The real cost of sequencing: scaling computation to keep pace with data generation

Paul Muir[1,2,3], Shantao Li[4], Shaoke Lou[4,5], Daifeng Wang[4,5], Daniel J Spakowicz[4,5], Leonidas Salichos[4,5], Jing Zhang[4,5], Farren Isaacs[1,2], Joel Rozowsky[4,5], Mark Gerstein[4,5,6*]

[1]Department of Molecular, Cellular, and Developmental Biology, Yale University, New Haven, CT 06520, USA.
[2]Systems Biology Institute, Yale University, West Haven, CT 06516, USA.
[3]Integrated Graduate Program in Physical and Engineering Biology, Yale University, New Haven, CT 06520, USA.
[4]Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT 06520, USA
[5]Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06520, USA
[6]Department of Computer Science, Yale University, New Haven, CT 06520, USA

*To whom correspondence should be addressed: pi@gersteinlab.org

Paul Muir – paul.muir@yale.edu
Shantao Li - shantao.li@yale.edu
Shaoke Lou - shaoke.lou@yale.edu
Daifeng Wang - daifeng.wang@yale.edu
Daniel J Spakowicz - daniel.spakowicz@yale.edu
Leonidas Salichos - leonidas.salichos@yale.edu
Jing Zhang - j.zhang@yale.edu
Farren Isaacs - farren.isaacs@yale.edu
Joel Rozowsky - joel.rozowsky@yale.edu
Mark Gerstein – pi@gersteinlab.org

Paul Muir 2/8/16 6:35 PM
Style Definition: Default Paragraph Font

Abstract: As the cost of sequencing continues to decrease and the amount of sequence data generated grows, new paradigms for data storage and analysis are increasingly important. The relative scaling behavior of these evolving technologies will impact genomics research moving forward.

Keywords: Bioinformatics; costs of sequencing; data analysis; next-generation sequencing;

Abbreviations: BAM: Binary Sequence Alignment/Map; CRAM: compression algorithm; GB: gigabyte; HIPAA: Health Insurance Portability and Accountability Act; NGS: next-generation sequencing; SRA: Sequence Read Archive; BLAST: Basic Local Alignment Search Tool; BLAT: BLAST-like Alignment Tool; STAR: Spliced Transcripts Alignment to a Reference; BWT: Burrows-Wheeler transform; BWA: Burrows-Wheeler Aligner; TCGA: The Cancer Genome Atlas; NIH: National Institutes of Health; MPI: Message Passing Interface

History from the 50s to NGS

In the 1950s, the contemporaneous development of biopolymer sequencing and the digital computer started a digital revolution in the biosciences. Then in the late 1970s the advent of the personal computer (PC) and Sanger sequencing led to an appreciable amount of sequence data being generated, stored in databases, and conceptualized within a computational framework (1-4). In the 1980s communal sequence databases were developed (5, 6); however, most investigators worked with data of a scale that allowed transfer to and processing on a local client. In the 1990s the rise of the Internet facilitated increased data sharing and analysis techniques began to shift to programs hosted on websites (7). In the mid 2000s the most recent big change occurred with the advent of cloud computing and next generation sequencing (NGS), which led to a dramatic increase in the scale of datasets (see box on increase in sequencing) (4, 8). This necessitated changes in the storage infrastructure; databases such as the European Nucleotide Archive (9) and the Sequence Read Archive (SRA) (10) were created to store and organize high-throughput sequencing data. The SRA has grown significantly since its creation in 2007, now containing almost four petabases, approximately half of which are open access (11). These datasets present a challenge because they are too large for the old sharing and analysis paradigms. However recent innovations in computational technologies and approaches, especially the rise of cloud computing, provide promising avenues for handling the vast amounts of sequence data being generated.

Organizing principles for biocomputing history

In relation to the coevolution of sequencing and computing there are a number of key concepts to keep in mind. First is the idea that scientific research and computing have progressed through a series of discrete paradigms driven by the technology and conceptual frameworks available at the time, a notion popularized by Jim Gray from Microsoft (12). Gray organized his views into four paradigms of scientific research. The first two paradigms are empirical observation and attempts to identify general theories. Gray's third paradigm describes the original type of scientific computing, epitomized by large supercomputer-based calculations and modeling – e.g. computing a rocket trajectory from a set of equations. This approach tends to favor differential equations and linear-algebraic types of computations.

The fourth paradigm is much more data intensive. Here the "capture, curation, and analysis" of large amounts of information fuels scientific research (12). One often tries to find patterns in "big data" and a premium is placed on resource interoperability and statistical pattern finding. In order to fully realize the potential of this approach to science, significant investment must be made in both the computational infrastructure to support data processing and sharing as well as providing training resources for researchers to better understand, handle, and compare large datasets.

Paul Muir 2/8/16 6:35 PM
Formatted: Font color: Auto

The second key concept is the interplay between fixed and variable costs, especially with regard to their impact on the scaling behavior. Much of the decrease in sequencing costs has been a result of a shift between these two cost structures. Next-generation sequencing introduced more efficient and complicated equipment, increasing the fixed cost. However, a reduction of the variable costs of sequencing via lower per sample costs has accompanied this increase in fixed cost and encouraged the sequencing of an ever-greater number of samples in order to reduce the average cost and achieve economies of scale.

Paul Muir 2/8/16 6:35 PM
Formatted: Font color: Auto

The opposite shift in cost structures is beginning to occur in the context of scientific computing. In the past, computing operated under a similar cost structure as seen for sequencing. This often involved a large fixed cost associated with purchasing a machine followed by low variable costs for actual running of the machine (e.g. usually power, cooling, systems administration time). Cloud computing and its associated concepts such as software, platform, and infrastructure as a service removes the need for a large initial fixed cost investment (13). However, the variable costs associated with cloud computing access can be significantly higher. This new regime in which costs scale with the amount of computational processing time places a premium on driving down the average cost by developing efficient algorithms for data processing.

The different cost structure of this new computing paradigm will significantly impact how funding agencies and researchers approach data analysis. Traditionally, in academic settings large computing equipment expenses have been exempt from additional indirect cost fees levied by universities on smaller consumption purchases. Furthermore, running costs for the hardware, such as electricity and cooling required, are supported by the university at little to no cost for the individual investigator (usually from the overall pool of indirect costs). However, in the case of cloud computing time, universities do not consider it an equipment purchase and levy the indirect cost fees on top of the "service" purchase. Additionally, the cloud computing cost often incorporates the additional costs (electricity, rent, etc.) directly into the price. These funding schemes add to the expense of purchasing cloud-computing time compared to large purchases of computing equipment.

BASE.

The cost of sequencing is frequently referred to as a dollar amount per genome. Whether this price includes all steps in the sequencing process (e.g. sample prep, downstream processing, etc.) or merely the sequencing run is often ambiguous. This single price also obscures the cost breakdown of sequencing projects. A more comprehensive approach in which the full economic cost (FEC) of sequencing is evaluated would enable both researchers and funding agencies to better understand and plan such projects. This approach breaks the cost of a sequencing project into its substituent parts and identifies the shared institutional resources used as well as the

indirect costs associated with the project. Such accounting practices would more explicitly call attention to the shift in cost structures described in the previous paragraph and better enable the adaptation of funding mechanisms to meet the changing needs of sequencing enabled research.

A comparison of more detailed cost breakdowns between sequencing projects can also reveal how different components of the sequencing pipeline scale with the size of the project. Figure 2 illustrates the cost breakdown of exome and a whole genome sequencing projects into the cost of labor, reagents and supplies, instrument depreciation and maintenance, administration, basic data processing and initial storage, and indirect fees. Noticeably, reagents and supplies represent a larger fraction of the cost for whole genome sequencing relative to that of exome sequencing, which requires sequencing a smaller number of bases. Labor costs follow the opposite trend, which indicates that labor scales well with the amount of sequence data generated. However, these analyses have one common drawback since they generally include only the cost of basic data processing and initial storage. As bioinformatics becomes increasingly important in the generation of biological insight from sequencing data the long-term storage and analysis of sequencing data will represent a larger fraction of project cost. Efforts to better incorporate detailed and realistic accounting of downstream bioinformatics analysis is essential to the development of accurate models of the full economic cost of sequencing projects.

The third key concept to take into account with these developments is the idea of scaling behavior in sequencing technology and its impact on biological research. The most prominent analogous example of this is Moore's law, which describes the scaling of integrated circuit development and its wide-ranging impact on the computer industry.

Paul Muir 2/8/16 6:35 PM
Formatted: Font color: Auto

Backdrop of the computer industry & Moore's law

Improvements in semiconductor technology have dramatically stimulated the development of integrated circuits during the last half-century. This spurred the development of the personal computer and the Internet era. Various scaling laws, which model and predict the rapid developmental progress in high-tech areas driven by the progress in integrated circuit technology, have been proposed. Moore’s law accurately predicted that the number of transistors in each square inch would double every two years (14). In fact, the integrated circuit industry has used Moore’s law to plan its research and development cycles. Besides Moore’s law, various other predictive laws have also been proposed for related high-tech trends. Rock’s law (also called Moore’s second law) predicted that the fixed cost of constructing an integrated circuit chip fabrication plant doubles about every four years (15). Additionally, Kryder’s law describes the roughly yearly doubling in the area storage density of hard drives over the last few decades (16).

The roughly exponential scaling described by these laws over a period of multiple decades is not simply the scaling behavior of a single technology but rather the superposition of multiple S-curve trajectories. These curves represent the scaling of different technological innovations that contribute to the overall trend (see Fig 1). The S-curve behavior of an individual technology is due to three main phases: development, expansion and maturity (17). For example, the near yearly doubling of hard drive storage density over the last two and a half decades is the superposition of the S-curves for five different basic storage technologies. This behavior is also seen for sequencing-based technologies.

The success of these predictive laws encouraged the development of forecasts for other emergent technologies including sequencing. The cost of sequencing roughly followed a Moore's law trajectory in the decade before 2008. However, the introduction of next generation sequencing technologies caused costs to drop faster than would be expected by Moore's law. Specifically, in the past five years the cost of a personal genome has dropped to $4,200 in 2015 from $340,000 in 2008 (18). This departure from Moore's law indicates that the transition between these technologies introduced a new cost-scaling regime.

Computational component of sequencing - what's happening in bioinformatics

The decreasing cost of sequencing and increasing number of sequence reads being generated are placing greater demand on the computational resources and knowledge necessary to handle sequence data. It is critically important that as the amount of sequencing data continues to increase it is not simply stored but organized in a manner that is both scalable as well as easily and intuitively accessible to the larger research community. We see a number of key directions of change in bioinformatics computing paradigms that are adapting in response to the ever-increasing amounts of sequencing data. The first is the evolution of alignment algorithms in response to larger reference genomes and sequence read datasets. The second involves the need for compression to handle large file sizes - especially the need for compression that takes advantage of domain knowledge more specific to sequencing data to achieve better outcomes than more generic compression algorithms. The third change involves the need for distributed and parallel cloud computing to handle the large amounts of data and integrative analyses. The fourth change is driven by the fact that much of the future sequencing data will be private data related to identifiable individuals; consequently, there is a need to put protocols in place to secure such data particularly within a cloud computing environment.

Innovations underlying scaling in alignment algorithms

Alignment tools co-evolved with sequencing technology to meet the demands placed on sequence data processing. The decrease in their running time approximately follows Moore's Law (see Fig 3). This improved performance is driven by a series of discrete algorithmic advances. In the early Sanger sequencing era, the Smith-Waterman (19) and Needleman-Wunsch (20) algorithms used dynamic programming to find a local or global optimal alignment. But the quadratic complexity of these approaches makes it impossible to map sequences to a large genome. Following this limitation many algorithms with optimized data structures were developed, employing either hash-tables (e.g. Fasta (21), BLAST (22), BLAT (23), MAQ (24), Novoalign (25)) or suffix arrays with the Burrows-Wheeler transform (BWT) (e.g. STAR (26), BWA (27), Bowtie (28)).

Paul Muir 2/8/16 6:35 PM
Deleted: 2

In addition to these optimized data structures, algorithms adopted different search methods to increase efficiency. Unlike Smith-Waterman and Needleman-Wunsch, which compare and align two sequences directly, many tools (e.g. FASTA, BLAST, BLAT, MAQ, STAR) adopt a two-step seed-and-extend strategy. While this strategy cannot be guaranteed to find the optimal alignment, speeds are significantly increased by not comparing sequences base by base. BWA and Bowtie further optimize by only searching for exact matches to a seed (25). The inexact match and extension approach can be converted into an exact match method by enumerating all combinations of mismatches and gaps.

In addition to changing search strategies, algorithms adjusted to larger datasets by first organizing the query, the database, or both. This involves an upfront computational investment but returns increased speed as datasets grow larger. For example, some algorithms (e.g. BLAST, FASTA, MAQ) first build indexes for query sequences before scanning the database. On the database side, some algorithms format the database into compact binary files (e.g. BLAST, MAQ), while others build an offline index (e.g. BLAT, Novoalign, STAR, BWA, Bowtie). In particular, STAR, BWA and Bowtie can significantly reduce the marginal mapping time (i.e. the time it takes to map a single read), but require a relatively large amount of time to build a fixed index. In general, we find a negative correlation between the marginal mapping time (i.e. the time to map a single read) and the time to construct the fixed index making BWA, Bowtie and STAR better suited to handle progressively larger NGS datasets (Fig 3).). However, many of these alignment algorithms are not suitable for longer reads because of the scaling behavior of their seed search strategies. As long-read technologies continue to improve there will be an ever greater need to develop new algorithms capable of delivering similar speed improvements as have been obtained for short read alignment (25).

Paul Muir 2/8/16 6:35 PM
Deleted: 2

Recently, new approaches have been been developed that substitute assembly for mapping. As such these are not directly comparable to the mappers above, but they have significant speed gains in certain contexts and may represent the next technological innovation in alignment. These approaches, including Salmon and Kallisto (29, 30), mostly focus on RNA-seq transcript identification and quantification and employ hashed k-mers and a De Bruijn graph for the task of RNA-Seq quantification. Moreover, instead of developing a base pair resolution alignment these approaches identify a 'pseudoalignment' that consists of the set of transcripts compatible with a given read.

In addition to read alignment the other main computationally intensive algorithmic issue associated with the analysis of sequencing reads is the de novo assembly of a genome sequence. The have been many tools developed for assembly using short read sequencing technology (31, 32). The time and memory requirements of these different algorithms vary significantly (see Fig. 3B). The advent of long read sequencing technologies such as Pacific Biosciences, Oxford Nanopore and Moleculo (33) promise high quality sequence assemblies with potentially reduced computational costs. However higher sequencing error rates for longer reads require novel assembly algorithms (34-37). The main benefit is that one is able to assemble contigs which are 10-100x larger as compared with the contigs from traditional short read technologies even with lower fold coverage (see (38) for a comparison in mammalian genomes).

ONLY TO SOME DEG. DO THEY REL TO GENOME SIZE

Compression

The explosion of sequencing data created a need for efficient methods of storage and transmission. General algorithms like Lempel-Ziv offer great compatibility, good speed and acceptable compression efficiency on sequencing data and are widely used (39). However, to further reduce the storage footprint and transmission time, customized algorithms are needed. For example, many researchers use the SAM/BAM (Sequence/Binary Alignment/Map) format to store reads. A widely accepted compression method, CRAM, is able to shrink BAM files by ~30% without any data loss ("losslessly") and more if one uses compression that loses some

Paul Muir 2/8/16 6:35 PM
Deleted: (31).

information ("lossy"), typically in the quality scores (40). CRAM only records the reference genome and applies Huffman coding to the result. Developing new and better compression algorithms is an active research field and we believe that high compatibility and the balance between usability and compression is key to moving forward.

Cloud computing

Scalable storage, query, and analysis technologies are necessary to handle the increasing amounts of genomic data being generated and stored. Distributed file systems greatly increase the storage I/O bandwidth, making distributed computing and data management possible. An example is the NoSQL database that provides excellent horizontal scalability, data structure flexibility, and support for high load interactive queries (41). Moreover, the parallel programming paradigm has evolved from fine-grained MPI/MP to robust, highly scalable frameworks such as MapReduce (42) and Apache Spark (43). This situation calls for customized paradigms specialized for bioinformatics study. We have already seen some exciting work in this field (44).

These distributed computing and scalable storage technologies naturally culminate in the framework of cloud computing, where data is stored remotely and analysis scripts are then uploaded to the cloud and the analysis is performed remotely. This greatly reduces the data transfer requirements since only the script and analysis results are transferred to and from data that resides permanently in the cloud.

Privacy

In a similar fashion to the way that the Internet gave rise "open source" software; the initial sequencing of the human genome (particularly that from the "public consortium") was associated with "open data." Researchers were encouraged to build upon existing publicly available sequence knowledge and contribute additional sequence data or annotations. However, as more genomes of individuals are sequenced concerns for the privacy of these subjects necessitates securing the data and only providing access to appropriate users (45).

As changing computing paradigms such as cloud computing are playing a role in managing the flood of sequencing data, privacy protection in the cloud environment becomes a major concern (46, 47). Research in this field can broadly be split into two layers: [1] sensitive data must be protected from leaking to a third party (48), and [2] the cloud service provider should be made as oblivious as possible to the computation (49). One possible culmination of these ideas could be the creation of a single, monolithic "biomedical cloud" that would contain all the protected data from genomics research projects. This would completely change the biomedical analysis ecosystem, with researchers gaining access to this single entry point and storing all their programs and analyses there. Smaller implementations of this strategy can be seen in the development of HIPAA compliant cloud resources where datasets can be stored and shared on remote servers (47).

The cost of sequencing and the changing biological research landscape

Paul Muir 2/8/16 6:35 PM
Deleted: (32).

Paul Muir 2/8/16 6:35 PM
Deleted: (33).

Paul Muir 2/8/16 6:35 PM
Deleted: (34)

Paul Muir 2/8/16 6:35 PM
Deleted: (35).

Unknown
Field Code Changed

Paul Muir 2/8/16 6:35 PM
Deleted: 36

Paul Muir 2/8/16 6:35 PM
Deleted: 37

Unknown
Field Code Changed

Paul Muir 2/8/16 6:35 PM
Deleted: (38, 39).

Paul Muir 2/8/16 6:35 PM
Deleted: (40)

Paul Muir 2/8/16 6:35 PM
Deleted: (41).

Paul Muir 2/8/16 6:35 PM
Deleted: (39).

The decrease in the cost of sequencing that has accompanied the introduction of NGS machines and the corresponding increase in the size of sequence databases has changed both the biological research landscape and common research methods. The amount of sequence data generated by the research community exploded over the past ten years. Decreasing costs enabled the formation of large consortia with broad goals (e.g. measuring human genetic variation, profiling cancer genomes), as well as individual labs to target more specific questions. These developments helped democratize and spread sequencing technologies and research, increasing the diversity and specialization of experiments. Using Illumina sequencing alone, nearly 150 different experimental strategies have been described, applying this technology to nucleic acid secondary structure, interactions with proteins, spatial information within a nucleus, and more (50).

Paul Muir 2/8/16 6:35 PM
Deleted: (42).

The changing cost structure of sequencing will significantly impact the social enterprise of genomics and bio-computing. Traditionally research budgets have placed a high premium on data generation. But now with sequencing prices falling rapidly and the size of sequence databases ever expanding, translating this data into biological insights is becoming increasingly important. Consequently, the analysis component of biological research is taking up a larger fraction of the real value in an experiment (8). This of course shifts the focus of scientific work and the credit in collaborations. As a corollary of this, job prospects for scientists with training in computational biology remain strong, despite squeezed budgets (51). Universities, in particular, have increased the number of hires in bioinformatics (see Fig 4).

Paul Muir 2/8/16 6:35 PM
Deleted: (43).

Paul Muir 2/8/16 6:35 PM
Deleted: 3

Paul Muir 2/8/16 6:35 PM
Formatted: Font color: Auto

Moreover, the falling price of sequencing and the growth of sequence databases reduced the cost of obtaining useful sequence information for analysis. Sequence data downloadable from databases is ostensibly free. However, costs arise in the need for computational storage and analysis resources as well as the training necessary to handle and interpret the data. Initial automated processing pipelines for sequence data have lower fixed costs but higher variable costs compared to sequence generation. Variable costs associated with data transfer, storage, and initial pipeline processing using the cloud (e.g. to call variants) all scale with the size of the sequence data being analyzed. In sequence data generation the high initial cost of a sequencing machine is offset by sequencing ever-greater amounts in order to distribute the cost of the initial capital investment over a larger number of sequenced bases. However, this approach merely increases the amount of computational time required for initial pipeline processing. In the context of cloud computing this translates into greater cost since the user is only charged for computational time used. This creates a mismatch, as the combination of costs in sequence data analysis doesn't provide the same economy of scale seen in the generation of sequence data.

There are two possible cost structures for the downstream analysis depending on how bioinformaticians are compensated. Bioinformaticians might be paid on a per project basis (in the extreme, an hourly wage) in which case they resemble the low initial fixed cost and higher variable cost structure of cloud computing. On the other hand, if bioinformaticians are salaried the cost structure of downstream analysis more closely resembles that of sequencing technologies with the salaries representing an initial fixed cost. However, bioinformaticians differ from sequencing machines in that they cannot be consistently replaced by more expensive versions capable of processing more sequencing information. Consequently, driving down the cost of sequence analysis follows a similar path regardless of cost structure. In order to drive down costs, downstream analysis should be made as efficient as possible. This will enable

Paul Muir 2/8/16 6:35 PM
Formatted: Font color: Auto

bioinformaticians to analyze as much sequence data as possible under given time constraints. Generating ever-greater amounts of sequence information will become futile if that data hits a bottleneck during processing and analysis.

This necessitates that many of the big projects in addition to having large amounts of sequencing data pay attention to making analysis and data processing efficient. This can often lead to a framework for large-scale collaboration where much of the analysis and processing of the data is done in a unified fashion. This enables the entire dataset after the fact to be used as a coherent resource without needing reprocessing. If the sequence data generated by individual labs is not processed uniformly and sequence databases are not made easily accessible and searchable, then analysis of aggregated datasets will be challenging. It might seem superficially cheaper to pool the results of many smaller experiments but the reprocessing costs for all of these datasets may be considerably larger than redoing the sequencing experiment itself. In addition to posing technical issues for data storage, the increasing volume of sequences being generated presents a challenge to integrate newly-generated information with the existing knowledge base. Hence, while people thought that the advent of next generation sequencing would democratize sequencing and spur a movement away from the large centers and consortia, in fact the opposite has been the case. The need for uniformity and standardization in very large datasets has, in fact, encouraged very large consortia such as 1000 Genomes (52) and TCGA (53).

In the future, one might like to see a way of encouraging uniformity and standardization without having an explicit consortium structure, letting many people aggregate small sequencing experiments and analyses together. Perhaps this could be done by open community standards in a similar manner to the way the Internet was built through pooling of many individual open source actors using community-based standards (54). It is imperative that such a standardization initiative accompany the development and implementation of new technologies including more efficient data processing and compression algorithms as well as secure cloud computing. A scalable biocomputing infrastructure is vital to a biological research ecosystem capable of integrating vast amounts of heterogeneous sequencing data.

Paul Muir 2/8/16 6:35 PM
Deleted: (44)

Unknown
Field Code Changed

Paul Muir 2/8/16 6:35 PM
Deleted: 45

Paul Muir 2/8/16 6:35 PM
Deleted: (46).

Figure Captions:

Figure 1:

A. The exponential increase in the number of gigabytes per dollar in hard drive storage technology is due in part to the sequential introduction and improvement of three technologies. B. Exponential scaling in technological cost improvement is often the superposition of multiple S-curve trajectories of individual technologies. At the beginning of a technology's life cycle, development costs keep cost reductions low. As the technology matures improvements in production are able to drive down per unit costs and establish an exponential regime. Eventually, the technology reaches maturity where technological limits are encountered and the cost improvements again slow down.

Figure 2:

Cost breakdown of exome and a whole genome sequencing projects. The total cost of these projects is split into the cost of labor, reagents and supplies, instrument depreciation and maintenance, administration, basic data processing and initial storage, and indirect fees.

Figure 3:

Multiple advances in alignment algorithms have contributed to an exponential decrease in running time over the past forty years. We synthesized one million single ended reads of 75 bp for both Human and Yeast. The comparison only considers the data structure, algorithms and speeds. There are many other factors, such as accuracy and sensitivity, which are not discussed here, but can be found elsewhere (25). Initial alignment algorithms based on dynamic programming were applicable to the alignment of individual protein sequences. However, they were too slow for efficient alignment at a genome scale. Advances in indexing helped reduce both running time. Additional improvements in index and scoring structures enabled next generation aligners to further improve alignment time. A negative correlation is also observed between the initial construction of an index and the marginal mapping time per read.

Figure 4:

Paul Muir 2/8/16 6:35 PM
Deleted: 3

The number of faculty position hires at 51 US universities in three-year bins. The recent increase in hiring coincides with the explosion in sequencing data. Data was obtained from (http://jeffhuang.com/computer_science_professors.html).

Box: Illustrations of the dramatic increase in rate and amount of sequencing

A. Next generation sequencing reads have become the dominant form of sequence data. This is illustrated in a graph of NIH funding related to the keywords "Microarray" and "Genome Sequencing", which shows increasing funding for next generation sequencing and decreases in the funding of previous technologies such as microarrays.

B. The size and growth rate of the SRA highlight the importance of efficiently storing sequence data for access by the broader scientific community. The SRA's centrality in the storage of DNA sequences from next-generation platforms means that it also serves as a valuable indicator of the scientific uses of sequencing. Furthermore, the rise in protected sequence data highlights the challenges facing genomics as ever-greater amounts of personally identifiable sequence data are being generated.

C. It is interesting to look at the contribution of large sequence depositions compared to smaller submissions. This provides an indication of the size distribution of sequencing projects. At one end of this size spectrum are large datasets generated through the collaborative effort of many labs. These include projects that have taken advantage of sequencing trends to generate population scale genomic data (1000 Genomes) or extensive characterization of cancer genomes by The Cancer Genome Atlas (TCGA). On top of generating vast amount of sequencing data to better understand human variation and disease, high throughput sequencing has dramatically expanded the number of species whose genomes are documented. The number of newly-sequenced genomes has exhibited an exponential increase in recent years. Entries with asterisks indicate projects that produce open access data.

D. A more detailed analysis of the SRA illustrates the pace at which different disciplines adopted sequencing. Plots depicting the cumulative number of bases deposited in the SRA and linked to papers appearing in different journals provide a proxy for sequencing adoption. More general journals such as Nature and Science show early adoption. Meanwhile, SRA data deposited by articles from more specific journals such as Nature Chemical Biology and Molecular Ecology remained low for a significantly longer time before increasing. These trends highlight the spread of sequencing to new disciplines.

E. Sequence data has also been distributed over the tree of life. In terms of size, the vast majority of sequence data generated has been for eukaryotes. This is due in part to the larger genome size of eukaryotes as well as efforts to sequence multiple individuals within a given species, especially humans. In terms of number of species sequenced prokaryotes are by far the best represented. Moving forward the continued decrease in the cost of sequencing will enable further exploration of the genetic diversity both within and across species.

1. Staden R. Automation of the computer handling of gel reading data produced by the shotgun method of DNA sequencing. Nucleic acids research. 1982;10(15):4731-51.
2. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proceedings of the National Academy of Sciences of the United States of America. 1977;74(12):5463-7.
3. Larson R, Messing J. Apple II computer software for DNA and protein sequence data. DNA. 1983;2(1):31-5.
4. Stevens H. Life out of sequence : a data-driven history of bioinformatics. Chicago: The University of Chicago Press; 2013. 294 pages p.
5. George DG, Barker WC, Hunt LT. The protein identification resource (PIR). Nucleic acids research. 1986;14(1):11-5.
6. Kanehisa MI. Los Alamos sequence analysis package for nucleic acids and proteins. Nucleic acids research. 1982;10(1):183-96.
7. Gouet P, Courcelle E, Stuart DI, Metoz F. ESPript: analysis of multiple sequence alignments in PostScript. Bioinformatics. 1999;15(4):305-8.
8. Sboner A, Mu XJ, Greenbaum D, Auerbach RK, Gerstein MB. The real cost of sequencing: higher than you think! Genome Biol. 2011;12(8):125.
9. Leinonen R, Akhtar R, Birney E, Bower L, Cerdeno-Tarraga A, Cheng Y, et al. The European Nucleotide Archive. Nucleic acids research. 2011;39(Database issue):D28-31.
10. Leinonen R, Sugawara H, Shumway M, International Nucleotide Sequence Database C. The sequence read archive. Nucleic acids research. 2011;39(Database issue):D19-21.
11. Sequence Read Archive : NCBI/NLM/NIH: NIH; 2015 [updated 10/13/2015; cited 2015 10/15/2015]. Available from: http://www.ncbi.nlm.nih.gov/Traces/sra/.
12. Hey AJG, Tansley S, Tolle KM. The Fourth Paradigm: Data-intensive Scientific Discovery: Microsoft Research; 2009.
13. Armbrust M, Fox A, Griffith R, Joseph AD, Katz R, Konwinski A, et al. A view of cloud computing. Commun ACM. 2010;53(4):50-8.
14. Brock DC, Moore GE. Understanding Moore's law : four decades of innovation. Philadelphia, Pa.: Chemical Heritage Foundation; 2006. 122 p. p.

Paul Muir 2/8/16 6:35 PM
Deleted: 1. Staden R. Automation of the computer handling of gel reading data produced by the shotgun method of DNA sequencing. Nucleic acids research. 1982;10(15):4731-51. ... [1]

15. Ross PE. 5 Commandments 2015 [updated 12/1/2003; cited 2015 10/15/2015]. Available from: http://spectrum.ieee.org/semiconductors/materials/5-commandments.
16. Walter C. Kryder's law. Sci Am. 2005;293(2):32-3.
17. Sood A, James GM, Tellis GJ, Zhu J. Predicting the Path of Technological Innovation: SAW vs. Moore, Bass, Gompertz, and Kryder. Market Sci. 2012;31(6):964-79.
18. KA. W. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP) Available at: http://www.genome.gov/sequencingcosts [cited 2015 September 14].
19. Smith TF, Waterman MS. Identification of common molecular subsequences. Journal of molecular biology. 1981;147(1):195-7.
20. Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of molecular biology. 1970;48(3):443-53.
21. Lipman DJ, Pearson WR. Rapid and sensitive protein similarity searches. Science. 1985;227(4693):1435-41.
22. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. Journal of molecular biology. 1990;215(3):403-10.
23. Kent WJ. BLAT--the BLAST-like alignment tool. Genome Res. 2002;12(4):656-64.
24. Li H, Ruan J, Durbin R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res. 2008;18(11):1851-8.
25. Li H, Homer N. A survey of sequence alignment algorithms for next-generation sequencing. Brief Bioinform. 2010;11(5):473-83.
26. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29(1):15-21.
27. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754-60.
28. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009;10(3):R25.
29. Bray N, Pimentel H, Melsted P, Pachter L. Near-optimal RNA-Seq quantification. arXiv:150502710. 2015.
30. Patro R, Duggal G, Kingsford C. Salmon: Accurate, Versatile and Ultrafast Quantification from RNA-seq Data using Lightweight-Alignment. bioRxiv. 2015.
31. Zhang W, Chen J, Yang Y, Tang Y, Shang J, Shen B. A practical comparison of de novo genome assembly software tools for next-generation sequencing technologies. PloS one. 2011;6(3):e17915.
32. Bradnam KR, Fass JN, Alexandrov A, Baranay P, Bechner M, Birol I, et al. Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. Gigascience. 2013;2(1):10.
33. Kuleshov V, Xie D, Chen R, Pushkarev D, Ma Z, Blauwkamp T, et al. Whole-genome haplotyping using long reads and statistical methods. Nature biotechnology. 2014;32(3):261-6.
34. English AC, Richards S, Han Y, Wang M, Vee V, Qu J, et al. Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. PloS one. 2012;7(11):e47768.
35. Koren S, Schatz MC, Walenz BP, Martin J, Howard JT, Ganapathy G, et al. Hybrid error correction and de novo assembly of single-molecule sequencing reads. Nature biotechnology. 2012;30(7):693-700.

36. Chin CS, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, et al. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat Methods. 2013;10(6):563-9.
37. Lee H, Gurtowski J, Yoo S, Marcus S, McCombie WR, Schatz M. Error correction and assembly complexity of single molecule sequencing reads. bioRxiv. 2014.
38. Chaisson MJ, Wilson RK, Eichler EE. Genetic variation and the de novo assembly of human genomes. Nature reviews Genetics. 2015;16(11):627-40.
39. Zhu Z, Zhang Y, Ji Z, He S, Yang X. High-throughput DNA sequence data compression. Brief Bioinform. 2015;16(1):1-15.
40. 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(5):734-40.
41. Cattell R. Scalable SQL and NoSQL data stores. SIGMOD Rec. 2011;39(4):12-27.
42. Dean J, Ghemawat S. MapReduce: simplified data processing on large clusters. Commun ACM. 2008;51(1):107-13.
43. Zaharia M, Chowdhury M, Franklin MJ, Shenker S, Stoica I. Spark: cluster computing with working sets. Proceedings of the 2nd USENIX conference on Hot topics in cloud computing. 2010;10:10.
44. Massie M, Nothaft F, Hartl C, Kozanitis C, Schumacher A, Joseph AD, et al. ADAM: Genomics Formats and Processing Patterns for Cloud Scale Computing. EECS Department, University of California, Berkeley, 2013 December 15. Report No.: UCB/EECS-2013-207.
45. Greenbaum D, Sboner A, Mu XJ, Gerstein M. Genomics and privacy: implications of the new reality of closed data for the field. PLoS Comput Biol. 2011;7(12):e1002278.
46. Greenbaum D, Du J, Gerstein M. Genomic anonymity: have we already lost it? Am J Bioeth. 2008;8(10):71-4.
47. Stein LD, Knoppers BM, Campbell P, Getz G, Korbel JO. Data analysis: Create a cloud commons. Nature. 2015;523(7559):149-51.
48. Popa RA, Redfield CMS, Zeldovich N, Balakrishnan H. CryptDB: protecting confidentiality with encrypted query processing. Proceedings of the Twenty-Third ACM Symposium on Operating Systems Principles; Cascais, Portugal. 2043566: ACM; 2011. p. 85-100.
49. Maas M, Love E, Stefanov E, Tiwari M, Shi E, Asanovic K, et al. PHANTOM: practical oblivious computation in a secure processor. Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security; Berlin, Germany. 2516692: ACM; 2013. p. 311-24.
50. Sequencing Library Preparation Methods: Illumina; 2015 [cited 2015 10/15/2015]. Available from: http://www.illumina.com/techniques/sequencing/ngs-library-prep/library-prep-methods.html.
51. Levine AG. An Explosion of Bioinformatics Careers. Science. 2014;344(6189):1303-4.
52. Genomes Project C, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. A global reference for human genetic variation. Nature. 2015;526(7571):68-74.
53. Cancer Genome Atlas Research N, Weinstein JN, Collisson EA, Mills GB, Shaw KR, Ozenberger BA, et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet. 2013;45(10):1113-20.

54. Isaacson W. The innovators : how a group of hackers, geniuses, and geeks created the digital revolution. First Simon & Schuster hardcover edition. ed. New York: Simon & Schuster; 2014. viii, 542 pages p.