The Open Bioinformatics Journal

Content list available at: www.benthamopen.com/TOBIOIJ/

DOI: 10.2174/1875036201811010231, 2018, *11*, 231-239

REVIEW ARTICLE

The Challenge of Genome Sequence Assembly

Andrew Collins[*]

Genetic Epidemiology and Bioinformatics Research Group, Faculty of Medicine, Duthie Building (MP 808), University of Southampton, Southampton General Hospital, Southampton, SO16 6YD, UK

Received: June 25, 2018 Revised: August 31, 2018 Accepted: September 17, 2018

Abstract:

Background:

Although whole genome sequencing is enabling numerous advances in many fields achieving complete chromosome-level sequence assemblies for diverse species presents difficulties. The problems in part reflect the limitations of current sequencing technologies. Chromosome assembly from 'short read' sequence data is confounded by the presence of repetitive genome regions with numerous similar sequence tracts which cannot be accurately positioned in the assembled sequence. Longer sequence reads often have higher error rates and may still be too short to span the larger gaps between contigs.

Objective:

Given the emergence of exciting new applications using sequencing technology, such as the Earth BioGenome Project, it is necessary to further develop and apply a range of strategies to achieve robust chromosome-level sequence assembly. Reviewed here are a range of methods to enhance assembly which include the use of cross-species synteny to understand relationships between sequence contigs, the development of independent genetic and/or physical scaffold maps as frameworks for assembly (for example, radiation hybrid, optical motif and chromatin interaction maps) and the use of patterns of linkage disequilibrium to help position, orient and locate contigs.

Results and Conclusion:

A range of methods exist which might be further developed to facilitate cost-effective large-scale sequence assembly for diverse species. A combination of strategies is required to best assemble sequence data into chromosome-level assemblies. There are a number of routes towards the development of maps which span chromosomes (including physical, genetic and linkage disequilibrium maps) and construction of these whole chromosome maps greatly facilitates the ordering and orientation of sequence contigs.

Keywords: Chromosome assembly, Cross-species synteny, Earth BioGenome Project, Linkage disequilibrium map, Sequence contigs, Whole genome sequencing .

1. INTRODUCTION

The construction of whole chromosome assemblies from sequencing data is recognised as one of the most challenging problems in modern genomics [1]. The scale of the problem is highlighted by recent developments, for example, the Earth BioGenome Project (EBP) which is a highly ambitious 10 year project to sequence, catalogue and characterise the genomes of 'all' of the Earth's eukaryotic biodiversity [2, 3]. The outcomes of the project will have the capacity to drive significant advances in agriculture, medicine and industry as the biological potential of sequences hidden within the world's hugely diverse set of genomes is realised. The three main project goals focus, firstly, on understanding evolutionary relationships between organisms and hence obtain insights into ecosystem composition, genome evolution and the acquisition of knowledge to accelerate the discovery of new species. The second goal

[*] Address correspondence to the author at the Genetic Epidemiology and Bioinformatics Research Group, Faculty of Medicine, Duthie Building (MP 808), University of Southampton, Southampton General Hospital, Southampton, SO16 6YD, UK; Tel: 44(0)2381206939; E-mail: arc@soton.ac.uk

1875-0362/18 2018 Bentham Open

considers conservation and strategies for the regeneration of biodiversity to develop a deeper understanding of the impact of climate change and human activities which will, in turn, facilitate evidence-based conservation. Thirdly the project will address the identification of novel medical resources, new routes to agricultural improvement, the identification of new biomaterials and strategies for enhanced environmental quality. However, the challenges to be addressed for completion of the project are considerable. To date, the genomes of only ~2500 eukaryotic species have been completely or partly sequenced and so only a tiny fraction of the predicted 10-15 million eukaryotic species is currently available for analysis at the genome level. Furthermore, only ~25 of these sequences meet the N50 standard for contig scaffold which quantifies how much of the genome is assembled into larger contigs (such that 50% of the assembled bases are in contigs of size N or greater [2, 4]. The EBP aims to start with the 1.5 million named eukaryotes with an initial target of obtaining quality reference sequences for a single member of each of the 9000 or so eukaryote families. The recognition that, with rapidly decreasing costs, whole genomes might be sequenced for just ~100 USD in the near future (for example on Illumina NovaSeq platforms) provides an indication that the target sequencing aims might be achieved cost-effectively.

However, despite the great success of numerous sequencing projects to date, *de novo* sequence assembly presents many difficulties. Most sequencing projects apply ‘short read’ shotgun sequencing and this approach has been highly effective for sequence analysis of many genomes, particularly in the recognition of disease-related variation in human data. Short-read sequence data are most frequently used for alignment against an established reference sequence to localise important DNA variants or polymorphisms which may be related to a phenotype of interest. This strategy has enabled the identification of mutations underlying many human diseases [5]. Success in the analysis of sequences from disease case samples, and subsequent recognition of disease causal variation is dependent on the alignment of sequences from the patient against the high-quality human reference sequence. The reference sequence was established using a combination of strategies including, relatively expensive and low-throughput, Sanger sequencing to produce long sequence reads of high quality. The assembly was supported through the use of dense genetic and physical maps as scaffolds and with further application of shotgun sequencing using a tiling path of short and long insert clones. The genetic and physical maps were essential to assign order and orientation to sequence contigs enabling the construction of whole chromosome assemblies [6 - 8].

In contrast to the development of the human genome sequence, high costs and technical complexities mean that achieving comparable high-quality sequence assemblies for most other species is unlikely to be possible. For most species draft genome sequences contain thousands of sequence contigs with limited information on how these can be assembled into representations of individual chromosomes. Because chromosome-level representations of the sequence are usually incomplete and may contain errors the value of these data for molecular and evolutionary studies is reduced. Therefore, a range of strategies is required to maximise the value of any sequence data obtained. This is important because the sequences of individual genes may be broken and/or badly annotated and lack the correct genomic context to facilitate detailed comparative analysis [9 - 11]. An overview of recent developments and a review of the range of strategies available to enhance chromosome-level assembly, along with the different potential of these methods for resolving some of these difficulties, are considered here.

2. THE STRENGTHS AND LIMITATIONS OF SEQUENCING TECHNOLOGIES

De novo assembly from Next-Generation Sequencing (NGS) data [12] requires the grouping of short sequence reads into contigs using regions of overlap [13]. Contigs are then assembled into scaffolds using ‘paired-end’ reads which are the two sequenced ends of a longer DNA insert from a library of cloned DNA fragments [14]. The success of contig assembly depends in part on the size of the sequenced DNA fragments. Paired-end sequencing of larger library fragments or longer sequence reads is more useful for covering genome regions where the contigs are separated by tracts of repetitive DNA in which accurate sequencing is difficult. The sequence scaffolds may then be further integrated into chromosome assemblies often through the use of a chromosome-spanning genetic or physical map. However, such maps may be unavailable, technically challenging to construct or too costly to generate. For this reason, chromosome-level assemblies can be particularly difficult to achieve.

The problems presented by the alternative sequencing technologies (Sanger, short read ‘next generation’ and longer read ‘third generation’ sequencing) relate to variation in cost, throughput, error rate and coverage of difficult-to-sequence regions. Broadly, Sanger sequencing is low-throughput and relatively expensive but with low error rates and cleaner resolution of some complex genomic regions. Short read sequencing is relatively inexpensive and high-throughput but alignment difficulties confound analysis of complex regions. Although using longer reads may resolve some of these regions the error rates may be higher and costs increased. Sanger technology using clones is much better at identifying segmental duplications and so relying on shotgun methods may miss important lineage-specific regions, as happened in earlier drafts of the mouse genome sequence. Refinement of the mouse genome yielded 140 megabases of new sequence which was missing from the earlier assembly. The new assembly included 1000s of evolutionary recent genes which are specific to mice, along with genome regions containing much duplication which were resistant to sequencing [15].

De novo assembly using short sequence reads exacerbate the problems because of the difficulty in distinguishing between, for example, paralagous loci which have limited sequence divergence. Progress in the development of algorithms which facilitate *de novo* assembly using short read sequences has not been sufficient to date to support the use of these methods for routine assembly of genomes [16]. In general, *de novo* assemblies using short reads require 5-10 fold more highly fragmented sequence data than was used for developing the finished human sequence [6].

A combination of short and long read technologies has been used to improve assembly [17]. The genome assembly of the Loblolly Pine (*Pinus taeda*) [18] was developed firstly through constructing a fragmented draft assembly which was based on Illumina short sequence reads. The assembly was subsequently improved using long sequence reads from Pacific Biosciences (PacBio) Single Molecule Real Time Sequencing [19]. However, in general, the longer reads (each ~10 kb) have a relatively high error rate (~15%) and therefore more sequencing (higher ‘read depths’), with increased costs, is required to achieve robust sequence assemblies.

3. SPECIES-SPECIFIC CONSIDERATIONS

The size, spacing and arrangement of repetitive DNA regions can present problems for assemblers but the extent to which these regions confound assembly can depend on genome composition for a particular species. The genomes of some highly inbred organisms, such as laboratory strains of worms, fish, or mice, may be particularly amenable to sequencing because of low levels of polymorphism which can greatly simplify assembly. The degree of genomic polymorphism can vary in unexpected ways, for example, the genome of the sea squirt (*Ciona savignyi*) [20] has a considerably higher density of polymorphic sites (1 per 50-100 base pairs) compared to the human genome (1 per 1000 base pairs) [2]. Sequence assembly is difficult because the differences between reads due to polymorphism may be misinterpreted by assemblers and errors introduced in the sequence. The alternative computer algorithms used for assembly have variable strengths and the choice of the best approach is likely to vary by organism [9]. For example, solutions which are effective for small bacterial genomes may be impossible and/or impractical for use with eukaryotic genomes which have far more extensive repetitive sequences.

4. A MAP IS NEEDED

Because sequencing generates contigs of limited length there are usually numerous gaps spanning regions which are resistant to short range sequencing. Badly located contigs, along with incorrect ordering within contigs, reduce the value of the assembly for comparative genomics because as false breakpoints may lead to incorrect evolutionary inferences. Lewin *et al*., [21] point out that draft, and even ‘finished’ sequences, can fail to provide enough fine scale ‘granularity’ to allow deep analysis of the expansion and adaptation of gene families across different evolutionary lineages. They propose continued development and application of high-resolution physical maps of chromosomes as essential frameworks for annotation and evolutionary analysis of genomes. Independent physical maps, when used together with sequence assemblies, facilitate the ordering and orientation of sequence contigs for whole genome assembly.

Framework maps include genetic linkage maps which were utilised successfully in the development of the scaffold of the human genome sequence [22]. Linkage data have been used for the correction of draft genomes [23] and when combined with *de novo* sequencing, have enabled improved sequence assemblies [11, 21]. Linkage maps depend on analysis of genetic recombination between related individuals using genotyped polymorphic markers positioned along the length of chromosomes. Map construction usually requires the establishment of a ‘mapping population’ which in turn depends on substantial investment of resources. Recombination patterns can enable recognition of linkage groups specifying polymorphism order and genetic distance ultimately forming whole chromosome maps as scaffolds for assembly of shorter sequence contigs. However, because there are a limited number of meiotic breaks along individual chromosomes within each generation of the mapping population the resolution of the linkage map may be insufficient to provide a framework for assembly of relatively short sequence contigs. Furthermore, the computational challenges associated with reliably ordering marker polymorphisms to form a linkage map should not be underestimated [11]. Because genetic linkage maps require analysis of meiotic events across pedigrees they may be too costly or impractical to construct for many species.

Radiation Hybrid (RH) maps [24] were used successfully as physical map scaffolds during the construction of the human reference genome sequence. These, along with Bacterial Artificial Chromosome (BAC) derived maps [25] and Zoo-FISH [26], have been used to anchor genome sequence assemblies of species such as macaque, dog, rat, horse, opossum and cattle [21]. More recent physical mapping developments include sequence motif maps made using optical mapping [27]. Optical mapping labels specific motifs in single DNA molecules and these molecules are then stretched uniformly allowing measurement of physical map distances between the labelled motifs. This is a high-throughput method which uses nanofluidic chips with nanochannels to maintain long DNA molecules in a uniformly elongated state. Fluorescently labelled DNA molecules are imaged after being drawn into nanochannels. These maps may inform contig order and orientation and highlight incorrectly assembled contigs, which can then be corrected iteratively to improve the assembly. The average spacing between motifs is of the order of ~9 kb in the Major Histocompatibility Complex (MHC) region example described [27]. Because long chromosome fragments are involved phased haplotypes may also be resolved. Sequence motif maps are proposed as scaffolds for the *de novo* assembly of sequence contigs including the assembly of sequences in complex genomic regions. However, the technique is technically challenging and only offers coverage at mid-range (multiple kilobase) resolution [6].

Genome-wide patterns of chromatin interaction offer another strategy for developing a long-range scaffold for chromosome assembly [6]. The Hi-C method relies on the three-dimensional folding of chromosomes which bring distant functional elements into close proximity. Proximity ligation enables loci which are interacting in the nucleus to be fixed such that the links between them remain stable once the DNA is fragmented. Fragments containing these junctions are sequenced to produce a catalog of interacting fragments [28, 29]. Analysis of these data reveals that the genome has components with open, accessible and active chromatin and a dense compartment with closed, inaccessible and inactive chromatin. Burton *et al*., [6] describe a method to extract long-range information for ordering and orienting genome sequences in chromosome assemblies. The technique relies on the expectation that intra-chromosomal contacts are much more frequent that inter-chromosomal. Specifically, although the probability of interaction decays rapidly as a function of distance, even loci separated by more than 200 Mb are more likely to show a higher frequency of interaction than loci on different chromosomes [28]. The program LACHESIS (ligating adjacent chromatin enables scaffolding *in situ*) [6] uses the signal due to genomic proximity for ultra-long-range scaffolding to support genome assembly. Given a set of contigs from a draft sequence assembly and a corresponding set of Hi-C links, LACHESIS recognises contigs from the same chromosome as having more Hi-C links enabling clustering of contigs into chromosome groups. Further scaffolding is achieved because contigs located in close proximity tend to have more links between them. This enables ordering and eventually precise positioning using the links between close contigs to facilitate orientation of fragments. The method has been shown effective for assigning, ordering and orienting chromosome contigs even including the spanning of centromere gaps which may be multiple megabases in size. In simulations, using human genome data, this approach achieved 98% accuracy for assigning scaffolds to chromosome groups and 99% accuracy in ordering and orienting scaffolds within chromosome groups. One technical difficulty is the requirement for substantial quantities of DNA. The authors recognise that, although the method enables chromosome-scale scaffolding, the contiguity needed for the initial *de novo* assembly (fragments size in the order of 50 kb) might be difficult to achieve for application to some organisms. Therefore the need for methods which provide ‘intermediate’ contiguity is pressing.

5. CONTIG ORDERING AND ORIENTATION USING CROSS-SPECIES SYNTENY

A number of methods use cross-species synteny to identify relationships between sequence fragments and this information can be useful to guide chromosome assembly. The sequenced genomes of closely related species are likely to have been subject to relatively few sequence inversions, translocations, fusions and fissions during evolution. Understanding the pattern of chromosome rearrangement throughout genome evolution is essential for resolving mechanisms of speciation and species adaptation. The application of synteny relationship data for understanding relationships between contigs is facilitated where there are existing reference genomes for closely-related species which have robustly established and reliable sequence assembly (for example comparison between human and chimpanzee genomes) [30].

Many of the larger differences between species arise because of the existence of a relatively small number of Evolutionary Breakpoint Regions (EBRs) which are ‘fragile’ chromosome sites which are prone to reorganisation [31, 32]. While broad relationships between larger chromosome segments might be readily seen the assembly of more complex sequence-based assembly data, which contain numerous genome reorganisation events, including duplications and deletions, presents analytical challenges.

Bourque and Pevzner [33] develop an algorithm for understanding rearrangements based on 'reversal' distance which enables phylogenetic tree construction. Genome rearrangement analysis constructs phylogenetic trees based on the orders of genes. Inversions are frequently observed and these are referred to as reversals: gene orders can be represented in the form of signed integers where the sign denotes orientation. If two alternative orders are considered as permutations then the reversal distance [34] is defined as the smallest number of reversals needed to convert one permutation into the other. Other methods apply graph theory to use information about modern species to infer segment order in the ancestral genome [35]. Kim *et al*., [36] describe 'reference-assisted chromosome assembly' (RACA) an algorithm to reliably order and orient sequence scaffolds into chromosome assemblies using paired-end read and comparative genome information. The algorithm considers comparisons between the target species sequences, the reference genome of a closely related species and one or more outgroup genomes which provide evolutionary information. The algorithm uses comparative genome information to compute scores which reflect the likelihood that two syntenic fragments are adjacent in the target genome considering adjacencies in the reference and outgroup genomes, along with evidence from paired-end reads which support the adjacency. The RACA algorithm constructs chains of syntenic fragments stepwise using adjacency scores as weights. RACA is shown to be capable of predicting genome-wide chromosome organisation for a *de novo* sequenced species without the need for a genetic or physical map. However, success depends on the closeness of the relationship between the target and reference species which impacts the alignability of the target sequences to the reference. Therefore the method is particularly useful for assemblies within the major phylogenetic clades of vertebrates where there are many existing genome assemblies [37]. For example, the authors were able to use RACA to reconstruct chromosomes from Tibetan antelopes (*Pantholops hodgsonii*) using homology with cattle chromosomes.

For sequence-based and genome-wide comparisons, where numerous rearrangements are likely, resolving the complex rearrangement between multiple species is extremely challenging. However, precise resolution of, for example, EBRs is important because properties of the underlying sequence within these regions might reflect the genes which determine differences between species lineages, justifying the effort for fine-scale chromosome reconstruction [21, 38]. As more genome sequences for diverse species are assembled new opportunities arise for exploiting syntenic relationships to guide chromosome level construction. Kim *et al*., [36] describe the DESCHRAMBLER algorithm which considers syntenic regions from whole genome sequence data. The algorithm is evolution-based evaluating adjacency probabilities for pairs of syntenic fragments in a target ancestor based on the order and orientation of the segments in related species. The most likely paths are identified to represent order and orientation of the fragments in the target ancestor. Evaluation of the algorithm determined 162 chromosomal breakpoints (at 300 Kb resolution) over 105 million years of mammalian evolution (from the eutherian ancestor's genome to humans). A reliable reference genome (the human genome was used for these evaluations) is critical for success in forming the backbone for alignment of orthologous chromosome regions in different species. Such a procedure is more easily achieved in, for example, the mammalian lineage, given the quality of the human reference sequence, but for less well studied lineages, such as many that will be represented in the EBP, there are likely to be greater difficulties.

6. ORDERING AND ORIENTATION BY LINKAGE DISEQUILIBRIUM

In human populations Linkage Disequilibrium (LD) usually extends further than the longer sequence reads (up to ~50 kb [39]), but extends further still in more inbred species. LD maps have a much higher resolution than linkage maps because LD structure depends in part on recombination events accumulated over many generations. For this reason LD structure may usefully inform the ordering and orientation of relatively short sequence contigs of up to a few kilobases in length. Since LD maps are constructed from population data (typically unrelated individuals from a population) there is no requirement to establish a mapping population in which recombination events are traced in families, as required for the construction of linkage maps. Ennis *et al*., [40] consider the use of LD for discriminating between draft locus orders in a small genomic region. They concluded that LD can be a powerful strategy for identifying the correct orders of polymorphisms in a genome region.

Khatkar *et al*., [41] and Jones *et al*., [42] employ the LODE (Locus Ordering by Linkage Disequilibrium) algorithm to position 'orphan' SNPs within genetic linkage map scaffolds. This approach has been shown to improve assembly by locally increasing marker coverage and therefore aiding the positioning of shorter sequence contigs. Through this method the resolution of the linkage map scaffold can be substantially increased. The method is also useful for validating the quality of an assembly by testing the strength of evidence for the positions of SNPs already located in the assembly. Jones *et al*., [42] applied this approach to develop an integrated linkage and LD map of the Pacific White Shrimp (*Litopenaeus vannamei*). A linkage scaffold and LD data from 75 individuals was used as the basis to position several hundred SNPs within the scaffold.

Linkage disequilibrium maps [39, 43] are constructed from population data but are closely analogous to the genetic linkage map because LD structure is determined to a large degree by accumulated recombination events. Pengelly and Collins [44] describe a method for ordering, orienting and positioning sequenced contigs using LD maps. The maps are constructed using SNP genotype data from unrelated individuals. The pattern of LD across a contig, where the internal order of SNPs is known, is quantified in Linkage Disequilibrium Units (LDUs) such that one LDU is the distance along the chromosome over which LD declines to 'background' levels. The LD structure of a chromosome includes regions with strong LD, which have low haplotype diversity (LD blocks), interrupted by narrow regions of intense LD breakdown which broadly align with recombination hotspots [45, 46].

The evaluation of the method considers alternative orders and orientations for a small number of sequenced contigs for which SNP genotype data are available. A simulation study considers alternative orders and orientations for three contigs in which the combined re-arranged assembly is treated as a single map. The map with the shortest LDU length corresponds to the correct order and orientation. Alternative incorrect orders and orientations have elevated LDU map lengths. The approach may be less able to resolve misorientated individual contigs but is more effective at determining the correct contig order. Because unrelated individuals in population samples are used to construct an LD map this approach is relatively inexpensive and much less technically demanding than construction of a linkage map through establishing a mapping population. However, the method requires genotyping of SNPs for multiple unrelated individuals to effectively 'tag' each contig. Such data are likely to be available for organisms where there is a focussed effort to establish covering SNP panels, for example for species of actual, or potential, commercial interest.

7. DISCUSSION

It is clear that short-read sequencing strategies alone will not be suitable for establishing reliable orders of contigs on scaffolds to enable detailed investigation of comparative genomes [2]. The presence of extensive repetitive sequences which are resistant to short read sequencing, along with closely similar representatives of large gene families and the presence of abundant copy number variation, present major difficulties. Longer reads (for example read lengths of the order of 200 base pairs) [47] may be effective for chromosome level assembly in the absence of a covering map but have had more limited application for assembly to date [9].

The EBP [2] aims to construct high quality reference genomes for a member of each eukaryotic family to guide the assembly of lower quality, but informative, sequence builds for other species in the same family. The authors recognise that the longer read techniques and optical mapping [27] and other methods to build robust scaffolds such as Hi-C [6], can be technically challenging and require large quantities of high-quality DNA with consequently increased costs and difficulties for DNA acquisition and processing. Pilot studies are underway to inform future efforts recognising that a multi-technique approach is likely to be essential to achieve the family-specific aim. The extent to which technically demanding approaches, such as optical mapping, can be scaled-up to achieve the project's aims is subject to further research. A major consideration is the relationship between sequence contig size and coverage and the resolution of the chromosome-wide map. Maps with low resolution might be too coarse to guide positioning and orientation of relatively short sequence contigs. Map-building strategies vary in their resolution. Within linkage maps, for example, polymorphic markers might be reliably ordered only where they are at least one megabase apart providing a very coarse framework only likely to be useful in locating larger sequence contigs. LD maps are much higher resolution and the two have been used successfully together to create a finer covering map.

CONCLUSION

The impact of differences in sequence composition across species, including variation in the proportion and size of repetitive regions alongside variability in polymorphism, depends on the family and species under consideration. Furthermore the requirement for high quality sequence assembly is likely to vary widely with species. Species of known or potential commercial value, or key species for use as reference genomes for a family, will be subject to higher requirement for sequence build 'polishing'. Because assembly algorithms are focussed on developing longer contigs incorrect joins and orientations are a possibility [9] therefore sequence polishing needs to expect this possibility and revision using alternative sources of information (such as LD structure) have the potential to resolve these errors.

A combination of strategies is likely to be the best way to optimise the value of the vast sequence datasets which are now being developed or planned. Tang *et al*., [48] describe a tool called ALLMAPS which recognises the importance of combined approaches for chromosomal assemblies. ALLMAPS is designed to integrate multiple maps within a unified framework. The primary focus is on integration of genetic maps but extension to other types of data (for example optical maps and synteny information) is described. Ongoing development of this and similar approaches aimed at recovering more useful information for assembly from species with diverse sequences is important. Although sequencing costs are reducing the scale of the developing sequencing projects mean that the value of low-coverage sequence data needs to be maximised. Optimal strategies will be fine-tuned as more information is gathered and in particular knowledge of differences in sequence composition across genomes emerges. This information will facilitate choice of mapping strategy including the need for and composition of a suitable framework map and ways that might be achieved. In this way the challenge of genome assembly might be met with subsequent dramatic yields in understanding of the potential of diverse genome sequences.

CONSENT FOR PUBLICATION

Not applicable.

CONFLICT OF INTEREST

The author confirms there are no conflicts of interest related to the article contents.

ACKNOWLEDGEMENTS

None declared.

REFERENCES

[1] Kim J, Larkin DM, Cai Q, *et al.* Reference-assisted chromosome assembly. Proc Natl Acad Sci USA 2013; 110(5): 1785-90. [http://dx.doi.org/10.1073/pnas.1220349110] [PMID: 23307812]

[2] Lewin HA, Robinson GE, Kress WJ, *et al.* Earth BioGenome Project: Sequencing life for the future of life. Proc Natl Acad Sci USA 2018; 115(17): 4325-33. [http://dx.doi.org/10.1073/pnas.1720115115] [PMID: 29686065]

[3] Pennisi E. Sequencing all life captivates biologists. Science 2017; 355(6328): 894-5. [http://dx.doi.org/10.1126/science.355.6328.894] [PMID: 28254891]

[4] Schatz MC, Delcher AL, Salzberg SL. Assembly of large genomes using second-generation sequencing. Genome Res 2010; 20(9): 1165-73. [http://dx.doi.org/10.1101/gr.101360.109] [PMID: 20508146]

[5] Turnbull C, Scott RH, Thomas E, *et al.* The 100000 genomes project: Bringing whole genome sequencing to the NHS. BMJ: British Medical Journal 2018; 361 K1687.

[6] Burton JN, Adey A, Patwardhan RP, Qiu R, Kitzman JO, Shendure J. Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. Nat Biotechnol 2013; 31(12): 1119-25. [http://dx.doi.org/10.1038/nbt.2727] [PMID: 24185095]

[7] Lander ES, Linton LM, Birren B, *et al.* Initial sequencing and analysis of the human genome. Nature 2001; 409(6822): 860-921. [http://dx.doi.org/10.1038/35057062] [PMID: 11237011]

[8] International human genome sequencing consortium. Finishing the euchromatic sequence of the human genome. Nature 2004; 431(7011): 931-45. [http://dx.doi.org/10.1038/nature03001] [PMID: 15496913]

[9] Baker M. *De novo* genome assembly: What every biologist should know. Nat Methods 2012; 9(4): 333-7. [http://dx.doi.org/10.1038/nmeth.1935]

[10] Denton JF, Lugo-Martinez J, Tucker AE, Schrider DR, Warren WC, Hahn MW. Extensive error in the number of genes inferred from draft genome assemblies. PLOS Comput Biol 2014; 10(12): e1003998. [http://dx.doi.org/10.1371/journal.pcbi.1003998] [PMID: 25474019]

[11] Fierst JL. Using linkage maps to correct and scaffold *de novo* genome assemblies: Methods, challenges, and computational tools. Front Genet 2015; 6: 220. [http://dx.doi.org/10.3389/fgene.2015.00220] [PMID: 26150829]

[12] Miller JR, Koren S, Sutton G. Assembly algorithms for next-generation sequencing data. Genomics 2010; 95(6): 315-27. [http://dx.doi.org/10.1016/j.ygeno.2010.03.001] [PMID: 20211242]

[13] Pevzner PA, Tang H, Waterman MS. An Eulerian path approach to DNA fragment assembly. Proc Natl Acad Sci USA 2001; 98(17): 9748-53. [http://dx.doi.org/10.1073/pnas.171285098] [PMID: 11504945]

[14] Chaisson MJP, Brinza D, Pevzner PA. De novo fragment assembly with short mate-paired reads: Does the read length matter? Genome Res 2008; gr-079053.
[PMID: 19056694]

[15] Church DM, Goodstadt L, Hillier LW, *et al.* Lineage-specific biology revealed by a finished genome assembly of the mouse. PLoS Biol 2009; 7(5): e1000112.
[http://dx.doi.org/10.1371/journal.pbio.1000112] [PMID: 19468303]

[16] Compeau PEC, Pevzner PA, Tesler G. How to apply de bruijn graphs to genome assembly. Nat Biotechnol 2011; 29(11): 987-91.

[17] Li R, Zhu H, Ruan J, *et al. De novo* assembly of human genomes with massively parallel short read sequencing. Genome Res 2010; 20(2): 265-72.
[http://dx.doi.org/10.1101/gr.097261.109] [PMID: 20019144]

[18] Zimin AV, Stevens KA, Crepeau MW, *et al.* An improved assembly of the loblolly pine mega-genome using long-read single-molecule sequencing. Gigascience 2017; 6(1): 1-4.
[http://dx.doi.org/10.1093/gigascience/giw016] [PMID: 28369353]

[19] Koren S, Schatz MC, Walenz BP, *et al.* Hybrid error correction and *de novo* assembly of single-molecule sequencing reads. Nat Biotechnol 2012; 30(7): 693-700.
[http://dx.doi.org/10.1038/nbt.2280] [PMID: 22750884]

[20] Small KS, Brudno M, Hill MM, Sidow A. Extreme genomic variation in a natural population. Proc Natl Acad Sci USA 2007; 104(13): 5698-703.
[http://dx.doi.org/10.1073/pnas.0700890104] [PMID: 17372217]

[21] Lewin HA, Larkin DM, Pontius J, O'Brien SJ. Every genome sequence needs a good map. Genome Res 2009; 19(11): 1925-8.
[http://dx.doi.org/10.1101/gr.094557.109] [PMID: 19596977]

[22] Dausset J, Cann H, Cohen D, Lathrop M, Lalouel JM, White R. Centre d'etude du polymorphisme humain (CEPH): Collaborative genetic mapping of the human genome. Genomics 1990; 6(3): 575-7.
[http://dx.doi.org/10.1016/0888-7543(90)90491-C] [PMID: 2184120]

[23] Hahn MW, Zhang SV, Moyle LC. Sequencing, assembling, and correcting draft genomes using recombinant populations. G3 (Bethesda) 2014; 4(4): 669-79.
[http://dx.doi.org/10.1534/g3.114.010264] [PMID: 24531727]

[24] Gyapay G, Schmitt K, Fizames C, *et al.* A radiation hybrid map of the human genome. Hum Mol Genet 1996; 5(3): 339-46.
[http://dx.doi.org/10.1093/hmg/5.3.339] [PMID: 8852657]

[25] McPherson JD, Marra M, Hillier L, *et al.* A physical map of the human genome. Nature 2001; 409(6822): 934-41.
[http://dx.doi.org/10.1038/35057157] [PMID: 11237014]

[26] Chowdhary BP, Raudsepp T, Frönicke L, Scherthan H. Emerging patterns of comparative genome organization in some mammalian species as revealed by Zoo-FISH. Genome Res 1998; 8(6): 577-89.
[http://dx.doi.org/10.1101/gr.8.6.577] [PMID: 9647633]

[27] Lam ET, Hastie A, Lin C, *et al.* Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly. Nat Biotechnol 2012; 30(8): 771-6.
[http://dx.doi.org/10.1038/nbt.2303] [PMID: 22797562]

[28] Lieberman-Aiden E, van Berkum NL, Williams L, *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science 2009; 326(5950): 289-93.
[http://dx.doi.org/10.1126/science.1181369] [PMID: 19815776]

[29] Van Berkum NL, Lieberman-Aiden E, Williams L, *et al.* Hi-C: A method to study the three-dimensional architecture of genomes. J Vis Exp 2010; 39: 1869.

[30] Murphy WJ, Agarwala R, Schäffer AA, *et al.* A rhesus macaque radiation hybrid map and comparative analysis with the human genome. Genomics 2005; 86(4): 383-95.
[http://dx.doi.org/10.1016/j.ygeno.2005.05.013] [PMID: 16039092]

[31] Ruiz-Herrera A, Castresana J, Robinson TJ. Is mammalian chromosomal evolution driven by regions of genome fragility? Genome Biol 2006; 7(12): R115.

[32] Larkin DM, Pape G, Donthu R, Auvil L, Welge M, Lewin HA. Breakpoint regions and homologous synteny blocks in chromosomes have different evolutionary histories. Genome Res 2009; 19(5): 770-7.
[http://dx.doi.org/10.1101/gr.086546.108] [PMID: 19342477]

[33] Bourque G, Pevzner PA. Genome-scale evolution: Reconstructing gene orders in the ancestral species. Genome Res 2002; 12(1): 26-36.
[PMID: 11779828]

[34] Sankoff D, Leduc G, Antoine N, Paquin B, Lang BF, Cedergren R. Gene order comparisons for phylogenetic inference: Evolution of the mitochondrial genome. Proc Natl Acad Sci USA 1992; 89(14): 6575-9.
[http://dx.doi.org/10.1073/pnas.89.14.6575] [PMID: 1631158]

[35] Ma J, Zhang L, Suh BB, *et al.* Reconstructing contiguous regions of an ancestral genome. Genome Res 2006; 16(12): 1557-65.

[http://dx.doi.org/10.1101/gr.5383506] [PMID: 16983148]

[36] Kim J, Farré M, Auvil L, *et al.* Reconstruction and evolutionary history of eutherian chromosomes. Proc Natl Acad Sci USA 2017; 114(27): E5379-88.
[http://dx.doi.org/10.1073/pnas.1702012114] [PMID: 28630326]

[37] Genome 10K Community of Scientists. Genome 10K: A proposal to obtain whole-genome sequence for 10,000 vertebrate species. J Hered 2009; 100(6): 659-74.
[http://dx.doi.org/10.1093/jhered/esp086] [PMID: 19892720]

[38] Groenen MA, Archibald AL, Uenishi H, *et al.* Analyses of pig genomes provide insight into porcine demography and evolution. Nature 2012; 491(7424): 393-8.
[http://dx.doi.org/10.1038/nature11622] [PMID: 23151582]

[39] Tapper W, Collins A, Gibson J, Maniatis N, Ennis S, Morton NE. A map of the human genome in linkage disequilibrium units. Proc Natl Acad Sci USA 2005; 102(33): 11835-9.
[http://dx.doi.org/10.1073/pnas.0505262102] [PMID: 16091463]

[40] Ennis S, Collins A, Tapper W, Murray A, MacPherson JN, Morton NE. Allelic association discriminates draft orders. Ann Hum Genet 2001; 65(Pt 5): 503-4.
[http://dx.doi.org/10.1017/S000348000100879X] [PMID: 11811150]

[41] Khatkar MS, Hobbs M, Neuditschko M, Sölkner J, Nicholas FW, Raadsma HW. Assignment of chromosomal locations for unassigned SNPs/scaffolds based on pair-wise linkage disequilibrium estimates. BMC Bioinformatics 2010; 11(1): 171.
[http://dx.doi.org/10.1186/1471-2105-11-171] [PMID: 20370931]

[42] Jones DB, Jerry DR, Khatkar MS, *et al.* A comparative integrated gene-based linkage and locus ordering by linkage disequilibrium map for the Pacific white shrimp, *Litopenaeus vannamei.* Sci Rep 2017; 7(1): 10360.
[http://dx.doi.org/10.1038/s41598-017-10515-7] [PMID: 28871114]

[43] Zhang W, Collins A, Maniatis N, Tapper W, Morton NE. Properties of Linkage Disequilibrium (LD) maps. Proc Natl Acad Sci USA 2002; 99(26): 17004-7.
[http://dx.doi.org/10.1073/pnas.012672899] [PMID: 12486239]

[44] Pengelly RJ, Collins A. Linkage disequilibrium maps to guide contig ordering for genome assembly. Bioinformatics 2018. in press.
[http://dx.doi.org/10.1093/bioinformatics/bty687] [PMID: 30101310]

[45] Jeffreys AJ, Kauppi L, Neumann R. Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. Nat Genet 2001; 29(2): 217-22.
[http://dx.doi.org/10.1038/ng1001-217] [PMID: 11586303]

[46] Lau W, Kuo TY, Tapper W, Cox S, Collins A. Exploiting large scale computing to construct high resolution linkage disequilibrium maps of the human genome. Bioinformatics 2007; 23(4): 517-9.
[http://dx.doi.org/10.1093/bioinformatics/btl615] [PMID: 17142813]

[47] Sundquist A, Ronaghi M, Tang H, Pevzner P, Batzoglou S. Whole-genome sequencing and assembly with high-throughput, short-read technologies. PLoS One 2007; 2(5): e484.
[http://dx.doi.org/10.1371/journal.pone.0000484] [PMID: 17534434]

[48] Tang H, Zhang X, Miao C, *et al.* ALLMAPS: Robust scaffold ordering based on multiple maps. Genome Biol 2015; 16(1): 3.
[http://dx.doi.org/10.1186/s13059-014-0573-1] [PMID: 25583564]