The Human Genome

1. Background

3 billion base pairs (3 Gigabases, Gb)
Completion announced in June 2000.
"Draft" sequence published by two independent groups in 2001:
a) Publicly-funded International Human Genome Project (IHGP; Francis Collins, Eric Lander). Published in *Nature* (409: 860-921). Made sequence and annotation freely available to public through GenBank
b) Celera Genomics (Craig Venter). Published in *Science* (291: 1304-1351). Made raw sequence with minimal annotation available on Celera website for free. Charged a subscription for full access to sequence with annotation.

Where did they get the DNA? Whose genome was sequenced?
IHGP - DNA collected from anonymous donors of both sexes and diverse ethnic backgrounds. A subset (≈10%) were used for library construction. Identity is untraceable.

Celera - 21 voluntary donors from diverse ethnic backgrounds, of these 5 were chosen for sequencing (3 women, 2 men), 1 African, 1 Chinese, 1 Hispanic, 2 Caucasian. Later it was revealed that most of the sequenced DNA was from Venter himself.

2. Genome Sequencing Strategy

IHGP - "Clone-by-clone" or hierarchical shotgun sequencing, distributed worldwide. BAC clones of ≈100-200 kb.
total clone sequence = 4.3 billion base pairs. Most clones in draft form (3-5x coverage). About 20% in finished form (8-12x) coverage. Total Raw sequence = 23 billion base pairs.

Celera - Whole Genome Shotgun (WGS) sequencing, done at Celera center
Inserts of 2 kb, 10 kb, 50 kb with paired reads for > 75%. Total sequence = 14.8 billion base pairs ≈ 5x coverage.

3. Assembly Strategy

IHGP - each large-insert clone assembled separately from shotgun reads, then entire genome ordered based on order of previously mapped clones.

Celera - Computational assembly similar to that used for Drosophila: Screener, Unitigger, Scaffolder, Repeat Resolver (Rocks, Stones). But what data did they use for assembly?

a) Whole Genome Shotgun Assembly: Celera shotgun reads + shredded IHGP bactig assemblies ("faux" shotgun reads of 550 bp of perfect 2X coverage of each bactig. Bactig = a contig of assembled reads from within a BAC clone. Due to BAC overlap, this is ≈ 3x complete (not random) coverage.
b) Compartmentalized Genome Assembly: Used above data, but scaffolds and bactigs were first separated into large chromosomal regions, and each region assembled separately. This assembly was used for annotation and analysis.
c) Celera did not do WGS assembly using only their own shotgun data!

"Final Sequence"
IHGP - 2.69 Gb of sequence, 145,514 gaps
Celera - 2.65 Gb of sequence, 116,442 gaps

It is hard to directly compare the two versions of the genome, but by most measures they appear to be quite similar in quality and quantity.

4. Annotation

IHGP - Trained Genscan with known human genes (splice signals, codon usage, exon and intron length) to predict ORFs. Confirmed ORFs with EST or protein match, or with Genie prediction.

Celera - Otto, automated gene prediction software. Uses combination of experimental evidence and *ab initio* prediction. (human or mouse EST, protein database match, mouse genome fragment match).

Total number of human protein-encoding genes (conservative predictions, supported by at least two pieces of evidence): IHGP = 31,778; Celera = 26,588

This is much lower than most people expected. The "standard" estimate was ≈ 100,000 genes, some other estimates based on ESTs were closer to 150,000. The current gene number estimate is around 20,500.

5. Follow up

After the initial publications, the two projects continued to argue about the different assembly methods. In particular, some IHGP researchers criticized Celera's use of IHGP bactigs as "faux reads" in their assembly. See the following papers:

Proc. Natl. Acad. Sci. USA vol. 99 (2002)
- Lander's criticism, pp. 3712-3716; Venter's reply, pp. 4145-4146

Proc. Natl. Acad. Sci. USA vol. 100 (2003)
- Lander's reply to the reply, pp. 3022-3024; Venter's response, pp. 3025-3026

In 2004, Celera published a human genome assembly using only their own data:
Istrail *et al*. (2004). Whole-genome shotgun assembly and comparison of human genome assemblies. Proc Natl Acad Sci USA 101:1916-1921.

In October 2007, the first diploid genome sequence of an individual human (Craig Venter) was published. This was the first time that a single genome (instead of a composite sequence) was available, and the first time that it was possible to identify both alleles (maternal and paternal) present over the whole genome:
Levy *et al*., 2007. The diploid genome sequence of an individual human. PLoS Biol. 5: e254. http://www.plosbiology.org/article/info%3Adoi%2F10.1371%2Fjournal.pbio.0050254

In 2007, 454 sequencing technology was used to sequence the genome of James Watson:
This was Published in 2008: Wheeler *et al*. (2008), Nature 452: 872-877.

Since then, many more "personal genomes" have been sequenced and a "1000 human genomes" project is underway (now expanded to 10,000 or 100,000 genomes).

The ENCODE (Encyclopedia of DNA elements) has attempted to map all functional elements in the human genome sequence. (http://www.genome.gov/10005107)