Venter et al (2001)

The sequence of the human genome.

***Science*, vol 291, issue 16 February**

General remarks:

- This is a landmark paper and that is the reason why it has been included in the "reading list" for this course
- It is beyond the scope of this course to let you understand all concepts discussed in this paper. Take the following questions as guidelines to make links with the "Bioinformatics Course" , to understand relevant concepts while searching for their meaning in the paper and on the internet or course notes, to get excited about the wealth of information there is out there and the complexity of biological systems ….

1. What is whole-genome shotgun sequencing ?
2. On how many individuals was the "human genome sequencing project" described in the paper based? Do you think this is representative? Do you think it is a problem that ethnicities are mixed when analyzing sequencing data afterwards? Think about the issues involved in population stratification …
3. What is a "contig"?
4. What is meant by gene annotation?
5. Why is knowledge about G+C content important?
6. What do CpG islands refer to? Where do they occur? What is their importance (i.e., is there a correlation with coding exons)?
7. How can genome structure present itself?
8. What is cytogenetic mapping and how can it be useful in learning something about a disease?
9. What is linkage mapping? What kind of study design is needed (family-based design or unrelated individuals)?
10. What is the distance metric in genetics? How is it defined and how is it used in linkage mapping?
11. Is the recombination rate stable across the genome? Is it the same between females and males? What are recombination hotspots? Can you make a link with "haplotype construction" and haplotype association analysis?
12. What are Copy Number Variations (or CNVs)? What are SNPs?
13. What is the SNP rate between chromosomes? What is the consequence for choosing genetic markers in genomewide association studies? The density of the marker distribution achieved with the genotyping platform of your choice will clearly matter. How may it affect the power of your study? What are other factors affecting power? What about multiple testing issues when increasing the number of markers?

14. During genetic association studies, one mostly ignores the dynamic properties of the genome. How can the dynamic nature of genome evolution be captured? What is the difference between gene duplication and segmental genomic duplication? Make a link with phylogenetic analysis and try to understand that DNA polymorphisms only carry a snapshot of the past operation of population genetic factors (name a few of these factors).
15. Name a few common molecular functions. Is the distribution of the molecular functions of the whole human protein set similar to the distribution of the molecular functions of the considered conserved protein set? Is it therefore useful to look at conserved regions when trying to interpret mapping results? Think for instance at the important role of apoptosis in developmental regulation.
16. The importance of interactions: are there good and bad genes? Is it as simple as that?