The real cost of sequencing: computational analysis

History from the 50s to NGS:

The contemporaneous development of biopolymer sequencing and the digital computer in the 1950s started a digital revolution in the biosciences. Adoption was slow at first; some historians of science, such as Hallam Stevens, have argued that the lack of computers in biology was partially due to the incompatibility of computational approaches and biological research [1]. The data generated by biological experiments was often not in a form that benefited from computational processing power. Early biopolymer sequencing in 60s & 70s started to shift the nature of biological data toward more quantifiable and computationally tractable sequence data. However, computational approaches were not often brought to bear on that much due to its relative infancy and the limited computational resource available at the time. This changed with the advent of the personal computer and Sanger sequencing in the late 1970's leading to the generation of ever-greater amounts of sequence data. Large amounts of sequence data could then be generated and stored in databases and conceptualized within a computational framework. As the computational and biological sciences have developed together they have spurred and reacted to innovations in each other.

The computing technologies used in the analysis of sequence data have helped shape how researchers approach such analysis and the structure of biological research more generally. The PC era, in which Sanger DNA sequencing developed, left its imprint on how sequence data is analyzed. In the 1980's, sequence databases were developed and filled with ever greater amounts of sequence. However, most of the data relevant to an investigator could be transferred to and processed on a local client. The rise of the internet encouraged sharing of sequence data and enabled new bioinformatics approaches in which analysis programs could be hosted on websites onto which data would then be uploaded and analyzed. These conditions coupled with the increasing availability of reference genomes for various species including humans created an ecosystem in which researchers could better query the existing sequencing knowledge base and situate their work within it [1].

The next big change occurred in the mid 2000s with the advent of cloud computing and next generation sequencing (NGS), which led to a dramatic increase in the scale of sequence datasets (see box on increase in sequencing) [CITE]. This necessitated changes in the sequence data storage infrastructure. Databases such as the European Nucleotide Archive and the Sequence Read Archive (SRA) were created to store and organize high throughput sequencing data generated for research purposes. The SRA has grown significantly since its creation in 2007. It now contains $3.9*10^{15}$ bases with approximately half of these being open access [CITE]. These datasets present a challenge as they are too large for the old sharing and analysis paradigms. However recent innovations in computational technologies and approaches, especially the rise of distributed and cloud computing, provide promising avenues for handling the vast amounts of sequence data being generated and stored in databases.

Key concepts to the interpret the 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 paradigms driven by the technology and conceptual frameworks available at the time. This has been popularized by eminent database researchers such as Jim Gray from Microsoft. In this view, empirical observation and attempts to identify general theories are seen

Deleted: ,

Deleted: compliant

Deleted: this data was

Deleted: computed

Deleted: this data

Deleted: be

Deleted: larger

Deleted: Key components of

Deleted: are databases

Deleted:), which

Deleted: ,

Deleted: which

Deleted:

Deleted: viewed

as the first two paradigms of scientific research. Gray's third paradigm describes traditional scientific supercomputing based on large calculations and modeling [CITE The fourth paradigm]. For instance, computing a rocket trajectory from a set of equations. This approach tends to favor differential equations and linear algebraic types of computations.

The fourth or new paradigm is much more data intensive. In this paradigm, scientific research is fueled by the "capture, curation, and analysis" of large amounts of information [CITE]. Improvements in computing and data collection methods have helped usher in this big data era. In the past one might have worked on simulating large amounts of mathematical calculations. Now, one is often trying to find patterns in very large datasets and here a premium is placed on data 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.

The second key concept is the interplay between fixed and variable costs. Much of the decrease in sequencing costs has been a result of trading the higher variable cost of reagents and sequencing technicians' time for larger fixed costs in terms of ever more efficient and complicated equipment. The large initial cost of a sequencing machine followed by low per sample costs has encouraged the sequencing of an ever greater number of samples. A different paradigm shift plays out in the context of scientific computing. In the past, computing often involved a large fixed cost associated with purchasing a machine followed by low variable costs. Cloud computing removes the need for a large initial fixed cost investment. However, the variable costs associated with cloud computing access are significantly higher. The different cost structure of these new computing paradigms can have a significant impact on how funding agencies and researchers approach data analysis. For example, budgeting for equipment and computational analysis will need to take the computing as a service nature of cloud computing into consideration.

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 that has had a wide-ranging impact on the computer industry.

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 has 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 that are driven by the progress in integrated circuit technology, have been proposed. Moore's law accurately predicted that the number of transistors integrated in each square inch would double every two years [CITE]. 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 [CITE] (http://spectrum.ieee.org/semiconductors/materials/5-commandments/2) Kryder's law describes the roughly yearly doubling in the area storage density of hard drives over the last few decades. Additionally, 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.

Deleted: e

Deleted: the

Deleted: much more

Deleted: ... [1]

Formatted: None, Space Before: 0 pt, After: 0 pt

Deleted: ... [2]

Formatted: Font:12 pt, Not Bold

Deleted: led to

Deleted: For instance, the well-known

Deleted: developments

Deleted: For instance

Deleted: s

Deleted: a

Deleted: r

Deleted: nd

Deleted: Similarly, Kryder's law describes the roughly yearly doubling in the area storage density of hard drives over the last few decades.

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 representing the scaling behavior of different technological innovations that contribute to the overall trend (see figure 1). The S-curve behavior of an individual technology is due to the three main phases (development, expansion and maturity) (ref Sood 2012). For example, the near yearly doubling scaling behavior of hard drive storage density over the last two and a half decades is the superposition of the S-curves for five different storage technologies. This behavior is also can also be seen for sequencing based technologies.

The success of predictive laws applied to various technologies in last half century has encouraged the development of laws to forecast trends in other emergent technologies including sequencing-based technologies. The cost of sequencing roughly followed a Moore's law trajectory in the decade before 2008 [CITE NIH cost-seq figure]. However, since then the cost of sequencing has deviated from this path, dropping faster than would be expected using Moore's law as a guide after the introduction of high-throughput next generation sequencing technologies NIH cost-seq figure [CITE]. In the past five years, the cost of a personal genome has dropped to $4,200 in 2015 from $340,000 in 2008 [CITE]. This departure from Moore's law reveals that the transition between these technologies introduced a new cost-scaling regime. Consequently, we think that the development of sequencing technology at this stage is far away from following a predictive trajectory.

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 done so 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 next change involves the need for more distributed and parallel cloud computing to handle the large amounts of data and integrative analysis. The fourth change is driven by the fact that much of the future sequencing data will be private data related to identifiable individuals and consequently the need to put protocols in place to secure such data particularly within a cloud computing environment.

Innovations underlying scaling in alignment algorithms:

Alignment tools have 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 figure 2). Underlying this improved performance is a series of discrete algorithmic advances. In the early Sanger sequencing age, the Smith-Waterman (SW) and Needleman-Wunsch (NW) 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. Many algorithms with optimized data structures were developed to resolve this problem. Fasta, BLAST, BLAT, MAQ and Novoalign utilize hash-tables to make large scale sequence alignment more time-efficient. STAR, BWA and Bowtie employ suffix arrays and the

Deleted: of
Deleted: true
Deleted: ve
Deleted:
Deleted: XXX
Deleted: 4
Deleted: XXXX
Deleted: amount
Deleted: s
Deleted: sequencing
Deleted: .
Deleted: E
Deleted: one
Deleted: re
Deleted: be
Deleted: demand of
Deleted: s

Burrows-Wheeler transform (BWT) to further advance ultrafast alignment. Unlike Smith-Waterman and Needleman-Wunsch, which compare two sequences directly, quite a few tools, including FASTA, BLAST, BLAT, MAQ and STAR adopt a two-step seed-and-extend strategy. They are not guaranteed to find the optimal alignment but can significantly speed up sequence alignment because they need not compare the query and target sequences base by base. BWA and Bowtie further optimize alignment by only searching for exact matches to the seed [2]. 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 algorithm improvement, database formatting and sequence indexing are widely utlized. BLAST and MAQ first format the sequence database into binary files. FASTA, BLAST and MAQ build online or offline indexes for query sequences each time and then scan the target sequences. However, BLAT, Novoalign, STAR, BWA and Bowtie only need to build the index offline once for the target databases and are then ready for batch queries. In particular, STAR, BWA and Bowtie can significantly reduce the marginal mapping cost but require a relatively large amount of time to build a fixed index. In general, we can find a negative trend between marginal mapping cost and the fixed index time (figure 2). Decreasing the marginal alignment cost by reasonably increasing the fixed index time makes BWA, Bowtie and STAR better suited to handle progressively larger NGS datasets. However, many of these short read alignment algorithms are not suitable for long reads. As long read technologies continue to improve there will be an ever greater need to develop new algorithms capable of delivering similar speed improvements seen in short read alignment to long read alignment.

Compression:

The explosion of sequencing data has created a need for efficient methods of storage and transmission. General algorithms like Lempel-Ziv offer great compatibility, good compression speed and acceptable compression efficiency on sequencing data and are widely used. However, to further reduce the storage footprint and transmission time, customized algorithms are needed. 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% losslessly and more if lossy on quality score [3]. CRAM only records the differences between reads and the reference genome and applies Huffman coding. Developing new and better compression algorithms is an active research field. We believe high compatibility and the balance between usability and compression ratio are the keys for compression methods moving forward. With the latter depending heavily on specific research purposes, there is perhaps no one-size-fit-all algorithm. Besides compression, there is also work on data representation formats to improve scalability in parallel computation and achieve better compatibility by defining an explicit data schema [CITE Massie: EECS-2013-207].

Parallel and distributed 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 which provides excellent horizontal scalability, data structure flexibility, and support for high load interactive queries [CITE]. Current bioinformatics research heavily uses statistical learning algorithms, user defined functions and semi-structured data. Moreover, the parallel programming paradigm has evolved from fine-grained MPI/MP to robust, highly scalable frameworks such as MapReduce and Apache Spark [CITE]. This situation calls

Deleted: [[SKL: or “because they don’t need to do a full alignment”]].

Deleted: s

Deleted: [[SKL: that are defined as offline index or database binary formatting time in our analysis]]

Deleted: increasing

Deleted: more pressing

Deleted: excellent

Deleted: .

Deleted: For example, distributed

Deleted:

Deleted: today

for customized paradigms specialized for bioinformatics study. We have already seen some exciting work in this field [CITE ADAM from AMP Berkeley].

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 the cloud.

Privacy:

In a similar fashion to the way that the Internet gave rise "open source" software, the human reference 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, now there is a change as more individual genomes are sequenced and concerns for the privacy of the sequenced subjects necessitates securing the data and only providing access to authenticated users [4].

As changing computing paradigms such as cloud computing are playing a role in managing the flood of sequencing data privacy protection in a cloud environment becomes a major concern. Researchers are interested in finding reliable and affordable solutions to minimize the risk of sensitive data leakage. Privacy protection in a cloud environment can be split into two layers: first sensitive data must be protected from leaking to a third party. Second, the computation should be made oblivious to the cloud service provider following methods such as homomorphic encryption [5, 6]. One possible culmination of these ideas could be the creation of a single, monolithic "biomedical-cloud" that would contain all the protected data from US or perhaps even global bioinformatics research projects. This would completely change the biomedical analysis ecosystem, with researchers simply gaining access to this single entry point and storing all their programs & analyses there. Smaller implementations of this strategy can be seen in the HIPAA compliant cloud resources being developed so that datasets can be stored and shared on remote servers [5].

The cost of sequencing and the changing biological research landscape:

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 has exploded over the past ten years. In some cases, the decreasing cost has enabled ambitious large-scale projects aimed at measuring human variation in large cohorts and profiling cancer genomes. On the other hand, as sequencing has become less expensive it has become easier for individual labs with smaller budgets to undertake sequencing projects. These developments have 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, yielding information about nucleic acid secondary structure, interactions with proteins, spatial information within a nucleus, and more [CITE] (ref poster "For all your Seq needs). Perhaps unsurprisingly, the market continues to expect growth from Illumina; their stock valuation outperforms other small-cap biotech, as well as similarly sized companies from other sectors (see figure 4).

Formatted: Font:(Default) Times New Roman, 12 pt

Deleted: This may represent an extension of Jim Gray's 4th paradigm in which data integration is better achieved through reliance upon large cloud-based aggregation of data.

Deleted: sequence

Deleted: [[CITE…some interesting work includes (limited) computation and query directly on encrypted database, isolating encrypted data etc.]].

The growth of sequence databases has 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. The analysis of sequence data has lower fixed costs but higher variable costs compared to sequence generation. Variable costs associated with data transfer, storage, and processing all scale with the amount of sequence data being analyzed. The combination of costs in sequence data analysis doesn't provide the same economy of scale seen in the generation of sequence data.

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, increased importance is being placed on translating this data into biological insights. Consequently, the analysis component of biological research is taking up a larger fraction of the real value in an experiment. This of course shifts the focus of scientific work and the credit in collaborations. In an era of squeezed budgets and fierce competition, job prospects for scientists with training in computational biology remain strong [CITE Explosion of Bioinformatics Careers Science 2014]. Universities have increased the number of hires in the areas of computer science, and specifically in bioinformatics (see figure 4).

Deleted: as a goal of great value.

Deleted: credit and collaboration and the

Deleted: . Furthermore, the cost structures associated with analysis are very different. However, in

Bioinformaticians fundamentally operate on a different cost structure than sequencing machines being essentially fixed costs with little variable nature with respect to projects. 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 collaborative fashion. This enables the entire dataset after the fact to be used as a coherent and consistent 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 integrated datasets will become increasingly challenging. It might seem superficially cheaper to pool and aggregate the results of many smaller experiments but the reprocessing costs of aggregating 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 machines would democratize sequencing and spur a movement away from the large 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 consortiums such as 1000 Genomes and TCGA.

Deleted: a very

Deleted: relative

Deleted: is

In the future, one might like to see a way of encouraging this 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.

Deleted: spool

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

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.

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 dramatic rise in private sequence data highlights the challenges facing genomics as ever-greater amounts of personally identifiable sequence data are being generated.

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 by 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 dramatically increasing. These trends highlight the spread of sequencing to new disciplines.

Additionally, 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 are documented. The number of newly sequenced genomes has exhibited an exponential increase in recent years.

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. Stevens, H., *Life out of sequence : a data-driven history of bioinformatics*. 2013, Chicago: The University of Chicago Press. 294 pages.
2. Li, H. and N. Homer, *A survey of sequence alignment algorithms for next-generation sequencing.* Brief Bioinform, 2010. **11**(5): p. 473-83.
3. Hsi-Yang Fritz, M., et al., *Efficient storage of high throughput DNA sequencing data using reference-based compression.* Genome Res, 2011. **21**(5): p. 734-40.
4. Greenbaum, D., et al., *Genomics and privacy: implications of the new reality of closed data for the field.* PLoS Comput Biol, 2011. **7**(12): p. e1002278.
5. Stein, L.D., et al., *Data analysis: Create a cloud commons.* Nature, 2015. **523**(7559): p. 149-51.

6. Greenbaum, D., J. Du, and M. Gerstein, *Genomic anonymity: have we already lost it?* Am J Bioeth, 2008. **8**(10): p. 71-4.

Page 2: [1] Deleted	Muir, Paul	9/27/15 6:56 PM

Fixed and variable costs in sequencing:

Page 2: [2] Deleted	Muir, Paul	9/27/15 6:56 PM

Backdrop of the computer industry & Moore's law: