Answers lecture 4: Regulatory variation & Precision medicine

1. Understanding the relationship between a geno- and a phenotype is still a black box. Explain why, and how recent genomic analyses have advanced our understanding of in what way most variants seem to be implicated.
The majority of trait-associated variants (SNPs) map to non-coding regions (which are a great majority of the genome ~95%) and it is therefore difficult to infer the associated function. Among these non-coding GWAS SNPs, the majority maps to regulatory DNA, since they are located in nucleosome-free regions surrounded by active chromatin marks.
This implies that most phenotypic variation that we observe every day or our susceptibility to develop disease is mostly driven by regulatory and not coding variation. So we need to understand therefore how a variant / variants affect molecular networks in a system and as such affect the ultimate trait.

2. Understand the importance of regulatory polymorphisms by discussing the HIV example.
Humans are heterozygous at more functional cis-regulatory sites than at amino acid positions, with 10,700 functional biallelic cis-regulatory polymorphisms in a typical human.
• Case study with the CC chemokine receptor 5, a major chemokine co-receptor of HIV-1 necessary for viral entry into cells. It has been shown that CCR5 promoter polymorphisms determine the density of CCR5 on macrophages and as such the magnitude of HIV-1 propagation.
• G to A SNP of CCR5 at –2459 nt
• CCR5 density – low (homozygous GG), intermediate (GA), and highest (homozygous – AA) (correlates with disease progression, i.e. fastest in AA individuals)

3. What is an eQTL? Describe in simple terms how we detect eQTLS.
eQTL – expression quantitative trait locus/loci - genomic loci that contribute to variation in expression levels of mRNAs. eQTL is a locus that induces a heritable change in gene expression.
Mapping eQTLs is done by testing the linkage between variation in expression and genetic polymorphisms. The only considerable difference with a GWAS trait study is that eQTL studies can involve a million or more expression microtraits.

4. What the GTEx Initiative revealed in terms of how abundant cis-eQTLs are and why they could be useful?
Cis-eQTLs are highly abundant as almost all protein-coding genes can be linked to at least one cis-eQTL in one tissue. They can be useful given that they are enriched in GWAS QTLs, meaning that they can provide a regulatory mechanism as to why the respective variant is linked to a specific trait or disease. This enrichment is however not huge (1-5 fold) indicating that we may still miss important cis-eQTLs, perhaps because this GTEx analysis was performed in a tissue and not cell type specific fashion and of course, we are somewhat missingca time component as some variants may only be active at one specific time during development or cell function.

5. What is a cis- versus a trans-eQTL? Which one of the two tends to be most significant?
Neighboring or overlap of the eQTL and the target gene = *cis*-acting transcriptional regulation or cis-eQTL (or local eQTL). Trans-eQTLs - those that map far from the location of

their gene of origin, often on different chromosomes, are referred to as distant or *trans*-acting eQTLs.
Trans effects are weaker (less significant) than those in cis, but are clearly present.
One also needs many more samples to detect *trans* eQTLs because thousands of SNPs need to be correlated with the expression level of each of 20,000 or so genes, so enormous multiple testing, which needs to be corrected for.

6. Describe in more general terms the nature of these trans-eQTLs (hint: two scenarios from Kreimer and Pe'er paper)
Trans-eQTLs are due to polymorphisms that alter the function (a) or expression (b) of a diffusible factor. The effect can act on many genes at once in trans, so there is no allelic imbalance since both alleles are equally affected. For example, a trans-eQTL can be located in a TF coding gene, or can itself be a cis-eQTL for a TF-coding gene since this would cause a change in TF concentration and hence in the expression levels of target genes of this TF.

In contrast, a *cis* eQTL tends to result in allelic expression imbalance in a heterozygous individual, since one allele will behave differently (in terms of expression) than the other allele (this is called "allele-specific expression")

7. What are tfQTLs?
tfQTLs are variants that induce a heritable variation in TF DNA binding because they alter the binding site for a TF by decreasing or increasing the binding affinity.

8. Explain how the mapping of eQTLs and tfQTLs can aid in uncovering the full molecular chain of causality underlying a decreased risk of myocardial infarction (note: the purpose is not to memorize the names of TFs or genes but rather understand the flow of molecular information). This would be a nice example of an integrative question.
The majority of GWAS hits are found in non-coding regions, thus it can be hard to infer the mechanism and target of these variants, especially since we know (based on chromosome conformation studies) that regulatory elements may affect the expression of genes that are far way (and thus not necessarily next to the respective regulatory element). By overlapping GWAS QTL with eQTL or tfQTL data, you can identify which gene(s) and / or TF binding site a certain GWAS variant is affecting and thus gain more knowledge on the biological pathway / process involved.

In this case, a variant in the 3'UTR of CELSR2 creates a novel TF binding site for the CCAAT/enhancer-binding protein (C/EBP). In liver cells, the binding of C/EBP leads to increased expression of sortilin 1 (SORT1) 40kb downstream of the variant. The increased expression of *SORT1* then further leads to a downregulation of LDL-C, and thus a reduced risk of myocardial infarction.

9. Provide 3 reasons why determining each person's genetic make-up may be important. Provide an example for each.
1. DNA-based risk assessment for common complex disease – for example: the chance of developing breast cancer: women with BRCA1 mutation ~65% chance of developing breast or ovarian cancer before the age of 70.

2. Identification of novel molecular signatures for disease diagnosis, prognosis, or drug design - Type II Diabetes: many possible molecular scenarios why a person developed diabetes. Identifying which is the causal process in every patient may thus result in a more targeted therapy on a per patient basis.
3. A DNA-guided therapy and dose selection - A person's genetic make-up significantly affects the efficacy of a drug: Polymorphisms in the CYP2D6 gene dictates the probability of relapse in women with breast cancer treated with Tamoxifen.

Also, a drug may be as efficient as an existing drug on a whole population level (and may therefore not become approved). However, this same drug may now turn out to be optimal for a specific subpopulation. In other words, as a function on genotype of the patient a drug might be more or less beneficial, efficient or toxic, or present diverse combinations of the above-mentioned phenomena. The response is patient-specific and that is why genotyping is now essential in clinical trials.

10) a) What important insight did recent cancer genome profiling analyses provide us? Explain the underlying concept (hint: the MRCA principle); b) chromotrypsis in a way opposes this insight →explain why by defining what chromotrypsis is and why it is dramatically different from canonical tumorigenesis.
Cancers are genomically diverse and dynamic entities: Unique clones arise because of accumulating driver mutations in the Most Recent Common Ancestor (MRCA) cell progeny. Ongoing linear and branching evolution results in multiple sub-clones which drive disease relapse and metastasis. The dynamic clonal architecture is shaped by mutation and competition between sub-clones given specific environmental selection pressures, including cancer treatments.

Chromotrypsis – phenomenon by which up to thousands of clustered chromosomal rearrangements occur in a single event in localised and confined genomic regions in one or a few chromosomes, and is known to be involved in both cancer and congenital diseases. It occurs through one massive genomic rearrangement during a single catastrophic event in the cell's history and is most common in patients with p53 (the guardian of the genome) mutations. It opposes the conventional theory that cancer is the gradual acquisition of genomic rearrangements and somatic mutations over time.

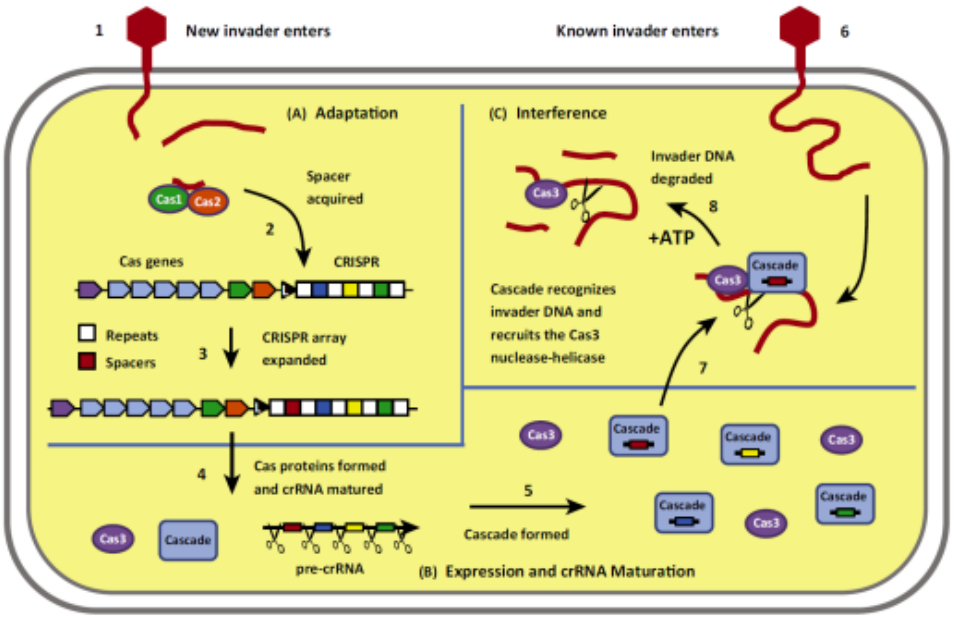
11) Explain the most commonly used gene therapy using bubble boy as an example.
The current most common gene therapy method of immune deficient patients like the bubble boy, involves the introduction of viral vectors expressing the gene of interest (e.g. the gene that is dysfunctional in the patient) into extracted hematopoietic cells (CD34+). These CD34 cells are then injected into the bone-marrow of the patient. The success of this treatment depends on the level of engraftment and proper expression of the gene in the modified cells.

12) Provide at least three possible problems with such kind of gene therapy.
There are several pitfalls.

- The transduction/transfection of the viral vectors may be inefficient resulting in low levels of gene expression.
- The viral vector may integrate into the genome at a random location, causing another severe disease or other unwanted side-effects.

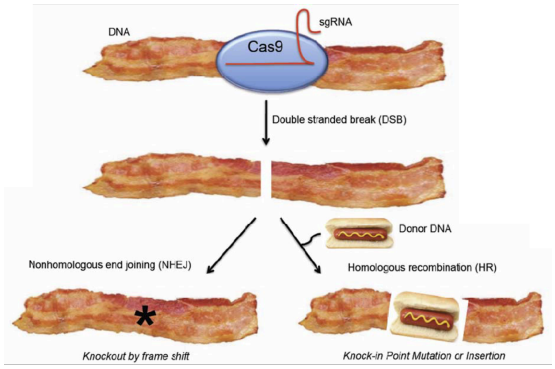
- Transcription of the viral vector may initiate an immune response resulting in the killing of the modified cells.

13) Briefly sketch the natural principle of CRISPR, then explain how it has been co-opted as a very promising genome editing tool. Finally, provide several actual applications of the technology.



CRISPR evolved as an "acquired immunity"-like system in bacteria. Double-stranded DNA from a bacteriophage is cleaved by nucleases into short segments and the latter are then integrated in the bacterial genome in so-called CRISPR regions = DNA loci that contain multiple, short, direct repetitions of base sequences. These regions are then expressed, the resulting RNA loaded in a CAS complex which then allows this complex to be targeted to newly inserted DNA from the same bacteriophage type through the complementarity principle (RNA-DNA recognition). The complex will then cut and destroy the DNA, thus preventing invasion from foreign viral DNA.

CRISPR has been co-opted as a very promising tool due to its low complexity and ease of use. It only requires the Cas9 endonuclease and guide RNA. The only thing you have to change is a short 20 nucleotide sequence in the guide RNA to alter the target of the Cas9. Thus, CRISPR is very efficient, versatile and easy to scale up for high throughput experiments.



CRISPR can be used to introduce indels, large deletions and insertions. Furthermore, by using a de-activated Cas9, the technology can be used to activate or repress gene expression, alter the chromatin modifications at the target site or simply allow for the imaging of specific genomic loci.

14) Describe the two anticipated CRISPR-based therapy strategies:
Ex vivo: cells are removed from patients after which the desired modification / genetic correction is achieved in vitro using CRISPR (which can be entered into cells using a protein mix) after which the "corrected" cells are expanded and infused back in to the patient. This is mostly useful for blood-related diseases (the hematopoietic system)

In vivo: if isolation of the "disease / affected" cells is not possible, then we may consider injecting the CRISPR machinery directly into the body using a viral delivery system. Alternatively, we may use lipid nanoparticles that contain the required CRISPR machinery. The challenge here is to provide these particles with the type of receptors that will allow these particles to specifically home toward the target organ (in the example, the liver for example).

Early embryo: the most radical strategy is to directly intervene at the level of the developing embryo or even in germline cells (e.g. fertilized oocyte) to correct "parental" mutations such as the resulting embryos are "clean", i.e. their cells do no longer contain the faulty gene.