THE FUNCTIONAL SPECTRUM OF HUMAN VARIATION NOTES

MICHAEL DEWITT

Caveat Emptor

1. Understanding purifying selection

The purpose of figure 4 from *An integreated map of genetic variation from 1,092 human genomes* [1] is to illustrate the role of purifying selecting within and between populations. Critically, this first requires a discussion of purifying selection. **Purifying selection** or negative selection(22) is a type of background selection resulting in lower genetic diversity [2]. When mutations occur in the genome which are highly deleterious, offspring do not survive long enough to pass on these mutations to subsequent generations, at least on longer timescales [2]. When this background selection occurs, the observed genetic diversity is lower than what would be expected under neutral substitition. However, these dynamics exist within the broader population level (longer) timescales and periodic deleterious mutations do appear to exist on shorter term time scales. Additionally, as described by Vitti et al, "selection operates at the level of the phenotype, alleles showing evidence of selection are likely to be of funtional relevance" [3].

2. Quantifying human variation

2.1. **Genomic Evolutionary Rate Profiling Score.**

The Genomic Evolutionary Rate Profiling Score (GERP(2)) scores provide a way of measuring which sites likely lead to delertious mutations. This statistical framework provides a way of estimating the difference between the expected number of mutations under a neutral substitution model (which assumes no impact on fitness) and the observed variation. Positive GERP(2) scores thus represent fewer mutations than would be expected, likely indicating a more conserved site, while negative scores would indicate more substitutions than expected. Put another way, we would expect that high GERP(2) scores will occur in regions which are important for survival to reproductive age and are largely conserved in the population (i.e., at higher levels in the population with fewer mutations in these regions).

2.2. **Derived Allele(1) Frequency.**

We can examine the derived allele(1) frequency (DAP) to assess the overall distribution of alleles within a population. As a reminder, a derived allele(1) are variants which have arisen since the last common ancestor. The derived allele(1) frequency is then a summarization of the pattern and frequency that these variant alleles appear. Taking the population sampled as a whole we can calculate the frequency with which each allele(1) variant appears.

3. Examination of Figure 3

3.1. **Panel A.**

In Figure 1 we see the following:

- X-axis: the GERP(2) score representing the evolutionary conservation where higher scores are more conserved.
- Y-axis: the proportion of variants with a DAF(3) $< 0.5\%$ where higher values indicate lower frequencies in the studied population
- Colored lines: the different functional elements
- cross on the x and y axes representing the average values for GERP(2) score and proportion of variants with a DAF(3) $< 0.5\%$, respectively.

From this figure we can conclude that:

- More generally, there are fewer mutation observed in more highly conserved sites (i.e., higher GERP(2) scores and higher proportions).
- Specifically, we see that additional Stop(4) codons, Splice(5) mutations, and non-synonymous (Nonsyn(6)) mutations appear very rare and are likely associated with deleterious effects. Later in the paper these are identified as "loss of function" mutations.
- The addition of Stop(4) codons continues to be relatively rare at most sites across GERP(2) scores. This implies that the addition of the codons likely results in a severe loss of function. As stop(4) codons stop(4) transcription prematurely, these additions will result in macromolecules not being transcribed more generally. This patern is similar amongst the splice(5) mutations which impact the assembly of said macromolecules.
- We see an interesting phenomena with splice(5) variants having higher mutation in less conserved locations.
- The authors note that rare variant loads are similar for synomynous and nonsynomynous locations suggesting weak selective constraints

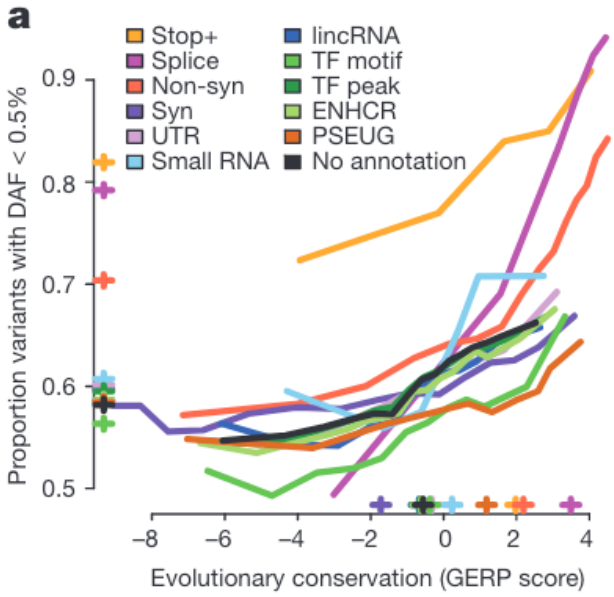

FIGURE 1. The relationaship between evolutionary conservation (measured by GERP(2) score) and rare variant proportion (fraction of all variants weith derived allele(1) frequency (DAF(3)) $< 5\%$)for variants occurring in different functional elements and with different coding consequences. Crosses indicate the average GERP(2) score at variant sites (x axis) and the proportion of rare variants (y axis) in each category.

3.2. **Panel B.**

In Figure 2 examines the CTCF-binding motif(15) within the CTCF-binding peaks. The transcription repressor CTCF has been characterized as playing a vital role in transcription regulation including the recombination of the antibody loci and the regulation of chromatin architecture [4, 5]. Intuitively, we would expect relatively low diversity in this gene as chromatin structural formation is vital for transcription (and cellular generation more generally).

The binding motif(15) is shown in the picto-graphic (called the "logo plot") for the actual nucleotide. In all cases, red represents the "out peak" and blue represents the "in peak" from the Chip-seq (ChipSeq(16)) analysis which is used to map binding sites. Those sites that are located within the peak are likely related to binding and associated with function. The in peak is the mapped functional/ active site of the

CTCF gene while the out peak represents the CTCF motif(15) , but not on the CTCF gene. This indicates that the conservation and lower diversity rates are active site conserving (preserving functionality of the gene).

3.2.1. *Upper panel.*

The y axis again represents the GERP(2) score in the different regions. Higher values represent more likely to be a conserved region as it changes less than expected under a natural substitution model.

3.2.2. *Lower panel.*

The y axis represents the average diversity as defined as the per-nucleotide pairwise distance with higher values representing more differences (i.e., more distance and differences)

3.2.3. *Figure conclusions.*

As suspected, in an important gene we see that those sites associated with function ("in peak") vary less than expected under natural substition as shown by the GERP(2) scores and lower pair-wise nucleotide distances in the lower panel. This is not to say that there isn't a more complex story as there is a hint of degeneracy in position 8.

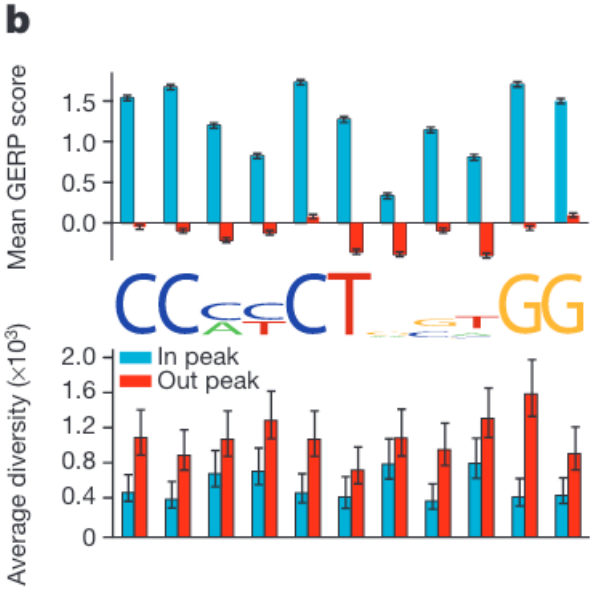

FIGURE 2. Levels of evolutionary conservation (mean GERP(2) score, top) and genetic diversity (per-nucleotide pairwise differences, bottom) for the sequences matching the CTCF-binding motif(15) within CTCF-binding peaks, as identified experimentally by ChIP-seq in the ENCODE project (blue) and in a matched set of motifs outside peaks (red). The logo plot shows the distribution of identified motifs within peaks.

3.3. **Key conclusions.**

There are some highly conserved regions and likely purifying selection is driven by the addition of stop(4) codons, splice(5) mutations, and non-synomynous mutations at these conserved sites due to their lower observed frequencies of occuring.

Glossary

allele(1) A variation in the same sequence of nucleotides at the same place on a long DNA molecule.
→ https://en.wikipedia.org/wiki/Allele(1)

GERP(2) The Genomic Evolutionary Rate Profiling is a statistical approach to quantify the deficit of substitutions in base pairs based on neutral the neutral rate of substitution. The neutral rate of substitution is an estimate of the number of new mutations per each generation multiplied by the probability that the substitution reaches fixations. In cases where neutral mutations exist, the rate of substitution will be equal to the rate of mutations. The GERP(2) Score thus indicates the deviation from the neutral rate of subsitution indicating that positive scores will indicate that fewer substitutions have taken place than would be expected under neutral substitution while negative scores indicate that more substitutions have taken place than would be expected under the neutral model.
→ https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1172034/

DAF(3) Derived allele(1) frequency which represents the distribution of alleles at a given loci.
→ https://en.wikipedia.org/wiki/Allele_frequency_spectrum

Stop(4) A nucleotide substitution leading to the addition of a stop(4) codon (UAA, UAG, UGA) and leads to premature truncation of a protein during transcription. These are associated with a loss of function.
→ https://illustrated-glossary.nejm.org/term/stop(4) -gain_variant

Splice(5) Often associated with post-translational changes in either the N-terminal region or a change in the availability of glycosylationsites through a change in the signal peptide. This can lead to secondary changes due to changes in the structure and is associated with a loss of function.
→ https://onlinelibrary.wiley.com/doi/10.1002/prot.10568

Nonsyn(6) A non-synonymous mutation is a change in the DNA sequence which results in a change in the encoded amino acid.
→ https://bmccancer.biomedcentral.com/articles/10.1186/s12885-019-5572-x

Syn(7) A synonymous mutation is a change in the DNA sequence which does not result in a change in the encoded amino acid.
→ https://bmccancer.biomedcentral.com/articles/10.1186/s12885-019-5572-x

UTR(8) The untranslated region. These appear on either side of the coding region with the 5' untranslated region called the leader sequence and the 3' untranslated region called the trailer sequence. The role of the untranslated region is not well described and may have a role in transcription regulation and mutations in this region may be associated with increased risk of certain cancers.
→ https://en.wikipedia.org/wiki/Untranslated_region

SmallRNA(9) Small RNA sequences are short, normally non-coding regions with a nucleotide length less than 200 bases. RNA silencing has been associated with these sequences.
→ https://en.wikipedia.org/wiki/Small_RNA

lincRNA(10) Large intergenic non-coding RNA are associated with remodelling of chromatin, genome architecture, RNA stabilisation, and transcription regulation.
→ https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5889127/

TFmotif(11) Transcription factor motif(15) is the sequence associated with the proteins which are involved in the transcription of DNA to mRNA.
→ https://en.wikipedia.org/wiki/Transcription_factor

TFpeak(12) Transcription factor motif(15)
→ https://en.wikipedia.org/wiki/Transcription_factor

ENHCR(13) Enhancers regulate the transcription of genes by increasing the levels than without them.
→ https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4445073/

PSEUG(14) Pseudogenes are nonfunctional segments of DNA which resemble genes. These could be artifacts from mutations of prior functioning genes.
→ https://en.wikipedia.org/wiki/Pseudogene

motif(15) A motif(15) is a nucleotide or amino acid sequence associated with a specific structure and are associated with functionally important sites.
→ https://www.ncbi.nlm.nih.gov/CBBresearch/Przytycka/download/lectures/PCB_Lect08_Bind_Motifs.pdf

ChipSeq(16) Chromatin immunoprecipitation(17) followed by sequencing (ChIP-seq) is extensively used to determine transcription factor binding sites (TFBSs). DNA binding proteins can be cross-linked to the DNA that they are binding and by using an antibody against said proteins, the entire DNA/protein complex can be precipitated out of solution.
→ https://academic.oup.com/bioinformatics/article/29/21/2705/195767

Immunoprecipitation(17) Use of an antibody to precipitate out a particular protein antigen.
→ https://en.wikipedia.org/wiki/Immunoprecipitation(17)

heterozygote advantage(18) A trend in which the fitness of a heterozygote is greater than that of either homozygote. Also referred to as overdominance.
→ https://pubmed.ncbi.nlm.nih.gov/24274750/

frequency dependent selection(19) A trend in which the fitness of a given genotype is correlated with its prevalence in the population (e.g., if an allele(1) is advantageous when it is rare)
→ https://pubmed.ncbi.nlm.nih.gov/24274750/

codominance(20) the condition in which multiple alleles are dominant; the heterozygote expresses phenotypes associated with both alleles.
→ https://pubmed.ncbi.nlm.nih.gov/24274750/

positive selection(21) in which an allele(1) is favoured and is propagated
→

negative selection(22) in which an allele(1) is disfavored (also called purifying selection)
→

References

[1] T. 1000 Genomes Project Consortium, "An integrated map of genetic variation from 1,092 human genomes," *Nature*, vol. 491, Nov. 2021, doi: 10.1038/nature11632.

[2] I. Cvijović, B. H. Good, and M. M. Desai, "The Effect of Strong Purifying Selection on Genetic Diversity," *Genetics*, vol. 209, no. 4, Aug. 2018, doi: 10.1534/genetics.118.301058.

[3] J. J. Vitti, S. R. Grossman, and P. C. Sabeti, "Detecting natural selection in genomic data," *Annu. Rev. Genetics*, vol. 47, 2013, doi: 10.1146/annurev-genet-111212-133526.

[4] G. N. Filippova, S. Fagerlie, et al., "An Exceptionally Conserved Transcriptional Repressor, CTCF, Employs Different Combinations of Zinc Fingers To Bind Diverged Promoter Sequences of Avian and Mammalian c-myc Oncogenes," *Mol. Cellular Biol.*, vol. 16, no. 6, Jun. 1996, doi: 10.1128/MCB.16.6.2802.

[5] G. M. Cooper, E. A. Stone, et al., "Distribution and intensity of constraint in mammalian genomic sequence," *Genome Res.*, vol. 15, no. 7, Jul. 2005, doi: 10.1101/gr.3577405.

Department of Biology, Wake Forest University, Winston Salem, NC 27101
Email address: dewime23@wfu.edu
URL: www.michaeldewittjr.com