Cost of Sequencing Draft PMv1

Introduction:

As sequencing prices have continued to drop over the past four years the relative contributions of the different stages of the sequencing pipeline 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 essentially 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. 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 shouldered by sequencing core facilities and not directly by individual investigators. Since these sequencing facilities must pay this fixed cost even in the event they shut down it shouldn't determine the minimum price at which they would provide sequencing services. This minimum price is instead determined by the variable costs of labor and reagents. 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 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. [Include download statistics for datasets]

Beyond structural improvements in data storage, 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.

Furthermore, the inclusion of sequencing data public access databases results in positive externalities that change the individual investigator's decision of how much to sequence. Externalities arise when researchers are not able to internalize the indirect costs of or benefits from their research. An investigator will likely only benefit from their initial analysis of the data. Meanwhile, the broader research community would benefit from the inclusion of additional sequencing data in public databases. This creates a disconnect between the incentives of the individual investigator and those of the larger community. The investigator's cost curve is shifted to the left of the community's. Consequently, if individual investigators are required to pay for sequencing projects, even though the decreasing cost of sequencing is leading to ever more sequencing data generation this may still undersupply the amount of sequence data optimal for the research community. Consortia can provide a mechanism to incorporate the social benefit from public datasets into the research funding calculation. Collectively the members of the consortia are able to exert influence on the research activities of each individual member. This ensures that genomic sequences are not undersupplied due to a discrepancy between the PI cost function and the larger research community's cost function.

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
- Bases in major journals over time
- Changes in # of sequencers and locations over time (from omicsmaps.com)
- Use of datasets by secondary analysts
    - Can we split this into reuse of consortia generated data vs data generated from individual labs.