Some Notes on Sequencing

Sebastian Magierowski

***Abstract*—Some things that I think are important for engineers working on high-speed sequencing**

I. Sequencing Applications

- Genome sequencing (traditional): Studies of entire genomes (whole-genome sequencing).
- Exome sequencing (traditional): Investigations of smaller functional portions of the genome.
- Functional genomics (traditional)
  - Transcriptome analysis (small and long RNA-seq): Analysis of the transcribed genome.
  - Chromatin Immunoprecipitation (ChIP-seq): Analysis of protein-DNA binding sites
- Array-capture followed by sequencing (e.g. enriching for transposable elements) [emerging]
- Fosmid pools sequencing (facilitate haplotype phasing) [emerging]
- Metagenomics sequencing [emerging]
- Bisulfite sequencing and methylated DNA immunoprecipitation (MeDIP-Seq) (study DNA methylation patterns) [emerging]

II. Cost

Good information on sequencing costs can be found in [1]. As of January 2014 the **"production cost"** (components of which are listed in [1]) is $4,008 for a whole human genome and 4.5¢ per Mb.

For the per Mb estimate, the cost of just generating raw unassembled data is used. For the genome cost, the cost of the effort to assemble the whole sequence is accounted for. That cost also assumes **"re-sequencing"** which means you have a reference human genome to work with that cuts the costs of the data analysis effort. In other words, if you were sequencing from scratch, **your cost would be higher than thise genome sequencing number.**

The haploid human genome consists of about 3000 Mb (variations of about 1% in the actual length between different humans is not unexpected [2, p. 370]).

With production costs plummeting there are notices that **"non-production cost"** will become a serious problem [3]–[5]. These include [1]:

- Quality assessment/control for sequencing projects
- Technology development to improve sequencing pipelines
- Development of bioinformatics/computations tools to improve sequencing pipelines or to improve downstream sequence analysis
- Management of individual sequencing projects
- Informatics equipment
- Data analysis downstream of initial data processing (e.g. sequence assembly, sequence alignments, identifying variants, and interpretation of results)

Many thanks to EMIL's friends.

III. Quality

How good is a sequencing result? The accepted "high-quality" standard for Sanger-based sequencing is an error probability of 1% or Phred_{20} (or Q_{20}) (like 20-dB?)

IV. Coverage

All mainstream genome sequencing techniques are unable to process a whole 3000 Mb human sequence in one go. They need to chop it up into smaller chunks and then use bioinformatics algorithms to figure out the whole strand again. These algorithms require **"sequence redundancy"** to work. In effect you need a number of overlapping sequence chunks. This redundancy is also called **"sequence coverage"**.

For the main techniques we have:

Sanger: Avg. read is 500-600 bases $\rightarrow$ 6-fold coverage

454: Avg. read is 300-400 bases $\rightarrow$ 10-fold coverage

Illumina and SOLiD: Avg. read is 50-100 bases $\rightarrow$ 30-fold coverage

V. Sequencing Methods

Second generation genome sequencing relies randomly cutting up the DNA molecule into many short overlapping pieces of the full DNA. Each piece is roughly 500-bp long. Once each of these pieces is sequenced the complete DNA is reconstructed from this information using bioinformatics techniques.

Capillary sequencers were used largely before 2005.

Second-generation methods can be thought of as "massively parallel sequencing". They also merge the sequencing (i.e. construction of DNA copies) and detection schemes. As you sequence your DNA segments (i.e. add nucleotides to them) you detect the nucleotides being added. This seems to only work for short sequences however, adding to the post-processing bioinformatics burden for genome construction.

Direct quantitative comparisons are possible.

VI. Key Steps

For NGS a key initial step is to assemble a sequence from the large number of separate reads.

1) Obtain a **draft sequence** (a reasonably confident assembly of reads into **contigs**). Gaps are likely at this stage, but at least crude predictions of the size and location of these gaps should be possible.
2) As part of the **finishing process** use more "targeted methods" to fill in the gaps and correct more obvious

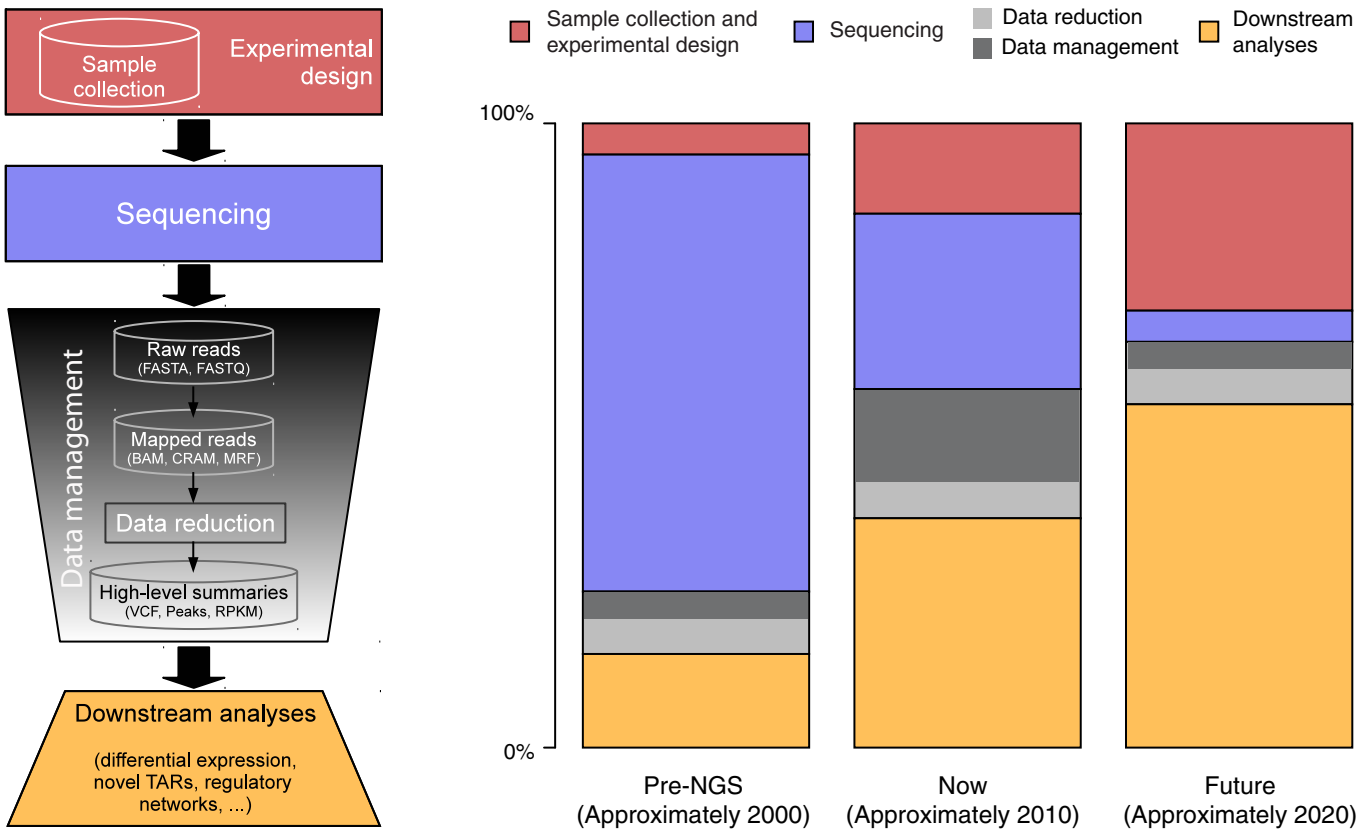

Fig. 1. Sequencing breakdown and cost approximation [5].

errors and uncertainties. No known gaps should remain and *finished* sequence accuracy can be stated to a defined level.

3) The **annotation stage** is the last step when coding sequences and their potential products are identified along with a number of other features.

It looks like this generally falls in the domain of **secondary analysis**, an **extremely important** step in finding the functional patterns in your genome beyond just some list of As, Gs, Cs, and Ts.

A. Analysis

1) Identify ORFs: "A statistical technique that has been invaluable in the identification of protein-coding sequences, and in the prediction of intron-exon boundaries, is known as a hidden Markov Model (HMM)." [6]

2) Identify Gene Function and their Products:

3) Other Features:

B. Repetitive Elements and Gaps

1) **repetitive elements**: Identical sequences that appear more than once (often many times) in the genome. Repetition ranges from a few to nearly 50%.
2) **dispersed/interspersed repeats**: A repeat sequence, mainly "mobile elements" (e.g. insertion sequences and transposons), that occur at different sites distributed around genome.
3) **tandem repeats**: A repeat sequence, usually short (even as low as 1-3 nucleotides), occurring many times in succession at one locus.
4) **copy number variants**: A segmental duplication, where a "block of perhaps several hundred kb has been copied to a different region of the genome".

Repetitive elements are dangerous because they can cause me to splice different regions together during contig formation that cut out large chunks of the genome in between. In the case of *tandem repeats* the number of copies of the tandem repeats can be easy to miscount.

REFERENCES

[1] K. Wetterstrand. (2014) DNA sequencing costs: Data from the NHGRI Genome Sequencing Program (GSP). National Human Genome Research Institute, NIH. [Online]. Available: www.genome.gov/sequencingcosts/
[2] L. H. Hartwell, L. Hood, M. L. Goldberg, A. E. Reynolds, and L. M. Silver, *Genetics: From Genes to Genomes*, 4th ed. New York: McGraw-Hill, 2011.
[3] J. D. McPherson, "Next-generation gap," *Nat. Methods Suppl.*, vol. 6, no. 11, pp. S2–S5, Nov. 2009.
[4] E. R. Mardis, "The $1,000 genome, the $100,000 analysis?" *Genome Medicine*, vol. 2, no. 11, pp. 1–3, 2010.
[5] A. Sboner, X. J. Mu, D. Greenbaum, R. K. Auerbach, and M. B. Gerstein, "The real cost of sequencing: Higher than you think!" *Genome Biology*, vol. 12, no. 8, pp. 1–10, 2011.
[6] J. W. Dale, M. von Schantz, and N. Plant, *From Genes to Genomes: Concepts and Applications of DNA Technology*, 2nd ed. Wiley-Blackwell, 2012.