Experimental Cell Research 358 (2017) 433–438

Contents lists available at ScienceDirect

Experimental Cell Research

journal homepage: www.elsevier.com/locate/yexcr

The value of new genome references

Kim C. Worley[a,b,*], Stephen Richards[a,b], Jeffrey Rogers[a,b]

[a] Human Genome Sequencing Center, Baylor College of Medicine, One Baylor Plaza, MS BCM226, Houston, TX 77030, USA
[b] Department of Molecular and Human Genetics, Baylor College of Medicine, One Baylor Plaza, Houston, TX 77030, USA

ARTICLE INFO

Keywords:
Genome assembly
Sequencing
Gene family analysis
Segmental duplications

ABSTRACT

Genomic information has become a ubiquitous and almost essential aspect of biological research. Over the last 10–15 years, the cost of generating sequence data from DNA or RNA samples has dramatically declined and our ability to interpret those data increased just as remarkably. Although it is still possible for biologists to conduct interesting and valuable research on species for which genomic data are not available, the impact of having access to a high quality whole genome reference assembly for a given species is nothing short of transformational. Research on a species for which we have no DNA or RNA sequence data is restricted in fundamental ways. In contrast, even access to an initial draft quality genome (see below for definitions) opens a wide range of opportunities that are simply not available without that reference genome assembly. Although a complete discussion of the impact of genome sequencing and assembly is beyond the scope of this short paper, the goal of this review is to summarize the most common and highest impact contributions that whole genome sequencing and assembly has had on comparative and evolutionary biology.

1. Introduction

One basic distinction is critical at the outset. In many circumstances, the phrase “sequencing a new genome” refers to the analysis of an individual, including the subsequent comparison of that one individual's genome to a reference genome assembly meant to represent that species. When the genome of a human patient with an undiagnosed clinical disorder is sent for “sequencing,” the typical procedure is to generate sufficient raw read data to compare some (the protein coding exome) or (nearly) all of the patient's genome to a standard human reference and look for differences that may be pathogenic and hence clinically relevant. Although it is likely that in the future it will be possible, and maybe even routine, to produce *de novo* whole genome assemblies for individual patients at acceptable cost this is not now practical. However, the re-sequencing of individuals (i.e. the analysis of a given individual by comparing that individual's DNA sequence to a curated and annotated reference genome) is not the focus of this review. Rather, we discuss the methods for and impact of producing *de novo* whole genome assemblies for species for which such information is not currently available. Thus, the question at issue is not “what do we learn about a specific individual from sequencing the genome of that specific individual” but rather “what do we learn about species X from sequencing an individual or pool of individuals that represent species X.” Producing a new reference genome facilitates analyses of other individuals from the same or closely related species, and this wider meaning and impact is our topic here.

2. How does one produce a new reference genome assembly?

The technology for sequencing DNA and assembling the raw sequence read data into a continuous representation of chromosomes continues to improve at a rapid pace. For this reason, any review of genomic methods and anticipated results will have a short shelf-life. Nevertheless, some general principles are likely to remain relevant for the foreseeable future. The current methods for producing the raw sequence data from which *de novo* whole genome assemblies can be constructed fall into two categories. The dominant short read technology comes from Illumina, Inc. and uses well-established chemistry to identify the sequence of nucleotides present in a given short segment of DNA. Current Illumina sequencing platforms produce basepair sequences of lengths up to 250 bp in a given DNA segment, and generally are used to read those sequences from both ends of a DNA fragment. This “next generation” technology was a dramatic improvement over older Sanger sequencing methods that produced longer reads but at much higher cost. The Illumina platforms have become the workhorses of genome sequencing but “third generation” technologies from Pacific Biosciences, Oxford Nanopore and other companies are gaining users

Abbreviations: NGS, next generation sequencing; LGT, lateral gene transfer

[*] Corresponding author at: Human Genome Sequencing Center, Baylor College of Medicine, One Baylor Plaza, MS BCM226, Houston, TX 77030, USA.
E-mail address: kworley@bcm.edu (K.C. Worley).

http://dx.doi.org/10.1016/j.yexcr.2016.12.014
Received 22 November 2016; Accepted 22 December 2016
Available online 23 December 2016

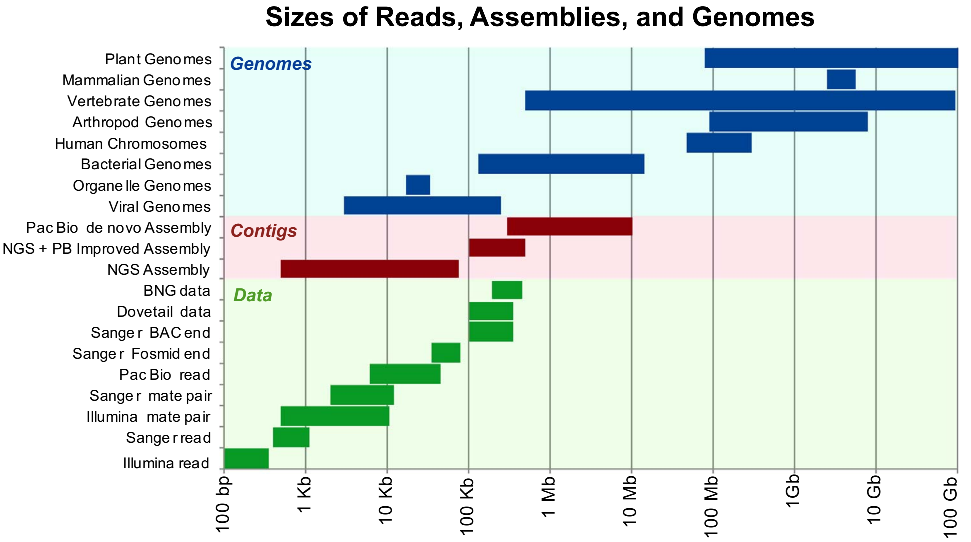

Fig. 1. This figure shows the challenges of creating complete genome representations. The relative sizes of different genomes, genome assembly contigs, and sequencing technologies are shown with the logarithmic scale across the bottom. At the top in blue are the sizes of genomes for different clades of organisms. Vertebrate genome sizes vary over 5 orders of magnitude, while mammalian genome sizes are more similarly sized (~3 Gb). In the middle, in red, are the size ranges for the contigs (measure by contig N50) for current sequencing technologies (Illumina next generation or NGS assemblies, NGS assemblies with PacBio improvement, and PacBio de novo assemblies). The lengths of sequence reads, mate-pair distances and mapping fragments of different technologies are shown in the green bars at the bottom.

and making impact. These “third generation” methods generate longer reads, up to tens of thousands of basepairs per read, but with higher error rates and other disadvantages.

The field of *de novo* whole genome assembly is currently grappling with the challenges of determining the optimal use of these various and constantly changing data types. As the methods for generating the raw sequence read data have evolved, so have the analytical strategies and software tools designed to assemble the large number of short or long reads into continuous sequences millions of bases (megabases) in length. The ultimate goal of genome assembly is production of error-free continuous sequences that span entire chromosomes. In contrast, currently attainable genome assemblies have tens to hundreds of thousands of gaps, though much of the genome is represented with good sequence quality in the assembled regions. Progress has been rapid and it is not outrageous to envision that essentially complete, highly accurate whole genome assemblies will be practical in the near future.

The data, problem and current results are illustrated in Fig. 1 where the size of elements are represented on the logarithmic scale range from 100 basepairs to 100 billion basepairs (100 Gb). At the top in Fig. 1 are the sizes of the targets of genome assembly (genomes and chromosomes). The sequence read lengths and mate-pair distances for the different technologies are on the bottom.

In the middle of Fig. 1, the lengths of the pieces in a *de novo* assembly are indicated using the contig N50 values. Contig N50 is a statistic commonly used to compare different genome assembly results in terms of their contiguity (the length of continuous DNA sequences without gaps, called contigs). Contig N50 is calculated by sorting all the contigs within an assembly from largest to smallest, then determining the size of the contig at which half of the total genome is in pieces bigger than that N50 value. Larger contig N50 values correlate with improved recovery of genomic features (see below). Note that the genome sizes and chromosome sizes are 10–1000 times longer than the contig N50 values of current *de novo* assemblies. Improving the contiguity by adding some long read data to an NGS assembly or using a *de novo* long read method is beneficial. Often it is genomic features such as repeats that limit the contiguity attained in a genome assembly and these methods resolve many of these repeats and recover missing data. As an example, the *de novo* PacBio gorilla genome recovered 87% of the exons missing in earlier assemblies [1]. There remains room for improvement, because even easier to assemble haploid samples need directed finishing efforts to address the last difficult regions [2,3], and the extensively studied “finished” human reference genome is still being improved [4]. Despite this potential for further progress, changing reference genome versions is a painful process of transferring genome analysis coordinates (the basepair locations) from the old version to the new, and there will always be sequences that are altered during this transfer that someone prefers as the old version. So even as genomic sequencing costs drop and *de novo* genome assemblies becomes more achievable, researchers should plan to use the current genome iteration for several years.

3. Genome analysis wish lists

As a foundational biological tool there are many analyses possible with a new genome (Table 1). We present here a short list of the most common analyses with example results leading to biological insight. First and foremost on this wish list is identification of the protein-coding genes in that genome and the comparative analysis of gene families. Gene family analyses begin with identifying the collection of protein coding genes in the organism, which is compared to the gene complement from other species already sequenced. Differences in the number and types of genes present as well as differences in the sequences of orthologous genes are evaluated.

The genome assemblies first produced by large-scale sequencing efforts allowed investigation of lineage-specific gene duplication and gene loss [5]. Expanding gene families are thought to be a birthplace for molecular innovation [6] and have repeatedly provided biological insight into for example photoreception [7] and rumination [8]. The

Table 1
Genome analysis wish list.

Protein coding genes	Structure and complete sequence
Gene families	Expansions and contractions
Rapidly evolving genes	Positively selected
Non-coding RNA genes	Structure and complete sequence
Ancestral gene content	
Repetitive elements	Transposable elements
Segmental duplications	
Population history	Phylogenetics, population size
Genomic history	Lateral gene transfer, admixture