Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Discovery Notes

Sequence Analysis

Linkage disequilibrium maps to guide contig ordering for genome assembly

Reuben J. Pengelly [1,*] **and Andrew Collins** [1]

[1] Genetic Epidemiology & Bioinformatics, Faculty of Medicine, University of Southampton, SO16 6YD

* To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Efforts to establish reference genome sequences by *de novo* sequence assembly have to address the difficulty of linking relatively short sequence contigs to form much larger chromosome assemblies. Efficient strategies are required to span gaps and establish contig order and relative orientation. We consider here the use of linkage disequilibrium (LD) maps of sequenced contigs and the utility of LD for ordering, orienting and positioning linked sequences. LD maps are readily constructed from population data and have at least an order of magnitude higher resolution than linkage maps providing the potential to resolve difficult areas in assemblies. We empirically evaluate a linkage disequilibrium map-based method using single nucleotide polymorphism genotype data in a 216 kilobase region of human 6p21.3 from which three shorter contigs are formed.

Results: LD map length is most informative about the correct order and orientation and is suggested by the shortest LD map where the residual error variance is close to one. For regions in strong LD this method may be less informative for correcting inverted contigs than for identifying correct contig orders. For positioning two contigs in linkage disequilibrium with each other the inter-contig distances may be roughly estimated by this method.

Availability: The LDMAP program is written in C for a linux platform and is available at https://www.soton.ac.uk/genomicinformatics/research/ld.page

Contact: R.J.Pengelly@soton.ac.uk

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Next generation sequencing has been transformational for understanding the genomes of many species including of course, humans, where these technologies are revolutionising medicine. However, for many organisms the difficulty of establishing reliable reference genome assemblies negatively impacts the application of NGS for genome research. Many draft genomes contain thousands of sequenced segments (contigs) but with very limited information on how these can be assembled into chromosome sequences. Incomplete assemblies are a problem for molecular and developmental studies since individual genes are likely to be broken, badly annotated and lack the correct genomic context to facilitate comparative genome evolutionary studies (Baker, 2012, Denton et al, 2014, Fierst, 2015).

A number of factors impact the success of *de novo* assembly and therefore reduce the utility of genomic tools which interpret genome sequences (Hahn et al, 2014). For example, the lack of a linkage map may mean there is no chromosome-wide 'scaffold' on which sequence contigs can be assembled. Complex genome features, including regions with repeat structures, can prevent completion of sequence assemblies because of substantial gaps preventing connection of shorter sequence contigs. Furthermore, short sequence reads used in NGS exacerbate the problems because of the difficulty in distinguishing between, for example, paralogous loci which have limited sequence divergence. In the absence of a strategy to anchor, connect, orient and position contigs constructed through NGS the value of such sequencing effort is much reduced.

A number of routes to improve assembly have been considered and applied. A recent study (Zimin et al, 2017) describes genome assembly for the Loblolly Pine (*Pinus taeda*). Initially a draft but fragmented assembly was built using short Illumina sequence reads (100-250 base-pairs per read). The assembly was improved using long sequence reads through Pacific

Biosciences (PacBio) Single Molecule Real Time Sequencing (Koren et al, 2012). However, the longer reads (each 10 kb) have a relatively high error rate (ca. 15%) and higher read depths (with increasing costs) are required to compensate. Furthermore, these longer sequence reads have limited utility for spanning larger gaps between contigs.

Linkage maps, combined with *de novo* sequencing have been used to enhance sequence assembly (Lewin et al, 2009, Fierst, 2015). Linkage maps rely on the analysis of genetic recombination between related individuals by genotyping those individuals each with the same set of markers. Establishing a 'mapping population' in which to quantify patterns of recombination requires significant investment of resources. Recombination patterns identify markers inherited together thereby enabling establishment of linkage groups which specify order and genetic distance, potentially spanning whole chromosomes. These maps may provide whole-chromosome scaffolds upon which shorter sequence contigs can be assembled. However, there may be a disconnect between the much lower-resolution linkage map (which has a resolution on multiple-megabase scales) and much smaller sequence contigs. Furthermore, the computational challenges associated with ordering of markers in linkage maps should not be underestimated (Fierst, 2015).

Linkage disequilibrium extends further than the longest sequence reads (to ca. 50 kb in humans (Tapper et al, 2005), but more in more inbred species) so has the potential to inform contig ordering and orientation over much larger distances than long sequence reads. Since LD maps are constructed from population data there is no requirement to establish a mapping population as is required for construction of linkage maps. Furthermore, LD has much higher resolution than linkage because it reflects, in part, historical recombination over the population's history. Ennis et al (2001) evaluated the use of linkage disequilibrium for discriminating between draft locus orders in a small genomic region. They concluded that LD can be a powerful strategy for distinguishing alternative orders of polymorphisms. We extend this concept to consider linkage disequilibrium maps (Zhang et al, 2002, Tapper et al, 2005) which are analogous to the genetic linkage map. We empirically evaluate the utility of linkage disequilibrium maps for ordering, orientation and positioning of three contigs to evaluate the strengths and limitations of this approach.

2 Methods

2.1 The linkage disequilibrium map algorithm

We empirically evaluated the utility of linkage disequilibrium maps for resolving sequence contig positioning, orientation and distances using the program LDMAP (Lau et al, 2006, Kuo et al, 2007). The program constructs LD maps according to the Malécot-Morton model:

$$\hat{\rho} = (1 - L)Me^{-\epsilon d} + L \qquad (1)$$

where $\hat{\rho}$ is the association between a pair of SNPs, the asymptote L is 'background' association which is not due to linkage (and is increased in small population samples and with residual population structure), M reflects association at zero distance with values *ca.* 1 consistent with monophyletic haplotypes and < 1 with polyphyletic inheritance, ϵ is the rate of LD decline, and d is the physical distance in kilobases between SNPs. For LD map construction parameters ϵ, L and M are estimated iteratively for each SNP interval in the map.

LDU distance is computed as the product ϵd and represented as cumulative map distances similar to the centimorgan scale. One LDU corresponds to the (highly variable) physical distance over which LD declines to background levels. LDUs plotted against chromosome location display 'steps' which reflect intense breakdown in LD (typically 'recombination hotspots'), and plateaus aligning with blocks of low haplotype diversity.

The LDU map is constructed iteratively with the fit of the pairwise SNP association data to the kilobase map established through composite likelihood. Pairwise data enter composite log likelihood as:

$$ln(lk) = -\sum_{i=1}^{n} \frac{K_\rho(\hat{\rho}_i - \rho_i)^2}{2} \qquad (2)$$

where the summation is over informative SNP pairs, ρ is the observed association between the i^{th} SNP pair, $\hat{\rho}$ is the fitted values from the Malécot-Morton model (equation 1) where the log likelihood is iteratively recalculated using the computed LDU distance for term d in place of the physical distance; K_ρ is the information about ρ for each SNP-SNP pair. The fit of the completed LDU map to the pairwise SNP association data is quantified through the error variance (EV) (Collins & Morton, 1998; Collins et al. 1999) as:

$$\mathrm{EV} = \frac{-2ln(lk)}{\nu} \qquad (3)$$

where degrees of freedom, ν is the number of SNP pairs. Because $-2ln(lk)$ has a χ^2 distribution, the residual EV is expected to be ca. 1 for a good fit of the LDU map and pairwise SNP association data.

2.2 SNP data sample

The LDMAP program accepts input SNP data in TPED format (Purcell et al, 2007) and each SNP has a known (or estimated) location in base pairs from an origin (for example, from the p-telomere for a whole chromosome map). We consider three sequenced contigs in which the internal (within-contig) relative locations of SNPs are known from sequencing. We illustrate and evaluate the utility of the LDMAP concept for ordering, orienting and positioning contigs using the human SNP data sample from Jeffreys et al (2001) for which the LDU map, which shows considerable variation in LD intensity across the region, was described by Zhang et al (2002). The chosen example considers three contigs; this method applies equally to any number of contigs, though becomes more computationally challenging with a greater number of contigs owing to the greater number of positional permutations.

Applying the LDMAP cut-offs for excluding rare SNPs (minor allele frequency <0.05, 47 SNPs excluded) and conformation to Hardy-Weinberg equilibrium (HWE, one SNP deviating with $P < 0.001$ excluded) the LD map contains 248 SNPs. We divided the map into three regions with roughly equal numbers of SNPs: region A from 0-119.57 kb, region B from 119.57-184.1 kb and region C from 184.1-215.6 kb (Figure 1). The breakpoint between regions A and B interrupts a region of rapid LD breakdown, aligning with a known recombination hotspot (Jeffreys et al, 2001, Zhang et al, 2002). This enables evaluation of the impact on the method for assembling sequences where two adjacent regions have relatively weak LD between them.

We compared alternative orders and orientations of these three regions: we considered the six alternative orders of the three segments (without changing their orientation) and, for each of these six orders we considered the impact of inverting each of the three segments in turn (Supplementary Table 1). The quality of the LDU map made assuming each order was evaluated through the error variance, which is expected to be ≥ 1 and total LDU map length, where the shorter map suggests a 'good' order. We also considered the impact on map quality for different assumed distances between the three segments.

3 Results

3.1 The LDU map of the whole region

The LD map of the region (Figure 1) spans 10.3 LDUs across 215.6 kb. This indicates an average extent of LD (kb/LDU = the 'swept radius') of 21 kb. There is substantial variability across the map contour showing regions of intense LD breakdown (recombination hotspots or ‘steps’ in the map) and regions with strong LD (blocks, plateaus on the map).Three regions were defined to include approximately the same number of SNPs in each region. These regions A–C have swept radii of, respectively, 22.7, 17.3 and 24.4 kb.

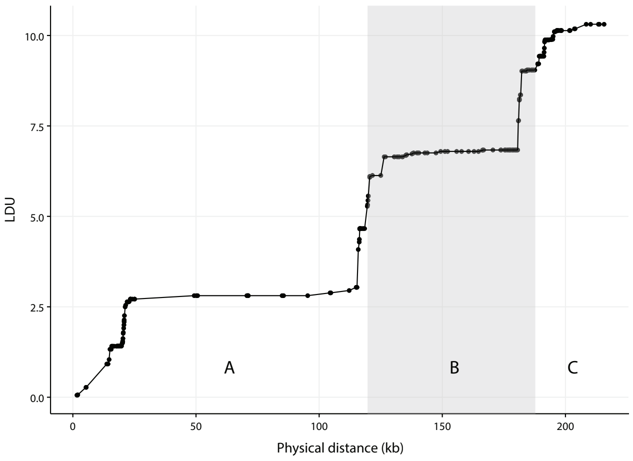

Fig. 1. LDU map of a 215.6 kb region of 6p21.3 (Jeffreys et al., 2001, Zhang et al, 2002). Points on the graph represent individual SNPs. The change in graph background colour indicates the transition between the three map segments (A, B and C) used in the evaluation. Map segments contain approximately equal numbers of SNPs.

3.2 Impact of permuted and inverted orders

We considered the impact on map length and on error variance of constructing LDU maps for the six different permuted orders and, for each of the six, we tested the effect of inverting segments A, B and C separately. We introduce the notation ‘a’ to represent an inversion of segment A (and ‘b’ and ‘c’ as inversions of B and C respectively). This scheme generates six different orders without inversions and 18 alternative orders which contain inverted segments (Supplementary Table 1, Figure 2). Orders which correspond to a complete reversal of the whole sequence give the same result so are excluded from the table (for example CBA is equivalent to abc). For this reason double and triple inversions are already captured in Supplementary Table 1. Small differences in the total number of SNP pairs informative for the maps (Supplementary Table 1) reflect the presence of a few SNPs with incomplete genotyping. The sliding window approach used by LDMAP captures all pairs informative for LD in a given SNP interval located 100 SNPs either side of that interval and so the SNP pairs selected depend partly on the assumed map order.

The distance between SNPs closest to the two introduced breakpoints was fixed at 114 base pairs as the average of the gap distances in the original data (87 and 140 base pairs respectively for the two breakpoints), thus the original total kb map length is preserved. Figure 2 plots error variance against total map length in LDU. The best predictor of the correct order is total LDU map length with the shortest map (10.3 LDUs) corresponding to the correct ABC order. Two alternative orders in which single segments are inverted show only small percentage increases in map length relative to the correct order. Hence order ABc has a map which is only ca. 2% longer than the correct order and AbC has a map which is only ca. 3% longer, suggesting the algorithm is less able to discriminate between maps with inversions in otherwise correctly ordered segments.

Alternative orders suggest map lengths in the range 10.5 LDUs (for order ABc) to 17.9 LDUs (for order aCB, a 73% increase in map length). Orders which show the greatest change in LDU length are ACB and the three versions of this order with inversions (mean LDU length 14.7) and BCA (mean LDU length 14.8).

A proportion of the variation in the error variance is related to LDU map length (correlation $R = -0.58$, $P = 0.0029$). Figure 2 shows that the LDMAP algorithm may minimise EV to values less than one by inflating the LDU length of the map. This suggests over-fitting of the data (Bevington, 1969) by inflating total LDU map length for poor orders. A combination of an EV close to, or marginally greater than 1, and minimal LDU map length is a strong indicator of ‘good’ contig order and orientation.

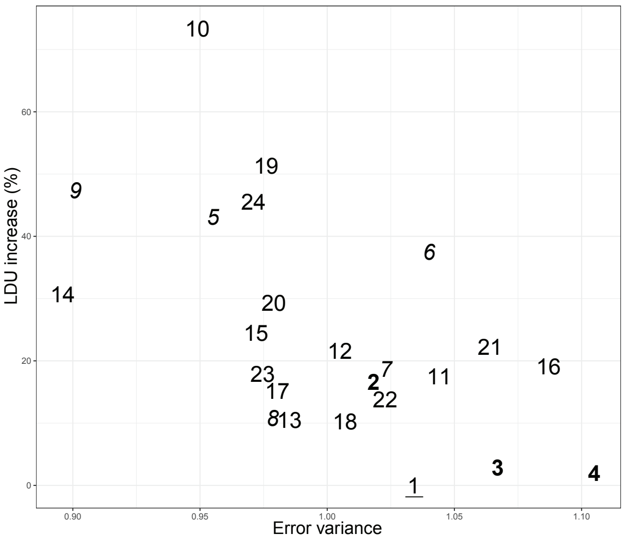

Fig. 2. Model error variance versus percentage increase in LDU map length for alternative orders and orientations. The correct order (1; underlined) produces the shortest LD map, with relatively minor increase in map length for single inversions in the correct order (2–4; bold). However, reordered (5–9; italics), as well as inverted and re-ordered segments (10-24) are associated with substantial increases in map length. Error variances below 1 suggest over-fitting achieved by inflating map length. Specific orders corresponding to number labels are found in Supplementary Table 1.

3.3 Effect of variable inter-segment distances

Because in sequence assembly the physical distance between individual contigs is likely to be unknown, then incorrect assumed gap distances may impact map assembly. We considered the impact of different assumed gap sizes between the three map segments (Supplementary Table 2, Figure 3). Introducing the ‘correct’ map distance (114 base pairs) as the gap between segments 1-2 and segments 2-3 preserves the total map length of ca. 215.6 Kb. Inserting larger gaps, hence increasing the total map length in the range 215.65-235.42 kb, generates LD maps with only fractionally increased length (from 10.3 to 10.4 LDU). This suggests that the algorithm is relatively robust to miss-specification of the inter-contig distance by at least 10 kb per gap. Since the swept radius for this map is 21 kb we might expect that an inter-gap distance > 21 kb for the two breakpoints (generating a total map length of ca. 258 kb or more) would effectively break the LD connection between the segments. This is suggested by Figure 3 which shows steep increase in error variance for maps > 250 kb in total length and declining LDU length suggesting the LDMAP algorithm sharply reduces LDU map length (but with increased error variance) where assumed physical distances between linked contigs are inflated beyond the swept radius.

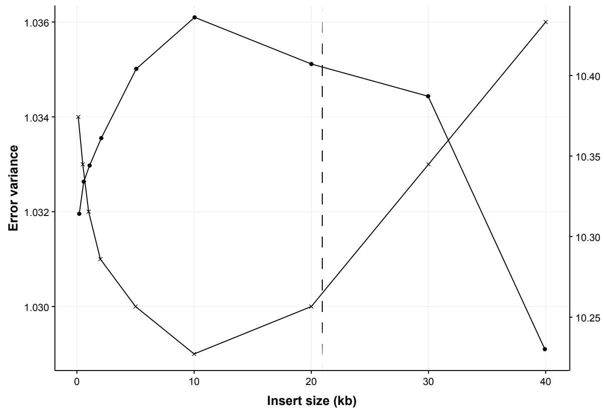

Fig. 3. Effect on error variance (crosses) and map length (dots) of varying the distance between segments for the correct (ABC) order. The algorithm is relatively robust to miss-specification of inter-segment gap size unless the inserted gap exceeds the swept radius beyond which the error variance increases substantially. The swept radius of 21 Kb, representing the distance at which LD declines to 'background' levels, is shown on the figure (dashed vertical line).

4 Discussion

Accurate sequence contig ordering, orientation and positioning remains a challenge for reference genome assembly despite the huge advances provided by NGS (Richards, 2018). The use of longer sequence reads and genetic linkage map scaffolds are invaluable but both are costly and may be difficult to obtain. Linkage disequilibrium maps offer an alternative approach to guide assembly and are relatively low cost, requiring only population SNP data. The LDMAP algorithm is shown to be useful for discriminating between alternative contig orders and orientations with optimal orders indicated by short LDU maps and residual error variance, representing the fit of the LDU map, close to one. The data suggest that the algorithm is less able to discriminate between assemblies with the correct order where single segments have been inverted. It is likely that factors such as sequence contig size and overall extent of LD will have an impact on the utility of this approach for specific studies. Three orders (ABC, AbC and ABc), in the example used, show the shortest overall LDU map lengths, but lengths differ by only a few percent suggesting weak discrimination between orders where a single segment inversion is the only change.

Evaluation of the significance of differences between closely similar maps may be achieved using a "delete-d" jackknife as described by Service et al (2006). They describe LDU maps for 12 populations with map lengths in the range 368–865 LDUs. The delete-d jackknife estimates the standard error of the mean (s.e.m) of the LDU length of each map achieved in this case by randomly selecting 100 samples of 180 individuals from the total sample size of 200 per population (delete-d of 20 samples in each run) and re-computing the LD map in each jackknife sample. The authors found s.e.m estimates in the range ca. 15–34 LDUs, an average of 4% of total map length. Such an approach might be usefully applied to approximate the significance of differences between closely similar maps. The evaluations indicate that the LDMAP model is susceptible to over-fitting (indicated by error variance less than one). However, these instances are associated with inflated LDU map distances as a signature of a 'poor' assembly. Atypically small error variances (<1) should therefore be interpreted with care. Misspecification of the inter-contig distance does not appear problematic unless, for linked segments, the assumed distance between contigs exceeds the swept radius.

Alternative approaches which exploit LD information to guide genome assembly include "Locus ordering by linkage disequilibrium" (LODE) (Khatkar et al, 2010, Jones et al, 2017). LODE uses LD data to place 'orphan' SNPs within genetic linkage map scaffolds, increasing marker coverage which facilitates positioning of short sequence contigs. The algorithm positions SNPs by first obtaining a candidate chromosome assignment for a SNP based on the pattern of association with SNPs already in the map. Once assigned to a chromosome the location with strongest association is selected as a candidate location for the orphan SNP. This approach is also successful for validating genome assembly by testing the strength of evidence for the positions of SNPs already located within the assembly (Khatkar et al, 2010). Applying this method, Jones et al (2017) constructed an integrated linkage/LD map for the Pacific white shrimp (*Litopenaeus vannamei*) for which a minimum of 75 individuals provided accurate LD estimates to place several hundred SNPs within the linkage scaffold.

Utsunomiya et al (2016) consider the use of LD to identify within-contig assembly errors. Their algorithm examines patterns of atypical LD decay. This approach identified 2906 poorly located SNPs in the bovine reference genome. LD information is therefore potentially valuable in three areas of genome assembly: 1. Increasing the resolution of linkage map scaffolds by locating orphan markers. 2. Identifying within-contig assembly errors by identifying wrongly located SNPs and 3. Ordering, orienting and positioning contigs assembled from de novo short read sequences.

The recently announced Earth BioGenome Project (EBP, Lewin et al, 2018) proposes to sequence, catalogue and characterise the genomes of all eukaryotes (1.5 million species) over 10 years. Quality reference sequence assemblies for a least one representative species from ca. 9000 eukaryotic families are envisaged. The authors recognise that the creation of traditional genetic (linkage) and physical maps to guide chromosome-scale assemblies is cost-prohibitive for a project of this scale. The EBP will be relying on short-read sequencing and longer–range scaffolding to approach chromosome-level assembly. More complete chromosome-level assembly might consider employ optical mapping (Lam et al, 2012) and Hi-C (Burton et al, 2013) to assemble contigs but these are currently both technically challenging and costly. The use of LD structure as described here might facilitate chromosome-level assembly where it is otherwise difficult or too expensive. The implementation of a range of techniques and strategies will be required for this ambitious and exciting project to succeed.

5 Conclusion

The interpretation of LD patterns does not provide a complete solution to the problems of sequence assembly but the use of LD information to guide sequence assembly is a strategy which is currently not widely exploited. The use of this method depends on the availability of a representative population sample and covering SNP data. In previous studies samples of >20 unrelated individuals have been effective (e.g. Tapper et al, 2003). However, much of the analysis thus far has considered the characterisation of the fine-scale LD structure of chromosomes and so larger sample sizes and high SNP coverage have been emphasized. For the purpose of ordering and orienting contigs a smaller number of unrelated individuals and low-density SNP coverage may be sufficient to facilitate draft assembly. Where feasible, LD maps might be used most effectively in combination with long read approaches (which have a resolution ca. 10–20 kb) and linkage maps (which, optimally, may have ca. 1,000 kb resolution) because LD extends in the 10s–100s kb range and so has an intermediate resolution.

References

Baker, M. (2012). De novo genome assembly: what every biologist should know. *Nature Methods*, **9**, 333.

Bevington, P. (1969). *Data reduction and error analysis for the physical sciences*. McGraw-Hill, New York.

Burton, J. N. *et al.* (2013). Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nature Biotechnology*, **31**(12), 1119–1125.

Collins, A. and Morton, N. E. (1998). Mapping a disease locus by allelic association. *Proceedings of the National Academy of Sciences of the United States of America*, **95**(4), 1741–5.

Collins, A. *et al.* (1999). Genetic epidemiology of single-nucleotide polymorphisms. *Proceedings of the National Academy of Sciences of the United States of America*, **96**(26), 15173–7.

Denton, J. F. *et al.* (2014). Extensive Error in the Number of Genes Inferred from Draft Genome Assemblies. *PLoS Computational Biology*, **10**(12), e1003998.

Ennis, S. *et al.* (2001). Allelic association discriminates draft orders. *Annals of human genetics*, **65**(Pt 5), 503–4.

Fierst, J. L. (2015). Using linkage maps to correct and scaffold de novo genome assemblies: methods, challenges, and computational tools. *Frontiers in genetics*, **6**, 220.

Hahn, M. W. *et al.* (2014). Sequencing, assembling, and correcting draft genomes using recombinant populations. *G3 (Bethesda, Md.)*, **4**(4), 669–79.

Jeffreys, A. J. *et al.* (2001). Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nature Genetics*, **29**(2), 217–222.

Jones, D. B. *et al.* (2017). A comparative integrated gene-based linkage and locus ordering by linkage disequilibrium map for the Pacific white shrimp, Litopenaeus vannamei. *Scientific reports*, **7**(1), 10360.

Khatkar, M. S. *et al.* (2010). Assignment of chromosomal locations for unassigned SNPs/scaffolds based on pair-wise linkage disequilibrium estimates. *BMC Bioinformatics*, **11**(1), 171.

Koren, S. *et al.* (2012). Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nature Biotechnology*, **30**(7), 693–700.

Kuo, T.-Y. *et al.* (2007). LDMAP. In *Linkage Disequilibrium and Association Mapping: Analysis and Applications*, pages 47–57. Humana Press.

Lam, E. T. *et al.* (2012). Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly. *Nature Biotechnology*, **30**(8), 771–776.

Lau, W. *et al.* (2007). Exploiting large scale computing to construct high resolution linkage disequilibrium maps of the human genome. *Bioinformatics*, **23**(4), 517–519.

Lewin, H. A. *et al.* (2009). Every genome sequence needs a good map. *Genome research*, **19**(11), 1925–8.

Lewin, H. A. *et al.* (2018). Earth BioGenome Project: Sequencing life for the future of life. *Proceedings of the National Academy of Sciences*, **115**(17), 4325–4333.

Purcell, S. *et al.* (2007). PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics*, **81**(3), 559–575.

Richards, S. (2018). Full disclosure: Genome assembly is still hard. *PLOS Biology*, **16**(4), e2005894.

Service, S. *et al.* (2006). Magnitude and distribution of linkage disequilibrium in population isolates and implications for genome-wide association studies. *Nature Genetics*, **38**(5), 556–560.

Tapper, W. *et al.* (2005). A map of the human genome in linkage disequilibrium units. *Proceedings of the National Academy of Sciences*, **102**(33), 11835–11839.

Tapper, W. J. *et al.* (2003). A Metric Linkage Disequilibrium Map of a Human Chromosome. *Annals of Human Genetics*, **67**(6), 487–494.

Utsunomiya, A. T. H. *et al.* (2016). Revealing misassembled segments in the bovine reference genome by high resolution linkage disequilibrium scan. *BMC Genomics*, **17**(1), 705.

Zhang, W. *et al.* (2002). Properties of linkage disequilibrium (LD) maps. *Proceedings of the National Academy of Sciences of the United States of America*, **99**(26), 17004–7.

Zimin, A. V. *et al.* (2017). An improved assembly of the loblolly pine mega-genome using long-read single-molecule sequencing. *GigaScience*, **6**(1), 1–4.