The *Haemophilus influenzae* genome

1. Background

The first complete genome sequence of a free-living organism.
Published in 1995 (*Science* 269: 496-512).
Genome size =1.8 Mb; 1 circular chromosome.
Previously only viruses or organelles had been sequenced (max. ≈ 200 kb).
Done by Craig Venter's group (TIGR, nonprofit) by WGS method.
Grant proposal rejected by NIH. Done with private funding from Human Genome Sciences (Biotech company) and Perkin-Elmer (maker of automated sequencers, later ABI).

The *H. influenzae* strain sequenced is a benign lab strain. There are strains that cause ear and lung infections (mainly in children) and can cause meningitis. However, it is not of major medical importance.

Why *H. influenzae*? Reasons given in paper:

a) genome size typical for bacteria
b) G+C content (38%) similar to human genome
c) physical map did not exist

Additional reason: Hamilton Smith (Nobel Prize Winner, Johns-Hopkins University) had worked on this bacterial strain for a long time. He discovered restriction enzymes working with *H. influenzae*. He is a senior author on the paper.

2. Genome Sequencing Strategy

For shotgun sequencing, it is important that the library (that is, the collection of DNA fragments inserted into plasmids) is random. How did they test this?
Sequence 4,000 templates (≈1 fold coverage) and check all sequences against each other for overlaps. Compare observed distribution to theoretical prediction (Poisson distribution).

Sequencing Strategy:

1. genomic DNA broken into 2-kb fragments and cloned into plasmids
2. insert DNA sequenced with universal primer (16,240 reads of 485 bp), 1/2 also with reverse primer (7,744 reads of 444 bp)
3. ≈300 large insert clones (15-20 kb in lambda, λ, phage) sequenced from both ends
4. total seq = 11,631,485 bp ≈ 6.5 fold coverage assembled by computer
5. targeted closure of gaps (after assembly)

3. Assembly Strategy

a) TIGR Assembler. Pairwise comparison of all reads for overlaps (build contigs = contiguous stretches of DNA sequence, 140 contigs, 140 gaps). Paired reads must point to each other and be separated by ≈1 kb.
b) use forward/reverse info to connect contigs (sequence gaps, 98 gaps, primer walking)
c) the remainder are physical gaps (no template, 42 gaps)

4. Gap Closure

a) Sequence gaps filled by primer walking
b) Physical gaps: Specific oligonucleotide primers designed to ends of each contig.

    1) DNA hybridization (Southern blotting) to develop "fingerprint". Genomic DNA cut with restriction enzyme(s) and pieces separated by gel electrophoresis. Then hybridized with labeled primer DNA. If 2 contig ends are close to each other, they should show the same (or similar) fingerprints. Filled 15 gaps.

2) Peptide links. Each contig end was used for a BLAST search against peptide database. If 2 contigs match different parts of the same protein, then they are likely adjacent. Filled 2 gaps.
3) λ clones. The λ libraries (10-15 kb inserts) were screened with the labeled primers. If a contig end hybridized with a λ clone, then the ends of the λ clone were sequenced and this was used to bridge the gap. Filled 23 gaps.
4) PCR. Pairwise combinations of contig end primers used for PCR reactions. An adjacent pair of primers will give a PCR product of the size of the gap between them. Filled 37 gaps (used to verify other methods).

The PCR method was most effective, but becomes much less practical for larger genome projects where there are many more physical gaps. This is because it requires pairwise primer combinations, while the others use one primer at a time. Thus, the other methods scale linearly. Pairwise comparisons increase as $n(n-1)/2$, where n is the number of primers.

Examples:			
	42 contigs	84 primers	3,486 PCR reactions
	420 contigs	840 primers	352,380 PCR reactions
	134,000 contigs	268,000 primers	35 billion PCR reactions

5. Annotation

ab initio gene prediction (GENEMARK) - predicted ORFs based on codon frequency matrices from 122 *H. influenzae* coding sequences in GenBank. Predicted coding sequences were compared with GenBank DNA sequences and translated protein sequences were compared with the SwissProt database.

Example of gene prediction using *H. influenzae* codon usage table:
AUG is not always start codon, it also encodes methionine. Thus there may be several conflicting ORFs that overlap. Which is most likely to be "real"?

Any given string of amino acids is unlikely, but some more than others.

AUG CGG UGU GUC AGG ACC .. =(1) (.0013) (.0047) (.0051) (.0008) (.014)= 3.5×10^{-15}

AUG UUA CGU UAU CCA AGU .. =(1) (.043) (.017) (.033) (.015) (.019)= 6.8×10^{-9}

Thus, the second sequence is over a million times more likely; the log likelihood ratio is ≈ 6

Total predicted protein-encoding genes =1743
736 were unique genes of unknown function
1007 were assigned functional roles based on homology to other bacterial proteins

Conclusion: Whole Genome Shotgun works! (at least for bacteria)