Cost of Sequencing Draft

The real cost of sequencing: higher than you think!

The real cost of sequencing: scaling behaviour

The real cost of sequencing: it's really now computer calc scale

The real cost of sequencing: processing, storage & data transfer

The real cost of sequencing: will it continue to scale?

Scaling of sequencing costs: a fixed & var cost perspective

Scaling of sequencing costs: Are we still in a moore's law regime

** Introduction

[[MG2PM: work on the intro and increase in seq.]]

- Number of species sequenced
- Bases in major journals over time
- Changes in # of sequencers and locations over time (from omicsmaps.com)

* NIH reporter graph

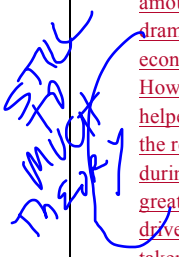
- Cost of sequencing on Genohub

*** TCGA_Data_Universe_2015-07-16 (1).pptx

Fixed and variable costs in sequencing and analysis:

The continued decrease in the cost of sequencing and 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 dramatic increase can be attributed to multiple factors. Sequencing has benefited from highly favorable economies of scale. The initial purchase of a high throughput sequencing machine is a large fixed cost. However, technological innovations that have improved high throughput sequencing machines have helped decrease the per base variable cost of sequencing. This variable cost of sequencing is a function of the reagents and labor required to run a sequencer as well as the amount of data that can be generated during a sequencing run. This high fixed cost and low variable cost encourages the generation of ever greater amounts of sequence data in order to spread the fixed cost over a greater of sequenced bases and drive down the average cost per base. The establishment and growth of sequencing core facilities have taken advantage of this economy of scale and helped increase the accessibility of sequencing technology by mitigating the upfront fixed cost of purchasing machines. (See Omicsmaps historical data figure).

Meanwhile, 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



Style Definition ... [1]
Formatted: Font:(Default) Arial
Deleted: **The real cost of sequencing: higher than you think!** ... [2]
Formatted: Font:(Default) Arial
Deleted: **The real cost of sequencing:**
Formatted: Font:(Default) Arial
Deleted: **The real cost of sequencing:**
Formatted: Font:(Default) Arial
Deleted: **The real cost of sequencing:**
Formatted: Font:(Default) Arial
Deleted:
Formatted: Font:(Default) Arial
Formatted: Font:(Default) Arial
Deleted: **The real cost of sequencing**: An amortized analysis
Formatted ... [3]
Formatted: Heading 2, None, Space Before: 0 pt
Formatted: Font:(Default) Arial
Formatted: Font:(Default) Arial
Formatted: Pattern: Clear, Highlight
Deleted:
Formatted ... [4]
Formatted: Pattern: Clear, Highlight
Deleted:
Formatted ... [5]
Formatted: Pattern: Clear, Highlight
Deleted:
Formatted ... [6]
Formatted ... [7]
Formatted: Pattern: Clear, Highlight
Deleted:
Formatted ... [8]
Formatted ... [9]
Deleted: **Low cost…sequencing and analysislarge databases changing biological research** ... [10]
Formatted: Font:(Default) Arial
Deleted: The establishment and growth of sequencing core facilities has
Deleted: [Add detail about omicsmaps.com historical data figure] The per base cost of sequencing has also been falling, allowing investigators to generate more sequence data. Furthermore… the growth of sequence databases has reduced the cost of obtaining useful sequence information for analysis. Sequence dataData ... [11]

compared to sequence generation. The training and salary of bioinformatics analysts is a key fixed cost in sequence analysis. 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.

Improvements in sequencing are creating an expanding diverse research ecosystem as both large standardized datasets and targeted experimental datasets are being generated. Investigators are increasingly able to contextualize their research by querying the wealth of existing sequence databases. However, to take full advantage of the vast amounts of data being generated efficient and scalable solutions for data transfer, storage, and analysis are required.

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

A key component of the sequence data infrastructure is the sequence read archive (SRA), which was created to store and organized 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 $2.1*10^{15}$ of these being open access. The size and growth rate of the SRA highlight the importance of efficiently storing sequence data for access by the broader scientific community.

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 remained low for a significantly longer time before dramatically increasing. 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). In addition to 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. (see Tree of Life sequencing figure)

**** Computers , backdrop of the computer industry & Moore's**

[[DW + JR]]
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 also used the

Formatted: Font:(Default) Arial
Deleted: These developments
Deleted: DNA
Deleted: a
Deleted: in which consortia are producing and disseminating
Deleted: individual investigators are
Deleted: As this research landscape changes we need
Deleted: reevaluate
Deleted: way in which research is performed
Deleted: how the associated costs
Deleted: calculated
Formatted: Font:(Default) Arial
Formatted: Font:(Default) Arial
Deleted: The dramatic drop in sequencing costs has been accompanied by an explosion of sequence data generation. This sequence generation is occurring in part through large consortia generating enormous datasets. Large consortia
Formatted: Font:(Default) Arial
Deleted: Meanwhile, an ever expanding set of seq related assays has taken advantage of inexpensive sequencing to serve as a readout in assays investigating a range of biological processes Additionally, ever larger amounts of sequence data are being generated from experimental protocols that utilize DNA sequence data generated from high throughput platforms as a readout. This data is more lab and experiment specific compared to the standardized creation of general use datasets addressing larger questions such as TCGA, 1000 Genomes, and ENCODE and can serve as a valuable addition to such larger standardized approaches. Individual lab generated and highly question specific datasets can provide valuable additional annotations and context to the larger general datasets. ... [12]
Formatted: Font:(Default) Arial
Formatted: Font:
Formatted: Heading 2, None, Space Before: 0 pt
Formatted: Font:(Default) Trebuchet MS, 13 pt
Deleted: [[MG2DW : edit
Formatted: Font color: R,G,B (255,252,246)
Formatted: Font color: R,G,B (255,252,246)
Deleted: background section]] ... [13]
Formatted: Font color: R,G,B (255,252,246)
Deleted: the development of integrated circuits; e.g.,
Formatted: Font color: R,G,B (255,252,246)
Deleted: almost
Formatted: Font color: R,G,B (255,252,246)
Formatted: Font color: R,G,B (255,252,246)

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://sourcetech411.com/2012/12/engineering-laws-moores-rocks-butters-and-others/). 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-curve behavior of different technology over the life of the scaling behavior (see figure XXX). 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 fives different technologies. This behavior is also true for sequencing based technologies. Because the success of the predictive laws in high tech areas in last half century, companies have used these laws to forecast the new related emergent technologies such as for sequencing based technologies. For example, the sequencing cost indeed roughly followed Moore's law type behavior in the decade before 2008 \cite{NIH cost-seq figure}.

The development of sequencing technology, however, is driven by both scientific questions and applications at the current stage unlike the integrated circuits which are motivated solely by applications. Thus, we think that the development of sequencing technology at this stage is far away from following a predictive trajectory. At the scientific side, the whole community strives for solving widely variety of biological questions that haven't been answered. For example, various next generation sequencing technologies such as RNA-seq, ChIP-seq, which have sprung up after 2008, try to find molecular activities at the genome scale. At the application side, the sequencing technology has already been widely used in biomedicine such as personalized genomics. For example, the sequencing cost has not followed a Moore's like law after 2008 \cite{NIH cost-seq figure}. The cost of sequencing has dropped faster than expected. In recent five years, the cost of sequencing a personal genomics dramatically dropped to XXX in 2014 from XXXX in 2008.

Comparison of sequencing technology's trajectory to the growth of the computer industry.

which has experienced a similar if less dramatic scaling in its capabilities, can yield insights into the future of sequencing. The exponential scaling of the number of transistors in a microprocessor reshaped both the computer industry and a host of other industries. This rate of technological improvement enabled increases in computer performance and decreases in cost. Higher performance machines allowed computers to address ever more challenging problems while decreases in cost drove their widespread adoption. [[STL: distributed computing cuts cost. a single beefy node is much expensive than 100 mediocre nodes]] Additionally, the development of intuitive interfaces and research on human-computer interaction helped harness these technological improvements.[[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]].

[[

Deleted: guide
Formatted: Font color: R,G,B (255,252,246)
Deleted: R&D
Deleted: After
Formatted: Font color: R,G,B (255,252,246)
Formatted: Font color: R,G,B (255,252,246)
Formatted: Font color: R,G,B (255,252,246)
Formatted: Font color: R,G,B (255,252,246)
Deleted: were
Formatted: Font color: R,G,B (255,252,246)
Formatted: Font color: R,G,B (255,252,246)
Deleted: the
Formatted: Font color: R,G,B (255,252,246)
Formatted: Font color: R,G,B (255,252,246)
Formatted: Font color: R,G,B (255,252,246)
Deleted: ones always tend to use those
Formatted: Font color: R,G,B (255,252,246)
Formatted: Font color: R,G,B (255,252,246)
Deleted: including the
Formatted: Font color: R,G,B (255,252,246)
Deleted: technology
Formatted: Font color: R,G,B (255,252,246)
Deleted: the
Formatted: Font color: R,G,B (255,252,246)
Formatted: Font color: R,G,B (255,252,246)
Deleted:
Formatted: Font color: R,G,B (255,252,246)
Formatted: Font color: R,G,B (255,252,246)
Deleted: does not simply copy the
Formatted: Font color: R,G,B (255,252,246)
Deleted: of computer industry. It even runs
Formatted: Font color: R,G,B (255,252,246)
Formatted: Font:(Default) Arial
Deleted: S-shaped curves contribution to scaling behavior .
Formatted: Font:(Default) Arial
Formatted: Pattern: Clear, Highlight
Formatted: Font:(Default) Arial
Formatted: Font:(Default) Arial
Formatted: Font:(Default) Arial

paradigms for sci computing
mainframe era
pc era
web era [bioinformatics grew up in]
cloud era
how things scale differently !
]]

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

Scientific computing and cloud computing

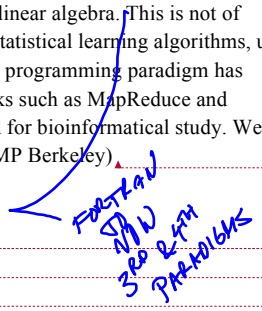
[[MG2STL : edit the bioinformatics section from a more CS perspective]]

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. 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 NoSQL database provides excellent horizontal scalability, data structure flexibility and support for 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. [[STL: also democratized research...no fixed/sunk cost]] [Include download statistics for datasets]

[[STL(Aug. 8): Should we mention commercial cloud server versus in-house hpc?]]

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 bioinformatical study. We have already seen some exciting work in this field (cite ADAM from AMP Berkeley)



** Alignment algorithms

[[MG2SKL : edit the alignment section]]

[[

the algorithmic effect -
dynamic programming
hashing - '92 - blast/fasta hash query

Formatted: Pattern: Clear, Highlight
Formatted: Font:(Default) Arial
Formatted: Pattern: Clear, Highlight
Formatted: Font:(Default) Arial
Formatted: Pattern: Clear, Highlight
Formatted: Font:(Default) Arial
Formatted: Pattern: Clear, Highlight
Formatted: Font:(Default) Arial
Formatted: Pattern: Clear, Highlight
Formatted: Font:(Default) Arial
Formatted: Pattern: Clear, Highlight
Formatted: Font:(Default) Arial
Formatted: Font:(Default) Arial
Formatted: Font:
Formatted: Font:(Default) Trebuchet MS, 13 pt
Formatted: Heading 2, None, Space Before: 0 pt
Formatted: Font:(Default) Arial
Formatted: Font:(Default) Arial
Formatted: Space After: 0 pt
Formatted: Font:(Default) Arial
Formatted: Font:(Default) Arial
Formatted: Font:(Default) Arial
Formatted: Space After: 0 pt
Formatted: Font:
Formatted: Heading 2, None, Space Before: 0 pt
Formatted: Font:(Default) Trebuchet MS, 13 pt
Formatted: Font:(Default) Arial
Formatted: Font:(Default) Arial
Formatted: Font:(Default) Arial
Formatted: Font:(Default) Arial
Formatted: Font:(Default) Arial
Formatted: Font:(Default) Arial

blat - hash the db

BWT - suffix tree

tradeoff from memory & speed - put in fixed v marginal cost
sacrificing of optimal alignment

http://bib.oxfordjournals.org/content/11/5/473.full.pdf

]]

Alignment algorithms

Beyond structural improvements in data storage, alignment tools have co-evolved with sequencing technology to meet demand of sequence data processing. The running time fulfill Moore's Law and decrease by half every 18 months. In the very early Sanger sequencing age, Smith-Waterman and Needleman-Wunsch algorithm use dynamic programming to find a local or global optimal alignment. But both O(mn) time and space complexity make it impossible to map sequences to a large genome. FASTA, BLAST and BLAT, as the successor, introduce seed-and-extend paradigm and use exact-matched K-mer as the seed. However, the original FASTA simply combine the K-mer and Smith-Waterman algorithm, can not make sure best alignments are seeded. BLAST uses a heuristic statistical method to find high-scoring segment pairs (HSPs) by hashing the query sequence and scan it against sequence database. In contrast, BLAT build index for the genome and scan the K-mer against query sequence, which can achieve 50 times faster than BLAST. Now, the challenge turn into rapidly aligning millions of short sequences (reads) to a reference genome for next generation sequencing (NGS) aligners. Among tools selected in our analysis, MAQ and Novoalign are based on hash table, STAR is based on suffix array, BWA and Bowtie are both based on BWT. Gapped-kmer is used by MAQ to improve the sensitivity of seed-and extension schema. And a category of data structure, such as suffix array and FM-index (Ferragina–Manzini index or Full-text index in Minute space) adopted by STAR and BWA respectively, are used to find perfect match instead of dynamic programming. In particular, Burrows-Wheeler Transform (BWT) can link suffix array/tree with FM-index to find exact match by enumerating all combinations of possible mismatches and gaps in the query sequence. The result turn out to be sacrificing optimal alignment and error tolerance for extremely fast retrieval of perfect matches.

On the other hand, more and more alignment tools try to reduce the mapping cost by building an index data structure, and generally the index time and alignment time are highly negative correlated. The hash table based tools: BLAT, MAQ and Novoalign take less time to build index structure, but require significant more time to do alignment. Though it take a long time to build the FM-index for BWA and Bowtie, and to construct an uncompressed suffix array for STAR, the index time cost is fixed and the marginal cost for reads alignment can be dramatically reduced, and make them become much more popular to handle progressively rising NGS data.

Data storage and algorithmic improvements also need to be packaged in intuitive and easily navigable formats to spur the wider adoption of sequencing information amongst the biological research

Formatted: Font:(Default) Arial
Formatted: Font:(Default) Arial
Formatted: Font:(Default) Arial
Formatted: Font:(Default) Arial
Formatted: Font:(Default) Arial
Formatted: Font:(Default) Arial
Deleted:
Formatted: Pattern: Clear, Highlight
Formatted: Pattern: Clear, Highlight
Formatted: Font:(Default) Arial
Deleted:
Deleted: um
Deleted: Both the
Deleted: is O(mn) and
Deleted:
Deleted: use
Formatted: Font color: Black, Pattern: Clear, Highlight
Deleted: pair
Formatted: Font color: Black, Pattern: Clear, Highlight
Deleted: For the NGS aligner
Deleted: is to
Deleted: .
Deleted: adopt gapped
Deleted: A
Deleted: ,suffix tree
Deleted:
Deleted: is
Deleted: sacrifice
Formatted: Font:(Default) Arial
Formatted: Justified
Formatted: Font:(Default) Arial
Deleted: a
Deleted: .
Deleted: maintain a
Deleted: are
Formatted: Font:(Default) Arial
Formatted: Space After: 0 pt

community. Illumina's BaseSpace takes a promising step in this direction by creating an environment that integrates everything from data transfer out of the sequencers to the app-like options for analysis programs.

**** How have reduced costs changed biological research:**

[[MG2DS: work on the end]]

- Bioinformatics jobs
- P/E ratio of illumina vs. other tech
- Use of datasets by secondary analysts
- ????? Cost of sequencing on Genohub

*** how will this shape sci. what does it mean for bioinformatics jobs, big v little

The dramatic drop in sequencing costs has changed the biological research landscape and spurred increased generation of sequencing data. However, the extent to which these increases in sequencing data can be translated into novel biological insights is dependent on scaling analysis methods to handle the flood of sequence data.

A recurring theme in high throughput sequencing is that of fixed and variable costs. The initial purchase of sequencing machines is a large initial fixed cost. However, the lowering of variable costs has enabled economies of scale to encourage the generation of vast amounts of sequence data. As newer sequencing machines are able to produce more reads, the average per base cost of sequencing decreases. Similar efforts are needed to drive down the per base cost of analysis.

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. However, such trends also run the risk of fragmenting the genomics research community. If the sequence data generated by individual labs is not processed properly and 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.

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. In the case of consortia, there are often required to ensure that their data is uniformly processed and easily accessible to the public. [[STL: probably we could talk a little about distributed database systems and how it realizes interactive querying in large scale]]

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

Formatted: Font:(Default) Arial

Formatted: Font:

Formatted: Font:(Default) Trebuchet MS, 13 pt

Formatted: Heading 2, None, Space Before: 0 pt

Formatted: Font:(Default) Arial

Formatted: Font:(Default) Arial

Deleted:

Formatted: Pattern: Clear, Highlight

Formatted: Pattern: Clear, Highlight

Formatted: Font:(Default) Arial

Formatted: Pattern: Clear, Highlight

Deleted:

Formatted: Pattern: Clear, Highlight

Formatted: Font:(Default) Arial

Deleted:

Formatted: Pattern: Clear, Highlight

Formatted: Pattern: Clear, Highlight

Formatted: Font:(Default) Arial

Formatted: Pattern: Clear, Highlight

Formatted: Font:(Default) Arial

Formatted: Font:(Default) Arial

Deleted: to what

Deleted: are the

Deleted: due

Deleted: large sequencing centers and established projects producing ever more sequencing

Deleted: as compared to adoption of sequencing approaches by labs which did not previously use sequencing data?

Formatted: Font:(Default) Arial

Formatted: Font:(Default) Arial

Formatted: Font:(Default) Arial

Deleted: A recurring theme in the topic of high throughput

Deleted: is that of fixed and variable costs. The initial purchase of

Deleted: machines is a large initial fixed cost. However, this cost is often largely shouldered by

technologies and research, increasing the diversity and specialization of experiments. Using Illumina sequencing alone, nearly 150 different experimental strategies have been described (Ref of the poster "For all your Seq needs) yielding information about nucleic acid secondary structure, interactions with proteins, spatial information within a nucleus, and more. For example, natural products research has seen a resurgence as it has become clear that the biosynthetic potential of organisms greatly exceeds what has been observed to be produced; efforts to heterologously express and characterize "silent biosynthetic clusters" holds much promise for the discovery of new antibiotics and other medicines. 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 P/E ratio of illumina vs. other tech).

In an era of squeezed budgets and fierce competition, job prospects for scientists with training in computational biology remain strong (Explosion of Bioinformatics Careers Science 2014). Universities have increased the number of hires in the areas of computer science, and specifically in bioinformatics (See Bioinformatics jobs).

====

Possible Figures:

- S-shaped curves contribution to scaling behavior
- Alignment algorithms
- Cost of sequencing on Genohub
- Bioinformatics jobs
- Number of species sequenced
- P/E ratio of illumina vs. other tech (investors expect strong growth of illumnia. They are willing to pay premium for stock price)
- Bases in major journals over time
- Changes in # of sequencers and locations over time (from omicsmaps.com)
- Use of datasets by secondary analysis (Can we split this into reuse of consortia generated data vs data generated from individual labs)

Deleted: core facilities and not directly by individual investigators. The fixed cost is amortized in accounting,…and specialization of experiments. Using Illuminaaffects pricing. Nevertheless, as newer…sequencing alone, nearly 150 different experimental strategies have been described (Ref of machines are able to produce more reads, the poster "For all your Seq needs) yieldingaverage total cost of sequencing decreases. Moreover, if the number of sequencing facilities increases, creating greater competition, economic theory predicts that the price of sequencing should be driven down and approach marginal cost. In an environment of perfect competition, the cost of sequencing should be equal to the marginal cost, and the fixed cost of purchasing a sequencing machine should not enter into the pricing function; rather, it should impact only the dec ... [14]

Formatted: Font:(Default) Arial

Deleted: ... [15]

Formatted: Space After: 0 pt

Formatted: Font:(Default) Arial

Formatted: Font:(Default) Arial

Deleted:

Formatted: Pattern: Clear, Highlight

Formatted ... [16]

Formatted: Pattern: Clear, Highlight

Deleted:

Formatted ... [17]

Formatted: Pattern: Clear, Highlight

Deleted:

Formatted ... [18]

Formatted: Pattern: Clear, Highlight

Deleted:

Formatted ... [19]

Formatted: Pattern: Clear, Highlight

Deleted:

Formatted ... [20]

Formatted: Pattern: Clear, Highlight

Deleted:

Formatted ... [21]

Formatted: Pattern: Clear, Highlight

Deleted:

Formatted ... [22]

Formatted: Pattern: Clear, Highlight

Deleted:

Formatted ... [23]

Deleted:

Formatted: Pattern: Clear, Highlight

Formatted ... [24]

Page 1: [2] Deleted | Muir, Paul | 8/15/15 10:03 AM

The real cost of sequencing: higher than you think!

The real cost of sequencing:

Page 1: [3] Formatted | Muir, Paul | 8/15/15 10:03 AM

Font:

Page 1: [3] Formatted | Muir, Paul | 8/15/15 10:03 AM

Font:

Page 1: [4] Formatted | Muir, Paul | 8/15/15 10:03 AM

Pattern: Clear, Highlight

Page 1: [4] Formatted | Muir, Paul | 8/15/15 10:03 AM

Pattern: Clear, Highlight

Page 1: [5] Formatted | Muir, Paul | 8/15/15 10:03 AM

Pattern: Clear, Highlight

Page 1: [5] Formatted | Muir, Paul | 8/15/15 10:03 AM

Pattern: Clear, Highlight

Page 1: [6] Formatted | Muir, Paul | 8/15/15 10:03 AM

Pattern: Clear, Highlight

Page 1: [6] Formatted | Muir, Paul | 8/15/15 10:03 AM

Pattern: Clear, Highlight

Page 1: [7] Formatted | Muir, Paul | 8/15/15 10:03 AM

Pattern: Clear, Highlight

Page 1: [7] Formatted | Muir, Paul | 8/15/15 10:03 AM

Pattern: Clear, Highlight

Page 1: [8] Formatted | Muir, Paul | 8/15/15 10:03 AM

Pattern: Clear, Highlight

Page 1: [8] Formatted | Muir, Paul | 8/15/15 10:03 AM

Pattern: Clear, Highlight

Page 1: [9] Formatted | Muir, Paul | 8/15/15 10:03 AM

Pattern: Clear, Highlight

Page 1: [9] Formatted | Muir, Paul | 8/15/15 10:03 AM

Pattern: Clear, Highlight

Page 1: [10] Deleted | Muir, Paul | 8/15/15 10:03 AM

Low cost

Page 1: [10] Deleted | Muir, Paul | 8/15/15 10:03 AM

Low cost

Page 1: [11] Deleted	Muir, Paul	8/15/15 10:03 AM

[Add detail about omicsmaps.com historical data figure] The per base cost of sequencing has also been falling, allowing investigators to generate more sequence data. Furthermore

Page 1: [11] Deleted	Muir, Paul	8/15/15 10:03 AM

[Add detail about omicsmaps.com historical data figure] The per base cost of sequencing has also been falling, allowing investigators to generate more sequence data. Furthermore

Page 2: [12] Deleted	Muir, Paul	8/15/15 10:03 AM

Meanwhile, an ever expanding set of seq related assays has taken advantage of inexpensive sequencing to serve as a readout in assays investigating a range of biological processes Additionally, ever larger amounts of sequence data are being generated from experimental protocols that utilize DNA sequence data generated from high throughput platforms as a readout. This data is more lab and experiment specific compared to the standardized creation of general use datasets addressing larger questions such as TCGA, 1000 Genomes, and ENCODE and can serve as a valuable addition to such larger standardized approaches. Individual lab generated and highly question specific datasets can provide valuable additional annotations and context to the larger general datasets.

Online vs. offline science:

In the past, hypothesis driven research provided a clear beginning and endpoint for a given experiment. Data generated during an experiment was designed to test a specific question. The increase in the rate of DNA sequencing and size of sequence databases has spurred a rise in hypothesis generating research. Many sequencing experiments are aimed at providing large standardized datasets for general use. Even in cases where sequencing is used to address a specific question the resultant data may later be employed for a purpose significantly different from that of the original investigator. Under these conditions, the computational storage and analysis component of a sequencing experiment will come to represent an increasing proportion of the costs associated with high throughput sequencing experiments relative to the initial cost of generating sequence data.

The difference between these two paradigms of scientific research is comparable to the distinction in computer science made between offline and online algorithms. Offline algorithms take in all inputs at once and then begin processing. Meanwhile, online algorithms process their input as it arrives without knowledge of all of the subsequent inputs.

The process of hypothesis-driven research is similar to that of the offline algorithm. Data is collected at once and processed afterwards resulting in a conclusion that represents the end of that specific data's use.

As decreased cost and the generation of new seq-based experimental protocols increase the deposition of sequences in public databases it is becoming increasingly difficult to follow the offline algorithm model. The scope of large consortia generated datasets such as TCGA makes waiting for all the data to be generated before performing the analysis prohibitive. The use of sequence from existing databases for secondary analysis creates a situation in which it is impossible to predict the full extent of a dataset's usage. Furthermore, if DNA sequencing and synthesis are to be combined in an iterative cycle to better

link genotypic variation to phenotype, then constant analysis of new data in light of the existing sequencing knowledge base is required.

These new modes of biological research necessitate conceptualizing the experimental process as an online algorithm in which both data generation and analysis are viewed as events in a larger research process of unknown duration.

Amortized analysis:

Amortized analysis is used by computer scientists to evaluate the worst-case scenario over a sequence of operations. This type of analysis is commonly applied to online algorithms. The analysis method relies on the idea that some computationally expensive processes may pay off over a series of operations because they enable speed improvements in subsequent operations. This approach contrasts with less holistic approaches in which the worst-case scenario of each operation in a process is determined independently. These two analysis methods may suggest different series of events as optimal.

It is becoming increasingly important to account for the entire lifespan of a sequence dataset as repeated use of sequence data in different analyses extend the life of a given dataset. As sequencing prices drop and the scale of generated sequence data increases the relative importance of initial data generation decreases. This situation is conducive to the holistic approach taken by amortized analysis.

Considerations in the sequencing pipeline:

[[PM: These next paragraphs might be useful to link between the different sections of the perspective. More work needs to be done applying an amortized analysis to each step in the sequencing pipeling. The amortized analysis might be a good analogy to follow through the perspective as it integrates the fixed and variable cost idea with a more CS perspective.]]

The first question an investigator is faced with is whether to purchase a sequencer or utilize a sequencing core facility or company. Often the costs involved in purchase, maintenance, and operation of a sequencer as well as the requisite expertise required of its operators makes it unfavorable for individual investigators to purchase a sequencer. However, this could change with the advent of low cost sequencer alternatives such as Oxford Nanopore's MinION.

[[The following links to cost on Genohub as well as elements in the intro and conclusion]]

The investigator then must decide on the protocol for sample preparation and which sequencing platform best conforms to their needs. Here a focus on the constraints of the scientific question being asked and the experimental setup available must be taken into account while optimizing for collection of the most informative sequence dataset.

[[The following links to data size, network bandwidth, and computational processing power scaling]]

Next the investigator must decide in where and in what format to store the sequence data. At this step it is important to consider both the volume of sequence generated as well as the frequency and mode of data access required for downstream analyses. Options at this step range from downloading the data on a local

machine where all analyses will also be performed to uploading the data to the cloud and similarly performing analyses in the cloud.

[[The following links to alignment algorithms]]
After storing the data, the investigator must decide what algorithms to use in order to initially process the data. These alignment algorithms can be viewed as a microcosm for the types of questions now being asked about sequencing more generally. The initial alignment algorithms developed in the 1970s would be hopelessly slow if confronted with the scale of modern datasets. Over time algorithmic innovations have enabled alignment algorithms to keep up with the dramatic increase in size and scale. These newer algorithms utilize practices encouraged by amortized analysis. They devote a significant amount of their runtime to a computationally expensive indexing operation that later provides significant improvements in the performance of the alignment operation.

[[The following links to the conclusion]]
Once the data has been initially processed, downstream analyses and interpretation must be performed to obtain scientific insights and knowledge. The search for biological meaning in these datasets will be helped by two trends. The first trend is an increase in the size and diversity of sequence-based datasets allowing for ever greater statistical power and new comparisons between datasets. This will help with large scale analyses of both labelled and unlabeled biological datasets. The second trend is a decrease in the cost of DNA synthesis combined with improvements in genome engineering at single nucleotide resolution. This will enable investigators to more easily experimentally follow up on findings derived from sequence analysis.

Page 2: [13] Deleted **Muir, Paul** **8/15/15 10:03 AM**

background section]]

Moore

Page 7: [14] Deleted **Muir, Paul** **8/15/15 10:03 AM**

core facilities and not directly by individual investigators. The fixed cost is amortized in accounting,

Page 7: [14] Deleted **Muir, Paul** **8/15/15 10:03 AM**

core facilities and not directly by individual investigators. The fixed cost is amortized in accounting,

Page 7: [14] Deleted **Muir, Paul** **8/15/15 10:03 AM**

core facilities and not directly by individual investigators. The fixed cost is amortized in accounting,

Page 7: [14] Deleted **Muir, Paul** **8/15/15 10:03 AM**

core facilities and not directly by individual investigators. The fixed cost is amortized in accounting,

Page 7: [14] Deleted **Muir, Paul** **8/15/15 10:03 AM**

core facilities and not directly by individual investigators. The fixed cost is amortized in accounting,

Page 7: [14] Deleted **Muir, Paul** **8/15/15 10:03 AM**

core facilities and not directly by individual investigators. The fixed cost is amortized in accounting,

Page 7: [14] Deleted **Muir, Paul** **8/15/15 10:03 AM**

core facilities and not directly by individual investigators. The fixed cost is amortized in accounting,

Page 7: [14] Deleted	Muir, Paul	8/15/15 10:03 AM

core facilities and not directly by individual investigators. The fixed cost is amortized in accounting,

Page 7: [14] Deleted	Muir, Paul	8/15/15 10:03 AM

core facilities and not directly by individual investigators. The fixed cost is amortized in accounting,

Page 7: [15] Deleted	Muir, Paul	8/15/15 10:03 AM

====

Page 7: [16] Formatted	Muir, Paul	8/15/15 10:03 AM

Pattern: Clear, Highlight

Page 7: [16] Formatted	Muir, Paul	8/15/15 10:03 AM

Pattern: Clear, Highlight

Page 7: [17] Formatted	Muir, Paul	8/15/15 10:03 AM

Pattern: Clear, Highlight

Page 7: [17] Formatted	Muir, Paul	8/15/15 10:03 AM

Pattern: Clear, Highlight

Page 7: [18] Formatted	Muir, Paul	8/15/15 10:03 AM

Pattern: Clear, Highlight

Page 7: [18] Formatted	Muir, Paul	8/15/15 10:03 AM

Pattern: Clear, Highlight

Page 7: [19] Formatted	Muir, Paul	8/15/15 10:03 AM

Pattern: Clear, Highlight

Page 7: [19] Formatted	Muir, Paul	8/15/15 10:03 AM

Pattern: Clear, Highlight

Page 7: [20] Formatted	Muir, Paul	8/15/15 10:03 AM

Pattern: Clear, Highlight

Page 7: [20] Formatted	Muir, Paul	8/15/15 10:03 AM

Pattern: Clear, Highlight

Page 7: [21] Formatted	Muir, Paul	8/15/15 10:03 AM

Pattern: Clear, Highlight

Page 7: [21] Formatted	Muir, Paul	8/15/15 10:03 AM

Pattern: Clear, Highlight

Page 7: [22] Formatted	Muir, Paul	8/15/15 10:03 AM

Pattern: Clear, Highlight

Page 7: [22] Formatted	Muir, Paul	8/15/15 10:03 AM

Pattern: Clear, Highlight

Page 7: [23] Formatted	Muir, Paul	8/15/15 10:03 AM

Pattern: Clear, Highlight

Page 7: [23] Formatted	Muir, Paul	8/15/15 10:03 AM

Pattern: Clear, Highlight

Page 7: [24] Formatted	Muir, Paul	8/15/15 10:03 AM

Pattern: Clear, Highlight

Page 7: [24] Formatted	Muir, Paul	8/15/15 10:03 AM

Pattern: Clear, Highlight