Supervisor: Yihong Jennifer Tan
Gaëlle Spack
Melina Lopez Sciarra

Identification of disease-associated processed pseudogenes

What is our project about?

The aim of our project is to investigate 36 different processed pseudogenes that are likely associated with trait-linked genetic variants. First of all, all SNPs (common SNPs and GWAS[1] SNPs) contained in the nearby regions (± 100 MB) of each pseudogenes will be analysed with a eQTL[2] in order then to determine which GWAS SNPs are correlated with a pseudogene and which traits they cause in association. In the second part of the work, the interaction between pseudogenes and their parental mRNAs will be investigated focusing on pseudogenes working as microRNA decoys.

Background

Processed pseudogenes

Pseudogenes are derivative of coding genes. Processed pseudogenes are a class of pseudogenes that result of a reintegration of reverse transcribed mRNA in the genome. Their signatures are that they only contain exons and a poly-A tail typical of mRNA. They have lost their ability to code protein generally due to an accumulation of mutations. Therefore they are generally viewed as "junk" DNA, but recently several indications suggest that some pseudogenes can act as regulatory elements.

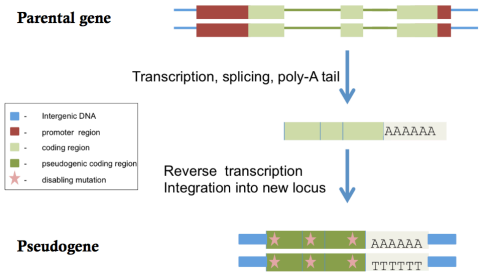

Figure 1: Creation of processed pseudogenes

MicroRNAs

MicroRNAs (miRNA) are a class of small non-coding RNAs that play a key role in regulation of gene expression by inducing gene silencing when binding to the target mRNA. They act at the post-transcriptional level. MicroRNA are short and single stranded RNA molecules of approximately 22 nucleotides long. They contain a seed sequence, from 2 to 7 nucleotides, through which they bind to sequence complementary sites with mRNA to regulate their expression.

One example of crosstalk regulation between pseudogene and its parental gene is the couple gene/pseudogene, PENT/PTENP1. As PTEN and PTENP1 contain a highly conserved region and because PTEN-trageting miRNA seed matches within the high homology region are conserved, PTENP1 play a role of protection for PTEN by attracting the miRNAs that normally would silence the gene. When studying this particular interaction, it has been shown that PTENP1 has a tumor suppressive activity and that when PTENP1 is lost it favours cancer.

Figure 2 : MicroRNA-dependent crosstalk regulation between pseudogene and its parental gene

[1] GWAS: Genome Wide Association Study is an analysis of common genetic variants coming from different individuals, in the purpose to see if a variant is associated with a trait. These kind of analysis generally focus on an association between SNPs and a disease.

[2] eQTL: Quantitative Trait Loci are sequence in the genome containing genes coding for traits or phenotypes. The eQTL analysis helps to understand the genetic traits behind a disease and the regulatory networks (which gene affects the expression on another gene).

Methods

Part 1: Identify GWAS SNPs associated with pseudogenes in cis-eQTL

The data used in our project come from GEUVADIS consortium[3]. These data, consist of lymphoblastoid cells (LCLs) derived from 373 samples of European individuals, provided us the RNAseq for the pseudogenes. The SNPs genotype was obtained from the 1000 genomes project.

The first step was to create the cis-eQTL region for each chromosome, by adding and subtracting 1MB from the transcriptional start site (TSS) of the pseudogene. The purpose was to identify eQTLs likely influence pseudogene expression in cis.

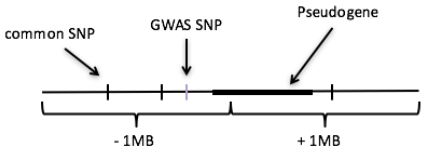

Figure 3: Cis-eQTL region

Correlation tests were performed between each pseudogene and all their nearby SNPs in the cis-regions, followed by a 1000 times permutation of the expression levels of the pseudogenes to determine the significance of the correlation. eQTL with p-value less than 5% ($p < 0.05$) were considered to be significant.

Next, significant pseudogene eQTL where SNPs were previously associated with complex traits (based on GWAS catalogue) was compared to all common SNP eQTLs to select for GWAS eQTLs with the highest impact on pseudogene expression. This was done using the Regulatory Trait Concordance (RTC) test. This allowed us to identify pseudogene significantly associated with complex traits.

Part 2: interaction between pseudogenes and their parental mRNAs – microRNA

Next, we investigated whether trait-associated pseudogenes can regulate their parental mRNAs through a microRNA-dependent mechanism. To do this, we first identified parental gene of the processed pseudogenes by blasting the sequence of the pseudogenes using the BLAT tool (UCSC genome browser). We then extracted the following information from the search: the sequence of the pseudogene, the identity of the parental gene and sequence of its 3'UTR (since microRNA are known to bind to 3' UTR of the mRNAs). The second step was to find the miRNA that may be involved. To do so, names and seed sequences of B cell (the most similar cell line to LCLs) expressed miRNAs were obtained from the database microrna.org. Next, miRNA binding sites within the pseudogenes and 3' UTRs of their parental genes were predicted using TargetScan.

Finally, we determined if the miRNA seed sequence was found in both the pseudogene sequence and the 3' UTR of the parental miRNA, which should be the case if the pseudogenes to work as miRNA decoys.

Results

Part 1: determination of GWAS SNPs associated with pseudogenes investigated.

Pseudogene		GWAS SNPs	Correlation coefficient	P-value	RTC	Trait
ENSG00000237624.1	OXCT2P1	rs3916164	0.2937	0	0.98736	- Red_blood_cell_traits
ENSG00000224497.1	RPL36P4	rs6060373	0.2768	0	0.9824	- Height
		rs6088813	-0.2815	0	0.9985	- Height
		rs224333	0.2817	0	0.93862	- Height
		rs6141600	0.1844	0.00033	1	- Height
		rs143384	0.2527	2.9810^{-6}	0.9837	- Height
		rs6060369	0.2805	0	0.9533	- Height

Table 1: Extract of pseudogenes associated with their GWAS SNPs causing a disease

[3] GEUVADIS consortium: Genetic European Variation in Health and Disease

For the purpose of the report, we investigate into two eQTL pseudogene in detail. First of all, it is known that GWAS SNPs can impact the expression of protein coding, so the GWAS SNPs may be associated with a trait. The purpose of the project is to see if GWAS SNPs[4] can also impact on the expression of the pseudogene.

On this table (Table 1), we can see that there is a high correlation between the pseudogene and the GWAS SNP, that the p-value is low, and that the RTCs are high, close to 1. We can also notice that pseudogene RPL36P4 has six GWAS SNPs that can be associated with the same trait. It is more likely the pseudogene is associated with the trait. It would be interesting to test whether the pseudogene has a functional role in the development of these traits.

Part 2: Interaction between pseudogenes and their parental mRNAs – microRNA

Pseudogene		Parental gene	Predicted shared miRNA binding sites	Supporting literature of miRNA involvement in eQTL trait
ENSG00000237624.1	OXCT2P1	OXCT2	miR-24; miR-33a	GATA factor-dependent genetic networks that control red blood cell development. - **miR-24** is important for erythropoiesis [pubmed: 24049083] - **miR-24** Is required for hematopoietic differentiation [25634354]
ENSG00000224497.1	RPL36P4	RPL36	None	-

Table 2: Extract of results

For some pseudogenes (i.e. OXCT2P1, Table 2) seed sequence of miRNAs were found in both pseudogene and parental gene. Because they share predicted miRNA binding sites, it is possible that the pseudogene acts as a miRNA decoy for the parental gene by competing the same pool of miRNAs. Only experimental work could confirm this prediction.

An interesting point to notice in the results is that miR-24 (Table 2) is reported to control red blood cell development, which related to the phenotype of OXCT2P1 eQTL, which is also in red blood cell traits. The hypothesis might be that it is through competing for miR-24 with its parental gene, and perhaps with other mRNAs, that the pseudogene is implicated in red blood cell traits.

On the other hand, it happened that no miRNA was predicted to be shared between the pseudogene RPL36P4 and its parental mRNA, RPL36 (Table 2). Two different hypotheses can be made. First it could be because the pseudogene is unlikely to act as a miRNA decoy for its parental mRNA, or it could also be that because the miRNA come from B-cells and not from Lymphoid cells this leads to the fact that some information is missing.

Conclusion

To conclude two main take home messages:

1) Pseudogenes in eQTL with complex trait-associated SNPs were identified and may have roles in human traits.
2) Some pseudogenes could act as miRNA decoys and regulate their parental gene expression.

[4] A GWAS SNP is a common SNP that has been previously reported to be associated with complex traits.

Bibliography

- Wikipedia
- https://www.quora.com/What-is-eQTL-Expression-quantitative-trait-loci-and-why-is-it-important-to-study-them
- https://en.wikipedia.org/wiki/Genome-wide_association_study
- http://www.geuvadis.org/web/geuvadis;jsessionid=1ECB0F347A857C2379AFCCBB8949B730
- http://www.exiqon.com/what-are-microRNAs
- L.POLISENO, L.SALMENA, J.ZHANG, B.CARVER, W.J. HAVEMAN AND P.P. PANDOLFI. (2010) A coding-independent function of gene and pseudogene mRNAs regulates tumor biology. *Nature*. 465, 1033-1040.