Questions lecture 2: Genomic variation

1) A) Define common, versus rare and very rare variants.
Refers to the frequency of the minor allele in the human population.
Common variants = minor allele frequency (MAF) >1% in the population (also described as polymorphisms).
Rare variants = MAF < 1%
Very rare variants = MAF < 0.1%

B) Are major blood type variants common and why are there so many distinct alleles?
Yes. Their distribution reflects our pathological past (i.e. exposure to infections). Both clinal (genetically inherited traits that gradually change in frequency from one geographic region to another, they change in clines) and discontinuous (geographic distribution of a trait such that it appears in high or low frequencies in various areas with little or no gradation between them) distributions exist, suggesting a complicated evolutionary history for humanity and their pathological record.

2) What is short read alignment and why does it represent an important computational challenge?
Short read alignment is a process of figuring out where in the reference genome a given DNA sequence, for example representing a short Illumina sequencing read, is from.
This is tricky for several reasons: 1) The reference genome is really big. Searching big things is harder than searching small things. 2) You aren't always looking for exact matches in the reference genome–or, at least, probably not (enormous number of possible matches, and must allow for some mismatches - for SNP calling or technical errors).

3) Which genome of which human population (Caucasian, Asian, Yoruban) tends to harbor the most variation, and why?
Yoruban (African) due to the higher diversity in the African population compared to Asian and European (due to the out-of-Africa bottleneck).

4) How many SNPs are circulating in the "human genome", how many in each human genome? Are most rare or common, clarify your answer.
Around 324 million in dbSNP 150 (doubled thanks to the latest whole genome sequencing studies). In each human, ~3-4 million variants with an average of ~8,500 novel variants.
~50% of all variants are very rare (allele count = 1)

5) What are LD bins? Explain why the HapMap Project deemed these bins useful for inferring missing genetic information, thereby discussing the implicit trade-off in settling for a threshold of 0.8 or greater linkage.
LD (linkage disequilibrium) bins are groups of highly correlated SNPs (inherited together). By knowing the LD structure of the genome, you can impute (infer) missing genetic information. Many variants are in perfect LD (r2=1), so it is cheaper and more convenient to tag some instead of all because by knowing one SNP, one has 100% confidence that the r2=1-linked SNP will be there as well. A threshold of less than 0.8 increases the false-positive rate too much.

6) What is population stratification? Why is this important when linking allele frequencies to disease prevalence, i.e. in GWAS studies?
Separation of the population into different ethnic groups based on differential SNP frequencies. Importantly, the frequency of many SNPs vary a lot between populations, so if by chance the disease group of the GWAS is also stratified compared to the control population, then any SNP whose frequency is naturally higher in the disease group would yield a positive correlation, even though it has nothing to do with the disease. Controlling for population stratification avoids this pitfall.

7) a) What is the major conclusion from the 1000 genome project in terms of how variants impact coding and non-coding function?
All humans have protein truncating variants as well as many missense variants without having diseases. Furthermore, we have fewer variants in the protein coding regions than the non-coding regions.
b) Are we will still evolving?
Yes. A convincing example is lactase persistence. Everywhere dairying started to emerge, we see a parallel selection for mutations that allow lactase to persist throughout adulthood. These are not necessarily the same mutations, but they have the same effect.

8) a) Is the principle of a "meta-genome" still valid? Clarify your answer and be able to discuss distinct mutation scenarios. What are the consequences of this finding and what is mosaicism? b) At which rate do we accumulate mutations and how does this relate to "Peto's paradox"? What key finding was recently made that at least partially solves Peto's paradox.
a) No, single cell sequencing has proven that even within one organ there exist genetic differences between cells. These genetic differences can arise during development or post-natally with early development mutations of course having a much larger distribution resulting in substantial mosaicism. Mosaicism is the presence of 2 or more populations of cells with different variants in one individual. Given that we have much higher variability in our cells than previously thought, this significantly complicates geno-phenotype relationship studies as these mosaic mutations can have large effects on many phenotypes.

b) Interestingly, the accumulation of somatic mutations appears to occur linearly throughout life. The clear consequence is that the chance of picking up a "bad" mutation (e.g. hitting a tumor-suppressor gene) also increases linearly with time. It is puzzling therefore that longer-living animals do not necessarily have a much greater risk to develop cancer than short-living ones (one part of Peto's paradox), However, a recent finding now reveals that the mutation rate is much greater in short- compared to longer-living animals, which explains the lack of cancer risk difference. That said, the other part of Peto's paradox, namely also a lack of correlation for cancer risk with body size, is still unsolved.

9) How is the exome typically captured and sequenced?
After fragmentation of the genomic DNA, the fragments are hybridized to exon-specific probes, thus corresponding to the (selected) coding sequences, so that only exons will be targeted. These are then isolated, sequenced and mapped to the human reference genome.

10) Explain at least three disease gene identification strategies using exome sequencing.

- Sequencing of multiple unrelated and affected individuals.
- Sequencing of pedigrees with multiple affected individuals.
- Sequencing of parent-child trios.
- Sequencing of the extreme phenotypes of a normal distribution.

11) Please explain the difference between the common versus the rare variant contribution to complex disease hypotheses.
Common variants tend to be responsible for common diseases. In reality, common variants usually have smaller odds ratios, but collectively they can induce substantial phenotypic variation.
The summation of low-frequency, high-penetrance variants as drivers of complex traits is driven by the fact that these variants each usually have higher effects on the phenotype (higher odds ratios).

12) Know how to calculate the odds ratio.
OR = odds ratio or $\frac{p_1/(1-p_1)}{p_2/(1-p_2)} = \frac{p_1/q_1}{p_2/q_2} = \frac{p_1 q_2}{p_2 q_1}$,

13) Why is the term "genome-wide" association studies a misnomer? (see also question 5)
Not all variants are tagged by the genotype chips and cannot be imputed afterwards (not all variants are in LD). Rare variants are not tagged at all and cannot be imputed either.

14) What's a Manhattan plot?
A plot showing the obtained p-values (in –log10) for all SNPs included in the GWAS across the whole genome. The skyscapers (hence, "Manhattan") depict highly associated variants with a certain trait or disease.

15) (Patients with = disease) ≠(Patients with = underlying biological disorder). Please explain this statement. How does this relate to personalized medicine?
Some genetic risk loci are also involved in other phenotypes, where they even can be protective. This might make personalized medicine more complicated, as it may mean that treating an individual for one disease by inhibiting a gene found to be responsible for the disease, may increase this individual's risk for another disease instead. Thus, the molecular underpinnings of the same "disease" may be very different between patients, hence why each patient requires a personalized treatment.

16) A. What is the "missing heritability" and where can we find it?
That for highly heritably traits like height, the observed genetic heritability (i.e. by combining all associated SNPs) is still much lower than the expected (which for height is around 80%), thus something is missing in the analysis. By also considering very weakly associated SNPs (thus having a very small contribution to height), we can now account for 68% of the height heritability (based on over 150 thousand subjects), but this is still lower than the predicted 80%. This missing genetic heritability part could be due to rare SVs and SNPs not tagged by the used genotyping chips. Thus, only whole genome sequencing will answer that question. Finally, genetic interactions between variants may also contribute. Of course, the other 20% for height could come from geneXenvironment interactions (see 16B).

B. What is meant by "nature via nurture" or "it takes two to tango"?
Using the example of longevity, only 25% of the observed variation among people can be directly linked to genes, the rest may be dictated by one's lifestyle / environment (e.g. smoking) and the interactions between your genes and the environment (GxE) (e.g. some people eat a lot and unhealthy, yet they never gain weight and remain metabolically fit). To really find genetic determinants, it may therefore be necessary to first take out certain causal factors out of the equation and then perform a GWAS on the remaining variation.