Investigating the relationship between somatic mutations in breast cancer and oestrogen receptor alpha binding sites

In this exam, you will be asked to investigate whether there is an association between locations in the genome where a specific transcription factor binds and genetic variants present in breast cancer cells, using what you have learned in the *Statistical Genomics* module of the course. You should follow the instructions under Part 1 and Part 2 in this document, and create a Galaxy page in the Genomic Hyperbrowser that will make it possible for someone else to reproduce your results. Also, make a short presentation (15-20 minutes) for the oral exam, where you explain the conclusions and results of your analysis and an overview of how you reached those conclusions, including reasons for choices you made (e.g. why you chose a specific test statistic or null model). You will not need to include such discussions and reasoning in the Galaxy page, as the main aim of the Galaxy page is to make your analysis reproducible. At the beginning of the oral exam, you will be asked to share the link to the Galaxy page you created. After the presentation, you will be asked further questions related to the analysis you have done. You will not need to make any written report other than the Galaxy page.

You will be evaluated based on the “Learning outcomes” specified for this module (see slide 4 in the lecture slides). This means that we will measure to what degree you have achieved the learning outcomes. We will do this based on your presentation, the Galaxy page you create and your answers to the questions we ask during the oral exam.

The following background section will provide you with the necessary biological knowledge that is required in order to solve the tasks given in this exam. You may need to use the information given here in order to make reasoned choices when doing the analysis. You are not required to read or use any other sources (e.g. articles) to make such choices.

If you have problems using the Genomic Hyperbrowser (e.g. cannot find a specific tool or how to do something specific), please contact us.

Background

Oestrogen Receptor Alpha (ESR1) is a transcription factor that is activated by the oestrogen hormone. Transcription factors, such as ESR1, are proteins that bind to specific locations in the genome. When binding, they might affect gene regulation, by upregulating nearby genes. A transcription factor typically binds to thousands of places in the genome and an experimental technique known as ChIP-seq can be used to estimate where the transcription factor binds.

These estimated binding locations are generally referred to as “peaks”. In this task, you will be provided a bed file containing peaks for ESR1 in a breast cell. It is assumed that the binding of a transcription factor to a specific location in the genome can be disrupted by a change in the genomic sequence at that location. This means that a change in underlying sequence (e.g. a SNP) can result in the transcription factor not binding to that position, and in turn result in a gene being down-regulated. We assume here that a change in genetic sequence anywhere within a predicted peak might affect binding to that area (in reality, predicted peaks are wider -- typically 100-200 base pairs -- than the exact binding site, which is typically 5-15 base pairs, but here we only know that the exact binding site is somewhere within the peak).

A cancer cell will have genetic variants (changes in the genomic sequence) that are not present in “normal” cells in the body. Such variants are known as *somatic mutations*. Here, we are interested in whether known somatic mutations in breast cancer tend to occur at places where ESR1 binds. The idea is that somatic mutations might alter the sequence leading to the transcription factor not binding less than it normally should. We will use a set of known somatic point mutations in breast cancer, i.e. mutations that are found to be common in breast cancer based on variant calling on cancer cells from many breast cancer patients.

Part 1

Use the Genomic Hyperbrowser at https://hyperbrowser.uio.no to investigate whether somatic mutations in breast cancer tend to co-localize with predicted peaks for ESR1. Use the two datasets found in this Galaxy history: https://hyperbrowser.uio.no/hb/u/ivar/h/home-exam-data-sets (you can import the history and continue your analysis with it as a starting point)

You should define and use a hypothesis test for determining whether there is an association or not. State and discuss the assumptions you make. You are encouraged to use descriptive statistics to describe and understand the data you are given. For this, you can use any appropriate tool, such as the Genomic Hyperbrowser, BEDtools or your own methods/scripts.

Part 2

In the previous part you investigated whether ESR1 co-localizes with somatic mutations in breast cancer more than expected by chance, by using some null model.

It can be useful to also check whether ESR1 co-localized with the somatic mutations more than what other transcription factors do. To do this, we will collect predicted peaks for other transcription factors for the same cell line that the ESR1 peaks were predicted on (T-47D, which is a breast cell line).

Create a GSuite file with peak data sets for transcription factors on the cell line *T-47D* by using the tool “Create a GSuite from an integrated catalog of genomic datasets” in the Genomic Hyperbrowser. Choose broad peaks. You should get 19 data sets using this tool, keep all of them.

Use the resulting GSuite file to investigate how the somatic mutations co-localize with transcription factors in T-47D. Use these results together with what you found in Part 1 to conclude on whether you believe there is a clear and significant association between locations of somatic mutations in breast cancer and ESR1 binding sites or not.

Tips and things to be aware of:

- Make sure you have created a user and that you are signed in when doing the analyses in the Genomic Hyperbrowser. This is necessary in order to keep the results and create a Galaxy page.
- If you are struggling with navigating the Genomic Hyperbrowser, e.g. finding a tool you are looking for or where to find results, please contact us. We don’t want you to waste time trying to find the correct tool doing what you want to do.
- If using Monte Carlo simulation in the Genomic Hyperbrowser, it should be sufficient to choose “Moderate resolution of global p-value” as Monte Carlo false discovery rate sampling depth. You can also choose “Fixed (10 000 samples)”, but that will be slower.
- All the data sets provided are on reference genome hg19.
- The implementation of Galaxy Pages contains a bug complicating the creation of pages (resulting in content not being displayed in the output). In our experience, some simple measures avoids the issue: First, when beginning a page, add several blank lines and end the document with a period: "." After having embedding histories or other elements, make sure to manually jump to the next pre-entered line (by clicking) instead of creating a new line from the end of the line with the embedded content.
- If you use the tool “Determine GSuite tracks coinciding with a target track”, you should select to include “target” in the “Select track attributes to display with the results page” section to get the transcription factor names in the output table. The target called “ERalpha_a” is equivalent to the ESR1 transcription factor.

Good luck!

Questions can be emailed to ivargry@ifi.uio.no.