The real cost of sequencing: computing

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. Some historians of science have argued that the lack of computers in biology was partially due to the incompatibility of computational approaches and biological research (cite Hallam Stevens "Life Out of Sequence"). 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 change the nature of biological data toward more quantifiable and computationally manipulatable sequence data. However, it was not computed on that much due to its relative infancy and the limited computational resource available at the time. However, 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 be stored in databases and conceptualized in 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 larger amounts of sequence. However, most of the data relevant to an investigator could be downloaded 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 and data would then be uploaded onto these sites for analysis. These conditions coupled with the increasing availability of reference genomes for various species including humans created an environment in which researchers could better query the existing sequencing knowledge base and situate their work within it.

The next big change occurred in the mid 2000s when the advent of next generation sequencing (NGS) led to a dramatic increase in the scale of sequence datasets (see box on increase in sequencing). A key component of the sequence data infrastructure is the sequence read archive (SRA), which was created to store and organize high throughput sequencing data generated for research purposes. The database has grown significantly since its creation in 2007. It now contains $3.9*10^{15}$ bases with approximately half of these being open access. These datasets are too large for the old sharing and analysis paradigms. However, the development of NGS has coincided with the rise distributed and cloud computing which provide promising avenues for handling the vast amounts of sequence data being generated and stored in databases.

Key concepts to the interpret the history:

Deleted: : higher than you think!

Formatted: Font:11 pt, Not Bold

Formatted: Font:11 pt, Not Bold

Formatted: Font:11 pt, Not Bold

Formatted: Font:11 pt, Not Bold

Formatted: Font:11 pt, Not Bold

Deleted: processing, storage & data transfer

Formatted: Normal, Centered

Deleted: Introduction: ... [1]

Deleted: [[bases & bits]]

Deleted: .

Deleted: However, this changed with the advent of Sanger sequencing and

Deleted: computational

Deleted: c

Moved down [1]: 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.

Deleted: However, now there's a change as more individual genomes are sequenced and concerns for the privacy of the sequenced subjects necessitates securing the sequence data and only providing access to authenticated users. [[plos cb article] ... [2]

Moved down [2]: This 4th paradigm holds the possibility of synthesizing the previous paradigms of empirical observation, theory, and computational simulation. However, 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.

Deleted: The advent of next generation sequencing (NGS) has

Deleted: approximately

In relation to 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 viewed as the first two paradigms of scientific research. Gray's third paradigm describes traditional scientific supercomputing based on large calculations and modelling. For instance computing a rocket trajectory from a set of equations. This 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 information. Computing is now confronting the 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 the premium is much more on data interoperability and statistical pattern finding. This 4th paradigm holds the possibility of synthesizing the previous paradigms of empirical observation, theory, and computational simulation. However, 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. 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 an impact on how funding agencies and researchers approach data analysis.

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 semiconductor development which 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 for more than the last half century, which has led to 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 semiconductor technology, have been proposed. For instance, the well-known Moore's law accurately

Formatted: Space Before: 0 pt

Moved (insertion) [2]

Deleted: However, this combination of technologies also presents new challenges. Distributed computing systems for storing and sharing this data must also account for the protected nature of some of these datasets. Additionally, the different cost structure of these new computing paradigms can have an impact on how funding agencies and researchers approach data analysis.

Formatted: Space Before: 0 pt

Deleted: These changes democratize access to high performance computing but also place a premium on highly efficient analysis methods. These two technologies are increasingly intertwined and have a significant impact on both the scale, scope, and methods of biological research.

Deleted: .

Formatted: Font:11 pt

Formatted: Font:11 pt, Not Bold

Formatted: Space Before: 18 pt, After: 4 pt

Deleted: . ... [3]

Deleted: s

Deleted: People have made observations of various

Deleted: es

Deleted: these

predicted that the number of transistors integrated in each square inch would double every two years \cite{}. In fact, the semiconductor 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 developments (http://spectrum.ieee.org/semiconductors/materials/5-commandments/2) For instance, from an economic point of view, Rock's law (also called Moore's second law) predicts that the cost of a semiconductor chip fabrication plant doubles around every four years. 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 the superposition of multiple S-curve graphs 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). 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 of five different storage technologies. This behavior is also true for sequencing based technologies.

The success of predictive laws in high tech fields in last half century have encouraged the development of laws to forecast trends in other emergent technologies including sequencing based technologies. The cost of sequencing did roughly follow a Moore's law behavior in the decade before 2008 \cite{NIH cost-seq figure}. However, the cost of sequencing has not followed a Moore's like law since 2008 after the introduction of new high throughput sequencing technologies \cite{NIH cost-seq figure}. Instead, it has dropped faster than would be expected using Moore's law as a guide. In the past five years, the cost of a personal genome has dropped to XXX in 2014 from XXXX in 2008. This departure from Moore's law is due to the dramatically different slopes for the s-curves representing Sanger sequencing and NGS. Consequently, transition between these technologies represented a new cost scaling regime. Thus, 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 amount of sequence reads being generated are placing greater demands 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 terms of how bioinformatics computing paradigms 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 sequencing datasets. The second involves the need for compression to handle large file sizes and especially the need for compression that takes advantage of domain knowledge more specific to sequencing data to achieve better

Deleted: The…semiconductor industry has used ... [4]
Formatted ... [5]
Deleted: was proposed to predict…the cost of a ... [6]
Deleted: yearly doubling…scaling of these …esc ... [7]
Deleted: the …redictive laws in high tech ... [8]
Moved down [3]: ... [9]
Deleted: [[SKL2MG: I didn't use 4 eras, for I found it is difficult to isolate MAQ and novoalign from BWA etc. And it is too strong if we claim NGS tools didn't use SW at all.]]
Moved down [4]: Alignment tools have co-evolved with sequencing technology to meet demand of sequence data processing.
Formatted: Font:(Default) Trebuchet MS, 13 pt, Bold
Deleted: In light of this computational time bottleneck and increasing dataset sizes, hash table based methods that use a seed-and-extend paradigm with a word of length k(k-mer) as the seed were developed to drive down alignment time. The original FASTA approach simply combines the K-mer to find the common ones between query and target sequences. However, it cannot make sure best alignments are seeded. To improve, BLAST adopts a heuristic statistical method to find high-scoring segment pairs (HSPs) by using substitution matrix and k-letter word, which can perform over 50 times faster than Smith-waterman algorithm. Not like BLAST hashing the query sequence and scanning it against sequence database, BLAT builds a k-mer index for the genome and scans against query sequence and is able to achieve run times 500 times faster than BLAST. ... [11]
Formatted: Font:(Default) Trebuchet MS, 13 pt
Formatted: Font:(Default) Trebuchet MS, 13 pt, Bold
Deleted: The
Moved down [5]: running time fulfills Moore's Law and decreases by half every 18 months (see figure 2). Underlying this improved performance are a series of discrete algorithmic advances.
Moved down [6]: But the quadratic complexity of these approaches make it impossible to map sequences to a large genome.
Formatted: Font:(Default) Trebuchet MS, 13 pt, Bold
Formatted: Space Before: 18 pt, After: 4 pt
Deleted: In the early Sanger sequencing age, the Smith-Waterman and Needleman-Wunsch algor ... [10]
Formatted ... [12]
Formatted: Justified
Deleted: easily and intuitively accessible to the larger research community.
Moved down [7]: Scalable storage, query and analysis technologies are necessary to handle the increa ... [13]

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

Innovations underlying scaling in alignment algorithms:
[[MG:

suffix array & derived structures

BWA, bowtie & STAR

* exact match v inexact - sw, fasta v bwa

hash match - bwa
extend from seed

one progression

generation 1 - sw, nw

gen 2 - fasta, blast, blat, maq - index & hashes
gen 3 - suffix array - bwa, bowtie, star

another progression - index larger things

fasta & blast - query or query + 1
blat, maq, nov... bwa, bowtie - index the db

3rd progression - optimal alignement, seed & extend & just seed
optimal alignment - sw
seed & extend - blast, fasta (blat??)
seed - bwa
]] [[SKL: rewrite done]]

Alignment tools have co-evolved with sequencing technology to meet demand of sequence data processing. Their running time fulfills Moore's Law and decreases by half every 18 months (see figure 2). Underlying this improved performance are a series of discrete algorithmic advances. In the early Sanger sequencing age, the Smith-Waterma(SW) and Needleman-Wunsch(NW) algorithms used dynamic programming to find a local or global optimal alignment. But the quadratic complexity of these approaches make it impossible to map sequences to a large genome. Many algorithms with optimized data structure are developed to resolve the problem. Fasta, BLAST, BLAT, MAQ and Novoalign utilize hash-table to make large scale sequence alignment time-efficiently. And STAR, BWA and Bowtie employ suffix array and BWT to further

Moved (insertion) [3]
Formatted: Font:11 pt, Bold
Formatted: Font:11 pt
Formatted: Normal

Moved (insertion) [4]
Moved (insertion) [5]
Moved (insertion) [6]
Deleted: ... [14]
Moved down [8]: 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.
Deleted:
Moved down [9]: ... [15]
Deleted: privacy protection in the cloud environment becomes a huge concern. Researchers are interested in finding reliable and affordable solution to minimize the risk of sensitive data leakage. Privacy protection in cloud environment can be split into two layers: a. protect sensitive data from leaking to a third party [[cite…some interesting work includes (limited) computation and query directly on encrypted database, isolating encrypted data etc.]]; b. make the computation oblivious to the cloud service provider [[cite…]]. ... [16]
Moved down [10]: heavily uses statistical learning algorithms, user defined functions and semi-structured data. Moreover, today the parallel programming paradigm has evolved from fine-grained MPI/MP to robust, highly scalable frameworks such as MapReduce and Apache Spark. This situation calls for customized paradigms specialized for bioinformatics study. We have already seen some exciting work in this field (cite ADAM from AMP Berkeley)

advance alignment ultrafastly. Unlike SW and NW that compare two sequences directly, quite a few tools, such as FASTA, BLAST, BLAT, MAQ and STAR adopt a two-step seed-and-extension strategy. They cannot guarantee to find optimal alignment but can significantly speed up sequence alignment because they need not compare query and target sequences base by base [[SKL: or “because they don’t need to do a full alignment”]]. BWA and Bowtie further optimize alignment to only keep an exact seed match step. Inexact match and extension step are converted into exact match by enumerating all combinations of mismatches and gaps.

Besides algorithm improvement, database formatting and sequence index are widely used. BLAST and MAQ firstly need to format sequence database into binary files. FASTA, BLAST and MAQ build online or offline index for query sequences each time and then scan to the target sequences. However, BLAT, Novoalign, STAR, BWA and Bowtie only need to build index offline once for the target databases and ready for batch queries. In particular, STAR, BWA and Bowtie can significantly reduce the marginal mapping cost in company with a relative high fixed index time[[SKL: that are defined as offline index or database binary formatting time in our analysis]]. In general, we can find a negative trend between marginal mapping cost with the fixed index time (figure 2). Decreasing “marginal” alignment cost by reasonably increasing “fixed” index time makes BWA, Bowtie and STAR more suitable to handle progressively increasing NGS data.

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 thus widely used. However, to further reduce the storage footprint and transmission time, customized algorithms are needed. Many researchers use SAM/BAM (Sequence/Binary Alignment/Map) format to store reads. A widely accepted compression method, CRAM, is able to shrink BAM file by ~30% losslessly and more if lossy on quality score (\cite 21245279). 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 excellent compatibility and the balance between usability and compression ratio are the keys for compression methods. 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 format to improve scalability in parallel computation and achieve better compatibility by defining an explicit data schema (\cite Massie: EECS-2013-207).

Parallel and distributed computing:

Scalable storage, query and analysis technologies are necessary to handle the increasing amounts of genomic data being generated and stored. For example, distributed file system greatly increases the storage I/O bandwidth, making distributed computing and data management possible. Another example is the NoSQL database which provides excellent horizontal scalability, data structure flexibility, and support for high load interactive queries.

Formatted: Font:Times New Roman, 14 pt

Deleted: posed

Deleted: of

Deleted: for

Deleted: gzip

Deleted: An extensively

Moved (insertion) [7]

Current bioinformatics research heavily uses statistical learning algorithms, user defined functions and semi-structured data. Moreover, today the parallel programming paradigm has evolved from fine-grained MPI/MP to robust, highly scalable frameworks such as MapReduce and Apache Spark. This situation calls for customized paradigms specialized for bioinformatics study. We have already seen some exciting work in this field (cite ADAM from AMP Berkeley)

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 sequence data and only providing access to authenticated users. [[plos cb article]]

However, 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 huge 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 [[cite…some interesting work includes (limited) computation and query directly on encrypted database, isolating encrypted data etc.]]. Second, the computation should be made oblivious to the cloud service provider. [[cite nature jk, Lincoln stein, gady getz, peter campell … DG&MG, Am J Bioethics]] One possible culmination of these thought processes could be to create 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 ecosystem of biomedicine, 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. 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. 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. [[paradigm 4.5]]

The cost of sequencing and the changing biological landscape:

The decrease in the cost of sequencing that has accompanied the introduction of new NGS machines and the corresponding increase in the size of sequence databases has changed both the biological research landscape and the common modes of research. The amount of sequence data generated by the research community has exploded over the past ten years. This data has come from a variety of sources. 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

Moved (insertion) [10]

Moved (insertion) [1]

Moved (insertion) [9]

Moved (insertion) [8]

Formatted: Font:Times New Roman, 14 pt

Formatted: Font:11 pt, Bold

Formatted: Font:11 pt

Formatted: Normal, Space Before: 0 pt

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 (ref poster “For all your Seq needs) yielding information about nucleic acid secondary structure, interactions with proteins, spatial information within a nucleus, and more. 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).

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. Meanwhile, the training and salary of bioinformatics analysts is a key fixed cost in sequence analysis. 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 as a goal of great value. 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 credit and collaboration and the focus of scientific work. Furthermore, the cost structures associated with analysis are very different. However, 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).

Bioinformaticians fundamentally operate on a different cost structure than sequencing machines being essentially fixed costs with a very little variable nature relative to projects. This of course 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 of having a large-scale collaboration where much of the analysis and processing of the data is done in a unified collaborative fashion meaning that the entire dataset after the fact can be used as a coherent and consistent resource without having to reprocess. 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 much cheaper to pool and aggregate the results of many smaller experiments but the reprocessing costs of aggregating all of these

Formatted: Space Before: 10 pt

Formatted: Font:Times New Roman, 14 pt

Moved down [11]: 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. In addition to posing

Moved down [12]: 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.

Deleted: These trends also run the risk of fragmenting the genomics research community.

Deleted: ... [17]

Deleted:

Formatted: Font:Times New Roman, 14 pt

Moved (insertion) [11]

datasets is actually 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 is 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.

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 spool small sequencing experiments and analyses together. Perhaps this could be done by open community standards in a very similar way to the way the internet was built through pooling of many individual open source actors using community-based standards.

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

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 (see figure 3). These trends highlight the spread of sequencing to new disciplines.

A graph of NIH funding related to the keywords "Microarray" and "Genome Sequencing" show the increasing prominence of next generation sequencing at the expense of previous technologies such as microarrays.

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.

Moved (insertion) [12]

Formatted: Font:Bold

Deleted:

Deleted: Cell

Deleted:

Deleted:

Page 1: [1] Deleted	Muir, Paul	9/14/15 4:09 AM

Introduction:

Introduction:

Page 1: [2] Deleted	Muir, Paul	9/14/15 4:09 AM

However, now there's a change as more individual genomes are sequenced and concerns for the privacy of the sequenced subjects necessitates securing the sequence data and only providing access to authenticated users. [[plos cb article]
Microsoft researcher Jim Gray argued that the use of computers to process large volumes is leading to a "fourth paradigm" in scientific research in which discovery is fueled by the "capture, curation, and analysis" of information.

Page 2: [3] Deleted	Muir, Paul	9/14/15 4:09 AM

Semiconductor

Page 3: [4] Deleted	Muir, Paul	9/14/15 4:09 AM

The

Page 3: [4] Deleted	Muir, Paul	9/14/15 4:09 AM

The

Page 3: [4] Deleted	Muir, Paul	9/14/15 4:09 AM

The

Page 3: [4] Deleted	Muir, Paul	9/14/15 4:09 AM

The

Page 3: [5] Formatted	Muir, Paul	9/14/15 4:09 AM

Font color: Black

Page 3: [5] Formatted	Muir, Paul	9/14/15 4:09 AM

Font color: Black

Page 3: [6] Deleted	Muir, Paul	9/14/15 4:09 AM

was proposed to predict

Page 3: [6] Deleted	Muir, Paul	9/14/15 4:09 AM

was proposed to predict

Page 3: [6] Deleted	Muir, Paul	9/14/15 4:09 AM

was proposed to predict

Page 3: [7] Deleted	Muir, Paul	9/14/15 4:09 AM

yearly doubling

Page 3: [7] Deleted	Muir, Paul	9/14/15 4:09 AM

yearly doubling

Page 3: [7] Deleted	Muir, Paul	9/14/15 4:09 AM

yearly doubling

Page 3: [7] Deleted	Muir, Paul	9/14/15 4:09 AM

yearly doubling

Page 3: [7] Deleted	Muir, Paul	9/14/15 4:09 AM

yearly doubling

Page 3: [7] Deleted	Muir, Paul	9/14/15 4:09 AM

yearly doubling

Page 3: [7] Deleted	Muir, Paul	9/14/15 4:09 AM

yearly doubling

Page 3: [7] Deleted	Muir, Paul	9/14/15 4:09 AM

yearly doubling

Page 3: [7] Deleted	Muir, Paul	9/14/15 4:09 AM

yearly doubling

Page 3: [8] Deleted	Muir, Paul	9/14/15 4:09 AM

the

Page 3: [8] Deleted	Muir, Paul	9/14/15 4:09 AM

the

Page 3: [8] Deleted	Muir, Paul	9/14/15 4:09 AM

the

Page 3: [8] Deleted	Muir, Paul	9/14/15 4:09 AM

the

Page 3: [8] Deleted	Muir, Paul	9/14/15 4:09 AM

the

Page 3: [8] Deleted	Muir, Paul	9/14/15 4:09 AM

the

Page 3: [8] Deleted	Muir, Paul	9/14/15 4:09 AM

the

Page 3: [8] Deleted	Muir, Paul	9/14/15 4:09 AM

the

Page 3: [8] Deleted	Muir, Paul	9/14/15 4:09 AM

the

Page 3: [9] Moved to page 4 (Move #3)	Muir, Paul	9/14/15 4:09 AM

Innovations underlying scaling in alignment algorithms:

Page 3: [10] Deleted	Muir, Paul	9/14/15 4:09 AM

In the early Sanger sequencing age, the Smith-Waterman and Needleman-Wunsch algorithms used dynamic programming to find a local or global optimal alignment.

Page 3: [11] Deleted	Muir, Paul	9/14/15 4:09 AM

In light of this computational time bottleneck and increasing dataset sizes, hash table based methods that use a seed-and-extend paradigm with a word of length k(k-mer) as the seed were developed to drive down alignment time. The original FASTA approach simply combines the K-mer to find the common ones between query and target sequences. However, it cannot make sure best alignments are seeded. To improve, BLAST adopts a heuristic statistical method to find high-scoring segment pairs (HSPs) by using substitution matrix and k-letter word, which can perform over 50 times faster than Smith-waterman algorithm. Not like BLAST hashing the query sequence and scanning it against sequence database, BLAT builds a k-mer index for the genome and scans against query sequence and is able to achieve run times 500 times faster than BLAST.

Now, the challenge has turned into rapidly aligning millions of short sequences (reads) to a reference genome for next generation sequencing (NGS) aligners. MAQ and Novoalign are both based on k-mer hash tables. Gapped-kmer is used by MAQ to improve the sensitivity of seed-and-extension schema. And then Suffix array/tree and its variant data structure are wide used in reads alignment. STAR employs the uncompressed suffix array to find a maximal exactly matched seeds and then extend based on seed clusters. BWA and Bowtie utilize Burrows-Wheeler Transform (BWT) to link suffix array with FM-index (Ferragina–Manzini index or Full-text index in Minute space) and find exact match by backward searching. They convert inexact match, which allow mismatches and gaps, into exact match by enumerating all combinations of mismatches and gaps. Finally, these tools sacrifice optimal alignment for extremely fast retrieval of exact matches.

Meanwhile, many of the algorithmic advancements employed by alignment tools try to reduce the marginal mapping cost by building an index data structure. In general, a negative correlated trend can be found between the index and alignment time. (see figure 2) The hash table based tools: BLAT, MAQ and Novoalign build index structure very fast, but relatively require more time to do alignment. BWA and STAR take much more time to build index data structure (FM-index and suffix array), but reads alignment of these tools are ultra fast. Decreasing “marginal” alignment cost by reasonably increasing “fixed” index time makes them more suitable to handle progressively rising NGS data.

Page 3: [12] Formatted	Muir, Paul	9/14/15 4:09 AM

Font:11 pt

Page 3: [12] Formatted	Muir, Paul	9/14/15 4:09 AM

Font:11 pt

Page 3: [13] Moved to page 4 (Move #7)	Muir, Paul	9/14/15 4:09 AM

Scalable storage, query and analysis technologies are necessary to handle the increasing amounts of genomic data being generated and stored. For example, distributed file system greatly increases the storage I/O bandwidth, making distributed computing and data management possible. Another example is the NoSQL database which provides excellent horizontal scalability, data structure flexibility, and support for high load interactive queries.

Page 4: [14] Deleted	Muir, Paul	9/14/15 4:09 AM

Changing computing paradigms such as cloud computing are playing a role in managing the flood of sequencing data. HIPAA compliant cloud resources are being developed so that datasets can be stored and shared on remote servers.

Page 4: [15] Moved to page 5 (Move #9)	Muir, Paul	9/14/15 4:09 AM

However,

Page 4: [16] Deleted	Muir, Paul	9/14/15 4:09 AM

privacy protection in the cloud environment becomes a huge concern. Researchers are interested in finding reliable and affordable solution to minimize the risk of sensitive data leakage. Privacy protection in cloud environment can be split into two layers: a. protect sensitive data from leaking to a third party [[cite…some interesting work includes (limited) computation and query directly on encrypted database, isolating encrypted data etc.]]; b. make the computation oblivious to the cloud service provider [[cite…]]].

Traditional scientific computing paradigm is aggressively optimized on linear algebra. This is not of much benefit to nowadays bioinformatics research, which

Page 7: [17] Deleted	Muir, Paul	9/14/15 4:09 AM

In