Using Galaxy to separate potentially functional and benign SNPs

*Belinda Giardine[1], Burhans R.[1], Riemer C.[1], Ratan A.[1], Harris R.[1], den Dunnen J.T.[2], Hardison R.C.[1], Zhang Y.[1], Miller W[1].

1. The Pennsylvania State University, Center for Comparative Genomics and Bioinformatics, University Park, PA, USA
2. Department of Human Genetics, Center for Human and Clinical Genetics, Leiden University Medical Center, Leiden, the Netherlands
Corresponding author: giardine@bx.psu.edu

Galaxy: usegalaxy.org
Supplement: https://main.g2.bx.psu.edu/u/Belinda/p/snp-classification
SNP tutorials: www.bx.psu.edu/miller_lab

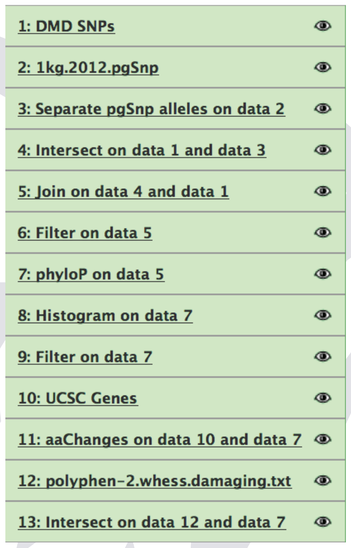
The Galaxy software framework provides a variety of tools that can be used to help distinguish SNPs that are potentially deleterious from those that are probably benign. Here we illustrate results from several of these tools, and show that with appropriate parameter selection they can produce a reasonable emulation of curated classification.

1. Galaxy is a free, powerful computational framework and service that allows users to run a wide variety of tools on their data in a highly interoperable manner. It provides a means for SNP analysis, along with a transparent setting for sharing both the methods used and results.

We illustrate the power of SNP tools in Galaxy by employing several of them to evaluate a large set of SNPs in the DMD gene and a smaller set in the LMNA gene. Some mutations in the DMD gene are known to cause Duchenne's muscular dystrophy, while mutations in the LMNA gene are associated with several diseases. We evaluate both coding and non-coding SNPs, predicting which are likely to be damaging. To assess the effectiveness of the tools, we compare these results to a knowledge-based determination from an expert. In both the test and reference sets, SNPs of unknown significance are classified as phenotype-associated.

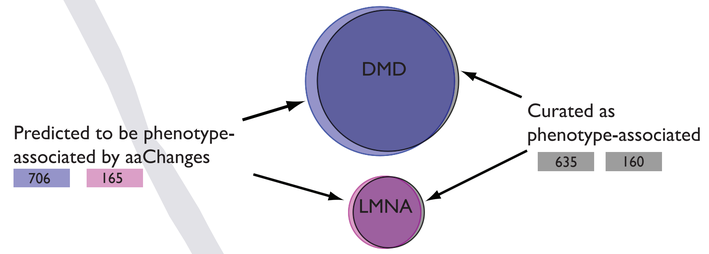
On the right is a Galaxy history showing the assortment of tools we used to classify SNPs obtained from the corresponding LMDp databases[(a)] for these two loci.

Example Galaxy History

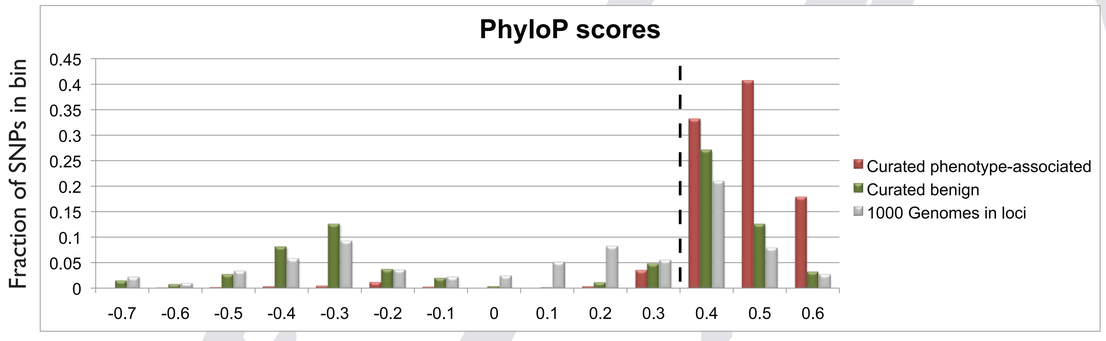


2. Some of the tools, such as aaChanges and PolyPhen2, examine the effect of SNPs on amino acids. For non-coding SNPs, we can use the Regulatory Potential (RP) score[(b)] to predict whether they are likely to be functional. And other approaches can be applied anywhere in the genome, such as PhyloP conservation scores[(c)] between species. Even the mere occurrence of a SNP in the genomes of more-or-less healthy people, such as those in the 1000 Genomes Project[(d)], can have predictive value for non-complex diseases.

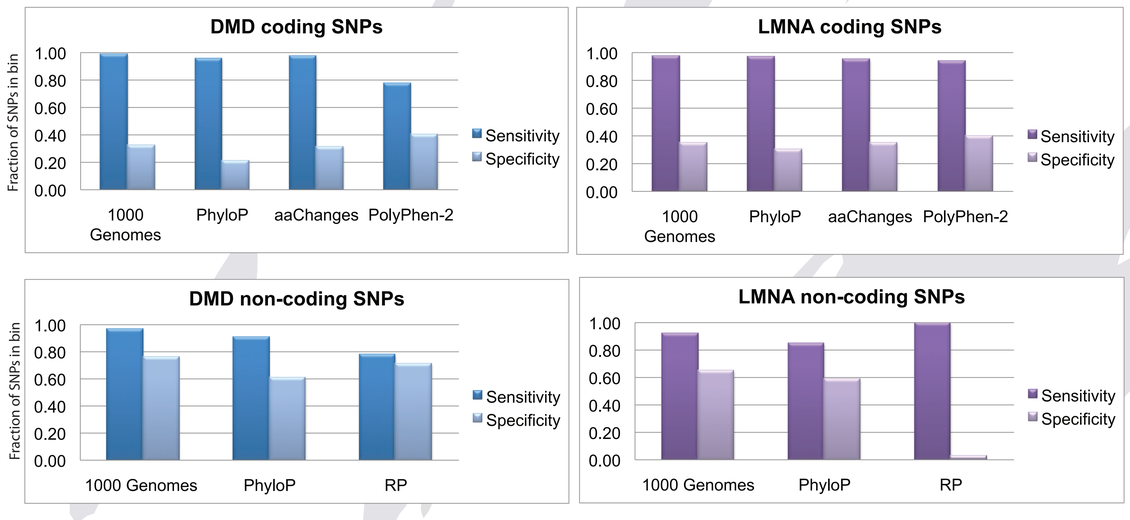
Example: the aaChanges tool shows good overlap between its phenotype-associated predictions and the curated SNP set.



3. Some of these tools have a score threshold that can be adjusted, i.e. the cutoff value used to separate the two classes of SNPs. To choose these, we plotted the score distributions of the curated SNPs, along with those of all the 1000 Genomes SNPs in our two loci (just as a background reference), and selected a value that appeared to best distinguish the SNPs curated as phenotype-associated from those curated as benign. We chose to slightly favor sensitivity over specificity. This led to PhyloP and RP score thresholds of 0.35 and -0.005, respectively. For the 1000 Genomes occurrence criterion, we used the frequency count as the score, and settled on a threshold of 1 (i.e. any SNP appearing in 1000 Genomes at all was classified as benign, while absent => phenotype-associated). The aaChanges and PolyPhen2 tools do not require any parameters.



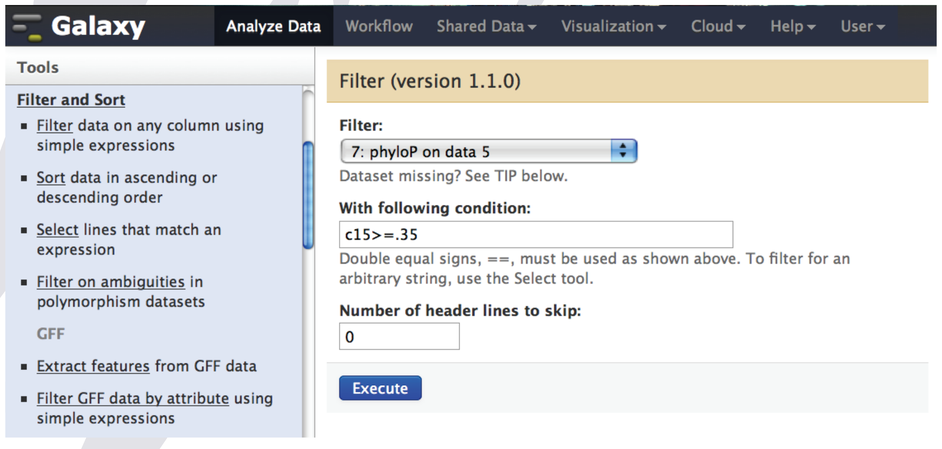
4. Sensitivity and specificity of each of these tools using the chosen thresholds. One next step would be to use Galaxy to intersect these classifications, since any SNPs placed repeatedly in the same category are more likely to be correct.



5. Similar analyses can be performed with other software, or by using these same tools outside of Galaxy. However, using Galaxy has a number of advantages:

1. Galaxy is free.
2. Galaxy is a do-it-yourself framework; you choose the tools and datasets you want.
3. With Galaxy it is easy to experiment and fine-tune the parameter settings as needed.
4. Workflows make it simple to repeat exactly the same steps on different datasets.

Screen shot showing how the above threshold for PhyloP is used in the condition box for filtering (step 9 in the history figure).



References:
(a) Leiden Muscular Dystrophy pages (www.dmd.nl)
(b) Diana Kolbe, James Taylor, Laura Elnitski, Pallavi Eswara, Jia Li, Webb Miller, Ross Hardison and Francesca Chiaromonte. Regulatory Potential Scores from Genome-Wide 3-way Alignments of Human, Mouse and Rat. Genome Research 14: 700-707.
(c) Cooper GM, Stone EA, Asimenos G, NISC Comparative Sequencing Program, Green ED, Batzoglou S, Sidow A. Distribution and intensity of constraint in mammalian genomic sequence. Genome Res. 2005 15(7):901-13.
(d) McVean et al, An integrated map of genetic variation from 1,092 human genomes, Nature 491, 56–65 (01 November 2012)

Acknowledgements:
➪ Galaxy Development Team
➪ Funding for our work on assembling and documenting the Phenotype Association tools was provided by Grant Number UL1 TR000127 from the National Center for Advancing Translational Sciences (NCATS). This project is funded, in part, under a grant with the Pennsylvania Department of Health using Tobacco CURE Funds. The Department and NIH specifically disclaim responsibility for any analyses, interpretations, or conclusions.