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 The Institute for Genomic Research (TIGR), a non-profit institute led by J. Craig Venter.
Sequenced using the Whole Genome Shotgun (WGS) method.
Venter's grant proposal for the project was rejected. The sequencing was done with private funding from Human Genome Sciences (Biotech company) and Perkin-Elmer (maker of automated sequencing machines, later ABI).

The *H. influenzae* strain that was sequenced is a benign lab strain. For short, it is often referred to as "H. flu", although it is not responsible for the flu in humans nor is it related to the influenza virus. There are some *H. influenzae* strains that cause ear and lung infections (mainly in children) and can cause meningitis. However, this bacterium is not of major medical importance.

Why sequence *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

Of course, these features apply to many bacterial species. An additional reason is that Hamilton Smith (Nobel Prize Winner, Johns-Hopkins University) worked on this species for many years. He discovered type II restriction enzymes while working with *H. influenzae*. Smith is a collaborator on the project and 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?
They sequenced 4,000 templates (≈1x coverage of the genome) and checked all sequences against each other for overlaps. They then compared the observed distribution of how many bases were sequenced *x* times to the theoretical prediction (Poisson distribution). The specific data were not reported in the publication, but were said to be very close to the expectation. Thus, they went ahead with additional shotgun sequencing.

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) to give "paired reads".
3. ≈300 large insert clones (15–20 Kb in lambda, λ, phage) sequenced from both ends
4. total seq = 11,631,485 bp ≈ 6.5x coverage assembled by computer
5. targeted closure of gaps ("finishing", after computational assembly)

Sequencing was done with early-generation automated sequencing machines (ABI 373), which used fluorescent detection and "slab" gels. With 8 people and 14 machines, the sequencing took about 3 months.

3. Assembly Strategy

a) TIGR Assembler. Pairwise comparison of all sequence reads for overlaps (build contigs = contiguous stretches of DNA sequence). Result = 140 contigs + 140 gaps. Required that paired reads point to each other and be separated by ≈1 Kb.
b) use paired-read information to connect contigs (sequence gaps: 98 gaps, filled by primer walking)
c) the remainder are physical gaps (no template available, total = 42 gaps)

4. Gap Closure

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

    1) DNA hybridization (Southern blotting) to develop a "fingerprint". Genomic DNA cut with restriction enzyme(s) and pieces separated by gel electrophoresis, then hybridized with labeled primer DNA. If two contig ends are close to each other, they should show the same (or very similar) fingerprints. This filled 15 gaps.
    2) Peptide links. Each contig end was used for a BLAST search against a protein database. If two contigs match different parts of the same protein, then they are likely to be adjacent. This 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. This 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. This filled 37 gaps (and was used to verify other methods).

The PCR method was most effective, but becomes much less practical for larger genome projects that have many more physical gaps. This is because it requires pairwise primer combinations. Total pairwise comparisons can be calculated 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

Thus, the PCR approach would not be feasible for larger genomes.

5. Annotation

ab initio gene prediction (program = GENEMARK): predicted open reading frames (ORFs) based on codon frequency matrices from 122 *H. influenzae* coding sequences in GenBank. Predicted coding sequences were compared with GenBank and SwissProt databases.

Gene prediction using the *H. influenzae* codon usage table:
The sequence ATG is not always a start codon, it can encode methionine within a protein, or could appear out-of-frame or in a non-coding region. There may be several potential ORFs that overlap. Which is most likely to be "real"?

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

Example: the first ATG is assigned a probability of 1 as it was chosen as the potential start codon. The frequencies of the other codons comes from the *H. influenzae* codon usage table.

ATG CGG TGT GTC AGG ACC ... = (1) (.0013) (.0047) (.0051) (.0008) (.014) = 3.5×10^{-15}

ATG TTA CGT TAT CCA AGT ... = (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 in *H. influenzae* = 1,743
736 were unique genes of unknown function
1,007 were assigned functional roles based on homology to other bacterial proteins

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