

Exceptional Longevity Exercise

August 3, 2010

In a July 2010 Scienceexpress publication (10.1126/science.1190532), Sebastiani et al. reported the results of a genome-wide association study (GWAS) of 1055 centenarians and 1267 control subjects. Based on the results of the study they built a genetic model of 150 single-nucleotide polymorphisms (SNPs) that could predict exceptional longevity (EL, as opposed to Average Longevity, AL) with 77% accuracy.

In this exercise, you are going to calculate the chance of living beyond 100 years old. Please note: the results of this calculation are based solely on the method described in the above paper, which has been under quite a bit of scrutiny for its claims of predictive value. Please take this to be an exercise in working with genomic data rather than a gold-standard prediction of your chances at longevity!

From the study the authors ordered the SNPs in descending order of Bayesian significance with respect to predicting EL. They then built nested SNP sets, starting with just the first SNP, then the first 2 SNPs, ..., up to all 150 SNPs in the last set, with the full set to be the best predictor of EL. We have implemented their Bayesian model in a python script. The model uses the prior probabilities of having each SNP variant given EL or AL, and starts with the prior probabilities $P(EL) = P(AL) = 0.5$. The main equation is given in Supplemental Figure S7 as:

$$\mathbf{K\ SNPs\ set\ } \Sigma_k = (\text{SNP}_1, \dots, \text{SNP}_k) \\ P(EL|\Sigma_k) = \frac{P(EL) \prod_{i=1}^k P(\text{SNP}_i|EL)}{P(EL) \prod_{i=1}^k P(\text{SNP}_i|EL) + P(AL) \prod_{i=1}^k P(\text{SNP}_i|AL)}$$

Using the script:

1. You may use your own genotype file, or one of the files downloaded from <http://gene210.stanford.edu/data/Genomes/>.
2. Open the Terminal application (Applications - Utilities - Terminal), and navigate to where you have place your genome file(s).
3. You should still have a link the the GENE210 scripts directory, gene210-scripts. Run the script using the following one of the following syntaxes (if

you don't give the population and codename arguments, you'll be asked to input them):

```
python el_test.py <genome_file>
```

```
python el_test.py <genome_file> [CEU | YRI] [codename]
```

4. We have also included a helper script `run-all.sh`, which will run the script on all genome files named with the convention `genome_*.txt` that have been placed in a directory named `data/` in your current directory.
5. Output will be printed to the Terminal, and an HTML page showing your chance of EL as well as some other data will be opened in your web browser as well as stored in your current directory with the same name as the genome file used to generate it, with `.html` appended. The page also shows a graph for the progressing $P(EL)$ for the nested SNP sets, starting at the prior probability $P(EL) = 0.5$, and then changing as the sets have an additional SNP added to each one up to the full 150 SNP set.