Genome Biology

COMMENT

So, you want to sequence a genome…

Derek L Stemple*

Anyone who has attempted to identify the responsible gene or mutation underlying a disease or mutant phenotype will know the importance of an accurate reference genome assembly. For complex vertebrate genomes, however, generating such an assembly is not trivial, even with new sequencing technologies. In 2007, a mid-point in the zebrafish genome sequencing project, I was asked to lead the project to completion. At that point we were faced with a highly fragmented physical assembly and lacked genetic maps of sufficient density and resolution to produce an accurate assembly.

A high-quality reference genome assembly is generally made up of a large set of minimally overlapping large-insert genomic clones, each of which has been sequenced to completion, with a minimal number of gaps and with no artificially duplicated regions. These high-quality reference genome assemblies, such as the current human reference genome (http://www.genomereference.org), are essential for modern molecular genetic studies. For many species, however, only lower quality whole-genome shotgun assemblies are available. When one considers, for example, only the protein-coding genes, this quality of genome sequence is often not sufficient to determine the complete gene count or comprehensive set of accurate gene models. It is important, for the best application of the genomic information, that the reference genome be complete and accurately assembled. While high-throughput short-read sequencing using the current generation of machines will yield high quality for bacterial, and other small, genomes, it is not possible to completely and accurately assemble the large, complex genomes of vertebrates without other long-range contiguity information. Experience with the zebrafish genome [1] may provide some useful guidance for anyone embarking on a genome-sequencing project for new species with a complex genome.

The human, mouse and zebrafish reference genomes were assembled using old-school approaches, where the long-range contiguity was derived from genetic or genomic mapping and not derived directly from sequencing reads or read-pairs. The maps used were accurate physical maps of overlapping genomic DNA fragments or high-resolution genetic maps with a high density of short sequence markers, but such maps are expensive and time-consuming to generate. There are some good possibilities for cheaper, easier way to generate accurate maps, but there are several issues to consider.

Use a single haplotype (if you can)

Any reference for a complex genome is necessarily going to be a compromise. The reference will, in the first instance, only represent one exemplar haploid copy of an arbitrarily selected individual or the fragmented agglomeration of several different individuals. For some species, such as mice, where recombinant inbred lines exist, the task was simple because both copies of the haploid genome are essentially identical. For most species, however, the degree of polymorphism between the two copies of the diploid genome can be high and will confound the assembly. Single nucleotide polymorphism (SNP) rates of 1 change every 200 bases on average are not uncommon, as we found with zebrafish, but will seriously limit the efficacy of most assembly algorithms. Although it is possible to obtain genomic DNA of a single haplotype for some species, it is rare. If single haplotype genomic DNA is not available, it is important to minimize the number of individuals used as a source of reference genome DNA for generation of the libraries. Ideally, a single individual would be used. For the zebrafish reference genome, most of the genomic libraries used were generated from a collection of different, but related, individuals. Consequently, the physical maps were highly fragmented and contained a large number of haplotypic duplications. Ultimately, we were forced to turn to an alternative strategy to produce an accurate assembly.

High-resolution genetic maps can provide clarity

From the beginning of the zebrafish genome-sequencing project in 2002, genome sequence was made publicly available through the Ensembl genome browser (http://www.ensembl.org/Danio_rerio). With each new assembly as the project progressed, inaccuracies were rapidly flagged up by zebrafish researchers from around the world. One major aspect of zebrafish genetic research is

*Correspondence: ds4@sanger.ac.uk
Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, UK

the use of very high-resolution genetic mapping data (several hundred to thousands of meioses) to pinpoint the location of mutated genes in the genome in positional cloning studies. The local accuracy of these mapping projects led us to realize that a solution to the genome assembly problem might lie in a genome-wide high-resolution meiotic map. We constructed such a map, now known as SATmap [1], which comprises over 140,000 SNP markers scored over more than 900 meioses and now underpins the reference genome assembly. Indeed, SATmap provides genetic resolution, allowing the accurate ordering and orientation of large insert clones over more than 90% of the genome.

How can this be put together for a *de novo* genome-sequencing project?

The number of meiotic events measured will set the resolution of a meiotic map derived from a fixed pedigree, such as that used to generate SATmap. Alternatively, a similarly useful genetic map can be generated from an arbitrary population of individuals. In the population-based map, however, one needs to set the mapping resolution through a mixture of the number of different individuals analyzed and the average size of linkage-disequilibrium (LD) blocks across the measured population. The human HapMap is a good example of such a population-based LD map (http://hapmap.ncbi.nlm.nih.gov/) and such maps could be generated for essentially any population. For a genome-sequencing project of new species, it is easy and relatively cheap to obtain a deep-coverage short-read assembly using new sequencing technologies. If one obtains an LD map from a select population of that species with an LD block size that is on the same order as the average assembled fragment size, then it may be possible to produce a very accurate assembly very quickly and cheaply by comparison with old-school methods.

Making LD maps

For LD map construction, it is important to accurately and consistently measure the genotype of each individual from many thousands of locations across the genome. The widespread use of hybridization-based enrichment methods, like those used for whole-exome enrichment to identify causal lesions in rare human genetic diseases, provides a good methodology for rapid genome-wide genotyping across many individuals. For this to work, one would start with a good collection of assembled fragments from the whole-genome short-read sequencing data and, from that, identify unique 'bait' sequences spread evenly across the genome. One could, for example, design baits such that each 100 kb chunk of short read assembly is covered. Then one would generate barcoded libraries from each individual in the population, carry out the enrichment hybridization and sequence the enriched genomic fragments using short-read high-throughput sequencing technology.

By analyzing the distribution of common SNPs among the genomic fragments across the individuals in the population, using mapping and variant calling programs such as SAMtools (http://samtools.sourceforge.net/), the raw genotyping data could be derived to generate a LD map. It would probably be necessary to over-sample the population in pilot studies to select a group of individuals that would have the appropriate SNP rates and LD block sizes. A large genetically diverse population should have a sufficiently small LD block size, and sub-selection from the population could be employed to match the appropriate LD block size with the average assembled fragment size. In general, the generation of an LD map would be appropriate for situations where production of crosses and pedigrees for a meiotic map are difficult, because the effort required to identify the appropriate population of individuals to generate the map may be substantial in these scenarios. Ultimately, the optimal approach may be obviated by new DNA sequencing technologies with the potential to produce very long reads.

New sequencing technologies to the rescue?

An ideal sequencing strategy would be to start sequencing at one end of each chromosome and determine the entire sequence in one long contiguous read. There are only a couple of sequencing platforms at present that do not require the amplification of DNA, and none of the existing technologies is capable of such fantastically long reads. There is some promise. Considering just the current high-throughput sequencers, there has been a gradual but fairly steady increase in raw read length, and new library generation methods are allowing contiguity assignments over greater and greater distances. Ultimately the single-molecule and pore-based methods may provide the very long read lengths necessary to accurately assemble complex genomes. One can only hope.

Published: 30 July 2013

Reference

1. Howe K, Clark MD, Torroja CF, Torrance J, Berthelot C, Muffato M, Collins JE, Humphray S, McLaren K, Matthews L, McLaren S, Sealy I, Caccamo M, Churcher C, Scott C, Barrett JC, Koch R, Rauch G-J, White S, Chow W, Kilian B, Quintais LT, Guerra-Assunção JA, Zhou Y, Gu Y, Yen J, Vogel J-H, Eyre T, Redmond S, Banerjee R, *et al.*: **The zebrafish reference genome sequence and its relationship to the human genome.** *Nature* 2013, **496**:498-503.

doi:10.1186/gb-2013-14-7-128
Cite this article as: Stemple DL: So, you want to sequence a genome…. *Genome Biology* 2013, **14**:128.