CellPress

Review

Genome Mapping in Plant Comparative Genomics

Lindsay Chaney,[1] Aaron R. Sharp,[1] Carrie R. Evans,[1] and Joshua A. Udall[1,@,*]

Genome mapping produces fingerprints of DNA sequences to construct a physical map of the whole genome. It provides contiguous, long-range information that complements and, in some cases, replaces sequencing data. Recent advances in genome-mapping technology will better allow researchers to detect large (>1 kbp) structural variations between plant genomes. Some molecular and informatics complications need to be overcome for this novel technology to achieve its full utility. This technology will be useful for understanding phenotype responses due to DNA rearrangements and will yield insights into genome evolution, particularly in polyploids. In this review, we outline recent advances in genome-mapping technology, including the processes required for data collection and analysis, and applications in plant comparative genomics.

Trends

Genomic structural variations (large DNA rearrangements, such as insertions, deletions, duplications, inversions, and translocations) can lead to phenotypic differences.

Genome mapping images individual DNA molecules, fluorescently labeled at restriction enzyme recognition sites, to create an ordered barcode hundreds of kilobase-pairs long. These barcodes are used to assemble a map that spans the genome.

Due to the continuity of information in genome mapping, it is an ideal tool for use in plants that commonly contain highly repetitive genomic regions and differences in genome sizes.

Recently, advances in genome-mapping technology have made this resource more widely available. Improved optics, advanced molecular biology, informatics tools, and creative innovations have been combined to create relatively low-cost mapping tools.

Origins of Genome Mapping

Despite advances in creating new genomic tools, in some cases revisiting old approaches to scientific questions can be fruitful. This retrospective strategy brings to mind the title of a popular show tune, 'Everything old is new again' [1]. In the case of genome mapping, something old is indeed new again. For many years, cytogeneticists looked at banding patterns of condensed chromosomes and made significant deductions and contributions to our understanding of plant genome organization. In recent years, improved optics, advanced molecular biology, and creative innovations have been combined to create higher-throughput genomic tools that have roots in, and similarities to, many older cytogenetic methods. These strategies produce maps of large individual DNA molecules.

One reason why these long-molecule maps are receiving attention is because of their ability to complement genome sequencing. The relative ease of genome sequencing often overshadows its shortcomings: a puzzle with many small pieces is difficult to solve without additional, long-range information. For example, genome maps can be combined with sequence assemblies comprising numerous scaffolds and contigs, in which case they provide the necessary structure for joining contigs and improving the *de novo* assembly of plant genomes. Aside from *de novo* genome assembly, **genome mapping** (see Glossary) provides some unique research opportunities for comparative plant genomics, which were previously closed because short sequencing reads cannot detect certain large structural variations. Here, we review genome mapping, including its limitations and capacities, and explore some of its potential applications in the field of plant comparative genomics.

Comparing Plant Genomes

Comparative plant genomics examines the similarities of, and differences in, genomes between plant species. By comparing genomes of evolutionarily divergent species, we can better understand the patterns and processes that underlie plant genome evolution as well as uncover functional regions of genomes [2]. **Structural variations** are large (>1 kbp in size)

[1]Plant and Wildlife Sciences Department, Brigham Young University, Provo, UT 84602, USA

*Correspondence: jaudall@byu.edu (J.A. Udall).
@Twitter: @udall

rearrangements of DNA that include insertions, deletions, duplications (also referred to as copy number variations), inversions, and translocations (Figure 1) [3]. These genomic alterations are an important source of genetic and phenotypic diversity. For example, structural variations in plants have been associated with stress tolerance, disease resistance, domestication, increase in yields, leaf size, fruit shape, reproductive morphology, adaptation, and speciation [4–8]. Through the use of **cytogenetics**, researchers have been able to identify large chromosomal changes (e.g., translocations, aneuploidy, and loss of repeats) [9–13], yet this method is labor

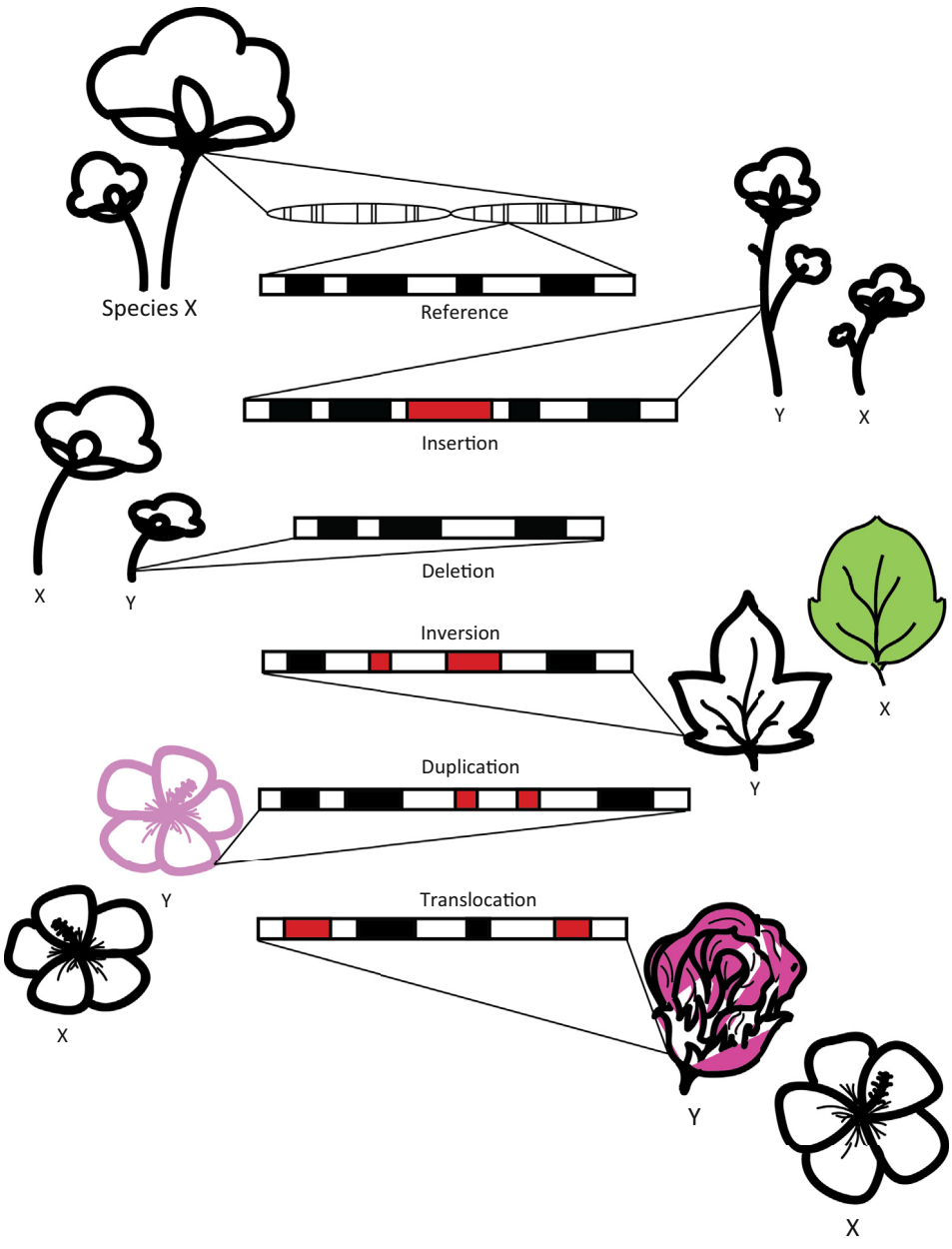

Figure 1. Conceptual Representation of Different Genomic Structural Variations to a Single Region of the Reference Genome. Structural variations are large (>1 kbp) rearrangements of DNA that frequently result in phenotypic differences. These variations include insertions, deletions, inversions, duplications, and translocations. By comparing genomes of different species, large chromosomal changes can be identified.

Glossary

Consensus genome map: a large genomic region represented by a set of contigs. Each contig comprises the total length and distribution of restriction enzyme recognition sites. A consensus map is constructed from single-molecule maps that share compatible distance patterns and, therefore, are likely to represent the same place in a genome.

Cytogenics: microscopic examination of chromosomes. One common molecular cytogenetic method is fluorescent *in situ* hybridization (FISH), where metaphase chromosomes are hybridized with a fluorescently labeled DNA probe. This method can examine a specific DNA sequence and can detect large DNA rearrangements ranging between a few mbp to 1 gbp.

Genome mapping: a variant of optical mapping commercialized as the BioNano lyrs system that uses modified restriction enzymes, fluorescent nucleotide incorporation, and automated imaging in parallel nanochannel arrays to increase the rate of data generation over traditional optical mapping.

High molecular weight (HMW) DNA: DNA molecules that are between 50 kbp and 2 mbp long. The length of the DNA molecules allows them to span large regions of the genome that are typically difficult to resolve with short read sequencing.

Molecule map: a single DNA molecule characterized by an ordered list of distances between restriction enzyme recognition sites. Also referred to as *rmaps*, these individual molecules are combined and/or assembled to construct consensus genome maps.

Optical map: a method that uses microscopic imaging to produce ordered restriction enzyme recognition site maps from a single linearized DNA molecule. Optical mapping allows detection of DNA with a resolution of 1 kbp to several mbp.

Physical map: a map of the physical distances between identifiable landmarks in DNA, generally produced using restriction enzyme digestion of bacterial artificial chromosomes (BACs). By contrast, genetic maps depict relative positions of loci based on the degree of recombination.

intensive, prone to error, and generally only captures differences with limited resolution. Thus, cytogenetic methods may greatly underestimate the number of diverse changes in architecture that are in fact found between plant genomes.

As sequencing technologies have become more accessible, the field of comparative genomics has greatly expanded our knowledge of plant genome structure. With the high throughput and low cost of next-generation sequencing (NGS), more than 100 plant genomes have been sequenced [14]. Although short-read sequencing is useful for detecting small-scale sequence variations (e.g., <1 kbp in size or a few nucleotides), it is unable to detect most large-scale structural variations [15,16]. Plant genomes are notorious for being large and highly repetitive, and many contain multiple copies of entire chromosomes (e.g., polyploidy) [8,17,18]. Most sequencing techniques are effective at detecting deletions, but have difficulty resolving sequence redundancies owing to short reads typical of NGS. Thus, sequencing advances alone may not provide sufficient resolution for comparing the organization and structure of genomes from evolutionarily divergent species [19].

Structural variant: DNA rearrangements of 1 kbp or more (Figure 2, main text). These include insertions, deletions, duplications (collectively called copy number variations), translocations, and inversions. Structural variations differ from their short counterparts, sequence variants, such as single-nucleotide polymorphisms and small (<1 kbp) insertions or deletions (indels).

Mapping Genomes

Physical Maps

One method to compare genome structures is **physical maps** [19]. Just as a cartographer would start with key landmarks when mapping a region, then later fill in the details of the location, physical maps provide molecular anchor points to link sequence contigs, bridge repetitive regions, and give a course-grain view of genome structure [20]. Physical maps have been key in completing high-quality genome assemblies (e.g., [21–24]) and are typically made using large insert clone libraries, such as bacteria artificial chromosomes (BACs). Although BAC-based physical maps are helpful in the completion of *de novo* genome sequencing, their widespread use in plant comparative genomics has been limited because they are expensive and time consuming, and require a great deal of experimental expertise. Additionally, BAC libraries are subject to clone amplification biases resulting in incomplete coverage, and some regions of BAC physical maps can be difficult to resolve due to the sequence redundancy typically found in plant genomes [25,26].

Optical Maps

An **optical map** is an ordered genome-wide physical map constructed from unamplified DNA molecules. A unique 'fingerprint' or 'barcode' is created by mapping the location of restriction enzyme recognition sites present in a long DNA molecule. Optical mapping has advantages over other genomic technologies because it uses long DNA molecules that are not cloned or amplified and preserves the order of restriction enzyme recognition sites. This allows a map to be created that accurately reflects long repetitive regions, and is free from cloning or amplification bias. It also allows the map to resolve complicated genome regions, including copy number variations and potentially homoeologous segments from polyploid genomes, more efficiently and unambiguously than unordered restriction fragment maps. Although this system has been useful in improving plant genome assemblies [27–31], it was initially applied to small genomes (e.g., fungi and bacteria) [32–36]. The low throughput of traditional optical mapping makes it difficult to use in large-scale plant comparative genomics projects (Box 1).

Genome Mapping

Through advances in labeling, imaging, automation, and nanofabrication, a higher-throughput mapping system has recently been developed. This mapping system has been commercialized by BioNano Genomics as the Irys platform (Box 1). Its capabilities to capture 50–200 gbp of data per day have led to its increasing popularity among researchers. The relative ease of quickly mapping large genomes at high coverage to identify structural variations with or without a

Box 1. Evolution of Mapping Technologies

Optical mapping was pioneered by the Schwartz Lab [32] during the mid-1990s. This system, upon which the OpGen Argus platform is built, immobilizes linearized DNA (~150 kbp–2 mbp) onto a charged imaging surface. DNA is cleaved by a sequence-specific restriction enzyme and stained with an intercalating dye, such as YOYO-1. DNA fragments are imaged, sized, and analyzed to create an ordered restriction map, where restriction sites are identified as gaps in a DNA segment. Although effective, these traditional methods are complex and have low throughput, primarily due to inefficiencies in extending the DNA, imaging, and data analysis. As a result, this method has primarily been used to identify pathogens or map small-genome organisms, such as bacteria, yeast, and other fungi [32–36]. The use of nanoconfinement DNA elongation techniques and the automation of imaging and data processing increased throughput, allowing the mapping of larger genomes [73]. In 2010, data for a genome-wide optical map of the human genome were collected at a rate of ~ 5 gbp per day [49].

The BioNano Genomics Irys system has improved throughput 10–40 times that of traditional optical mapping, capturing 50–200 gbp of data per day. This system uses a modified restriction endonuclease or nickase (e.g., Nt.BspQI) that introduces single-strand breaks or nicks in a DNA molecule at a sequence-specific recognition site (Figure 2, main text). The breaks are then labeled using a DNA polymerase enzyme to incorporate fluorescent nucleotide analogs at the nick sites. The DNA backbone is stained to allow for accurate measurement of DNA molecules. An automated system uses electrophoresis to pull DNA into massively parallel arrays of linear nanochannels (13 000). In the nanochannels, DNA is immobilized and imaged, then new DNA molecules are pulled into the nanochannels for imaging. The use of microfluidic nanochannels allows hundreds of DNA molecules to flow through the field-of-view of the microscope and to be mapped rapidly in parallel. These differences in throughput can have a significant impact on experimental time; for example 50× coverage of a human genome was completed in about a month using OpGen by Teague *et al.* [49], while generating the same coverage with the Irys system took just over a day [74]. Nevertheless, Opgen has improved automation: in 2013, improvements enabled the goat genome to be assayed at ~10 gbp/h [46]. For a more detailed review of the different optical mapping technologies, see [60,72,75].

reference sequence assembly suggests that this system has potential for use in plant comparative genomics. In this review, we focus on data collection and analysis on the recently developed Irys platform and its use in comparing plant genomes.

Data Collection

The first step to creating a genome map is to collect high-quality data. Data quality is partly determined by individual molecule length and the accuracy of the measured distances between nick-repaired florescent labels. The use of **high-molecular-weight (HMW) DNA** allows a single DNA molecule to cover long genomic regions, often spanning problematic regions of the genome that have been difficult to resolve using NGS. However, physical isolation of long, high-quality DNA in plants can be challenging due to the variety of organic compounds plant tissues contain. Plants harbor large amounts of polyphenols, polysaccharides, and other secondary metabolites, which normally aid in functions such as plant defenses, but can also contaminate DNA for laboratory purposes [37]. Different species, different parts of a plant, and even the same plant at different development stages can all vary in their chemical composition, making protocols problematic to generalize. One way to avoid excess secondary metabolites is to use very young tissue; new tissue has the highest DNA content:mass ratio possible, and the lowest level of metabolites. Dark treating seedlings for 24–72 h significantly reduces the amount of polysaccharides and has shown favorable results [38].

Although only a small amount of DNA is needed (as little as 300 ng), isolating HMW DNA from plants can be challenging and methods are not as well established as conventional DNA extractions. The isolation of intact nuclei can keep the DNA from shearing after the disruption of the cell wall. Isolated nuclei can also be cleaned extensively to remove contaminants that are generally in the cytoplasm [39]. Contaminants are removed through use of a strong reducing agent (e.g., 2-mercaptoethanol), detergents (e.g., Triton-X 100), and polymers [e.g., polyvinylpyrrolidone (PVP)] that prevent the oxidation of polyphenols, solubilize lipids and enzymes, and bind polyphenols, all of which can reduce DNA quality [38–41]. Furthermore, a Percoll density gradient is used to help separate nuclei from particles of a different density [42]. After isolation, nuclei are washed several times before being embedded in low-melting point agarose plugs

before membrane lysis. The physical matrix of agarose plugs allows the naked DNA to be further treated with RNase and proteinases, and to be washed to remove contaminants and residual reagents without excessive physical shearing of the DNA. Extraction of DNA from the plugs requires melting the agarose plug and subsequent drop dialysis, which further removes low-molecular-weight contaminants and salts, as well as concentrating the DNA [43].

Once enough sufficiently pure DNA has been isolated, the actual mapping can begin. Sequence-specific, single-strand breaks or nicks are introduced into the DNA with a modified restriction enzyme or nickase (Figure 2). A polymerase then incorporates fluorescent nucleotide analogs at the break sites (Box 1). Labeled DNA is loaded onto a nanofluidic chip, where an automatically applied electric field draws iterative samples of DNA through a series of columns for linearization, and then into nanochannels for imaging [44]. Electric currents are applied in such a way that the

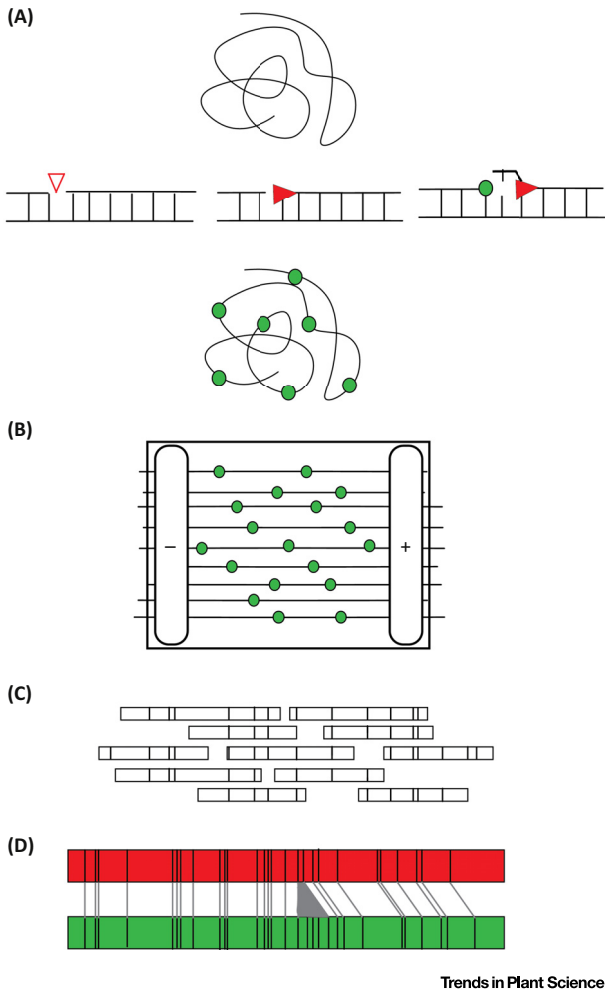

Figure 2. The Workflow of Data Collection and Analysis Using the Irys System to Create Genome Optical Maps. (A) Data collection: high-molecular-weight (HMW) DNA is extracted and a single-stranded break or nick is introduced at a sequence-specific recognition site on individual DNA molecules. DNA is fluorescently labeled at the sequence-specific site using a DNA polymerase enzyme and nucleotide. The DNA backbone is stained to allow for accurate measurement of the DNA molecule. (B) On a specialized chip, automated electrophoresis pulls DNA into arrays of nanochannels where linearized DNA is imaged. (C) Data analysis: each imaged molecule is digitally measured for length and distances between labeled sites to create a molecule map. Molecule maps that overlap and whose distance patterns match, are assembled into a consensus contig. (D) Consensus contig maps from different plant samples are then compared to identify large structural variations, such as an insertion in the green contig depicted here.

large pool of HMW DNA is repeatedly sampled through a different series of run cycles. Once labeled DNA from the first cycle is positioned in the channel, fragments are imaged through an automated system on the Irys instrument [45]. A series of raw images are converted to single **molecule maps**, digital measurements of molecule length and intensity, as well as physical distances between and intensity of incorporated labels (Figure 2).

Data Analysis

Creating a High-Quality Consensus Map

Once data have been collected, single-molecule maps are assembled into consensus map that spans a large genomic region. Each imaged molecule is characterized by its total length estimate and a linear series of fluorescently labeled nick sites that represent physical distances between endonuclease recognition sites (Figure 2), and each 'fragment pattern' matches a distinct region of the genome. These distinct series of kilobase-sized distances are analogous to fragments from digested BACs, except that they are already arranged in linear order. Molecule maps that overlap are identified through a heuristic alignment algorithm that first matches partial distance patterns. Consensus contigs are created using an overlap-layout-consensus algorithm (Figure 2). The end result of the assembly process is a set of contigs with unique distance patterns, each of which represents a certain region of a chromosome within the plant genome, and is referred to as a **consensus genome map**. Typical of other genomic approaches, contigs representing intact entire chromosomes are not usually achieved; however, a low number of long contigs that match the expected genome size and chromosome number would suggest complete assembly.

There are several factors that may impact assembly quality, including DNA quality, nick efficiency, imaging artifacts, and genome complexity. Common errors in these data include: (i) inaccurate sizing of molecules due to non-uniform fluorescent staining or stretching; (ii) spurious enzyme cut sites due to random breakage of the DNA molecule or star activity (false-positive label sites); and (iii) missing label sites due to missing enzyme cut sites, incomplete digestion, or labeling errors (false-negative label sites). Traditional optical mapping, OpGen, also has the added error of small fragment loss [46]. While careful laboratory techniques aim to minimize the impact of these factors, they cannot be entirely removed. Thus, part of the assembly quality depends on the effectiveness of the assembly algorithm at compensating for noise in the input data.

Algorithms and Resources

To align single-molecule maps into a consensus genome map, dynamic programming algorithms are used to account for the inherent error characteristics of the molecule maps. Many of the methods and approaches used by common DNA sequence assembly programs are not useful for imaged HMW molecules due to differences in how data are generated (e.g., not amplified, single 'dimension' of fragment length, etc.) [47]. Compared with the analytical advances made in sequencing data, there are relatively few methods that exist for analyzing and utilizing genome map data. There have been a series of software tools developed specifically for the BioNano Irys system to improve mapping quality. The IrysSolve software addresses noise in data by allowing users to customize many input parameters that describe the error profile for their data, such as false positive and false negative nicks, molecule stretch, and florescence intensity, which can be estimated empirically using a genome sequence assembly [48]. The algorithm then makes compensatory decisions based on those input parameters. Furthermore, IrysSolve has been developed to detect a variety of structural variants in large genomes (e.g., human [49]) (Figure 2). However, to date, there has only been experimental validation of its ability to detect insertions and deletions; other structural variants, such as inversions or translocations, have yet to be validated. Sharp *et al.* developed a method that selects the best input parameters by running multiple assemblies with permutations of input parameters with minimal resource

(computing) usage [50]. Tools developed by Shelton *et al.* primarily focus on complementing genome sequence assembly. One tool maps a subset of the molecules to an *in silico* digest of a reference sequence assembly at a variety of error profiles, selecting the profile that maximizes the mapping efficiency [51]. Additionally, they present software called Stitch that automatically parses and interprets the output from a comparison between a consensus map and a reference genome to super scaffold a sequence assembly [51]. Another piece of software, ALLMAPS, performs a similar computational task as Stitch to link map data with draft genome sequence assemblies, although it is not written specifically for genome map data [52].

Most other software are not written specifically for the Irys genome-mapping system, but can be used on both traditional optical-mapping and Irys genome-mapping data. Some existing algorithms, which were originally developed for small genomes, are unable to be ported to large genome assemblies because of computational limitations. For example, Gentig, a proprietary software, has been able to successfully assemble small consensus optical maps, but is unable to scale to large genomes [53]. Opgen's MapSolver software allows for map visualization, comparisons between maps or between a map and an *in silico* digest of a sequence assembly, and aids in assembly improvement, but only on genomes up to 100 mbp, ruling out most plants. Its Genome-Builder software uses an iterative Bayesian maximum-likelihood, a modified Smith-Waterman dynamic algorithm, plus heuristic filtering process. This software is capable of performing assembly improvement (super-scaffolding) but does not generate a consensus map; neither does it facilitate direct comparisons between optical maps from two plants [46].

There have been several programs that have the computational feasibility for large genome map assemblies (for a detailed review, see [54]). For example, Valouev *et al.* developed an algorithm that is able to align two optical maps and also align an optical map to a reference map [47]. This is the first algorithm capable of producing accurate maps of large genomes in a feasible timeframe. SOMA and TWIN are both open source software that align optical map data to a reference sequence, but the latter is highly sensitive to false positive and false negative label sites in the input data [55,56]. AGORA, is a proof of concept that optical map data can be used to constrain a de Bruijn graph used for genome sequence assembly, which was successfully used to resolve highly repetitive regions in the genome [57].

Utilizing Optical Map Assemblies

Once high-quality consensus maps have been created, several downstream analyses may be performed. Current applications of genome mapping have primarily been used to improve or validate sequence assemblies (e.g., to improve the resolution of contigs from BAC pools of wheat 7D short arm [58] and a one of the most contiguous *de novo* assemblies of a human genome [59]). However, genome mapping could be used to replace several well-established but low-throughput technologies for comparative genomics (e.g., [9,10]); furthermore, it is superior to sequencing to detect large structural variants because of large input molecules (Box 2). The use of genome-mapping comparisons have not been fully utilized in plant genomes due to historically high costs [60].

Structural Variations between Species

Genome mapping has the potential to be used for comparing structural variations between species. Structural variations are thought to be the major contributors of phenotypic variation in plants, leading to an increased focus on characterizing structural variations between plant genomes [61]. One example of the potential use of optical mapping to compare structural variation across plant species is in *Brassica napus*. Segregating populations of *B. napus* doubled haploid lines and codominant RFLP markers detected pairs of homoeologous loci on N7 and N16 for which the annual and biennial parents had identical alleles in regions expected to be homoeologous [62]. High-throughput genome mapping could replace labor-intensive

Box 2. The Length of Genomic Technologies

Genomic technologies are rapidly evolving. Differences in each technology stem from variations in biochemistry and technical aspects that contribute to differences in read length (i.e., the length of continuous DNA it is able to accommodate), a major determinant of the utility of each technology.

The first major player in genome sequencing was the Sanger technology. This technology is expensive, time consuming, and labor intensive, but its use resulted in some remarkable genome-sequencing achievements [76]. It has maximum read lengths of approximately 1 kbp [77]. With the advent of NGS technologies, an enormous volume of genomic data was able to be produced cheaply and quickly using short reads, realizing the ability to sequence whole genomes of numerous organisms. A variety of NGS platforms exist (e.g. Illumina, Roche 454, SOLiD, and Helicos) with differences in read length (from 10 to 500 bp), accuracy, speed, and cost that result in advantages with respect to each specific application (reviewed in [78,79]).

Now emerging are long-read sequencing technologies, also known as the third generation of sequencers (Figure I). These new single-molecule sequencing technologies can produce average read lengths exceeding 10 kbp (and >100 kbp maximum), spanning greater genomic distances than NGS. These technologies include PacBio SMRT, Illumina TruSeq Synthetic Long Reads, and the Oxford Nanopore MinION. Two other technologies, 10X Genomics and Dovetail, take an alternative approach to long-read sequencing. Through the use of microfluidic sorting, they produce jumping mate-pair libraries of short reads connected along longer blocks. Long-read technologies also face the challenge of extracting high-quality HMW DNA from plants, as we have discussed for genome mapping. More detailed reviews of the long-read technologies are available elsewhere [80–82].

Genome mapping using the Irys platform is able to produce the longest map lengths, yet at the lowest resolution. Genome mapping is able to determine the large-scale sequence structure of DNA but does not sequence every base: it can identify the presence and the location of structural variations between genomes, but it fails to distinguish the exact sequence and breakpoint. Genome mapping is best used for questions that require a large genomic picture and do not need high resolution. Each of these evolving technologies has its own unique list of pros and cons, making no one platform a perfect technology to use in answering every research question.

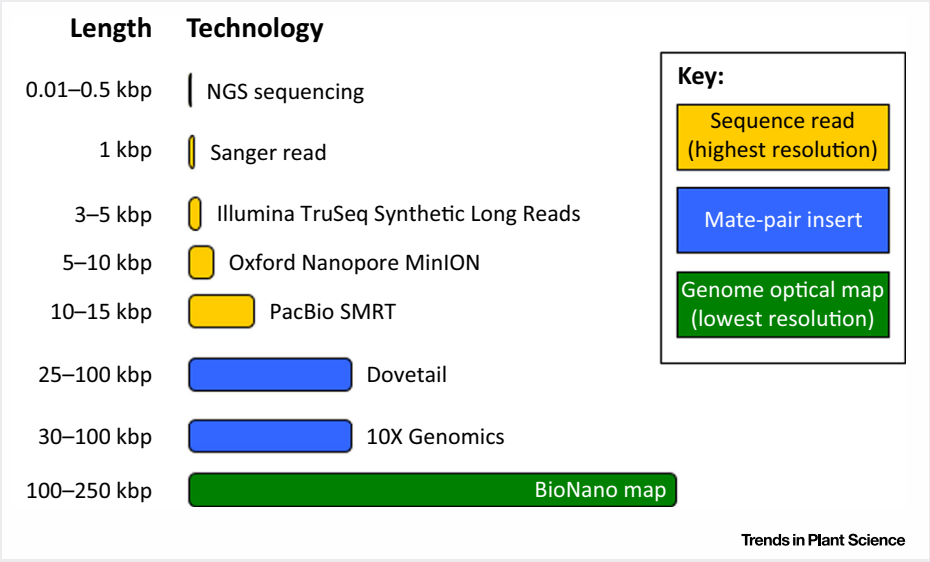

Figure I. A Comparison of Average Read Lengths, Mate-Pair Distances, and Map Lengths Commonly Obtained Using Different Genomic Technologies. The sequencing provides the highest resolution (yellow), yet short reads have trouble spanning repetitive regions and detecting large rearrangements. The long maps produced by optical genome mapping (green) are able to span large structural variations, but resolution is limited to 1 kbp to several mbp. Abbreviation: NGS, next-generation sequencing.

RFLP or *in situ* hybridization experiments to more quickly uncover such large genomic rearrangements. Analysis of intentionally derived genetic stocks, for example the wheat nullisomic-tetrasomic lines or chromosome arm deletion stocks, have been used to physically locate the position of genes on chromosome arms [63]. Genome mapping could be used on these same genetic stocks to further identify breakpoints and missing genomic regions.

Understanding the Evolution of Genomes in Polyploids

Polyploidy, the doubling of all the chromosomes in a cell, is ubiquitous in the evolution of plant species. Most, if not all, angiosperm species have gone through multiple rounds of polyploidization [64]. At the onset of polyploidization, a period of rapid genomic reorganization and massive gene loss occurs and structural variations arise [65]. Structural variations can also arise through local duplication events and the activity of transposons, resulting in the differential loss of genes between lineages [8]. Little is known about how long chromosomal variation may persist and how it might influence the establishment and evolution of polyploids in nature [66,67]. Genome mapping could be used to characterize chromosomal composition before and after polyploidization events. For example, cultivated cotton is an allotetraploid that evolved following a polyploidization event involving two diploid cottons (the A- and D-genomes) 1–2 million years ago [68]. The cotton system serves as an excellent model for identifying structural variations between species (Figure 1), more specifically, examining nonreciprocal homoeologous recombination, intergenomic spread of transposable elements, and alterations and biases in gene duplicate expression. Research has shown some evidence of chromosomal rearrangements between the homoeologous genomes in polyploids [65,67,69], but it is still not completely understood how genome variations compare across different cotton species and genomes. Genome mapping could be useful in pinpointing segmental losses and exchanges among homoeologous chromosomes, which are important aspects of polyploidy genome evolution [60].

Use in Crop Improvement

One common use of comparative genomics is in plant breeding for crop improvement. Alleles that differ between lines can be correlated with favorable agronomic traits. While it is theoretically feasible to incorporate genome mapping for structural variant genotyping into this system, there are some practical problems that make the idea less tenable. The coverage required for structural-variant calling using map fingerprints is likely lower than that required for whole-genome assembly, but it is unclear what the required coverage would be. Also, the number of different lines that are normally required for a large-scale breeding project would currently be prohibitive, because genome-mapping technology has not yet been optimized to run several samples concurrently. However, the future of crop improvement will likely be centered on a deep functional understanding of the genome of a species, including structural variants [70], in which case genome mapping could help address key biological questions for crop improvement.

Concluding Remarks and Future Perspectives

Through the use of genome mapping, large gains in the field of plant comparative genomics are likely. The long-range information (single molecules spanning hundreds of kbp) that is able to span complex genomic regions will greatly improve our understanding of relations between different genomes. Furthermore, the ability to capture large quantities of data in a relatively short time frame and at low cost will allow researchers to compare whole genomes of multiple plant species with relative ease. These qualities of genome mapping make it an extremely useful tool in situations where a low-resolution genomic picture is sufficient, such as identifying structural variations between plant species and identifying phylogenetic patterns in genome evolution (Box 2).

However, for this technology to reach its full potential, obstacles must be overcome. HMW DNA extractions from plants continue to be challenging, and software tools will need to be further developed to overcome the inherent errors in the data. Future developments to genome mapping include multicolored labeling that will allow the recognition of multiple sequence motifs in a single sample [58]. The ability to map the epigenome through labeling DNA methylation will allow comparison of the genetic and epigenetic composition of genomes [71]. If genome mapping is to complement DNA sequencing, parallelizing data collection by both technologies is required [72].

Outstanding Questions

What improvements can be made to data collection protocols and data analysis algorithms for genome mapping?

Are restriction enzyme recognition motif distributions, as detected by genome mapping, sufficiently conserved in evolutionary time to make effective comparisons between highly diverged taxa?

To what extent is phenotypic variation and speciation explained by structural variation?

To what extent do the structural variations caused by genome instability contribute to allopolyploidy establishment and early evolution?

At what rate are structural variations formed and fixed in a population? How does this rate vary by taxa, mating system, and environment?

Do structural variations contain phylogenetic signals?

Past plant comparative genomic studies have investigated differences in genome size, gene number, transposable elements, and syntenic relations, yet their methods underestimate the diverse architecture found in plant genomes [2]. Many studies that address changes in genomic content focus on single-nucleotide polymorphisms or short indels as markers of association genetics, yet this research largely ignores the large structural variations that often have significant impacts. Direct comparisons of large genomic structural variations have so far been lacking in plants, and genome mapping shows great promise for revealing genomic regions that are not easily accessible through conventional sequencing methods. Genome mapping could become an integral tool in the study of plant domestication, polyploid evolution, and trait development [60] (see Outstanding Questions).

Author contributions

L.C. wrote the manuscript. A.R.S. contributed to manuscript writing. L.C., A.R.S., and J.A.U. conceptualized the ideas and framework for the manuscript. C.R.E. produced Figures 1 and 2. J.A.U. provided supervision and contributed to manuscript writing. All authors read and approved the manuscript.

Acknowledgments

We thank members of the Udall Lab for feedback on this manuscript. We appreciate comments on optical mapping from J. Mudge. This work was supported by National Science Foundation Plant Genome Research Program (Award #1339412) and by Cotton Incorporated.

References

1. Allen, P. (1974) *Everything Old is New Again.* A&M Recods
2. Caicedo, A.L. and Purugganan, M.D. (2005) Comparative plant genomics. Frontiers and prospects. *Plant Physiol.* 138, 545–547
3. Gaut, B.S. *et al.* (2007) Recombination: an underappreciated factor in the evolution of plant genomes. *Nat. Rev. Genet.* 8, 77–84
4. Lowry, D.B. and Willis, J.H. (2010) A widespread chromosomal inversion polymorphism contributes to a major life-history transition, local adaptation, and reproductive isolation. *PLoS Biol.* 8, e1000500
5. Rieseberg, L.H. (2001) Chromosomal rearrangements and speciation. *Trends Ecol. Evol.* 16, 351–358
6. Strasburg, J.L. *et al.* (2012) What can patterns of differentiation across plant genomes tell us about adaptation and speciation? *Philos. Trans. R. Soc. B Biol. Sci.* 367, 364–373
7. Zhang, Z. *et al.* (2015) Genome-wide mapping of structural variations reveals a copy number variant that determines reproductive morphology in cucumber. *Plant Cell.* Published online May 22, 2015. http://dx.doi.org/10.1105/tpc.114.135848
8. Saxena, R.K. *et al.* (2014) Structural variations in plant genomes. *Brief. Funct. Genomics* 13, 296–307
9. Tang, Z. *et al.* (2014) New types of wheat chromosomal structural variations in derivatives of wheat-rye hybrids. *PLoS ONE* 9, e110282
10. Xiong, Z. *et al.* (2011) Homoeologous shuffling and chromosome compensation maintain genome balance in resynthesized allopolyploid *Brassica napus*. *Proc. Natl. Acad. Sci.* 108, 7908–7913
11. Lim, K.Y. *et al.* (2008) Rapid chromosome evolution in recently formed polyploids in *Tragopogon* (Asteraceae). *PLoS ONE* 3, e3353
12. Pontes, O. *et al.* (2004) Chromosomal locus rearrangements are a rapid response to formation of the allotetraploid *Arabidopsis suecica* genome. *Proc. Natl. Acad. Sci.* 101, 18240–18245
13. Skalická, K. *et al.* (2005) Preferential elimination of repeated DNA sequences from the paternal, *Nicotiana tomentosiformis* genome donor of a synthetic, allotetraploid tobacco. *New Phytol.* 166, 291–303
14. Michael, T.P. and VanBuren, R. (2015) Progress, challenges and the future of crop genomes. *Curr. Opin. Plant Biol.* 24, 71–81
15. Barrick, J.E. *et al.* (2014) Identifying structural variation in haploid microbial genomes from short-read resequencing data using breseq. *BMC Genomics* 15, 1039
16. Hormozdiari, F. *et al.* (2009) Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes. *Genome Res.* 19, 1270–1278
17. Bolger, M.E. *et al.* (2014) Plant genome sequencing — applications for crop improvement. *Curr. Opin. Biotechnol.* 26, 31–37
18. Doležel, J. *et al.* (2014) Advances in plant chromosome genomics. *Biotechnol. Adv.* 32, 122–136
19. Lewin, H.A. *et al.* (2009) Every genome sequence needs a good map. *Genome Res.* 19, 1925–1928
20. Oeveren, J. and van *et al.* (2011) Sequence-based physical mapping of complex genomes by whole genome profiling. *Genome Res.* 21, 618–625
21. Chen, M. *et al.* (2002) An integrated physical and genetic map of the rice genome. *Plant Cell Online* 14, 537–545
22. Fang, Z. *et al.* (2003) iMap: a database-driven utility to integrate and access the genetic and physical maps of maize. *Bioinformatics* 19, 2105–2111
23. Mozo, T. *et al.* (1999) A complete BAC-based physical map of the *Arabidopsis thaliana* genome. *Nat. Genet.* 22, 271–275
24. Wu, C. *et al.* (2004) A BAC- and BIBAC-based physical map of the soybean genome. *Genome Res.* 14, 319–326
25. Meyers, B.C. *et al.* (2004) Mapping and sequencing complex genomes: let's get physical! *Nat. Rev. Genet.* 5, 578–588
26. Paux, E. *et al.* (2008) A physical map of the 1-gigabase bread wheat chromosome 3B. *Science* 322, 101–104
27. Tang, H. *et al.* (2014) An improved genome release (version Mt4.0) for the model legume *Medicago truncatula*. *BMC Genomics* 15, 312
28. Chamala, S. *et al.* (2013) Assembly and validation of the genome of the nonmodel basal angiosperm *Amborella*. *Science* 342, 1516–1517
29. Zhang, Q. *et al.* (2012) The genome of *Prunus mume*. *Nat. Commun.* 3, 1318
30. Kawahara, Y. *et al.* (2013) Improvement of the *Oryza sativa* Nipponbare reference genome using next generation sequence and optical map data. *Rice* 6, 4

31. Zhou, S. *et al.* (2009) A single molecule scaffold for the maize genome. *PLoS Genet.* 5, e1000711
32. Schwartz, D.C. *et al.* (1993) Ordered restriction maps of *Saccharomyces cerevisiae* chromosomes constructed by optical mapping. *Science* 262, 110–114
33. Haas *et al.* (2009) Genome sequence and analysis of the Irish potato famine pathogen *Phytophthora infestans*. *Nature* 461, 393–398
34. Lai *et al.* (1999) A shotgun optical map of the entire *Plasmodium falciparum* genome. *Nat. Genet.* 23, 309–313
35. Lin *et al.* (1999) Whole-genome shotgun optical mapping of *Deinococcus radiodurans*. *Science* 285, 1558–1562
36. Zhou *et al.* (2004) Single-molecule approach to bacterial genomic comparisons via optical mapping. *J. Bacteriol.* 186, 7773–7782
37. Sahu, S.K. *et al.* (2012) DNA extraction protocol for plants with high levels of secondary metabolites and polysaccharides without using liquid nitrogen and phenol. *Int. Sch. Res. Not.* 2012, e205049
38. Hein, I. *et al.* (2005) Isolation of high molecular weight DNA suitable for BAC library construction from woody perennial soft-fruit species. *BioTechniques* 38, 69–71
39. Zhang, M. *et al.* (2012) Preparation of megabase-sized DNA from a variety of organisms using the nuclei method for advanced genomics research. *Nat. Protoc.* 7, 467–478
40. Kim, C.S. *et al.* (1997) A simple and rapid method for isolation of high quality genomic DNA from fruit trees and conifers using PVP. *Nucleic Acids Res.* 25, 1085–1086
41. Levi, A. *et al.* (1992) A rapid procedure for the isolation of RNA from high-phenolic-containing tissues of pecan. *HortScience* 27, 1316–1318
42. Pay, A. and Smith, M.A. (1988) A rapid method for purification of organelles for DNA isolation: self-generated percoll gradients. *Plant Cell Rep.* 7, 96–99
43. Marusyk, R. and Sergeant, A. (1980) A simple method for dialysis of small-volume samples. *Anal. Biochem.* 105, 403
44. Das, S.K. *et al.* (2010) Single molecule linear analysis of DNA in nano-channel labeled with sequence specific fluorescent probes. *Nucleic Acids Res.* 38, e177
45. Lam, E.T. *et al.* (2012) Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly. *Nat. Biotechnol.* 30, 771–776
46. Dong, Y. *et al.* (2013) Sequencing and automated whole-genome optical mapping of the genome of a domestic goat (*Capra hircus*). *Nat. Biotechnol.* 31, 135–141
47. Valouev, A. *et al.* (2006) Alignment of optical maps. *J. Comput. Biol.* 13, 442–462
48. Mak, A.C. *et al.* (2015) Genome-wide structural variation detection by genome mapping on nanochannel arrays. *Genetics* 202, 351–362
49. Teague, B. *et al.* (2010) High-resolution human genome structure by single-molecule analysis. *Proc. Natl. Acad. Sci.* 107, 10848–10853
50. Sharp, A.R. and Udall, J.A. OMWare: a tool for efficient assembly of genome-wide physical maps. *BMC Bioinformatics* (in press)
51. Shelton, J.M. *et al.* (2015) Tools and pipelines for BioNano data: molecule assembly pipeline and FASTA super scaffolding tool. *BMC Genomics* 16, 734
52. Tang, H. *et al.* (2015) ALLMAPS: robust scaffold ordering based on multiple maps. *Genome Biol.* 16, 3
53. Anantharaman, T.S. *et al.* (1999) Genomics via optical mapping III: contiging genomic DNA and variations. In *The Seventh International Conference on Intelligent Systems for Molecular Biology* (Lengauer, T., Schneider, R., Bork, P., Brutlad, D., Glasgow, J., Mewes, H.-W. and Zimmer, R., eds), pp. 18–27, AAAI Press
54. Mendelowitz, L. and Pop, M. (2014) Computational methods for optical mapping. *GigaScience* 3, 33
55. Nagarajan, N. *et al.* (2008) Scaffolding and validation of bacterial genome assemblies using optical restriction maps. *Bioinformatics* 24, 1229–1235
56. Muggli, M.D. *et al.* (2014) Efficient indexed alignment of contigs to optical maps. In *Algorithms in Bioinformatics* (Brown, D. and Morgenstern, B., eds), pp. 68–81, Springer
57. Lin, H.C. *et al.* (2012) AGORA: assembly guided by optical restriction alignment. *BMC Bioinformatics* 13, 189
58. Hastie, A.R. *et al.* (2013) Rapid genome mapping in nanochannel arrays for highly complete and accurate de novo sequence assembly of the complex Aegilops tauschii genome. *PLoS ONE* 8, e55864
59. Pendleton, M. *et al.* (2015) Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nat. Methods* 12, 780–786
60. Tang, H. *et al.* (2015) Optical mapping in plant comparative genomics. *GigaScience* 4, 3
61. Marroni, F. *et al.* (2014) Structural variation and genome complexity: is dispensable really dispensable? *Curr. Opin. Plant Biol.* 18, 31–36
62. Osborn, T. *et al.* (2003) Detection and effects of a homeologous reciprocal transposition in Brassica napus. *Genetics* 165, 1569–1577
63. Law, C.N. *et al.* (1987) Aneuploidy in wheat and its uses in genetic analysis. In *Wheat Breeding* (Lupton, D.F.G.H., ed.), pp. 71–108, Springer
64. Cui, L. *et al.* (2006) Widespread genome duplications throughout the history of flowering plants. *Genome Res.* 16, 738–749
65. Zhang, T. *et al.* (2015) Sequencing of allotetraploid cotton (*Gossypium hirsutum* L. acc. TM-1) provides a resource for fiber improvement. *Nat. Biotechnol.* 33, 531–537
66. Chester, M. *et al.* (2015) Patterns of chromosomal variation in natural populations of the neoallotetraploid *Tragopogon mirus* (Asteraceae). *Heredity* 114, 309–317
67. Paterson, A.H. *et al.* (2000) Comparative genomics of plant chromosomes. *Plant Cell* 12, 1523–1539
68. Grover, C.E. *et al.* (2015) Re-evaluating the phylogeny of allopolyploid *Gossypium* L. *Mol. Phylogenet. Evol.* 92, 45–52
69. Wang, S. *et al.* (2015) Sequence-based ultra-dense genetic and physical maps reveal structural variations of allopolyploid cotton genomes. *Genome Biol.* 16, 108
70. Morrell, P.L. *et al.* (2012) Crop genomics: advances and applications. *Nat. Rev. Genet.* 13, 85–96
71. Levy-Sakin, M. *et al.* (2014) Towards single-molecule optical mapping of the epigenome. *ACS Nano* 8, 14–26
72. Neely, R.K. *et al.* (2011) Optical mapping of DNA: single-molecule-based methods for mapping genomes. *Biopolymers* 95, 298–311
73. Jo, K. *et al.* (2007) A single-molecule barcoding system using nanoslits for DNA analysis. *Proc. Natl. Acad. Sci.* 104, 2673–2678
74. Cao, H. *et al.* (2014) Rapid detection of structural variation in a human genome using nanochannel-based genome mapping technology. *GigaScience* 3, 34
75. Levy-Sakin, M. and Ebenstein, Y. (2013) Beyond sequencing: optical mapping of DNA in the age of nanotechnology and nanoscopy. *Curr. Opin. Biotechnol.* 24, 690–698
76. Consortium, I.H.G.S. (2004) Finishing the euchromatic sequence of the human genome. *Nature* 431, 931–945
77. Sanger, F. *et al.* (1977) DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci.* 74, 5463–5467
78. Metzker, M.L. (2010) Sequencing technologies: the next generation. *Nat. Rev. Genet.* 11, 31–46
79. Shendure, J. and Ji, H. (2008) Next-generation DNA sequencing. *Nat. Biotechnol.* 26, 1135–1145
80. Lee, H. *et al.* (2016) Third-generation sequencing and the future of genomics. *bioRxiv* Published online April 13, 2016. http://dx.doi.org/10.1101/048603
81. Quail, M.A. *et al.* (2012) A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics* 13, 1
82. Eisenstein, M. (2015) Startups use short-read data to expand long-read sequencing market. *Nat. Biotechnol.* 33, 433–435