., Li, Z., Lu, X., Cai, H., Hao, H. *et al.* (2022) Genomic footprints of sorghum domestication and breeding selection for multiple end uses. *Molecular Plant*.

39. Galperin, M.Y. and Fernández-Suárez, X.M. (2011) The 2012 Nucleic Acids Research Database Issue and the online Molecular Biology Database Collection. *Nucleic Acids Research*, **40**, D1-D8.

40. Paladin, L., Bevilacqua, M., Errigo, S., Piovesan, D., Mičetić, I., Necci, M., Monzon, A.M., Fabre, M.L., Lopez, Jose L., Nilsson, J.F. *et al.* (2020) RepeatsDB in 2021: improved data and extended classification for protein tandem repeat structures. *Nucleic Acids Research*, **49**, D452-D457.

41. Members, C.-N. and Partners. (2020) Database Resources of the National Genomics Data Center, China National Center for Bioinformation in 2021. *Nucleic Acids Research*, **49**, D18-D28.

42. Liu, Y., Wang, Z., Wu, X., Zhu, J., Luo, H., Tian, D., Li, C., Luo, J., Zhao, W., Hao, H. *et al.* (2021) SorGSD: updating and expanding the sorghum genome science database with new contents and tools. *Biotechnology for Biofuels*, **14**, 165.

43. The Arabidopsis Genome, I. (2000) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature*, **408**, 796-815.

44. Paterson, A.H., Bowers, J.E., Bruggmann, R., Dubchak, I., Grimwood, J., Gundlach, H., Haberer, G., Hellsten, U., Mitros, T., Poliakov, A. *et al.* (2009) The Sorghum bicolor genome and the diversification of grasses. *Nature*, **457**, 551-556.

45. McCormick, R.F., Truong, S.K., Sreedasyam, A., Jenkins, J., Shu, S., Sims, D., Kennedy, M., Amirebrahimi, M., Weers, B.D., McKinley, B. *et al.* (2018) The *Sorghum bicolor* reference genome: improved assembly, gene annotations, a transcriptome atlas, and signatures of genome organization. *The Plant Journal*, **93**, 338-354.

46. Hirsch, C.N. and Buell, C.R. (2013) Tapping the Promise of Genomics in Species with Complex, Nonmodel Genomes. *Annual Review of Plant Biology*, **64**, 89-110.

47. Earl, D., Bradnam, K., St. John, J., Darling, A., Lin, D., Fass, J., Yu, H.O.K., Buffalo, V., Zerbino, D.R., Diekhans, M. *et al.* (2011) Assemblathon 1: A competitive assessment of de novo short read assembly methods. *Genome Research*, **21**, 2224-2241.

48. Yuan, Y., Bayer, P.E., Batley, J. and Edwards, D. (2017) Improvements in Genomic Technologies: Application to Crop Genomics. *Trends in Biotechnology*, **35**, 547-558.

49. Eisenstein, M. (2015) Startups use short-read data to expand long-read sequencing market. *Nature Biotechnology*, **33**, 433-435.

50. Kalhor, R., Tjong, H., Jayathilaka, N., Alber, F. and Chen, L. (2012) Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. *Nature Biotechnology*, **30**, 90-98.

51. Lam, E.T., Hastie, A., Lin, C., Ehrlich, D., Das, S.K., Austin, M.D., Deshpande, P., Cao, H., Nagarajan, N., Xiao, M. *et al.* (2012) Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly. *Nature Biotechnology*, **30**, 771-776.

52. Falconer, E. and Lansdorp, P.M. (2013) Strand-seq: A unifying tool for studies of chromosome segregation. *Seminars in Cell & Developmental Biology*, **24**, 643-652.

53. Jiao, W.-B. and Schneeberger, K. (2017) The impact of third generation genomic technologies on plant genome assembly. *Current Opinion in Plant Biology*, **36**, 64-70.

54. Deschamps, S., Zhang, Y., Llaca, V., Ye, L., Sanyal, A., King, M., May, G. and Lin, H. (2018) A chromosome-scale assembly of the sorghum genome using nanopore sequencing and optical mapping. *Nature Communications*, **9**, 4844.

55. Belser, C., Istace, B., Denis, E., Dubarry, M., Baurens, F.-C., Falentin, C., Genete, M., Berrabah, W., Chèvre, A.-M., Delourme, R. *et al.* (2018) Chromosome-scale assemblies of plant genomes using nanopore long reads and optical maps. *Nature Plants*, **4**, 879-887.

56. Wenger, A.M., Peluso, P., Rowell, W.J., Chang, P.-C., Hall, R.J., Concepcion, G.T., Ebler, J., Fungtammasan, A., Kolesnikov, A., Olson, N.D. *et al.* (2019) Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nature Biotechnology*, **37**, 1155-1162.

57. Hon, T., Mars, K., Young, G., Tsai, Y.-C., Karalius, J.W., Landolin, J.M., Maurer, N., Kudrna, D., Hardigan, M.A., Steiner, C.C. *et al.* (2020) Highly accurate long-read HiFi sequencing data for five complex genomes. *Scientific Data*, **7**, 399.

58. Zhou, Q., Tang, D., Huang, W., Yang, Z., Zhang, Y., Hamilton, J.P., Visser, R.G.F., Bachem, C.W.B., Robin Buell, C., Zhang, Z. *et al.* (2020) Haplotype-resolved genome analyses of a heterozygous diploid potato. *Nature Genetics*, **52**, 1018-1023.

59. Chen, H., Zeng, Y., Yang, Y., Huang, L., Tang, B., Zhang, H., Hao, F., Liu, W., Li, Y., Liu, Y. *et al.* (2020) Allele-aware chromosome-level genome assembly and efficient transgene-free genome editing for the autotetraploid cultivated alfalfa. *Nature Communications*, **11**, 2494.

60. Jupe, F., Rivkin, A.C., Michael, T.P., Zander, M., Motley, S.T., Sandoval, J.P., Slotkin, R.K., Chen, H., Castanon, R., Nery, J.R. *et al.* (2019) The complex architecture and epigenomic impact of plant T-DNA insertions. *PLoS Genetics*, **15**, e1007819.

61. Michael, T.P., Jupe, F., Bemm, F., Motley, S.T., Sandoval, J.P., Lanz, C., Loudet, O., Weigel, D. and Ecker, J.R. (2018) High contiguity Arabidopsis thaliana genome assembly with a single nanopore flow cell. *Nature Communications*, **9**, 541.

62. Naish, M., Alonge, M., Wlodzimierz, P., Tock, A.J., Abramson, B.W., Schmücker, A., Mandáková, T., Jamge, B., Lambing, C., Kuo, P. *et al.* (2021) The genetic and epigenetic landscape of the *Arabidopsis* centromeres. *Science*, **374**, eabi7489.

63. Michael, T.P. and VanBuren, R. (2020) Building near-complete plant genomes. *Current Opinion in Plant Biology*, **54**, 26-33.

64. Li, K., Jiang, W., Hui, Y., Kong, M., Feng, L.-Y., Gao, L.-Z., Li, P. and Lu, S. (2021) Gapless indica rice genome reveals synergistic contributions of active transposable elements and segmental duplications to rice genome evolution. *Molecular Plant*, **14**, 1745-1756.

65. Song, J.-M., Xie, W.-Z., Wang, S., Guo, Y.-X., Koo, D.-H., Kudrna, D., Gong, C., Huang, Y., Feng, J.-W., Zhang, W. *et al.* (2021) Two gap-free reference genomes and a global view of the centromere architecture in rice. *Molecular Plant*, **14**, 1757-1767.

66. Liu, J., Seetharam, A.S., Chougule, K., Ou, S., Swentowsky, K.W., Gent, J.I., Llaca, V., Woodhouse, M.R., Manchanda, N., Presting, G.G. *et al.* (2020) Gapless assembly of maize chromosomes using long-read technologies. *Genome Biology*, **21**, 121.

67. Liu, H., Wang, X., Wang, G., Cui, P., Wu, S., Ai, C., Hu, N., Li, A., He, B., Shao, X. *et al.* (2021) The nearly complete genome of *Ginkgo biloba* illuminates gymnosperm evolution. *Nature Plants*, **7**, 748-756.

68. Michael, T.P. (2014) Plant genome size variation: bloating and purging DNA. *Briefings in Functional Genomics*, **13**, 308-317.

69. Stevens, K.A., Wegrzyn, J.L., Zimin, A., Puiu, D., Crepeau, M., Cardeno, C., Paul, R., Gonzalez-Ibeas, D., Koriabine, M., Holtz-Morris, A.E. *et al.* (2016) Sequence of the Sugar Pine Megagenome. *Genetics*, **204**, 1613-1626.

70. Zhang, J., Zhang, X., Tang, H., Zhang, Q., Hua, X., Ma, X., Zhu, F., Jones, T., Zhu, X., Bowers, J. *et al.* (2018) Allele-defined genome of the autopolyploid sugarcane Saccharum spontaneum L. *Nature Genetics*, **50**, 1565-1573.

71. Bertioli, D.J., Jenkins, J., Clevenger, J., Dudchenko, O., Gao, D., Seijo, G., Leal-Bertioli, S.C.M., Ren, L., Farmer, A.D., Pandey, M.K. *et al.* (2019) The genome sequence of segmental allotetraploid peanut Arachis hypogaea. *Nature Genetics*, **51**, 877-884.

72. VanBuren, R., Man Wai, C., Wang, X., Pardo, J., Yocca, A.E., Wang, H., Chaluvadi, S.R., Han, G., Bryant, D., Edger, P.P. *et al.* (2020) Exceptional subgenome stability and functional divergence in the allotetraploid Ethiopian cereal teff. *Nature Communications*, **11**, 884.

73. Zou, C., Li, L., Miki, D., Li, D., Tang, Q., Xiao, L., Rajput, S., Deng, P., Peng, L., Jia, W. *et al.* (2019) The genome of broomcorn millet. *Nature Communications*, **10**.

74. Cao, J., Schneeberger, K., Ossowski, S., Günther, T., Bender, S., Fitz, J., Koenig, D., Lanz, C., Stegle, O., Lippert, C. *et al.* (2011) Whole-genome sequencing of multiple Arabidopsis thaliana populations. *Nature Genetics*, **43**, 956-963.

75. Gan, X., Stegle, O., Behr, J., Steffen, J.G., Drewe, P., Hildebrand, K.L., Lyngsoe, R., Schultheiss, S.J., Osborne, E.J., Sreedharan, V.T. *et al.* (2011) Multiple reference genomes and transcriptomes for Arabidopsis thaliana. *Nature*, **477**, 419-423.

76. Tettelin, H., Masignani, V., Cieslewicz, M.J., Donati, C., Medini, D., Ward, N.L., Angiuoli, S.V., Crabtree, J., Jones, A.L., Durkin, A.S. *et al.* (2005) Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: Implications for the microbial "pan-genome". *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 13950-13955.

77. Lei, L., Goltsman, E., Goodstein, D., Wu, G.A., Rokhsar, D.S. and Vogel, J.P. (2021) Plant Pan-Genomics Comes of Age. *Annual Review of Plant Biology*, **72**, 411-435.

78. Khan, A.W., Garg, V., Roorkiwal, M., Golicz, A.A., Edwards, D. and Varshney, R.K. (2020) Super-Pangenome by Integrating the Wild Side of a Species for Accelerated Crop Improvement. *Trends in Plant Science*, **25**, 148-158.

79. Liu, Y., Du, H., Li, P., Shen, Y., Peng, H., Liu, S., Zhou, G.-A., Zhang, H., Liu, Z., Shi, M. *et al.* (2020) Pan-Genome of Wild and Cultivated Soybeans. *Cell*, **182**, 162-176.e113.

80. Della Coletta, R., Qiu, Y., Ou, S., Hufford, M.B. and Hirsch, C.N. (2021) How the pan-genome is changing crop genomics and improvement. *Genome Biol*, **22**, 3.

81. Willson, J. (2020) Resolving the roles of structural variants. *Nature Reviews Genetics*, **21**, 507-507.

82. Sedlazeck, F.J., Rescheneder, P., Smolka, M., Fang, H., Nattestad, M., von Haeseler, A. and Schatz, M.C. (2018) Accurate detection of complex structural variations using single-molecule sequencing. *Nature Methods*, **15**, 461-468.

83. Lin, K., Smit, S., Bonnema, G., Sanchez-Perez, G. and de Ridder, D. (2014) Making the difference: integrating structural variation detection tools. *Briefings in Bioinformatics*, **16**, 852-864.

84. Becker, T., Lee, W.-P., Leone, J., Zhu, Q., Zhang, C., Liu, S., Sargent, J., Shanker, K., Mil-homens, A., Cerveira, E. *et al.* (2018) FusorSV: an algorithm for optimally combining data from multiple structural variation detection methods. *Genome Biology*, **19**, 38.

85. Wijfjes, R.Y., Smit, S. and de Ridder, D. (2019) Hecaton: reliably detecting copy number variation in plant genomes using short read sequencing data. *BMC Genomics*, **20**, 818.

86. Tao, Y., Luo, H., Xu, J., Cruickshank, A., Zhao, X., Teng, F., Hathorn, A., Wu, X., Liu, Y., Shatte, T. *et al.* (2021) Extensive variation within the pan-genome of cultivated and wild sorghum. *Nature Plants*.

87. Cheng, S., Melkonian, M., Smith, S.A., Brockington, S., Archibald, J.M., Delaux, P.-M., Li, F.-W., Melkonian, B., Mavrodiev, E.V., Sun, W. *et al.* (2018) 10KP: A phylodiverse genome sequencing plan. *GigaScience*, **7**.

88. Kress, W.J., Soltis, D.E., Kersey, P.J., Wegrzyn, J.L., Leebens-Mack, J.H., Gostel, M.R., Liu, X. and Soltis, P.S. (2022) Green plant genomes: What we know in an era of rapidly expanding opportunities. *Proceedings of the National Academy of Sciences of the United States of America* **119**, e2115640118.

89. Lewin, H.A., Robinson, G.E., Kress, W.J., Baker, W.J., Coddington, J., Crandall, K.A., Durbin, R., Edwards, S.V., Forest, F., Gilbert, M.T.P. *et al.* (2018) Earth BioGenome Project: Sequencing life for the future of life. *Proceedings of the National Academy of Sciences of the United States of America*, **115**, 4325-4333.

90. De Coster, W., Weissensteiner, M.H. and Sedlazeck, F.J. (2021) Towards population-scale long-read sequencing. *Nature Reviews Genetics*, **22**, 572-587.

91. Marx, V. (2021) Long road to long-read assembly. *Nature Methods*.

92. Sedlazeck, F.J., Lee, H., Darby, C.A. and Schatz, M.C. (2018) Piercing the dark matter: bioinformatics of long-range sequencing and mapping. *Nature Reviews Genetics*, **19**, 329-346.

93. Consortium, T.C.P.-G. (2016) Computational pan-genomics: status, promises and challenges. *Briefings in Bioinformatics*, **19**, 118-135.

94. Yuan, Y., Bayer, P.E., Batley, J. and Edwards, D. (2021) Current status of structural variation studies in plants. *Plant Biotechnology Journal*, **19**, 2153-2163.

95. Gabur, I., Chawla, H.S., Snowdon, R.J. and Parkin, I.A.P. (2019) Connecting genome structural variation with complex traits in crop plants. *Theoretical and Applied Genetics*, **132**, 733-750.

96. Eizenga, J.M., Novak, A.M., Sibbesen, J.A., Heumos, S., Ghaffaari, A., Hickey, G., Chang, X., Seaman, J.D., Rounthwaite, R., Ebler, J. *et al.* (2020) Pangenome Graphs. *Annual Review of Genomics and Human Genetics*, **21**, 139-162.

97. Garrison, E., Sirén, J., Novak, A.M., Hickey, G., Eizenga, J.M., Dawson, E.T., Jones, W., Garg, S., Markello, C., Lin, M.F. *et al.* (2018) Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nature Biotechnology*, **36**, 875-879.

98. Eggertsson, H.P., Kristmundsdottir, S., Beyter, D., Jonsson, H., Skuladottir, A., Hardarson, M.T., Gudbjartsson, D.F., Stefansson, K., Halldorsson, B.V. and Melsted, P. (2019) GraphTyper2 enables population-scale genotyping of structural variation using pangenome graphs. *Nature Communications*, **10**, 5402.

99. Sheikhizadeh, S., Schranz, M.E., Akdel, M., de Ridder, D. and Smit, S. (2016) PanTools: representation, storage and exploration of pan-genomic data. *Bioinformatics*, **32**, i487-i493.

100. Tørresen, O.K., Star, B., Mier, P., Andrade-Navarro, M.A., Bateman, A., Jarnot, P., Gruca, A., Grynberg, M., Kajava, A.V., Promponas, V.J. *et al.* (2019) Tandem repeats lead to sequence assembly errors and impose multi-level challenges for genome and protein databases. *Nucleic Acids Research*, **47**, 10994-11006.

101. Brinkerhoff, H., Kang, A.S.W., Liu, J., Aksimentiev, A. and Dekker, C. (2021) Multiple rereads of single proteins at single-amino acid resolution using nanopores. *Science*, **374**, 1509-1513**.**

102. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A. *et al.* (2021) Highly accurate protein structure prediction with AlphaFold. *Nature*.

103. Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G.R., Wang, J., Cong, Q., Kinch, L.N., Schaeffer, R.D. *et al.* (2021) Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, eabj8754.

104. Marçais, G., Delcher, A.L., Phillippy, A.M., Coston, R., Salzberg, S.L. and Zimin, A. (2018) MUMmer4: A fast and versatile genome alignment system. *PLoS Computational Biology*, **14**, e1005944.

105. Sirén, J., Monlong, J., Chang, X., Novak, A.M., Eizenga, J.M., Markello, C., Sibbesen, J.A., Hickey, G., Chang, P.-C., Carroll, A. *et al.* (2021) Pangenomics enables genotyping of known structural variants in 5202 diverse genomes. *Science*, **374**, abg8871.

106. Sirén, J., Garrison, E., Novak, A.M., Paten, B. and Durbin, R. (2019) Haplotype-aware graph indexes. *Bioinformatics*, **36**, 400-407.

107. Varshney, R.K., Bohra, A., Yu, J., Graner, A., Zhang, Q. and Sorrells, M.E. (2021) Designing Future Crops: Genomics-Assisted Breeding Comes of Age. *Trends in Plant Science*, **26**, 631-649.

108. Yu, H., Lin, T., Meng, X., Du, H., Zhang, J., Liu, G., Chen, M., Jing, Y., Kou, L., Li, X. *et al.* (2021) A route to *de novo* domestication of wild allotetraploid rice. *Cell*.

109. Yang, W., Feng, H., Zhang, X., Zhang, J., Doonan, J.H., Batchelor, W.D., Xiong, L. and Yan, J. (2020) Crop Phenomics and High-Throughput Phenotyping: Past Decades, Current Challenges, and Future Perspectives. *Molecular Plant*, **13**, 187-214.

110. Bhatta, M., Sandro, P., Smith, M.R., Delaney, O., Voss-Fels, K.P., Gutierrez, L. and Hickey, L.T. (2021) Need for speed: manipulating plant growth to accelerate breeding cycles. *Current Opinion in Plant Biology*, **60**, 101986.

111. Xu, Y., Ma, K., Zhao, Y., Wang, X., Zhou, K., Yu, G., Li, C., Li, P., Yang, Z., Xu, C. *et al.* (2021) Genomic selection: A breakthrough technology in rice breeding. *The Crop Journal*.

112. Zhang, D., Tang, S., Xie, P., Yang, D., Wu, Y., Cheng, S., Du, K., Xin, P., Chu, J., Yu, F. *et al.* (2022) Creation of fragrant sorghum by CRISPR/Cas9. *Journal of Integrative Plant Biology*, **64**, 961-964.

113. Purugganan, M.D. and Jackson, S.A. (2021) Advancing crop genomics from lab to field. *Nature Genetics*, **53**, 595-601.

114. van Dijk, A.D.J., Kootstra, G., Kruijer, W. and de Ridder, D. (2021) Machine learning in plant science and plant breeding. *iScience*, **24**, 101890.

115. Eraslan, G., Avsec, Ž., Gagneur, J. and Theis, F.J. (2019) Deep learning: new computational modelling techniques for genomics. *Nature Reviews Genetics*, **20**, 389-403.

116. Zou, J., Huss, M., Abid, A., Mohammadi, P., Torkamani, A. and Telenti, A. (2019) A primer on deep learning in genomics. *Nature Genetics*, **51**, 12-18.

Summary

The quest for analyzing ever growing volumes of molecular biological data has been a prominent endeavor for bioinformatics for decades. With the advances in sequencing technology, more abundant omics data became accessible. This was accompanied by progress in bioinformatics methodology, to study these data more efficiently. This thesis is dedicated to developing bioinformatics approaches to analyze and manage large-scale genome, proteome and variome data for functional genomics and population genetics studies.

In Chapter 1, I give a brief overview of the history of bioinformatics, followed by a detailed discussion on the analysis of sequences of proteins and genomes, as well as the mining algorithms and managing principles for biological data. Then I discuss the prospects of applying next-generation sequencing (NGS) technology and bioinformatics approaches in crop omics research, combined with the formulation of scientific questions in crop population genomics and breeding.

I start from the mining and comparative genomics analysis of proteome and genome data across multiple species, explore the biological significance of amino acid repeats widely spread in protein sequences across different life kingdoms. In Chapter 2, I introduce the biological context of amino acid repeats and list well-studied repeat- containing proteins (RCP) and their functional roles. Then I review and classify amino acid repeat detection strategies, discuss their algorithmic framework and application context. Since no single algorithm can detect all repeat patterns, I propose that multiple algorithms should be combined to identify different repeat patterns. In Chapter 3, I implement several amino acid repeat detection algorithms, develop integrated data mining and annotation procedures, extract amino acid repeat sequences, annotations, and cross-references from public protein and DNA databases, and construct a specialized database (ProRepeat) to manage and explore protein repeats and do comparative analyses.

In the following chapters, I focus on research in crop plant population genetics and functional genomics promoted by NGS technology. After reviewing the progress of NGS technology and its application to studying sorghum omics and breeding in Chapter 4, I describe the whole-genome resequencing and variation detection across hundreds of sorghum germplasms and the development of a sorghum SNP database (SorGSD) to store the sorghum variome data in Chapter 5. Using the sorghum genomics data, I apply multiple bioinformatics approaches, combined with population genetics and experimental methods, to study the sorghum key gene for a vital breeding trait in Chapter 6. The aim of this analysis is to disclose crucial genes functional and evolutionary roles during sorghum domestication and improvement. In Chapter 7, I first summarize the main results and contributions of this thesis, then discuss the limitations in the thesis that should be improved or are worthy for further study, and finally propose future research directions for bioinformatics in crop research.

Acknowledgements

I would like to express my most sincere gratitude to the following people. I couldn't have gotten here without them.

Professor Jack Leunissen, my initial promotor of the PhD project, who passed away on May 14, 2012. Jack was so vigorous a person. His smiles will always be in my memory.

Professor Dick de Ridder, my promotor. I first met Dick in the winter of 2009 when I took his PhD course "Algorithms for Biological Networks" in Delft. At that moment, both Dick and I could not possibly anticipate that we will discuss my propositions together ten years later. It is destiny.

Dr. Harm Nijveen, my co-promotor, the "programming problem terminator". There is a saying circulating among his colleagues "Nobody needs the manual with Harm at hand."

Professor Hai-Chun Jing, my cooperation instructor in Chinese Academic of Science, who has also been in Wageningen for years. I learned much from him.

My friends and colleges. Linke, Jifeng, Arni, Pieter, Anand, Blaise, Ernest, Judith. The kindest young people with the best brains.

Our secretary, Maria, who is a life saver for paper work, invoice, visa, and everything.

My parents, Jingchu and Honglan, who always support me unconditionally.

My wife, Xin, and my new born daughter, Anlan. You are the source of my happiness.

About the Author

Hong Luo was born on May 27, 1977 in Beijing, China.

He learned engineering and computer science in college and later turned to bioinformatics in the Center of Bioinformatics, Peking University.

In 2004, he started his MSc in Bioinformatics at Wageningen University and became a PhD candidate in the Bioinformatics lab, Wageningen University two years later.

From 2012 to now, he joined the Energy Crop Molecular Breeding Lab, Institute of Botany, Chinese Academy of Science.

From 2017, he was invited as the outside scientific advisor and director of bioinformatics, China Golden Marker company.

List of Publications

Luo, H., Lin, K., David, A., Nijveen, H. and Leunissen, J.A.M. (2011) ProRepeat: an integrated repository for studying amino acid tandem repeats in proteins. *Nucleic Acids Research*, **40**, D394-D399.

Luo, H. and Nijveen, H. (2014) Understanding and identifying amino acid repeats. *Briefings in Bioinformatics*, **15**, 582-591.

Luo, H. Mocoeur ARJ, Jing H-C. (2014) Next-generation sequencing technology for genetics and genomics of sorghum. *Genetics, genomics and breeding of sorghum.* CRC Press, pp. 226-250.

Luo, H., Zhao, W., Wang, Y., Xia, Y., Wu, X., Zhang, L., Tang, B., Zhu, J., Fang, L., Du, Z. *et al.* (2016) SorGSD: a sorghum genome SNP database. *Biotechnology for Biofuels*, **9**, 6.

Zhang, L.M., Leng, C.Y., **Luo, H.**, Wu, X.Y., Liu, Z.Q., Zhang, Y.M., Zhang, H., Xia, Y., Shang, L., Liu, C.M. *et al.* (2018) Sweet sorghum originated through selection of *Dry*, a plant-specific NAC transcription factor gene. *Plant Cell*, **30**, 2286-2307.

Tao, Y., **Luo, H.**, Xu, J., Cruickshank, A., Zhao, X., Teng, F., Hathorn, A., Wu, X., Liu, Y., Shatte, T. *et al.* (2021) Extensive variation within the pan-genome of cultivated and wild sorghum. *Nature Plants*.

Wu, X., Liu, Y., **Luo, H**., Shang, L., Leng, C., Liu, Z., Li, Z., Lu, X., Cai, H., Hao, H. *et al.* (2022) Genomic footprints of sorghum domestication and breeding selection for multiple end uses. *Molecular Plant*.

The research described in this thesis was financially supported by the Netherlands Bioinformatics Centre and the Chinese Academy of Science.

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

Printed by ProefschriftMaken || www.proefschriftmaken.nl