

Fine Scale Regulatory Annotation of Cancer Genes

29 June – 7 August 2009

1. Introduction

The purpose of this 6-week project is the fine scale regulatory annotation of a list of genes that have been specifically shown to be active in tumour (cancer) cells – we will refer to these as “cancer” genes. Understanding under what circumstances these genes get turned on, that is, what factors regulate the transcription of cancer genes is of utmost importance in cancer research. Bioinformatics provides computational tools that facilitate the discovery of such factors.

During this project you will have a chance to familiarize yourself with the wide range of methods available for locating transcription factor binding sites (TFBSs) and you will be required to utilise some of these to characterise as precisely as possible the regulation of a number of genes that are recently published targets of cancer biology.

We will provide you with a list of references, papers, results of past related projects and you are always welcome to ask questions or request help on how to proceed with the analysis but we also encourage you to use your own ideas and communicate with all the other project members and work on solutions to arising problems together.

Time permitting you can also get your share in the development of our group’s own, novel tool for TFBS detection by comparative analysis, called BigFoot [3].

2. Background

2.1. *Cancer biology*

Cancer is a class of diseases in which a group of cells display uncontrolled growth (division beyond the normal limits), invasion (intrusion on and destruction of adjacent tissues), and sometimes metastasis (spread to other locations in the body via lymph or blood). These three malignant properties of cancers differentiate them from benign tumors, which are self-limited, and do not invade or metastasize.

Nearly all cancers are caused by abnormalities in the genetic material of the transformed cells. These abnormalities may be due to the effects of carcinogens, such as tobacco smoke, radiation, chemicals, or infectious agents. Other cancer-promoting genetic abnormalities may be randomly acquired through errors in DNA replication, or are inherited, and thus present in all cells from birth. The heritability of cancers are usually affected by complex interactions between carcinogens and the host’s genome. New aspects of the genetics of cancer pathogenesis, such as DNA methylation, and microRNAs are increasingly recognized as important.

Genetic abnormalities found in cancer typically affect two general classes of genes. Cancer-promoting oncogenes are typically activated in cancer cells, giving those cells new properties, such as hyperactive growth and division, protection against programmed cell death, loss of respect for normal tissue boundaries, and the ability to become established in diverse tissue environments. Tumor suppressor genes are then inactivated in cancer cells, resulting in the loss of normal functions in those cells, such as accurate DNA replication, control over the cell cycle, orientation and adhesion within tissues, and interaction with protective cells of the immune system.

2.2. *Gene regulation*

Transcriptional regulation plays a vital role in the development and functioning of each cell of any organism. It is the key link between components of regulatory networks that ultimately determine

the expression level of each individual gene in a cell. Several genetic diseases have been shown to be caused by a defect in gene regulation [1, 2].

There are a number of different factors involved in the regulation of transcription. Some of these work by binding to the RNA polymerase itself and some do bind to the DNA chain, to specific sites upstream, downstream or at introns within the gene's coding sequence. In this project we are not addressing protein-protein type interactions but rather concentrate on detecting binding sites that could potentially play a role in the regulation of a target gene (i.e. cis-regulatory elements), including binding sites of transcription factors, activators, repressors and – within the promoter – of the RNA polymerase itself.

3. Methods

One approach to identifying putative cis-regulatory elements is through methods that look for known motifs. These motifs are a few base-pair long consensus sequences of experimentally verified binding sites usually given in the form of position-specific weight matrices (PSWMs). A number of tools exist that can perform such a search.

Another approach based on completely different principles is the comparative one, which has the advantage of being able to predict novel binding sites [3, 4]. The key observation that lies behind this family of methods is that sites involved in regulation are significantly more conserved across closely related species than other non-coding regions lacking biological function. By detecting highly conserved sites in a homologous set of sequences it is possible to identify most of the regulatory elements. Of course, binding sites can appear and disappear over time, so the success of the analysis largely depends on the choice of the species involved.

Genes are also analysed by approaches that can go in at different places in this ladder:

A. Non-homologous signal search. It could be a major advantage to allow non-homologous analysis. This provides a very different framework, since tools such as alignment and evolutionary model cannot be used. However, such comparison provides important additional information since signals can be common for functional reasons and this can only be revealed by comparing non-homologous genes. Lawrence et al. (1993) provided a non-homologous method. Wang et al. (2003) combined homologous with non-homologous methods, but in a non-statistical fashion. For a suggestion of how to do this in a more statistical fashion, look at the project description “A Unified Approach to Signal Detection”.

B. Functional analysis of a group of non-homologous genes. Grouping genes as co-regulated, having the same function (for instance according to Gene Ontology) or according to some other criteria is important in formulated hypotheses about the genes.

1) can be done if the appropriate data is available (which it normally is in the form of many closely related genomes), but still with some error. It might seem hopefully unambitious, but is in reality very valuable due to the “needle in haystack” problem that genomic analysis present as what is functionally important for a given problem can be a very small fraction of the complete genome.

No biologist believes methods promising to make predictive dynamic models from the genome). But even attempting to devise models that can do 4) can be valuable, since this is what a biologist will try to do manually. In several cases (2-4) it is possible to sketch a method, but there presently exists no programs that can be used.

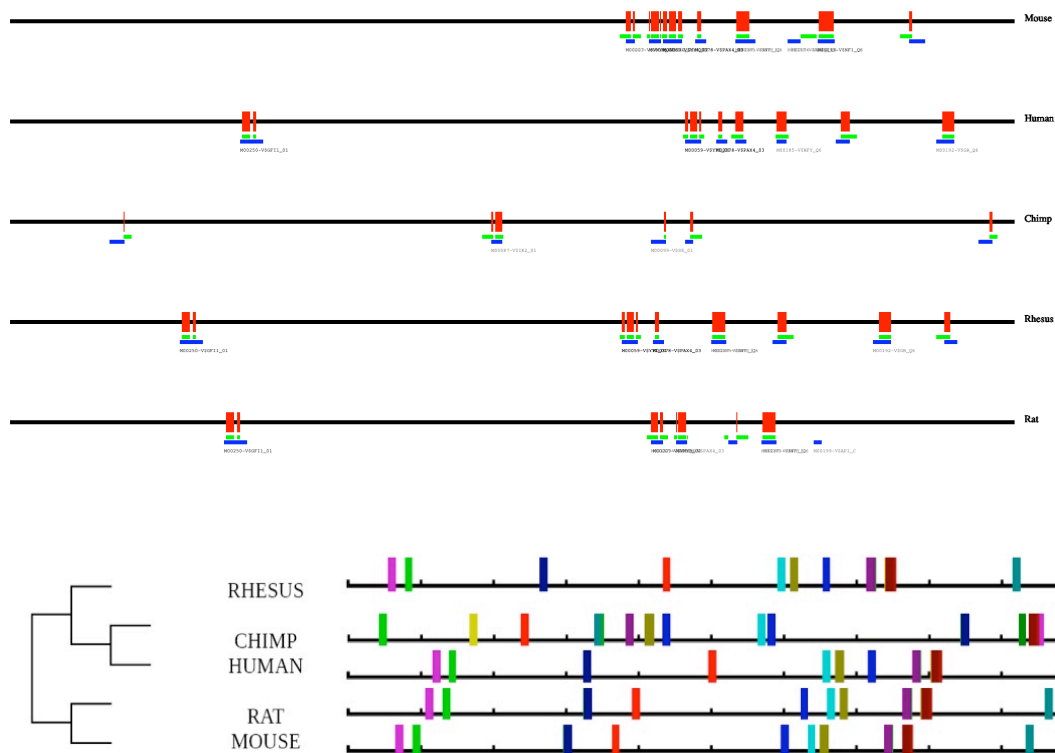


Figure 1: Found conserved regions (motifs) with FootPrinter for ORM DL2 upstream nucleotide sequences. Different colours indicate different motifs. The phylogenetic tree next to the motifs was provided by FootPrinter. There are consistent motifs, which means that the same series of these motifs can be found in all five ORM DL2 sequences. However, there are many artifacts as well, that cannot be highly conserved sequence part. (FootPrinter 3.0, motif size: 11, maximum number of mutations: 3, maximum number of mutations per branch: 2)

3.1. Using BigFoot

Given a selected gene, your task is to identify all of the conserved motifs in the upstream region that could potentially correspond to a binding site and thus take part in the regulation of the gene.

1. Select a gene of interest. This could be any gene that is related to cancer.
2. Download the upstream region of the selected gene and its homologs from 4-5 species that are relatively closely related to the target one. The length of the upstream sequences should be at least 1000 for a human and 300 for a bacterial gene. Collect all your sequences in a FASTA file.
3. Download BigFoot, a recently developed tool for TFBS detection from <http://www.stats.ox.ac.uk/~satija/BigFoot/>
4. Run BigFoot on your data, making sure you add your tree beforehand. The default number of MCMC cycles is 1 million which is usually needed to get optimal results. Note that several hours of computation will be required so leave enough time for the analysis to finish.
5. Use the MPD view to identify highly conserved motifs: the red curve will show the level of conservation at each site and the candidate TFBS motifs will be written in capitals.
6. Ideally you will get short conserved regions embedded into long variable regions. However, if too much or too little of the sequences is conserved you chose too closely/distantly related species, so you will have to replace some of your sequences and re-run the analysis.

3.2. Other methods

When you have successfully identified a few putative regulatory motifs try to compare your results to experimentally verified data. Find regulatory databases for your chosen target species – for human sequences, CisRED is a good place to start, for bacterial genes EcoCyc is recommended.

You should try PSWM-based tools such as MotifScanner to search for known binding sites in your sequences and other methods that detect statistically overrepresented oligonucleotides [5].

4. Project details

4.1. Genes of interest

We received a list of 'interesting' genes from Dr Thorunn Rafnar (our Icelandic collaborator who is an expert in cancer biology) together with a list of recently published papers on them. You can download the pdf files of the papers from here:

<http://phylogeny-cafe.elte.hu/OxfordSummer2009/>

The first suggestion is the TERT gene - it is one of the most important "cancer" genes. It is turned on in the majority of cancers and variants in the region are associated with predisposition to multiple cancers. See Hung et al.

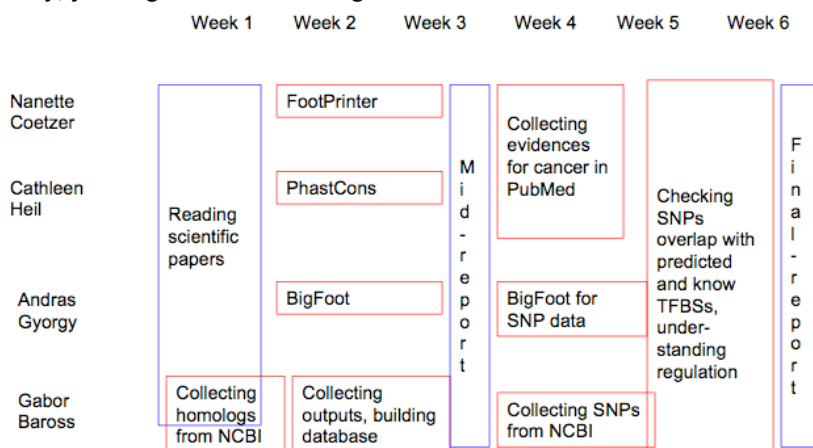
Then we could look at EGFR2 which is often amplified or mutated in cancer and has been used as target for focused cancer treatment. Being a tyrosine kinase receptor, it is only present in multicellular organisms, unlike TERT. (Check it out on Wikipedia http://en.wikipedia.org/wiki/Epidermal_growth_factor_receptor) Another gene is FGFR which was mentioned in a review paper (Rafnar, Nat. Gen. 2009)

Then there is probably a polymorphism in the estrogen alpha receptor that is associated with risk of breast cancer - here again is a very important "cancer" gene. See Zheng et al, 2009.

Then there are the pigment genes that have polymorphisms that are also associated with skin cancer (ASIP, Tyrosinase) – these could be very interesting because the frequency of these SNPs varies greatly between ethnicities and also on a north-south gradient within Europe. See Sulemet et al. and Gudbjartsson et al.

4.2. Project timing

After carefully analyzing the background of each student participating in this project, hereby we suggest a division of work together with project timing. However, the division of work is not mandatory, you might revise it with a good reason.



References

- [1] M. F. Moffatt et al. Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma. *Nature*, July 2007.
- [2] Hung et. al. A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25. *Nature*, April 2008.
- [3] R. Satija, I. Miklos, A. Novak, R. Lyngsø, and J. Hein. BigFoot: Bayesian alignment and phylogenetic footprinting with MCMC. *BMC Evolutionary Biology*, 2008. (submitted).
- [4] A. Siepel, G. Bejerano, J. S. Pedersen, A. S. Hinrichs, M. Hou, K. Rosenbloom, H. Clawson, J. Spieth, L. W. Hillier, S. Richards, G. M. Weinstock, R. K. Wilson, R. A. Gibbs, W. J. Kent, W. Miller, and D. Haussler. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res*, 15(8):1034–1050, August 2005.
- [5] S. Sinha and M. Tompa. Discovery of novel transcription factor binding sites by statistical overrepresentation. *Nucleic Acids Research*, 30(24):5549–5560, 2002.