The real cost of sequencing: higher than you think!
The real cost of sequencing: processing, storage & data transfer

Introduction:

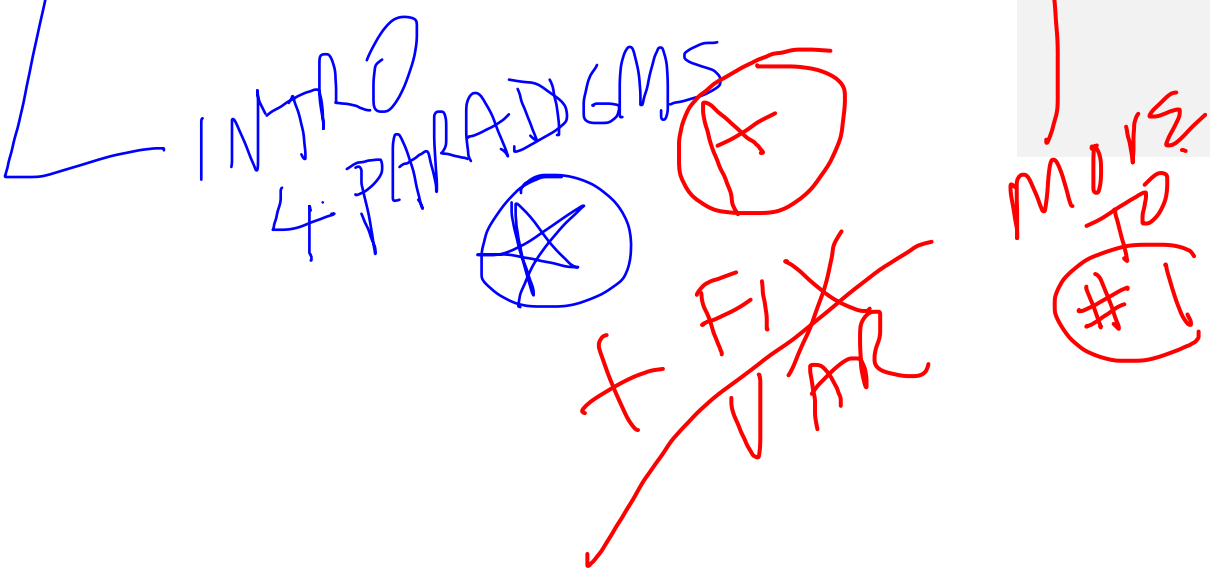
Introduction:

The contemporaneous development of biopolymer sequencing and the digital computer started a digital revolution in the biosciences. [[based & like]] 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. The data generated by biological experiments was often not in a form that benefited from computational processing power. However, this changed with the advent of Sanger sequencing and generation of ever greater amounts of sequence data. Large amounts of sequence data could be stored in computational 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 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 could 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.

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'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. This 4th paradigm holds the possibility of synthesizing the previous paradigms of empirical observation, theory, and computational simulation. However, in



Style Definition: Heading 1
Style Definition: Heading 2
Style Definition: Heading 3
Style Definition: Heading 4
Style Definition: Heading 5
Style Definition: Heading 6
Style Definition: Title
Style Definition: Subtitle

Deleted: Sanger
Deleted: personal
Deleted: Prior to the introduction of the PC, the use of computers in biological research was sparse. This was due in part to the expense and inaccessibility of mainframe computers. Additionally, some
Deleted: are easily
Deleted: the
Deleted: The internet further encouraged sharing of sequence data. The completion of the human genome project coincided with the rise of the internet. These conditions
Deleted: Additionally, efforts to improve sharing of biological information have influenced the development of the internet. The CGI.pm module instrumental in the development of interactive websites was written by Lincoln Stein, a computational biologist looking to develop a method to better share genomic information over the web.
Moved (insertion) [1]

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 advent of next generation sequencing (NGS) has 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 approximately $3.9*10^{15}$ 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. 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. 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. 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.

Backdrop of the computer industry & Moore's law:

Semiconductor technology has 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. People have made observations of various laws which model and predict the rapid developmental progresses in these high-tech areas that are driven by the progress in semiconductor technology. For instance, the well-known Moore's law accurately predicted that the number of transistors integrated in each square inch would double every two year \cite{}. The semiconductor industry has used the Moore's law to plan its research and development progress. Besides Moore's law, various other corresponding predictive laws have also been proposed for related high-tech development (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) was proposed to predict the cost of a semiconductor chip fabrication plant doubles around every four years. Similarly, Kryder's law describes the related roughly yearly doubling of the area storage density of hard drives over the last few decades.

The roughly yearly doubling scaling of these described by these laws over the period of multiple decades is not simply the scaling behavior of a single technology but the superposition of the S-

Deleted: high throughput

Moved (insertion) [2]

Deleted: high throughput sequencing

Deleted: The scale of NGS has enabled the collection of a large amount of personally identifiable sequence information.

Deleted: these datasets.

Moved (insertion) [3]

Deleted: also

Deleted: (http://sourcetech411.com/2012/12/engineering-laws-moores-rocks-butters-and-others/).

curve behavior of different technology over the life of the scaling behavior (see figure 1). The S-curve behavior of an individual technology is due to the three main phases (development, expansion and maturity). For example, Kyder's Law, yearly doubling scaling behavior over the last two and a half decades is the superposition of the S-curves of five different technologies. This behavior is also true for sequencing based technologies.

The success of the predictive laws in high tech areas in last half century have encouraged the development of laws to forecast trends in related 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 sequencing cost 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, the cost of sequencing has dropped faster than would be expected using Moore's law as a guide. In recent five years, the cost of sequencing a personal genomics dramatically dropped to XXX in 2014 from XXXX in 2008. This departure from Moore's law is due to the dramatically different S-curve slopes for 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.

Innovations underlying scaling in alignment algorithms:

[[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.]]

Alignment tools have co-evolved with sequencing technology to meet demand of sequence data processing. The 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-Waterman and Needleman-Wunsch 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. 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

Deleted: [[Moore's is baked into the computer industry.... will it become baked to illumina? Cern ... [1]
Deleted:
Formatted: Font:Arial, 11 pt, Not Bold
Formatted: Heading 2
Formatted: Font:Arial, 11 pt, Not Bold, No underline
Formatted: Font:Arial, 11 pt, Not Bold
Deleted: an exact-matched K
Formatted: Font:Arial, 11 pt, Not Bold
Formatted: Font:Arial, 11 pt, Not Bold
Deleted: FASTA, BLAST and its successor BLAT ... [2]
Formatted: Font:Arial, 11 pt, Not Bold
Deleted: , which
Formatted: Font:Arial, 11 pt, Not Bold
Deleted: and Smith-Waterman algorithm,
Formatted: Font:Arial, 11 pt, Not Bold
Formatted: Font:Arial, 11 pt, Not Bold
Deleted:), which
Formatted: Font:Arial, 11 pt, Not Bold
Deleted: by
Formatted: Font:Arial, 11 pt, Not Bold
Deleted: . In contrast
Formatted: Font:Arial, 11 pt, Not Bold
Deleted: an
Formatted: Font:Arial, 11 pt, Not Bold
Deleted: the K-mer
Formatted: Font:Arial, 11 pt, Not Bold
Deleted: . By hashing the genome instead of the ... [3]
Formatted: Font:Arial, 11 pt, Not Bold
Deleted: BLAT
Formatted: Font:Arial, 11 pt, Not Bold
Deleted: [[SKL2PM: 500 for mRNA/DNA alignme ... [4]
Formatted: Font:Arial, 11 pt, Not Bold
Deleted: The design of these alignment algorit ... [5]
Moved up [3]: . In
Deleted: order to better operate at the scale o ... [6]
Formatted: Font:Arial, 11 pt, Not Bold
Deleted: based on building hash tables for the ... [7]
Formatted: Font:Arial, 11 pt, Not Bold
Deleted: BWT.
Formatted: Font:Arial, 11 pt, Not Bold
Deleted: K
Formatted: Font:Arial, 11 pt, Not Bold

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.

Computational component of sequencing - what's happening in bioinformatics:

[[STL(Sep4): The flow here is "data is getting large > hard to share > we need to move to cloud > we can even do computation in the cloud and only transfer the result to the user, saving transmission cost etc. " It describes a data warehouse that can do some minimal computation The scope is too narrow. Cloud is more perceived as an economic and flexible way to do computing rather than data distribution. Computing is like utility nowadays. Users pay as they go, very flexible, the economy also scales…]]

The decreasing cost of sequencing and increasing amount of sequence reads 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 easily and intuitively accessible to the larger research community. 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.

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. 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.

Formatted ... [8]
Deleted: a category of
Formatted ... [9]
Deleted: , such as
Formatted ... [10]
Deleted: and FM-index (Ferragina-Manzini ind ... [11]
Formatted ... [12]
Deleted: perfect match instead of dynamic ... [13]
Formatted ... [14]
Deleted: can
Formatted ... [15]
Deleted: /tree
Formatted ... [16]
Deleted: to
Formatted ... [17]
Deleted: possible
Formatted ... [18]
Deleted: in the query sequence. These trend ... [19]
Formatted ... [20]
Deleted: and error tolerance
Formatted ... [21]
Deleted: perfect
Formatted ... [22]
Deleted: ... [23]
Formatted ... [24]
Deleted: , and generally
Formatted ... [25]
Deleted: time
Formatted ... [26]
Deleted: are highly negative correlated
Formatted ... [27]
Deleted:). To optimize and improve the align ... [28]
Formatted ... [29]
Deleted: significant
Formatted ... [30]
Deleted: or
Formatted ... [31]
Deleted: require
Formatted ... [32]
Deleted: or
Formatted ... [33]
Deleted: Dramatically decreasing
Formatted ... [34]
Formatted ... [35]
Formatted ... [36]
Deleted:
Deleted: [[STL: also democratized research...no ... [37]

However, 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 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)

The explosion of sequencing data has posed a need of efficient methods for storage and transmission. General algorithms like gzip offer great compatibility, good compression speed and acceptable compression efficiency on sequencing data and are thus widely used. However, to further reduce storage footprint and transmission time, customized algorithms are needed. Many researchers use SAM/BAM (Sequence/Binary Alignment/Map) format to store reads. An extensively 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 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).

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 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,

Moved down [4]: Illustrations of the dramatic increase in rate and amount of sequencing:

Moved up [2]: 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 approximately $3.9*10^{15}$ bases with approximately half of these being open access. The size and growth rate of the SRA

Moved down [5]: 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.

Moved down [6]: 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 Cell and Molecular Ecology remained low for a significantly longer time before dramatically increasing (see figure 3).

Deleted:

Moved down [7]: 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.

Deleted: ... [38]

Deleted:

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

Deleted:

Deleted:

Deleted: high throughput

Deleted: Ref of the

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.

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

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).

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.

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 Cell and Molecular Ecology remained low for a significantly longer time before dramatically increasing (see figure 3).

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

Moved down [8]: ... [39]

Deleted:

Moved up [1]: 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.

Deleted: 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.

Moved (insertion) [8]

Deleted: In order to fully utilize of large sequence databases bioinformaticians are needed to both identify meaningful patterns within the data and to develop methods and APIs for easily navigating and interacting with the ever increasing amounts of data. Furthermore, in the genomic era it is vital to train computationally literate biologists who are capable of leveraging the ever growing digital biological knowledge base and situating their work in this larger context.

Moved (insertion) [4]

Moved (insertion) [5]

Formatted: Space Before: 10 pt

Moved (insertion) [6]

Moved (insertion) [7]

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.

Deleted:

Page 3: [1] Deleted **Author** **9/4/15 6:46 PM**

[[Moore's is baked into the computer industry.... will it become baked to illumina? Cern thing - how has moore's law affected sci - & Moore's 2nd law]]

Page 3: [2] Deleted **Author** **9/4/15 6:46 PM**

FASTA, BLAST and its successor BLAT try to improve and optimize the Smith-waterman algorithm while losing as little alignment accuracy as possible.

Page 3: [3] Deleted **Author** **9/4/15 6:46 PM**

. By hashing the genome instead of the query sequence

Page 3: [4] Deleted **Author** **9/4/15 6:46 PM**

[[SKL2PM: 500 for mRNA/DNA alignment, 50 times faster than protein alignment \cite{11932250}]].

Page 3: [5] Deleted **Author** **9/4/15 6:46 PM**

The design of these alignment algorithms changes from previous tools and two distinct trends can be found. On the one hand, NGS alignment tools have adopted alternative alignment methods instead of Smith-water algorithm to avoid high computing cost.

Page 3: [6] Deleted **Author** **9/4/15 6:46 PM**

order to better operate at the scale of NGS datasets alignment algorithms have made use of additional techniques such as suffix arrays and the Burrows-Wheeler transform. Among tools selected in our analysis,

Page 3: [7] Deleted **Author** **9/4/15 6:46 PM**

based on building hash tables for the reads and reference respectively, STAR is based on suffix array, BWA and Bowtie are

Page 4: [8] Formatted **Author** **9/4/15 6:46 PM**

Font:Arial, 11 pt, Not Bold

Page 4: [9] Formatted **Author** **9/4/15 6:46 PM**

Font:Arial, 11 pt, Not Bold

Page 4: [10] Formatted **Author** **9/4/15 6:46 PM**

Font:Arial, 11 pt, Not Bold

Page 4: [11] Deleted **Author** **9/4/15 6:46 PM**

and FM-index (Ferragina-Manzini index or Full-text index in Minute space) adopted by STAR and BWA respectively, are used

Page 4: [12] Formatted **Author** **9/4/15 6:46 PM**
Font:Arial, 11 pt, Not Bold

Page 4: [13] Deleted **Author** **9/4/15 6:46 PM**
perfect match instead of dynamic programming. In particular,

Page 4: [14] Formatted **Author** **9/4/15 6:46 PM**
Font:Arial, 11 pt, Not Bold

Page 4: [15] Formatted **Author** **9/4/15 6:46 PM**
Font:Arial, 11 pt, Not Bold

Page 4: [16] Formatted **Author** **9/4/15 6:46 PM**
Font:Arial, 11 pt, Not Bold

Page 4: [17] Formatted **Author** **9/4/15 6:46 PM**
Font:Arial, 11 pt, Not Bold

Page 4: [17] Formatted **Author** **9/4/15 6:46 PM**
Font:Arial, 11 pt, Not Bold

Page 4: [18] Formatted **Author** **9/4/15 6:46 PM**
Font:Arial, 11 pt, Not Bold

Page 4: [19] Deleted **Author** **9/4/15 6:46 PM**
in the query sequence. These trends tend to

Page 4: [20] Formatted **Author** **9/4/15 6:46 PM**
Font:Arial, 11 pt, Not Bold

Page 4: [21] Formatted **Author** **9/4/15 6:46 PM**
Font:Arial, 11 pt, Not Bold

Page 4: [22] Formatted **Author** **9/4/15 6:46 PM**
Font:Arial, 11 pt, Not Bold

Page 4: [23] Deleted **Author** **9/4/15 6:46 PM**

Many

Page 4: [24] Formatted **Author** **9/4/15 6:46 PM**
Font:Arial, 11 pt, Not Bold

Page 4: [24] Formatted **Author** **9/4/15 6:46 PM**
Font:Arial, 11 pt, Not Bold

Page 4: [25] Formatted	Author	9/4/15 6:46 PM

Font:Arial, 11 pt, Not Bold

Page 4: [26] Formatted	Author	9/4/15 6:46 PM

Font:Arial, 11 pt, Not Bold

Page 4: [27] Formatted	Author	9/4/15 6:46 PM

Font:Arial, 11 pt, Not Bold

Page 4: [28] Deleted	Author	9/4/15 6:46 PM

). To optimize and improve the alignment performance, the index data structure becomes more complex and costs more time to build. We found, In general, the index time and alignment time are highly negative correlated (see figure 2[[SKL2PM: actually, figure 2 doesn't show the clear correlation, I calculated the correlation(-1) using the data for both pearson and spearman]]).

Page 4: [29] Formatted	Author	9/4/15 6:46 PM

Font:Arial, 11 pt, Not Bold

Page 4: [29] Formatted	Author	9/4/15 6:46 PM

Font:Arial, 11 pt, Not Bold

Page 4: [30] Formatted	Author	9/4/15 6:46 PM

Font:Arial, 11 pt, Not Bold

Page 4: [31] Formatted	Author	9/4/15 6:46 PM

Font:Arial, 11 pt, Not Bold

Page 4: [32] Formatted	Author	9/4/15 6:46 PM

Font:Arial, 11 pt, Not Bold

Page 4: [33] Formatted	Author	9/4/15 6:46 PM

Font:Arial, 11 pt, Not Bold

Page 4: [34] Formatted	Author	9/4/15 6:46 PM

Font:Arial, 11 pt, Not Bold

Page 4: [35] Formatted	Author	9/4/15 6:46 PM

Font:Arial, 11 pt, Not Bold

Page 4: [36] Formatted	Author	9/4/15 6:46 PM

Heading 2

Page 4: [37] Deleted	Author	9/4/15 6:46 PM

[[STL: also democratized research...no fixed/sunk cost]]

Page 5: [38] Deleted	Author	9/4/15 6:46 PM

[[STL: distributed computing cuts cost. a single beefy node is much expensive than 100 mediocre nodes]]

Page 6: [39] Moved to page 6 (Move #8)	Author	9/4/15 6:46 PM

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