Ancestry adjustment improves genome-wide estimates of regional intolerance

Tristan J. Hayeck[1-2], Nicholas Stong[3], Evan Baugh[3], Ryan Dhindsa[3], Tychele N. Turner[4], Ayan Malakar, [3] Yuncheng Duan,[5] Iuliana Ionita-Laza,[6] David Goldstein,[3] Andrew S. Allen[5]

1. Department of Pathology and Laboratory Medicine, Children's Hospital of Philadelphia, Philadelphia, PA.

2. Department of Pathology and Laboratory Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA

3. Institute for Genomic Medicine, Columbia University, New York, NY

4. Department of Genetics Washington University St. Louis, MO

5.Department of Biostatistics and Bioinformatics, Duke University, Durham, NC

6. Department of Biostatistics Columbia University, New York, NY

Genomic regions subject to purifying selection are of greater importance to the health and survival of an organism than regions not under such selection and therefore more likely to carry disease causing mutations in humans. Methods for identifying such regions can roughly be divided into those using cross species conservation and those using intolerance to standing genetic variation within a species. Cross species conservation relies on identifying regions with fewer differences than expected given divergence times between species. This makes regions that are under purifying selection in only one species difficult to detect. In contrast, intolerance looks for depletion of variation relative to expectation within a given species, allowing species specific features to be identified. When investigating the intolerance of coding sequence, this depletion is often focused on variation that affects the amino acid sequence. However, in noncoding sequence, the functional consequence of variation is less well defined, and methods strongly leverage variant frequency distributions. As the expected distributions depend on demography, if not properly controlled for, ancestral population source may obfuscate signals of selection. We demonstrate that properly incorporating demography in intolerance estimation results in greatly improved variant classification (14% increase in AUC relative to comparison constraint test, CDTS; and 7% relative to conservation). We provide a genome-wide intolerance map that is condition on demographic history that is likely to be particularly valuable in variant prioritization.

Introduction

Understanding the functional impact of noncoding sequence on protein coding sequence is one of the largest challenges in human genomics. Our ability to call variation in noncoding sequence has greatly outpaced our ability to interpret that variation and, currently, even studies employing whole genome sequencing (WGS) often restrict analyses to coding sequence. Previously, cross-species conservation has been used to identify genomic regions of likely importance. These methods are effective at identifying genomic regions that retain their functional importance across different species,[1–4] but are not effective at identifying genomic regions that have emerged as important in a given species.[5] However, emerging WGS datasets present an opportunity to address this problem as they provide a mechanism for detecting signatures of purifying selection within noncoding sequence by looking for intolerance in large standing human populations,[6,7] where up until recently this has been difficult to detect due to relatively small sample sizes.

Methods for estimating genetic intolerance have previously been applied to noncoding sequence by either by comparing the observed local distribution of variation to expectation under neutrality given a sequence context informed mutation rate,[6] or by comparing local sequence context dependent distributions of variation to genome-wide sequence context dependent distributions.[7] One such method, Orion,[6] was shown to be highly discriminative of known classes of regulatory elements and in a recent machine learning based classifier[8], it was shown to be the most informative feature of variant pathogenicity among a set of features that characterize intolerance, conservation, 3D structure, expression, and other combined metrics. Here, we improve upon existing methods in several ways. First, instead of comparing the observed SFS to that expected under neutrality, we compute the expectation empirically, by stochastically sampling from putatively neutral regions making the method less sensitive to demographic factors that may distort the SFS, as these factors will affect both the observed SFS and its neutral expectation. Second, we stratify the analysis by ancestry, computing both the observed SFS and its

expectation within each ancestry and then combining these contrasts into the final Population Conditional Intolerance Test (PCIT). It is well known that genetic diversity varies across ancestries and natural selection drives population differences in disease response.[9] By stratifying the analysis on ancestry we effectively eliminate variability in genetic diversity between human subpopulations in estimating intolerance, leading to greater precision, as we demonstrate below.

Results

Differences in Neutral SFS Across Ancestry

We begin by investigating whether the neutral SFS varies across ancestry. Since our approach contrasts the observed SFS within a region to an empirical estimate of the neutral SFS, if there are differences in the neutral SFS across ancestry groups, stratifying on ancestry could eliminate an important source of variability leading to increased power. To this end, we used 15,496 whole genome sequenced samples from the genome aggregation database (gnomAD) across 8 different ancestries,[10] including: African, American Latino/Admixed American, Ashkenazi Jewish, East Asian, Finnish, Non-Finnish European, South Asian, and other.

We estimate the neutral SFS by sampling variation from an intergenic sequence not annotated to be functional (see Quality Control, Defining Intergenic Regions, and Annotations). To test whether a given subpopulation's neutral SFS differs significantly from another subpopulation we employ the following approach: begin by randomly sampled a million positions from neutral sequence and then randomly assigned a given position to be part of the SFS estimation for one of the two subpopulations being compared. This gives half a million positions on which to estimate the neutral SFS for each subpopulation, then quantify the difference between the two distributions using the log-rank test. This process was repeated a thousand times, the subpopulation label was then randomly shuffled at each site and the log rank test was recalculated after assignment. This gives us a null distribution of SFS differences between subpopulation. This was done with each pairwise ancestry, within each ancestry, and using the entire combined gnomAD population (Fig. 1c).

The African/African-American (AFR) SFS does not differ significantly from the pooled gnomAD SFS sample population (average logrank across thousand iterations: 0.8: p-value = 0.21), whereas every other population demonstrates a significant difference in SFS relative to the full population. It is to be suspected that the AFR population's neutral SFS would demonstrate the most genetic diversity (Fig. 1b). When looking at the largest ancestries, 4,368 AFR and 7,509 Non-Finish European (NFE) samples, there is a significant difference in their SFSs (average logrank test -34: $p<1e-8$). The differences observed between the neutral SFSs across gnomAD ancestry groups suggest that conditioning on ancestry in intolerance estimation may control an important source of variability, leading to improved precision. We investigate this approach in the next section.

Ancestry Adjusted Intolerance Estimation

We conducted a genome wide scan of regional intolerance by looking for differences in the SFS within a given genomic region with an ancestry specific estimate of the neutral SFS. Specifically, we estimated the neutral SFS from a million random intergenic positions that were stochastically sampled across the genome, within a given ancestry and across all ancestries. We then compared this neutral SFS to the SFS distribution estimated from a given hundred and one base pair window (fifty bases on each side of the index position) using a log-rank test. This contrast was applied within each population and then combined to get the population conditional intolerance test (PCIT). We also conducted an unadjusted analysis where both the neutral SFS and the observed SFS within window were estimated from a pooled sample comprised of all ancestries. We refer to this as the unadjusted intolerance test (UIT). We then slide the hundred base query window to cover every position across the genome, with the restriction that the regions considered pass coverage and QC criteria (see Quality Control, Defining Intergenic Regions, and Annotations section). Since the populations have different sample sizes and relative neutral distributions, the within ancestry scores are then standardized to mean zero variance one and then the average across these scores is used to create a population conditional test statistic. The

p-value is then empirically calculated for the population conditional intolerance test using permutation. To improve computational efficiency, we take advantage of the fact that as you move from window to window very little of the data actually change and thus we can leverage our previous calculations in updating to a new window, avoiding the need to reload and recalculate across all data elements.

We investigated the performance of the various approaches by looking at how they classify non-coding ClinVar pathogenic variants versus a million randomly sampled common variants (MAF>5%) taken from 62,784 whole genome Trans-Omics for Precision Medicine (TOPMed projects) samples. As can be seen from figure 1, areas under the receiver operator curves (AUCs) are significantly improved when ancestry is accounted for (PCIT: AUC=90%) relative to when it is not (UIT: AUC=80%). PCIT also substantially outperforms two previously proposed approaches for measuring constraint and conservation in noncoding sequence (CDTS: AUC=76% and GERP: AUC=83%) (Fig. 2). Similar results are seen in classifying ClinVar coding variants. Specifically, we found with when ancestry is accounted for (PCIT: AUC=93%) relative to a similar unadjusted approach (UIT: AUC=85%). It also substantially outperforms a previously proposed approach for classifying noncoding sequence (CDTS: AUC=80% and GERP: AUC=90%) (Supp. Fig. 1). The same comparison was done with Orion, which also uses a sliding window approach to compare the observed SFS with a theoretical expectation under neutrality, computed given the sequence context dependent mutation rate found within the window. Orion slightly outperformed CDTS but underperformed relative to both the new population unadjusted intolerance test and the population conditional intolerance test (Orion non-coding AUC=78%, coding AUC =80%). Orion was run using a different quality control criteria and, as a result, used a significantly different set of variants. For this reason, we left it out of the comparisons found in the main text, but still note it in the supplementary materials (Supp. Fig. 2).

We next investigated how sequences showing extreme PCIT scores (in the top 10%) are distributed across different genomic regions, such as exons: introns, enhancers, promotors, etc.

by comparing how often sequence with a given annotation lies in the top decile of PCIT relative to sequence found in common intergenic regions (details of annotations described in Quality Control, Defining Intergenic Regions, and Annotations). We found that ultra-conserved regions have a 34.81 fold enrichment of extreme PCIT scores relative to common intergenic regions (Fig. 3). Exons show a 21.07 fold enrichment ;16.19 for UTRs; 15.4 for introns; 14.99 for Hi-C experimental data that was taken from di Iulio used to validate CDTS,[7] and 14.98 for enhancers (Supp Table 1). Non-coding RNA elements also showed an enrichment of intolerance relative to common intergenic regions (Fig. 3): miRNA showed a 14.99 fold enrichment; and a 13.19 fold enrichment for lincRNA (Fig. 3).

We then looked at how the intolerance of regulatory elements correlates with the genic intolerance of the genes they regulate. We began by connecting specific enhancer regions to the genes they regulate in various tissue types using Roadmap (http://www.biolchem.ucla.edu/labs/ernst/roadmaplinking/).[11–13] As the PCIT is a nucleotide level score, we took the average of such scores across each enhancer, to create a single score per enhancer. Using the roadmap links for a given cell line, we linked a target gene for each enhancer. We binned the intolerance scores for enhancers targeting genes in a given gene set into 20 equal sized bins and took the median enhancer score within each bin. We also computed median RVIS score for the genes being targeted by the enhancers within each bin. Finally, we correlated the bins' median enhancer and median target gene RVIS scores (Fig. 4). Consistent patterns of correlation between genic intolerance and regional regulatory intolerance were seen in OMIM genes (Fig. 4A-B correlation brain = 0.85, neurosphere 0.92), haploinsufficient genes (Fig. 4C-D correlation brain = 0.78, neurosphere 0.53), and neurodevelopmental autosomal dominant genes (Fig. 4E-F correlation brain = 0.62, neurosphere 0.43).).

Interestingly, some of the gene sets (e.g., haploinsufficient genes with enhancers linked both via brain cells and via neurosphere cells) show non-monotonicity in the relationship between enhancer and target gene intolerance, with low enhancer scores corresponding to genes with

relatively high genic intolerance scores. One possible explanation is that small genes that are subject to purifying selection may have poorly estimated genic intolerance, due to their short coding sequence and limited potential for variation[14]. To investigate further, we computed the average length of the target genes of the enhancers found within each bin (Supp. Fig. 3). Indeed, those bins with relatively high RVIS scores but with very intolerant ehancers tend to be comprised of relatively short target genes.

The strong correlations seen in the OMIM gene sets for brain and neurosphere linked enhancers were also seen when linked through other cell types with very strong correlations being seen when enhancers were linked via fibroblast lung (IMR90 correlation = 0.95) and muscle (Myosat correlation = 0.93, muscle correlation = 0.92) (Supp. Fig. 4).

Discussion

Clear patterns of correlation were observed between intolerant regulatory elements and the genic intolerance scores of the target genes those elements regulate. Interestingly, the most striking violations of this pattern, where enhancers were strongly intolerant but the target genes those enhancers regulated were relatively tolerant, were comprised of relatively short genes. A plausible explanation for this is that the genes' coding and regulatory sequence is actually under relatively strong purifying selection but that the short coding sequence of the gene limits the genic intolerance signa we are able to observe, but this warrants further investigation.

Sample size is a key factor in being able to precisely identify intolerant regions. Given the current limited number of publicly available whole genome sequences, some ancestries are under-represented in this sample. For example, there are only 151 Ashkenazi Jewish individuals represented in the current analysis. However, with emerging population-based sequencing programs such as All of Us[15] and the UK biobank[16] we expect the number of sequences to increase dramatically in the coming years and to better represent diverse ancestries. This, in turn, will allow more precise regional intolerance estimation of smaller and smaller subunits of the

genome. In the current study a sliding window of 100 bases was chosen to capture enough variation to precisely estimate intolerance while also balancing the localization of the scores. With larger and larger samples sizes, smaller windows will provide the same precision while better identifying finer and finer local structure in intolerance signal.

An alternative to the sliding window approach is to use predefined regional definitions based on known biology, e.g., enhancers, promotors, exons, introns, etc. However, the sizes of such regions can vary dramatically and, as a result, the precision of intolerance estimates will vary dramatically as well, with smaller regions' intolerance often being very poorly estimated. A possible solution to this is to develop hierarchical models in this context that allow borrowing of information across similar regions, potentially stabilizing estimates. Such an approach has been successfully applied to intolerance estimation of subregions in coding sequence.[17]

Other methods for estimating regional intolerance consider sequence context variation rates, either to estimate an expected number of variants within a given region using a sequence context mutation rate,[10,18,19] or by comparing a local rate of variation within a given sequence context to that observed in that context genome-wide.[7] Though we found that our SFS-based approach outperformed CTDS, it is clear that two regions that are subject to the same level of purifying selection may have very different SFSs simply due to sequence context mutation rate differences between the regions. Thus, there is an opportunity to further refine the PCIT framework by developing sequence context informed SFS estimates within ancestry groups and to contrast that with an ancestry specific sequence context informed neutral SFS.

We do not directly compare to other methods that aggregate annotations[8,20–23] from other rich feature sets[24–26]; however, we are encouraged that the approach proposed here will be useful in this context by the fact that Orion, a previously proposed SFS-based approach for estimating genome-wide regional intolerance, was recently demonstrated to be the most informative feature in a recent predictive model of variant pathogenicity.[8] Improving such predictions, especially for

variants in non-coding sequence, has implications for the interpretation of variation across genetics studies from genetic discovery to the diagnostic interpretation of patient genomes.

Online Resources

https://github.com/tris-10/PopCondIntolTest

https://macarthurlab.org/2017/02/27/the-genome-aggregation-database-gnomad/

https://egg2.wustl.edu/roadmap/web_portal/chr_state_learning.html#core_15state

https://www.nhlbi.nih.gov/science/trans-omics-precision-medicine-topmed-program

http://www.biolchem.ucla.edu/labs/ernst/roadmaplinking/

Quality Control, Defining Intergenic Regions, and Annotations

Variants included in SFS calculations for the PCIT had to meet gnomAD PASS criteria. In addition, indels were excluded and only autosomal chromosomes were used. Low coverage regions where all positions in a window did not have 10X coverage in 70% of samples, SEGDUP regions, and recent repeat regions defined as having sequence < 10% diverged from the consensus in RepeatMasker[27]were removed. Aggarwala and Voight characterized sequence context and modeled heptamere mutation rates focusing on intergenic regions to explicitly avoid the potential impact of negative selection.[9] We build on this definition to create an empirical sample of within ancestry intergenic SFS spectrums defined as the full set of genomic sequence filtering out centromeric, telomeric, repetitive regions, gene deserts of length greater than 2MB, sequence not present in the combined accessibility mask of 1000 genomes. Additionally, we restricted regions to be least 1KB away from any gene.

The annotations in Fig. 3 were predominantly taken from Ensembl[27] including: CDS exon, CDS intron, CCCTC-binding factor (CTCF), promoter flanking regions, open chromatin regions, transcription factor binding sites (TFBS), enhancer, promoter, untranslated regions (UTR),

transcript nonsense mediated decay, lincRNA, miRNA, snoRNA, miscRNA, rRNA, and snRNA. Additional annotations that were used included: Human accelerated regions (HAR),[28] ultra conserved elements (UCE) (https://www.ultraconserved.org/),[29] Hi-C experimental data,[7] DNase I hypersensitive (DHS),[6] and a million randomly sampled variants from TOPMed (https://www.nhlbi.nih.gov/science/trans-omics-precision-medicine-topmed-program)[30] with MAF greater than 1%.Cell specific enhancers were defined based on Roadmap Core 15-state model (5 marks, 127 epigenomes https://egg2.wustl.edu/roadmap/web_portal/chr_state_learning.html#core_15state).

Online Methods

We use a log rank test to test for differences between the SFS within a given query window and the SFS estimated from intergenic neutral sequence. To estimate the intergenic neutral SFS, we sampled a million random intergenic positions, and computed the SFS from the variation at those positions within each population. Our testing approach was optimized to take advantage of the fact that the data is sparse and does not change dramatically when moving from query window to query window. Thus, when moving from one window to another we can avoid fully reloading the data structures by simply removing the counts from the old position and then adding those for the new position. This greatly increases computational efficiency. Since the different ancestries have different sample sizes, the log rank statistics across each population are standardized to have mean zero and variance one and then the average across the statistics are taken at each position to get a stratified analysis. We characterize an empirical null by sampling a million random intergenic positions, using the statistic calculated over the 101 base window at each of those positions, to generate an empirical distribution function. The empirical p-value for each position

corresponds to any statistic that is as extreme or more extreme than that observed statistic in the query window.

1. Davydov, E. V. *et al.* Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput. Biol.* **6**, (2010).
2. Petrovski, S. *et al.* The Intolerance of Regulatory Sequence to Genetic Variation Predicts Gene Dosage Sensitivity. *PLoS Genet.* **11**, 1–25 (2015).
3. Schrider, D. R. & Kern, A. D. Inferring selective constraint from population genomic data suggests recent regulatory turnover in the human brain. *Genome Biol. Evol.* **7**, 3511–3528 (2015).
4. Sundaram, L. *et al.* Predicting the clinical impact of human mutation with deep neural networks. *Nat. Genet.* **50**, 1161–1170 (2018).
5. Rands, C. M., Meader, S., Ponting, C. P. & Lunter, G. 8.2% of the Human Genome Is Constrained: Variation in Rates of Turnover across Functional Element Classes in the Human Lineage. *PLoS Genet.* **10**, (2014).
6. Gussow, A. B. *et al.* Orion : Detecting regions of the human non- coding genome that are intolerant to variation using population genetics. 1–17 (2017).
7. di Iulio, J. *et al.* The human noncoding genome defined by genetic diversity. *Nat. Genet.* **50**, 333–337 (2018).
8. Wells, A. *et al.* Ranking of non-coding pathogenic variants and putative essential regions of the human genome. *Nat. Commun.* **10**, 5241 (2019).
9. Nédélec, Y. *et al.* Genetic Ancestry and Natural Selection Drive Population Differences in Immune Responses to Pathogens. *Cell* **167**, 657-669.e21 (2016).
10. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
11. Fujita, P. A. *et al.* The UCSC genome browser database: Update 2011. *Nucleic Acids Res.* **39**, 876–882 (2011).
12. Ernst, J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**, 43–49 (2011).
13. Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–329 (2015).
14. Bartha, I., di Iulio, J., Venter, J. C. & Telenti, A. Human gene essentiality. *Nat. Rev. Genet.* **19**, 51–62 (2017).
15. All, T. *et al.* The "All of Us" Research Program. *N. Engl. J. Med.* **381**, 668–676 (2019).
16. Sudlow, C. *et al.* UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLoS Med.* **12**, 1–10 (2015).
17. Hayeck, T. J. *et al.* Improved Pathogenic Variant Localization via a Hierarchical Model of Sub-regional Intolerance. *Am. J. Hum. Genet.* **104**, 299–309 (2019).
18. Karczewski, K., Francioli, L. & Karczewski, K. The genome Aggregation Database (gnomAD) | MacArthur Lab. 1–10 (2017).
19. Karczewski, K. J. *et al.* Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of- function intolerance across human protein-coding genes. (2019).
20. Zhou, J. & Troyanskaya, O. G. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat. Methods* **12**, 931–934 (2015).
21. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310–315 (2014).
22. Huang, Y.-F., Gulko, B. & Siepel, A. Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. *Nat. Genet.* **49**, 618–624 (2017).
23. Ionita-Laza, I., McCallum, K., Xu, B. & Buxbaum, J. D. A spectral approach integrating

functional genomic annotations for coding and noncoding variants. *Nat. Genet.* **48**, 214–220 (2016).
24. Hunt, S. E. *et al.* Ensembl variation resources. *Database (Oxford).* **2018**, 1–12 (2018).
25. Ardlie, K. G. *et al.* The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science (80-.).* **348**, 648–660 (2015).
26. Harrow, J. *et al.* GENCODE: The reference human genome annotation for the ENCODE project. *Genome Res.* **22**, 1760–1774 (2012).
27. Turner, T. N. *et al.* Genomic Patterns of De Novo Mutation in Simplex Autism. *Cell* **171**, 710-722.e12 (2017).
28. Doan, R. N. *et al.* Mutations in Human Accelerated Regions Disrupt Cognition and Social Behavior Article Mutations in Human Accelerated Regions Disrupt Cognition and Social Behavior. 341–354 (2016). doi:10.1016/j.cell.2016.08.071
29. Mccole, R. B. *et al.* Ultraconserved Elements Occupy Specific Arenas of Three-Dimensional Mammalian Genome Article Ultraconserved Elements Occupy Specific Arenas of Three-Dimensional Mammalian Genome Organization. 479–488 (2018). doi:10.1016/j.celrep.2018.06.031
30. Zachary, A. *et al.* Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Biorxiv* 1–46 (2019).

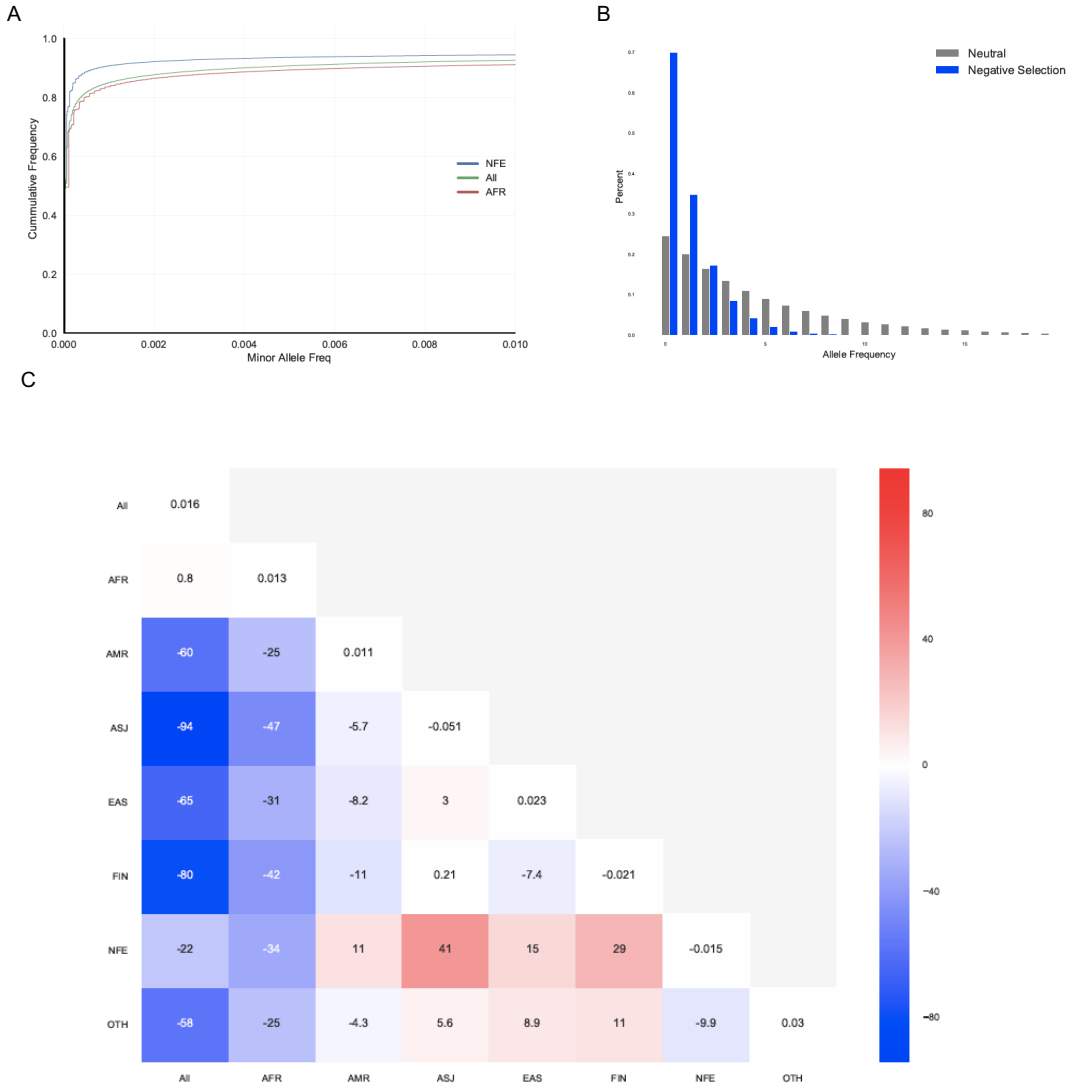

Fig. 1 Characterizing cross population shifts in the site frequency spectrum. *Understanding cross population shifts in site frequency spectrum, first is an illustrative diagram a) adapted from Sawyer and Hartl 1992 Genetics (remove Hartl version and put in ours) showing how negative selection is expected to effect the SFS. Then an empirical sample of b) the cumulative SFS for million intergenic bases taken across the two largest populations in gnomAD and the combined cumulative SFS. In the lower plot is c) a heatmap of the average over a thousand random permutations of half a million intergenic positions in one population versus another half a million intergenic positions from another population.*

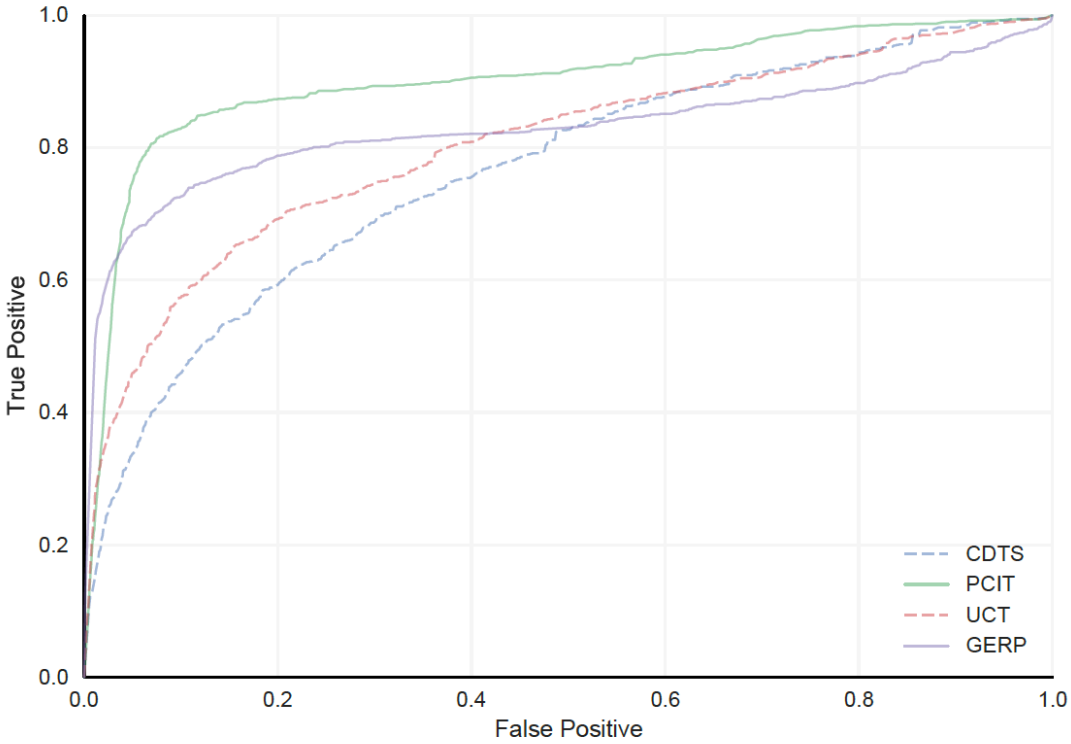

Fig. 2 Predictive utility of population conditional constraint test relative to other constraint and conservation metrics for ClinVar non-coding variants. *Non-coding ClinVar pathogenic variants versus a million randomly sampled variants from TOPMed with MAF greater than 5%.*

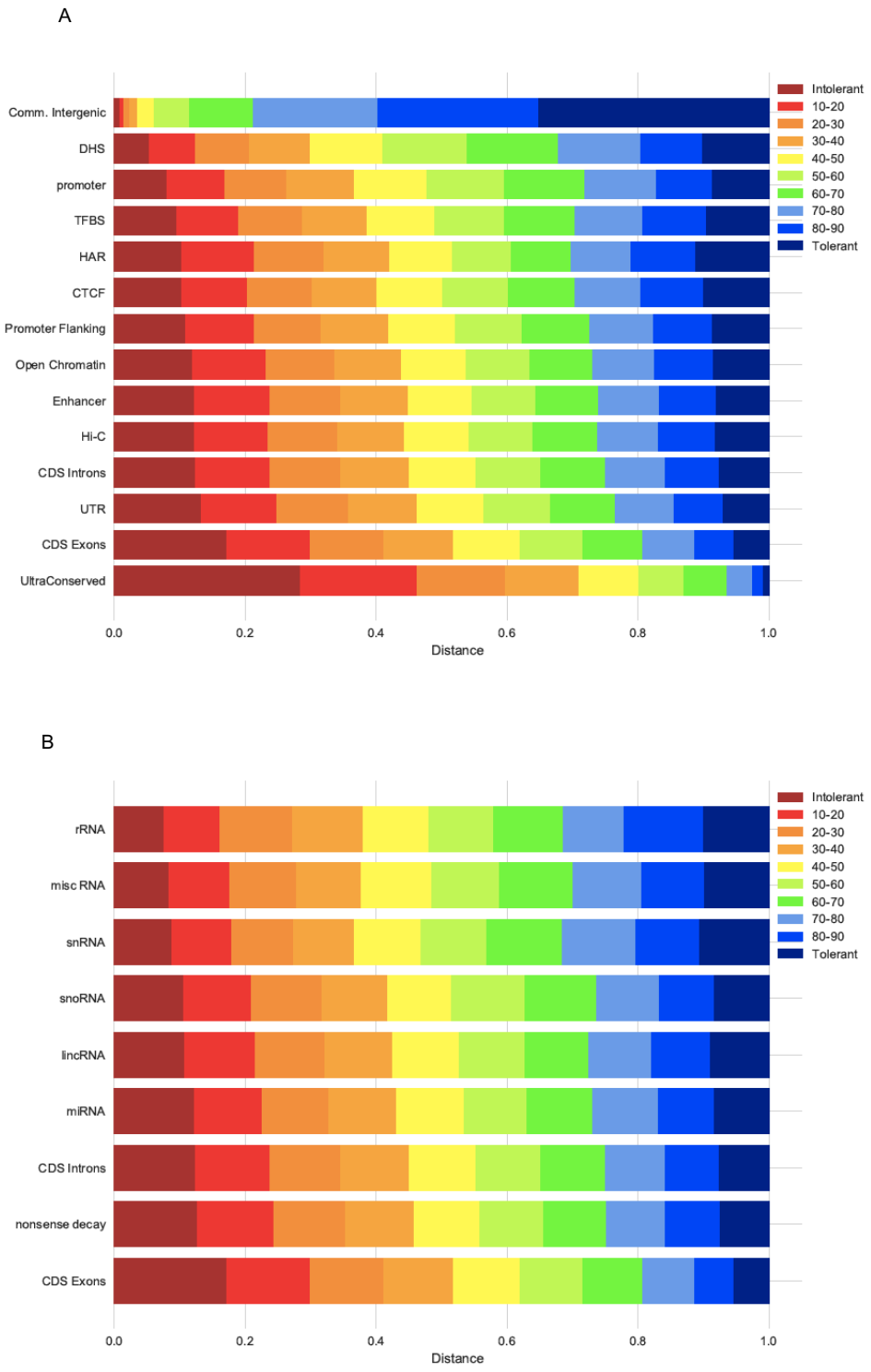

Fig. 3 Relative constraint across different annotations from Ensembl[27] and other resources.[5,6,31,32] *Decile break downs of PCIT scores across different functional annotations from and a million randomly sampled intergenic variants from TOPMed with MAF greater than 5%.*