Single-nucleotide evolutionary constraint scores highlight disease-causing mutations

To the Editor: Identifying disease-causing genetic variants in individual human genomes is a major challenge, even in protein-coding exons (the 'exome'). Analysis of nucleotide-level sequence conservation may help address this challenge, on the assumption that purifying selection 'constrains' evolutionary divergence at phenotypically important nucleotides. In contrast to functional classifiers (for example, nonsynonymous mutations), constraint scores are quantitative and applicable to any genomic position[1]. However, it remains unclear if constraint scores can facilitate causal variant discovery, as statistical power is estimated to be marginal at the single-nucleotide level given current genome alignments[1,2].

We therefore applied and assessed a nucleotide-level evolutionary metric to prioritize causal variants in genomes of 16 individuals. We analyzed exomes from four individuals with Freeman-Sheldon syndrome (FSS; Online Mendelian Inheritance in Man (OMIM) database identifier 193700), a dominant disease caused by mutations in *MYH3* (ref. 3); four individuals with Miller syndrome (OMIM identifier 263750), a recessive disease caused by mutations in *DHODH*[4]; and eight HapMap samples[3]. We generated constraint scores by genomic evolutionary rate profiling (GERP)[1] on the mammalian subset of the 44-way Multiz threaded blockset aligner alignments (for details see ref. 2). For each aligned site, GERP defines a 'rejected substitution' (RS) score by estimating the actual number of substitutions at that site and subtracting it from the number expected assuming neutrality (~5.82 substitutions per site). Selectively constrained sites tolerate fewer substitutions than neutral sites and have positive RS scores[1,2].

We first defined the consensus nucleotide from the chimpanzee, gorilla, orangutan and macaque genomes as ancestral and determined the derived allele frequency (DAF) for each variant in the eight HapMap exomes. We found a significant inverse correlation between DAF and RS score (**Fig. 1**; $P < 0.0001$; $R^2 = 1.4\%$, slope estimated as –1% DAF per RS). No correlation existed between DAF and the RS score for the nucleotide adjacent to the variant (**Supplementary Fig. 1**). Although the DAF-RS correlation resulted partly from enrichment for singletons at sites with high RS scores, it was significant even within common variants ($P < 0.0001$; **Supplementary Fig. 2**). We also found that segregating sites, regardless of DAF, were enriched at sites with low RS scores and progressively depleted as the RS scores increased (**Supplementary Fig. 3**). Consistent with previous data[2], these results suggest that RS scores enrich site-specifically for deleterious variants and nonvariant positions at which new mutations would be deleterious.

Next, we tested whether constraint scores could enrich for FSS or Miller syndrome causal variants. We identified candidate

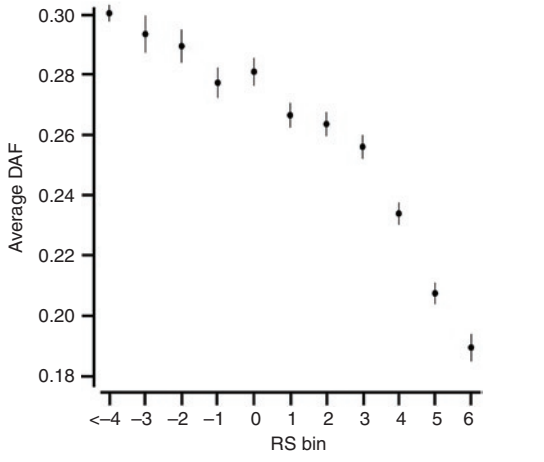

Figure 1 | RS scores inversely correlate with DAF of single-nucleotide variants in eight HapMap exomes. The average DAF was plotted for all variants at a site within a given RS bin. Error bars, 1 s.e. unit ($n = 48,750$).

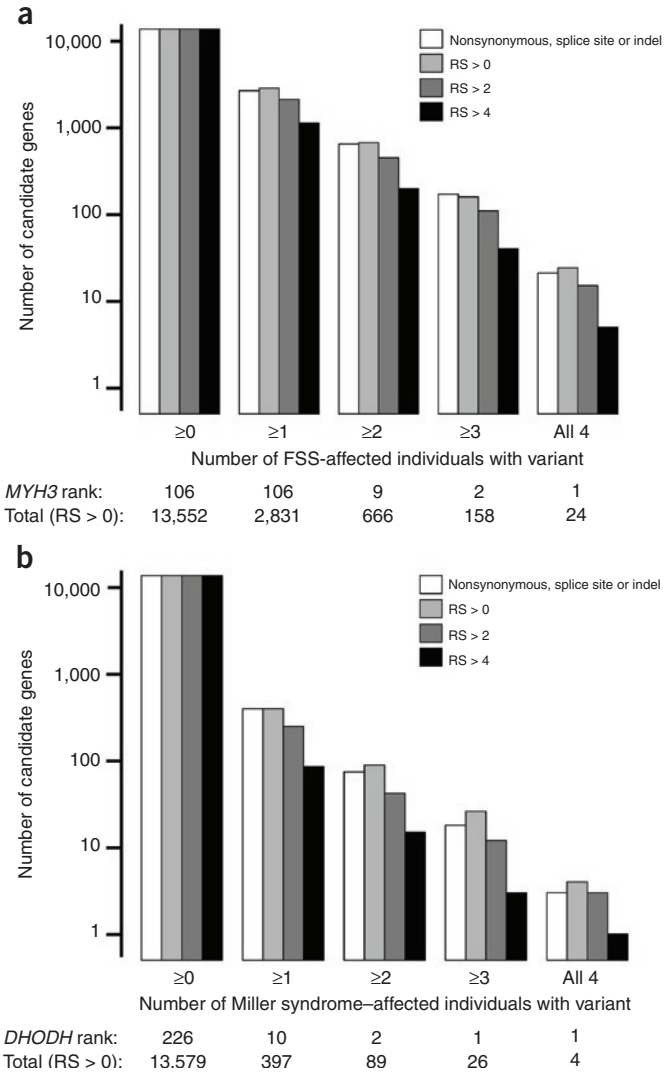

Figure 2 | Constraint scores enrich for disease-causing genes. (**a**,**b**) Number of candidate genes (log scale) in which at least the given number of individuals with FSS (**a**) or with Miller syndrome (**b**) has a rare variant that is functionally defined (nonsynonymous, splice site or indel) or with RS scores as indicated. Total numbers of candidate genes defined at RS > 0 and the rank of the indicated gene among those genes are also given.

disease genes as those in which the affected individuals had variants not seen in the HapMap exomes that affected a nucleotide with a high constraint score. For a comparison, we used functional definitions of deleteriousness, namely nonsynonymous, splice site or insertion-deletion (indel)[3,4]. We first used a threshold of RS > 0 (fewer substitutions than expected) and found that this narrowed candidate gene lists nearly as effectively as functional annotations. For example, there were 21 genes in which all FSS samples had a rare, functionally annotated variant[3,4] versus 24 genes in which all FSS samples had a rare variant with RS > 0 (**Fig. 2a**). Increasing the RS threshold, which cannot readily be done with functional annotations, reduced candidate gene lists. At a threshold of RS > 4, for example, *MYH3* was one of only five FSS candidate genes, and *DHODH* was the only Miller syndrome candidate (**Fig. 2b**).

We note that protein-based approaches could similarly be used to reduce candidate gene lists. For example, there were only seven genes in which all FSS individuals harbored a rare variant annotated by PolyPhen[5] as 'possibly' or 'probably' damaging. However, PolyPhen (and related approaches) is restricted to nonsynonymous variants, does not facilitate ranking of candidates and excluded *DHODH* as a Miller syndrome candidate[4].

Finally, we exploited the quantitative nature of constraint scores and ranked genes by the average score of all rare and deleterious (RS > 0) variants in the affected individuals. *MYH3* and *DHODH* ranked highly for their associated diseases, even under models allowing for the possibility of multiple causal genes. For example, requiring only that at least two individuals shared the same causal gene, *MYH3* ranked ninth among 666 genes. If we assumed FSS and Miller syndrome to be monogenic, *MYH3* and *DHODH* ranked as the top candidates, respectively (**Fig. 2**).

RS scores for known or user-defined variants can be obtained from the Genome Variation Server (http://gvs.gs.washington.edu/GVS/) or SeattleSeq annotation pipeline (http://gvs.gs.washington.edu/SeattleSeqAnnotation/). Constraint scores facilitate threshold flexibility and candidate ranking and do not require functional annotations. Even in exomes, this allows for the possibility that synonymous variants contribute to disease[6]. More importantly, this independence offers exciting potential for the discovery of causal variation in arbitrary genomic segments (for example, linkage peaks) and ultimately resequenced genomes.

Note: Supplementary information is available on the Nature Methods website.

ACKNOWLEDGMENTS

G.M.C. is grateful for support from a Merck, Jane Coffin Childs Memorial Fund postdoctoral fellowship. D.L.G. is supported by a Lucille P. Markey Biomedical Research Stanford Graduate Fellowship. S.B.N. is supported by the Agency for Science, Technology and Research, Singapore. This work was also supported by grants from the US National Institutes of Health: U01 HL66682 (D.A.N.), 5R01HL094976-02 (D.A.N. and J.S.), 5R01HD048895 (M.J.B.) and 1R21HG004749-01 (J.S.).

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Gregory M Cooper[1], David L Goode[2], Sarah B Ng[1], Arend Sidow[2,3], Michael J Bamshad[1,4], Jay Shendure[1] & Deborah A Nickerson[1]

[1]Department of Genome Sciences, University of Washington, Seattle, Washington, USA. [2]Department of Genetics and [3]Department of Pathology, Stanford University, Stanford, California, USA. [4]Department of Pediatrics, University of Washington, Seattle, Washington, USA.
e-mail: coopergm@u.washington.edu

1. Cooper, G.M. *et al. Genome Res.* **15**, 901–913 (2005).
2. Goode, D.L. *et al. Genome Res.* **20**, 301–310 (2010).
3. Ng, S.B. *et al. Nature* **461**, 272–276 (2009).
4. Ng, S.B. *et al. Nat. Genet.* **42**, 30–35 (2010).
5. Sunyaev, S. *et al. Hum. Mol. Genet.* **10**, 591–597 (2001).
6. Cartegni, L., Chew, S.L. & Krainer, A.R. *Nat. Rev. Genet.* **3**, 285–298 (2002).