Chapter 14

Population Genomics of Fungal Plant Pathogens and the Analyses of Rapidly Evolving Genome Compartments

Christoph J. Eschenbrenner, Alice Feurtey, and Eva H. Stukenbrock

Abstract

Genome sequencing of fungal pathogens have documented extensive variation in genome structure and composition between species and in many cases between individuals of the same species. This type of genomic variation can be adaptive for pathogens to rapidly evolve new virulence phenotypes. Analyses of genome-wide variation in fungal pathogen genomes rely on high quality assemblies and methods to detect and quantify structural variation. Population genomic studies in fungi have addressed the underlying mechanisms whereby structural variation can be rapidly generated. Transposable elements, high mutation and recombination rates as well as incorrect chromosome segregation during mitosis and meiosis contribute to extensive variation observed in many species. We here summarize key findings in the field of fungal pathogen genomics and we discuss methods to detect and characterize structural variants including an alignment-based pipeline to study variation in population genomic data.

Key words Fungal pathogens, Genome compartments, Transposable elements, De novo assembly, Multiple genome alignments

1 Introduction

The kingdom Fungi comprises a diverse group of pathogens that infect animals and plants. Understanding the evolution and infection biology of fungal pathogen species is evidently necessary to know how to combat the diseases caused by these organisms. Primary objectives to be addressed in population genomic studies of fungal pathogens relate to the origin of the pathogen, routes of migration, and epidemiology. Moreover, genome data can shed light on the underlying determinants of pathogenicity, which may be new targets in disease control. Finally, as we will outline in this chapter, fungal pathogens provide interesting model systems to study the evolution of genome architecture.

In this chapter, our focus will be on fungi that cause disease on plants. Genome data permitted the reconstruction of the

Julien Y. Dutheil (ed.), *Statistical Population Genomics*, Methods in Molecular Biology, vol. 2090,
https://doi.org/10.1007/978-1-0716-0199-0_14, © The Author(s) 2020

evolutionary histories of some of the most important fungal plant pathogens. For example, the speciation history of the ascomycete wheat pathogen *Zymoseptoria tritici* has been reconstructed by whole genome coalescence analyses revealing that this pathogen emerged with the onset of wheat domestication in the Middle East during the Neolithic Revolution 10–12,000 years ago [1, 2]. Population genetic analyses of isolates representing a world-wide collection of *Z. tritici* was applied to infer the migration history of the pathogen and showed a subsequent dispersal of the pathogen with the spread of wheat cultivation to Europe, Asia and later to New World countries [3]. Another important and recently emerged wheat pathogen is the wheat blast fungus *Magnaporthe oryzae*. The wheat blast disease first emerged in South America and strict quarantine strategies were employed to contain the pathogen within one region and avoid dispersal to other continents. However, the disease was recently reported in Bangladesh. Islam and colleagues were able to track the origin of the wheat blast outbreak in Bangladesh to South America using a genome-wide SNP dataset from 20 isolates collected from different host species in Brazil and Bangladesh [4]. This type of phylogenomic studies and “genomic surveillance” has proven of great relevance to monitor plant disease outbreaks and support the design of improved disease management strategies.

Genome data from fungal plant pathogens has also been a resource for the discovery of genes encoding virulence determinants. In particular quantitative trait locus (QTL) mapping and genome-wide association studies (GWAS) have proven powerful in this field. QTL mapping, based on phenotypic analyses and marker segregation in progeny populations, have been applied to identify the avirulence gene *AvrStb6* in *Z. tritici* [5, 6]. However, QTL analysis has several drawbacks: it relies on the analyses of crosses between two strains. This limits the resolution of the study, depending on the amount of variation between the two strains. Moreover, many fungi propagate primarily by asexual reproduction and many sexual species cannot be crossed under laboratory conditions excluding the possibility of QTL analysis. GWAS on the other hand uses outbred population and polymorphisms that represent the standing genetic variation in a population, providing a higher resolution along the genome [7]. GWAS analyses have been used to identify polymorphisms associated with fungicide sensitivity, mycotoxin production and aggressiveness of the wheat pathogen *Fusarium graminearum* [8], virulence determinants of the pine tree pathogen *Heterobasidion annosum* [9], and toxin production of another wheat pathogen *Parastagnospora nodorum* [10].

Another way to detect genes relevant for pathogenicity in fungi, is to apply evolutionary predictions to identify signatures of recent or past selection. Genes involved in host–pathogen interactions are expected to evolve by antagonistic selection, either following an "arms-race" or a "trench-warfare" scenario of coevolution [11, 12]. The "arms-race" scenario refers to positive selection that repeatedly fixes new advantageous alleles at the locus under selection. The trench-warfare scenario on the other hand refers to the continuous maintenance of different alleles in the population by balancing and diversifying selection. Thus, identifying genes with signatures of positive or balancing selection in pathogen genome will likely uncover genes playing a role in host–pathogen interaction. Evolutionary predictions have been used to identify a number of virulence determinants in fungal plant pathogens and confirm the prediction that virulence determinants indeed often exhibit a signature of positive selection and accelerated evolution [13, 14].

Genome sequencing of hundreds of pathogenic fungal species has revealed extensive variation in genome structure and size [12]. Sequenced genomes range in size from 2 Mb in the Microsporidia to 2 Gb in Pucciniales species, and comprise different levels of ploidy and in some species even aneuploidy [15]. A consistent finding from comparative studies of pathogenic fungi is an extreme extent of genome plasticity whereby closely related species or individuals of the same species can have highly different genome structure and size, and vary in gene content and gene organization [12]. There is evidence that this genome plasticity is crucial for the pathogenic lifestyle. Indeed, variation is essential for pathogens to rapidly adapt to changes in their environment, in particular changes in host resistances, and a highly flexible genome composition appears to be an adaptive mechanism for pathogens to rapidly generate new genetic variation.

The field of fungal pathogen genomics has focused on the sources and patterns of genomic variation, and the contribution of this variation to gene evolution, in particular the evolution of virulence related genes, so called effectors. Effector genes encode secreted proteins that are involved in the suppression of host defenses and these genes are located in genomic segments exhibiting structural variation, including accessory chromosomes and islands of repetitive DNA (e.g., [16–19]). The challenge of studying patterns of evolution in these regions lies in the difficulty of assembling and comparing structurally different sequences.

Population genomic analyses, taking structural variation into account, have been instrumental in determining the underlying drivers of rapid evolution and genome variation in most pathogenic fungal species. This chapter will summarize some of the key discoveries from population genomics analyses of fungal pathogens.

2 Key Discoveries from Population Genomics in Plant Fungal Pathogens

2.1 High Recombination Rates and Population Admixture Contribute to Rapid Adaptation of Fungal Plant Pathogen Genomes

Population genomic data has been applied in a few studies to address the rapid evolution of fungal plant pathogens (reviewed in [12, 20]). Mechanisms that generate genetic variation in a population include mutational processes, recombination and gene flow. The fate of this variation is then determined by selection, genetic drift and the effective population size of the organism. Many aspects make it difficult to study the population genetics and demography of fungi and to assess the contribution of different evolutionary mechanisms to evolution. Most population genetic analyses rely on evolutionary models that make assumptions about the underlying genetic structure of the population (e.g., random mating, infinite site model, a low and constant recombination rate, clonality, skewed offspring, and constant population size). In fungi, many species reproduce both asexually and sexually. More generally, the reproductive mode of fungi can be considered as a continuum ranging from predominantly clonal to strictly out-crossing. Furthermore, the reproductive mode of a particular taxa may change over time. For example, a species may propagate asexually for a certain time followed by a time of more frequent sexual reproduction. Extensive differences in the content of transposable elements between closely related species may support the occurrence of prolonged periods of asexual reproduction in many individual lineages [21, 22].

In population genetic analyses, it is often necessary to have a clear definition of generation time, in order to convert relative time to actual years. However, the generation time of a fungal individual that produces both by asexual and sexual reproduction is difficult to define. In these organisms, not only sexual generations can contribute to novel genetic variation but also asexual generations where high mutations rates generate clonal variation. Furthermore, little is known about the variation in sexual or asexual generations per year. However, the frequency of sexual mating and spore formation may vary from year to year according to environmental conditions and the availability of compatible sexual partners. In summary, analyses of fungal population genomic data, based on existing population models involve many uncertainties caused by our limited understanding of the population biology of fungal pathogen species and by the inadequacy of classic population genetic assumptions to the life history traits of these organisms.

Despite these limitations, population genomic data has, for a few model species, provided new insight into genome the evolution and population biology of the plant pathogens. For example, the impact of recombination on genome evolution has been studied in both ascomycete and basidiomycete pathogens. Badouin and coworkers used population genomic data to infer linkage disequilibrium (LD) along the genome of two closely related species of the anther smut fungus *Microbotryum* [23]. Using information about the extent of LD and the site frequency spectrum (SFS), the authors could determine the distribution and frequency of selective sweeps along the genomes and thereby demonstrate the recent impact of natural selection on gene and genome evolution in the two species. While recombination in *Microbotryum* has been crucial to fix adaptive mutations, suppression of recombination in other parts of the genome has shaped evolution of mating type chromosomes. On these chromosomes recombination suppression has contributed to the generation and maintenance of "super genes" comprising the genes responsible for pre- and postmating compatibility [24].

The impact of recombination has also been studied in *Z. tritici* and its close relative *Zymoseptoria ardabiliae* using population genomic data. These analyses revealed exceptionally high rates of recombination, including recombination hotspots localizing in protein coding genes [25]. Furthermore a strong correlation of recombination with both positive and negative selection was recently demonstrated [26]. Thereby, a negative correlation of recombination and p*N*/p*S*, the proportion of nonsynonymous to synonymous polymorphisms, demonstrates an important role of recombination in removing nonadaptive mutations. On the other hand, a positive correlation of recombination with the rate of adaptive nonsynonymous mutations, $\omega_{A,}$ was reported, showing that recombination likewise contributes to the efficient fixation of advantageous mutations in this species.

The impact of intra-specific gene flow on the population genetic structure and dynamic was elegantly demonstrated by a transcriptome sequencing of wheat leaves infected with the yellow stripe rust pathogen *Puccinia striiformis* [27]. *P. striiformis* is an obligate pathogen and difficult to culture on artificial media. Direct sequencing of infected leaf material thus provides a powerful approach to capture the genetic diversity of isolates in the field. Bueano-Sancho and colleagues used data from 246 infected leaves of wheat, triticale, and rye collected in 2 years and at different geographical locations. They used population genetic analyses to infer the population structure and recent patterns of gene flow and admixture of the European rust population and demonstrate extremely diverse populations and rapid seasonal shifts of the rust populations [27]. A significant impact of gene flow on the population genetic structure of fungal pathogens has been demonstrated in other studies also using population genomic data, for example, in the rice blast pathogen *Magnaporthe oryzae* [28] and the ash dieback pathogen *Hymenoscyphus fraxineus* [29].

The impact of new mutations has also been extensively studied in fungal plant pathogen genomes. This is because many species show exceptionally high rates of mutational changes in some segments of their genomes, and the ability to rapidly generate new genetic variation by mutations likely represents an adaptive trait. In the next section we outline the peculiarity of many plant pathogen genomes with respect to genome architecture of the distribution of mutation-prone genome regions.

3 Fungal Plant Pathogen Genomes Are Often Compartmentalized, A Trait Driven by Transposable Elements

The origin of genome compartments in fungal pathogens is still poorly understood, but can only be studied with well-assembled and aligned genome sequences that allow us to study patterns of nucleotide variation within and around these particular genomic regions. Improved genome assemblies have provided insight into the repetitive fraction of fungal pathogen genomes. Repeat contents can vary from less than 1% in *Fusarium graminearum* to more than 80% in some rust and mildew species [30, 31]. The factors determining repeat accumulation are poorly understood, but can include sexual versus asexual reproduction and different genome defense mechanisms such as DNA methylation and Repeat Induced Point mutations (RIP). Transposable elements may accumulate during prolonged asexual reproduction in the absence of recombination; however, some of the sequenced species with the highest repeat content, such as many rust fungi, are sexual, suggesting that other factors likewise are important determinants of transposable element activities.

In some fungal pathogen species a large portion of the repetitive elements are found in particular accessory segments or entire chromosomes that are nonessential but in some species important for virulence. The genome of the asexual fungus *Verticillium dahliae* comprises particular islands enriched with transposable element and encoding effector genes [16, 32]. These islands are present in different lineages of the pathogen and contribute to variation in virulence. Interestingly, these genomic islands harbor little nucleotide variation among individuals that share a particular island, possibly reflecting the strong impact of natural selection on the genes encoded by these regions. Variation in virulence phenotypes is thus given by the presence–absence polymorphism of an entire genomic fragment.

The genome of the fungus *Leptosphaeria maculans* infecting oil seed rape also comprises repeat rich compartments that encode effector proteins [17]. These regions show a particular mutation pattern conferred by RIP. RIP acts to inactivate transposable elements by introducing mutations in repetitive sequence. RIP produces cytosine to thymine (C to T) mutations and can thereby locally impacts the GC content of the sequence [33]. This is the case for *L. maculans* where the repeat-rich islands have become AT isochores with highly distinct GC content compared to the remaining genome.

Genome compartments can also be contained in the genome as accessory chromosomes. The wheat pathogen *Zymoseptoria tritici* has a large number of such accessory chromosomes, eight of them have been sequenced in the reference isolate. These chromosomes can be lost and rearranged during mitosis as well as meiosis [34, 35]. Beside this large complement of accessory chromosomes, *Z. tritici* also exhibits a considerable amount of chromosome length polymorphisms of the core chromosomes as demonstrated by electrophoretic separation of chromosomes and PacBio sequencing [36, 37]. In the soil-borne pathogen *Fusarium oxysporum* lineage-specific chromosomes encode virulence determinants that enable the fungus to be pathogenic on specific host species by the defeat of host defenses [22].

How are accessory chromosomes lost and how are they maintained in populations? A few studies mainly focusing on *F. oxysporum* and *Z. tritici* have started to address these questions. These studies have demonstrated the exceptionally fast rate of accessory chromosomes loss during mitosis [38, 39]. In *F. oxysporum* amplification and maintenance of the chromosomes likely depend on the horizontal exchange of these chromosomes by vegetative fusion of hyphae. In *Z. tritici* however, the accessory chromosomes can be amplified during meiosis by a meiotic drive mechanism [40]. In both species, mechanisms that allow the loss of chromosomes as well as mechanisms that reamplify the chromosomes may have evolved to rapidly generate new genetic variation in the populations of pathogens.

4 Interspecific Hybridization Contributes to Genome Evolution of Fungal Plant Pathogens

Reproductive barriers between fungal species are in many cases poor predictors of species boundaries. Sexual mating and fusion of hyphae between nonconspecific individuals have been frequently described and demonstrate a pathway of gene exchange across species boundaries in the kingdom Fungi. We have recently reviewed the literature on fungal hybridization [41] and will here only mention a few prominent examples of hybridization and gene exchange between fungal species.

Hybridization has been shown to be responsible for the rapid emergence of new virulent lineages of different fungal plant pathogens, including *Ophiostoma nova-ulmi*, the causal agent of Dutch Elm disease and the powdery mildew pathogen *Blumeria graminis-triticale* on crop species Triticale [42, 43]. For the Dutch Elm disease fungus, occasional hybridization events have played a role in the exchange of virulence determinants between otherwise distinct lineages. *B. graminis-triticale*, on the other hand, is the product of few hybridization events between powdery mildew species infecting wheat and rye, respectively. The evidence for a hybridization event is a particular mosaic distribution of genetic variation that clearly reflects the two parental genomes recombined in one genome [43]. The two examples demonstrate very different outcome of hybridization ranging from a few signatures of introgression to entirely mixed parental genomes and hybrid speciation.

The exchange of genetic material can also occur as horizontal gene transfer where only a fragment of DNA is integrated into the genome of one species from another organism. The wheat pathogens *Parastagonospora nodorum* and *Pyrenophora tritici-repentis* are two distantly related ascomycete pathogens. However, their genomes comprise one region of exceptionally high sequence identity [44]. This region that is flanked by transposable elements includes a gene that encodes a proteinaceous toxin, ToxA. ToxA is a virulence factor that confers necrosis in susceptible wheat cultivars and the acquisition of the *ToxA* gene by *P. tritici-repentis* from *P. nodorum* by horizontal gene transfer, allowed the emergence a new virulent lineage of *P. tritici-repentis* infecting wheat. Interestingly, genome sequencing revealed that the *ToxA* gene also is present in another wheat pathogen *Bipolaris sorokiniana* suggesting that this gene may be carried by a bacterial or viral vector frequently associated with wheat [45].

Multiple signatures of hybridization and interspecific gene exchange supports a high extent of flexibility in terms of genome content and structure in fungal plant pathogens. The finding that introgression and horizontal gene transfer in some cases involve virulence determinants underlines the importance of studying not only these regions, but also the processes whereby they occur. However, hybridization events between more distantly related species may be challenging to identify with population genomic data. This is because outlier loci in the genome that comprise highly diverged haplotypes can be difficult to assemble by reference-based assembly approaches. Below we discuss how to circumvent this issue by alignment of de novo assembled genomes.

5 Discovering Variation in Population Genomic Data

5.1 Variant Calling Through Short-Read Mapping: Methods and Limits

Most population genomics approaches are based on the mapping of short sequencing reads, using software such as bwa or bowtie to a well-assembled reference genome [46–48]. Tools such as *GATK*, *SAMtools mpileup*, or *FreeBayes* can be used to call single nucleotide variants and small indels from the mapping file and output this information in a Variant Call Format (VCF) file [49–52]. Here, we will not go further into details about these methods for SNP discovery as these have been extensively reviewed elsewhere (e.g., [53–55]).

Variant discovery through mapping of short reads to a reference is supported by a large number of well-documented tools. However, these methods have drawbacks, some of them especially relevant in nonmodel organisms such as most fungal pathogens. As mentioned above, many pathogenicity-related genes locate in repeat rich compartments of fungal pathogen genomes, and mapping based approaches may not be ideal for the characterization of genetic variation in these regions. Alignment in low-complexity or repetitive regions, although facilitated by paired or mated reads, is often challenging due to the difficulty of correctly mapping the sequence to the reference [55]. Dependence on a reference genome can also be an issue in nonmodel organisms for which a complete reference genome is not always available. Indeed, any misassembly or single nucleotide error in the reference genome could be reflected in the final variants. Poor assembly quality would also lead to structural variation being impossible to discover. Finally, mapping of short reads will not perform efficiently in presence of high genetic variability. Such high variability may be found locally in genomes that have experienced introgression or in some regions have a higher mutation rate. In either case, reads containing multiple alternative alleles might not map correctly, resulting in the under-representation of the diverging haplotypes [55].

Another limitation to mapping-based approaches is the detection of structural variation. To detect translocations or inversions, genomes can be de novo assembled and compared in a multiple genome alignment (Fig. 1). Fungal genomes are convenient for this approach as they often are relatively small and can be sequenced in the haploid phase, therefore preventing issues with heterozygosity and phasing.

Sequencing technologies based on longer reads (e.g., PacBio SMRT or Nanopore sequencing) provide improved resources for de novo assembly. These technologies have proven valuable in the improved detection of structural variation in plant pathogen genomes, including repeat-rich accessory segments on core chromosomes [56, 57]. Below, we describe methods to use de novo genome assembly based on both long and short read sequencing and give the details of a pipeline which allows variants calling from these assembled genomes.

6 De Novo Assembly and the Rise of Long-Read Sequencing

A number of assemblers are available for the different types of sequencing reads available including short reads produced by Illumina sequencing and long reads produced, for example, by SMRT

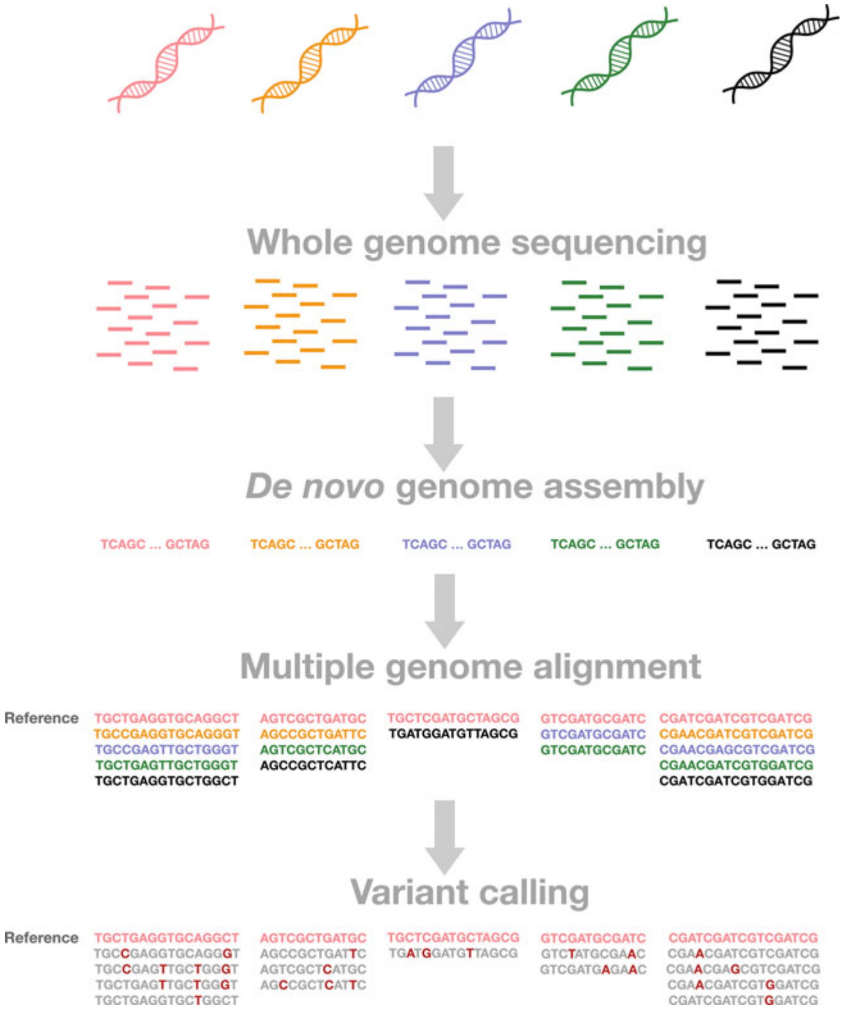
Whole genome sequencing
De novo genome assembly
TCAGC ... GCTAG
TCAGC ... GCTAG
TCAGC ... GCTAG
TCAGC ... GCTAG
TCAGC ... GCTAG
Multiple genome alignment
Reference
TGCTGAGGTGCAGGCT AGTCGCTGATGC TGCTCGATGCTAGCG GTCGATGCGATC CGATCGATCGTCGATCG
TGCCGAGGTGCAGGGT AGCCGCTGATTC TGATGGATGTTAGCG GTCGATGCGATC CGAACGATCGTCGATCG
TGCCGAGTTGCTGGGT AGTCGCTCATGC GTCGATGCGATC CGAACGAGCGTCGATCG
TGCTGAGTTGCTGGGT AGCCGCTCATTC CGAACGATCGTGGATCG
TGCTGAGGTGCTGGCT CGATCGATCGTGGATCG
Variant calling
Reference
TGCTGAGGTGCAGGCT AGTCGCTGATGC TGCTCGATGCTAGCG GTCGATGCGATC CGATCGATCGTCGATCG
TGCCGAGGTGCAGGGT AGCCGCTGATTC TGATGGATGTTAGCG GTCTATGCGAAC CGAACGATCGTCGATCG
TGCCGAGTTGCTGGGT AGTCGCTCATGC GTCGATGAGAAC CGAACGAGCGTCGATCG
TGCTGAGTTGCTGGGT AGCCGCTCATTC CGAACGATCGTGGATCG
TGCTGAGGTGCTGGCT CGATCGATCGTGGATCG

Fig. 1 Generation of population genomic datasets using multiple genome alignment (MGA). Genomes of multiple individuals are generated by short or long read sequencing and assembled de novo. De novo genome assemblies are aligned to generate a MGA. The alignment consists of alignment blocks of different sizes (number of sequences) and lengths (base-pair of alignment). The MGA is projected against a single reference sequence (here shown in red). The projection rearranges each alignment block so that the reference sequence represents the positive strand of the genome. Variable positions can be called directly from the MGA and summarized in a VCF file

sequencing. For de novo assembly of short read data programs like *SPAdes* [58], *SOAPdenovo2* [59] or *IDBA-UD* [60] based on de Brujn graph assemblies are available [61]. De Brujn graph-based assemblers work by splitting the short reads into even shorter units of uniform size, the so-called k-mers. These k-mers provide the basis for the reconstruction of the genome sequence based on overlap of different k'mers while information about the local connectivity of each k'mer is preserved by a De Bruijn graph structure (*see*, e.g., ref. 62, 63). To properly handle repetitive regions the De

Bruijn graph assembler masks repetitive and low-complexity regions and assemble the remaining genome into many contigs and scaffolds.

De novo genome assemblies of long read data is based on other algorithms and build on the alignment of overlapping reads [64, 65]. Long read sequencing with SMRT technology provides an average read length of 10 kb that can be assembled with assemblers like *Canu* [64], *Falcon* [65], or *SMRTAssembly* (©Pacific Biosciences). Nanopore sequencing is producing even longer reads, mostly dependent on the length of the extracted DNA fragment, by a MinION instrument. Methods to assemble genomes based on this technology partially overlap with the ones used with SMRT technology, for example, with the *Canu* assembler [64] and are reviewed in de Lannoy et al. [66]. Nanopore reads have been used to improve the N50 of the maize pathogen *Rhizoctonia solani* by an order of magnitude compared to previous efforts [67]. This improvement is even more pronounced in genomes with high repetitive content (e.g., [56, 68]). In the oat crown rust fungus *Puccinia coronata* f. sp. *avenae* with a genome-wide repeat content of more than 50%, long read sequencing has enabled detailed characterization of structural variants [69]. Moreover, assembly of long read data provided a map of SNPs not only between individuals but also between nuclei in the dikaryotic hyphae of *P. coronata* f. sp. *avenae*, a level of variation so far poorly studied in fungi.

The main inconvenience with long-read sequencing methods so far is the high error rate. To circumvent this issue, it is necessary to either increase the sequencing depth or to combine the advantages of short and long reads. Indeed, assemblies of long and short read data from the same genome is also possible with "hybrid assemblers" like *hybridSPAdes* whereby the long read data ensures the assembly of long scaffolds and the short read data provides high coverage of individual nucleotides in the assembly [70]. Instead of using both types of reads during the assembly process, it is also possible to correct long-read de novo assembly with short-read data, using software like *Pilon* [71]. Such an approach was recently used to assemble genomes of the species *Leptosphaeria* and *Zymoseptoria* [37, 72].

It is important to note that different assemblers (for short-read data as well as long-read data) may perform differently with different genome datasets depending on the repeat content and sequence complexity. Moreover, the long-read technologies are improving at a very rapid rate and new tools and methods are constantly developed. We therefore advise reviewing the latest methods, testing different assemblers with a given dataset and comparing the resulting assemblies with tools such as *Quast* to determine the best performance [73]. To evaluate the quality of the assemblies, key parameters to compare are the total length of the assembly, the number of contigs and the overall size of the assembled fragments which can be summarized by the N50 value (defined as the largest contig length, *L*, whereby contigs of length superior or equal to *L* accounts for at least 50% of the bases of the assembly).

For population genomics analyses of the fungal wheat pathogen *Z. tritici* we have developed a pipeline based on de novo genome assembly and multi genome alignments (Fig. 1). This method has allowed us to quantify and characterize accessory regions in the genome of *Z. tritici* and to identify hitherto overseen signatures of introgression along the genome of the pathogen [74]. Following de novo assembly of either short or long read sequencing, the next step in our pipeline is the generation of a multiple genome alignment with a multiple genome aligner such a *TBA* [75], *Mugsy* [76], or *progessiveMauve* [77]. These aligners first generate pairwise alignments of all genomes and next combine these into a multiple genome alignment. The resulting alignment, for example, in "multiple alignment format" *maf* file consists of a large number of local alignment blocks, that differ in their length and the number of sequences included in the block (*see* Chapter 2). The variation in sequence numbers per block along the genome may reflect actual presence/absence variation in genome segments, but can also reflect the parts of the genome that is prone to assembly and/or alignment errors. A thorough filtering and realignment of the alignment blocks is therefore necessary to ensure that the observed patterns are biologically relevant. Programs like *Mafft* or *T-Coffee* are available for realignment of alignment blocks to ensure the optimal comparison of sequences [78, 79].

Filtering and variant detection from a multiple genome alignment can be done with programs like *Maffilter* [80] (*see* also Chapter 2). *Maffilter* allows the list of variant sites identified across the aligned genomes to be outputted as a VCF file. This format is identical to the one used by classic variant calling following a mapping approach. This is especially convenient, as it will allow for these variants to be used as input by any population genomics programs designed to work on this well-known format. Another advantage of this pipeline is that it allows to detect variants simultaneously using sequencing data produced by different technologies, for example, in the case here some genomes are obtained by Illumina sequencing and other by PacBio SMRT sequencing (Feurtey et al. unpublished).

7 Detection of Structural Variation in Genomes

Structural variation is increasingly being recognized as an important level of genetic variation to study. In a study of a single human genome, Pang and colleagues found that the genome differed from the reference human genome by only 0.1% when considering SNPs but by approximately 1.2% when considering other source of genetic variation such as insertions, deletions, or copy number variations [81]. In fungal plant pathogens, structural variation is recognized as an important type of variation as highlighted in some of the examples summarized above.

Methods based on read mapping can be applied to characterize structural variation along genomes [82]. These methods rely on several types of information to detect structural variants including read-depth, and the distribution of paired-end and split reads. Read depth in a mapping, that is, the number of sequencing reads aligning to a specific locus, can give information about copy number variations and deletions. For example, a locus with a higher depth than expected could indicate a duplication and a lower depth (close to 0 in a haploid genome, half the expected depth in a diploid genome) a deletion [83].

Deletions in the resequenced genome compared to the reference genome will cause the insert size of paired-end reads (the DNA fragment including the sequenced reads and the gap sequence between the reads) to appear larger than expected, while an insertion will make the insert look smaller than expected. Furthermore, pairs in which one read aligns to the genome while the other does not may reflect an insertion of a TE if the second read aligns to a repeated element somewhere else in the genome. Likewise translocations, inversions, and other kinds of structural variants can be inferred from pairs of reads. Aligning DNA genomic sequencing reads using an aligner created for RNAseq and thus able to split a read sequence (usually, due to intronic sequences being spliced out of the read) would allow detecting deletions since the deleted sequence will look like a splicing junction site. Software that can detect such structural variants include *Pindel*, *Delly*, or *LUMPY* [84–86]. More details about these methods and software can be found for instance in [83, 87, 88].

Although these methods can uncover many structural variants from short and long reads, they do have their limits. Some of these methods make strict assumptions about the sequencing data, which are not always met in real data. Methods based on read depth assume that the sequencing depth is uniform across the genome and that variation mainly is explained by structural variants. However, variation in GC content and sequence composition along the genome can also cause variation in sequencing efficiency and thereby sequencing depth [83, 89]. Moreover, genomic segments such as accessory chromosomes or large insertions which do not always exist in a reference genome cannot be detected by mapping of short reads to a reference [88].

Whole genome assembly is able to uncover all types of structural variation, including large DNA fragments, which are not present in the reference genome. Another advantage of whole genome assembly is that, if the quality of the assembly is good, it provides strong evidence that no structural variant has gone undetected [88]. When the number of genomes is low, structural variants can be identified visually using, for instance, *Symap* or *circos*, which provide easily interpretable visualization of genome alignments [90, 91]. Specific software able to detect structural variants from de novo assemblies have also been developed such as Assemblytics and AsmVar, an automatization step that accounts for structural variants a population level [92, 93]. In summary several tools are available to detect and characterize structural variants in population genomic data. In organisms, like fungal pathogens, with highly variable genomes, accounting for structural variants is essential in order to understand genome evolution and the impact of mutation and recombination along the genome.

8 Conclusion

Analyses of genetic variation in fungal plant pathogen genomes have to a large extent focused on highly variable regions, on species-specific traits and presence–absence variation. More detailed analyses in a few species point to these regions being of particular interest as they can encode important pathogenicity factors. Variation in these regions is therefore considered to be adaptive in accordance with rapid host–pathogen coevolution. Population genomic studies that aim to characterize genetic variation in highly variable regions rely on high quality assemblies and alignments. De novo assemblies of long read sequence data provide an important new resource to capture variation in these regions, including variation in transposable element sequences.

The processes that drive genome evolution in fungal pathogen genomes is still poorly understood. We have demonstrated exceptionally high rates of recombination and particular mechanisms that introduce new mutations at high a rate. Furthermore, we know that fungal pathogens can exchange genes with other species either by sexual mating or fusion of asexual structures. However, the underlying mechanism of these processes, as well as the impact of natural selection on genetic variation generated by these is still to be unraveled.

With their small genome size and in many cases particular genome architecture, fungal pathogens, however, provide excellent models organisms for fundamental studies of genome evolution. Moreover, a better understanding of evolutionary processes occurring in pathogen populations is crucial for the development of agricultural ecosystems with higher disease resistance [94].

Acknowledgments

Research in the lab of EHS is funded by the German Research Council, the Max Planck Society and the State of Schleswig Holstein. The preparation of this book chapter was, furthermore, supported by Deutsche Forschungsgemeinschaft (DFG), in the context of the priority program SPP1819—"Rapid evolutionary adaptation—Potential and constraints."

References

1. Stukenbrock EH, Banke S, Javan-Nikkhah M, McDonald BA (2007) Origin and domestication of the fungal wheat pathogen Mycosphaerella graminicola via sympatric speciation. Mol Biol Evol 24:398–411
2. Stukenbrock EH, Bataillon T, Dutheil JY, Hansen TT, Li R, Zala M, McDonald BA, Wang J, Schierup MH (2011) The making of a new pathogen: insights from comparative population genomics of the domesticated wheat pathogen Mycosphaerella graminicola and its wild sister species. Genome Res 21:2157–2166
3. Banke S, McDonald BA (2005) Migration patterns among global populations of the pathogenic fungus Mycosphaerella graminicola. Mol Ecol 14:1881–1896
4. Islam MT, Croll D, Gladieux P et al (2016) Emergence of wheat blast in Bangladesh was caused by a South American lineage of Magnaporthe oryzae. BMC Biol 14:84
5. Zhong Z, Marcel TC, Hartmann FE, Ma X, Plissonneau C, Zala M, Ducasse A, Confais J, Compain J, Lapalu N (2017) A small secreted protein in Zymoseptoria tritici is responsible for avirulence on wheat cultivars carrying the Stb6 resistance gene. New Phytol 214 (2):619–631
6. Kema GHJ, Mirzadi Gohari A, Aouini L et al (2018) Stress and sexual reproduction affect the dynamics of the wheat pathogen effector AvrStb6 and strobilurin resistance. Nat Genet 50:375–380
7. Genissel A, Confais J, Lebrun M-H, Gout L (2017) Association genetics in plant pathogens: minding the gap between the natural variation and the molecular function. Front Plant Sci 8:1301
8. Talas F, McDonald BA (2015) Genome-wide analysis of Fusarium graminearum field populations reveals hotspots of recombination. BMC Genomics 16:996
9. Dalman K, Himmelstrand K, Olson Å, Lind M, Brandström-Durling M, Stenlid J (2013) A genome-wide association study identifies genomic regions for virulence in the non-model organism Heterobasidion annosum s.s. PLoS One 8:e53525
10. Gao Y, Liu Z, Faris JD, Richards J, Brueggeman RS, Li X, Oliver RP, McDonald BA, Friesen TL (2016) Validation of genome-wide association studies as a tool to identify virulence factors in Parastagonospora nodorum. Phytopathology 106:1177–1185
11. Tellier A, Moreno-Gámez S, Stephan W (2014) Speed of adaptation and genomic footprints of host–parasite coevolution under arms race and trench warfare dynamics. Evolution 68:2211–2224
12. Möller M, Stukenbrock EH (2017) Evolution and genome architecture in fungal plant pathogens. Nat Rev Microbiol 15(12):756–771
13. Poppe S, Dorsheimer L, Happel P, Stukenbrock EH (2015) Rapidly evolving genes are key players in host specialization and virulence of the fungal wheat pathogen Zymoseptoria tritici (Mycosphaerella graminicola). PLoS Pathog 11:e1005055
14. Schweizer G, Münch K, Mannhaupt G, Schirawski J, Kahmann R, Dutheil JY (2018) Positively selected effector genes and their contribution to virulence in the smut fungus Sporisorium reilianum. Genome Biol Evol 10:629–645
15. Stajich JE (2017) Fungal genomes and insights into the evolution of the kingdom.
16. de Jonge R, Bolton MD, Kombrink A, van den Berg GCM, Yadeta KA, Thomma BPHJ (2013) Extensive chromosomal reshuffling drives evolution of virulence in an asexual pathogen. Genome Res 23:1271–1282
17. Rouxel T, Grandaubert J, Hane JK et al (2011) Effector diversification within compartments of the Leptosphaeria maculans genome affected by repeat-induced point mutations. Nat Commun 2:202
18. Enkerli J, Bhatt G, Covert SF (1997) Nht1, a transposable element cloned from a

dispensable chromosome in Nectria haematococca. Mol Plant Microbe Interact 10 (6):742–749. https://doi.org/10.1094/MPMI.1997.10.6.742

19. Schirawski J, Mannhaupt G, Münch K et al (2010) Pathogenicity determinants in smut fungi revealed by genome comparison. Science 330:1546–1548
20. McDonald BA, Linde C (2002) Pathogen population genetics, evolutionary potential, and durable resistance. Annu Rev Phytopathol 40:349–379
21. Grandaubert J, Lowe RGT, Soyer JL et al (2014) Transposable element-assisted evolution and adaptation to host plant within the Leptosphaeria maculans-Leptosphaeria biglobosa species complex of fungal pathogens. BMC Genomics 15:891
22. Ma L-J, van der Does HC, Borkovich KA et al (2010) Comparative genomics reveals mobile pathogenicity chromosomes in Fusarium. Nature 464:367–373
23. Badouin H, Gladieux P, Gouzy J, Siguenza S, Aguileta G, Snirc A, Le Prieur S, Jeziorski C, Branca A, Giraud T (2016) Widespread selective sweeps throughout the genome of model plant pathogenic fungi and identification of effector candidates. Mol. Ecol. 26 (7):2041–2062
24. Branco S, Carpentier F, Rodríguez de la Vega RC et al (2018) Multiple convergent supergene evolution events in mating-type chromosomes. Nat Commun 9:2000
25. Stukenbrock EH, Dutheil JY (2018) Fine-scale recombination maps of fungal plant pathogens reveal dynamic recombination landscapes and intragenic hotspots. Genetics 208:1209–1229
26. Grandaubert J, Dutheil JY, Stukenbrock EH (2019) The genomic determinants of adaptive evolution in a fungal pathogen. Evolution Letters 3(3):299–312
27. Hubbard A, Lewis CM, Yoshida K, Ramirez-Gonzalez RH, de Vallavieille-Pope C, Thomas J, Kamoun S, Bayles R, Uauy C, Saunders DGO (2015) Field pathogenomics reveals the emergence of a diverse wheat yellow rust population. Genome Biol 16:1–15
28. Gladieux P, Condon B, Ravel S, Soanes D, Maciel JLN, Nhani A, Chen L, Terauchi R, Lebrun MH, Tharreau D (2018) Gene flow between divergent cereal-and grass-specific lineages of the rice blast fungus Magnaporthe oryzae. mBio 9:e01219–e01217
29. McMullan M, Rafiqi M, Kaithakottil G et al (2018) The ash dieback invasion of Europe was founded by two genetically divergent individuals. Nat Ecol Evol 2:1000–1008
30. Wicker T, Oberhaensli S, Parlange F et al (2013) The wheat powdery mildew genome shows the unique evolution of an obligate biotroph. Nat Genet 45:1092–1096
31. Duplessis S, Cuomo CA, Lin Y-C et al (2011) Obligate biotrophy features unraveled by the genomic analysis of rust fungi. Proc Natl Acad Sci 108:9166–9171
32. de Jonge R, Peter van Esse H, Maruthachalam K et al (2012) Tomato immune receptor Ve1 recognizes effector of multiple fungal pathogens uncovered by genome and RNA sequencing. Proc Natl Acad Sci U S A 109:5110–5115
33. Gladyshev E (2017) Repeat-induced point mutation (RIP) and other genome defense mechanisms in fungi. Microbiol Spectr 5(4). https://doi.org/10.1128/microbiolspec.FUNK-0042-2017
34. Wittenberg AHJ, van der Lee TAJ, Ben M'Barek S, Ware SB, Goodwin SB, Kilian A, Visser RGF, Kema GHJ, Schouten HJ (2009) Meiosis drives extraordinary genome plasticity in the haploid fungal plant pathogen Mycosphaerella graminicola. PLoS One 4:e5863
35. Habig M, Quade J, Stukenbrock EH (2017) Forward genetics approach reveals host genotype-dependent importance of accessory chromosomes in the fungal wheat pathogen Zymoseptoria tritici. MBio 8:e01919–e01917
36. Mehrabi R, Taga M, Kema GHJ (2007) Electrophoretic and cytological karyotyping of the foliar wheat pathogen Mycosphaerella graminicola reveals many chromosomes with a large size range. Mycologia 99:868–876
37. Plissonneau C, Hartmann FE, Croll D (2018) Pangenome analyses of the wheat pathogen Zymoseptoria tritici reveal the structural basis of a highly plastic eukaryotic genome. BMC Biol 16:5
38. Vlaardingerbroek I, Beerens B, Schmidt SM, Cornelissen BJC, Rep M (2016) Dispensable chromosomes in Fusarium oxysporum f. sp. lycopersici. Mol Plant Pathol 17:1455–1466
39. Moeller M, Habig M, Freitag M, Stukenbrock EH (2018) Extraordinary genome instability and widespread chromosome rearrangements during vegetative growth. bioRxiv 304915
40. Habig M, Kema G, Holtgrewe Stukenbrock E (2018) Meiotic drive of female-inherited supernumerary chromosomes in a pathogenic fungus. Elife 7:pii: e40251
41. Feurtey A, Stukenbrock EH (2018) Interspecific gene exchange as a driver of adaptive evolution in fungi. Annu Rev Microbiol 72:377–398
42. Brasier CM (2001) Rapid evolution of introduced plant pathogens via interspecific hybridization. Bioscience 51:123

43. Menardo F, Praz CR, Wyder S et al (2016) Hybridization of powdery mildew strains gives rise to pathogens on novel agricultural crop species. Nat Genet 48:201–205

44. Friesen TL, Stukenbrock EH, Liu Z, Meinhardt S, Ling H, Faris JD, Rasmussen JB, Solomon PS, McDonald BA, Oliver RP (2006) Emergence of a new disease as a result of interspecific virulence gene transfer. Nat Genet 38:953–956

45. Mcdonald MC, Ahren D, Simpfendorfer S, Milgate A, Solomon PS (2018) The discovery of the virulence gene ToxA in the wheat and barley pathogen Bipolaris sorokiniana. Mol Plant Pathol 19(2):432–439

46. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25(14):1754–1760. https://doi.org/10.1093/bioinformatics/btp324

47. Lunter G, Goodson M (2011) Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. Genome Res 21:936–939

48. Langmead B (2010) Aligning short sequencing reads with Bowtie. Curr Protoc Bioinforma Chapter 11:Unit 11.7

49. Li H (2011) A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 27:2987–2993

50. Garrison E, Marth G (2012) Haplotype-based variant detection from short-read sequencing. arXiv preprint arXiv:1207.3907.

51. DePristo MA, Banks E, Poplin R et al (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 43:491–498

52. McKenna A, Hanna M, Banks E et al (2010) The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20:1297–1303

53. Mielczarek M, Szyda J (2016) Review of alignment and SNP calling algorithms for next-generation sequencing data. J Appl Genet 57:71–79

54. Altmann A, Weber P, Bader D, Preuß M, Binder EB, Müller-Myhsok B (2012) A beginners guide to SNP calling from high-throughput DNA-sequencing data. Hum Genet 131:1541–1554

55. Pfeifer SP (2017) From next-generation resequencing reads to a high-quality variant data set. Heredity 118:111–124

56. Plissonneau C, Stürchler A, Croll D (2016) The evolution of orphan regions in genomes of a fungal pathogen of wheat. MBio 7: e01231-16

57. Faino L, Seidl MF, Datema E, Berg GC, Janssen A, Wittenberg AH (2015) Single-molecule real-time sequencing combined with optical mapping yields completely finished fungal genome. MBio. https://doi.org/10.1128/mBio.00936-15

58. Bankevich A, Nurk S, Antipov D et al (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol 19:455–477

59. Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y (2012) SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. Gigascience 1(1):18

60. Peng Y, Leung HCM, Yiu SM, Chin FYL (2012) IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics 28:1420–1428

61. Pevzner PA, Tang H, Waterman MS (2001) An Eulerian path approach to DNA fragment assembly. Proc Natl Acad Sci 98:9748–9753

62. Zerbino DR, Birney E (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 18:821–829

63. Chaisson M, Pevzner P, Tang H (2004) Fragment assembly with short reads. Bioinformatics 20:2067–2074

64. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM (2016) Canu: scalable and accurate long—read assembly via adaptive k—mer weighting and repeat separation, pp 1–35

65. Chin C-S, Peluso P, Sedlazeck FJ et al (2016) Phased diploid genome assembly with single-molecule real-time sequencing. Nat Methods 13:1050–1054

66. de Lannoy C, de Ridder D, Risse J (2017) The long reads ahead: de novo genome assembly using the MinION. F1000Research 6:1083

67. Datema E, Hulzink RJM, Blommers L, et al. (2016) The megabase-sized fungal genome of Rhizoctonia solani assembled from nanopore reads only. bioRxiv 084772

68. Faino L, Seidl MF, Shi-Kunne X, Pauper M, van den Berg GCMM, Wittenberg AHJJ, Thomma BPHJHJ (2016) Transposons passively and actively contribute to evolution of the two-speed genome of a fungal pathogen. Genome Res 26:1091–1100

69. Miller ME, Zhang Y, Omidvar V et al (2018) De novo assembly and phasing of dikaryotic

genomes from two isolates of Puccinia coronata f. sp. avenae, the causal agent of oat crown rust. mBio 9

70. Antipov D, Korobeynikov A, McLean JS, Pevzner PA (2016) HybridSPAdes: an algorithm for hybrid assembly of short and long reads. Bioinformatics 32:1009–1015
71. Walker BJ, Abeel T, Shea T et al (2014) Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One 9:e112963
72. Dutreux F, Da Silva C, Couloux A et al (2018) De novo assembly and annotation of three Leptosphaeria genomes using Oxford Nanopore MinION sequencing. Sci Data 5:180235
73. Gurevich A, Saveliev V, Vyahhi N, Tesler G (2013) QUAST: quality assessment tool for genome assemblies. Bioinformatics 29:1072–1075
74. Feurtey A, Stevens DM, Stephan W, Stukenbrock EH (2019) Interspecific gene exchange introduces high genetic variability in crop pathogen. Genome Biol Evol. In Press
75. Blanchette M, Kent WJ, Riemer C et al (2004) Aligning multiple genomic sequences with the threaded blockset aligner. Genome Res 14:708–715
76. Angiuoli SV, Salzberg SL (2011) Mugsy: fast multiple alignment of closely related whole genomes. Bioinformatics 27:334–342
77. Darling AE, Mau B, Perna NT (2010) Progressivemauve: multiple genome alignment with gene gain, loss and rearrangement. PLoS One 5(6):e11147. https://doi.org/10.1371/journal.pone.0011147
78. Katoh K, Asimenos G, Toh H (2009) Multiple alignment of DNA sequences with MAFFT. Bioinforma DNA Seq Anal 537:39–64
79. Notredame C, Higgins DG, Heringa J (2000) T-coffee: a novel method for fast and accurate multiple sequence alignment. J Mol Biol 302:205–217
80. Dutheil JY, Gaillard S, Stukenbrock EH (2014) MafFilter: a highly flexible and extensible multiple genome alignment files processor. BMC Genomics 15:53
81. Pang AW, MacDonald JR, Pinto D et al (2010) Towards a comprehensive structural variation map of an individual human genome. Genome Biol 11:R52
82. Wala JA, Bandopadhayay P, Greenwald NF, et al (2018) SvABA: genome-wide detection of structural variants and indels by local assembly. bioRxiv 105080. https://doi.org/10.1101/gr.221028.117
83. Escaramís G, Docampo E, Rabionet R (2015) A decade of structural variants: description, history and methods to detect structural variation. Brief Funct Genomics 14:305–314
84. Ye K, Schulz MH, Long Q, Apweiler R, Ning Z (2009) Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics 25:2865–2871
85. Rausch T, Zichner T, Schlattl A, Stutz AM, Benes V, Korbel JO (2012) DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics 28: i333–i339
86. Layer RM, Chiang C, Quinlan AR, Hall IM (2014) LUMPY: a probabilistic framework for structural variant discovery. Genome Biol 15: R84
87. Tattini L, D'Aurizio R, Magi A (2015) Detection of genomic structural variants from next-generation sequencing data. Front Bioeng Biotechnol *3*:92. https://doi.org/10.3389/fbioe.2015.00092
88. Sedlazeck FJ, Lee H, Darby CA, Schatz MC (2018) Piercing the dark matter: bioinformatics of long-range sequencing and mapping. Nat Rev Genet 19:329–346
89. Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP (2014) Sequencing depth and coverage: key considerations in genomic analyses. Nat Rev Genet 15:121–132
90. Soderlund C, Bomhoff M, Nelson WM (2011) SyMAP v3.4: a turnkey synteny system with application to plant genomes. Nucleic Acids Res 39:e68
91. Krzywinski M, Schein J, Birol İ, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA (2009) Circos: an information aesthetic for comparative genomics. Genome Res 19:1639–1645. https://doi.org/10.1101/gr.092759.109
92. Nattestad M, Schatz MC (2016) Assemblytics: a web analytics tool for the detection of variants from an assembly. Bioinformatics 32:3021–3023
93. Liu S, Huang S, Rao J, Ye W, Krogh A, Wang J (2015) Discovery, genotyping and characterization of structural variation and novel sequence at single nucleotide resolution from de novo genome assemblies on a population scale. Gigascience 4(1):64
94. McDonald BA, Stukenbrock EH (2016) Rapid emergence of pathogens in agro-ecosystems: global threats to agricultural sustainability and food security. Phil Trans R Soc B 371:20160026