Cost of Sequencing Draft

Introduction:

As sequencing prices have continued to drop over the past four years the relative contributions of the different stages of sequencing data generation and analysis to economic calculations surrounding sequencing projects have changed. The establishment and growth of sequencing core facilities has helped increase the accessibility of sequencing technology by reducing the upfront fixed cost of purchasing machines. 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. 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. Over time the computational component of sequencing will come to represent an increasing proportion of the costs associated with high throughput sequencing experiments.

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.

A recurring theme in the topic of high throughput sequencing is that of fixed and variable costs. The initial purchase of sequencing machines is a large initial fixed cost. However, this cost is often largely shouldered by sequencing core facilities and not directly by individual investigators. The fixed cost is amortized in accounting, and affects pricing. Nevertheless, as newer sequencing machines are able to produce more reads, the average 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 decision of whether or not to operate. If we think about the use of previously generated sequencing information there are almost no fixed costs in obtaining sequence information. This condition would suggest a significant increase in market (sequence-based research) entry. What is keeping researchers out of this area? The variable costs of computational resources and training.

Computational component of sequencing:

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 and search [[STL: storage, query/traverse (primitives) and computing (analysis). Could consider to make this part more CSish by discussing distributed computing, file systems and new database management technologies]], technologies are necessary to handle the increasing amounts of genomic data being generated and stored.

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 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]

[[SKL:1 paragraph summary about alignment, Done, I have written three paragraph and try to clearly explain more details, just for reference, some of them need be put into method]]

Beyond structural improvements in data storage, alignment tools have co-evolved with sequencing technology to meet demand of sequence data processing. In the very early Sanger sequencing age, the algorithms (Smith-Waterman and Needleman-Wunsch) were designed to compare pairwise sequences and obtain a local or global optimum alignment. Later, as ever larger amounts of sequences accumulated and the human genome project was completed, the major task was to quickly find homologous regions or subsequences from massive sequence databases. With the advent of high throughput sequencing, the challenge became to rapidly align millions of short sequences (reads) to a reference genome. To explore the evolution of sequence alignment algorithms, we compared the alignment efficiency of tools developed from the 1970's to the 2010's.

Since more recent tools mainly focus on short sequences (50-200bp) and usually contain an index step to consolidate auxiliary data structure onto the hard disk and the index time will bias our estimation if the number of sequence is too few, we have simulated 1 Million short sequences in 75 bp. On the other hand, the very early algorithms, Smith-Waterman and Needle-Wunsch may need huge memory to construct a backtracking matrix for human genome and some algorithms will also allocate a large memory to achieve best performance when doing alignment. To balance this, we allocate rational enough RAM(at most 60GB) for all tools, and only test Smith-Waterman and Needle-Wunsch on Yeast genome with 1000 random selected short sequences from 1 Million simulated reads. For alignment tools that need to build index, the total running time is the summation of index and alignment time, and all the alignment time are scale up to the same 1 Million level. The total running time for all the algorithms reduce exponentially from 1970's to 2010's for both human and yeast dataset. The running time fulfill Moore's Law and decrease by half every 18 months [[SKL: Moore law has different versions, here I use 18month because based on linear regression, each year will drop to about 0.7, and every 18 month drop by half, and over 1000 and 400000 fold decrease between the fastest and slowest algorithms for human and yeast respectively]]. We also compared the indexing time and alignment time for those have. Quite interestingly, index time ratio is negatively correlated with alignment time ratio with correlation coefficient = -1. In general, algorithms only need index the reference genome once (except Maq that will also need index reads), the time cost for index can be thought as constant, and the alignment time can be dramatically decreased for those alignment algorithms, such as BWA, STAR, which maintain a more complex, well-designed auxiliary index structure.

[[SKL: suggest to delete the highlight region]] new algorithms are needed to more efficiently handle and process sequence data. The impact and importance of improvements in the algorithmic component of sequence analysis can be seen in the advances in alignment algorithms over time. Older alignment algorithms are hopelessly slow when confronted with something the size of the human genome. A graph of the running time of alignment algorithms over time emphasizes the decrease in running time as new algorithms have been released over the years. Another interesting feature in the graph is the relative contribution of indexing and

alignment to the total time of an algorithm. The relative importance of the fixed cost of building an efficient index relative to the variable cost of alignment can be seen changing as the data volume increases.

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 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:

The dramatic drop in sequencing costs has changed the biological research landscape and spurred increased generation of sequencing data. However, to what extent are the increases in sequencing data due to large sequencing centers and established projects producing ever more sequencing data as compared to adoption of sequencing approaches by labs which did not previously use sequencing data? Large consortia have taken advantage of sequencing trends to generate population scale genomic data (1000 Genomes) or extensive characterization of cancer genomes (TCGA). 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.

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]]