UNIVERSITY OF CALIFORNIA
RIVERSIDE

Methods for Comparative Genome Analysis With Applications to Pan-genomics and Genome Annotation

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Genetics, Genomics and Bioinformatics

by

Qihua Liang

December 2020

Dissertation Committee:

Dr. Stefano Lonardi, Chairperson
Dr. Timothy Close
Dr. Tao Jiang

The Dissertation of Qihua Liang is approved:

Committee Chairperson

University of California, Riverside

Acknowledgments

Life is an adventure. Five years ago, I stepped into a new country on a new continent to start a new adventure as a PhD student at University of California Riverside. I have met a lot of wonderful persons, gained important research and life experiences, and grown into an independent research scientist. Five years later, I am proud to present my PhD dissertation to the world.

I would like to express my deepest appreciation to my advisor Prof. Stefano Lonardi for being supportive, patient, enthusiastic and the best advisor. Every time I introduce myself as a member of Lonardi Lab to other fellow students from the department, everyone congratulates me on having such a great advisor. He has set an example of excellence as an instructor, mentor, researcher, and role model. I hope that I could be as passionate and energetic as Stefano and someday be able to make scientific impacts to the community as well as he can. I would also like to extend my sincere thanks to my committee, Prof. Tao Jiang and Prof. Timothy Close, who provide scientific advice and insightful discussions about the research projects and thesis. I gratefully acknowledge the assistance from all collaborators whom I have worked with over the past several years for showing me what it means to be a dedicated scientist.

I have spent five long yet precious and wonderful years at UCR. I would like to recognize all the help and support I received from my genius lab-mates as well as close friends, Abid, Weihua, Abbas, Dipankar, Hind, Rachid and Saleh. They have provided emotional support during all the moments with self-doubt and guided me through the research difficulties along the way. Longtime friends from neighbor research groups, Huong

and Tin, extend a great amount of assistance in helping me surviving grad school. I also want to thank all my friends (too many to list here but you know who you are!) for their relentless support and sincere friendship.

I especially thank my family for all their support during my study. My hard-working parents provide unconditional support and care whenever I need them. Special thanks to the new additions to my family, my husband Jin and my son Aiden. The best outcome from these past several years is finding my best friend, soul-mate, and husband, Jin. I married the best person out there for me. I also dedicate this thesis to my lovely child, Aiden, who has made me stronger, better and more fulfilled than I could have ever imagined. I love all my family to the moon and back.

These past years have not been an easy journey, both academically and personally. There were times when I did not have faith in myself, when I was irritable and depressed, when I was up all night worried about future. The nights were long, but the years were short. This dissertation would not have been possible without the inspiration and support of all the wonderful persons I am lucky enough to meet. Thank you again to all of you.

To my family.

ABSTRACT OF THE DISSERTATION

Methods for Comparative Genome Analysis With Applications to Pan-genomics and Genome Annotation

by

Qihua Liang

Doctor of Philosophy, Graduate Program in Genetics, Genomics and Bioinformatics
University of California, Riverside, December 2020
Dr. Stefano Lonardi, Chairperson

Comparative genomics is a powerful analytical tool for understanding the structure of genomes and their evolution. The tenet of comparative genomics is that evolutionarily conserved (thus functionally important) genomic features between two species share significant similarity at the DNA or protein level. Recent technological advancement in DNA sequencing instruments enabled the number of sequenced genomes for different species to increase exponentially. The expanded set of available genomes has provided new opportunities to carry out comparative genome analyses at unprecedented scale.

In this dissertation, we discuss and investigate a set of comparative genomics methods relevant to genome assembly, genome annotation and pan-genome analysis. Comparative genomics can assist *de novo* genome assembly during the scaffolding phase and in the evaluation of assembly quality. The annotation phase takes advantage of comparative genomics by leveraging annotations from related species to predict coding and non-coding gene boundaries, intron/exon boundaries, repetitive elements, and many other genomic features. Functional annotation also relies on comparative genomics to assign putative functions to

annotated genes using known functions of evolutionarily-conserved genes and proteins. Finally, intraspecies comparative genomics is the cornerstone of pan-genome analyses that allows one to determine which portions of the genome are common to all individuals, and which portions are variable among the individual of a species. A new pan-genome representation and visualization method is introduced here to elucidate complex structural genomic variations.

Experimental results on the genomes of (1) *Vigna unguiculata* (cowpea or black-eyed pea) which provides a valuable source of protein to millions of people in developing countries, (2) *Phytophthora infestans* which is an oomycete which causes a potato and tomato disease known as late blight, and (3) *Babesia duncani* which is tick-transmitted protozoan parasites that causes severe infection in immunocompetent individuals, demonstrate the effectiveness and utility of these comparative genomics methods.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Life science research was revolutionized by the invention of DNA sequencing in the 1970s, which led to a new era of genomic scientific investigation. The first complete bacterial genome, *Haemophilus influenzae*, was fully sequenced in 1995. The 1.83Mb genome sequence revealed 1742 protein-coding genes along with a small complement of transfer RNAs (tRNAs) and ribosomal RNAs [23]. Since then, advancement in DNA sequencing technology has allowed Life scientists to obtain the primary DNA genomic sequence of tens of thousands of bacteria and viruses, thousands of individual humans, and thousands of other eukaryotic species [98].

1.1 Genome Sequencing and Assembly

Deoxyribonucleic acid, or DNA, is the primary hereditary material in all living organisms. DNA is a double helix in which *nucleotides* (or bases) pair up with each other, A with T and C with G, to form units called *base pairs*. The genome is the complete set of

DNA in an organism and it contains all the genetic information needed to "build" an entire individual and maintain all necessary metabolic activities [1].

DNA sequencing instruments are used to obtain the primary DNA sequence of an organism, but they are unable to read chromosomes from their beginning to their end. The most popular type of sequencing strategy in use today, called *second generation* sequencing or next generation sequencing (NGS), is based on sequencing by synthesis (Illumina). DNA polymerase, which is the enzyme in cells that synthesizes DNA, is used to generate a new strand of DNA based on a target strand to be sequenced. During the sequencing reaction, DNA polymerase utilizes fluorescently labeled nucleotides into synthesizing a new complementary strand of the target DNA strand. Four nucleotides are used separately in order to react with DNA polymerase. When the paired nucleotide is present to incorporate with target sequence, the fluorescent signal is emitted during such incorporation and detected by a camera. The signal is different for four nucleotides and thus the current nucleotide is determined based on the detected signal. NGS can generate massive amounts of short reads few hundreds nucleotides long, and for this reason it is called *high-throughput sequencing*.

The third generation of sequencing technologies (i.e., Pacific Biosciences and Oxford Nanopore) can generate reads with size up to 100,000bp. Such read length is a significant improvement from second generation sequencing but the throughput is much lower. The first commercially available long read sequencing platform was introduced by Pacific Biosciences' (PacBio), called single molecule real-time (SMRT) sequencing technology [5]. SMRT sequencing also takes advantage of sequencing by synthesis and utilizes fluorescently labeled nucleotides as NGS. SMRT employs a zero-mode waveguide where a DNA poly-

merase enzyme is attached at the bottom of the flow-cell. Oxford Nanopore is the most recent third generation sequencing technology on the market. A nanopore is a nano-scale hole through which an ionic current is passed through. As the DNA passes through the pore, different nucleotides along the negatively charged DNA strand cause different electric current changes. The sensor detects ionic current fluctuations and determines the nucleotide passing through the pole. No DNA is synthesized during Nanopore sequencing process, which is a significant departure from NGS or PacBio technonologies.

Due to the limitations of sequencing instruments, a strategy called *whole genome sequencing* (WGS) has been developed to obtain the primary sequence of large eukaryotic genomes. In WGS, the genome is broken down into a collection of smaller DNA fragments by a random process called *shotgun*, and then each fragment is read by the sequencing instruments to get the order of nucleotides (i.e., each fragment generates a *read*). After the sequencing of the DNA fragments, the whole genome needs be assembled into the most contiguous and complete sequences based on the overlap between sequencing reads. The process of combining fragmented reads into longer fragments of the genome is called *assembly*. Currently, assembly algorithms can be divided in two general classes: overlap–layout–consensus (OLC) and de-Bruijn graph (DBG). OLC was introduced by Staden [103] and later adopted by several long-reads based assembly algorithm such as the Celera Assembler and more recently Canu [42], Falcon, etc. The OLC approach consists of three steps: (i) the overlaps phase (O), in which all overlaps among all the reads are detected; (ii) the layout phase (L), in which all the read overlaps are represented on a graph; (iii) the consensus phase (C), in which the assembled sequence is inferred [54]. The DBG approach first breaks all the reads

into shorter k-mers (which is a string of length k). Then, a de-Bruijn graph is constructed and used to infer the genome sequence. Many short-reads based assembly tools are based on DBG, such as ABySS [99] and SOAPdenovo [52]. DBG assemblers provides an efficient and effective genome assembly ideal for NGS short reads.

Even though genome assembly has considerably improved due to improvement in sequencing technologies and algorithmic innovations, assemblies produced entirely on sequencing reads are typically fragmented and they often cannot span entire chromosomes. To further improve quality of assemblies, several long-range technologies such as optical mapping and genetic maps have been developed [91]. An optical map is a sequence of lengths of fragmented DNA sequence resulting from restriction enzymes cutting at restriction sites. A genetic map is a type of species-specific chromosome map that shows the relative genetic distance of genomic marker (e.g., SNPs) [121]. Optical and genetic maps together with new assembly methods have enabled the generation of complete and high-quality genome assemblies for large eukaryotic genomes.

The cowpea genomes in Chapter 2 and Chapter 3 were sequenced with different technologies of second and third generation sequencing. Such sequencing reads together with genetic maps and optical maps are used in assembling the genomes of different cowpea accessions.

1.2 Genome Annotation

While the research community in genomics is mostly focused on genome sequencing assembly, genome annotation is starting to attract more and more attention [105]. Genome

annotation can be classified into two major types: (1) structural annotation, that is the identification of all elements in a genome, such as repeats, genes/pseudo-genes, 5'/3' UTRs, introns/exons; (2) functional annotation, that is the prediction of the biological functions of the elements detected in (1). Genome annotation is necessary because genome sequencing and assembly only produces the primary DNA sequence of the genome devoid of any interpretable information [2].

Structural annotation primarily aims to find repeats and genes within the genome. A significant portion of eukaryotic genomes is characterized as repetitive or interspersed repeat regions. In these genomes, it is challenging to identify repetitive regions prior to gene finding because of the complex multi-scale structure of repeats and high proportion of repetitive elements present. In many instances of genome annotation, repeat masking is the first annotation step, prior to gene finding. For instance, over 20% of the *Arabidopsis thaliana* genome is composed of repetitive elements [66]. These percentages are much higher in legumes (which are one of the focuses of this dissertation): it is 53.9% in soybean [122], 45.2% in common bean, 50.1% in mung bean [40], and 44.5% in adzuki bean [123]. As said, masking repetitive elements is carried out before the identification of other structures along the genome. Unfortunately, curated repeat libraries are available only for a limited number of species. Using repeat libraries for distantly-related species can lead to missing repeats thus negatively affecting gene prediction accuracy. In order to address this challenge, species-specific repeat libraries species can be built based on the sequence information of each genome. In small prokaryotic genomes, gene finding is mostly about identifying long open reading frames (ORFs). For instance, in the *Haemophilus influenzae* genome, 85%

of the primary DNA sequence consists of coding regions. The fraction of the genome that encodes for proteins is much lower in eukaryotic genomes, e.g., it is 70% in yeast, less than 25% in fruit fly and worm, and only about 1% in the human genome. For eukaryotic genomes, structures like 5'/3' UTRs and introns/exons are also critical and they are usually identified during gene finding.

Gene prediction methods can be classified broadly into two classes, namely *de novo* and homologous-based. *De novo* gene predictors use probabilistic models trained on known genes to detect putative genes. For example, predictors that use Hidden Markov models (HMM) may explicitly calculate how individual probabilities of a sequence of features are combined into probability estimate for the whole gene [105]. Several well-established *de novo* predictors have been used in gene finding for many species, including Augustus [104], GeneMark [59] and SnapHMM [43]. *De novo* predictors increase the possibility to annotate genes that are rarely expressed and would be missed by homologous-based methods (e.g., when no transcript evidence is available).

Homologous-based or evidence-based methods annotate genes from transcript evidences (e.g., ESTs, expressed proteins, RNA-Seq) and/or protein evidences either for the species under consideration or from related species. Genes annotated or supported by homologous comparison are expected to be more reliable than *de novo* prediction. Nevertheless, deriving a complete gene model from sequence similarity is not as straightforward as extracting sequence alignment results. Most genes in eukaryotic genomes are spliced into multiple exons, which significantly complicates the sequence alignment analysis. Besides, the homologous evidences may not be of high quality. For example, cDNA sequences may

contain repetitive sequences that may cause incorrect genomic sequence match. Protein sequences from related species can introduce unnecessary evolutionary divergence problems. To address such issues one can combine multiple homologous resources from different species as well as *de novo* predicted gene models. Annotation pipelines such as Maker2 [13], Funannotate [79] and EVM [31] combine both *de novo* and homologous approaches using adjustable weights applied to different predictors or evidence.

The final step after genome assembly and gene/repeat annotation is to evaluate the overall quality of the assembly and the annotation. As said, the objective is the obtain the most contiguous and error-free assembly, and the most complete catalogue of all the genes. A popular evaluation pipeline is BUSCO, which was designed to assess the completeness of genomes, gene sets, and transcriptomes [92]. BUSCO compares a given assembly and annotation to a set of universal single-copy genes that are expected to be present in any genome in that category of organisms.

After collecting the complete set of genes, the functional annotation steps aims at providing the most accurate biological functions to the annotated genes. For instance, *Haemophilus influenzae* has only 1,709 genes, yeast has about 5,600 genes and human has more than 30,000 genes [105]. A large fraction of the genes for many organisms still do not have well a characterized biological function. The function of a protein encoded by a gene is tied to its 3D structure (folding), which is hard to obtain experimentally. Functional annotation is typically carried out by seeking proteins that have similar sequence, and have a known functions. For instance, OrthoMCL [51] is designed to identify ortholog groups for eukaryotic genomes based on protein-level sequence similarity. To distinguish functional

redundancy from divergence, OrthoMCL detects recent paralogs to be considered in ortholog groups as within-species reciprocally hits. Protein databases SWISS-PROT TrEMBL [7], PFAM [9], PRINTS [6], PROSITE [34] and InterPro [4] are commonly used to identify gene functions based on sequence similarity.

Another important step in functional annotation is associate genes to biological processes such as cell cycle, metabolism, and cell component localization. One of the standard classification scheme is called Gene Ontology (GO) [76]. GO contains standardized terms to describe eukaryotic gene functions along three "components", namely molecular function, biological process and cellular component. GO organizes the functional terms hierarchically in a directed acyclic graph, making it easy to determine the dependencies among terms.

In Chapter 2 and Chapter 3 the genomes of cowpea, *Phytophthora infestans* and *Babesia duncani* were annotated with a combination of above mentioned *de novo* and homologous-based methods. Genome annotations of such species provide complete and accurate structure and function of genome features in order to advance in further genomic study.

1.3 Pan-genome

As sequencing costs continue to decrease, there is an exponentially increasing number of available genome for different species. These genome sequencing project are aimed to produce a *reference* genome for a species, i.e., they are based on single individual for that species. Such reference genomes serve as the basis of many genetic analyses includ-

ing genome comparisons and variation studies within and across species. A high-quality and well-annotated reference genome facilitates comparative genomic studies with its related species. For example, our cowpea reference genome reported in Chapter 2 allows the comparative study with other legumes such as soybean and common bean, revealing both similarities and differences of the genome characteristics and can help researchers to better understand traits and gene functions.

However, recent studies have shown that genomic sequences of one individual cannot fully representing the full range of genetic diversity of a species (see, e.g., [125, 56, 94, 10]). For example, recent sequencing of 910 humans of African descent reports up to 10% of the total genome size missing from the reference genome [93]. As said, the cost of whole genome sequencing has been decreasing rapidly and this has enabled sequencing for multiple individuals within the same species in order to capture the genetic diversity of a species [69, 65, 33, 110].

The term *pan-genome* was first used by Sigaux [96] to describe a public database containing an assessment of genome and transcriptome alterations in the major types of tumors as well as in relevant normal cells and experimental models. Later, Tettelin et al. [112] defined a microbial pan-genome with eight different strains of *Streptococcus agalactiae*, a pathogenic species isolated from human, as the combination of a *core genome*, composed of genes shared by all strains, and a *dispensable genome* (also known as *flexible* or *accessory genome*) consisting of partially shared and strain-specific. (Figures 1.1) A generalization of such a representation could contain not only the genes, but also other variations present in the collection of genomes. Here we adopt the defini-

tion proposed in [28] where a pan-genome consists of three components, namely, i) the core genome, formed by genes that are shared by all individual genomes (these genes are usually involved in essential cellular processes, like house-keeping genes); ii) accessory or dispensable genome, composed of genes that absent in some individuals; and iii) individual-specific genes, which are those genes that are present in a single individual genome. The main goal of pan genome analysis is to identify these three genomic components by comparing the genomes of different individuals of the same species. We should note that the number and the quality of the genome assemblies directly affects these notions. For instance, genes belong to very repetitive regions of the genome (i.e., centromers) are likely to be absent from the assembly because assembly tools struggle with repetitive regions. As a consequence, a gene could be declared dispensable or individual-specific only because it is missing from most the assemblies, not because it is truly dispensable.

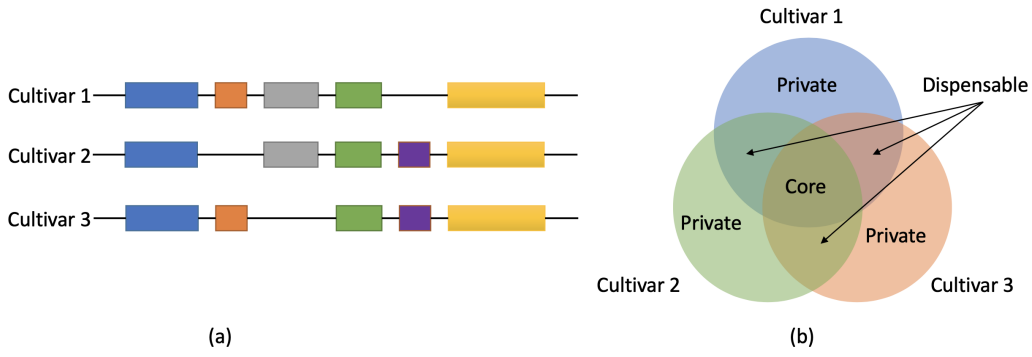

Figure 1.1: Pan-genome composition

In plants the notion of a pan-genome was first used to describe short variable regions of transposable elements in the rice and maize genomes [72]. It took almost ten years for plant pan-genomes to be available after the initial bacterial pan-genome study [10]. The first published plant pan-genome was based on whole genome assemblies of seven wild soybean accessions [53]. Genes present in wild *Glycine soja* but not identified in domesticated *Glycine max* were related to seed composition, flowering and maturity time, organ size and biomass. Such present-absent variation analyses also showed an increase of copy number of disease-resistance genes in wild *Glycine soja*. Later, pan-genomes have been constructed for other crop species, including rice [127], tomato [25], soybean [53, 56], among others. The motivation for building a pan-genome for a crop species is to understand the genetic diversity within different accessions and eventually identify key variations which are linked to agricultural production phenotypes in order to improve breeding.

Although crops have been selectively bred since their domestication, the genes underlying the selected phenotypes often remain unknown and are sometimes linked to genes with undesirable phenotypes. For example, a cultivar that produces larger fruit might be lacking in disease-resistance genes. Discovering these phenotype-causing genes can help both to breed and to genetically modify plants so as to create crops that are more disease-resistant, are more productive, have a longer shelf life or taste better, without sacrificing other desired phenotypes. Pan-genomic approaches in plants have already uncovered numerous associations between agronomic phenotypes and the presence or absence of specific genes. Inspired by previous plant pan-genome studies, we embarked on a project to build the pan-genome of cowpea (*Vigna unguiculata [L.] Walp.*), also known as black eye pea.

Cowpea is a major crop for worldwide food and nutritional security, providing a valuable source of protein to millions of people in developing countries. The cowpea pan genome incorporates multiple cowpea accessions from around the world, and will provides new biological insights into the genetic diversity of this important legume. Chapter 3 of this thesis reports on our the analysis of the cowpea pan-genome.

In response to the availability of the pan-genomes for many species, new analysis tools have been developed to answer some of the main questions, namely (1) how to find core/dispensable/unique portions of the pan-genome, (2) how to visualize the pan-genome in an effective/intuitive way. So far, no analytical pipeline has emerged that can satisfactorily address all these issues. The majority of the available pan-genome analysis tools either: (i) focus only on the genes, or (ii) they can only handle small genomes (e.g., bacterial genomes) and are unable to scale to larger eukaryotic genomes, or (iii) they require users to arbitrarily label one of the individual genomes as *the* reference. Most pan-genome studies to date have focused on the genic portion of the genome. However, study in maize genome has revealed that genomic regions in open chromatin also play a significant role in molecular phenotypes such as gene expression and recombination [86]. This suggests that many important agronomic traits may be determined by variations in intergenic regions through gene regulation rather than present-absent variation (PAV) at the gene level. Combined with further functional annotations of regulators, pan-genomes provide a rich resource for regulatory sequence variations that can be harnessed in breeding. Pan-genome studies focusing on genes may ignore variations in intergenic regions and thus may underestimate agronomic traits related to such intergenic variations.

For intergenic sequences, defining what is core versus dispensable/private becomes more important and challenging, especially in species in which a large fraction of genome is composed of divergent repeats. For instance, PanX first identifies orthologous gene clusters from a set of individual genomes, then allows users to interactively explore the relationships between genes via a web-based visualization tool [21]. Similarly, PanWeb [80] is a web-based front-end for PGAP (Pan-Genome Analysis Pipeline) [129]. PGAP provides several types of gene-level analysis, including gene cluster analysis, pan-genome profile analysis, variation analysis, evolution analysis and function enrichment analysis. PPanGGOLiN models a microbial pan-genome using a graph in which nodes represent gene families and edges represent genomic neighborhood [26]. The Genome Context Viewer is a genome browser that can identify and visualize micro-synteny regions, i.e., co-linear arrangement of homologous genes, in a pan-genome [18]. Other tools provide genome-wide insights by comparing the whole genomes. PanSeq identifies core, accessory and novel regions of genome-level by carrying out a pairwise alignment against one of the individual genome which need to be considered *the* reference [46]. PGAP-X is an extension of PGAP which uses whole genome sequence alignment to distinguish core, dispensable and strain-specific genes [128]. The Genome Context Viewer allows the exploration of precomputed macro-synteny blocks in pan-genomes [18].

In order to address the limitations of current tools mentioned above, in Chapter 4 we introduce a novel pan-genome representation and visualization method called PGV. The PGV representation: (i) is reference-agnostic (i.e., there is no need to artificially declare one of the individual genomes to be the reference), (ii) can handle large eukaryotic genomes

including human genomes, and (iii) is very intuitive and simple for users to use and understand.

1.4 Acknowledgments

The material in this thesis is the product of several collaborations. Section 2.1 and 2.2 is the product of a collaboration with the co-authors of [60]. Section 2.3 is unpublished, and it is the product of a collaboration with Prof. Howard Judelson's lab (UC Riverside). Section 2.4 is unpublished, and it is the product of a collaboration with Prof. Karine Le Roch's lab (UC Riverside) and Prof. Ben Mamoun's lab (Yale). The work in Chapter 3 is the product of a collaboration with Prof. Close's lab (UC Riverside), Prof. Maria Munoz Amatriain (Colorado State), Dr. Sassoum Lo (UC Davis), and its content is currently unpublished. Chapter 4 is also currently unpublished.

Chapter 2

Assembly and annotation of *V. unguiculata*, *P. infestans* and *B. duncani*

In this chapter we report on the *de novo* assembly and annotation of *Vigna unguiculata*, *Phytophthora infestans* and *Babesia duncani*. All genomes were sequenced using long reads (Pacific Biosciences for cowpea and phytophthora, Oxford Nanopore for babesia). For an introduction to sequencing technologies and assembly, please refer to Chapter 1.

2.1 Assembly and annotation of *Vigna unguiculata*

Cowpea (*Vigna unguiculata [L.] Walp.*), also known as black eye pea, is a major crop for worldwide food and nutritional security, providing a valuable source of protein to millions of people in developing countries. One of the strengths of cowpea is its resilience

to harsh conditions, including hot and dry environments, and poor soils [12]. Cowpea is a diploid ($2n = 22$) member of the family *Fabaceae* tribe *Phaseoleae*, closely related to mung bean, common bean, soybean and several other protein-rich warm-season legumes. In spite of its importance in food security, modest progress has been made in the generation of high quality genomic sequences for more effective breeding and development of improved varieties.

2.1.1 Genome Assembly

A highly fragmented draft assembly and BAC sequence assemblies of the cowpea variety called IT97K-499-35 were previously published [74]. Although this draft assembly has enabled significant progress on cowpea genomic studies, e.g., [126, 15, 71, 57, 35], it lacked the contiguity and completeness required for accurate genome annotation, detailed investigation of candidate genes or whole-genome comparative analyses.

Here two different sequencing strategies were used in order to obtain a high quality genome for accession IT97K-499-35, namely Pacific Biosciences (PacBio) long reads and 10x Genomics linked reads. About 56.8 Gb of PacBio data were generated with ∼91.7x coverage and reads N50 of 14,595 bp. 331.42M 10x linked reads were generated with ∼77.1x coverage. Three different assemblers including CANU [42], Falcon [17] and ABruijn [55] were used to assemble the PacBio reads into draft contigs. With the help of two optical maps and ten genetic maps a final assembly at the pseudo-molecule level was generated. 10x reads were assembled using superNova [119]. A summary of the basic statistics for these two assemblies is shown in Table 2.1. Observe that both assemblies had a comparable assembly size over 500Mb. The PacBio assembly was more contiguous with a N50 of over 41Mb and had a

smaller number of scaffolds/contigs, and thus this assembly was used in further downstream analyses.

Even though PacBio sequencing technology provided a higher quality assembly, it was not economically feasible for the planned pan-genome project that required the sequencing of several additional accessions. In order to evaluate a cheaper alternative in pan-genome sequencing, cowpea accession CB5-2 was sequenced and assembled by Dovetail Genomics from Illumina short reads (150x2). Dovetail Genomics used Meraculous [16] to assemble the reads, then Chicago and Hi-C libraries (using their proprietary chromatin proximity-ligation technology) to resolve possible mis-assemblies and to increase contiguity. The final assemblies were processed using ALLMAPS [108] using ten high-density genetic linkage maps previously generated [74, 60] to further increase contiguity. The CB5-2 Dovetail assembly was compared to the previous two assemblies of IT97K-499-35 in terms of assembly contiguity and completeness (Table 2.1 and Figure 2.1). Genome completeness was calculated using BUSCO v3 [97] with embryophyta-odb9 dataset in terms of genome, transcripts and proteins. Observe that the assembly contiguity and genome level BUSCO completeness of CB5-2 were similar to that of IT97K-499-35 using PacBio reads. The CB5-2

Table 2.1: Assembly Statistics

	IT97K-499-35(PacBio)	IT97K-499-35(10x)	CB5-2(Dovetail)
Total (bp)	519,435,864	506,741,239	448,043,751
# contigs	686	27,997	6,534
# contigs $\geq$ 100 kb	177	653	28
# contigs $\geq$ 1 Mb	61	110	11
# contigs $\geq$ 10 Mb	11	0	11
N50(bp)	41,684,185	765,309	36,897,245
L50	6	173	6
GC (%)	33.0	32.4	32.5

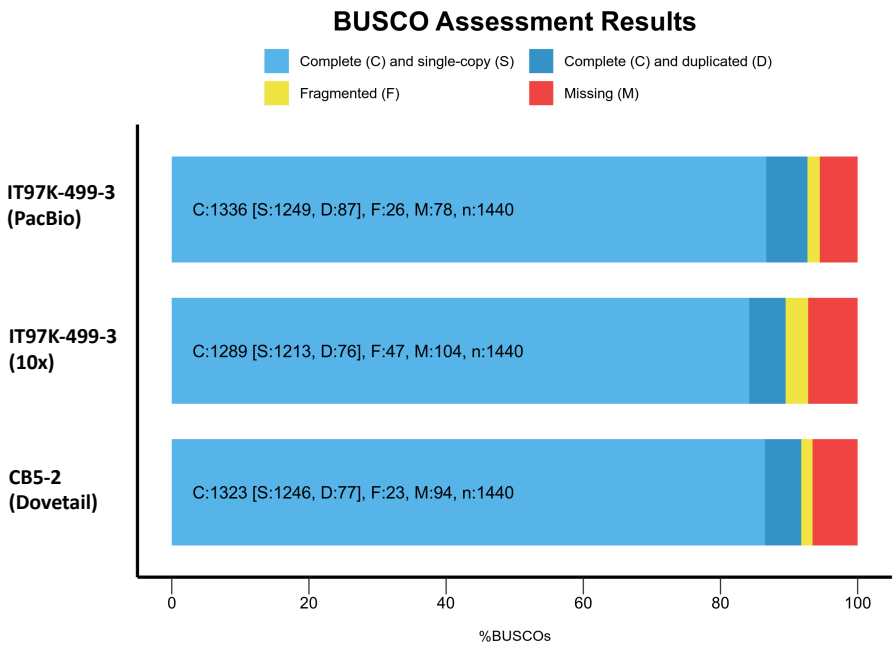

Figure 2.1: BUSCO completeness analysis for the three assemblies in Table 2.1

assembly was composed of eleven pseudo-molecules, and its high BUSCO gene completeness indicated high quality of the assembly.

2.1.2 Chromosome Numbering

The numbering of chromosomes of *Phaseoleae* tribe has been assigned by different research groups independently within and across species. The common bean (*vulgaris*) genome sequence appeared to be a reasonable model for a standardized chromosome numbering system for *Phaseoleae* tribe [90]. A synteny view between cowpea (Vu) and common bean (Pv) chromosomes was made using the previous cowpea chromosome numbering adopted in a previously published draft assembly [73, 74]. Extensive synteny was identified between cowpea and common bean, as shown in Figure 2.2(A), which provided a fundamen-

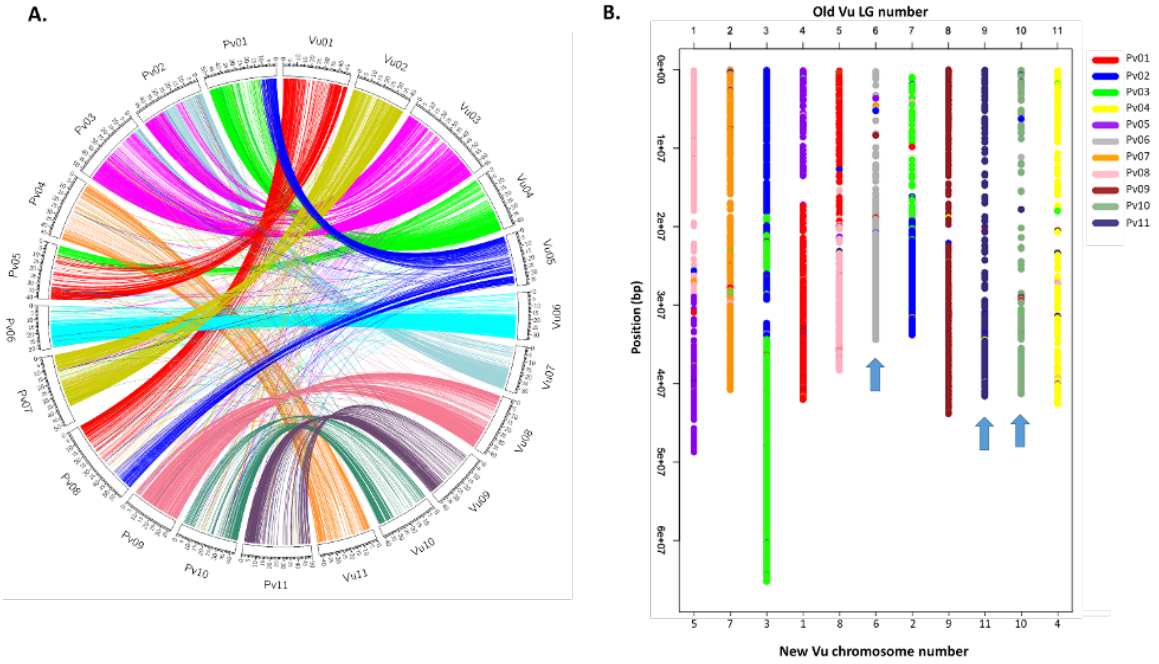

Figure 2.2: Synteny view between cowpea and common bean using the previous chromosome nomenclature. (A) Circos illustration of synteny. (B) Cowpea chromosomes painted based on syntenic relationships with common bean chromosomes (in different colors)

tal basis for a revised chromosome numbering system for cowpea. The total length of the syntenic matches (exact match of least 100bp, alignment length of least 1kb) with the top two *P. vulgaris* (Pv) chromosomes was shown. Chromosomes that were inverted to meet the "short arm on top" convention were indicated in parenthesis. The asterisk indicated the optimal solution to the assignment problem. Chromosomes indicated with arrows were inverted to meet the convention "short arm on top" based on previous BAC-FISH analysis [38].

As summarized in Figure 2.2 and Table 2.2, six cowpea chromosomes, including chromosome 2, 6, 8, 9, 10, 11, were highly syntenic with six common bean chromosomes in one-to-one correspondence relationships. Chromosome numbering conversion of these

Table 2.2: Cross-reference between old and new chromosome numbers for cowpea (Vu)

Old Vu Chr.	Pv Chr (kb)	Pv Chr (kb)	New Vu Chr.
1	8 (671.2)	5 (485.1)	5*
2	7 (1390.0)		7
3	3 (1493.9)	2 (899.8)	3
4	1 (932.0)	5 (245.0)	1
5	8 (573.3)	1 (309.0)	8*
6	6 (996.8)		6 (inverted)
7	2 (736.9)	3 (163.8)	2
8	9 (1439.5)		9
9	11 (751.8)		11 (inverted)
10	10 (593.9)		10 (inverted)
11	4 (564.1)		4

six chromosomes was straightforward. Each of the remaining five cowpea chromosomes was related to parts of two *P. vulgaris* chromosomes. Generally, numbering for most of these chromosomes were converted based on the common bean chromosome sharing largest syntenic region with cowpea. For example, old cowpea chromosome 3 had extensive synteny with both common bean chromosome 2 and 3, specifically Pv chromosome 3 shared 1493kb with Vu chromosome 3; Pv chromosome 2 shared 899kb. Therefore, Pv chromosome 3 was used to assign for cowpea old chromosome 3. However two cowpea chromosomes (old chromosome 1 and chromosome 5) both shared their largest block of synteny with *P. vulgaris* chromosome 8. The optimum solution was to assign Vu08 to previous cowpea chromosome 5 and assign Vu05 to previous chromosome 1.

2.1.3 Comparisons with other warm-season legumes

Synteny analyses were performed between cowpea and its close relatives including adzuki bean (*Vigna angularis*) [123], mung bean (*Vigna radiata*) [40] and common bean (*P.*

vulgaris) [89]. Extensive synteny was observed between cowpea and the other three diploid warm-season legumes although, as expected, a higher conservation was observed with the two *Vigna* species (Figure 2.3a–c) than with common bean. Six cowpea chromosomes (Vu04, Vu06, Vu07, Vu09, Vu10 and Vu11) largely had synteny with single chromosomes in all three other species. Cowpea chromosomes Vu02, Vu03 and Vu08 also had one-to-one relationships with the other two *Vigna* species but one-to-two relationships with *P. vulgaris*, suggesting that these chromosome rearrangements were characteristic of the divergence of *Vigna* from *Phaseolus*. The remaining cowpea chromosomes Vu01 and Vu05 had variable syntenic relationships, each with two chromosomes in each of the other three species, suggesting that these chromosome rearrangements were more characteristics of speciation within the *Vigna* genus. It should be noted also that most chromosomes that have a one-to-two relationship across these species or genera were consistent with translocations involving the centromeric regions (Figure 2.3a–c). On the basis of these synteny relationships, adoption of the revised cowpea chromosome numbering for adzuki bean, mung bean and presumably other Vigna species is straightforward. This will facilitate reciprocal exchange of genomic information on target traits from one Vigna species to another.

Figure 2.3(c) showed major structural variations including inversions and translocation between cowpea and common bean on chromosome 2 and 3, which indicated possible chromosome rearrangements between *Vigna* and *Phaseolus*. In order to further investigate such inversions and translocation among *Vigna* and *P. vulgaris*, more synteny analyses were done on the two corresponding chromosomes for *P. vulgaris-V. unguiculata*, *P. vulgaris-V. angularis* and *V. unguiculata-V. angularis*. (Figure 2.4)

Sequence synteny data identified additional complex micro rearrangements, including several inversions and translocations. Approximate regions of long arm Vu2 and short arm Vu3 mapped to short and long arms of Pv2 (green lines) respectively, while regions of short arm Vu2 and long arm Vu3 mapped at short and long arms of Pv3 (red lines) respectively (Figure 2.4a). About 47.3% of the chromosomal total length of Vu2 corresponded to Pv2. Similar but not equal rearrangement patterns were observed from Va10 to Pv2/Pv3 (Figure 2.4b). Over 70% of Va10 had synteny to Pv2, which was much higher than Vu2 to Pv2. Despite the gap at pericentromeric region, the Va2 centromere seemed to be mapped to Pv2 sequences (green lines), while the Vu2 centromere seemed to be mapped to Pv3 sequences (red lines).

Va1 and Vu3 had similar rearrangement patterns: two inversions and one translocation. Short arm regions of Va1 and Vu3 had an inversion to the long arm Pv2 individually (green lines). The long arm of Va1 and Vu3 had another inversion compared to the long arm Pv3 (red lines).

Sequence synteny from Vu2/Vu3 to Va10/Va1 showed that these two chromosomes were macrosyntenic between Vigna species, with micro translocations found, especially between Va1 and Vu3 (Figure 2.4c). It was interesting to find that *V. angularis* and *V. unguiculata* share the translocation and inversions when aligned with *P. vulgaris*, with similar patterns. Synteny links were colored by blocks as in Pv2/Pv3, with red links representing Pv2 blocks and green links representing Pv3 blocks.

Table 2.3: Predicted centromeric positions in cowpea

Chromosome	Start (bp)	End (bp)	Range (bp)
1	14,698,036	16,525,496	1,827,460
2	10,238,236	14,020,258	3,782,022
3	30,476,981	31,470,261	993,280
4	19,069,641	21,130,843	2,061,202
5	25,704,431	33,885,354	8,180,923
6	9,156,830	9,235,637	78,807
7	16,587,031	16,604,960	17,929
8	14,914,119	15,164,402	250,283
9	20,802,610	22,685,597	1,882,987
10	18,917,563	19,028,450	110,887
11	17,283,961	18,283,861	999,900

2.1.4 Genome Annotations

Centromeric Region Identifications. Centromeric positions were predicted based on sequence similarity to a previously identified 455bp tandem repeat [37]. Start and end positions of centromeric regions were calculated based on first and last clustering BLAST alignment of this tandem repeat on each chromosome. Centromeric position predictions are summarized in Table 2.3. Centromeric regions on chromosome 6 and 7 were significantly smaller than other chromosomes, which may indicate an incomplete assembly around such regions.

Gene Annotations. Transcript assemblies were produced from ≈1.5B 2x100 paired-end Illumina RNA-seq reads from leaf, stem, root, flower, pod and seed tissue [126, 88] using Cufflinks [116]. The set of 120,745 assembled transcripts produced by Cufflinks, 29,728 EST sequences from [73] and transcripts from common bean and Arabidopsis were used in the gene annotation pipeline as available transcript evidences. Gene models were predicted

by MAKER [13]. Genes were *de novo* predicted by Augustus [104], GeneMark [59] and SnapHMM [43] on the repeat-masked cowpea genome using RepeatMasker [77]. GeneMark was trained using CEGMA [81]; Augustus and SnapHMM were trained by MAKER [14] using transcript evidences listed above. Trained gene predictors together with above assembled transcripts were used in MAKER to identify the final gene models. In total, 22,683 protein-coding genes were annotated, with a BUSCO v3 [97] completeness of 85.9% against the plant data set embryophyta odb9.

An alternative annotation pipeline was also used for the cowpea genome to provide a comparative viewpoint. In this alternative pipeline the set of ESTs and RNA-Seq data listed above was used together with protein sequences of arabidopsis, common bean, soybean, medicago, poplar, rice and grape. In total, 29,773 protein-coding loci were annotated, along with 12,514 alternatively spliced transcripts. Most (95.9%) of the 1440 expected plant genes in BUSCO v3 were identified in the cowpea gene set, indicating high completeness of genome assembly and annotation. The average gene length was 3881 bp, the average exon length was 313 bp, and there were 6.29 exons per gene on average. The GC content in coding exons was higher than in introns plus UTRs (40.82% versus 24.2%, respectively). Intergenic regions had an average GC content of 31.84%. Due to its higher completeness, this second annotated gene set was used in subsequent gene analysis.

Recombination rate was calculated as a polynomial curve fit of cM position for each of the eleven linkage groups from ten genetic maps obtained from biparental RIL populations. Together with the above annotated repeat and gene density, Figure 2.6 shows a comparisons between gene density (green line), repeat density (blue line) and recombi-

nation rate (red line) across the 11 cowpea chromosomes. Gene and repeat density were measured in 1 Mb non-overlapping windows, while recombination rate was measured in non-overlapping windows of 100kb. Vertical lines delimit the predicted centromeric regions. Observe that cowpea centromeric and pericentromeric regions were highly repetitive in sequence composition, and exhibited low gene density and low recombination rates, while both gene density and recombination rate increased as the physical position became more distal from the centromeres. Contrasting examples include Vu04, where the recombination rate near the telomeres of both arms of this metacentric chromosome were roughly ten times the rate across the pericentromeric region, versus Vu02 and Vu06, where the entire short arm in each of these acrocentric chromosomes had a low recombination rate (Figure 2.6). These patterns were also observed in other plant genomes including legumes [89, 90], and have important implications for genetic studies and plant breeding. For example, a major gene for a trait that lies within a low recombination region can be expected to have high linkage drag when introgressed into a different background. Knowledge of the recombination rate can be integrated into decisions on marker density and provide weight factors in genomic selection models to favor rare recombination events within low recombination regions.

Gene Family Clustering. Cowpea gene families were analyzed using two approaches with different related species, namely (i) with four other closely related species including common bean (*P. vulgaris*), adzuki bean (*V. angularis*), mung bean (*V. radiata*) and soybean (*G. max*); (ii) with four other further related species including soybean (*G. max*), arabidopsis (*A. thaliana*, TAIR10) [47], rice (*O. sativa*) [78] and medicago (*M. truncatula*) [107]. A total of 173,383 protein sequences from five closely related legumes included in

(i) and 277,700 protein sequences from five related plants in (ii) were collected. All protein sequences for genes in such selected species were used to perform all-to-all BLAST. Alignments with E-value $\leq$ E-30 were chosen to group orthologous protein sequences in OrthoMCL [51].

There were 19,539 OrthoMCL gene families clustered in five legumes comparisons among cowpea, adzuki bean, mungbean, common bean and soy bean, whereas 17,600 gene clusters were consisted of cowpea genes. Exactly 13,192 gene clusters were shared by all five legumes, indicating common orthologs in such legumes. Exactly 236 gene clusters were specific to cowpea, containing 686 cowpea genes (Figure 2.7-a).

For gene clustering among cowpea, soybean, arabidopsis, rice and *Medicago truncatula*, a total of 27,027 clusters were identified with 9,057 orthologous clusters. The total number of gene clusters were higher than the clustering on the five related legumes and the number of orthologs were lower, mainly because genes were less conserved than in legume species. Exactly 651 gene clusters containing 2110 cowpea genes were specific to cowpea (Figure 2.7-b).

2.2 Structural Variations on IT97K-499-35

2.2.1 SNPs

Whole-genome shotgun data from an additional 36 diverse accessions relevant to Africa, China and USA were previously used to identify 957,710 single-nucleotide polymorphisms (SNPs; hereinafter referred as the "1M list") [74]. Almost all (99.83%) of the

Table 2.4: Number and location of SNPs relative to annotated cowpea genes.

	1M list	iSelect
# SNPs	957,710	51,128
# SNPs in genes (%)	336,285 (35%)	31,708 (62%)
# SNPs in exons (%)	138,892 (15%)	16,898 (33%)
# SNPs in or within 1 kb from gene (%)	460,709 (48%)	38,286 (75%)
# SNPs in or within 2 kb from gene (%)	540,773 (56%)	39,856 (78%)
# SNPs in or within 10 kb from gene (%)	792,318 (83%)	45,648 (89%)
# unique genes containing or near SNPs (< 1 kb) (%)	25,433 (85%)	19,319 (65%)
# unique genes containing or near SNPs (< 2 kb) (%)	26,130 (88%)	19,818 (67%)
# unique genes containing or near SNPs (< 10 kb) (%)	27,021 (91%)	21,205 (71%)

957,710 discovered SNPs were detected in the reference genome sequences, including 49,697 SNPs that can be assayed using the Illumina iSelect Consortium Array [74]. About 35% of the SNPs in the 1M list were associated with genes (336,285 SNPs), while that percentage increased to 62% in the iSelect array (31,708 SNPs; Table 2.4). This indicates that the intended bias towards genes in the iSelect array design [74] was successful. The number of annotated cowpea gene models containing a SNP was 23,266 (78% of total) or 27,021 (91% of total) when considering genes within 10 kb of a SNP (Table 2.4). In general, SNP density was lowest near centromeric regions (Figures 2.5 and Figure 2.8). This information enables formula-based selection of SNPs, including distance to gene and recombination rate. When these metrics are combined with minor allele frequency and nearness to a trait determinant, one can choose an optimal set of SNPs for a given constraint, for example cost minimization, on the number of markers.

2.2.2 4.2Mb Inversion on Chromosome 3

Ten genetic maps were used to anchor and orient scaffolds into pseudochromosomes. Plots of genetic locations against physical positions for SNPs on seven of those

genetic maps showed a relatively large region in inverted orientation relative to IT97K-499-35 (Figures 2.9(a)). The other three genetic maps showed no recombination in this region, suggesting that the two parents in the cross had opposite orientations. The genotype data from all of the parental lines showed that one of the parents from each of those three populations, but not the other parent, had the same haplotype as IT97K-499-35, and hence presumably the same orientation. To define the inversion breakpoints, WGS data available from some of these accessions [74] were used. In both break-point regions, contigs from accessions that presumably had the same orientation as the reference (type A) showed good alignments, while those from accessions with the opposite orientation (type B) aligned only until the breakpoints. An additional *de novo* assembly of a 'type B' accession enabled a sequence comparison with the reference genome for the entire genomic region containing the inversion (Figure 2.9(b)). This provided a confirmation of the chromosomal inversion and the position of the two breakpoints in the reference sequence: 36,118,991 bp (break-point 1) and 40,333,678 bp (break-point 2) for a 4.21-Mb inversion containing 242 genes. PCR amplifications of both break-point regions further validated this inversion.

2.3 Assembly and annotation of *Phytophthora infestans*

Phytophthora infestans is an oomycete which causes a potato and tomato disease known as late blight or potato blight. *P. infestans* has an estimated genome size of ≈240Mb with about eighteen thousand genes. However, the genome released in 2009 was quite fragmented with scaffold-level assembly [30]. Here a new pseudomolecule-level assembly of *P. infestans* is presented to provide a more contiguous and complete genome for further

Table 2.5: Statistics for our assembly of *Phytophthora infestans*

	P. infestans	*P. infestans* [30]	*P. sojae*	*P. ramorum*
estimated genome size (Mb)	240	240	95	65
assembly size (Mb)	247.23	228.54	86.0	66.7
N50 length (kb)	13450	44.5	105.7	47.5
# scaffolds	671	4,921	1,810	2,576
GC content (%)	51.54	51	54.4	53.9
# annotated genes	23,880	17,797	16,988	14,451

analysis. This genome was assembled using PacBio long reads, with a total assembly size of 247.23Mb and N50 of 13.45Mb (see Table 2.5). Fifteen chromosomes were constructed with sequence length range from 10.1Mb to 22.9Mb. Overall, this assembly provided a significant improvement compared to the assembly previously published in [30].

P. infestans, *P. sojae* and *P. ramorum* are the three major phylogenetic clades of phytophthora [11] and thus the genome assembly and annotation statistics of *P. sojae* and *P. ramorum* were also included in Table 2.5 for cross species comparisons.

The repeat library was built using RepeatModeler [24] based on *P. infestans* genome. High-frequency sequences were used to identify interspersed repeat seeds in multiple alignment extension in order to discover the youngest and most abundant repeat families; older transposable element (TE) families were detected using a clustering and relationship determination approaches. Structural LTR elements identification results was also included in the repeat library. A total of 780 custom repeats for *P. infestans* genome were used by RepeatMasker [77] to identify repetitive elements. An estimated 66.6% of *P.infestans* genome was composed of repeats: 45.51% was composed of long terminal repeat (LTR), 8.44% of DNA transposons and 11.26% of unclassified repeats (Table 2.6).

Table 2.6: Repeat analysis in our assembly of *P. infestans*

		Number of elements	Total length (bp)	Genome fraction
SINEs		47	7,030	≈0%
LINEs		1,339	3,484,059	1.41%
	LINE1	164	286,552	0.12%
	L3/CR1	219	240,430	0.10%
LTR elements		48,378	112,514,363	45.51%
DNA elements		28,031	20,867,109	8.44%
Unclassified		40,022	27,833,020	11.26%
Total interspersed repeats			164,705,581	66.62%
Small RNA		40,022	27,833,020	1.35%
Satellites		16	1,716	≈0%
Simple repeats		6,112	285,238	0.12%
Low complexity		751	38,485	0.02%

A total of 110,667 transcripts were assembled from about 346 million reads of single-end Illumina RNA-Seq data using PASA [29]. These transcripts together with protein sequences from *Phytophthora infestans*, *Phytophthora parasitica*, *Phytophtora sojae*, *Pythium ultimum*, and Uniprot sequences were employed to annotate the genome using the Funannotate pipeline [79]. In this context, transcripts were aligned to genome by Minimap2 [49]; proteins were aligned by Exonerate [102]; RNA-Seq reads were aligned by BWA [50] with mapping quality ≥ 20. All such alignments allowed us to train Augustus for *de novo* gene predictions. Augustus predicted genes were then used to train Snap [43] and GlimmerHMM [63]. Genes predicted from Augustus, Snap, GlimmerHMM and previously trained GeneMark's genes were passed to EvidenceModeler [31] to generate comprehensive gene models. tRNA genes were predicted using tRNAscan-SE [61]. This pipeline identified 20,195 protein-coding genes and 11,272 tRNA genes. Previously annotated protein sequences of *Phytophthora infestans* [30] were mapped to these newly annotated protein sequences. Sequences with an alignment of less than 70% of their length or less than 70% identity were lifted over to this genome assembly. Exactly 4620 genes (4420 protein coding

genes and 200 pseudogenes) were lifted by Flo [85], 117 genes (115 protein coding genes and two pseudogenes) were manually lifted. In total, 36,204 genes were identified, where 24,730 were protein-coding genes.

The assembled genome was annotated using *de novo* gene prediction and transcript evidence based on RNA-Seq data from different tissue, and protein sequences of protein sequences from *Phytophthora infestans*, *Phytophthora parasitica*, *Phytophtora sojae*, *Pythium ultimum*, and UNIPROT. In total 35,152 genes were annotated where 23,880 genes were protein-coding genes and 11,272 tRNA genes. There were significantly higher number of genes annotated by this method than what was previously reported [30], which indicates a better and more complete gene set for *P. infestans* (Table 2.5). Most (95.8%) of the 215 protist genes in BUSCO v3 were found in this phytophthora gene set, indicating high completeness of genome assembly and annotations. The average length of protein-coding genes was 1815 bp; the average exon length was 578; there were 2.76 exons per gene on average. The average length of tRNA genes was 73 bp. The GC content in protein coding gene regions was 52.88%, higher than in intergenic regions of 50.30%. The GC content in coding exons was 53.85%, also higher than 49.50% in introns and UTRs.

2.3.1 Effectors

Phytophthora species, like many pathogens, secrete effector proteins that alter host physiology and facilitate colonization [115]. The genome of *P. infestans* revealed large complex families of effector genes encoding secreted proteins that are implicated in pathogenesis [39]. Among different categories of effector proteins, most notable are the RXLR and Crinkler (CRN) cytoplasmic effectors. All oomycete avirulence genes (encoding

products recognized by plant hosts and resulting in host immunity) discovered so far encode RXLR effectors, modular secreted proteins containing the amino-terminal motif Arg-X-Leu-Arg (in which X represents any amino acid) that defines a domain required for delivery inside plant cells [120]. CRN cytoplasmic effectors were originally identified from *P. infestans* transcripts encoding putative secreted peptides that elicit necrosis in plants, a characteristic of plant innate immunity [115].

Several computational strategies have been previously used to identify candidate effectors within genome sequences, relying on matches to an HMM profile or to a sequence pattern. Here we used effectR [106] to search for the motifs of interest (RxLR-EER motif for RxLR effectors and LFLAK motif for CRN effectors) using regular expressions.

We exploited known motifs and other conserved sequence features to predict 588 RXLR genes in the *P. infestans* genome, slightly higher than previously reported RXLR genes in [30]. RXLR genes are notably expanded in *P. infestans*, with 60% more predicted than in *P. sojae* and *P. ramorum* (Table 2.7). Analysis of the *P. infestans* genome sequence revealed an enormous family of 133 CRN genes of unexpected complexity and diversity, that is heavily expanded in *P. infestans* relative to *P. sojae* (100 CRNs) and *P. ramorum* (19 CRNs) (Table 2.7). Like RXLRs, CRNs are modular proteins. CRNs are defined by a highly conserved N-terminal 50-amino-acid LFLAK domain.

Bacterial proteins that contain PAAR repeat sequences are associated with the VgrG-like spikes found in the type VI secretion system of bacteria and have been shown to be essential in target cell killing by the bacterial species *Vibrio cholerae* and *Acitenobacter baylyi* [95]. The PAAR proteins have a homonymous amino acid sequence motif (PAAR)

with one or more repeats. Here we also use effectR to identify 884 proteins with PAAR repeats.

	P.infestans	*P.infestans* in [30]	*P.sojae*	*P.ramorum*
CRN	133	196	100	19
RxLR	588	563	350	350
PAAR	884			

Table 2.7: Effectors of Phytophthora

2.4 Assembly and annotation of *Babesia duncani*

Babesia species are tick-transmitted parasites that infect red blood cells and can cause babesiosis, a malaria-like disease with major health impacts [64]. Of the different species of *Babesia* that infect humans, *Babesia duncani* is identified as a zoonotic pathogen and causes severe infection in immunocompetent individuals [117]. Here, the genome of *B. duncani* was assembled using PacBio long reads. Different long-read assembly tools were used, namely CANU [42], Minimap+MiniAsm [48] and wtdbg2 [87]. The assemblers were used to generate three preliminary *de novo* assemblies. Then, nanopolish [58] and pilon [118] were used to polish the assemblies independently, producing a six additional assemblies. In total, nine preliminary assemblies were generated. Table 2.8 includes the genome statistics for these nine assemblies. Observe that Minimap-based assembly had a larger assembly size than other two tools, in unpolished and two polished results. CANU had the smallest results in all three categories. Nanopolish seemed to increase the assembly size slightly, while pilon reduced it. All assemblies had similar GC content of around 37%.

Table 2.8: Statistics of various *de novo* assemblies of *Babesia duncani*

	CANU	minimap	wtdbg2	CANU+nano	minimap+nano	wtdbg2+nano	CANU+pilon	minimap+pilon	wtdbg2+pilon
Assembly Size (bp)	8,983,470	9,635,426	9,080,066	9,030,406	9,637,100	9,142,069	8,737,144	9,345,322	8,878,105
N50(bp)	1,667,772	1,389,405	1,679,388	1,676,306	1,390,359	1,692,087	1,608,929	1,327,910	1,632,243
GC content(%)	37.1	37.0	37.3	37.1	37.1	37.4	37.2	37.2	37.5
#Scaffolds	25	21	52	25	21	52	25	21	52
Repeats(%)	1.27	1.2	1.1	1.26	1.16	1.11	1.3	1.21	1.1
#Genes	6,286	6,540	5,564	4,998	5,122	4,920	5,033	4,704	4,956
Avg Gene Length	625.05	795.16	524.04	1160.90	1185.42	972.93	1049.29	1233.64	972.93
Avg Exon Length	267.49	314.53	261.55	403.70	407.05	736.36	747.07	798.47	736.36
#Exon per Gene	2.00	2.20	1.76	2.61	2.60	1.27	1.33	1.44	1.27

Transcripts were assembled using Trinity [27], Cufflinks [116], StringTie [84] and PASA [29] from ≈5.8M pairs of 2x100 paired-end Illumina RNA-seq reads extracted from blood tissue. These transcripts, together with PASA generated genes and protein sequences of *Babesia bigemina*, *Babesia bovis*, *Babesia microti*, *Plasmodium falciparum*, *Toxoplasma gondii* and Swiss-Prot proteomes, were used in the FUNAnnotate pipeline to generate gene models. Transcripts were aligned to genome by Minimap2; proteins were aligned by Exonerate. All such alignments were used to train and run Augustus for *de novo* gene predictions. Augustus predicted genes were then employed to train Snap and GlimmerHMM. Genes predicted from Augustus, Snap, GlimmerHMM and previously trained GeneMark's genes, as well as input PASA genes were passed to EvidenceModeler to generate comprehensive gene models. Genes were filtered by length, spanning gaps, and transposable elements match. tRNA genes were predicted using tRNAscan-SE [61].

Observe in Table 2.8 that the number of annotated genes were lower in polished assemblies, but those genes were longer in length with longer exons, on average. These results indicated that the polishing step may correct base-pairs which encoded stop codon incorrectly and thus be able to combine fragmented genes in unpolished assemblies into longer genes.

Table 2.9: Genome statistics of species related to *B. duncani*

	P. falciparum	*B. bovis*	*B. microti*	*B. bigemina*
Assembly size (bp)	23,326,872	8,179,706	6,434,485	13,840,936
# chromosomes/scaffolds	15	14	6	483
N50 (bp)	1,687,656	1,797,577	1,766,409	2,541,256
GC content(%)	19.3	41.6	36.2	50.6
# genes	5,392	3,076	3,601	5,079
average gene length (bp)	2483.33	1635.14	1438.77	1810.91
average # exons per gene	2.59	2.74	7.45	2.62
average exon length (bp)	868.91	556.09	356.89	652.73

A comparison of genome statistics for species related to *Babesia duncani* is shown in Table 2.9, which includes *Plasmodium falciparum*, *Babesia bovis*, *Babesia microti* and *Babesia bigemina*. Observe that the sizes of our nine assemblies of *Babesia duncani*, ranging from 8.7Mb to 9.6Mb, were similar to assembly size of *B. bovis*, and larger than *B. microti* which also infects human. The number of genes in our nine assemblies ranged from 4704 to 6540, which was similar or slightly higher than gene counts in *B. bigemina* and *P. falciparum*.

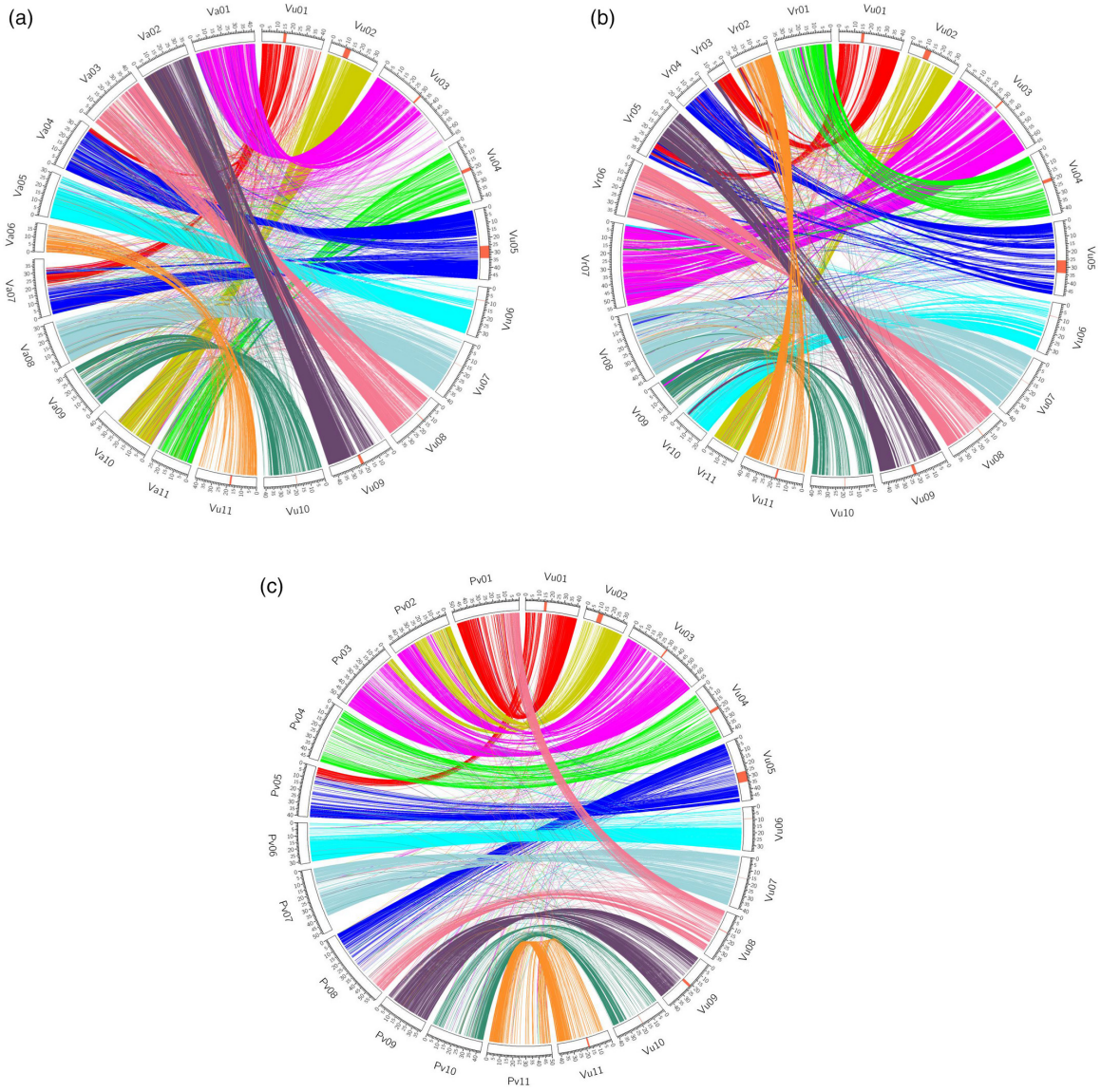

Figure 2.3: Synteny analysis between cowpea and other closely related legumes; (a) adzuki bean (Va; *V. angularis*); (b) mung bean (Vr; *V. radiata*); and (c) common bean (Pv; *P. vulgaris*); the cowpea (Vu; *V. unguiculata*) genome uses the revised chromosome numbering system.

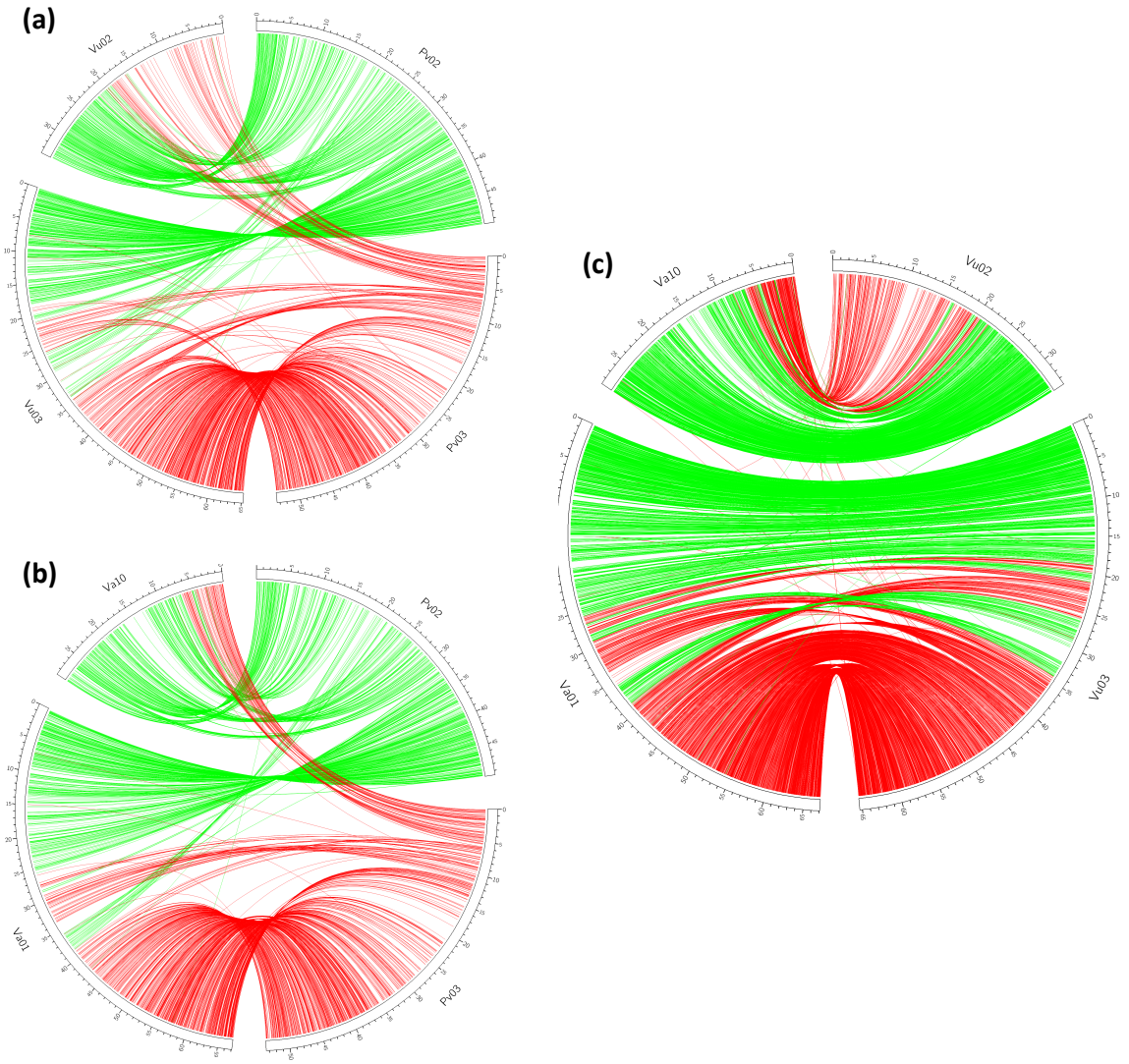

Figure 2.4: Synteny analysis between cowpea (Vu; *V. unguiculata*) and other closely related species; (a) common bean (Pv; *P. vulgaris*) to cowpea (Vu; *V. unguiculata*); (b) common bean to adzuki bean (Va; *V. angularis*); and (c) cowpea to adzuki bean.

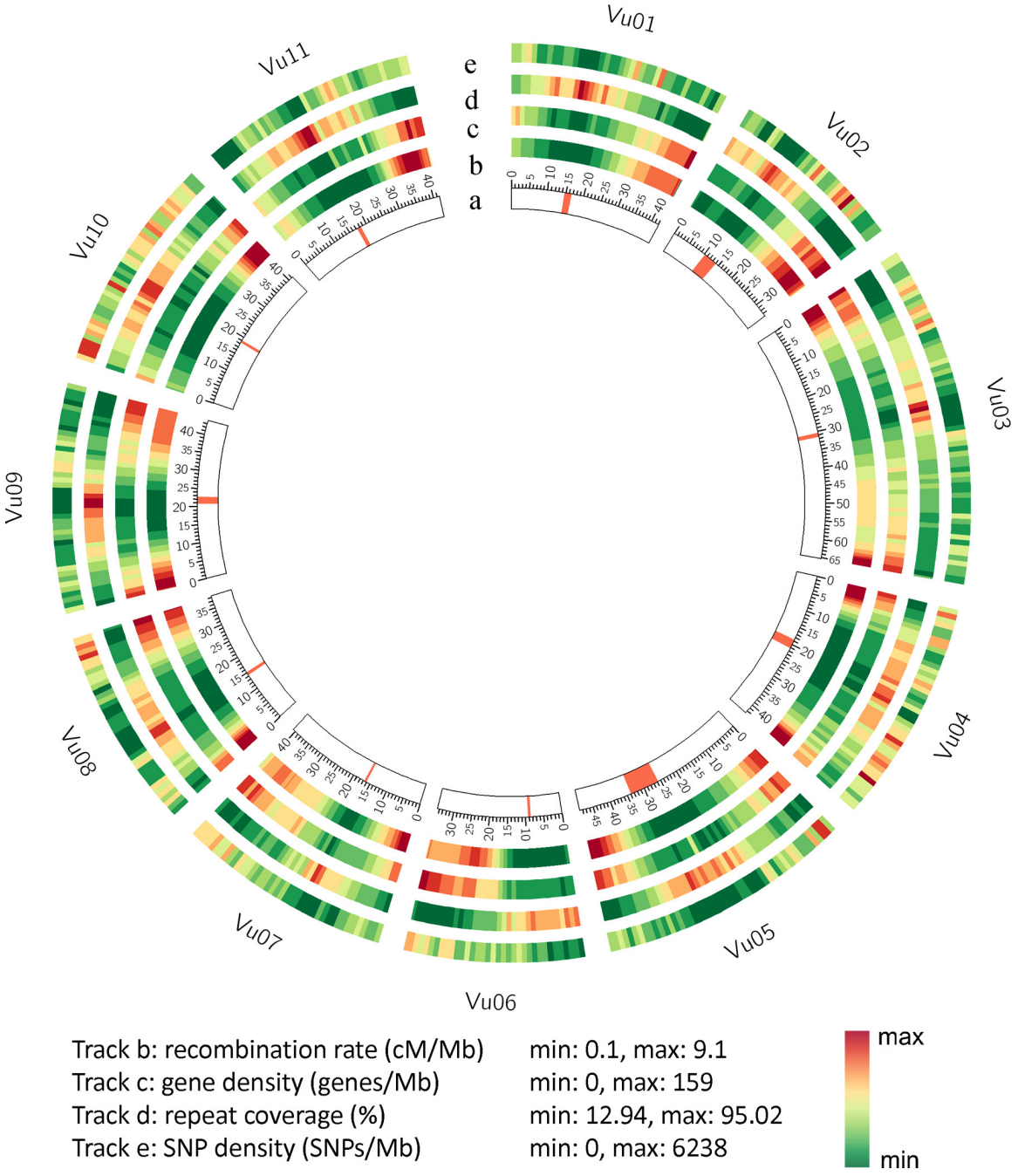

Figure 2.5: Summary of recombination rate (b), gene density (c), repeat coverage (d) and SNP density (e) along the eleven chromosomes of the cowpea genome (see text for details); orange blocks in track (a) represent predicted centromeric positions

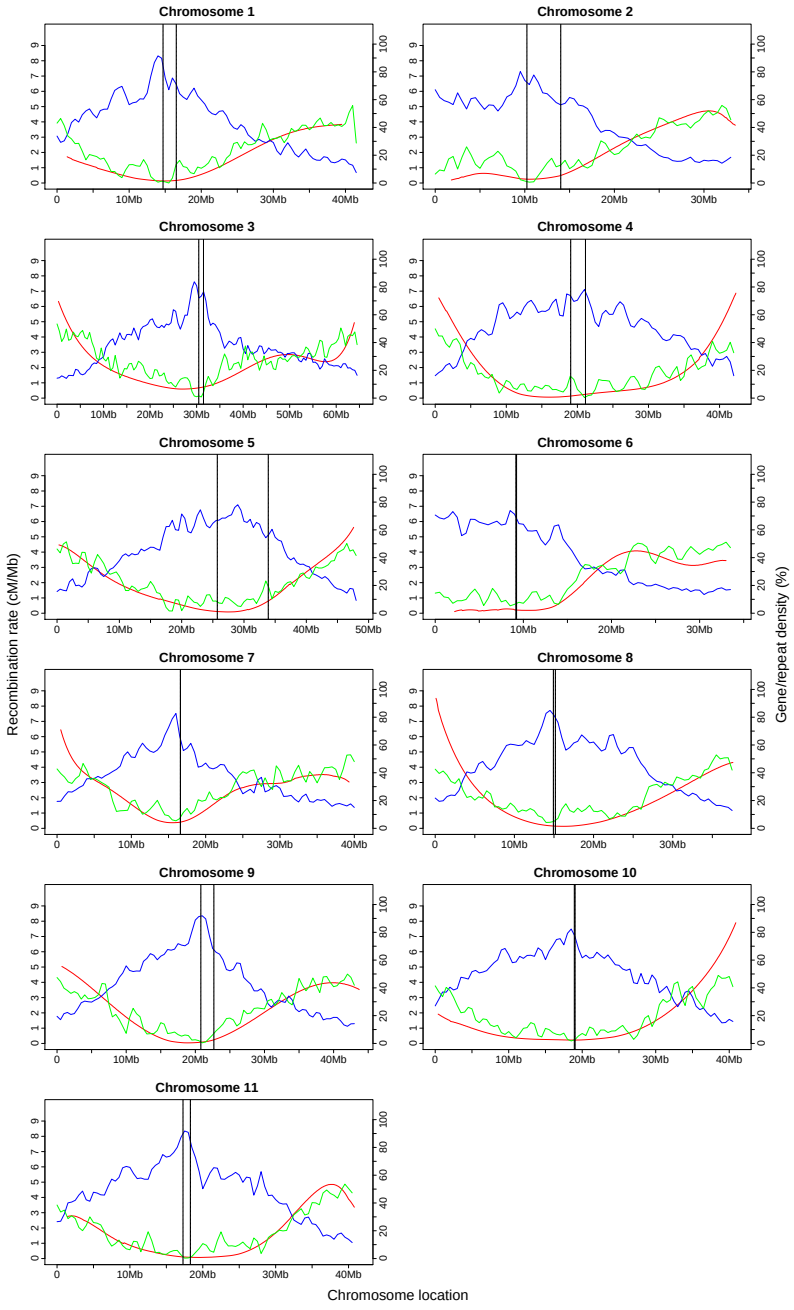

Figure 2.6: Gene density, repeat density, and recombination rate in the cowpea genome.

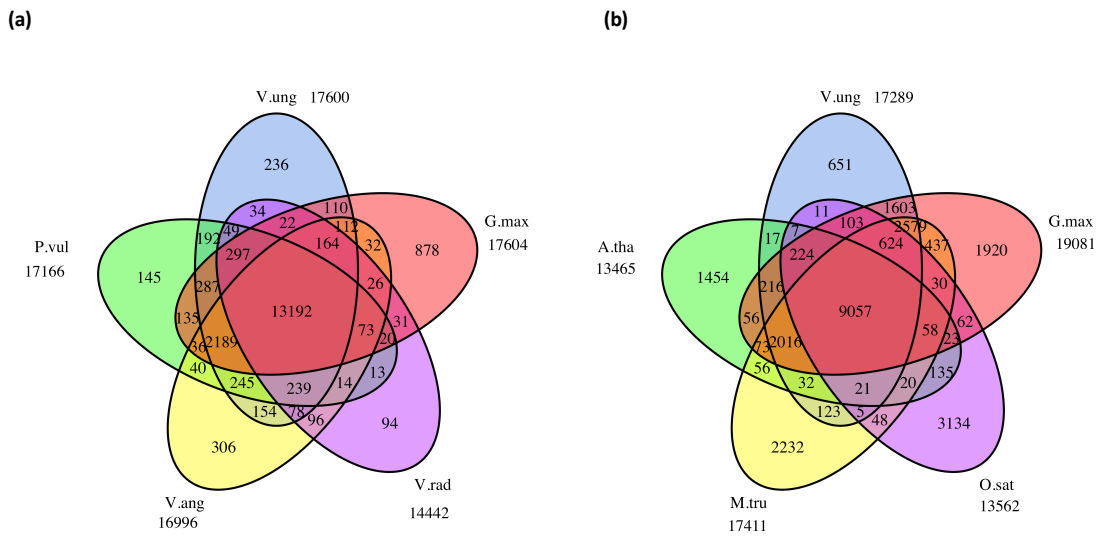

Figure 2.7: Venn diagram for the gene families shared by these five species

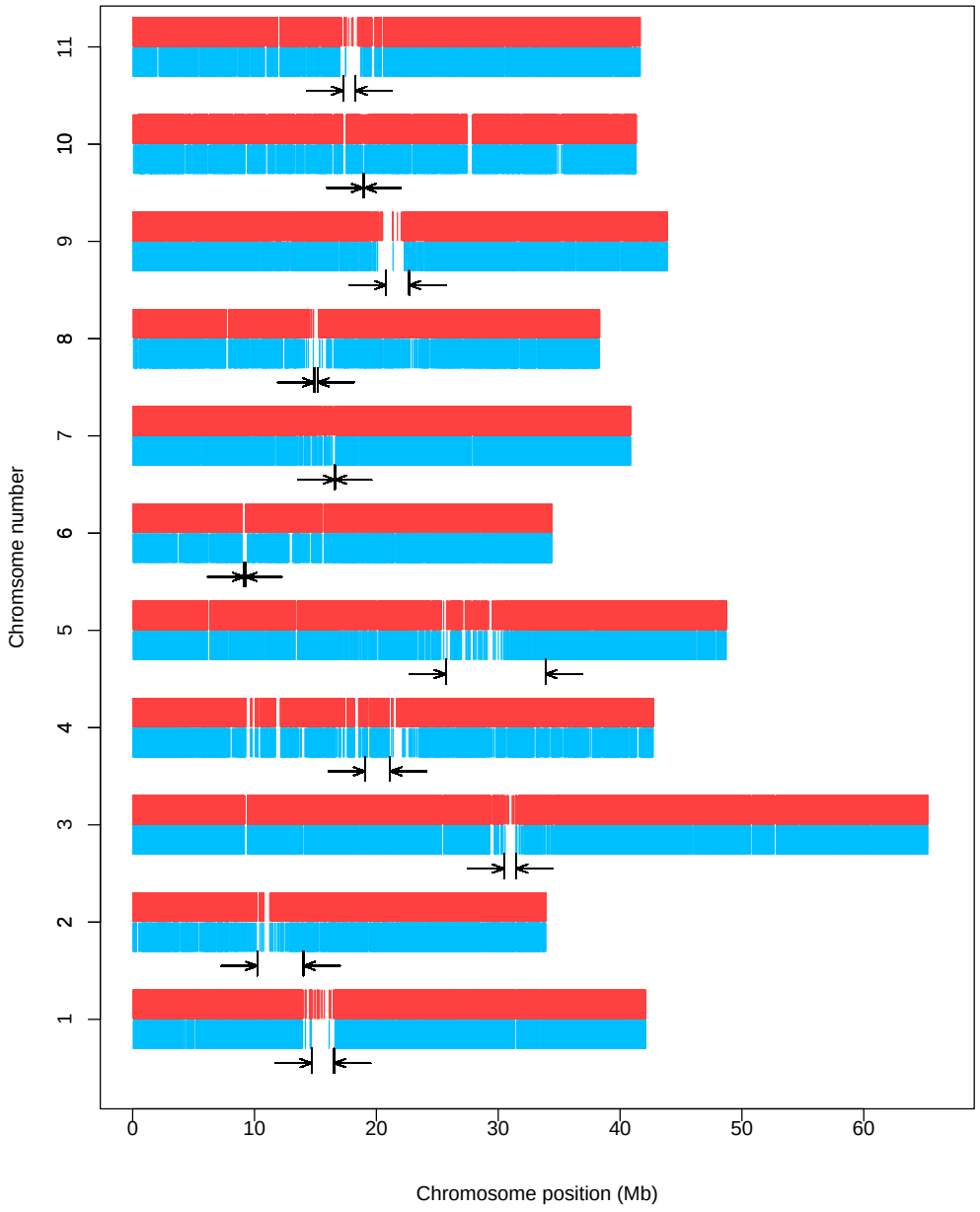

Figure 2.8: SNP distribution in the cowpea genome. SNPs from the "1M list" (red) and the Illumina iSelect Consortium Array (blue). Arrows delimit the predicted centromeric regions.

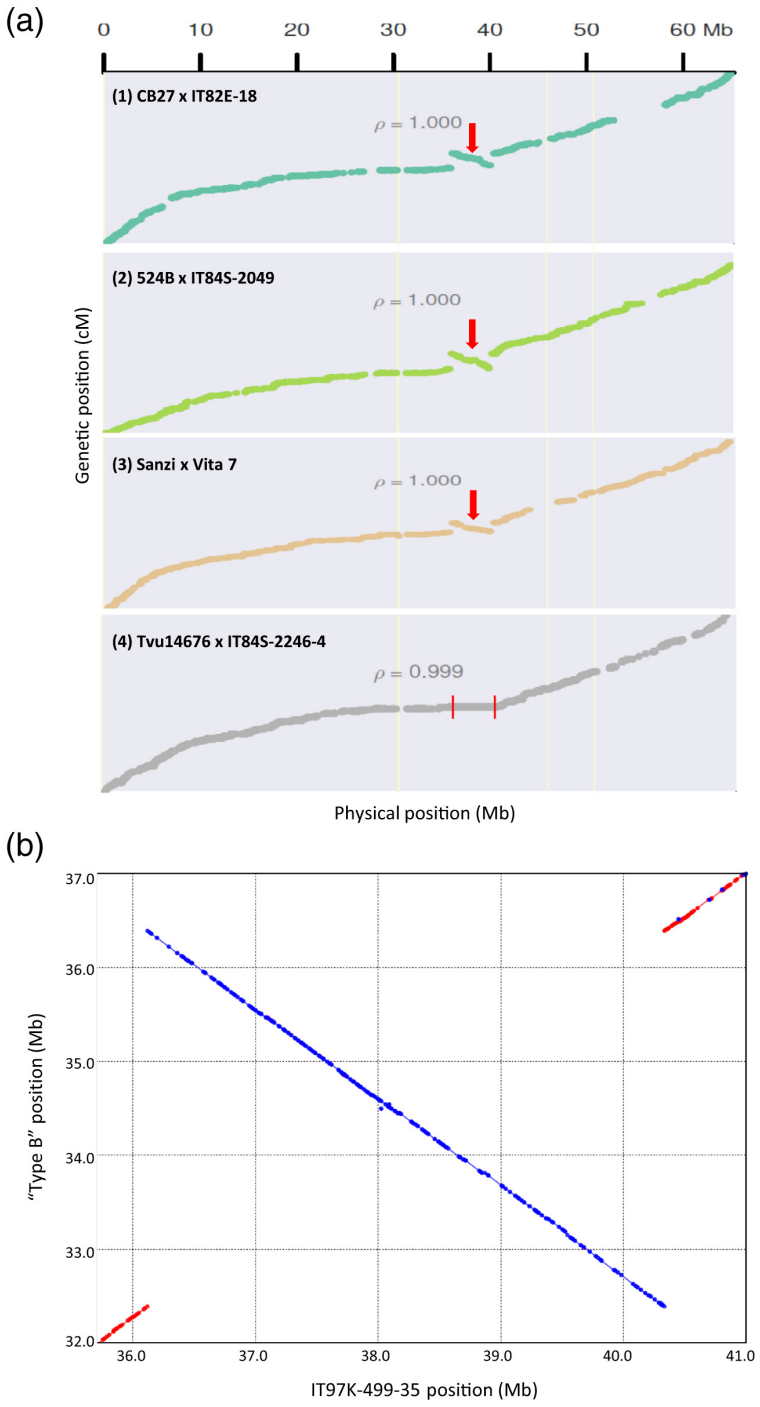

Figure 2.9: A large chromosomal inversion detected on chromosome 3 in cowpea.

Chapter 3

Cowpea Pan-genome Analysis

Cowpea (*Vigna unguiculata L. Walp.*) is a diploid warm-season legume, also known as black-eyed pea, among other common names. Cowpea is relevant as a grain legume in the USA, Europe and Latin America, and as a fresh vegetable in China and elsewhere in Asia. Cowpea presently serves as a major source of calories and protein for many people, especially in developing countries.

The phylogenetic distribution of several hundred domesticated cowpea accessions from SNP data generated using the Cowpea iSelect Consortium array was described in [74], defining six major sub-populations. Representative within the *unguiculata* cultivar group include (1) IT97K-499-35 from the breeding program at the International Institute of Tropical Agriculture in Nigeria, (2) CB5-2 from early breeding activities for California blackeyes, (3) Suvita2 as a landrace from Burkina Faso, (4) Sanzi as a landrace from Ghana, and (5) UCR779 as a landrace from Botswana. In addition, two longbean accessions were included from the Asian sub-population, which represents the sesquipedalis cultivar group,

namely (6) elite TZ30, and (7) landrace ZN016. Elite African variety IT97K-499-35 is a Striga- and Alectra-resistant, high yielding blackeye variety developed at the International Institute of Tropical Agriculture in Ibadan, Nigeria[100]. CB5-2 is a fully inbred stock that is closely related to CB5, which was the predominant Blackeye of the US Southwest for several decades. CB5 (Blackeye 8415) was bred by WW Mackie at the University of California [62] to add resistances to Fusarium wilt and nematodes to a previously preferred California Blackeye landrace. Suvita2, also known as Gorom Local (IITA accession TVu-15553, US NPGR PI 583259), is resistant to bruchids and some races of Striga, and is relatively drought tolerant. This landrace was collected from a local market by VD Aggarwal at the Institut de l'Environnement et de Recherches Agricoles (INERA) in Burkina Faso [3]. Sanzi is an early flowering, small-seeded landrace from Ghana with resistance to flower bud thrips [101]. UCR779 (PI 583014) is a landrace from Botswana, provided to UC Riverside as BOTS 19A in 1987 by CJ DeMooy of Colorado State University. Yardlong bean or asparagus bean (*V. unguiculata L. Walp. ssp. sesquipedalis*), the vegetable type of cowpea, is widely grown in Asian countries for consumption of tender long pods. TZ30 is an elite Chinese variety with the pod length of around 60 cm. ZN016 is a landrace originating from southeastern China with the pod length of around 35 cm and showing resistances to multiple major diseases of cowpea.

As explained in Chapter 1, the pan-genome of a species has been defined as the full complement of genes contained within that species. Initial results from *Zea mays*, *Brachypodium distachyon* and a few other plants have shown that each pan-genome is much larger than any individual genome. We have been characterizing the pan-genome of

domesticated cowpea to facilitate the use of genome information when choosing parents and progeny in cowpea breeding and for decisions related to germplasm management, while also to support and encourage basic research on this crop plant of such historical and current-day importance for worldwide food and nutritional security. For example, large inversions in certain regions of the genome when comparing one accession to another, and the relative rates of genetic variation and recombination as a function of position along a chromosome, must be considered in the design of breeding strategies. Additionally, the list of well-mapped major loci controlling traits relevant to agriculture, which are ripe for mechanistic and population diversity studies, has been rapidly expanding, including biotic and abiotic stress resilience, time to flowering, plant architecture, and seed and pod characteristics, among others. Cowpea has a compact genome of about 641 Mb [60] and the advantage of being usually an inbreeder, which has facilitated the establishment of fully inbred (single haplotype) stocks for sequencing.

To capture the genomic diversity of this important legume and build a cowpea pan-genome with full complement of genes, we performed *de novo* genome assemblies for representatives of the six major sub-populations of cowpea: CB5-2, Suvita2, Sanzi, UCR779, ZN016 and TZ30.

3.1 *De Novo* Genome Assembly and Annotation

The IT97K-499-35 genome was previously sequenced and assembled from Pacific Biosciences long reads, two Bionano Genomics optical maps and ten genetic linkage maps [60]. Six new *de novo* assemblies for CB5-2, Suvita2, Sanzi, UCR779, TZ30 and ZN016

Table 3.1: Genome Statistics of Pan-genome

	IT97K-499-35	CB5-2	Suvita2	Sanzi	UCR779	ZN016	TZ30
Assembly size (bp)	519,435,864	448,043,751	447,585,192	447,277,261	453,970,486	451,130,807	451,468,680
N50 (bp)	41,684,185	36,897,245	36,142,647	34,759,918	35,700,653	37,764,243	36,906,789
#Contigs/scaffolds	686	6,534	9,123	11,268	12,939	7,032	6,771
#Contigs/scaffolds ¿= 100kbp	103	28	28	17	13	28	48
#contigs/scaffolds ¿= 1Mbp	13	11	11	11	11	11	11
#contigs/scaffolds ¿= 10Mp	11	11	11	11	11	11	11
Longest contig (bp)	65,292,630	60,086,998	58,539,223	58,655,738	58,369,212	60,653,587	59,481,915
Repetitive content	47.25%	45.52%	45.43%	45.50%	45.89%	45.68%	45.76%
Annotated genes (#)	31,948	28,297	28,545	28,461	28,562	27,723	27,742
BUSCO completeness							
Genome	1595 98.8%	1574 97.5%	1580 97.8%	1581 97.9%	1574 97.6%	1589 98.5%	1583 98.1%
Transcripts	1594 98.8%	1570 97.2%	1582 98.0%	1585 98.2%	1581 97.9%	1584 98.1%	1580 97.8%
Proteins	1595 98.8%	1569 97.3%	1584 98.2%	1587 98.3%	1585 98.2%	1584 98.1%	1582 98.0%

were produced by Dovetail Genomics from Illumina short reads (150x2). Dovetail Genomics used Meraculous [16] to assemble the reads, then Chicago and Hi-C libraries (using their proprietary pipeline) to resolve mis-assemblies and increase contiguity. The final assemblies were processed using AllMaps [108] using ten high-density genetic linkage maps previously generated [74, 60]. A summary of the main statistics for the seven assemblies is reported in Table 3.1.

The contiguity of the new six assemblies, as indicated by their N50s, is comparable to the PacBio assembly for IT97K-499-35 despite being based on short-read sequences. In all six assemblies each of the eleven chromosomes of cowpea is represented by a single scaffold. These six newly assembled genomes ranged very narrowly in size from 447.58 Mb to 453.97 Mb, with a mean of 449.91 Mb. IT97K-499-35 had a 15% larger (more complete) assembled size (519.44 Mb) than these six accessions due to long-read sequencing and optical mapping. The difference between the two sequencing methods is partially attributable to the centromeric regions of some chromosomes of the six new accessions appearing to be missing from the assemblies (Figure 3.5, Figure 3.6 and Table 3.2). The new six accessions share

the same percentage of repetitive content of about 45-46%. IT97K-499-35 had a slightly higher repetitive content, which may be a result of higher completeness of the centromeric regions.

All cowpea genomes of different accessions were annotated using the JGI plant genome annotation pipelines that integrated gene call and gene model improvement. RNA was prepared from each accession to support gene annotation from young leaves and roots of well-watered and water limited plants, flower buds, seeds at the color break stage of maturation, and pods from 2-5 days after pollination. The number of genes annotated in the six new assemblies ranged from 27,723 to 28,562 with a mean of 28,222 (Table 3.1). IT97K-499-35 had ∼13% more annotated genes with a total of 31,948, reflecting deeper transcriptome sequencing and to some extent the more complete assembly of the IT97K-499-35 genome using PacBio long-read sequences. The number of alternative transcripts in the six new assemblies ranged from 15,088 to 17,115. Again, IT97K-499-35 had a higher number alternative transcripts, a total of 22,536. The average number of exons was 5.4 in each of the six new assemblies, and 5.2 in IT97K-4899-35, with a median length ranging from 162 to 169 bp. Gene density and repeat density were computed in 1 Mb non-overlapping sliding windows along each chromosome (Figure 3.7b and Figure 3.2), and in each accession (Figure 3.1). All chromosomes have a higher gene density in their telomeric regions, while repeat density peaks in the centromeric regions. Also, all accessions have similar gene and repeat density. All accessions have high BUSCO v4 completeness at the genome, transcript and protein levels, again with somewhat higher numbers for IT97K-499-35 than the six new assemblies.

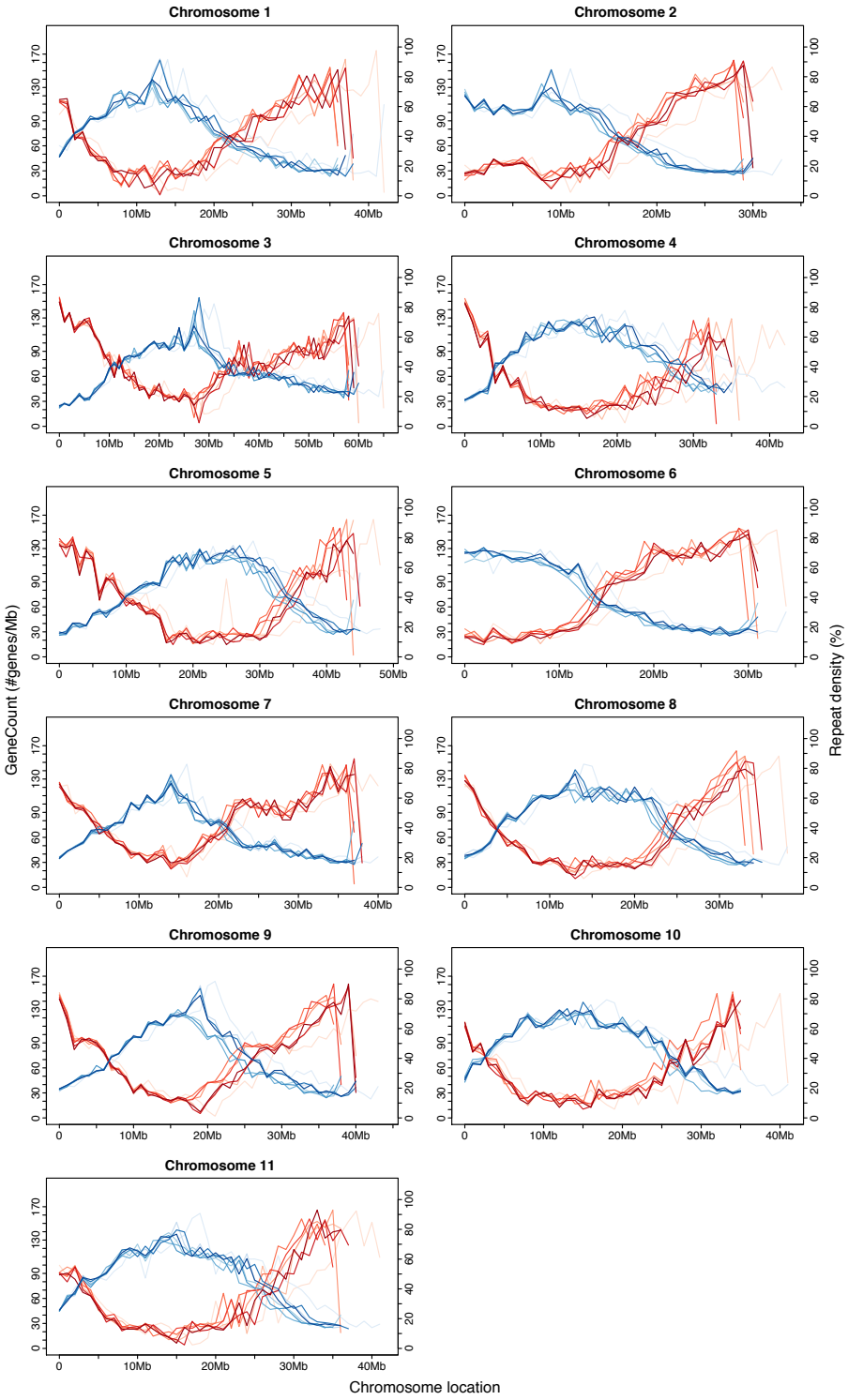

Figure 3.1: Gene density (red) and repeat density (blue)

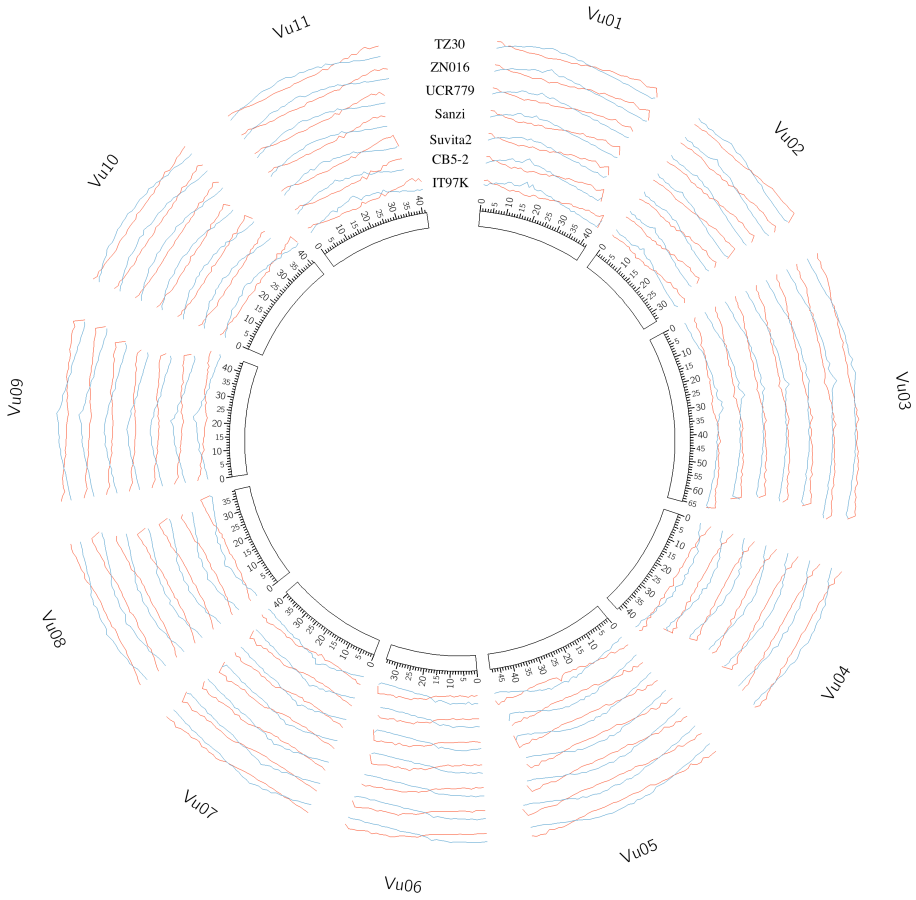

Figure 3.2: Gene density (red) and repeat density (blue)

Gene density and repeat density were calculated in 1Mb non-overlapping sliding windows. In Figure 3.1, different shade of red/blue indicated different accessions, from light to dark as IT97K-499-35, CB5-2, Suvita2, Sanzi, UCR779, ZN016, TZ30 in order. The peaks of IT97K-499-35 in both gene density and repeat density for some chromosomes shifted towards right in a few Mb, mainly because of the completeness in centromeric regions in IT97K-499-35 and thus a larger chromosome size with a larger position of long arm ends of chromsomes.

Table 3.2: Putative centromeric region coordinates (all numbers are bp)

Assemblies	IT97K-499-35			CB5-2			Sanzi			ZN016			TZ30		
chr	start	end	size	start	end	size	start	end	size	start	end	size	start	end	size
1	14,698,036	16,525,496	1,827,460	12,796,513	14,236,341	1,439,828			-	13,289,165	17,690,285	4,401,120			-
2	10,238,236	14,020,258	3,782,022			-			-			-			-
3	30,476,981	31,470,261	993,280			-			-			-			-
4	19,069,641	21,130,843	2,061,202	16,669,712	18,502,124	1,832,412	15,648,675	16,247,976	599,301			-			-
5	25,704,431	33,885,354	8,180,923	26,553,015	26,811,298	258,283			-	25,404,785	27,305,040	1,900,255			-
6	9,156,830	9,235,637	78,807			-			-			-			-
7	16,587,031	16,604,960	17,929	14,903,264	14,933,285	30,021			-	14,573,279	14,601,618	28,339	13,530,446	14,858,230	1,327,784
8	14,914,119	15,164,402	250,283	12,820,933	14,867,106	2,046,173			-	13,476,113	14,282,079	805,966			-
9	20,802,610	22,685,597	1,882,987			-			-			-			-
10	18,917,563	19,028,450	110,887			-			-			-			-
11	17,283,961	18,283,861	999,900			-			-			-			-
total			20,185,680			5,606,717			599,301			7,135,680			1,327,784

Centromeric regions were defined based on the presence of a 455-bp tandem repeat that was previously identified by FISH [38]. Table 3.2 shows the coordinates of the putative centromeric regions in IT97K-499-35 for all eleven chromosomes for total span of 20.18 Mb, in CB5-2 on five chromosomes for a total span of 5.6 Mb, in Sanzi on one chromosome for a total span of 0.59 Mb, in ZN016 on four chromosomes for a total of 7.13 Mb and in TZ30 on one chromosome for 1.32 Mb. The tandem repeat was not found in any assembled chromosome in Suvita2 or UCR779, nor in the other chromosome assemblies where coordinates are not listed in Table 3.2.

3.2 Pairwise Whole Genome Comparisons

Pairwise whole genome alignment (WGA) identified synteny relations as well as large structural variations between two genome assemblies. To perform WGA for pan-genomes, each of the six newly sequenced accession were compared to the published cowpea reference genome of IT97K-499-35 independently. Alignments were generated using MUMmer3 [45], with a minimum length of an exact match set to 100 bp. Alignments with a length less than 10 kb were filtered out. The output alignments between genomes were visualized using Circos v0.69-3 [44] Synteny relations identified by WGA detect whole chromosome

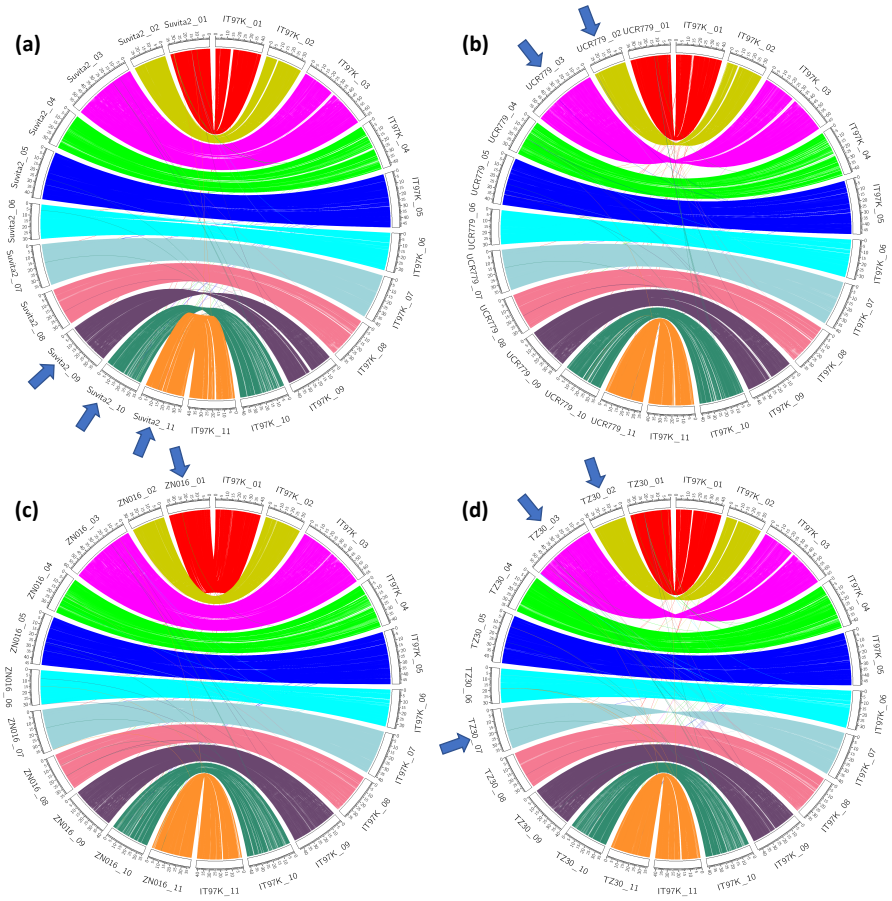

Figure 3.3: Whole Chromosome Inversions in Suvita2(a), UCR779(b), ZN016(c) and TZ30(d)

inversion in Suvita2, UCR779, ZN016 and TZ30 draft genome assembly. In Figure 3.3, Suvita2 chromosome 9, 10, 11 were reversed compared to corresponding chromosomes in IT97K-499-35, as well as UCR779 chromosome 2, 3, ZN016 chromosome 1 and TZ30 chromosome 2, 3, 7. Such chromosomes were reversed in order to keep the same orientations across pan-genome. Reversed chromosomes were used in following analyses.

Extensive synteny had been observed between IT97K-499-35 and other accessions. (Figure 3.4) There were gaps with missing synteny near centromeric regions on some chro-

mosomes of IT97K-499-35, which indicated incompleteness of centromeric regions on some other accessions.

In order to better visualize the structural variations between IT97K-499-35 and other six accessions, we carried out a pairwise whole genome comparison using dot plots for each chromosome. Each individual genome for the new six accessions was aligned against the IT97K-499-35 genome assembly using Minimap2 [49]. The resulting pairwise whole-genome alignments were visualized chromosome by chromosome with a modified dotplotly script. Figure 3.5 and Figure 3.6 provided direct visualization for large structural variations among cowpea pan-genome, including centromeric region deletions, inversions and translocations.

3.3 Variation Analysis

3.3.1 Present-Absent Variations

To answer the question of whether adding more accessions is likely to significantly change the numbers and proportions of core, dispensable and private genes, we carried out the same homology analysis above by using subsets of accessions of size $s = 1, 2, \ldots, 7$. For each value of s, we computed the number of core, dispensable and private genes for all possible choices of s accessions. In Figure 3.7-A, the green points represent the cumulative number of distinct genes in all accessions (i.e., in which homologous genes across accessions are counted once, or the set of pan-genes), and the yellow points are the number of core genes, as a function of s (on the x-axis). Observe that the number of green/yellow points for each value of s is $\binom{7}{s}$, that is 7 for $s = 1$ or $s = 6$, 21 for $s = 2$ or $s = 5$, 35

for $s = 3$ or $s = 4$. The green curve is a polynomial fit on the pan-genes, while the yellow curve is a polynomial fit on the core genes. As expected, the number of pan-genes increased as additional accessions were "added" to pan-genome, while the number of core genes decreased. However, the fact that the yellow curve is flattening considerably for $s = 6$ or $s = 7$ indicates that the vast majority of the core genes have been identified with these seven diverse accessions. In contrast, the green curve is not flattening, indicating that there are many more dispensable and private genes not included among these seven accessions, which means also that it would be necessary to sequence more genomes to thoroughly represent the entire pan-genome of cowpea.

Present-absent variation analyses were performed, both at the gene level and at the genome level, by comparing the seven cowpea genome assemblies to each other, along with their corresponding annotations.

For gene-level analysis, homologous genes were identified using BLAT. Coding sequences (CDS) of genes in IT97K-499-35 were mapped to CDS from the other six accessions using BLAT [41]. Alignments with 95% identity and 90% length coverage were counted as positive matches. If there was more than one gene hit, neighboring genes were used to determine one-to-one correspondence. After finding the correspondences between IT97K-499-35 genes to genes from the other accessions, the same strategy was repeated to seek correspondences for unpaired genes from CB5-2 by mapping to unpaired genes from remaining five accessions, in the order of Suvita2, Sanzi, UCR779, ZN016 and TZ30. Genes that were present in all seven accessions were called core, genes present in two to six accessions were called dispensable and genes present in only one accession were called private.

The total number of distinct genes (i.e., counting homologous genes only once) was 44,861, of which 21,330 were core, 10,065 were dispensable and 13,466 were private. Figure 3.7B shows the cumulative length of core, dispensable and private genes. Figure 3.7D shows the cumulative length of core, dispensable, and private genes in each individual accession. Figure 3.7F shows the fraction of each accession's transcriptome into core, dispensable, and private genes. IT97K-499-35 stood out as having higher number of private genes than the other accessions, which is likely due to the higher total number of genes annotated in IT97K-499-35. The other six accessions had a comparable number of private genes. The total lengths of dispensable genes in each accession were also similar.

For the genome-level analysis, seven assemblies were aligned using progressive-Mauve [20], to identify core, dispensable and private genomic blocks using a pan-genome representation called PGV. Genomic blocks that were present in all seven accessions were called core, blocks present in two to six accessions were called dispensable and block present in only one accession were called private. Genomic blocks were classified based on a reference-agnostic pan-genome representation called PGV. Briefly, PGV carries out a genome-wide multiple sequence alignment on all genomes using progressiveMauve, then computes the consensus ordering for the common blocks, which constitute the backbone of the pan-genome. Blocks in each genome are then ordered according to the consensus ordering. PGV detected 2,863 core blocks (comprising 77.41% of the cowpea genome), 11,856 dispensable blocks and 42,484 private blocks. Core genomic blocks covered about 360 Mb across all seven accessions on average; dispensable blocks were less than 50 Mb; and the collective total of private genomic blocks covered over 480 Mb (Figure 3.7C). Figure 3.7E

shows the cumulative length of core, dispensable and private blocks. IT97K-499-35 has more than 100 Mb of private blocks, which is much higher than other accessions and again is likely to be due to IT97K-499-35 having a more complete assembly than the other six accessions. The other accession had a comparable fraction of their genome to be private (Figure 3.7G), except for UCR779 which was significantly higher.

By comparing genes against genomic blocks (i.e., Figure 3.7B vs. C, D vs. E or F vs. G) it is clear that private genomic blocks constitute a much higher fraction of the genome, compared to the total length of private genes relative to the entire transcriptome. This indicates that private genomic regions have a lower gene density than other two types of regions.

3.3.2 Paralogous Genes

OrthoMCL was used to cluster the 200,235 protein sequences for all genes in the seven accessions. The total of 200,235 protein sequences for all genes in the seven accessions were used to perform all-to-all BLAST. Alignments with E-value less than E-30 were chosen to group orthologous protein sequences in OrthoMCL [51]. Resulting gene clusters had genes from all seven accessions were core, dispensable clusters had genes from 2-6 accessions, and private clusters had genes only from one accession. OrthoMCL produced 25,436 clusters of paralogous genes in the cowpea pan genome, of which 20,071 were core (i.e., they had genes from all seven accessions), 5155 were dispensable (i.e., they had genes from 2-6 accessions) and 147 were private to one accession (see Table 2). Exactly 140 of the private clusters were present in IT97K-499-35, while CB5-2, Sanzi and TZ30 did not have any private clusters. Exactly 16,028 of the core clusters were composed of genes

Table 3.3: Summary of OrthoMCL clusters

	IT97K-499-35	CB5-2	Suvita2	Sanzi	UCR779	ZN016	TZ30
Core			20,071				
Dispensable	2,989	2,806	2,925	2,911	2,719	2,282	2,277
Private	140	0	2	0	4	1	0

from the same core gene correspondence, and 9408 core clusters had genes from multiple correspondences. Exactly 3223 of dispensable clusters had genes from the same dispensable gene correspondence and 1932 had genes from multiple dispensable and/or private gene correspondence.

3.3.3 Structural Variations (SNPs, indels, and large SVs)

To investigate the structural variations in the cowpea pan-genome small variations, namely (a) SNPs and short indels, and large structural variations (b) inversions, translocations, deletions and inversions, were identified.

For SNPs and short indels the genome of each accession was used in turns as the "reference", mapping the reads for all other accessions against that genome. SNPs were detected from the mapped reads of each accession individually using GATK [67] as well as collectively. SNPs and indels were called using one reference genome versus the reads from six other accessions using two complementary approaches, namely (i) reads from the other six accessions than the specific reference were used and combined to call all small variations of this reference accession, and (ii) reads from each of other six accessions were used independently to call small variations. Selected reads were mapped to a specific reference genome using BWA. Alignments were merged (for first approach of combined small

Table 3.4: Summary of cowpea SNPs across the accessions

	IT97K-499-35	CB5-2	Suvita2	Sanzi	UCR779	ZN016	TZ30	Merged (reads)	Merged (by position)
IT97K-499-35		1,607,267	1,660,122	1,692,267	1,974,970	2,092,617	1,839,113	3,472,245	4,963,630
CB5-2	1,747,504		1,528,608	1,499,900	2,501,756	1,812,756	1,466,499	3,210,892	5,293,181
Suvita2	1,766,011	1,489,850		1,512,626	2,625,678	2,056,752	1,847,818	3,252,450	5,292,933
Sanzi	1,813,050	1,485,875	1,539,212		2,468,787	2,002,988	1,811,927	3,232,415	5,303,979
UCR779	2,029,638	2,427,619	2,605,364	2,417,122		2,440,589	2,424,480	3,438,834	5,349,217
ZN016	2,091,894	1,692,125	1,980,368	1,896,387	2,382,090		1,338,143	3,265,483	5,278,946
TZ30	1,939,167	1,442,388	1,865,974	1,802,432	2,472,527	1,422,297		3,228,656	5,302,674

Table 3.5: Summary of cowpea indels across the accessions

	IT97K-499-35	CB5-2	Suvita2	Sanzi	UCR779	ZN016	TZ30	All
IT97K-499-35		500,184	493,649	516,920	586,252	580,477	548,279	1,066,944
CB5-2	397,913		367,662	377,767	592,079	425,572	373,435	853,116
Suvita2	393,125	363,343		372,008	619,009	472,306	446,080	853,899
Sanzi	413,050	373,961	372,460		583,945	460,767	440,759	850,746
UCR779	462,494	575,960	610,235	575,146		556,606	572,875	898,249
ZN016	466,574	415,167	470,203	457,046	561,920		330,037	869,572
TZ30	435,661	364,014	442,477	436,458	579,168	331,440		852,546

variations using reads from six accessions) and removed duplicates using Picard [36]. The union of all SNPs for one accession was collected by merging SNPs identified by approach (i) by their locations on this accession. All SNP calling used GATK pipeline with same parameters and filters of '$QD < 2.0||FS > 200.0||ReadPosRankSum < -20.0||SOR > 10.0$'. Indels were identified using same pipeline with different filters of '$QD < 2.0||FS > 200.0||ReadPosRankSum < -20.0||SOR > 10.0$'.

SNPs from the six accessions were combined using two different methods: in the first, mapped reads (i.e., the BAM files) from each of the six accessions were merged together before re-calling the SNPs using GATK; in the second, the six SNP sets (i.e., the GATK outputs) were merged by taking the union of the SNPs based on their location on the target genome (i.e., a SNPs in two accessions was counted only once if it appeared in the same genomic position). Table 3.4 summarizes the number of SNPs detected, where the reference

genome is listed on each row. For instance, using Suvita2 as the reference, 1,489,850 SNPs were detected using mapped reads from CB5-2, compared to 2,625,678 SNPs using the reads from UCR779. Combining the SNPs by counting all distinct SNPs in the union of the six sets of SNPs, the number of SNPs for Suvita2 was 5,292,933. Observe in Table 3.4, that IT97K-499-35 had the highest number of combined SNPs (about 3.4 million), while the other six accessions had a lower, but similar, number of SNPs (about 3.2 million). For pairwise SNPs calling (i.e., using reads from only one accession), observe that when UCR779 was used as the reference, a much higher number SNPs was detected, indicating that UCR779 is the most "different" among all accessions. Also, CB5-2 has a relatively lower number of SNPs with respect to TZ30 and ZN016 than other accessions. This suggests that CB5-2 is somewhat closer to these two accessions than to the other four accessions. Table 3.5 reports a similar analysis for indels, where again UCR 779 stands out as being the most different from the other accessions. In Table 3.5, "all" refers to the number of indels obtained by combining the read mapping for all six accessions against that reference. Figure 3.9A shows a phylogenetic tree of the seven accessions that was constructed using SNPhylo based on the SNPs of IT97K-499-35.

The SNP frequency ranged from one in 309 bp to one in 139 bp; and the indel frequency ranged from one in 529 bp to one in 486 bp. Circos plots for SNP density (SNPs per Mb) on each chromosome using each accessions as the reference are in Figure 3.8. Chromosome 4 and 10 had the highest SNP frequency, chromosome 5 and 9 had the lowest. Also, when using UCR779 as the "reference" (Figures 3.8), the number of SNPs on chromosome 4 and 10 was significantly amplified.

Structural variations were identified by aligning each individual genome against the IT97K-499-35 genome and then visualizing the corresponding pairwise whole-genome alignments with a dot-plot (Figure 3.5 and Figure 3.6). A total of fifteen translocations and inversions larger than 1 Mbp were identified (Table 3.6). In the table, each row indicates a structural variation, with an approximate start and end positions on the chromosome. Two rows are used to describe a translocation, the genomic region in the first set of coordinates is swapped with the genomic region in the second set of coordinates. The inversion and translocation on chromosome 1 between IT97K-499-35 and TZ30 refer to the same regions, which indicates that this inversion is also part of a translocation. Similar situation is observed for inversion and translocation on chromosome 10 between IT97K-499-35 and ZN016. Several large variations appear within the IT97K-499-35 centromere regions. It is noted that that the ∼4.2Mb region on chromosome 3 that was previously reported in [60] occurs in the same orientation in six accessions and in the opposite orientation only in IT97K-499-35. On chromosomes 4, 5 and 7, several inversions within the centromeric regions defined from the IT97K-499-35 assembly are shared by the majority of accessions. The ∼9.0Mb inversion on chromosome 6 is the largest structural variation found and its orientation is private to Suvita2.

Table 3.6: Large structural variations (larger than 1 Mb). Coordinates are in Mb.

chr	type	IT97K	CB5-2	Suvita2	Sanzi	UCR779	ZN016	TZ30	comments
		start end	start end	start end	start end	start end	start end	start end	
1	inversion	7.8 11.8		6.8 10.5			7.0 10.7	9.4 12.5	shared with translocation
1	translocation	8.1 11.5						9.4 12.5	shared with inversion
		11.9 13.9						7.5 9.4	
3	inversion	36.1 40.3	32.4 36.4	31.2 35.0	31.0 35.2	30.8 34.8	32.8 36.8	31.5 35.5	Inversion on IT97K
4	inversion	17.7 21.1	14.8 17.8	14.0 17.0	14.8 15.6	13.9 16.2	14.8 16.2		Overlaps centromeric region
5	inversion	25.7 27.0	23.5 24.8	23.3 24.6		23.0 24.0	24.0 25.3	23.4 24.7	Overlaps centromeric region
6	inversion	0.01 9.0		0.06 8.1					Only in one accession
6	inversion	29.0 30.1					26.3 27.4	26.6 27.6	
7	inversion	14.0 15.0		13.6 14.0					Only in one accession
7	inversion	15.7 16.7	14.8 15.4	14.8 15.1		14.5 15.0	14.5 15.3	14.6 15.1	Overlaps centromeric region
10	inversion	14.5 16.3					13.6 15.6		Shared with translocation
10	inversion	16.5 17.2	13.9 14.7	13.9 14.7	13.3 13.7			13.9 14.7	
10	inversion	17.5 18.7	14.8 15.9	14.8 15.8		15.0 16.0		14.7 15.8	
10	translocation	14.5 16.3					13.6 15.6		Shared with inversion
		17.5 18.8					12.0 13.1		
11	inversion	16.1 16.7					14.8 15.4	14.3 14.9	
11	inversion	30.5 33.6	26.3 29.3						Only in one accession

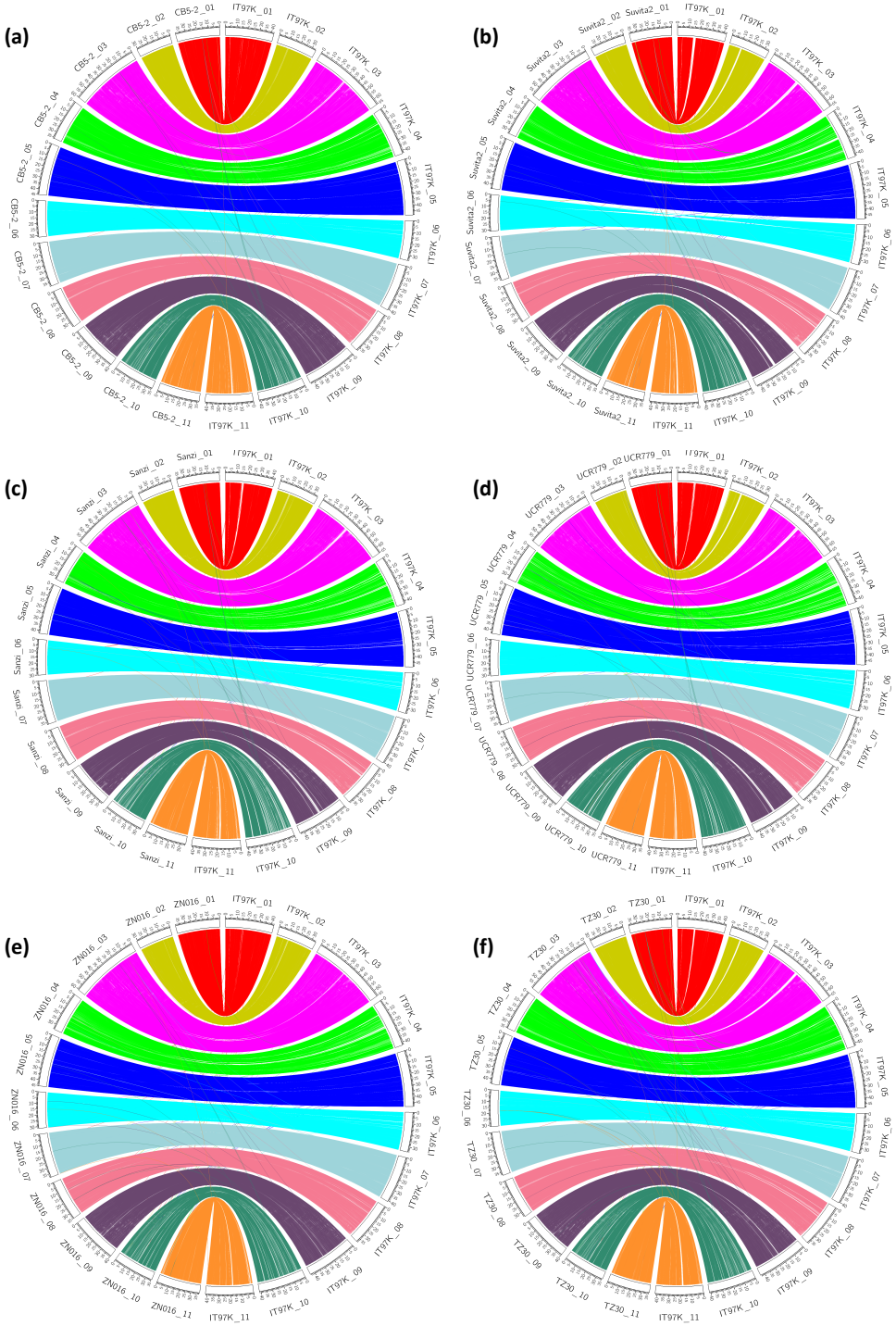

Figure 3.4: Synteny view between IT97K-499-35 and other accessions

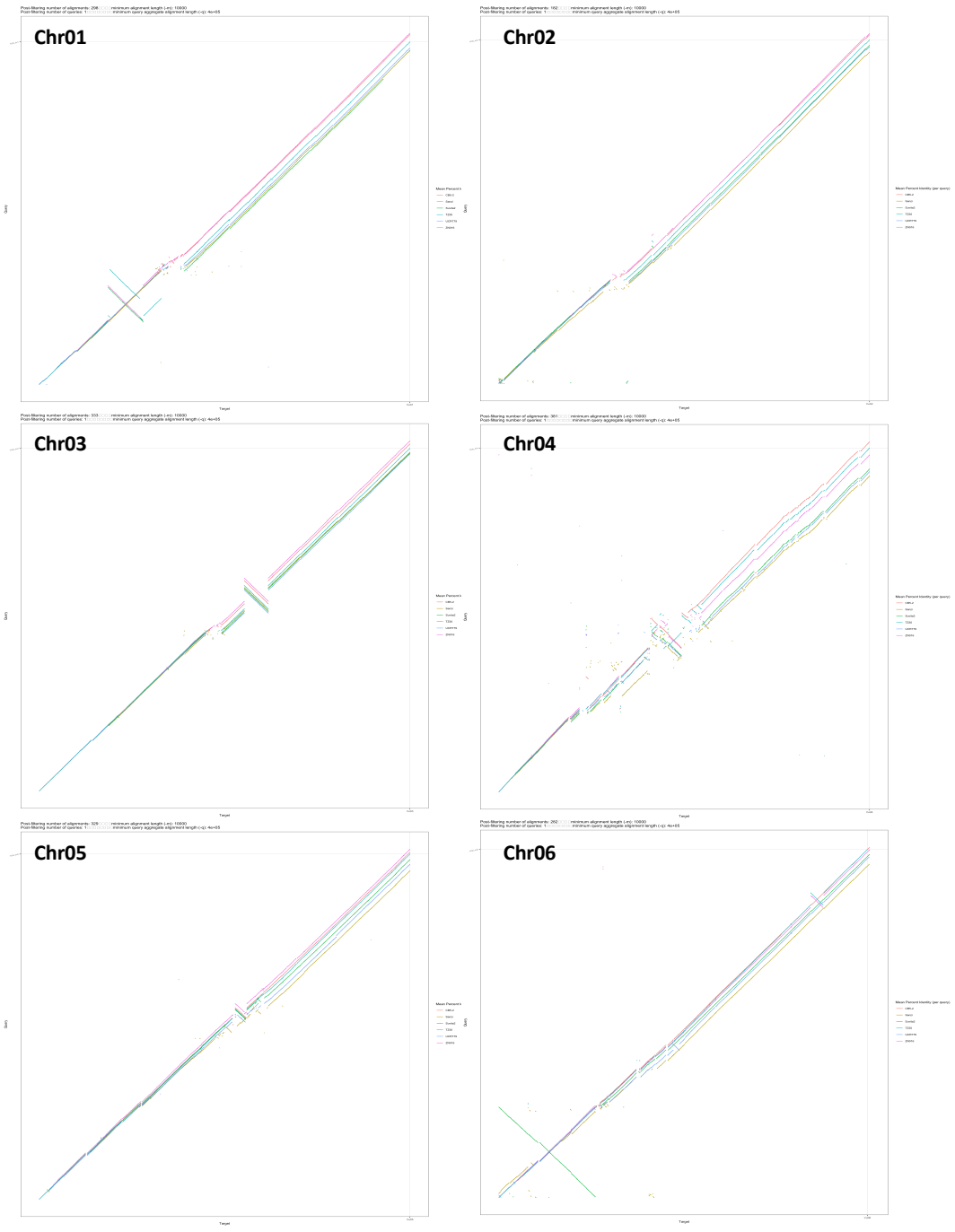

Figure 3.5: Pairwise comparison for chromosomes 1-6

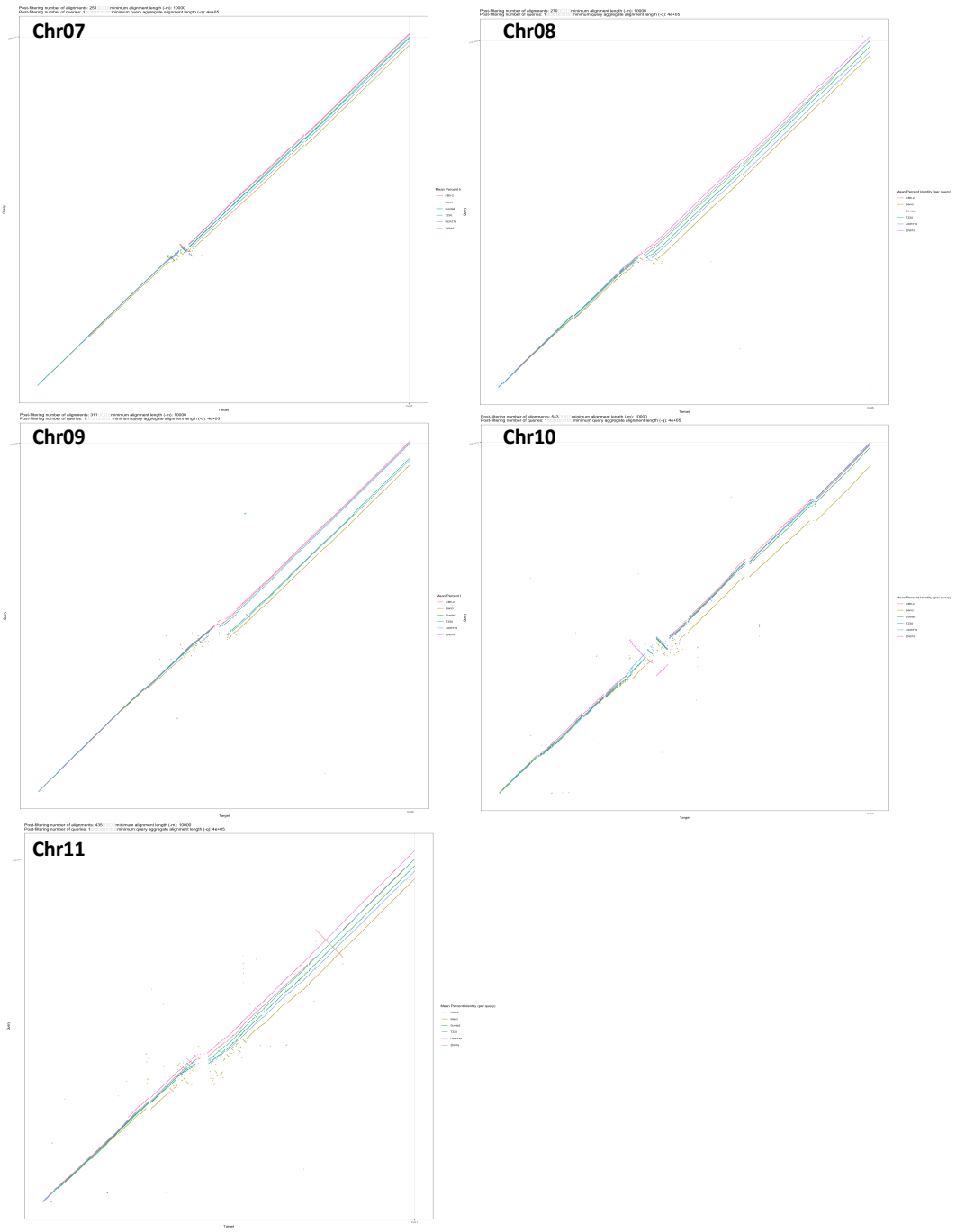

Figure 3.6: Pairwise comparison for chromosomes 7-11

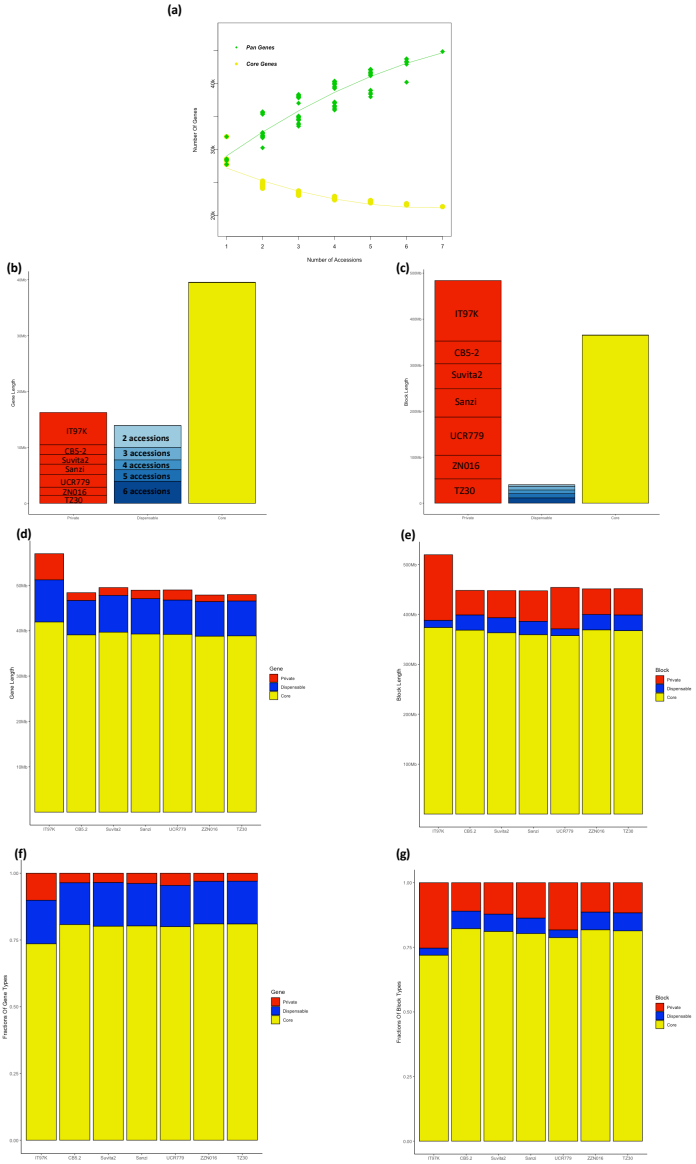

Figure 3.7: Pan-Genome Analyses; (a) number of pan-genes and core genes as a function of the number of accessions analyzed; (b) cumulative length for gene classified as core, dispensable or private; (c) cumulative length for genome blocks classified as core, dispensable or private; (d) cumulative length of core, dispensable or private gene in each individual accession; (e) cumulative length of core, dispensable or private genome blocks in each accession; (f) fraction of the transcript for core, dispensable or private gene in each individual accession; (g) fraction of the genome for core, dispensable or private genomic blocks in each individual accession

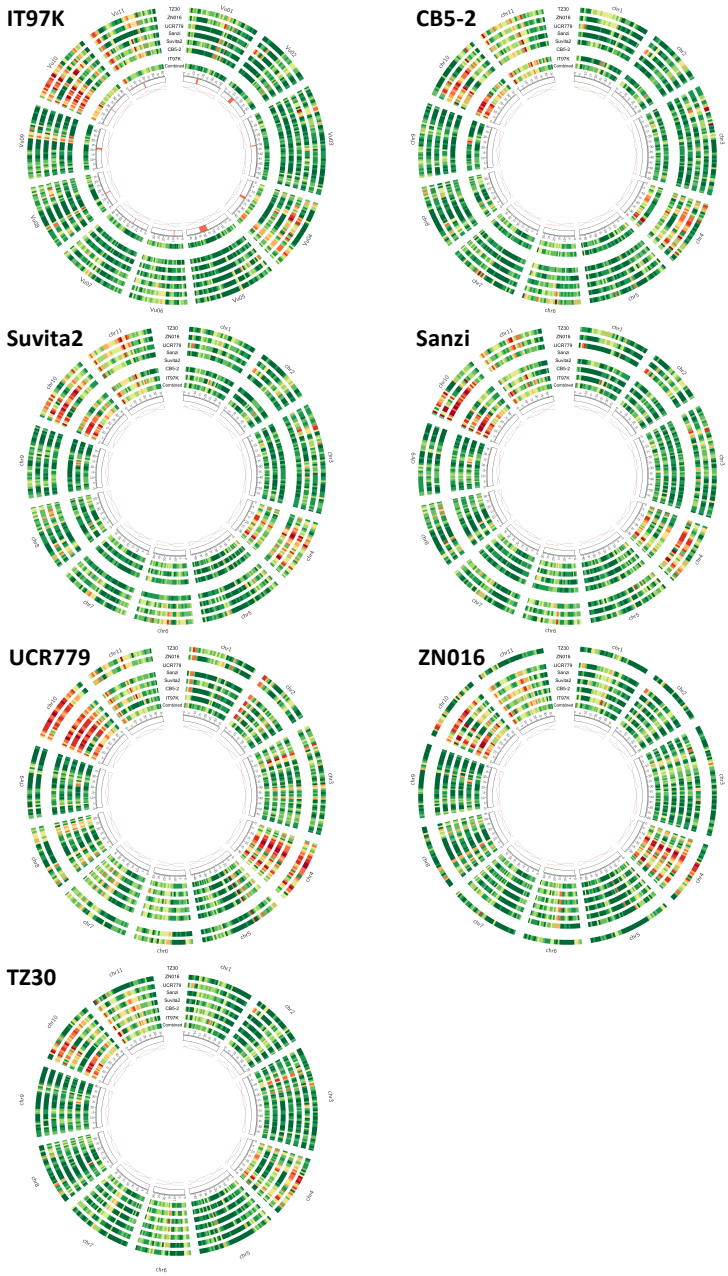

Figure 3.8: SNP density (number of SNPs per Mb) of different accessions

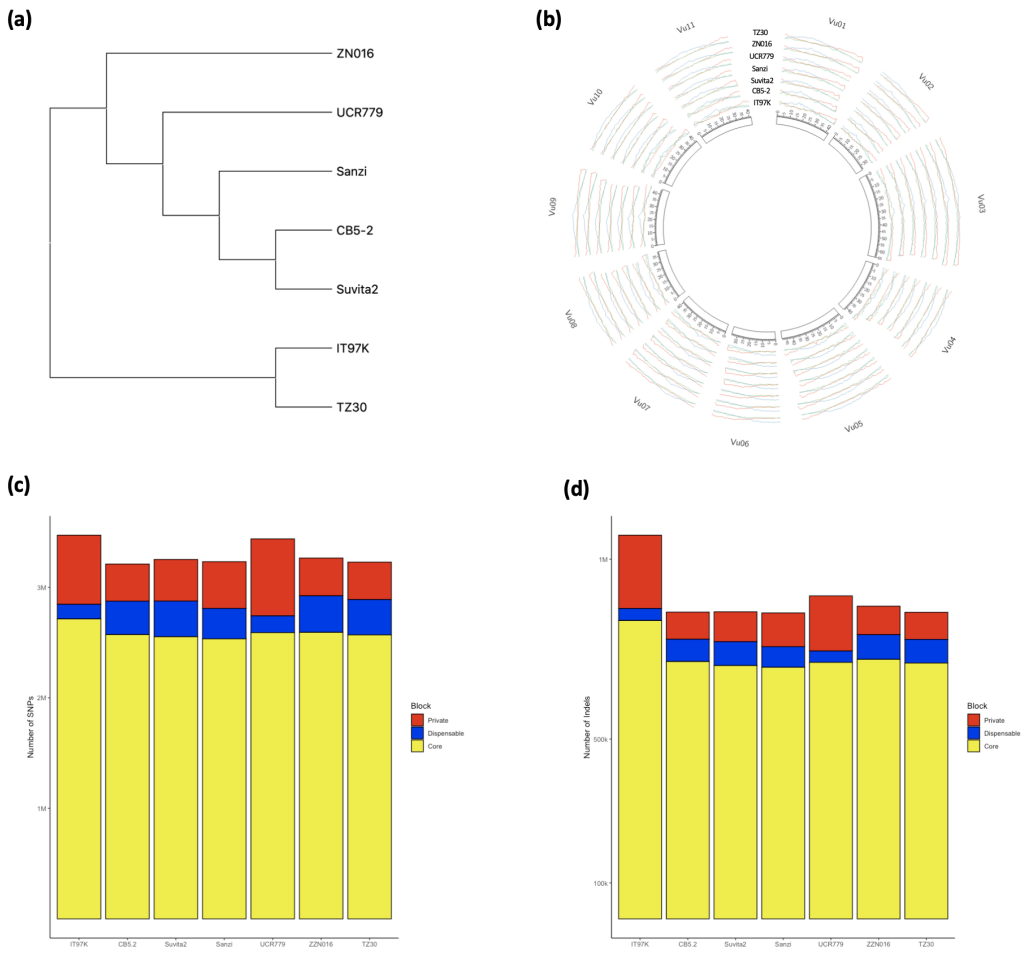

Figure 3.9: (a) phylogenetic tree for the seven accession based on IT97K-499-35 SNPs, (b) circos plot of gene density (red) and repeat density (blue) in 1Mb non-overlapping sliding windows, (c) number of SNPs in private, dispensable and core genomics blocks in each accession, (d) number of indels in private, dispensable and core genomics blocks in each accession

Chapter 4

Pan-genome Representation and Visualization

4.1 Introduction

As more and more individuals (cultivars, accessions, or strains) of a given species are sequenced and made available, the adequacy of the accepted notion of *the reference* genome for a species represented by a single DNA sequence is being challenged [8]. Declaring one individual as *the reference* for a species introduces a representational bias in downstream analyses, including SNP discovery, structural analysis, genome-wide association studies, etc. [8]. Recently, a large number of genomes from different individuals of the same species have become publicly available, and such genomes have significantly increase the possibility of analyzing multiple genomic features. One way to investigate these features is through the pan-genomic approach. The pan-genome captures the entire genetic diversity of a species

	gene based	genome based	alignment method to reference	largest genome that it was tested on	reference
PanX	✓		pairwise alignment using Diamond	microbial genome	[21]
PGAP	✓		all-pairs alignment and BLAST all	microbial genome	[129]
PGAP-X	✓	✓	reference-agnostic alignment using progressiveMauve	microbial genome	[128]
PPanGGOLiN	✓		uses gene families	tens of 1000s of microbial genomes	[26]
PanSeq		✓	pairwise alignment to reference	not mentioned	[46]

Table 4.1: Comparison of pan-genome analysis tools

by cataloging all the structural variants of its genome [19]. As explained in Chapter 1, a pan-genome is composed of (i) the *core* genome containing DNA sequences present in all individuals within the species, (ii) the *dispensable* genome containing DNA sequences present in a subset of the individuals, which includes *unique* individual-specific DNA sequences [110, 68, 28]. An effective representation (and its visualization) of a pan-genome is particularly challenging due to the complex rearrangements that can be observed when comparing multiple genome of the same species [114].

As discussed in Chapter 1, existing methods for pan-genome analysis mainly focus on either small genomes like bacterial genomes or genes only instead of whole genome sequences. Table 4.1 summarizes the main features and limitations for these tools. Most of such available methods requires a pre-selected individual as "reference". In response to such limitations, we propose here a novel pan-genome representation called PGV. The PGV representation is (i) reference-agnostic (i.e., there is no need to artificially declare one of the individual genome to be the reference), (ii) can handle large eukaryotic genomes, and (iii) is very intuitive and simple to understand. The PGV representation can be visualized by a dot-plot or using our genome browser, in which each block is colored depending on whether it is a core, dispensable or unique. Structural variations such as inversions and translocations are highlighted, and shared core/dispensable blocks are linked to illustrate

how the different accessions relate to each other. Users are also allowed to upload annotation tracks (e.g., gene annotations).

4.2 Methods

4.2.1 PGV Pipeline

The input to PGV is a set of n individual genomes for the same species, or a set of genomes from very closely-related species. To obtain the best results, input genomes must have a similar level of assembled contiguity. First, PGV carries out a genome-wide multiple sequence alignment on all the inputs using progressiveMauve [20]. Based on the output of the multiple sequence alignment, PGV classifies each alignment block into three types. A *core* genome block, or C-block, corresponds to an alignment that contains all n individuals. A *dispensable* genome block (also called *accessory*), or D-block, corresponds to an alignment which contains at least two individuals and at most $n - 1$. A *unique* genome block (also called *strain-specific*), or U-block, is a block that belongs exclusively to one individual genome. Please note that in the literature a unique block is a special type of dispensable block, while PGV distinguishes them. Next, PGV converts each individual genome into an ordered sequence of C, D, and U-blocks, each with its corresponding identifier (represented by a unique integer). In the example in Figure 4.3, the alignment of the five input genomes has produced seven common, four dispensable and nine unique blocks.

After the conversion of each genome into blocks, PGV computes the *consensus ordering* for the C-blocks, which will constitute the "back-bone" of the pan-genome. If we only consider C-blocks, observe that each genome can be represented by a permutation

σ of the C-block identifiers $\{1, 2, \ldots, m\}$, where m is the number of C-blocks. Let σ^i be the permutation for the i-th genome, where $i \in [1, n]$. We define the *consensus ordering* of the C-blocks as the ordering σ^* that minimizes the quantity $\sum_{i=1}^{n} L(\sigma^i, \sigma^*)$ where L is the Levenshtein (edit) distance between the permutations. In the literature, the string σ^* is called the *median string* of the set σ^i. The problems of finding the median for a set of strings under the Levenshtein distance is known to be NP-complete [32]. Similar theoretical results have been derived from more complex metrics [109]. A similar notion of consensus ordering for homology blocks was proposed by [75], but their pan-genome is captured by general bidirectional sequence graphs instead of paths.

PGV uses an efficient greedy algorithm to compute an approximation of the optimal ordering σ^*. The algorithm is described in the online Methods, including a detailed example that illustrates it step-by-step. Once the consensus ordering is computed, PGV produces a set of `.bed` tracks (one for each genome, plus the consensus track) that can be visualized off-line or on-line. In the off-line option, PGV generates a dot-plot between the ordering of C-blocks in each genomes and the consensus ordering (Figure 4.4). This option allows users to identify major structural variations in each genome compared to the consensus ordering, and to produce figures to be shared in reports or manuscripts. The on-line option is a genome browser which allows users to visually inspect genome rearrangements (see Figure 4.1). For the browser, PGV allows to generate an alternative type of `.bed` tracks in which gaps are introduced so that C-blocks are aligned vertically (see Figure 4.2). Users can upload in the browser any subset of the `.bed` tracks for individual genomes or the consensus ordering. Each genome is represented as a set of blocks whose sizes are pro-

portional to the underlying sequence length. Light blue blocks are core blocks with same relative ordering and orientation compared to the consensus ordering (e.g., not reversed or translocated); dark blue blocks are core blocks that are translocated compared to the consensus ordering; pink blocks are core blocks that are inverted compared to the consensus ordering. Green blocks are dispensable blocks and red blocks are unique blocks. Tracks can be reordered by clicking on the track names and dragging them with the mouse. The usual navigation tools are available (zoom in/out, pan left/right, select a chromosome, search for a block). Clicking on a block highlights the identifier of that block, namely U for unique, D for dispensable and C for core, followed by a unique ID. Clicking on a D or C-block generates a link that connects corresponding blocks in other genomes (if they are within the current zoom window). The browser also allows users to upload GFF3 containing gene annotations, which are shown as grey blocks.

Figure 4.1 shows a sample screenshot of cowpea pan-genome representation. In this example, the coordinates of each block match their original position in the genome, and thus users can upload annotation tracks to compare genes in core and dispensable blocks. The red lines between different accessions show the linkage information of a block across different genomes. Such linkage information also provides information whether a genomic block is core, dispensable or private and whether it is related to potential structural variations. Figure 4.2 shows the PGV mode in which the pan-genome is visualized so that the core blocks are aligned vertically.

PGV was implemented using Python. The PGV genome browser was implemented using Javascript and HTML. It is a light implementation that can either be used locally or

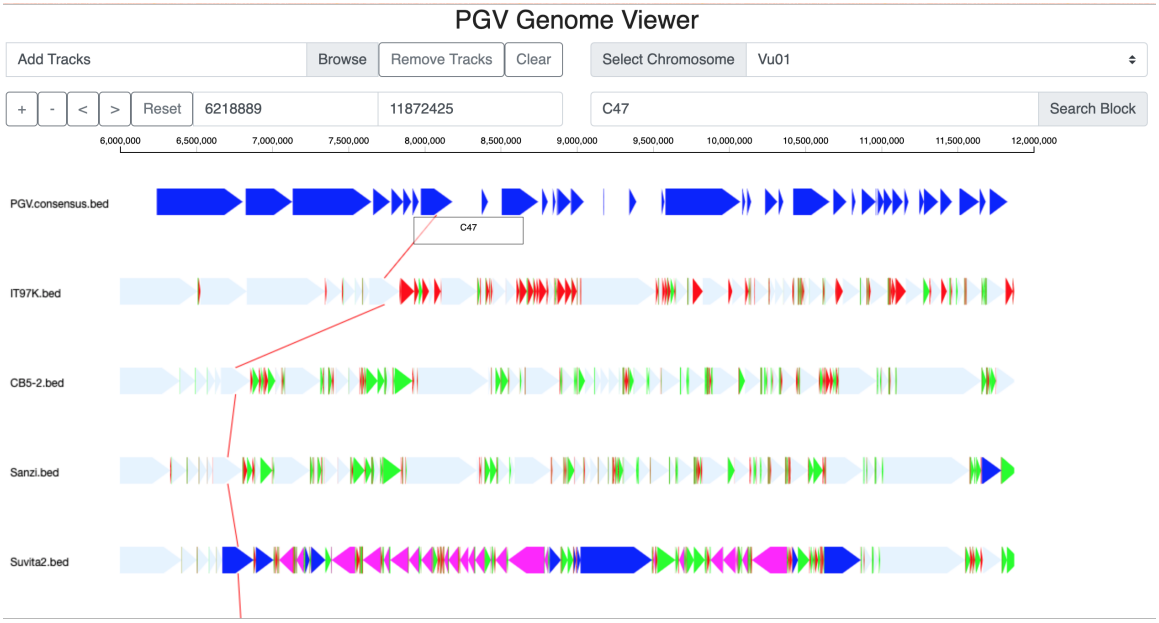

Figure 4.1: A screenshot of the PGV Genome Browser on four cowpea accessions; the first track represents the consensus ordering; IT97K, CB5-2 and Suvita2 and Sanzi are cowpea genomes; light blue blocks are core blocks with same relative ordering and orientation compared to the the consensus ordering; dark blue blocks are core blocks that are translocated compared to the consensus ordering; pink blocks are core blocks that are inverted compared to the consensus ordering; green blocks are dispensable blocks; red blocks are unique blocks.

be vendored by a remote server. The current version is running on Google Firebase. The D3.js library was used for data binding. Canvas was used for plotting the main elements, such as blocks, genes and links. An SVG layer was added to the diagram for plotting axes and tool tips. The Bootstrap library was used for the front-end cosmetics. The genome browser can be accessed at `http://pgv.cs.ucr.edu` The source code for PGV and the genome browser are available at `https://github.com/ucrbioinfo/PGV`. The github page offers some sample data to test the software installation.

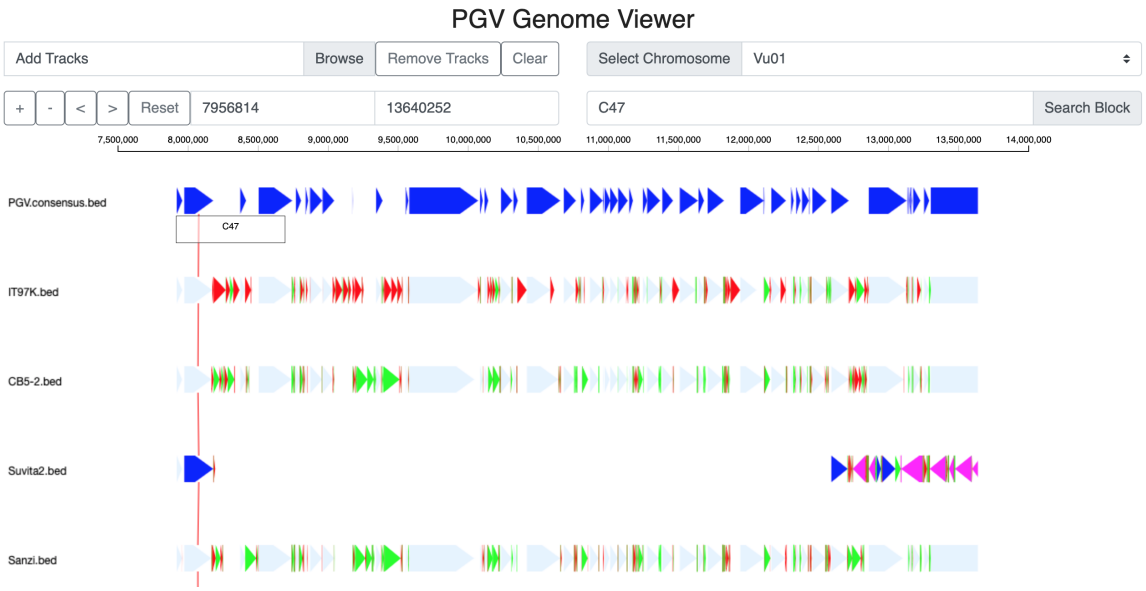

Figure 4.2: A screenshot of the PGV Genome Browser on cowpea accessions using aligned bed tracks; the first track represents the consensus ordering; IT97K, CB5-2 and Suvita2 and Sanzi are cowpea genomes; light blue blocks are core blocks with same relative ordering and orientation compared to the the consensus ordering; dark blue blocks are core blocks that are translocated compared to the consensus ordering; pink blocks are core blocks that are inverted compared to the consensus ordering; green blocks are dispensable blocks; red blocks are unique blocks.

4.2.2 PGV's Consensus Algorithm

While the ideal outcome is to produce a single linear (consensus) ordering for each chromosome, in some situations PanViz can only compute a partial ordering of the C-blocks. For this reason, PanViz maintains a set of linear orderings O, which is initially empty. PanViz starts from an arbitrarily C-block C_i which is added to O as the "seed" of a new linear ordering (or a path). Then, PanViz determines the list C_i's neighbors in the n input genomes and their frequency. Let C_1, C_2 and C_3 be the three neighbors of C_i with the highest frequency, and let f_1, f_2 and f_3 be their frequency. If either C_1, C_2 or C_3 are already in O, they are not considered for the next step. Several cases are possible, (i)

$f_1 \geq f_2 > f_3$, (ii) $f_1 > f_2 = f_3$, (iii) $f_1 = f_2 = f_3$. In case (i), blocks C_1 and C_2 become the candidates neighbors of C_i in the consensus ordering. The consensus ordering is extended as $C_1 \rightarrow C_i \rightarrow C_2$. Then, PanViz repeats the same process on C_1 and C_2, first extending to the left as much as possible, then extending to the right as much as possible. In case (ii), only block C_1 is added to the ordering and the process is repeated from C_1. In case (iii), the current consensus ordering is suspended and a new ordering starts from another arbitrary block that has not be processed yet (i.e., not in O).

Once PanViz has processed all C-blocks, PanViz aligns each path in O to the n genome orderings of the C-blocks to decide its orientation and determine whether it contains mis-joins. Each path and its reversed path are aligned to the original genome (C-block) orderings and an alignment scores is calculated. The alignment score is +1 for an aligned block and -1 for a gap or a mismatch. The local alignment with highest score is used to determine the correct orientation and possible mis-joins. If the alignment score of the reversed path is higher than the score of the alignment for the forward path, the path is reversed. After the orientation is decided, if the overall alignment score is lower than a minimum threshold (i.e., 80% of the highest possible alignment score for that genome length), (i) the path is removed from O, (ii) the path is broken into two or three pieces, namely a central region with the highest alignment, a left overhang (possibly empty), and a right overhang (possibly empty), (iii) the two/three sub-paths are added to O and processed individually through another round of alignments. When all paths are in the correct orientations and have an overall alignments score with the input genomes of at least 80% of the maximum, PanViz obtains the coordinates of each path by taking a majority

vote on their best alignment against the input genomes. PanViz uses these coordinates to order the paths, and produce the final consensus ordering (ideally composed of a single path). In the next section we provide a step-by-step explanation of the algorithm using the example in Figure 4.3(b).

4.2.3 An Example of PGV's Consensus Algorithm

In the example in Figure 4.3(b) we assume that PanViz arbitrarily starts from $C6$, and sets the consensus ordering $O = \{C6\}$. Then, PanViz collects the frequencies for the neighbors of $C6$: $C3$ occurs three times, $C1$ occurs three times, and $C3$ occurs three times. Since there is a tie for the second position, we are in case (ii) discussed above, and only $C6$ is added to O, resulting in $O = \{C6 \rightarrow C2\}$. In Step 2, PanViz collects the frequencies for the neighbors of $C6$. In the top two, only $C5$ is not in O, and thus we extend $O = \{C5 \rightarrow C6 \rightarrow C2\}$. Similarly, in Step 3, 4 and 5, $C5$, $C4$ and $C3$ are appended to the consensus. In Step 6, $C3$ cannot be extended because both its top two neighbors are already in O. Thus PanViz starts a new path by arbitrarily picking $C1$, thus now $O = \{C1, C3 \rightarrow C4 \rightarrow C5 \rightarrow C6 \rightarrow C2\}$. The top two neighbors of C_1 are H and C_2. Since C_2 is in O, only H is appended to C_1. PanViz cannot extend H because H is a chromosome boundary. In Step 8, a new path is created which is extended in Step 9 to the other chromosome boundary. The preliminary set of consensus ordering is thus $O = \{p_1 = C3 \rightarrow C4 \rightarrow C5 \rightarrow C6 \rightarrow C2, p_2 = C1 \rightarrow H, p_3 = C7 \rightarrow T\}$. At this point, PanViz aligns p_1, p_2 and p_3 to the individual genome orderings to decide their orientations. For instance, the alignment score of $H \rightarrow C1$ is higher than the alignment score of $C1 \rightarrow H$,

so p_2 is reversed. Paths p_1 and p_3 are left as is. Then, PanViz checks whether p_1, p_2 and p_3 have a good agreement with the input genomes. For instance, the alignment score of p_1 against the five input genomes is 4, 3, 5, 3, and 3, respectively. The total score for p_1 is 18, which is lower than 80% of the highest possible score, which is 0.8*5*5=20. Based on this, PanViz considers p_1 not to be a good ordering and it breaks it, as follows. The highest scoring sub-path of p_1 is $C3 \rightarrow C4 \rightarrow C5 \rightarrow C6$ on the majority of the input genomes, so PanViz splits p_1 into $p_4 = C3 \rightarrow C4 \rightarrow C5 \rightarrow C6$ and $p_5 = C2$, thus now $O = \{p_2, p_3, p_4, p_5\}$. PanViz again checks the alignments of p_2, p_3, p_4, p_5 in O. If any of them is not sufficiently high (i.e., at least 80% of the maximum score), it will be broken again. Once this iterative process is concluded, each path in O is aligned against the genomes and the starting position of its best alignment is recorded. The position with most votes (majority) determines the coordinate of each path. For instance, the best alignment of p_4 on the input genomes are at position 4,3,4,6 and 5, respectively. Thus, p_4 is given coordinate 4. Similarly, PanViz assigns p_2 position 1, p_3 position 8 and p_5 position 3. Based on these coordinate, PanViz orders the paths as $p_2 \rightarrow p_5 \rightarrow p_4 \rightarrow p_3$ which provides the final consensus ordering $H \rightarrow C_1 \rightarrow C_2 \rightarrow C_3 \rightarrow C_4 \rightarrow C_5 \rightarrow C_6 \rightarrow C7 \rightarrow T$.

4.3 Results

Human. PGV was used on four *Homo sapiens* assemblies, namely GCA_1405.28, GCA_3634875.1, GCA_2180035.3, and GCA_1292825.2. PGV identified 3,548 core blocks (comprising 94.8% of the human genome), 2,390 dispensable blocks and 11,807 unique blocks. Upon inspection of the initial PGV's dot plot we determined that ten chromosomes in GCA_003634875.1

were inverted. Figure 4.4(a) shows the dot-plot after reorienting those chromosomes. The four assemblies show a very high degree of consistency for the core blocks, with very few translocations indicated by the isolated dots.

Arabidopsis. PGV was run on three *Arabidopsis thaliana* assemblies, namely TAIR10.1, Ler, and Ath.Ler-0.MPIPZ. PGV identified 144 core blocks (comprising 96.17% of Arabidopsis genome), 31 dispensable blocks and 352 unique blocks. The higher fraction of the genome in core blocks compared to other species indicated that the three accessions are very closely related. Figure 4.4(b) shows high consistency between the three accessions and the consensus ordering, with very few translocations mostly on chromosome three.

Rice. PGV was used on four *Oryza sativa* assemblies, namely Japonica, Japonic HEG4, Indica, and Aus cultivar. PGV identified 2,632 core blocks (comprising 90.11% of genome), 2,531 dispensable blocks and 12,396 unique blocks. Figure 4.4(c) shows a significant amount of translocations (shown as single dots), and a (centromeric) inversion on chromosome 6 in Indica (orange anti-diagonal in the plot) which was previously reported ([22]).

Cowpea. PGV was run on eight *Vigna unguiculata* genome assemblies namely IT97K [60], CB5-2, Suvita2, Sanzi, UCR779, ZN016, TZ30 and G98. PGV detected 2,863 core blocks (77.41% of the cowpea genome), 11,856 dispensable blocks and 42,484 unique blocks. Figure 4.4(d) shows several inversions (anti-diagonals): (i) two large inversions at the beginning of chromosome 1 and 2 in G98 (further analysis showed that they were mis-assemblies), (ii) one inversion near the center of chromosome 3 of IT97K, which was previously reported by [60], (iii) an inversion shared by Suvita2, ZN016 and TZ30, previously unreported.

4.4 Conclusion

We introduced a representation of the pan-genome based on the notion of consensus ordering, which is reference-agnostic. Experimental results on several species demonstrate the utility of our representation.

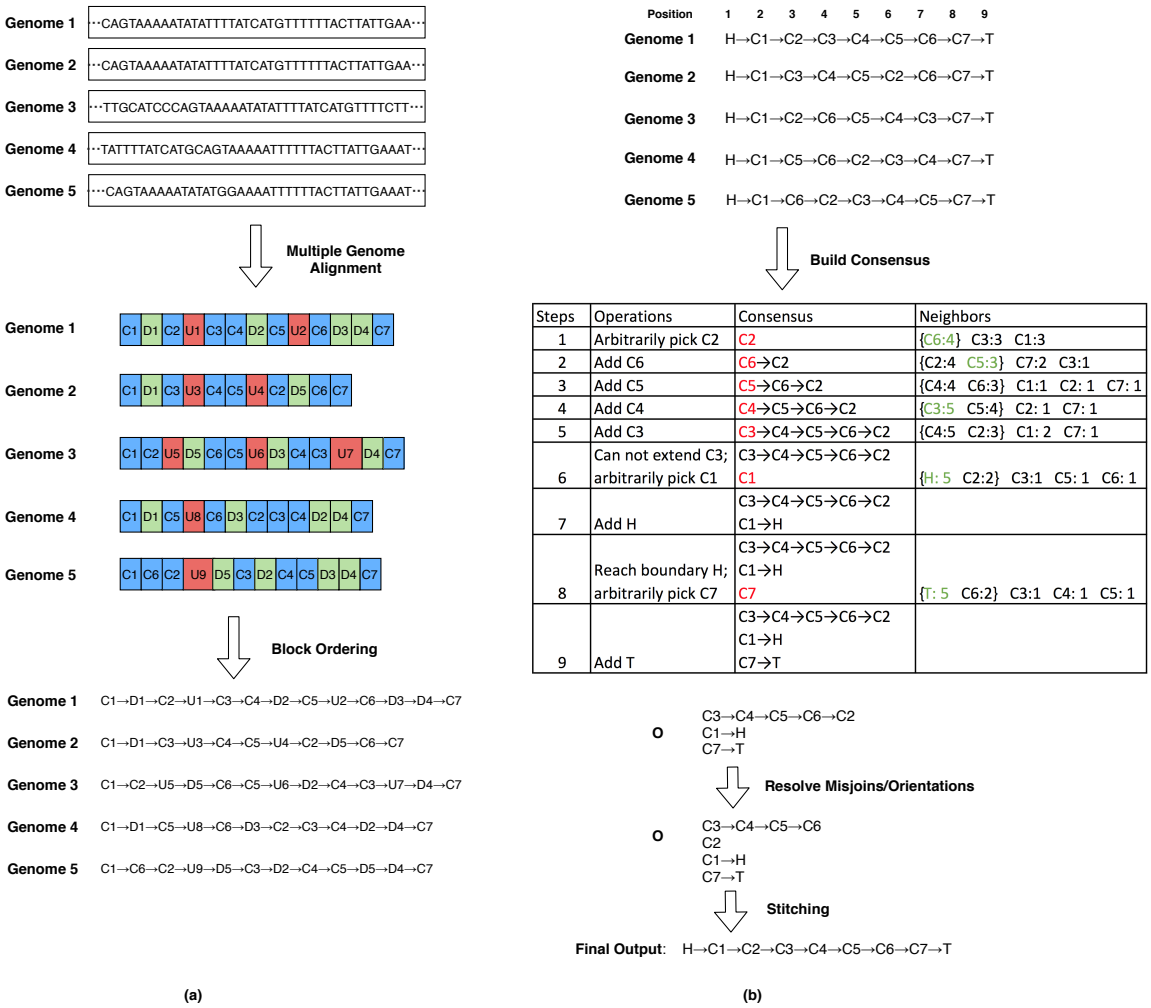

Figure 4.3: A detailed example of PGV's processing steps. (a) the input to PGV is a set of $n = 5$ genomes; PGV first carries out a multiple sequence alignment, then classifies each alignment block into core blocks (C), dispensable block (D) and unique block (U); each genome is then converted in an ordered sequence of C-, D-, and U-blocks, each with its corresponding identifier; (b) in the second phase, PGV computes the consensus ordering of the common blocks; red C-nodes are the active nodes; green C-nodes are the neighbors selected to be added to the linear ordering

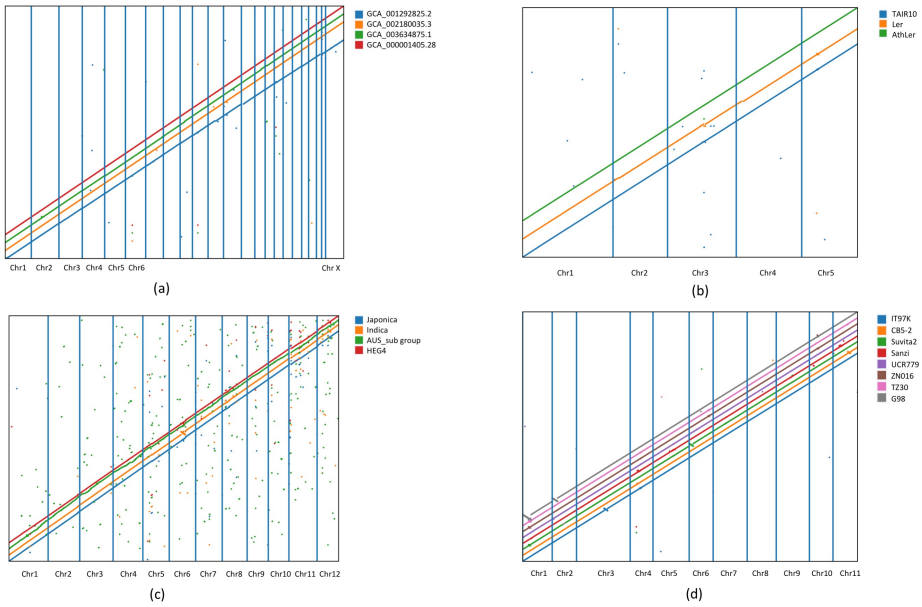

Figure 4.4: Human, arabidopsis, rice, and cowpea pan genome analysis using PGV. The x-axis represents the coordinates of the consensus ordering of core blocks computed by PGV. Genomes coordinates for the core blocks are used on the y-axis (staggered to avoid overlapping lines).

Chapter 5

Conclusions

In this new era of genomic research, more and more genomes for different species (and different individuals within the same species) are becoming available. Due to this availability, genomic analysis increasingly relies on the knowledge accumulated on well-studied species, e.g., model organisms. The principle behind comparative genomics is to exploit the genomic knowledge of evolutionarily related organism that share high similarity at the sequence level with the genome of interest. In most genomic analysis pipeline, several comparative genomic methods are combined in order to provide the most complete and comprehensive results. Comparative genomics analyses have a broad application in genome assembly, genome annotation as well as pan-genome studies. In this dissertation, multiple methods for comparative genomics analysis were used in our analysis pipeline for genome annotations, variants calling and pan-genome analysis and visualization.

Genome annotation is the process of identifying the structures and functions of all functional elements in genome, such as repeats, genes, intron/exons, promoters, enhancers,

etc. The typical annotation pipeline combines prediction tools with pre-trained libraries. In this dissertation, we introduced an improved annotation workflow which combines different predictors that use trained libraries and transcript evidence from RNA-Seq, ESTs and protein databases. This pipeline was applied on genome annotation of three different species, namely *Vigna unguiculata* (cowpea), *Phytophthora infestans* and *Babesia duncani*. The comparative evaluation on these genome assemblies and annotations have provided clear evidence that our assemblies have high-quality (high completeness, high contiguity, etc). During the genome annotation and structural variation analysis of cowpea, interspecific comparisons were applied in order to gain new insights from evolutionarily related legumes such as *Phaseolus vulgaris* (common bean). Such comparisons allowed us: (1) the establishment of a uniform numbering system for cowpea chromosomes, (2) the identification of large chromosomal inversions and translocations in cowpea (one of which was experimentally validated), (3) the identification of repeat and gene family changes among legumes. In the newly sequenced and assembled genome of *Phytophthora infestans*, our gene annotation pipeline was able to detect an additional six thousand genes compared to the previously published genome. In our *de novo* assembly of *Babesia duncani*, the gene annotations were significantly improved compared to the annotations carried out in a previous draft assembly. The annotation pipeline used on these genomes could be adapted and generalized for other eukaryotic genomes with minor modification.

As we have shown in Chapter 3, a pan-genome can be obtained from the comparative genome analysis of a set of individuals genomes within the same species. In our cowpea pan-genome, these comparisons were carried out both at the genome level and at

the gene level in order to provide a comprehensive report of the distribution of present-absent variations. We showed that the size of the core genome in the cowpea pan-genome (composed of seven accessions) approaches a plateau, which indicates a satisfactory completeness of the core genome. Other structural variations were also identified in the cowpea pan-genome, such as large inversions and translocation, and small variations, such as SNPs and indels. A total of fifteen large inversions and translocations with size greater than 1 Mb were reported, including an experimentally validated 4.2 Mb inversion on chromosome 3 for cowpea accession IT97K-449-35.

In order to carry out whole genome comparison for large genomes, a new pan-genome representation and visualization pipeline called PGV was developed. The PGV pipeline is reference-agnostic, which allows one to eliminate the representational bias introduced by arbitrarily using one of the genomes to be the reference. PGV uses a multiple genome alignment for classifying genomic blocks into core, dispensable or private. PGV then computes a consensus ordering of core genomic blocks which provides the "backbone" of the pan-genome. PGV also provides an intuitive linear representation of consensus and the assembled genomes. The interactive web-based visualization companion tool allows users to easily explore the structures and the variations across the pan-genome.

In summary, this thesis combined multiple comparative genomics methods for the genome annotation of three different species, the pan-genome construction of cowpea and a novel pan-genome representation and visualization. These analytical methods have provided new biological insights for these species and can serve as a general guidance for genome annotation, variation analysis and pan-genome studies in other eukaryotic organisms.

Bibliography

[1] A brief guide to genomics.

[2] Josep F Abril and S Castellano Hereza. Genome annotation. Elsevier, 2019.

[3] VD Aggarwal, N Muleba, I Drabo, J Souma, and M Mbewe. Inheritance of striga gesnerioides resistance in cowpea. In *Proceedings of the 3rd International Symposium on Parasitic Weeds, Aleppo, Syria*, pages 7–11, 1984.

[4] Rolf Apweiler, Terri K Attwood, Amos Bairoch, Alex Bateman, Ewan Birney, Margaret Biswas, Philipp Bucher, Lorenzo Cerutti, Florence Corpet, Michael DR Croning, et al. The interpro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic acids research*, 29(1):37–40, 2001.

[5] Simon Ardui, Adam Ameur, Joris R Vermeesch, and Matthew S Hestand. Single molecule real-time (smrt) sequencing comes of age: applications and utilities for medical diagnostics. *Nucleic acids research*, 46(5):2159–2168, 2018.

[6] Terri K Attwood, Michael DR Croning, Darren R Flower, AP Lewis, JE Mabey, Philip Scordis, JN Selley, and W Wright. Prints-s: the database formerly known as prints. *Nucleic Acids Research*, 28(1):225–227, 2000.

[7] Amos Bairoch and Rolf Apweiler. The swiss-prot protein sequence database and its supplement trembl in 2000. *Nucleic acids research*, 28(1):45–48, 2000.

[8] Sara Ballouz, Alexander Dobin, and Jesse A. Gillis. Is it time to change the reference genome? *Genome Biology*, 20(1):159, 2019.

[9] Alex Bateman, Ewan Birney, Richard Durbin, Sean R Eddy, Kevin L Howe, and Erik LL Sonnhammer. The pfam protein families database. *Nucleic acids research*, 28(1):263–266, 2000.

[10] Philipp E Bayer, Agnieszka A Golicz, Armin Scheben, Jacqueline Batley, and David Edwards. Plant pan-genomes are the new reference. *Nat. Plants*, 6:914–920, 2020.

[11] Jaime E Blair, Michael D Coffey, Sook-Young Park, David M Geiser, and Seogchan Kang. A multi-locus phylogeny for phytophthora utilizing markers derived from complete genome sequences. *Fungal Genetics and Biology*, 45(3):266–277, 2008.

[12] Ousmane Boukar, Nouhoun Belko, Siva Chamarthi, Abou Togola, Joseph Batieno, Emmanuel Owusu, Mohammed Haruna, Sory Diallo, Muhammed Lawan Umar, Olusoji Olufajo, et al. Cowpea (vigna unguiculata): Genetics, genomics and breeding. *Plant Breeding*, 138(4):415–424, 2019.

[13] Michael S Campbell, Carson Holt, Barry Moore, and Mark Yandell. Genome annotation and curation using maker and maker-p. *Current protocols in bioinformatics*, 48(1):4–11, 2014.

[14] Brandi L Cantarel, Ian Korf, Sofia MC Robb, Genis Parra, Eric Ross, Barry Moore, Carson Holt, Alejandro Sánchez Alvarado, and Mark Yandell. Maker: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome research*, 18(1):188–196, 2008.

[15] Márcia Carvalho, Teresa Lino-Neto, Eduardo Rosa, and Valdemar Carnide. Cowpea: a legume crop for a challenging environment. *Journal of the Science of Food and Agriculture*, 97(13):4273–4284, 2017.

[16] Jarrod A Chapman, Isaac Ho, Sirisha Sunkara, Shujun Luo, Gary P Schroth, and Daniel S Rokhsar. Meraculous: de novo genome assembly with short paired-end reads. *PloS one*, 6(8):e23501, 2011.

[17] Chen-Shan Chin, Paul Peluso, Fritz J Sedlazeck, Maria Nattestad, Gregory T Concepcion, Alicia Clum, Christopher Dunn, Ronan O'Malley, Rosa Figueroa-Balderas, Abraham Morales-Cruz, et al. Phased diploid genome assembly with single-molecule real-time sequencing. *Nature methods*, 13(12):1050–1054, 2016.

[18] Alan Cleary and Andrew Farmer. Genome Context Viewer: visual exploration of multiple annotated genomes using microsynteny. *Bioinformatics*, 34(9):1562–1564, 11 2017.

[19] Computational Pan-Genomics Consortium. Computational pan-genomics: status, promises and challenges. *Brief. Bioinform.*, 19(1):118–135, January 2018.

[20] Aaron E. Darling, Bob Mau, and Nicole T. Perna. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PLOS ONE*, 5(6):1–17, 06 2010.

[21] Wei Ding, Franz Baumdicker, and Richard A Neher. panX: pan-genome analysis and exploration. *Nucleic Acids Res.*, 46(1):e5, January 2018.

[22] Huilong Du, Ying Yu, Yanfei Ma, Qiang Gao, Yinghao Cao, Zhuo Chen, Bin Ma, Ming Qi, Yan Li, Xianfeng Zhao, Jing Wang, Kunfan Liu, Peng Qin, Xin Yang, Lihuang Zhu, Shigui Li, and Chengzhi Liang. Sequencing and de novo assembly of a near complete indica rice genome. *Nat. Commun.*, 8:15324, May 2017.

[23] Robert D Fleischmann, Mark D Adams, Owen White, Rebecca A Clayton, Ewen F Kirkness, Anthony R Kerlavage, Carol J Bult, Jean-Francois Tomb, Brian A Dougherty, Joseph M Merrick, et al. Whole-genome random sequencing and assembly of haemophilus influenzae rd. *Science*, 269(5223):496–512, 1995.

[24] Jullien M Flynn, Robert Hubley, Clément Goubert, Jeb Rosen, Andrew G Clark, Cédric Feschotte, and Arian F Smit. Repeatmodeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences*, 117(17):9451–9457, 2020.

[25] Lei Gao, Itay Gonda, Honghe Sun, Qiyue Ma, Kan Bao, Denise M Tieman, Elizabeth A Burzynski-Chang, Tara L Fish, Kaitlin A Stromberg, Gavin L Sacks, et al. The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. *Nature genetics*, 51(6):1044–1051, 2019.

[26] Guillaume Gautreau, Adelme Bazin, Mathieu Gachet, Rémi Planel, Laura Burlot, Mathieu Dubois, Amandine Perrin, Claudine Médigue, Alexandra Calteau, Stéphane Cruveiller, Catherine Matias, Christophe Ambroise, Eduardo P. C. Rocha, and David Vallenet. PPanGGOLiN: depicting microbial diversity via a partitioned pangenome graph. *PLOS Computational Biology*, 16(3):1–27, 03 2020.

[27] Manfred G Grabherr, Brian J Haas, Moran Yassour, Joshua Z Levin, Dawn A Thompson, Ido Amit, Xian Adiconis, Lin Fan, Raktima Raychowdhury, Qiandong Zeng, et al. Full-length transcriptome assembly from rna-seq data without a reference genome. *Nature biotechnology*, 29(7):644–652, 2011.

[28] Luis Carlos Guimarães, Jolanta Florczak-Wyspianska, Leandro Benevides de Jesus, Marcus Vinícius Canário Viana, Artur Silva, Rommel Thiago Jucá Ramos, Siomar de Castro Soares, and Siomar de Castro Soares. Inside the pan-genome - methods and software overview. *Curr. Genomics*, 16(4):245–252, August 2015.

[29] Brian J Haas, Arthur L Delcher, Stephen M Mount, Jennifer R Wortman, Roger K Smith Jr, Linda I Hannick, Rama Maiti, Catherine M Ronning, Douglas B Rusch, Christopher D Town, et al. Improving the arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic acids research*, 31(19):5654–5666, 2003.

[30] Brian J Haas, Sophien Kamoun, Michael C Zody, Rays HY Jiang, Robert E Handsaker, Liliana M Cano, Manfred Grabherr, Chinnappa D Kodira, Sylvain Raffaele, Trudy Torto-Alalibo, et al. Genome sequence and analysis of the irish potato famine pathogen phytophthora infestans. *Nature*, 461(7262):393–398, 2009.

[31] Brian J Haas, Steven L Salzberg, Wei Zhu, Mihaela Pertea, Jonathan E Allen, Joshua Orvis, Owen White, C Robin Buell, and Jennifer R Wortman. Automated eukaryotic gene structure annotation using evidencemodeler and the program to assemble spliced alignments. *Genome biology*, 9(1):R7, 2008.

[32] Morihiro Hayashida and Hitoshi Koyano. Finding median and center strings for a probability distribution on a set of strings under levenshtein distance based on integer linear programming. In Ana Fred and Hugo Gamboa, editors, *Biomedical Engineering Systems and Technologies*, pages 108–121, Cham, 2017. Springer International Publishing.

[33] Daniel G Hert, Christopher P Fredlake, and Annelise E Barron. Advantages and limitations of next-generation sequencing technologies: a comparison of electrophoresis and non-electrophoresis methods. *Electrophoresis*, 29(23):4618–4626, December 2008.

[34] Nicolas Hulo, Amos Bairoch, Virginie Bulliard, Lorenzo Cerutti, Edouard De Castro, Petra S Langendijk-Genevaux, Marco Pagni, and Christian JA Sigrist. The prosite database. *Nucleic acids research*, 34(suppl_1):D227–D230, 2006.

[35] Bao-Lam Huynh, Jeffrey D Ehlers, Bevan Emma Huang, María Muñoz-Amatriaín, Stefano Lonardi, Jansen RP Santos, Arsenio Ndeve, Benoit J Batieno, Ousmane Boukar, Ndiaga Cisse, et al. A multi-parent advanced generation inter-cross (magic) population for genetic analysis and improvement of cowpea (vigna unguiculata l. walp.). *The Plant Journal*, 93(6):1129–1142, 2018.

[36] Broad Institute. Picard tools. `http://broadinstitute.github.io/picard/`, 2018. Accessed: 2018-MM-DD; version X.Y.Z.

[37] Aiko Iwata-Otsubo, Jer-Young Lin, Navdeep Gill, and Scott A Jackson. Highly distinct chromosomal structures in cowpea (vigna unguiculata), as revealed by molecular cytogenetic analysis. *Chromosome Research*, 24(2):197–216, 2016.

[38] Aiko Iwata-Otsubo, Brittany Radke, Seth Findley, Brian Abernathy, C Eduardo Vallejos, and Scott A Jackson. Fluorescence in situ hybridization (fish)-based karyotyping reveals rapid evolution of centromeric and subtelomeric repeats in common bean (phaseolus vulgaris) and relatives. *G3: Genes, Genomes, Genetics*, 6(4):1013–1022, 2016.

[39] Sophien Kamoun. A catalogue of the effector secretome of plant pathogenic oomycetes. *Annual review of phytopathology*, 44, 2006.

[40] Yang Jae Kang, Sue K Kim, Moon Young Kim, Puji Lestari, Kil Hyun Kim, Bo-Keun Ha, Tae Hwan Jun, Won Joo Hwang, Taeyoung Lee, Jayern Lee, et al. Genome sequence of mungbean and insights into evolution within vigna species. *Nature communications*, 5:5443, 2014.

[41] W James Kent. Blat—the blast-like alignment tool. *Genome research*, 12(4):656–664, 2002.

[42] Sergey Koren, Brian P Walenz, Konstantin Berlin, Jason R Miller, Nicholas H Bergman, and Adam M Phillippy. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome research*, 27(5):722–736, 2017.

[43] Ian Korf. Gene finding in novel genomes. *BMC bioinformatics*, 5(1):59, 2004.

[44] Martin Krzywinski, Jacqueline Schein, Inanc Birol, Joseph Connors, Randy Gascoyne, Doug Horsman, Steven J Jones, and Marco A Marra. Circos: an information aesthetic for comparative genomics. *Genome research*, 19(9):1639–1645, 2009.

[45] Stefan Kurtz, Adam Phillippy, Arthur L Delcher, Michael Smoot, Martin Shumway, Corina Antonescu, and Steven L Salzberg. Versatile and open software for comparing large genomes. *Genome biology*, 5(2):R12, 2004.

[46] Chad Laing, Cody Buchanan, Eduardo N Taboada, Yongxiang Zhang, Andrew Kropinski, Andre Villegas, James E Thomas, and Victor P J Gannon. Pan-genome sequence analysis using Panseq: an online tool for the rapid analysis of core and accessory genomic regions. *BMC Bioinformatics*, 11:461, September 2010.

[47] Philippe Lamesch, Tanya Z Berardini, Donghui Li, David Swarbreck, Christopher Wilks, Rajkumar Sasidharan, Robert Muller, Kate Dreher, Debbie L Alexander, Margarita Garcia-Hernandez, et al. The arabidopsis information resource (tair): improved gene annotation and new tools. *Nucleic acids research*, 40(D1):D1202–D1210, 2012.

[48] Heng Li. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. *Bioinformatics*, 32(14):2103–2110, 2016.

[49] Heng Li. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 34(18):3094–3100, 2018.

[50] Heng Li and Richard Durbin. Fast and accurate long-read alignment with burrows–wheeler transform. *Bioinformatics*, 26(5):589–595, 2010.

[51] Li Li, Christian J Stoeckert, and David S Roos. Orthomcl: identification of ortholog groups for eukaryotic genomes. *Genome research*, 13(9):2178–2189, 2003.

[52] Ruiqiang Li, Hongmei Zhu, Jue Ruan, Wubin Qian, Xiaodong Fang, Zhongbin Shi, Yingrui Li, Shengting Li, Gao Shan, Karsten Kristiansen, et al. De novo assembly of human genomes with massively parallel short read sequencing. *Genome research*, 20(2):265–272, 2010.

[53] Ying-hui Li, Guangyu Zhou, Jianxin Ma, Wenkai Jiang, Long-guo Jin, Zhouhao Zhang, Yong Guo, Jinbo Zhang, Yi Sui, Liangtao Zheng, et al. De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. *Nature biotechnology*, 32(10):1045–1052, 2014.

[54] Zhenyu Li, Yanxiang Chen, Desheng Mu, Jianying Yuan, Yujian Shi, Hao Zhang, Jun Gan, Nan Li, Xuesong Hu, Binghang Liu, et al. Comparison of the two major classes of assembly algorithms: overlap–layout–consensus and de-bruijn-graph. *Briefings in functional genomics*, 11(1):25–37, 2012.

[55] Yu Lin, Jeffrey Yuan, Mikhail Kolmogorov, Max W Shen, Mark Chaisson, and Pavel A Pevzner. Assembly of long error-prone reads using de bruijn graphs. *Proceedings of the National Academy of Sciences*, 113(52):E8396–E8405, 2016.

[56] Yucheng Liu, Huilong Du, Pengcheng Li, Yanting Shen, Hua Peng, Shulin Liu, Guo-An Zhou, Haikuan Zhang, Zhi Liu, Miao Shi, et al. Pan-genome of wild and cultivated soybeans. *Cell*, 182(1):162–176, 2020.

[57] Sassoum Lo, María Muñoz-Amatriaín, Ousmane Boukar, Ira Herniter, Ndiaga Cisse, Yi-Ning Guo, Philip A Roberts, Shizhong Xu, Christian Fatokun, and Timothy J Close. Identification of qtl controlling domestication-related traits in cowpea (vigna unguiculata l. walp). *Scientific reports*, 8(1):1–9, 2018.

[58] Nicholas J Loman, Joshua Quick, and Jared T Simpson. A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nature methods*, 12(8):733–735, 2015.

[59] Alexandre Lomsadze, Vardges Ter-Hovhannisyan, Yury O Chernoff, and Mark Borodovsky. Gene identification in novel eukaryotic genomes by self-training algorithm. *Nucleic acids research*, 33(20):6494–6506, 2005.

[60] Stefano Lonardi, María Muñoz-Amatriaín, Qihua Liang, Shengqiang Shu, Steve I Wanamaker, Sassoum Lo, Jaakko Tanskanen, Alan H Schulman, Tingting Zhu, Ming-Cheng Luo, Hind Alhakami, Rachid Ounit, Abid Md Hasan, Jerome Verdier, Philip A Roberts, Jansen R P Santos, Arsenio Ndeve, Jaroslav Doležel, Jan Vrána, Samuel A Hokin, Andrew D Farmer, Steven B Cannon, and Timothy J Close. The genome of cowpea (Vigna unguiculata [L.] Walp.). *Plant J.*, 98(5):767–782, June 2019.

[61] Todd M Lowe and Sean R Eddy. trnascan-se: a program for improved detection of transfer rna genes in genomic sequence. *Nucleic acids research*, 25(5):955–964, 1997.

[62] William Wylie Mackie et al. Blackeye beans in california. 1946.

[63] William H Majoros, Mihaela Pertea, and Steven L Salzberg. Tigrscan and glimmerhmm: two open source ab initio eukaryotic gene-finders. *Bioinformatics*, 20(16):2878–2879, 2004.

[64] Choukri B Mamoun and David R Allred. Babesiosis. *eLS*, pages 1–8, 2018.

[65] Elaine R Mardis. Next-generation DNA sequencing methods. *Annu. Rev. Genomics Hum. Genet.*, 9:387–402, 2008.

[66] Florian Maumus and Hadi Quesneville. Deep investigation of arabidopsis thaliana junk dna reveals a continuum between repetitive elements and genomic dark matter. *PLoS One*, 9(4):e94101, 2014.

[67] Aaron McKenna, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, David Altshuler, Stacey Gabriel, Mark Daly, et al. The genome analysis toolkit: a mapreduce framework for analyzing next-generation dna sequencing data. *Genome research*, 20(9):1297–1303, 2010.

[68] Duccio Medini, Claudio Donati, Hervé Tettelin, Vega Masignani, and Rino Rappuoli. The microbial pan-genome. *Curr. Opin. Genet. Dev.*, 15(6):589–594, December 2005.

[69] Michael L Metzker. Emerging technologies in DNA sequencing. *Genome Res.*, 15(12):1767–1776, December 2005.

[70] Michael L Metzker. Sequencing technologies - the next generation. *Nat. Rev. Genet.*, 11(1):31–46, January 2010.

[71] Vikram A Misra, Yu Wang, and Michael P Timko. A compendium of transcription factor and transcriptionally active protein coding gene families in cowpea (vigna unguiculata l.). *BMC genomics*, 18(1):1–24, 2017.

[72] Michele Morgante, Emanuele De Paoli, and Slobodanka Radovic. Transposable elements and the plant pan-genomes. *Current opinion in plant biology*, 10(2):149–155, 2007.

[73] Wellington Muchero, Ndeye N Diop, Prasanna R Bhat, Raymond D Fenton, Steve Wanamaker, Marti Pottorff, Sarah Hearne, Ndiaga Cisse, Christian Fatokun, Jeffrey D Ehlers, et al. A consensus genetic map of cowpea [vigna unguiculata (l) walp.] and synteny based on est-derived snps. *Proceedings of the national academy of sciences*, 106(43):18159–18164, 2009.

[74] María Muñoz-Amatriaín, Hamid Mirebrahim, Pei Xu, Steve I Wanamaker, MingCheng Luo, Hind Alhakami, Matthew Alpert, Ibrahim Atokple, Benoit J Batieno, Ousmane Boukar, et al. Genome resources for climate-resilient cowpea, an essential crop for food security. *The Plant Journal*, 89(5):1042–1054, 2017.

[75] Ngan Nguyen, Glenn Hickey, Daniel R Zerbino, Brian Raney, Dent Earl, Joel Armstrong, W James Kent, David Haussler, and Benedict Paten. Building a pan-genome reference for a population. *Journal of Computational Biology*, 22(5):387–401, 05 2015.

[76] Gene Ontology. tool for the unification of biology. the gene ontology consortium. *Nature Genet*, 25(1):25–29, 2000.

[77] RepeatMasker Open. 4.0 [http://www. repeatmasker. org], 2015.

[78] Shu Ouyang, Wei Zhu, John Hamilton, Haining Lin, Matthew Campbell, Kevin Childs, Françoise Thibaud-Nissen, Renae L Malek, Yuandan Lee, Li Zheng, et al. The tigr rice genome annotation resource: improvements and new features. *Nucleic acids research*, 35(suppl_1):D883–D887, 2007.

[79] Jon Palmer and Jason Stajich. nextgenusfs/funannotate: funannotate v1.5.3. Mar 2019.

[80] Yan Pantoja, Kenny Pinheiro, Allan Veras, Fabrício Araújo, Ailton Lopes de Sousa, Luis Carlos Guimarães, Artur Silva, and Rommel T J Ramos. PanWeb: A web interface for pan-genomic analysis. *PLoS One*, 12(5):e0178154, May 2017.

[81] Genis Parra, Keith Bradnam, and Ian Korf. Cegma: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics*, 23(9):1061–1067, 2007.

[82] Benedict Paten, Mark Diekhans, Dent Earl, John St. John, Jian Ma, Bernard Suh, and David Haussler. Cactus graphs for genome comparisons. *Journal of Computational Biology*, 18(3):469–481, 2011. PMID: 21385048.

[83] Benedict Paten, Dent Earl, Ngan Nguyen, Mark Diekhans, Daniel Zerbino, and David Haussler. Cactus: Algorithms for genome multiple sequence alignment. *Genome Research*, 21(9):1512–1528, 2011.

[84] Mihaela Pertea, Geo M Pertea, Corina M Antonescu, Tsung-Cheng Chang, Joshua T Mendell, and Steven L Salzberg. Stringtie enables improved reconstruction of a transcriptome from rna-seq reads. *Nature biotechnology*, 33(3):290–295, 2015.

[85] Rodrigo Pracana, Ilya Levantis, Carlos Martínez-Ruiz, Eckart Stolle, Anurag Priyam, and Yannick Wurm. Fire ant social chromosomes: Differences in number, sequence and expression of odorant binding proteins. *Evolution letters*, 1(4):199–210, 2017.

[86] Eli Rodgers-Melnick, Daniel L Vera, Hank W Bass, and Edward S Buckler. Open chromatin reveals the functional maize genome. *Proceedings of the National Academy of Sciences*, 113(22):E3177–E3184, 2016.

[87] Jue Ruan and Heng Li. Fast and accurate long-read assembly with wtdbg2. *Nature methods*, 17(2):155–158, 2020.

[88] Jansen Rodrigo Pereira Santos, Arsenio Daniel Ndeve, Bao-Lam Huynh, William Charles Matthews, and Philip Alan Roberts. Qtl mapping and transcriptome analysis of cowpea reveals candidate genes for root-knot nematode resistance. *PloS one*, 13(1):e0189185, 2018.

[89] Jeremy Schmutz, Steven B Cannon, Jessica Schlueter, Jianxin Ma, Therese Mitros, William Nelson, David L Hyten, Qijian Song, Jay J Thelen, Jianlin Cheng, et al. Genome sequence of the palaeopolyploid soybean. *nature*, 463(7278):178–183, 2010.

[90] Jeremy Schmutz, Phillip E McClean, Sujan Mamidi, G Albert Wu, Steven B Cannon, Jane Grimwood, Jerry Jenkins, Shengqiang Shu, Qijian Song, Carolina Chavarro, et al. A reference genome for common bean and genome-wide analysis of dual domestications. *Nature genetics*, 46(7):707–713, 2014.

[91] Fritz J Sedlazeck, Hayan Lee, Charlotte A Darby, and Michael C Schatz. Piercing the dark matter: bioinformatics of long-range sequencing and mapping. *Nature Reviews Genetics*, 19(6):329–346, 2018.

[92] Mathieu Seppey, Mosè Manni, and Evgeny M Zdobnov. Busco: assessing genome assembly and annotation completeness. In *Gene Prediction*, pages 227–245. Springer, 2019.

[93] Rachel M Sherman, Juliet Forman, Valentin Antonescu, Daniela Puiu, Michelle Daya, Nicholas Rafaels, Meher Preethi Boorgula, Sameer Chavan, Candelaria Vergara, Victor E Ortega, et al. Assembly of a pan-genome from deep sequencing of 910 humans of african descent. *Nature genetics*, 51(1):30–35, 2019.

[94] Rachel M Sherman and Steven L Salzberg. Pan-genomics in the human genome era. *Nature Reviews Genetics*, pages 1–12, 2020.

[95] Mikhail M Shneider, Sergey A Buth, Brian T Ho, Marek Basler, John J Mekalanos, and Petr G Leiman. Paar-repeat proteins sharpen and diversify the type vi secretion system spike. *Nature*, 500(7462):350–353, 2013.

[96] François Sigaux. Cancer genome or the development of molecular portraits of tumors. *Bulletin de l'Academie nationale de medecine*, 184(7):1441–7, 2000.

[97] Felipe A Simão, Robert M Waterhouse, Panagiotis Ioannidis, Evgenia V Kriventseva, and Evgeny M Zdobnov. Busco: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, 31(19):3210–3212, 2015.

[98] Jared T Simpson and Mihai Pop. The theory and practice of genome sequence assembly. *Annual review of genomics and human genetics*, 16, 2015.

[99] Jared T Simpson, Kim Wong, Shaun D Jackman, Jacqueline E Schein, Steven JM Jones, and Inanç Birol. Abyss: a parallel assembler for short read sequence data. *Genome research*, 19(6):1117–1123, 2009.

[100] BB Singh, OO Olufajo, MF Ishiyaku, RA Adeleke, HA Ajeigbe, and SG Mohammed. Registration of six improved germplasm lines of cowpea with combined resistance to striga gesnerioides and alectra vogelii. *Crop science*, 46(5):2332–2333, 2006.

[101] Mohar Singh, Hari D Upadhyaya, and Ishwari Singh Bisht. *Genetic and genomic resources of grain legume improvement*. Newnes, 2013.

[102] Guy St C Slater and Ewan Birney. Automated generation of heuristics for biological sequence comparison. *BMC bioinformatics*, 6(1):31, 2005.

[103] R Staden. A mew computer method for the storage and manipulation of dna gel reading data. *Nucleic acids research*, 8(16):3673–3694, 1980.

[104] Mario Stanke, Mark Diekhans, Robert Baertsch, and David Haussler. Using native and syntenically mapped cdna alignments to improve de novo gene finding. *Bioinformatics*, 24(5):637–644, 2008.

[105] Lincoln Stein. Genome annotation: from sequence to biology. *Nature reviews genetics*, 2(7):493–503, 2001.

[106] Javier F Tabima and Niklaus J Grünwald. effectr: An expandable r package to predict candidate rxlr and crn effectors in oomycetes using motif searches. *Molecular Plant-Microbe Interactions*, 32(9):1067–1076, 2019.

[107] Haibao Tang, Vivek Krishnakumar, Shelby Bidwell, Benjamin Rosen, Agnes Chan, Shiguo Zhou, Laurent Gentzbittel, Kevin L Childs, Mark Yandell, Heidrun Gundlach, et al. An improved genome release (version mt4. 0) for the model legume medicago truncatula. *BMC genomics*, 15(1):312, 2014.

[108] Haibao Tang, Xingtan Zhang, Chenyong Miao, Jisen Zhang, Ray Ming, James C Schnable, Patrick S Schnable, Eric Lyons, and Jianguo Lu. Allmaps: robust scaffold ordering based on multiple maps. *Genome biology*, 16(1):3, 2015.

[109] Eric Tannier, Chunfang Zheng, and David Sankoff. Multichromosomal median and halving problems under different genomic distances. *BMC Bioinformatics*, 10(1):120, 2009.

[110] Hervé Tettelin, Vega Masignani, Michael J Cieslewicz, Claudio Donati, Duccio Medini, Naomi L Ward, Samuel V Angiuoli, Jonathan Crabtree, Amanda L Jones, A Scott Durkin, Robert T Deboy, Tanja M Davidsen, Marirosa Mora, Maria Scarselli, Immaculada Margarit y Ros, Jeremy D Peterson, Christopher R Hauser, Jaideep P Sundaram, William C Nelson, Ramana Madupu, Lauren M Brinkac, Robert J Dodson, Mary J Rosovitz, Steven A Sullivan, Sean C Daugherty, Daniel H Haft, Jeremy Selengut, Michelle L Gwinn, Liwei Zhou, Nikhat Zafar, Hoda Khouri, Diana Radune, George Dimitrov, Kisha Watkins, Kevin J B O'Connor, Shannon Smith, Teresa R Utterback, Owen White, Craig E Rubens, Guido Grandi, Lawrence C Madoff, Dennis L Kasper, John L Telford, Michael R Wessels, Rino Rappuoli, and Claire M Fraser. Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome". *Proc. Natl. Acad. Sci. U. S. A.*, 102(39):13950–13955, September 2005.

[111] Hervé Tettelin, Vega Masignani, Michael J Cieslewicz, Claudio Donati, Duccio Medini, Naomi L Ward, Samuel V Angiuoli, Jonathan Crabtree, Amanda L Jones, A Scott Durkin, Robert T Deboy, Tanja M Davidsen, Marirosa Mora, Maria Scarselli, Immaculada Margarit y Ros, Jeremy D Peterson, Christopher R Hauser, Jaideep P Sundaram, William C Nelson, Ramana Madupu, Lauren M Brinkac, Robert J Dodson, Mary J Rosovitz, Steven A Sullivan, Sean C Daugherty, Daniel H Haft, Jeremy Selengut, Michelle L Gwinn, Liwei Zhou, Nikhat Zafar, Hoda Khouri, Diana Radune, George Dimitrov, Kisha Watkins, Kevin J B O'Connor, Shannon Smith, Teresa R Utterback, Owen White, Craig E Rubens, Guido Grandi, Lawrence C Madoff, Dennis L Kasper, John L Telford, Michael R Wessels, Rino Rappuoli, and Claire M Fraser. Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome". *Proc. Natl. Acad. Sci. U. S. A.*, 102(39):13950–13955, September 2005.

[112] Hervé Tettelin, Vega Masignani, Michael J Cieslewicz, Claudio Donati, Duccio Medini, Naomi L Ward, Samuel V Angiuoli, Jonathan Crabtree, Amanda L Jones, A Scott Durkin, et al. Genome analysis of multiple pathogenic isolates of streptococcus agalactiae: implications for the microbial "pan-genome". *Proceedings of the National Academy of Sciences*, 102(39):13950–13955, 2005.

[113] Hervé Tettelin, David Riley, Ciro Cattuto, and Duccio Medini. Comparative genomics: the bacterial pan-genome. *Curr. Opin. Microbiol.*, 11(5):472–477, October 2008.

[114] The Computational Pan-Genomics Consortium. Computational pan-genomics: status, promises and challenges. *Briefings in Bioinformatics*, 19(1):118–135, 10 2016.

[115] Trudy A Torto, Shuang Li, Allison Styer, Edgar Huitema, Antonino Testa, Neil AR Gow, Pieter Van West, and Sophien Kamoun. Est mining and functional expression

assays identify extracellular effector proteins from the plant pathogen phytophthora. *Genome research*, 13(7):1675–1685, 2003.

[116] Cole Trapnell, Brian A Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J Van Baren, Steven L Salzberg, Barbara J Wold, and Lior Pachter. Transcript assembly and quantification by rna-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology*, 28(5):511–515, 2010.

[117] Edouard Vannier, Benjamin E Gewurz, and Peter J Krause. Human babesiosis. *Infectious disease clinics of North America*, 22(3):469–488, 2008.

[118] Bruce J Walker, Thomas Abeel, Terrance Shea, Margaret Priest, Amr Abouelliel, Sharadha Sakthikumar, Christina A Cuomo, Qiandong Zeng, Jennifer Wortman, Sarah K Young, et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PloS one*, 9(11):e112963, 2014.

[119] Neil I Weisenfeld, Vijay Kumar, Preyas Shah, Deanna M Church, and David B Jaffe. Direct determination of diploid genome sequences. *Genome research*, 27(5):757–767, 2017.

[120] Stephen C Whisson, Petra C Boevink, Lucy Moleleki, Anna O Avrova, Juan G Morales, Eleanor M Gilroy, Miles R Armstrong, Severine Grouffaud, Pieter Van West, Sean Chapman, et al. A translocation signal for delivery of oomycete effector proteins into host plant cells. *Nature*, 450(7166):115–118, 2007.

[121] Kelly L Williams. Gene mapping. In *Encyclopedia of bioinformatics and computational biology*, pages 242–250. Academic Press, 2019.

[122] Min Xie, Claire Yik-Lok Chung, Man-Wah Li, Fuk-Ling Wong, Xin Wang, Ailin Liu, Zhili Wang, Alden King-Yung Leung, Tin-Hang Wong, Suk-Wah Tong, et al. A reference-grade wild soybean genome. *Nature communications*, 10(1):1–12, 2019.

[123] Kai Yang, Zhixi Tian, Chunhai Chen, Longhai Luo, Bo Zhao, Zhuo Wang, Lili Yu, Yisong Li, Yudong Sun, Weiyu Li, et al. Genome sequencing of adzuki bean (vigna angularis) provides insight into high starch and low fat accumulation and domestication. *Proceedings of the National Academy of Sciences*, 112(43):13213–13218, 2015.

[124] Shaolun Yao, Chuan Jiang, Ziyue Huang, Ivone Torres-Jerez, Junil Chang, Heng Zhang, Michael Udvardi, Renyi Liu, and Jerome Verdier. The vigna unguiculata gene expression atlas (vu gea) from de novo assembly and quantification of rna-seq data provides insights into seed maturation mechanisms. *The Plant Journal*, 88(2):318–327, 2016.

[125] Wen Yao, Guangwei Li, Hu Zhao, Gongwei Wang, Xingming Lian, and Weibo Xie. Exploring the rice dispensable genome using a metagenome-like assembly strategy. *Genome biology*, 16(1):1–20, 2015.

[126] Xiwen Yao, Kaili Xu, and Yu Liang. Comparing the thermo-physical properties of rice husk and rice straw as feedstock for thermochemical conversion and characterization of their waste ashes from combustion. *BioResources*, 11(4):10549–10564, 2016.

[127] Qiang Zhao, Qi Feng, Hengyun Lu, Yan Li, Ahong Wang, Qilin Tian, Qilin Zhan, Yiqi Lu, Lei Zhang, Tao Huang, et al. Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. *Nature genetics*, 50(2):278–284, 2018.

[128] Yongbing Zhao, Chen Sun, Dongyu Zhao, Yadong Zhang, Yang You, Xinmiao Jia, Junhui Yang, Lingping Wang, Jinyue Wang, Haohuan Fu, Yu Kang, Fei Chen, Jun Yu, Jiayan Wu, and Jingfa Xiao. PGAP-X: extension on pan-genome analysis pipeline. *BMC Genomics*, 19(Suppl 1):36, January 2018.

[129] Yongbing Zhao, Jiayan Wu, Junhui Yang, Shixiang Sun, Jingfa Xiao, and Jun Yu. PGAP: pan-genomes analysis pipeline. *Bioinformatics*, 28(3):416–418, February 2012.