Genome Biology

OPINION

The real cost of sequencing: higher than you think!

Andrea Sboner[1,2], Xinmeng Jasmine Mu[1], Dov Greenbaum[1,2,3,4,5], Raymond K Auerbach[1] and Mark B Gerstein*[1,2,6]

Abstract

Advances in sequencing technology have led to a sharp decrease in the cost of 'data generation'. But is this sufficient to ensure cost-effective and efficient 'knowledge generation'?

Keywords Bioinformatics, costs of sequencing, data analysis, experimental design, next-generation sequencing, sample collection.

In recent years, a technological revolution has enabled a series of advances in our knowledge of complex biological processes from development to disease, largely fueled by massively parallel sequencing technology, or next-generation sequencing (NGS) [1,2]. Applications of NGS include studies of entire genomes (whole-genome sequencing), investigations of smaller functional portions of the genome (exome sequencing), and analysis of the transcribed genome (RNA-seq) and protein-DNA binding sites (ChIP-seq). We are still in the early days of this 'revolution', with new technological improvements expected to be introduced in the next few years: longer reads, faster processing, and a growing number of more 'exotic' experimental applications that can benefit from sequencing technologies, such as the chromatin conformation capture family of experiments to reveal a structural view of the genome, and methods such as GRO-seq and CLIP-seq, and several others, to study protein-RNA binding [3-5].

Advances in sequencing technologies paved the way for launching the $1,000 genome challenge in 2005 [6-8], an almost impossible goal to imagine at the time. In fact, the cost of sequencing the first human genome was about $3 billion [9], and it took several international institutes, hundreds of researchers and 13 years to complete. However, in the past few years the cost of sequencing has declined exponentially: James Watson's genome was completed for less than $1 million [10]; by 2009 the cost for a whole-genome sequence dropped to $100,000 [11]. Hence, today, a mere 10 years after the completion of the first draft of the human genome, the goal of the $1,000 genome seems surprisingly close, and it is now conceivable that this will be a step towards even cheaper genomes. Nevertheless, it has become clear that the act of sequencing DNA (or cDNA) is only one aspect of a more complex story (Figure 1).

Cost of sequencing versus cost of computation

The National Human Genome Research Institute (NHGRI) has tracked the cost of sequencing in the centers it funds. Analyzing the data revealed a stunning picture [12]. From 2008, the cost of sequencing dropped faster than what would have been expected from Moore's law (a term used to describe a trend in the computer industry). Moore's law states that the number of transistors of an integrated circuit doubles approximately every 2 years, but it is also applicable to several other digital electronic devices [13,14]. This implies that while we will be able to generate more and more sequence bases at a fixed cost, we will soon lack the facilities to store, process, analyze and maintain the data generated (Figure 2). However, the NHGRI survey ignores other cost components, including the following 'non-production' activities: quality assessment/control for sequencing projects, technology development to improve sequencing pipelines, development of bioinformatics/computational tools to improve sequencing pipelines or to improve downstream sequence analysis, management of individual sequencing projects, informatics equipment, and data analysis downstream of initial data processing (for example, sequence assembly, sequence alignments, identifying variants and interpretation of results) [12]. These are obviously integral to any sequencing project, and need to be accounted for in the overall costs. Moreover, the economic impact of these activities is likely to play a different role depending on the users: the costs of streamlining the sequencing pipeline can affect data providers in different ways, whereas downstream analysis costs have a similar impact irrespective of the investigators.

Here we highlight the relevance and important of these often unaccounted costs using examples from a typical whole-genome sequence and RNA-seq project (Table 1) We consider the associated costs incurred during a

*Correspondence: mark.gerstein@yale.edu
[1]Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT 06520, USA
Full list of author information is available at the end of the article

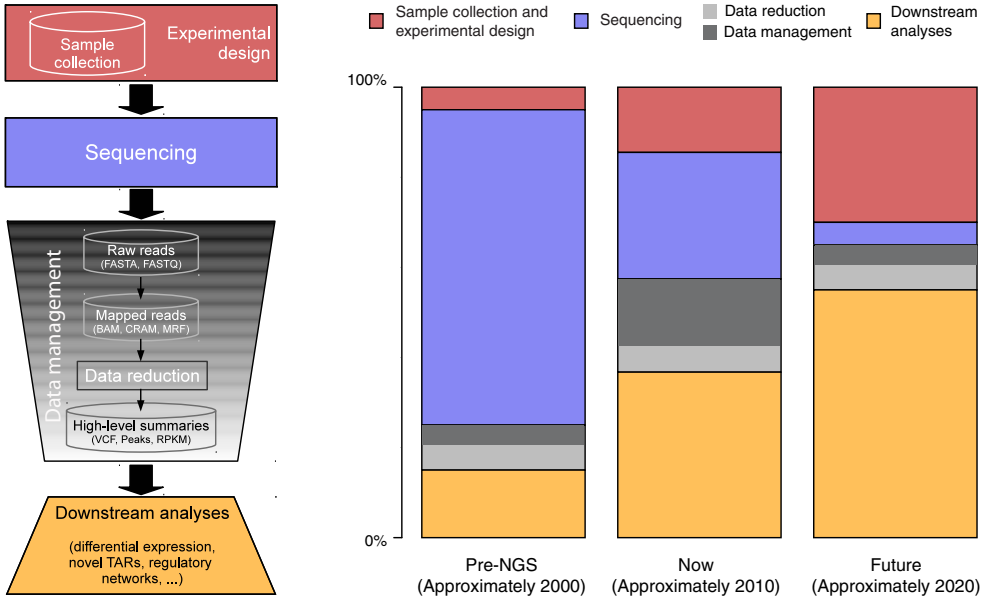

Figure 1. Contribution of different factors to the overall cost of a sequencing project across time. Left, the four-step process: (i) experimental design and sample collection, (ii) sequencing, (iii) data reduction and management, and (iv) downstream analysis. Right, the changes over time of relative impact of these four components of a sequencing experiment. BAM, Binary Sequence Alignment/Map; BED, Browser Extensible Data; CRAM, compression algorithm; MRF, Mapped Read Format; NGS, next-generation sequencing; TAR, transcriptionally active region; VCF, Variant Call Format.

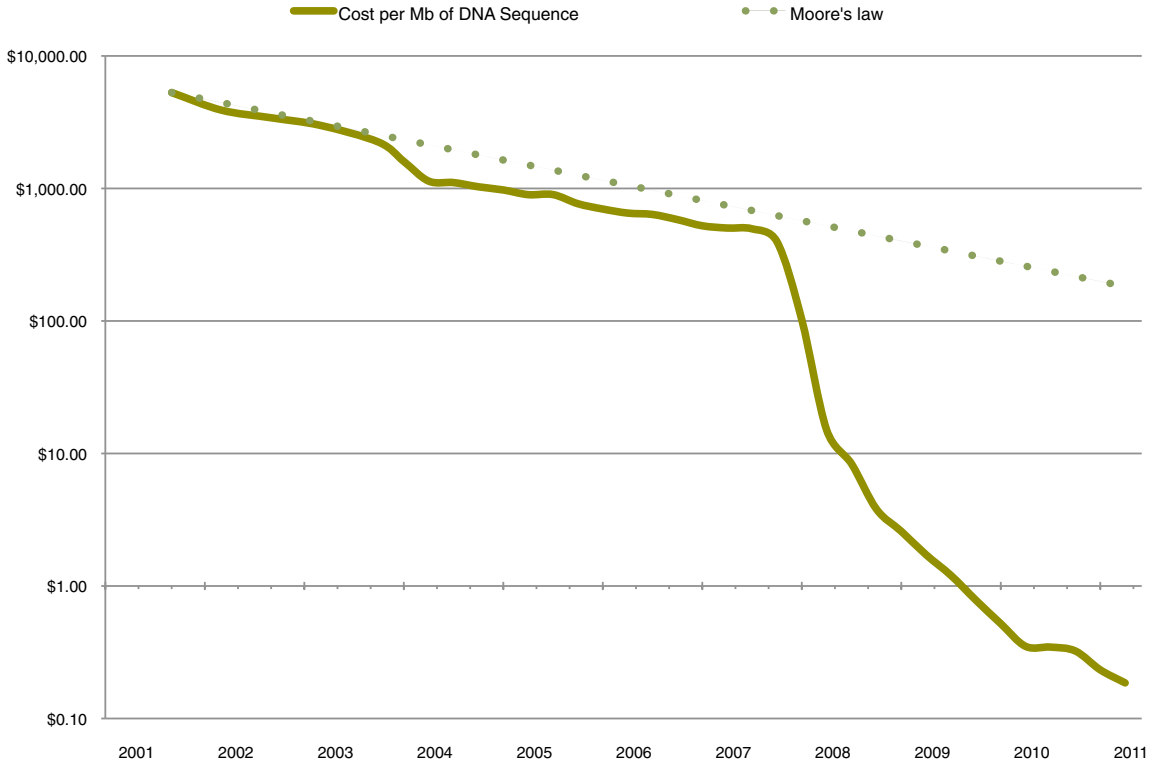

Figure 2. Cost of 1 MB of DNA sequencing. Decreasing cost of sequencing in the past 10 years compared with the expectation if it had followed Moore's law. Adapted from [11]. Cost was calculated in January of each year. MB, megabyte.

Table 1. Estimates of a whole-genome sequence and an RNA-seq experiment using Illumina HiSeq 2000 machine

	Whole genome sequencing				RNA-Seq			
		2011 cost	output	2011 time		2011 cost	output	2011 time
Sample collection and experimental design	from blood samples (easy to collect) to brain tissue (hard to collect)	~$100 onwards		from a few hours to several days	same as for whole genome sequencing			
Sequencing	library preparation + running the sequencer (whole dual flow cell)	~$6500 = ~$500 + ~$6000	~380M reads/lane; 1 individual: ~1140M total reads (~3 lanes for a 30x coverage); ~250Gb (intermediate files)	~11-12 day	library preparation + running the sequencer (whole dual flow cell)	~$3300 = ~$300 + ~$3000	~380M reads/lane	~12-14 day
	Data storage, low-level processing							
	Alignment (transfer* and storing raw data + mapping)	~$40 = ~$33 + ~$7	300Gb (BAM file)	~1/2 day *** (including transferring 250Gb FASTQ ~7.5 hrs)	Alignment (transfer* and storing raw data + mapping)	~$5 = ~$3 + ~$2	~30Gb (BAM); ~22Gb (MRF)	< 2 hrs ***
	(data transfer and storage for 10 days)*; **	~$40		~8.5 hrs	(data transfer and storage for 10 days)*; **	<$4		< 1hr
Data reduction and management	*High-level summaries****							
	SNP calling (compute + transfer out)	<$5 = ~$4 + ~$0.60	< 1Gb	~3 hrs	Gene and exon expression quantification	<$1	<1M	<1hr (1 CPU)
	Indel calling (compute + transfer out)	<$35 = ~$32 + ~$0.60	< 1Gb	~1 day	Isoform quantification	~$6	<1M	~4h
	SV calling (compute + transfer out)	<$35 = ~$32 + ~$0.60	< 1Gb	~1 day				
Downstream analyses		>$100K	~310Gb	months		>$100K	~30Gb	months
Total of sequencing, data management and reduction	~$6500	~310Gb	~15 days		~3500	~30Gb	~12-14 days	

[a]Assuming a 10 MB/s transfer rate. [b]The cost of transferring 300 GB (BAM file) if the mapping is performed locally. [c]Sixteen CPUs were used for all calculations, unless specified. BAM, Binary Sequence Alignment/Map; CPU, central processing unit; GB, gigabyte; MB, megabyte; SNP, single nucleotide polymorphism; ~, approximately. The table gives an estimation of the current costs of a whole-genome sequencing experiment and a functional genomic experiment (RNA-seq) using an Illumina HiSeq 2000 machine. The cost of sequencing is based on that reported by the Center for Cancer Research [53]. It is assumed that the same human sample is used for the genome sequence as well as for the RNA-seq experiment. In this scenario, a group of four technicians and bioinformaticians can run the entire pipeline. The costs related to data processing are estimated by considering all tasks performed in the Amazon Web Services 'cloud' environment. Pricing is based on a 'US standard' of the S3 Amazon Services (storage) [54] and a cost of $0.68 per hour (US, East Virginia) for the use of the Amazon EC2 (computation) [55] (July 2011).
See Additional file 1 for a version of this table in a colored layout for easier reading.

four-step process: (i) sample collection and experimental design, (ii) sequencing the sample, (iii) data reduction and management, and (iv) downstream analyses (that is, the complex analyses enabled by NGS and by the experimental design) (Figure 1, Table 1).

Experimental design and sample collection

As the cost of producing sequence reads decreased, more and more applications have been designed to exploit this technology. In addition to 'traditional' whole genome and exome sequencing, and traditional functional genomics experiments, such as transcriptome analysis (small and long RNA-seq) and chromatin immunoprecipitation (ChIP-seq), other applications of NGS have been designed: array-capture followed by sequencing (for example, enriching for transposable elements), fosmid pools sequencing (to facilitate haplotype phasing) [15], metagenomics sequencing, bisulfite sequencing and methylated DNA immunoprecipitation (MeDIP-Seq) to study DNA methylation patterns [16,17]. Some of these cutting-edge experimental designs require complex molecular and cellular biology experiments, such as chromosome sorting, to prepare the library for sequencing, thus adding considerably to the overall experimental costs.

In addition to the experimental design, the nature of the sample used can significantly alter the overall budget for a sequencing project. In the past, cell lines were the preferred biological material to sequence for obvious reasons: unlimited availability, well-known properties, and they are ideal for calibration and comparison with more traditional techniques. The advantages of these cell lines outweighed the noise introduced by the artifacts of immortalized cell lines. With an improved and more reliable sequencing technology, the questions addressed by researchers have also become more challenging, ranging from assessing human genetic variability to understanding the underlying biology of complex diseases, necessitating more time-consuming collection of clinically relevant samples, For example, studying 'trios' (that is, father, mother and child) to search for rare variants requires a considerable effort in terms of coordinating the experimental design compared with collecting blood samples from normal healthy volunteers. Similarly, the collection of matched cancer and normal frozen tissues needs proper standardized protocols to obtain high-quality biological material for sequencing. Working with a standard cell line, such as HeLa, requires less effort than working with brain tissue samples (where the dissection takes about 8 h and requires at least two people; Table 1).

If clinically relevant samples from patients are to be used, then we must consider the costs associated with protecting patients' rights. As more genomes are sequenced and we become better at identifying personal information from them, the protection of the raw and analyzed data is crucial to avoid compromising even traditional research studies because patients are wary of donating their samples for research. Therefore, appropriate sample collection, informed consent procedures and the development of infrastructure and security required to protect patients' data need to be considered. Institutional review boards (IRBs) or ethics committees are often charged with overseeing these tasks. However, IRBs may constrain or limit the nature of whole-genome sequencing projects in light of concerns over downstream usage of genomic data. When faced with high-risk studies, IRBs may prevent a researcher from conducting a particular study, or request that the investigator seek approval from governmental agencies, potentially resulting in bureaucratic delays, which should be factored in to project budgets in advance.

Furthermore, technological advancements over the course of the research project, while providing exciting new avenues of research, might effectively render the initial informed consent agreement moot, either requiring the acquisition of a new consent or the curtailment of a promising research program due to lack of consent.

Once private data are collected there are significant costs in maintaining a secure and functioning archival system. Compared with a similar heavy-data scenario, such as picture archiving and communication systems (PACS) in radiology departments, the costs could easily reach a few million US dollars [18]. This becomes even more complicated if the laboratory storing the data also wants to provide secure access to a wider audience over the web in terms of the actual cost of the infrastructure and software, and also the man-hours necessary to maintain the system updated [19].

Sequencing the sample

Sequencing entails the preparation of the library from the sample and running the sequencer to generate sequence reads. Streamlining the sequencing pipeline is an important aspect, especially for large-scale data-producing centers (for example, genomics centers, core sequencing centers at universities), which can benefit the most by leveraging economies of scale. However, the solutions implemented by large-scale data producers are different from those required by smaller centers. Standardized sample collection and identification procedures, and library preparation according to best practices, are the mainstay of large-scale data producers, where the main effort is to optimize and automate the entire process for many sequencing projects. This approach ensures high-quality and reproducible data, but it also requires a large-scale infrastructure. The main challenges are to accommodate the needs of multiple users, to achieve a low down-time of the sequencers, to adapt to new reagents

and experimental protocols quickly, and to reorganize the efforts of technicians based on newer, faster machines in an efficient manner. These issues can be addressed in large-scale centers by using robots to automate the process. It is likely that the initial costs of implementing this infrastructure will be partially offset by the sequencing 'running' costs, which are decreasing.

At the opposite end of the spectrum are smaller groups and individual investigators, perhaps with only one sequencer. They do not need a complex infrastructure to successfully complete their sequencing process, and the current solutions offered by the vendors are sufficient. In this scenario, a major burden is represented by the acquisition of the technical expertise to run the sequencer and keep up with the new protocols.

Data reduction and management

The basic output of a sequencing run is the set of sequence reads (in our example, about 6 billion sequence reads corresponding to approximately 600 billion nucleotides - approximately 187× coverage of the human genome; Table 1). Typically, this information is captured, rather inefficiently, in FASTQ files, which include the quality scores of each base. Three types of computation need to be distinguished: the first aims to efficiently manage this wealth of data by generating compressed data structures of the low-level raw data that can be easily accessed; the second aims at automatically extracting high-level summaries from these low-level raw data; and the third includes all downstream analyses that one can devise.

Data management: storage and transfer

In the first category, the alignment of those reads to a reference is the initial computational step. Different strategies can be adopted, but the final result is the assignment of the reads to their genomic location (while mapping to a known reference is the simplest procedure, some strategies may perform *de novo* assembly of the reads; we believe that this comprises a small fraction of experiments). Thus, files including mapped and unmapped reads can be considered the basic archival unit of an NGS experiment, since one can perform all the downstream analyses starting from the mapped reads, as well as extract all the reads and use a different alignment approach. Clearly, efficient compression strategies (such as Binary Sequence Alignment/Map (BAM) [20]) are desperately needed to reduce the size of those files. The amount of raw data generated will increase as sequencers will provide longer reads in the future. Indeed, the massive amounts of data constitute a major burden from several viewpoints. The capacity for storing and archiving the data is increasing at a slower pace (that is, following Moore's law) than sequencing throughput, suggesting that re-sequencing a sample is more cost-effective than keeping the data archived. However, this solution is justifiable only when working with immortalized cell lines or model organisms. It is not an option with very valuable samples, such as clinical tissues or limited amounts of antibodies.

Large datasets also pose a problem in terms of data transfer. Computational biologists have been taking full advantage of the 'open-world' paradigm, where datasets are freely shared among investigators via centralized repositories, such as Gene Expression Omnibus, ArrayExpress, UniProt and others [21-23]. However, this practice is challenged by the vast amount of data that would need to be piped through the existing networking infrastructure for uploading and downloading the data to and from those repositories. For example, at a network bandwidth of approximately 10 megabytes (MB)/s, it would require approximately 8.5 h to transfer 300 gigabytes (GB) [24]. Recently, a reference-based compression algorithm (CRAM) has been proposed that addresses both issues of storing and transferring data [25]. Inspired by video compression, this method currently achieves a considerable reduction in storage space required: 10-fold to 30-fold. Briefly, it keeps only the differences with respect to a reference and considers the relative location of a read based on the location of the previous read. Moreover, unmapped reads are assembled 'on the fly' to achieve even higher compression rates; this is a strategy that might also lead to novel discoveries. An irreducible fraction of reads, however, will not be mapped. This observation leads to a paradox: most of the disk space is required to store 'noise' (that is, the unmapped reads; obviously, this is an oversimplification, as not all unmapped reads can be considered as noise and some might actually lead to new discoveries). In the future, by identifying what information we can afford to lose, we could achieve higher levels of compression (approximately 200-fold), helping us to catch up with the decreased cost of sequencing [26].

Following our example, data storage, transfer and mapping would cost about $40 for about 12 h of compute work for whole-genome sequencing, or about $5 and a few hours for RNA-seq (Table 1).

Data reduction: high-level summaries

The second computational category regards the processing of the reads in order to obtain meaningful high-level summaries. For example Variant Call Format (VCF) files [27] describe genomic variants, including SNPs, short indels, structural variations (deletions, duplications and, more recently, inversions and translocations), Browser Extensible Data (BED) files report binding sites for ChIP-seq experiments, and tab-delimited files can provide RPKM values quantifying gene and exon expressions

[28]. Here, it is clear what types of output one aims to get (VCF files, BED files, and so on), although there is not yet a fixed set of computational algorithms to achieve this goal. The creation of these high-level summaries is analogous to what is commonly performed in other fields. For example, protein structures are now encoded in Protein Data Bank (PDB) files [29]. Very few people would analyze the raw diffraction images from which the PDB files were generated. Similarly, raw digital images acquired with modern smart phones undergo a tremendous level of processing to produce an MPEG4 compressed video that can be immediately shared with friends; and no one needs the raw high-definition images to enjoy their friends' pictures.

These high-level summaries are likely to become the most valuable piece of information from sequencing experiments. Many researchers interested in genomic variants would need only the VCF files, which are considerably smaller than the whole set of mapped reads: about 170 MB for approximately 3 million SNPs, about 8 MB for 300,000 indels, and about 0.2 MB for approximately 1,500 structural variations. Similarly, the files with expression levels from RNA-seq or binding sites from ChIP-seq are quite small: about 0.5 MB for approximately 25,000 genes (RNA-seq), about 4 MB for approximately 230,000 exons (RNA-seq), and about 3 MB for approximately 80,000 binding sites. The generation of these files takes about 1 day and about $35 each for whole-genome sequencing and a few hours and about $6 for RNA-seq (Table 1).

Complexity of data reduction: hybrid solutions

In addition to their smaller, almost negligible size, high-level summaries also have the advantage of reducing the potential privacy issues related to NGS data, which are more revealing of the underlying individual and include the potential for a full characterization. Although this solution seems attainable for most final users, it is not viable for researchers developing the computational tools to mine the data, who would still need the raw sequence data. A hybrid solution is to use compressed intermediate summaries, such as Mapped Read Format (MRF) [30]. Similar to CRAM, MRF focuses on encoding the location of the reads. In addition, it can protect the most sensitive information (that is, the sequences) by separating them from the alignment information. This has the advantage of enabling all the primary analyses where only the read locations are needed, such as quantifying expression levels, and of considerably reducing the file size.

The lack of agreed standards is a problem for data reduction approaches. One of the main benefits of centralized repositories has been to define standardized metadata and data formats that proved to be quite useful for the final users of those data. However, it has also introduced an additional burden to the data providers to generate those metadata and prepare the data submission in the proper format. A 'broker' who has expertise in data processing can efficiently deal with all these aspects, and its costs should thus be taken into account, especially for smaller groups.

To the cloud!

With the majority of computational tools for NGS data analyses being open source, computational-related costs are typically only hardware infrastructure (data storage and computational power) and human resources to properly set up the computational environment and run the processing pipeline. There are currently three solutions for this type of computational analysis: subscribing to sequencing-as-a-service offerings by sequencing companies, processing the data in-house, and using cloud computing services.

A few sequencing companies, such as Complete Genomics, DNANexus and Spiral Genetics among others, provide data reduction services, some also following whole-genome DNA sequencing. These data reduction pipelines are proprietary NGS data processing facilities and their cost to the customer is minimal in addition to the sequencing service (a few thousand dollars per genome). The nature of the data processing pipeline confers customers less control on the quality, reproducibility and other details of the data reduction pipeline.

A common aspect for analyzing the vast amount of data is the use of parallel computing. Typically, this would entail the use of a computer cluster, a facility not available to everyone, which requires a significant amount of investment that easily reaches thousands or millions of dollars. While large-scale centers might benefit from economies of scale, smaller institutions or individual laboratories might be priced out of these facilities.

An alternative to building in-house hardware infrastructures is to resort to commercialized cloud computing services or computation-as-a-service [24,31]. In this scenario, users 'rent' computational power to perform the analysis over the Internet. Amazon.com was one of the first to offer these general-purpose cloud services with the Amazon Web Services. Several service providers now offer solutions based on this computing paradigm, including Microsoft Windows Azure, The Rackspace Cloud and the US Federal Government, among others [31-34]. This solution might appeal to smaller groups as well as large-scale genomics centers. Smaller groups can delegate the burden of implementing, maintaining and running complex infrastructures and security protocols to the service providers while concentrating on obtaining meaningful results from their data. Similar to our illustrative example reported in Table 1, Langmead and colleagues [35] were able to re-align a human genome

and detect SNPs for a cost of the cloud use of approximately $113: $28 for transferring and storing the raw data, the remaining $85 to run the analyses on 320 central processing units in about 2.5 h. Larger organizations, which might already have their own computer clusters, can still make use of these services. Indeed, some service providers allow 'hybrid' solutions where private resources can be securely extended into the cloud. This solution is particularly useful to address 'spikes' in computer usage (that is, when simultaneous requests for computational resources exceed the local cluster capacity).

Another advantage of cloud computing is that standardized computational tools can be made available through the cloud such that even casual users can execute their analysis using state-of-the-art approaches (Galaxy, Myrna, JCVI BioLinux, and so on [31,36-39]). Although cloud computing seems quite attractive, there are other aspects that one must consider. First, cloud computing is particularly effective when a lot of computation, such as molecular dynamic simulations, is required. However, in genomics, processing is tightly linked to the large amount of data. Transferring massive amounts of data to the cloud may be time consuming and prohibitively expensive for small laboratories in research institutions with standard network bandwidth (approximately 10 MB/s). Moreover, by transferring the data to the cloud, one would create a redundancy that is not efficient, unless the output from the sequencers is directly transferred to the cloud. Reference-based compression can ameliorate some of these issues. A 10-fold reduction of a 300 GB BAM file (whole-genome sequence; Table 1) will reduce the overall cost to approximately $4 for transferring the data to the cloud and storing them for 10 days, in addition to reducing the transfer time considerably (approximately 1 h). Second, ensuring that data from human subjects remains private is important. Guaranteeing that the sequence data are handled according to the individual consent and protected in accordance with the various regulations is mandatory, particularly when using government funding. Although some companies already provide tools to deal with these issues in the cloud, the legal aspects of handling genomic data are still in flux. For example, in the USA, depending on the source of genetic information, Health Insurance Portability and Accountability Act (HIPAA) rules may or may not fully apply (that is, direct-to-consumer genomics companies probably do not need to comply with HIPAA privacy protections). Moreover, it is not clear whether the Fourth Amendment right against search and seizure for genomic data is maintained in the cloud.

A solution to the privacy issues may be that governmental funding agencies, such as the National Institutes of Health in the USA, create a 'private' cloud environment where a researcher can conduct his/her genomics research in accordance with the legal framework. This approach may also help in lowering the share of the agency's budget spent on funding the computational infrastructure of various research organizations. In this scenario, an important aspect is the definition of a shared common legal and ethical framework if data are shared internationally.

Downstream analyses

Whether the sequencing and initial data analyses are carried out externally or in house, researchers have to face the downstream analyses of NGS tailored to specific research projects. Hundreds of tools have been developed to unravel the complexity of biological mechanisms hidden in the sequence reads. Whole-genome DNA sequencing enables analyses such as looking for natural selection in genomic elements [40], investigating demographic history changes by comparing with the Neanderthal genome [41], and studying cross-species conservation, recombination hotspots and gene conversion [42]. It also allows us to perform phasing (family-, population- or physical-based) and construct diploid personal genomes [43]. A single RNA-seq experiment can be exploited to investigate not only gene expression levels, but also transcript expression levels, to identify differentially expressed 'elements' [44]. It can further provide information about novel transcriptionally active regions or novel splice sites [45], identify the presence of chimeric transcripts [46,47], and investigate the extent of RNA editing or allele-specific expression [48-50]. Moreover, the integration of multiple assays allows researchers to address more complicated questions with both a resolution and a breadth that was unforeseeable only a few years ago. For example, combining genotyping information (perhaps obtained from whole-genome sequencing) and RNA-seq data, it is possible to investigate expression quantitative trait loci [49,50]. Analyzing and combining ChIP-seq experiments allows the determination of putative functional elements, such as enhancers and promoters, and the generation of regulatory networks to study their static and dynamic aspects.

In contrast to the cost of hardware, the cost of human resources is hardly quantifiable at the current state of the art. No streamlined, standardized approach is yet available for the users, either an experienced or a casual one. Hence, considerable efforts are required to properly install, configure and run the computational pipeline to perform the data reduction as well as the more complex data analyses described here.

Some computational frameworks, such as Galaxy, provide users with several tools to perform data analysis [36]. However, as pointed out by the Galaxy developers in the 'Know what your doing' section of the framework [51]: 'There is no such thing (yet) as an automated

gearshift in (...)' (replace (...) with your favorite analysis: for example, short read mapping, splice-junction identification). Moreover, the time and effort considered for the data analysis pipeline in Table 1 includes only the estimation of actual computing time (that is, 'machine time'). There is an unpredictable amount of extra 'human time' spent on comparing and choosing suitable and reasonable software tools, learning how to install, configure and execute them, estimating the effects of tuning the parameters, interfacing input/output formats for serial modules in the pipeline, and debugging and streamlining. Sometimes, this latter phase is more time consuming than the actual processing time.

Finally, as pointed out by Elaine Mardis, the interpretation of the results, in many cases, requires a multidisciplinary team to make sense of all the primary data [52]. Linking a genetic variation to a phenotype, especially a clinically relevant one, requires a lot of expertise and effort, first to identify highly confident variants; second, to estimate their functional impact; third, to select from among the functional ones those that are correlated to the phenotype. None of these steps can be carried out in an automated fashion. This, in turn, further supports why no streamlined data analysis pipeline currently exists and the needs to customize and tailor the process to each research project. A team of bioinformaticians, statisticians, geneticists, biologists and physicians is required to translate the information in the primary data into useful knowledge to understand the impact of genomic variants in biological systems; this can often take weeks or months of extensive experimental validations using animal models or cell lines.

Future scenarios

The impressive advances made possible by the introduction of NGS are rapidly pushing the scientific community toward uncharted territories. NGS is currently quite expensive and mainly tailored to a research setting. However, the rapid decrease in its basic operational costs will make their utilization affordable by a larger group of investigators and in a diagnostic setting. New experimental protocols, new equipment tailored to interrogate smaller portions of the genome (for example, Ion Torrent Personal Genome Machine PGM, Illumina MiSeq) and new applications of sequencing will further broaden the user base of these technologies. In turn, this means that more 'challenging' samples will be analyzed with more complex experiments thus increasing the fraction of cost due to sample collection and experimental design (Figure 1).

The cost of sequencing is decreasing rapidly (that is, faster than Moore's law; Figure 2) in contrast to storage, which is decreasing in line with Moore's law. The bottleneck represented by data storage, maintenance and transfer should be addressed by better data compression algorithms, and by standardized high-level summary data. These summaries should be sufficiently expressive as to enable additional analyses and to be easily integrated with other data sets for comparative analysis. However, this approach would tackle only one aspect of the NGS data complexity. The full interpretation of the primary data is going to constitute a major expense (Figure 1).

In conclusion, the rapid decrease in the cost of 'data generation' has not been matched by a comparable decrease in the cost of the computational infrastructure required to mine the data. Cloud computing is one promising direction for smaller groups and large-scale organizations to address some of the underlying issues, with the former clearly benefiting the most. However, careful considerations about privacy of the data and network bandwidth must be taken into account. The major burden, however, will be represented by the downstream analysis and interpretation of the results. The bioinformatics and computational biology community needs to design and develop better computational algorithms and approaches to speed up the 'knowledge generation' pipeline, taking advantage of the possibilities opened up by the cloud. More automated and more reliable tools to process and analyze NGS data are certainly needed and constitute essential steps towards the realization of personalized genomics and medicine, one of the main challenges of the 21st century.

Additional file 1

Table S1. Estimates of a whole-genome sequence and an RNA-seq experiment using Illumina HiSeq 2000 machine.
A version of Table 1 with colored layout for improved ease of reading.

Abbreviations

BAM, Binary Sequence Alignment/Map; BED, Browser Extensible Data; CRAM, compression algorithm; GB, gigabyte; HIPAA, Health Insurance Portability and Accountability Act; IRB, institutional review board; MB, megabyte; MRF, Mapped Read Format; NGS, next-generation sequencing; NHGRI, National Human Genome Research Institute; PDB, program database; SNP, single nucleotide polymorphism; VCF, Variant Call Format.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

We would like to thank Naoki Kitabayashi, Cristina Sisu, Joel Rozowsky, Alexej Abyzov, Yuka Imamura Kawasawa and Jing Leng for useful discussions. We would like to acknowledge the National Institutes of Health for funding.

The opinions of Dov Greenbaum expressed in this piece are his own and do not necessarily represent, nor should they be imputed to represent, the opinion of his law firm, any of its employees, or its clients.

Author details

[1]Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT 06520, USA. [2]Department of Molecular Biology and Biophysics, Yale University, New Haven, CT 06520, USA. [3]Sanford T Colb & Co. Intellectual Property Law, 4 Shaar Hagai, Marmorek, Rehovot, 76122, Israel. [4]Center for Law

and the Biosciences, Stanford Law School, Stanford University, Stanford, CA 94305, USA. [5]Sholar in Residence Center for Health Law, Bioethics and Health Policy Kiryat Ono College, 55000, Israel. [6]Department of Computer Science, Yale University, New Haven, CT 06520, USA.

Published: 25 August 2011

References

1. Metzker ML: **Sequencing technologies – the next generation.** *Nat Rev Genet* 2010, **11:**31-46.
2. Mardis ER: **The impact of next-generation sequencing technology on genetics.** *Trends Genet* 2008, **24:**133-141.
3. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, Sandstrom R, Bernstein B, Bender MA, Groudine M, Gnirke A, Stamatoyannopoulos J, Mirny LA, Lander ES, Dekker J: **Comprehensive mapping of long-range interactions reveals folding principles of the human genome.** *Science* 2009, **326:**289-293.
4. Core LJ, Waterfall JJ, Lis JT: **Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters.** *Science* 2008, **322:**1845-1848.
5. Licatalosi DD, Mele A, Fak JJ, Ule J, Kayikci M, Chi SW, Clark TA, Schweitzer AC, Blume JE, Wang X, Darnell JC, Darnell RB: **HITS-CLIP yields genome-wide insights into brain alternative RNA processing.** *Nature* 2008, **456:**464-469.
6. Bennett ST, Barnes C, Cox A, Davies L, Brown C: **Toward the 1,000 dollars human genome.** *Pharmacogenomics* 2005, **6:**373-382.
7. Service RF: **Gene sequencing. The race for the $1000 genome.** *Science* 2006, **311:**1544-1546.
8. Mardis ER: **Anticipating the 1,000 dollar genome.** *Genome Biol* 2006, **7:**112.
9. **National Human Genome Research Institute** [http://www.genome.gov/11006943]
10. Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, He W, Chen Y-J, Makhijani V, Roth GT, Gomes X, Tartaro K, Niazi F, Turcotte CL, Irzyk GP, Lupski JR, Chinault C, Song X-zhi, Liu Y, Yuan Y, Nazareth L, Qin X, Muzny DM, Margulies M, Weinstock GM, Gibbs RA, Rothberg JM: **The complete genome of an individual by massively parallel DNA sequencing.** *Nature* 2008, **452:**872-876.
11. Brown C: **Guest post by Clive Brown: The disruptive power of cheap DNA sequencing** [http://www.genomesunzipped.org/2011/06/guest-post-by-clive-brown-the-disruptive-power-of-cheap-dna-sequencing.php]
12. Wetterstrand K: **DNA sequencing costs: data from the NHGRI large-scale genome sequencing program** [http://www.genome.gov/sequencingcosts/]
13. Walter C: **Kryder's Law.** *Sci Am* 2005, **293:**32-33.
14. Therani R: **As we may communicate** [http://www.tmcnet.com/articles/comsol/0100/0100pubout.htm]
15. Kitzman JO, MacKenzie AP, Adey A, Hiatt JB, Patwardhan RP, Sudmant PH, Ng SB, Alkan C, Qiu R, Eichler EE, Shendure J: **Haplotype-resolved genome sequencing of a Gujarati Indian individual.** *Nat Biotechnol* 2011, **29:**59-63.
16. Cokus SJ, Feng S, Zhang X, Chen Z, Merriman B, Haudenschild CD, Pradhan S, Nelson SF, Pellegrini M, Jacobsen SE: **Shotgun bisulphite sequencing of the *Arabidopsis* genome reveals DNA methylation patterning.** *Nature* 2008, **452:**215-219.
17. Down TA, Rakyan VK, Turner DJ, Flicek P, Li H, Kulesha E, Graf S, Johnson N, Herrero J, Tomazou EM, Thorne NP, Backdahl L, Herberth M, Howe KL, Jackson DK, Miretti MM, Marioni JC, Birney E, Hubbard TJP, Durbin R, Tavare S, Beck S: **A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis.** *Nat Biotechnol* 2008, **26:**779-785.
18. **PACS ROI** [http://www.healthimaging.com/index.php?option=com_articles&view=article&id=3527:Case%20Studies]
19. Smith A, Greenbaum D, Douglas SM, Long M, Gerstein M: **Network security and data integrity in academia: an assessment and a proposal for large-scale archiving.** *Genome Biol* 2005, **6:**119.
20. 1000 Genome Project Data Processing Subgroup: **The Sequence Alignment/Map format and SAMtools.** *Bioinformatics* 2009, **25:**2078-2079.
21. Barrett T, Troup DB, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Muertter RN, Holko M, Ayanbule O, Yefanov A, Soboleva A: **NCBI GEO: archive for functional genomics data sets - 10 years on.** *Nucleic Acids Res* 2010, **39:**D1005-D1010.
22. Parkinson H, Sarkans U, Kolesnikov N, Abeygunawardena N, Burdett T, Dylag M, Emam I, Farne A, Hastings E, Holloway E, Kurbatova N, Lukk M, Malone J, Mani R, Pilicheva E, Rustici G, Sharma A, Williams E, Adamusiak T, Brandizi M, Sklyar N, Brazma A: **ArrayExpress update - an archive of microarray and high-throughput sequencing-based functional genomics experiments.** *Nucleic Acids Res* 2010, **39:**D1002-D1004.
23. The UniProt Consortium: **Ongoing and future developments at the Universal Protein Resource.** *Nucleic Acids Res* 2010, **39:**D214-D219.
24. Stein LD: **The case for cloud computing in genome informatics.** *Genome Biol* 2010, **11:**207.
25. Hsi-Yang Fritz M, Leinonen R, Cochrane G, Birney E: **Efficient storage of high throughput DNA sequencing data using reference-based compression.** *Genome Res* 2011, **21:**734-740.
26. Birney E: **Compressing DNA: the future plan** [http://genomeinformatician.blogspot.com/2011/05/compressing-dna-future-plan.html]
27. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, Depristo MA, Handsaker R, Lunter G, Marth G, Sherry ST, McVean G, Durbin R: **The Variant Call Format and VCFtools.** *Bioinformatics* 2011, **27:**2156-2158.
28. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B: **Mapping and quantifying mammalian transcriptomes by RNA-Seq.** *Nat Methods* 2008, **5:**621-628.
29. **Worldwide Protein Data Bank (PDB)** [http://www.wwpdb.org/docs.html]
30. Habegger L, Sboner A, Gianoulis TA, Rozowsky J, Agarwal A, Snyder M, Gerstein M: **RSEQtools: a modular framework to analyze RNA-Seq data using compact, anonymized data summaries.** *Bioinformatics* 2011, **27:**281-283.
31. Schatz MC, Langmead B, Salzberg SL: **Cloud computing and the DNA data race.** *Nat Biotechnol* 2010, **28:**691-693.
32. **Windows Azure** [http://www.microsoft.com/windowsazure/]
33. **The Rackspace Cloud** [http://www.rackspace.com/cloud/]
34. **Apps.gov** [https://www.apps.gov/cloud/main/start_page.do]
35. Langmead B, Schatz M, Lin J, Pop M, Salzberg S: **Searching for SNPs with cloud computing.** *Genome Biol* 2009, **10:**R134.
36. Goecks J, Nekrutenko A, Taylor J, Galaxy Team T: **Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences.** *Genome Biol* 2010, **11:**R86.
37. Afgan E, Baker D, Coraor N, Chapman B, Nekrutenko A, Taylor J: **Galaxy CloudMan: delivering cloud compute clusters.** *BMC Bioinformatics* 2010, **11:**S4.
38. Langmead B, Hansen KD, Leek JT: **Cloud-scale RNA-sequencing differential expression analysis with Myrna.** *Genome Biol* 2010, **11:**R83.
39. **CloudBioLinux** [http://cloudbiolinux.org/]
40. Mu XJ, Lu ZJ, Kong Y, Lam HYK, Gerstein MB: **Analysis of genomic variation in non-coding elements using population-scale sequencing data from the 1000 Genomes Project.** *Nucleic Acids Res*, in press.
41. Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, Patterson N, Li H, Zhai W, Fritz MH-Y, Hansen NF, Durand EY, Malaspinas A-S, Jensen JD, Marques-Bonet T, Alkan C, Prüfer K, Meyer M, Burbano HA, Good JM, Schultz R, Aximu-Petri A, Butthof A, Höber B, Höffner B, Siegemund M, Weihmann A, Nusbaum C, Lander ES, Russ C, *et al.*: **A draft sequence of the Neandertal genome.** *Science* 2010, **328:**710-722.
42. The 1000 Genomes Consortium: **A map of human genome variation from population-scale sequencing.** *Nature* 2010, **467:**1061-1073.
43. Rozowsky J, Abyzov A, Wang J, Alves P, Raha D, Harmanci A, Leng J, Bjornson R, Kong Y, Kitabayashi N, Bhardwaj N, Rubin M, Snyder M, Gerstein M: **AlleleSeq: analysis of allele-specific expression and binding in a network framework.** *Mol Syst Biol*, in press.
44. Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB: **Alternative isoform regulation in human tissue transcriptomes.** *Nature* 2008, **456:**470-476.
45. Sultan M, Schulz MH, Richard H, Magen A, Klingenhoff A, Scherf M, Seifert M, Borodina T, Soldatov A, Parkhomchuk D, Schmidt D, O'Keeffe S, Haas S, Vingron M, Lehrach H, Yaspo M-L: **A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome.** *Science* 2008, **321:**956-960.
46. Maher CA, Palanisamy N, Brenner JC, Cao X, Kalyana-Sundaram S, Luo S, Khrebtukova I, Barrette TR, Grasso C, Yu J, Lonigro RJ, Schroth G, Kumar-Sinha C, Chinnaiyan AM: **Chimeric transcript discovery by paired-end transcriptome sequencing.** *Proc Natl Acad Sci U S A* 2009, **106:**12353-12358.
47. Pflueger D, Terry S, Sboner A, Habegger L, Esgueva R, Lin P-C, Svensson MA, Kitabayashi N, Moss BJ, MacDonald TY, Cao X, Barrette T, Tewari AK, Chee MS, Chinnaiyan AM, Rickman DS, Demichelis F, Gerstein MB, Rubin MA: **Discovery of non-ETS gene fusions in human prostate cancer using next-generation RNA sequencing.** *Genome Res* 2011, **21:**56-67.
48. Wulff B-E, Sakurai M, Nishikura K: **Elucidating the inosinome: global**

approaches to adenosine-to-inosine RNA editing. *Nat Rev Genet* 2011, **12:**81-85.
49. Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, Nkadori E, Veyrieras J-B, Stephens M, Gilad Y, Pritchard JK: **Understanding mechanisms underlying human gene expression variation with RNA sequencing.** *Nature* 2010, **464:**768-772.
50. Montgomery SB, Sammeth M, Gutierrez-Arcelus M, Lach RP, Ingle C, Nisbett J, Guigo R, Dermitzakis ET: **Transcriptome genetics using second generation sequencing in a Caucasian population.** *Nature* 2010, **464:**773-777.
51. **Galaxy** [http://main.g2.bx.psu.edu/]
52. Mardis ER: **The $1,000 genome, the $100,000 analysis?** *Genome Med* 2010, **2:**84.
53. **Center for Cancer Research** [https://ccrod.cancer.gov/confluence/display/CCROSTP/Cost+of+Sequencing+Projects]
54. **Amazon Web Services: S3 pricing** [http://aws.amazon.com/s3/#pricing]
55. **Amazon Web Services: E2 pricing** [http://aws.amazon.com/ec2/#pricing]

doi:10.1186/gb-2011-12-8-125
Cite this article as: Sboner A, *et al.*: The real cost of sequencing: higher than you think! *Genome Biology* 2011, **12:**125.