## Whole-genome patterns of nucleotide diversity & recombination in the model legume *Medicago truncatula*

Upload your report with max 2p of text (excluding tables and illustrations) and the R script you used to do the analysis **by December 12th 2017.**
Note that for administrative reasons, this deadline is final and non-negotiable and returning your assignment is mandatory to go to the exam.
Please, **Mark clearly your name and student ID number on both documents (R script and report).**

***Background:*** *Medicago truncatula* is a model plant for investigating legume genetics, including the genetics and evolution of legume–rhizobia symbiosis. You will explore a dataset where whole-genome sequence data was obtained on a diverse sample of 26 *M. truncatula* genotypes. The goals of this assignment are:

1. To produce a number of graphs to visualize patterns of genomic diversity and recombination,
2. Report chromosome-wide estimates of diversity and recombination together with appropriate measures of statistical uncertainty (Standard Errors, Confidence Interval)
3. Test a couple of well-defined hypotheses that make a priori predictions how the genomic environment affects the local amount of diversity in the genome. In doing so we will check some of the assertion made by the authors in their paper (see below)

DATA Source: Branca et al PNAS 2011 (> 130 citations)
http://www.pnas.org/content/108/42/E864.abstract

### 1. Import and filter the dataset in R.

Read about the metadata: how is the data organized, what do variables in the data mean roughly. ASK questions on the forum if needed.
How many chromosomes, how many genomic windows are available in the dataset?
Are obvious windows that are "big" outliers?
**Choose a subset of the dataset by excluding windows that have very few SNPs (LdH_SNPs) and/or a window size that is very small (examine the distribution of ldH_win_size).**
**Check what percentage of the windows you keep / discard.**
**Keep that filtered subset fixed for the remaining analysis.**

### 2. Displaying genomic data

Make a graph that illustrates patterns of variation in recombination (as measured by rho) and polymorphism (as measured by qp.site) along chromosome 5 (see for instance Fig 4 of the paper).

(NB: the data displayed above is taken Branca et al PNAS 2011, and the one you analyze were filtered in a slightly different way so minor discrepancies might exist in the graphs you do and the ones reported in the paper).

### 3. Summarizing levels of Recombination and polymorphism in M. truncatula

Use a boxplot or any other graphically more convenient way to illustrate the distribution of nucleotide diversity (qp.site) and

recombination rate (rho per kb) among the different chromosomes of *M. truncatula.*
Are means or median better summaries for these variables?
For each chromosome calculate the (mean) (or median) recombination rate and its associated 95% CI. Present these results as a table
The authors state that "The population-scaled recombination rate is approximately one-third of the mutation rate, consistent with expectations for a species with a high selfing rate." Using the ratio calculated by the authors (rho_theta) discuss if you support that assertion.
Is there a lot of variation between chromosomes for this ratio?

**4. A test of an evolutionary hypothesis**
One of the conclusion of the paper made by the is that "Nucleotide diversity in 100-kb windows was negatively correlated with gene density, which is expected if diversity is shaped by selection acting against slightly deleterious mutations. " We want to reproduce analysis and discuss if our analysis support this conclusion?

The so called "background selection hypothesis" states that genomic regions that are experiencing more deleterious mutations may exhibit overall less polymorphism because selection as it "weeds out" delteriousmutations also eliminates other SNP variation in their vicinity.
This hypothesis makes the prediction that "everything else being equal":

- Gene-dense regions (because they are more prone to produce deleterious mutation) should be exhibiting comparatively less polymorphism than regions that are gene poor.
- Background selection will have more influence in regions that have a low recombination rate.

So far testing this hypothesis and its predictions was not easy (except for a few model species) because it requires enormous amounts of polymorphism data. We want to test that hypothesis in *M truncatula*.

**4.1** Formulate expectations for what genomic covariate will explain variation in nucleotide diversity (qp.site) of a window.
Using linear models, explore if there is support for the predictions of the background selection hypothesis in the patterns of M truncatula genetic diversity.

**4.2** Are all chromosome of the M truncatula genome exhibiting similar patterns ?
(Explore this visually and using appropriate linear models that include a chromosome factor and how chromosome can "interact" with the factors of your anlaysis)
Can other covariate potentially confounding the background selection hypothesis predictions?

**5. Tracking footprints of recent and intense selection in the *Medicago truncatula* genome.**

**5.1** According to the so-called "selective sweep" scenario, if a region contains a mutation that was recently selected, this recent

and intense selection is likely to wipe out most of the polymorphism in that region.
Do you find footprints for a recent major selective sweep as evidenced by 2 or more consecutive windows with very low diversity (qp.site) relative to the rest of the genome?

**Hint:** explore graphically where the least polymorphic windows (use the 5% least polymorphic windows within the genome) are located.

**5.2** Is the location of least polymorphic windows randomly distributed on chromosomes?

**Hint:** To make such a test, you can group windows into larger segments comprising 20 windows and count the number of windows containing least polymorphic windows in each segment. "On average" you expect 1 of such 5% least polymorphic windows per segment.
Think about what probability distribution you expect can capture the fact that windows with low level of polymorphisms are occurring "at random" along the chromosome in each segment. Then decide on a statistic you can use to test the null hypothesis that the "location of least polymorphic windows is randomly distributed along the chromosomes". Get a null distribution for such test statistic and state whether you reject the null or not for each chromosome by discussing the p-values you obtain.

**5.3** Another typical footprint of a selective sweep is, besides having lower polymorphism, the fact that genomic regions that have experienced a recent selective sweep are expected to have an excess of rare variants (i.e. the only mutations that contribute to polymorphism are recent and therefore still in low frequency). The excess of rare variants in a window can be measured through the summary statistics "Tajima's D". Negative Tajima's D values mean an excess of rare variants (SNPs) while positive Tajima's D means an excess of frequent SNPs relative to a neutrally evolving region.
Are windows exhibiting abnormally low amounts of polymorphism —as measured by qp.site are also having more negative Tajima's D values? Use some statistical inference method to examine the relationship. Could other confounding variable obscure this relationship?