Schatz *et al. Genome Biology* 2012, **13**:243
http://genomebiology.com/2012/13/4/243

Genome Biology

REVIEW

Current challenges in *de novo* plant genome sequencing and assembly

Michael C Schatz*, Jan Witkowski and W Richard McCombie

Abstract

Genome sequencing is now affordable, but assembling plant genomes *de novo* remains challenging. We assess the state of the art of assembly and review the best practices for the community.

Keywords DNA sequencing, genome assembly, plant genomics.

The plant kingdom is filled with amazing diversity and significance. Plants form the base of the food chain that provides food for all living organisms, and just 15 crop plants provide 90% of the world's food intake [1]. Plant species are responsible for maintaining the balance of the carbon cycles [2], for developing and maintaining soil from erosion [3], and are promising sources of renewable energy [4]. Plant byproducts are used in many human medicines [5], and plants have been essential model organisms for studying biological systems such as the role of transposons and epigenetics [6]. For all these reasons and many more, there is great interest in sequencing plant genomes, but relatively few plant species have been sequenced compared with the hundreds of thousands of species around the world.

The first free-living organisms were sequenced less than 20 years ago, starting with simple microbial genomes [7], and increasing in complexity to the first eukaryotic genomes [8], the first multicellular species [9], and then on to plant genomes, including *Arabidopsis thaliana* (thale cress) [10], *Oryza sativa* (rice) [11], *Carica papaya* (papaya) [12] and *Zea mays* (maize) in 2009 [13], using first-generation capillary sequencing. Since then many others have been sequenced leveraging second-generation sequencing, including *Fragaria vesca* (strawberry) [14], *Solanum lycopersicum* (tomato) [15] and *Cajanus cajan* (pigeonpea) [16], and dozens more are nearing completion [17]. This increase in sequenced plant genomes has largely been driven by technological improvements: whereas the first generation of automated DNA sequencing instruments could sequence thousands of base pairs per day, current state-of-the-art second-generation sequencing instruments can sequence many billions of bases per day for hundreds or thousands of dollars per gigabase instead of millions or billions of dollars per gigabase [18]. These technologies have been applied to study thousands of genomes across the tree of life, enabling rich annotation of their gene networks [19], the development of comparative genomics approaches to infer evolutionary and domestication forces [13], the cataloging of genomic markers to optimize plant breeding [20], and numerous other studies that use the genome sequence as the backbone of the analysis [21].

In contrast to the tremendous advances in throughput, assembling sequencing reads remains a substantial endeavor, much greater than the sequencing efforts alone would suggest [22-24]. Large complex plant genomes remain a particularly difficult challenge for *de novo* assembly for a variety of biological, computational and biomolecular reasons. Plant genomes can be nearly 100 times larger [25] than the currently sequenced bird [26], fish [27] or mammalian genomes [28]. In addition they can have much higher ploidy, which is estimated to occur in up to 80% of all plant species [29], and higher rates of heterozygosity and repeats [30] than their counterparts in other kingdoms. Furthermore, the gene content in plants can be very complex, as shown by the presence of large gene families and abundant pseudogenes with nearly identical sequences derived from recent whole genome duplication events and transposon activity [13]. Plants tend to have high copy chloroplasts and mitochondria organelles, which complicate assembly of their remnants in the nuclear genome and skew coverage levels [12]. Finally, it is often very difficult to extract large quantities of high-quality DNA from plant material, making it difficult to prepare proper libraries for sequencing.

For all of these reasons, sequencing and *de novo* assembling a plant genome can create a highly fragmented result. Instead of large contigs and scaffolds spanning large chromosome regions seen in recent

*Correspondence: mschatz@cshl.edu
Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724, USA

vertebrate genome assemblies [31], there is a greater chance to assemble the sequencing reads into isolated gene islands among the background of high copy repeats [13]. Furthermore, the gene sequences may not always be correct, considering that nearly identical gene families are notoriously difficult to assemble and may collapse into a mosaic sequence without necessarily representing any member of the family [32]. If the level of fragmentation and mis-assembly is too great, downstream analysis will be noisy, and could even lead to false conclusions of the biology [33].

Knowing how to assemble these genomes accurately, how to best make use of the potentially highly fragmented assemblies and how to perform these applications at the lowest cost are important in today's funding environment. Genome assembly has always been an incremental process, and there are only a handful of truly finished large genomes today - even the latest release of the 'finished' human reference genome has millions of unresolved nucleotides [34]. Therefore, we need to assess when an assembly is good enough to be useful to the community, and how the agencies can get the most out of the available funding. Finally, how can researchers stay afloat in the rapidly evolving landscape with technology evolving so quickly it is challenging to know what the guidelines for plant assembly will be in 12 months or beyond. Here we assess the state of the art of *de novo* assembly, assess what can be expected to develop, and review the best practices for the plant community.

Assessing the needs

Assembling any genome requires the proper combination of coverage, read length and read quality [22]. If any of these factors are not met, then it is a mathematical certainty that the assembly will be fragmented into many small contigs. The Lander-Waterman model offers an analytic, if optimistic, prediction on the minimum coverage needed to assemble large contigs [35]. Using this model, a minimum of 15-fold coverage is required to assemble 100 bp reads into large contigs. However, once coverage has been equalized for errors, ploidy, sequence biases and other complicating factors, the minimum required coverage level may be much higher and sequencing to at least 100-fold coverage is recommended [31].

This statistical model also does not consider repeat composition, and short reads alone may never have the information content to resolve complex repetitive sequences. Resolving large or complex repeats fundamentally requires longer spanning information to bridge across the repeats back to unique sequence in the form of longer reads, mate-pairs, long-range mapping information or a method for fragment localization [32]. Read quality is also not directly considered in the Lander-Waterman model, but low-quality reads will reduce effective coverage and obscure true overlaps between sequencing reads, thus fragmenting the assembly and risking collapsing more repeats.

Overcoming these challenges depends on advances in both sequencing technology and assembly technology. Sequencing technology needs: (1) instrumentation improvements, including improvements in throughput, cost, read lengths and accuracy; and (2) molecular protocols, including developing new types of libraries and also new techniques for multiplexing samples to take advantage of the tremendous throughput available per instrument run. Assembly technology needs: (1) improved algorithms for accurately assembling complex genomes at scale; and (2) improved analytics to record, manipulate, analyze and visualize features to translate the salient assembly information to the broader plant biology community.

Sequence technology

The highest capacity sequencing instruments available today, such as the Illumina HiSeq 2000, can sequence nearly 100 Gbp per day, and make it possible to sequence a 3 Gbp genome to high coverage for less than US$10,000 [36]. Using these technologies, it is also possible to sequence paired-end or mate libraries ranging in size up to a few thousand base pairs. As such, even large plant genome projects can count on relatively inexpensive, deep coverage with approximately 100 bp reads and 1 to 5 kbp mate libraries. However, these short reads and small libraries have substantial limitations for large genomes with large repetitive content. Constructing high-quality draft genome assemblies for the largest plant genomes absolutely requires enhanced sequencing approaches to generate longer reads and mate-pair libraries, and protocols for localizing the sequencing and assembly problem.

One of the strongest needs is for protocols for efficiently generating a mix of larger libraries, such as 10 kbp, 40 kbp or 150 kbp in addition to standard 5 kbp libraries. Currently available protocols for these larger sizes, such as with fosmids [37], or bacterial artificial chromosome (BAC)-end sequencing [38], are effective but are laborious, costly and time consuming relative to the sequencing itself. Furthermore, the larger libraries inevitably have increased size variance and less reliable mate information. The sequencing itself needs to be improved to reduce the biases from GC composition, chimeric reads and mates, and other effects so that the coverage along the genome will be uniform and complete [39].

One promising approach for substantially longer reads and unbiased coverage is the rise of third-generation sequencing technologies such as that from Pacific Biosciences [40] and the newly announced instruments

from Oxford Nanopore [41]. These platforms promise to generate longer reads that can be used for sequencing through complex repeats, link gene islands and phase haplotypes. However, these technologies are relatively immature for immediate widespread application to all large genomes of interest. Sequencers from Roche/454 make it possible to sequence approximately 700 bp reads, but at greater cost than short read sequencing, and it may not be sufficient to span the largest repeats [42].

Optical mapping technologies are another possibility for generating very long range linking information between sequence contigs and have a successful history in plant genomics [43,44], although the current worldwide capacity is also below the demand. New technologies such as nanocoding [45], and new instruments from commercial vendors, including OpGen [46] and BioNanoGenomics [47], are expected in the next couple of years and they could expand the capacity for optical mapping similar to that seen in sequencing.

A complementary approach to improved sequencing and mapping is to develop methods for localizing sequencing and thus simplifying the assembly problem. There is a successful history of BAC-by-BAC sequencing of plant genomes [10,11], and this is effective in the sense that assembling an isolated BAC is far simpler than assembling the entire genome. However, this technology is now prohibitively expensive without significant enhancement. For example, sequencing large genomes such as maize using a BAC-by-BAC approach costs tens of millions of dollars and hundreds of thousands of BAC clones. While next-generation sequencing would certainly reduce this cost, it is not readily possible to efficiently use next-generation sequencing on the number of BAC clones needed. This, coupled with the high cost of making and storing the large numbers of libraries needed, greatly limits the feasibility of BAC-by-BAC sequencing in the next-generation world.

Versions of BAC-by-BAC using pools of BAC or pools of fosmids is an attractive option for localizing the problem, assuming such libraries can be efficiently made and barcoding protocols can be effectively applied to tag the molecules [48]. However, to utilize the capacity of current sequencers fully, so many BACs need to be pooled in a lane that it would not effectively localize the assembly problem unless the BACs can be multiplexed and barcoded to a very high degree. Furthermore, preparing and storing these libraries will still require a substantial cost unless they can be made in a fully automated fashion. Alternative molecular isolation technologies that can be used for localizing individual chromosomes in the sample, such as flow sorting, are promising alternatives and are starting to become more widely available [49,50].

Assembly technology

Genome assembly has been metaphorically described as the process of assembling a jigsaw puzzle from the individual reads [22]. In the case of the largest, most repetitive plant genomes, it could be metaphorically described as assembling a large jigsaw consisting of blue sky separated by nearly indistinguishable wisps of white clouds of genes - seemingly an impossible task. Assembly generally follows a hierarchical approach of comparing the individual reads to form an assembly graph of the overlapping reads or kmers, then simplifying the graph to form the initial contigs, and finally using mate-pairs and marker information to order and orient the initial contigs into scaffolds (Figure 1). Assembling a large genome is operationally complicated in that it demands extensive error correction and filtering, and large computational resources, and is often highly sensitive to the parameters used. Even beyond these complications, assembly is fundamentally complicated because repeats introduce ambiguity in how the reads should be ordered so that no perfect algorithm exists for reconstructing entire genomes even if every base of the genome has been sequenced to high depth.

Several short-read assembly packages have been proven for mammalian-sized genomes up to the 3 Gbp human genome, including ABySS [51], ALLPATHS-LG [31], the Celera Assembler [52,53], Newbler [54], SGA [55] and SOAPdenovo [56]. These assemblers can produce high-quality assemblies from short reads, although they generally require servers or clusters with 512 gigabytes of RAM and many terabytes of disk space available for a gigabase-sized genome [31]. However, these servers are decreasing in costs and can be purchased for under US$35,000 from several major computer vendors [57], and supercomputing centers make them available without any cost [58]. This is promising, but assembling the largest plant genomes currently being sequenced, such as the loblolly pine genome of approximately 21 Gbp [59], will increase the computational demands by nearly an order of magnitude, for which there is no proven technology. Enhanced algorithms for compression and distributing the computation are actively being researched [55].

Two major efforts to evaluate the state-of-the-art in assembly technology were published last year: the Assemblathon [24] and the Genome Assembly Gold-Standard Evaluation (GAGE) [23]. Both projects evaluated the performance of various genome assemblers in a competitive framework with both simulated and real datasets. They showed there was great difference in the quality of the results depending on the assembler and pipelines used. Researchers planning to assemble a genome of any size are encouraged to study their results, such as the needs for error correction, recommended

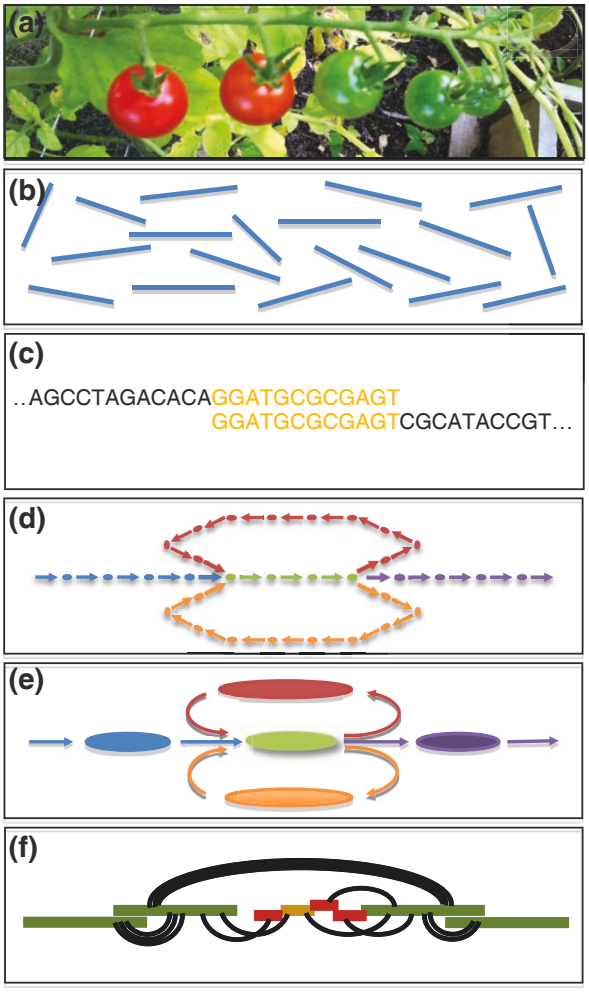

Figure 1. Schematic overview of genome assembly. (a) DNA is collected from the biological sample and sequenced. **(b)** The output from the sequencer consists of many billions of short, unordered DNA fragments from random positions in the genome. **(c)** The short fragments are compared with each other to discover how they overlap. **(d)** The overlap relationships are captured in a large assembly graph shown as nodes representing kmers or reads, with edges drawn between overlapping kmers or reads. **(e)** The assembly graph is refined to correct errors and simplify into the initial set of contigs, shown as large ovals connected by edges. **(f)** Finally, mates, markers and other long-range information are used to order and orient the initial contigs into large scaffolds, as shown as thin black lines connecting the initial contigs.

assemblers and evaluation criterion. However, the genomes studied in these projects were relatively small and simple compared with the most complex plant genomes. The plant community would be well served by hosting regular competitions with plant genomes, especially since all of the major assemblers have been developed targeting vertebrate genomes, and no assembler has been proven with higher levels of ploidy or heterozygosity.

Related to the *de novo* assembly problem, research is greatly needed to help improve the representation of assembled genomes, including creating graph-centric and population-aware formats that can represent the complexities of plant genomes, particularly those that are only partially assembled [60-62]. Incremental algorithms that can update the assembly and annotation as new data become available would also be extremely useful [33]. Finally, continued research into assembly validation is necessary for determining when an assembly is correct and conclusions can be trusted [32,63].

Analytics

Sequencing and assembling a genome are often just the first stages of a larger study. Immediately following the assembly, the genome will need to be annotated to catalog genes and other features of interest [64], or aligned to other genomes to enable comparative genomics studies [65]. Several sequencing-based assays, such as RNA-seq [66] and Methyl-seq [67], can be used with the assembly to study transcriptionally or epigenetically active regions of the genome, and population studies will often attempt to build higher-order relationships, such as gene networks, or relate genotype to phenotype.

Currently, pipelines are available for carrying out these operations and displaying results in a 'genome browser,' but continued research is needed to make the pipelines and results more accessible to different types of user. Systems such as Galaxy [68], Gramene [69] and Drupal [70] are among the leading graphical systems for executing workflows, visualizing sequencing assay results, and enabling collaborative discussions, respectively, but they operate as separate systems. A fully integrated system such as has been proposed by iPlant [71], and the DOE Systems Biology Knowledgebase [72] initiatives would lower the barrier for learning to operate these functions. In either case it is critical that the community enhance these systems and the underlying algorithms to better support the complexity of plant genomes and their evolving assemblies.

Trends and recommendations

The plant kingdom has incredible variation and diversity, and as a result each plant sequencing project seems to have its own unique analysis needs. Sequencing and assembly technologies are evolving so rapidly it is impossible to predict what will be available even one year in the future. Despite these complexities, certain trends are emerging as best practices.

Mixed library, high-coverage sequencing

Because of economic and technological reasons, the majority of sequence produced in the next 18 months will continue to originate from short reads of approximately

100 to 200 bp. Fortunately, sequences of this length can be assembled into high-quality draft assemblies for genomes as complex as human when sequenced in a mixture of libraries. In particular, Gnerre *et al.* [31] recommend 45× paired-end (2 × 100 bp at 180 bp), 45× short jump (2 × 100 bp at 3 kbp), 5× long jump (2 × 100 bp at 6 kbp) and 1× fosmid (2 × 26 bp at 40 kbp) to generate high-quality draft assemblies. Since the paired-end reads designed in this way overlap by approximately 20 bp, they can be preassembled into pseudo-long reads of approximately twice the original length using the built-in capabilities of ALLPATHS-LG [31] or by a standalone preassembler such as FLASH [73]. Assemblers that do not include built-in error correction greatly benefit from then applying software such as Quake [74] to identify and fix sequencing errors before assembly. The larger libraries are then needed for ordering the initial contigs into progressively larger scaffolds.

For the largest and most complex plant genomes, even these libraries may not be sufficient to span the largest or more complex repeats, and it may be necessary to employ a hybrid approach using a combination of short and long reads, and even long-range mapping technologies or localization methods. Long reads over 800 bp are available today from Roche/454, albeit at higher cost than short read sequencing, and third-generation sequencing technologies promise to provide even longer reads. As sequencing costs and instrument runtimes continue to drop, researchers are also recommended to sequence a low coverage 'genome snapshot' to evaluate the genome and library composition before attempting to sequence the genome to high coverage.

Bioinformatics partnerships

Assembling and analyzing raw sequence data still require substantial bioinformatics effort and expertise. Before attempting a complex assembly, plant biologists are strongly encouraged to develop partnerships with bioinformatics laboratories that have sufficient skills and resources to handle the onslaught of data and diagnosis problems as they occur. Fortunately, the funding agencies are aware of these challenges, and it is our hope they would be responsive to requests for appropriate bioinformatics funding.

Bioinformatics laboratories are encouraged to enhance, expand and refine their algorithms and analytics specifically for the complexities of plant genomes. In particular, because of high diversity, heterozygosity and ploidy not found in other kingdoms, there is a strong need to develop a plant-specific genome assembler that can overcome these challenges and represent the plant genome assemblies in more versatile graph-based formats along with the supporting tools for analyzing these graphs

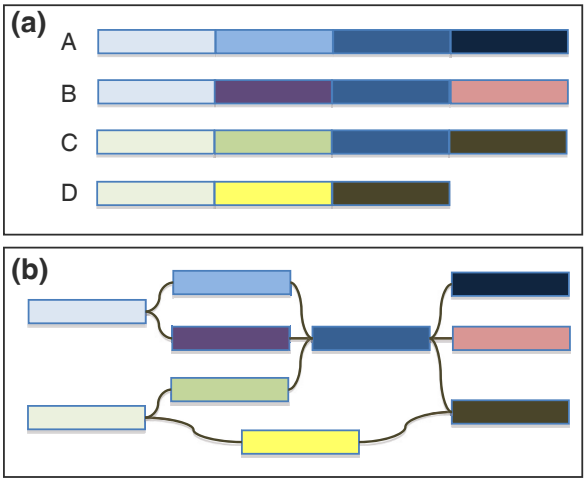

Figure 2. Ploidy, heterozygosity and the assembly graph. **(a)** Schematic representation of a tetraploid genome, such as apple, cotton or cabbage, consisting of haploid chromosomes A to D with homozygosity/heterozygosity shown as different colored blocks. **(b)** Even without repeats or sequencing error, the assembly graph of the homozygous and heterozygous segments of the genome branch and intertwine in complex patterns. A plant-specific assembler would need to recognize these branching patterns and attempt to reconstruct the individual sequences for chromosomes A to D.

(Figure 2). Furthermore, the trend in bioinformatics software development is to develop only enough of a user interface to support the needs of a particular project. If this trend continues, many groups will reinvent the same software over and over again, wasting time and resources. Instead, funding agencies would be better served by requiring software to be developed with a high-quality user-friendly interface or integrated into a graphical system such as Galaxy, even if it requires modestly more upfront funding.

Awareness, training and education

Principal investigators need to become better informed to the current best practices for genome assembly and develop a better understanding of the effort involved to sequence, assemble, annotate and analyze a new genome. More classes and training are needed for graduate and undergraduate students to learn the fundamentals of sequence analysis and quantitative techniques. Better training is needed to teach non-experts to use the software packages, and to educate everyone about the resources that are available. The plant sequencing community would benefit by forming and hosting plant genome analysis competitions in the spirit of the Assemblathon or GAGE to evaluate the state-of-the-art for assembly, annotation and other assays. The best practices of today are certain to change as new sequencing, mapping and computational technologies

are introduced, and this will be the only way to monitor these developments.

Final thoughts

We are still many years away from push-button sequencing and assembly of complex plant genomes into completely finished genomes at low cost. Nevertheless, it is now possible and affordable to sequence and assemble great numbers of interesting plant genomes into highly useful draft genome assemblies if one is mindful of the biotechnology and algorithmic challenges involved. The next frontier for plant genomics is to characterize the diversity of genomic variations across large populations, deeply annotate their functional elements, and develop predictive quantitative models relating genotype to phenotype. Improved sequencing technology and sequencing assays are certain to play a large role in these studies as well, and we envision a tight relationship between biology, biotechnology and analytics for years to come.

Abbreviations

BAC, bacterial artificial chromosome; bp, base pair; GAGE, Genome Assembly Gold-Standard Evaluation.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

We thank all of the participants of the meeting on the future of plant genome sequencing and analysis held at the Banbury Conference Center at Cold Spring Harbor in the summer of 2010. This work was funded, in part, by NSF award IOS-1135736, the US Department of Energy, Office of Biological and Environmental Research under Contract DE-AC02-06CH11357, and NIH RO1 HG006677-12.

Published: 27 April 2012

References

1. United Nations Food and Agriculture Organization: **Dimensions of Need - An atlas of food and agriculture. Staple foods: What do people eat.** [http://www.fao.org/docrep/u8480e/u8480e07.htm]
2. Falkowski P, Scholes RJ, Boyle E, Canadell J, Canfield D, Elser J, Gruber N, Hibbard K, Högberg P, Linder S, Mackenzie FT, Moore B 3rd, Pedersen T, Rosenthal Y, Seitzinger S, Smetacek V, Steffen W: **The global carbon cycle: a test of our knowledge of earth as a system.** *Science* 2000, **290:**291-296.
3. Pimentel D, Harvey C, Resosudarmo P, Sinclair K, Kurz D, McNair M, Crist S, Shpritz L, Fitton L, Saffouri R, Blair R: **Environmental and economic costs of soil erosion and conservation benefits.** *Science* 1995, **267:**1117-1123.
4. Fairley P: **Introduction: Next generation biofuels.** *Nature* 2011, **474:**S2-5.
5. Mann J: **Natural products in cancer chemotherapy: past, present and future.** *Nat Rev Cancer* 2002, **2:**143-148.
6. Lippman Z, Gendrel AV, Black M, Vaughn MW, Dedhia N, McCombie WR, Lavine K, Mittal V, May B, Kasschau KD, Carrington JC, Doerge RW, Colot V, Martienssen R: **Role of transposable elements in heterochromatin and epigenetic control.** *Nature* 2004, **430:**471-476.
7. Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM: **Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd.** *Science* 1995, **269:**496-512.
8. Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M, Louis EJ, Mewes HW, Murakami Y, Philippsen P, Tettelin H, Oliver SG: **Life with 6000 genes.** *Science* 1996, **274:**546, 563-567.
9. C. elegans Sequencing Consortium: **Genome sequence of the nematode *C. elegans*: a platform for investigating biology.** *Science* 1998, **282:**2012-2018.
10. Arabidopsis Genome Initiative: **Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*.** *Nature* 2000, **408:**796-815.
11. International Rice Genome Sequencing Project: **The map-based sequence of the rice genome.** *Nature* 2005, **436:**793-800.
12. Ming R, Hou S, Feng Y, Yu Q, Dionne-Laporte A, Saw JH, Senin P, Wang W, Ly BV, Lewis KL, Salzberg SL, Feng L, Jones MR, Skelton RL, Murray JE, Chen C, Qian W, Shen J, Du P, Eustice M, Tong E, Tang H, Lyons E, Paull RE, Michael TP, Wall K, Rice DW, Albert H, Wang ML, Zhu YJ, *et al.*: **The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya Linnaeus*).** *Nature* 2008, **452:**991-996.
13. Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, Liang C, Zhang J, Fulton L, Graves TA, Minx P, Reily AD, Courtney L, Kruchowski SS, Tomlinson C, Strong C, Delehaunty K, Fronick C, Courtney B, Rock SM, Belter E, Du F, Kim K, Abbott RM, Cotton M, Levy A, Marchetto P, Ochoa K, Jackson SM, Gillam B, *et al.*: **The B73 maize genome: complexity, diversity, and dynamics.** *Science* 2009, **326:**1112-1115.
14. Shulaev V, Sargent DJ, Crowhurst RN, Mockler TC, Folkerts O, Delcher AL, Jaiswal P, Mockaitis K, Liston A, Mane SP, Burns P, Davis TM, Slovin JP, Bassil N, Hellens RP, Evans C, Harkins T, Kodira C, Desany B, Crasta OR, Jensen RV, Allan AC, Michael TP, Setubal JC, Celton JM, Rees DJ, Williams KP, Holt SH, Ruiz Rojas JJ, Chatterjee M, *et al.*: **The genome of woodland strawberry (*Fragaria vesca*).** *Nat Genet* 2011, **43:**109-116.
15. Sol Genomis Network: **International Tomato Genome Sequencing Project** [http://solgenomics.net/organism/Solanum_lycopersicum/genome]
16. Varshney RK, Chen W, Li Y, Bharti AK, Saxena RK, Schlueter JA, Donoghue MT, Azam S, Fan G, Whaley AM, Farmer AD, Sheridan J, Iwata A, Tuteja R, Penmetsa RV, Wu W, Upadhyaya HD, Yang SP, Shah T, Saxena KB, Michael T, McCombie WR, Yang B, Zhang G, Yang H, Wang J, Spillane C, Cook DR, May GD, Xu X, *et al.*: **Draft genome sequence of pigeonpea (*Cajanus cajan*), an orphan legume crop of resource-poor farmers.** *Nat Biotechnol* 2012, **30:**83-89.
17. **NCBI Entrez Genome: Plant Genome Central** [http://www.ncbi.nlm.nih.gov/genomes/PLANTS/PlantList.html]
18. Zhou X, Ren L, Meng Q, Li Y, Yu Y, Yu J: **The next-generation sequencing technology and application.** *Protein Cell* 2010, **1:**520-536.
19. Park SJ, Jiang K, Schatz MC, Lippman ZB: **Rate of meristem maturation determines inflorescence architecture in tomato.** *Proc Natl Acad Sci U S A* 2012, **109:**639-644.
20. Moose SP, Mumm RH: **Molecular plant breeding as the foundation for 21st century crop improvement.** *Plant Physiol* 2008, **147:**969-977.
21. Morrell PL, Buckler ES, Ross-Ibarra J: **Crop genomics: advances and applications.** *Nat Rev Genet* 2011, **13:**85-96.
22. Schatz MC, Delcher AL, Salzberg SL: **Assembly of large genomes using second-generation sequencing.** *Genome Res* 2010, **20:**1165-1173.
23. Salzberg SL, Phillippy AM, Zimin A, Puiu D, Magoc T, Koren S, Treangen TJ, Schatz MC, Delcher AL, Roberts M, Marçais G, Pop M, Yorke JA: **GAGE: A critical evaluation of genome assemblies and assembly algorithms.** *Genome Res* 2012, **22:**557-567.
24. Earl D, Bradnam K, St John J, Darling A, Lin D, Fass J, Yu HO, Buffalo V, Zerbino DR, Diekhans M, Nguyen N, Ariyaratne PN, Sung WK, Ning Z, Haimel M, Simpson JT, Fonseca NA, Birol İ, Docking TR, Ho IY, Rokhsar DS, Chikhi R, Lavenier D, Chapuis G, Naquin D, Maillet N, Schatz MC, Kelley DR, Phillippy AM, Koren S, *et al.*: **Assemblathon 1: a competitive assessment of de novo short read assembly methods.** *Genome Res* 2011, **21:**2224-2241.
25. Pellicer J, Fay MF, Leitch IJ: **The largest eukaryotic genome of them all?** *Bot J Linn Soc* 2010, **164:**10-15.
26. Dalloul RA, Long JA, Zimin AV, Aslam L, Beal K, Blomberg LA, Bouffard P, Burt DW, Crasta O, Crooijmans RP, Cooper K, Coulombe RA, De S, Delany ME, Dodgson JB, Dong JJ, Evans C, Frederickson KM, Flicek P, Florea L, Folkerts O, Groenen MA, Harkins TT, Herrero J, Hoffmann S, Megens HJ, Jiang A, de Jong P, Kaiser P, Kim H, *et al.*: **Multi-platform next-generation sequencing of the domestic turkey (*Meleagris gallopavo*): genome assembly and analysis.** *PLoS Biol* 2010, **8:**e1000475.
27. Star B, Nederbragt AJ, Jentoft S, Grimholt U, Malmstrøm M, Gregers TF, Rounge TB, Paulsen J, Solbakken MH, Sharma A, Wetten OF, Lanzén A, Winer R, Knight J, Vogel JH, Aken B, Andersen O, Lagesen K, Tooming-Klunderud A, Edvardsen RB, Tina KG, Espelund M, Nepal C, Previti C, Karlsen BO, Moum T, Skage M, Berg PR, Gjøen T, Kuhl H, *et al.*: **The genome sequence of Atlantic cod reveals a unique immune system.** *Nature* 2011, **477:**207-210.
28. The International Human Genome Sequencing Consortium: **Initial sequencing and analysis of the human genome.** *Nature* 2001, **409:**860-921.
29. Meyers LA, Levin DA: **On the abundance of polyploids in flowering plants.**

Evolution 2006, **60:**1198-1206.
30. Gore MA, Chia JM, Elshire RJ, Sun Q, Ersoz ES, Hurwitz BL, Peiffer JA, McMullen MD, Grills GS, Ross-Ibarra J, Ware DH, Buckler ES: **A first-generation haplotype map of maize.** *Science* 2009, **326:**1115-1117.
31. Gnerre S, Maccallum I, Przybylski D, Ribeiro FJ, Burton JN, Walker BJ, Sharpe T, Hall G, Shea TP, Sykes S, Berlin AM, Aird D, Costello M, Daza R, Williams L, Nicol R, Gnirke A, Nusbaum C, Lander ES, Jaffe DB: **High-quality draft assemblies of mammalian genomes from massively parallel sequence data.** *Proc Natl Acad Sci U S A* 2011, **108:**1513-1518.
32. Phillippy AM, Schatz MC, Pop M: **Genome assembly forensics: finding the elusive mis-assembly.** *Genome Biol* 2008, **9:**R55.
33. Florea L, Souvorov A, Kalbfleisch TS, Salzberg SL: **Genome assembly has a major impact on gene content: a comparison of annotation in two Bos taurus assemblies.** *PloS One* 2011, **6:**e21400.
34. International Human Genome Sequencing Consortium: **Finishing the euchromatic sequence of the human genome.** *Nature* 2004, **431:**931-945.
35. Lander ES, Waterman MS: **Genomic mapping by fingerprinting random clones: a mathematical analysis.** *Genomics* 1988, **2:**231-239.
36. Illumina: **HiSeq 2500 Sequencing System Specifications** [http://www.illumina.com/Documents/%5Cproducts%5Cappnotes%5Cappnote_hiseq2500.pdf]
37. Ammiraju JS, Yu Y, Luo M, Kudrna D, Kim H, Goicoechea JL, Katayose Y, Matsumoto T, Wu J, Sasaki T, Wing RA: **Random sheared fosmid library as a new genomic tool to accelerate complete finishing of rice (*Oryza sativa spp. Nipponbare*) genome sequence: sequencing of gap-specific fosmid clones uncovers new euchromatic portions of the genome.** *Theor Appl Genet* 2005, **111:**1596-1607.
38. Kelley JM, Field CE, Craven MB, Bocskai D, Kim UJ, Rounsley SD, Adams MD: **High throughput direct end sequencing of BAC clones.** *Nucleic Acids Res* 1999, **27:**1539-1546.
39. Aird D, Ross MG, Chen WS, Danielsson M, Fennell T, Russ C, Jaffe DB, Nusbaum C, Gnirke A: **Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries.** *Genome Biol* 2011, **12:**R18.
40. Rasko DA, Webster DR, Sahl JW, Bashir A, Boisen N, Scheutz F, Paxinos EE, Sebra R, Chin CS, Iliopoulos D, Klammer A, Peluso P, Lee L, Kislyuk AO, Bullard J, Kasarskis A, Wang S, Eid J, Rank D, Redman JC, Steyert SR, Frimodt-Møller J, Struve C, Petersen AM, Krogfelt KA, Nataro JP, Schadt EE, Waldor MK: **Origins of the *E. coli* strain causing an outbreak of hemolytic-uremic syndrome in Germany.** *New Engl J Med* 2011, **365:**709-717.
41. Oxford Nanopore Technologies: **The GridION system** [http://www.nanoporetech.com/technology/the-gridion-system/the-gridion-system]
42. Roche: **454 GS FLX+** [http://my454.com/products/gs-flx-system/index.asp]
43. Zhou S, Bechner MC, Place M, Churas CP, Pape L, Leong SA, Runnheim R, Forrest DK, Goldstein S, Livny M, Schwartz DC: **Validation of rice genome sequence by optical mapping.** *BMC Genomics* 2007, **8:**278.
44. Zhou S, Wei F, Nguyen J, Bechner M, Potamousis K, Goldstein S, Pape L, Mehan MR, Churas C, Pasternak S, Forrest DK, Wise R, Ware D, Wing RA, Waterman MS, Livny M, Schwartz DC: **A single molecule scaffold for the maize genome.** *PLoS Genet* 2009, **5:**e1000711.
45. Jo K, Schramm TM, Schwartz DC: **A single-molecule barcoding system using nanoslits for DNA analysis : nanocoding.** *Methods Mol Biol* 2009, **544:**29-42.
46. **OpGen** [http://www.opgen.com/]
47. **BioNanoGenomics** [http://www.bionanogenomics.com/]
48. Kitzman JO, Mackenzie AP, Adey A, Hiatt JB, Patwardhan RP, Sudmant PH, Ng SB, Alkan C, Qiu R, Eichler EE, Shendure J: **Haplotype-resolved genome sequencing of a Gujarati Indian individual.** *Nat Biotechnol* 2011, **29:**59-63.
49. Ng BL, Carter NP: **Factors affecting flow karyotype resolution.** *Cytometry A* 2006, **69:**1028-1036.
50. Sudbery I, Stalker J, Simpson JT, Keane T, Rust AG, Hurles ME, Walter K, Lynch D, Teboul L, Brown SD, Li H, Ning Z, Nadeau JH, Croniger CM, Durbin R, Adams DJ: **Deep short-read sequencing of chromosome 17 from the mouse strains A/J and CAST/Ei identifies significant germline variation and candidate genes that regulate liver triglyceride levels.** *Genome Biol* 2009, **10:**R112.
51. Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I: **ABySS: A parallel assembler for short read sequence data.** *Genome Res* 2009, **19:**1117-1123.
52. Myers EW, Sutton GG, Delcher AL, Dew IM, Fasulo DP, Flanigan MJ, Kravitz SA, Mobarry CM, Reinert KH, Remington KA, Anson EL, Bolanos RA, Chou HH, Jordan CM, Halpern AL, Lonardi S, Beasley EM, Brandon RC, Chen L, Dunn PJ, Lai Z, Liang Y, Nusskern DR, Zhan M, Zhang Q, Zheng X, Rubin GM, Adams MD, Venter JC: **A whole-genome assembly of *Drosophila*.** *Science* 2000, **287:**2196-2204.
53. Miller JR, Delcher AL, Koren S, Venter E, Walenz BP, Brownley A, Johnson J, Li K, Mobarry C, Sutton G: **Aggressive assembly of pyrosequencing reads with mates.** *Bioinformatics* 2008, **24:**2818-2824.
54. Margulies M., Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer ML, Jarvie TP, Jirage KB, Kim JB, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, *et al*.: **Genome sequencing in microfabricated high-density picolitre reactors.** *Nature* 2005, **437:**376-380.
55. Simpson JT, Durbin R: **Efficient *de novo* assembly of large genomes using compressed data structures.** *Genome Res* 2012, **22:**549-556.
56. Li R, Zhu H, Ruan J, Qian W, Fang X, Shi Z, Li Y, Li S, Shan G, Kristiansen K, Li S, Yang H, Wang J, Wang J: ***De novo* assembly of human genomes with massively parallel short read sequencing.** *Genome Res* 2010, **20:**265-272.
57. **Dell** [http://www.dell.com]
58. **Pittsburgh Supercomputing Center** [http://www.psc.edu/]
59. **Loblolly Pine Genome Project** [http://dendrome.ucdavis.edu/NealeLab/lpgp/]
60. Ye Y, Godzik A: **Multiple flexible structure alignment using partial order graphs.** *Bioinformatics* 2005, **21:**2362-2369.
61. Lee C, Grasso C, Sharlow MF: **Multiple sequence alignment using partial order graphs.** *Bioinformatics* 2002, **18:**452-464.
62. Paten B, Diekhans M, Earl D, John JS, Ma J, Suh B, Haussler D: **Cactus graphs for genome comparisons.** *J Comput Biol* 2011, **18:**469-481.
63. Schatz MC, Phillippy AM, Sommer DD, Delcher AL, Puiu D, Narzisi G, Salzberg SL, Pop M: **Hawkeye and AMOS: visualizing and assessing the quality of genome assemblies.** *Brief Bioinform*, in press.
64. Ouyang S, Thibaud-Nissen F, Childs KL, Zhu W, Buell CR: **Plant genome annotation methods.** *Methods Mol Biol* 2009, **513:**263-282.
65. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL: **Versatile and open software for comparing large genomes.** *Genome Biol* 2004, **5:**R12.
66. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B: **Mapping and quantifying mammalian transcriptomes by RNA-Seq.** *Nat Methods* 2008, **5:**621-628.
67. Cokus SJ, Feng S, Zhang X, Chen Z, Merriman B, Haudenschild CD, Pradhan S, Nelson SF, Pellegrini M, Jacobsen SE: **Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning.** *Nature* 2008, **452:**215-219.
68. Goecks J, Nekrutenko A, Taylor J: **Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences.** *Genome Biol* 2010, **11:**R86.
69. Liang C, Jaiswal P, Hebbard C, Avraham S, Buckler ES, Casstevens T, Hurwitz B, McCouch S, Ni J, Pujar A, Ravenscroft D, Ren L, Spooner W, Tecle I, Thomason J, Tung CW, Wei X, Yap I, Youens-Clark K, Ware D, Stein L: **Gramene: a growing plant comparative genomics resource.** *Nucleic Acids Res* 2008, **36(Database issue):**D947-953.
70. **Drupal™** [http://drupal.org/]
71. **iPlant Collaborative™** [http://www.iplantcollaborative.org/]
72. **Genomic Science Program: DOE Systems Biology Knowledgebase** [http://genomicscience.energy.gov/compbio/]
73. Magoc T, Salzberg SL: **FLASH: fast length adjustment of short reads to improve genome assemblies.** *Bioinformatics* 2011, **27:**2957-2963.
74. Kelley DR, Schatz MC, Salzberg SL: **Quake: quality-aware detection and correction of sequencing errors.** *Genome Biol* 2010, **11:**R116.

doi:10.1186/gb-2012-13-4-243
Cite this article as: Schatz MC, *et al*.: Current challenges in *de novo* plant genome sequencing and assembly. *Genome Biology* 2012, **13:**243.