Scan to visit

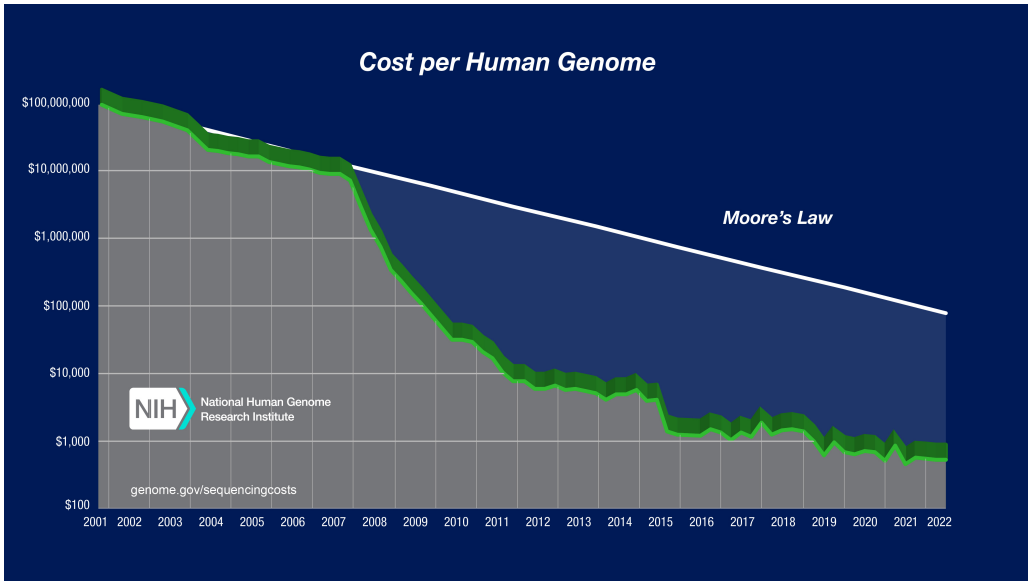
The Cost of Sequencing a Human Genome

Advances in the field of genomics over the past quarter-century have led to substantial reductions in the cost of genome sequencing. The underlying costs associated with different methods and strategies for sequencing genomes are of great interest because they influence the scope and scale of almost all genomics research projects.

Overview

Significant scrutiny and attention have been given to genome-sequencing costs and how they are calculated since the beginning of the field of genomics in the late 1980s. For example, NHGRI has carefully **tracked costs per genome** at its funded 'genome sequencing centers' for many years (see Figure 1). With the growing scale of human genetics studies and the increasing number of clinical applications for genome sequencing, even greater attention is being paid to understanding the underlying costs of generating a human genome sequence.

Accurately determining the cost for sequencing a given genome (e.g., a human genome) is not simple. There are many parameters to define and nuances to consider. In fact, it is difficult to cite precise genome-sequencing cost figures that mean the same thing to all people because, in reality, different researchers, research institutions, and companies typically track and account for such costs in different fashions.



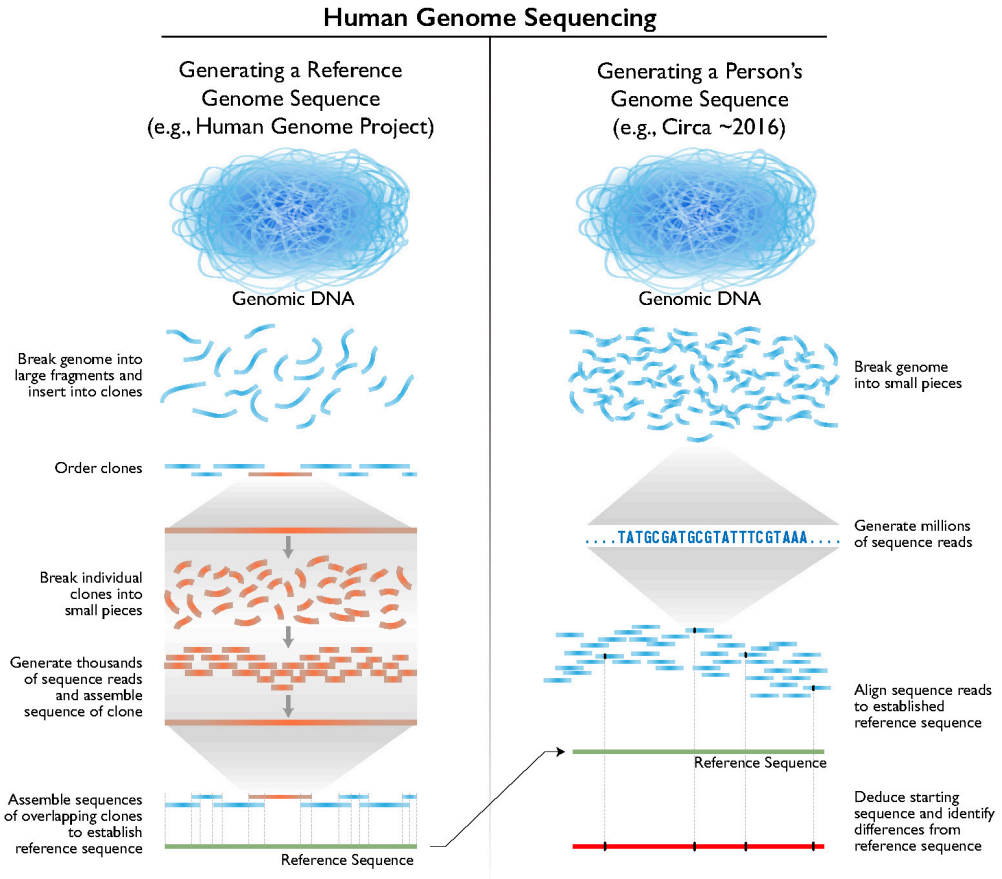
Cost per genome data - 2022

A Primer on Genome Sequencing

A genome consists of all of the DNA contained in a cell's nucleus. DNA is composed of four chemical building blocks or "bases" (for simplicity, abbreviated G, A, T, and C), with the biological information encoded within DNA determined by the order of those bases. Diploid organisms, like humans and all other mammals, contain duplicate copies of almost all of their DNA (i.e., pairs of chromosomes; with one chromosome of each pair inherited from each parent). The size of an organism's genome is generally considered to be the total number of bases in one representative copy of its nuclear DNA. In the case of diploid organisms (like humans), that corresponds to the sum of the sizes of one copy of each chromosome pair.

Organisms generally differ in their genome sizes. For example, the genome of E. coli (a bacterium that lives in your gut) is ~5 million bases (also called megabases), that of a fruit fly is ~123 million bases, and that of a human is ~3,000 million bases (or ~3 billion bases). There are also some surprising extremes, such as with the loblolly pine tree - its genome is ~23 billion bases in size, over seven times larger than ours. Obviously, the cost to sequence a genome depends on its size. The discussion below is focused on the human genome; keep in mind that a single 'representative' copy of the human genome is ~3 billion bases in size, whereas a given person's actual (diploid) genome is ~6 billion bases in size.

Genomes are large and, at least with today's methods, their bases cannot be 'read out' in order (i.e., sequenced) end-to-end in a single step. Rather, to sequence a genome, its DNA must first be broken down into smaller pieces, with each resulting piece then subjected to chemical reactions that allow the identity and order of its bases to be deduced. The established base order derived from each piece of DNA is often called a 'sequence read,' and the collection of the resulting set of sequence reads (often numbering in the billions) is then computationally assembled back together to deduce the sequence of the starting genome. Sequencing human genomes are nowadays aided by the availability of available 'reference' sequences of the human genome, which play an important role in the computational assembly process. Historically, the process of breaking down genomes, sequencing the individual pieces of DNA, and then reassembling the individual sequence reads to generate a sequence of the starting genome was called 'shotgun sequencing' (although this terminology is used less frequently today). When an entire genome is being sequenced, the process is called 'whole-genome sequencing.' See Figure 2 for a comparison of human genome sequencing methods during the time of the Human Genome Project and circa ~ 2016.



An alternative to whole-genome sequencing is the targeted sequencing of part of a genome. Most often, this involves just sequencing the protein-coding regions of a genome, which reside within DNA segments called 'exons' and reflect the currently 'best understood' part of most genomes. For example, all of the exons in the human genome (the human 'exome') correspond to ~1.5% of the total human genome. Methods are now readily available to experimentally 'capture' (or isolate) just the exons, which can then be sequenced to generate a 'whole-exome sequence' of a genome. Whole-exome sequencing does require extra laboratory manipulations, so a whole-exome sequence does not cost ~1.5% of a whole-genome sequence. But since much less DNA is sequenced, whole-exome sequencing is (at least currently) cheaper than whole-genome sequencing.

Another important driver of the costs associated with generating genome sequences relates to data quality. That quality is heavily dependent upon the average number of times each base in the genome is actually 'read' during the sequencing process. During the Human Genome Project (HGP), the typical levels of quality considered were: (1) 'draft sequence' (covering ~90% of the genome at ~99.9% accuracy); and (2) 'finished sequence' (covering >95% of the genome at ~99.99% accuracy). Producing truly high-quality 'finished' sequence by this definition is very expensive; of note, the process of 'sequence finishing' is very labor-intensive and is thus associated with high costs. In fact, most human genome sequences produced today are 'draft sequences' (sometimes above and sometimes below the accuracy defined above).

There are thus a number of factors to consider when calculating the costs associated with genome sequencing. There are multiple different types and quality levels of genome sequences, and there can be many steps and activities involved in the process itself. Understanding the true cost of a genome sequence therefore requires knowledge about what was and was not included in calculating that cost (e.g., sequence data generation, sequence finishing, upfront activities such as mapping, equipment amortization, overhead, utilities, salaries, data analyses, etc.). In reality, there are often differences in what gets included when estimating genome-sequencing costs in different situations.

Below is summary information about: (1) the estimated cost of sequencing the first human genome as part of the HGP; (2) the estimated cost of sequencing a human genome in 2006 (i.e., roughly a decade ago); and (3) the estimated cost of sequencing a human genome in 2016 (i.e., the present time).

Timeline of Costs

How much did it cost to generate the first human genome sequence as part of the Human Genome Project?

The HGP generated a 'reference' sequence of the human genome - specifically, it sequenced one representative version of all parts of each human chromosome (totaling ~3 billion bases). In the end, the quality of the 'finished' sequence was very high, with an estimated error rate of <1 in 100,000 bases; note this is much higher than a typical human genome sequence produced today. The generated sequence did not come from one person's genome, and, being a 'reference' sequence of ~3 billion bases, really reflects half of what is generated when an individual person's ~6-billion-base genome is sequenced (see below).

The HGP involved first mapping and then sequencing the human genome. The former was required at the time because there was otherwise no 'framework' for organizing the actual sequencing or the resulting sequence data. The maps of the human genome served as 'scaffolds' on which to connect individual segments of assembled DNA sequence. These genome-mapping efforts were quite expensive, but were essential at the time for generating an accurate genome sequence. It is difficult to estimate the costs associated with the 'human genome mapping phase' of the HGP, but it was certainly in the many tens of millions of dollars (and probably hundreds of millions of dollars).

Once significant human genome sequencing began for the HGP, a 'draft' human genome sequence (as described above) was produced over a 15-month period (from April 1999 to June 2000). The estimated cost for generating that initial 'draft' human genome sequence is ~**$300 million** worldwide, of which NIH provided roughly 50-60%.

The HGP then proceeded to refine the 'draft' and produce a 'finished' human genome sequence (as described above), which was achieved by 2003. The estimated cost for advancing the 'draft' human genome sequence to the 'finished' sequence is ~**$150 million** worldwide. Of note, generating the final human genome sequence by the HGP also relied on the sequences of small targeted regions of the human genome that were generated before the HGP's main production-sequencing phase; it is impossible to estimate the costs associated with these various other genome-sequencing efforts, but they likely total in the tens of millions of dollars.

The above explanation illustrates the difficulty in coming up with a single, accurate number for the cost of generating that first human genome sequence as part of the HGP. Such a calculation requires a clear delineation about what does and does not get 'counted' in the estimate; further, most of the cost estimates for individual components can only be given as ranges. At the lower bound, it would seem that this cost figure is at least $500 million; at the upper bound, this cost figure could be as high as $1 billion. The truth is likely somewhere in between.

The above estimated cost for generating the first human genome sequence by the HGP should not be confused with the total cost of the HGP. The originally projected cost for the U.S.'s contribution to the HGP was $3 billion; in actuality, the Project ended up taking less time (~13 years rather than ~15 years) and requiring less funding - ~**$2.7 billion**. But the latter number represents the total U.S. funding for a wide range of scientific activities under the HGP's umbrella beyond human genome sequencing, including technology development, physical and genetic mapping, model organism genome mapping and sequencing, bioethics research, and program management. Further, this amount does not reflect the additional funds for an overlapping set of activities pursued by other countries that participated in the HGP.

As the HGP was nearing completion, genome-sequencing pipelines had stabilized to the point that NHGRI was able to collect fairly reliable cost information from the major sequencing centers funded by the Institute. Based on these data, NHGRI estimated that the hypothetical 2003 cost to generate a 'second' reference human genome sequence using the then-available approaches and technologies was in the neighborhood of $50 million.

How much did it cost to sequence a human genome in 2006 (i.e., roughly a decade ago)?

Since the completion of the HGP and the generation of the first 'reference' human genome sequence, efforts have increasingly shifted to the generation of human genome sequences from individual people. Sequencing an individual's 'personal' genome actually involves establishing the identity and order of ~6 billion bases of DNA (rather than a ~3-billion-base 'reference' sequence; see above). Thus, the generation of a person's genome sequence is a notably different endeavor than what the HGP did.

Within a few years following the end of the HGP (e.g., in 2006), the landscape of genome sequencing was beginning to change. While revolutionary new DNA sequencing technologies, such as those in use today, were not quite implemented at that time, genomics groups continued to refine the basic methodologies used during the HGP and continued lowering the costs for genome sequencing. Considerable efforts were being made to the sequencing of nonhuman genomes (much more so than human genomes), but the cost-accounting data collected at that time can be used to estimate the approximate cost that would have been associated with human genome sequencing at that time.

Based on data collected by NHGRI from the Institute's funded genome-sequencing groups, the cost to generate a high-quality 'draft' human genome sequence had dropped to ~$14 million by 2006. Hypothetically, it would have likely cost upwards of $20-25 million to generate a 'finished' human genome sequence - expensive, but still considerably less so than for generating the first reference human genome sequence.

How much does it cost to sequence a human genome in 2016 (i.e., today)?

The decade following the HGP brought revolutionary advances in DNA sequencing technologies that are fundamentally changing the nature of genomics. So-called 'next-generation' DNA sequencing methods arrived on the scene, and their effects quickly became evident in terms of **lowering genome-sequencing costs**; note that these NHGRI-collected data are 'retroactive' in nature, and do not always accurately reflect the 'projected' costs for genome sequencing going forward).

In 2015, the most common routine for sequencing an individual's human genome involves generating a 'draft' sequence and comparing it to a reference human genome sequence, so as to catalog all sequence variants in that genome; such a routine does not involve any sequence finishing. In short, nearly all human genome sequencing in 2015 yields high-quality 'draft' (but unfinished) sequence. That sequencing is typically targeted to all exons (whole-exome sequencing) or aimed at the entire ~6-billion-base genome (whole-genome sequencing), as discussed above. The quality of the resulting 'draft' sequences is heavily dependent on the amount of average base redundancy provided by the generated data (with higher redundancy costing more).

Adding to the complex landscape of genome sequencing in 2015 has been the emergence of commercial enterprises offering genome-sequencing services at competitive pricing. Direct comparisons between commercial versus academic genome-sequencing operations can be particularly challenging because of the many nuances about what each includes in any cost estimates (with such details often not revealed by private companies). The cost data that NHGRI collects from its funded genome-sequencing groups includes information about a wide range of activities and components, such as: reagents, consumables, DNA-sequencing instruments, certain computer equipment, other equipment, laboratory pipeline development, laboratory information management systems, initial data processing, submission of data to public databases, project management, utilities, other indirect costs, labor, and administration. Note that such cost-accounting does not typically include activities such as quality assurance/quality control (QA/QC), alignment of generated sequence to a reference human genome, sequence assembly, genomic variant calling, or annotation. Almost certainly, companies vary in terms of which of the items in the above lists get included in any cost estimates, making direct cost comparisons with academic genome-sequencing groups difficult. It is thus important to consider these variables - along with the distinction between retrospective versus projected costs - when comparing genome-sequencing costs claimed by different groups. Anyone comparing costs for genome sequencing should also be aware of the distinction between 'price' and 'cost' - a given price may be either higher or lower than the actual cost.

Based on the **data** collected from NHGRI-funded genome-sequencing groups, the cost to generate a high-quality 'draft' whole human genome sequence in mid-2015 was just above $4,000; by late in 2015, that figure had fallen below $1,500. The cost to generate a whole-exome sequence was generally below $1,000. Commercial prices for whole-genome and whole-exome sequences have often (but not always) been slightly below these numbers.

Looking Ahead

Innovation in genome-sequencing technologies and strategies does not appear to be slowing. As a result, one can readily expect continued reductions in the cost for human genome sequencing. The key factors to consider when assessing the 'value' associated with an estimated cost for generating a human genome sequence - in particular, the amount of the genome (whole versus exome), quality, and associated data analysis (if any) - will likely remain largely the same. With new DNA-sequencing platforms anticipated in the coming years, the nature of the generated sequence data and the associated costs will likely continue to be dynamic. As such, continued attention will need to be paid to the way in which the costs associated with genome sequencing are calculated.

Contact

Kris A. Wetterstrand, M.S.

Program Operations Lead

Division of Extramural Operations