The Human Genome Fight

A. The Background

In 2001, two independent groups simultaneously published "draft" versions of the human genome. One was a public effort by the International Human Genome Project (IHGP) led by Francis Collins, Eric Lander, Robert Waterston, and John Sulston. The other was from a private company, Celera Genomics, led by Craig Venter and Eugene Myers.

In 2002, members of the IHGP published a paper in Proceedings of the National Academy of Sciences, USA (PNAS) that strongly criticized the methods that Celera used to assemble their version of the genome. In the same issue, the Celera team responded to the criticism and defended their method.

In 2003, the IHGP published their response to Celera's response. This was accompanied by Celera's "response to the reponse to the response".

B. The Exercise

Read the following exchange from the two sequencing groups, then answer the questions below.

1. Waterston et al., 2002. On the sequencing of the human genome. PNAS 99: 3712-3716.
2. Myers et al., 2002. On the sequencing and assembly of the human genome. PNAS 99: 4145-4146.
3. Waterston et al., 2003. More on the sequencing of the human genome. PNAS 100: 3022-3024.
4. Adams et al., 2003. The independence of our genome assemblies. PNAS 100: 3025-3026.

What data did Celera use in their genome assembly?
What are "faux reads"?
What is "shredding"?
What is "perfect tiling"?
What is the "N50 contig length"?
According to Waterston et al. (2002), how does the N50 of an assembly of chromosome 22 with 2X perfect tiling compare to that of 2X (or 5X) random shotgun coverage?
How did Celera's 2X perfect tiling data differ from that simulated by Waterston et al. (2002)?
According to Myers et al. (2002), how does an assembly of 2X tiling of chromosome 22 differ from that of the whole genome?
According to Waterston et al. (2003), what was a major problem with the Celera assembler's simulated reconstruction of chromosome 22 (or the genome) from 2X tiling data?
According to Waterston et al. (2003), how does the N50 of the published Celera genome compare to the IHGP genome or to that expected from 5X shotgun coverage?
What is the "scaffold N50 length"? How does it compare between the Celera and the IHGP genomes?
Which version of the genome had more "ordering errors"?
What aspect of Celera's data allowed for construction of large scaffolds and correction of ordering errors?