IMMUNOLOGY2013™ Travel Grants
from R&D Systems
$1000 travel grants will be awarded to
10 immunology researchers
Apply Now!
R&D SYSTEMS

An Abundance of Rare Functional Variants in 202 Drug Target Genes Sequenced in 14,002 People

Matthew R. Nelson *et al.*

Science **337**, 100 (2012);
DOI: 10.1126/science.1217876

Science (print ISSN 0036-8075; online ISSN 1095-9203) is published weekly, except the last week in December, by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005. Copyright 2012 by the American Association for the Advancement of Science; all rights reserved. The title *Science* is a registered trademark of AAAS.

essentially inactive (Fig. 4H), suggesting that *MPC1* function is evolutionarily conserved from yeast to humans.

The data presented here demonstrate that the Mpc1-Mpc2 complex is an essential component of the mitochondrial pyruvate carrier in yeast, flies, and mammals. This is consistent with experiments performed in rat liver, heart, and castor beans, which implicated proteins of 12 to 15 kD in mitochondrial pyruvate uptake (*15*)—similar to the molecular masses of Mpc1 (15 kD), Mpc2 (14 kD), and Mpc3 (16 kD). Although these individual sizes are relatively small, Mpc1 and Mpc2 form a complex of ~150 kD, suggesting that an oligomeric structure mediates pyruvate transport. The demonstration that Mpc1 and Mpc2 are sufficient to promote pyruvate uptake in a heterologous system provides further evidence that they constitute an essential pyruvate transporter (*16*). Finally, the degree to which carbohydrates are imported into mitochondria and converted into acetyl-CoA is a critical step in normal glucose oxidation as well as the onset of diabetes, obesity, and cancer. Thus, like PDH, which is controlled by allostery and posttranslational modification (*17*), the mitochondrial import of pyruvate is likely to be precisely regulated (*18*, *19*). The identification of Mpc1 and Mpc2 as critical for mitochondrial pyruvate transport provides a new framework for understanding this level of metabolic control, as well as new directions for potential therapeutic intervention.

References and Notes

1. D. Hanahan, R. A. Weinberg, *Cell* **144**, 646 (2011).
2. S. E. Kahn, R. L. Hull, K. M. Utzschneider, *Nature* **444**, 840 (2006).
3. A. P. Halestrap, R. M. Denton, *Biochem. J.* **138**, 313 (1974).
4. A. P. Halestrap, *Biochem. J.* **172**, 377 (1978).
5. A. P. Halestrap, *Biochem. J.* **148**, 85 (1975).
6. S. Todisco, G. Agrimi, A. Castegna, F. Palmieri, *J. Biol. Chem.* **281**, 1524 (2006).
7. J. C. Hildyard, A. P. Halestrap, *Biochem. J.* **374**, 607 (2003).
8. M. Jiang *et al.*, *Mol. Biol. Rep.* **36**, 215 (2009).
9. D. J. Pagliarini *et al.*, *Cell* **134**, 112 (2008).
10. A. Sickmann *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **100**, 13207 (2003).
11. H. Y. Steensma, L. Holterman, I. Dekker, C. A. van Sluis, T. J. Wenzel, *Eur. J. Biochem.* **191**, 769 (1990).
12. E. Boles, P. de Jong-Gubbels, J. T. Pronk, *J. Bacteriol.* **180**, 2875 (1998).
13. S. Papa, G. Paradies, *Eur. J. Biochem.* **49**, 265 (1974).
14. M. Brivet *et al.*, *Mol. Genet. Metab.* **78**, 186 (2003).
15. A. P. Thomas, A. P. Halestrap, *Biochem. J.* **196**, 471 (1981).
16. S. Herzig *et al.*, *Science* **337**, 93 (2012).
17. R. A. Harris, M. M. Bowker-Kinley, B. Huang, P. Wu, *Adv. Enzyme Regul.* **42**, 249 (2002).
18. F. M. Zwiebel, U. Schwabe, M. S. Olson, R. Scholz, *Biochemistry* **21**, 346 (1982).
19. R. Rognstad, *Int. J. Biochem.* **15**, 1417 (1983).
20. Materials and methods are available as supplementary materials on *Science* Online.

Acknowledgments: We thank members of the Rutter, Thummel, Winge, Stillman, Shaw, and Metzstein laboratories for helpful discussions. We thank the Shaw and Winge labs for the antibodies against Fzo1, Cyb2, and Mge1 and for the mito-RFP constructs. We thank J. M. Saudubray, L. Burglen, and H. Tevissen for referring patients and C. Thibault and J. L. Mandel (IGBMC, Strasbourg, France) for assistance in single-nucleotide polymorphism array hybridization. This research was supported by NIH grants R01GM083746 (J.R.), RC1DK086426 (C.S.T.), and R24DK092784 (J.R. and C.S.T.) and a pilot grant from P30DK072437 (J.R.). D.K.B. and C.M.C. were supported by the NIH Genetics Predoctoral Training Grant T32GM007464. E.B.T. was supported by NIH Pathway to Independence award K99AR059190. D.K.B., T.O., C.S.T., and J.R. are inventors on a patent application by the University of Utah covering the discovery of the MPC complex.

Supplementary Materials
www.sciencemag.org/cgi/content/full/science.1218099/DC1
Materials and Methods
Figs. S1 to S10
Table S1
References

19 December 2011; accepted 9 May 2012
Published online 24 May 2012;
10.1126/science.1218099

An Abundance of Rare Functional Variants in 202 Drug Target Genes Sequenced in 14,002 People

Matthew R. Nelson,[1]*† Daniel Wegmann,[2]* Margaret G. Ehm,[1] Darren Kessner,[2] Pamela St. Jean,[1] Claudio Verzilli,[3] Judong Shen,[1] Zhengzheng Tang,[4] Silviu-Alin Bacanu,[1] Dana Fraser,[1] Liling Warren,[1] Jennifer Aponte,[1] Matthew Zawistowski,[5] Xiao Liu,[6] Hao Zhang,[6] Yong Zhang,[6] Jun Li,[7] Yun Li,[4] Li Li,[1] Peter Woollard,[3] Simon Topp,[3] Matthew D. Hall,[3] Keith Nangle,[1] Jun Wang,[6,8] Gonçalo Abecasis,[5] Lon R. Cardon,[9] Sebastian Zöllner,[5,10] John C. Whittaker,[3] Stephanie L. Chissoe,[1] John Novembre,[2]†‡ Vincent Mooser[9]‡

Rare genetic variants contribute to complex disease risk; however, the abundance of rare variants in human populations remains unknown. We explored this spectrum of variation by sequencing 202 genes encoding drug targets in 14,002 individuals. We find rare variants are abundant (1 every 17 bases) and geographically localized, so that even with large sample sizes, rare variant catalogs will be largely incomplete. We used the observed patterns of variation to estimate population growth parameters, the proportion of variants in a given frequency class that are putatively deleterious, and mutation rates for each gene. We conclude that because of rapid population growth and weak purifying selection, human populations harbor an abundance of rare variants, many of which are deleterious and have relevance to understanding disease risk.

Understanding the genetic contribution to human disease requires knowledge of the abundance and distribution of functional genetic diversity within and among populations. The "common-disease rare-variant" hypothesis posits that variants affecting health are under purifying selection and thus should be found only at low frequencies in human populations (*1*–*3*). This hypothesis has become increasingly credible because very large genome-wide association studies of common variants have explained only a fraction of the known heritability of most traits (*4*, *5*). Investigating the role of rare variants for complex trait mapping has led to tests that aggregate rare variants (*6*) and determine the abundance, distribution, and phenotypic effects of rare variants in human populations (*7*, *8*).

Population genetic models predict that mutation rates, the strength of selection, and demography affect the abundance of rare variants, although the relative importance of each is a long-standing question (*9*–*11*). To understand rare variant diversity in humans, we sequenced 202 genes in a sample of 14,002 well-phenotyped individuals (table S1). These genes represent approximately 1% of the coding genome and approximately 7% of genes considered current or potential drug targets (*12*) and are enriched for cell-signaling proteins and membrane-bound transporters (table S2). A total of 864 kb were targeted, including 351 kb of coding and 323 kb of untranslated (UTR) exon regions (database S1). More than 93% of target bases were successfully

[1]Department of Quantitative Sciences, GlaxoSmithKline (GSK), Research Triangle Park, NC 27709, USA. [2]Department of Ecology and Evolutionary Biology, University of California–Los Angeles, Los Angeles, CA 90095, USA. [3]Department of Quantitative Sciences, GSK, Stevenage SG1 2NY, UK. [4]Department of Genetics and Biostatistics, University of North Carolina–Chapel Hill, Chapel Hill, NC 27599, USA. [5]Department of Biostatistics, University of Michigan–Ann Arbor, Ann Arbor, MI 48109, USA. [6]BGI, Shenzhen 518083, China. [7]Department of Human Genetics, University of Michigan–Ann Arbor, Ann Arbor, MI 48109, USA. [8]Department of Biology, Novo Nordisk Foundation Center for Basic Metabolic Research, University of Copenhagen, Copenhagen 3393 9524, Denmark. [9]Department of Quantitative Sciences, GSK, Upper Merion, PA 19406, USA. [10]Department of Psychiatry, University of Michigan–Ann Arbor, Ann Arbor, MI 48109, USA.

*These authors contributed equally to this work.
†To whom correspondence should be addressed. E-mail: matthew.r.nelson@gsk.com (M.R.N); jnovembre@ucla.edu (J.N.)
‡These authors contributed equally to this work.

sequenced, at a median depth of 27 reads per site (*13*). Because rare variant discovery can easily be confounded with sequencing errors, we performed numerous experiments to demonstrate high data quality (table S3) (*13*). The sequenced subjects include two population samples (n = 1322 and 2059 subjects) and 12 disease collections (n = 125 to 1125 cases) (table S4). The self-reported ancestry of the sample was predominantly European (12,514), African American (594), and South Asian (567). Some of the following analyses focus on the European subset, which is well-powered to investigate rare variants. On the basis of our sample size, we expect that 94% of variant alleles with minor allele frequency (MAF) of 0.01% in Europeans were sampled at least once.

Sequencing revealed an abundance of rare (MAF < 0.5%) single-nucleotide variants (SNVs) compared with common variants (Fig. 1, A and B). We observed on average 1 variant per 17 base pair (bp) in the overall sample and 1 variant per 21 bp in the Europeans (table S5). Among all variants, more than 95% were rare (MAF ≤ 0.5%), and more than 74% were observed in only one or two subjects. Approximately 90% of rare variants were not previously reported, as opposed to ~5% of common variants (MAF > 0.5%) (fig. S1). For the large European subset, Watterson's θ_W—a metric of genetic diversity (Table 1)—was much larger (40.38×10^{-4}) than in previous smaller-scale studies and an order of magnitude larger than the pairwise metric θ_π (3.96×10^{-4}). We observed a third allele at 2.0% of variable sites, and among those, 1.6% had a fourth allele. We found between 1.2 and 1.9 nondiallelic SNVs per kilobase of sequence (fig. S2), which tended to occur at sites under lower evolutionary conservation (fig. S3) (*13*). The rate of variant discovery remained nearly constant with increasing sample size (Fig. 2A). We project 111 to 153 variants per kilobase in a sample of 100,000 Europeans and 337 to 452 variants per kilobase in a sample of 1 million (Fig. 2, A and B).

These patterns are at odds with notions that human genetic diversity can be summarized by use of an effective population size (N_e) of 10,000 individuals (*14*). An N_e of 10,000 individuals is predictive of the average pairwise differences between human sequences (Table 1, θ_π) and is reflective of our emergence from a small population in Africa (*15*). However, the excess of rare variants observed here ($\theta_W >> \theta_\pi$) is a signature of the rapid growth and large population sizes that typify more recent human demographic history (*8*). When we fit a demographic model to the fourfold degenerate synonymous (S) variants in Europeans, we obtained a maximum-likelihood estimate for a recent growth rate of 1.7% [95% confidence interval (CI) = 1.2 to 2.3%] and a recent European effective population size

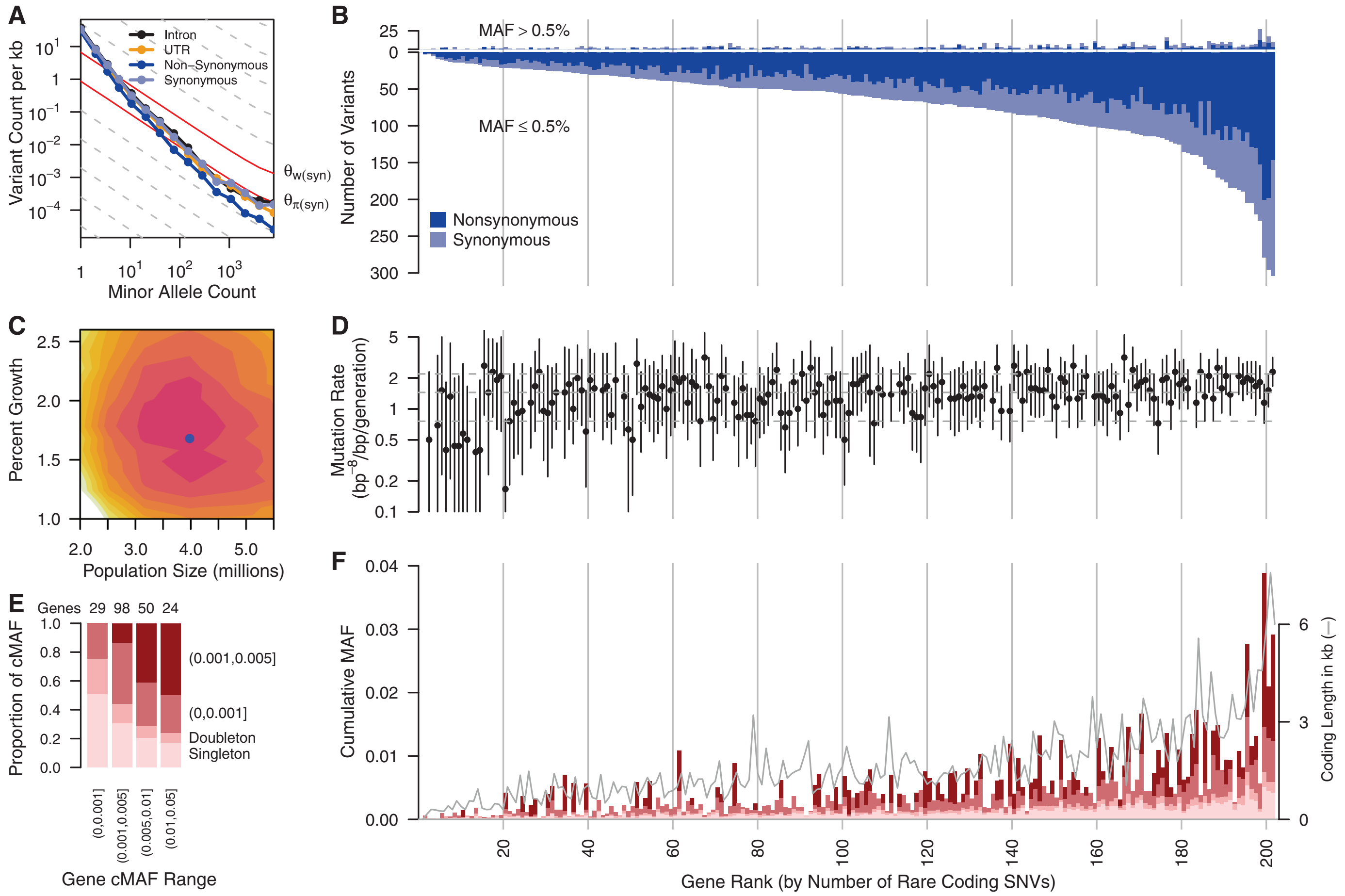

Fig. 1. (**A**) Frequency spectrum of variants relating the number of variants per kilobase within minor allele counts. Solid red lines provide expectations from nucleotide diversity (θ_π) and the number of segregating sites (θ_W). (**B**) The number of common (MAF > 0.5%, above the origin) and rare (MAF ≤ 0.5%, below the origin) coding variants observed in each gene are shown as stacked bars of NS and S variants. (**C**) Log-likelihood surface of European population growth (r) and population size (N_e) in a demographic model. Colored contours correspond to 2 log-likelihood intervals. The blue point is the maximum likelihood estimate of r and N_e. (**D**) Per-gene mutation rates with 2 log-likelihood intervals. Horizontal lines are 10th, 50th and 90th mutation rate percentiles. Seven genes on the X chromosome and four genes with low target coverage or yielding too few common variants for inference (*ADRB3*, *CCR5*, *MIF*, and *PTGER1*) were excluded. (**E**) Proportion of rare cMAF accounted for by SNVs of increasing frequency. (**F**) Proportion of rare variants in four cMAF ranges falling within the MAF categories shown in (E). The successfully sequenced coding length of each gene (in kilobase) is overlaid as a gray line. cMAFs in (E) and (F) are for amino acid–changing variants in each gene predicted to be damaging or are evolutionarily conserved (phyloP ≥ 2). Genes in (B), (D), and (F) are ordered by number of rare coding variants per gene, and vertical lines correspond to rank deciles.

of 4.0 million (95% CI = 2.5 million to 5.0 million) (Fig. 1C).

Taking advantage of the large size of this study for population genetics inference (*8*, *16*), we estimated mutation rates for each gene (Fig. 1D) (*13*) and obtained a median estimate of 1.38×10^{-8} per base pair per generation, with 90% of estimates falling between 1.7×10^{-9} and 2.4×10^{-8}. Incorporating singleton discovery false negative rates from 2 to 8% resulted in median estimates no greater than 1.45×10^{-8}. These population-genetic–based rate estimates are similar to recent pedigree-based mutation rate estimates of 1.36×10^{-8} per base pair per generation (*17*) and 1.17×10^{-8} per base pair per generation (*13*, *18*). Further, these data reject a model of uniform mutation rates across genes ($P < 2 \times 10^{-8}$) and show synonymous mutation rates are correlated with the number of nonsynonymous (NS) rare variants ($P = 0.04$) and guanine-cytosine content ($P < 2.4 \times 10^{-9}$) (*13*).

The excess of rare variants observed in coding regions is also due to an abundance of NS variants segregating at low frequencies that are not seen at more common variant frequencies as a result of purifying selection. Summing across all frequencies of variant sites, S and intronic variants occurred more frequently (~70 variants per kilobase each) as compared with UTR and NS variant sites (~55 and ~45 per kilobase of UTR or NS sequence, respectively) (Fig. 2A). Yet, examining the abundance of rare variants across functional categorizations of variant sites reveals little difference among classes when minor allele count is low (Fig. 1A). These patterns are likely due to an equal input of mutations for each category followed by purifying selection preventing deleterious NS and UTR variants from reaching higher frequencies (*13*, *19*). The ratio of NS:S in singletons is close to that expected among new mutations and then decreases with increasing frequency (Fig. 2C). Using the approach of (*2*), we estimate that although ~70% of all NS singletons in our sample are sufficiently deleterious that they will never reach frequencies >5%, only 13% of new NS mutations appear so deleterious that they would not be observed even as singletons in a sample of this size (*13*), putting an upper bound on the frequency of dominant lethal mutations (*15*). The output of functional prediction algorithms (Fig. 2, D and E) also suggest that rare variants are enriched for damaging variants.

On average, each subject carried a rare minor allele at 0.02% of all NS sites, of which ~56% are expected to be deleterious enough to never be fixed. More than 0.3% of sequenced subjects

Table 1. Comparison of classical population genetic measures of sequence diversity across studies.

Study	Number of genes*	Sample size	Sample†	Length (Mb)	θ_π ($\times 10^{-4}$)	θ_W ($\times 10^{-4}$)
Akey (*27*)	132	23	EU	2.50	3.41	7.35
SeattleSNPs (*28*)	213	23	EU	7.26	6.81	6.36
Ahituv (*29*)	58	757	EU	0.13	4.32	10.11
Current study	202	500‡	EU	0.74	3.96	8.79
		11,000	EU	0.74	3.96	40.38
		500‡	SA	0.69	4.04	10.67
Akey *et al.*	132	24	AA	2.50	4.49	12.10
SeattleSNPs	213	24	AA	7.26	8.97	10.15
Current study	202	500‡	AA	0.70	4.89	13.78

*Studies differ in the relative proportion of coding and noncoding sequences. †Ancestry is indicated as EU, European; AA, African-American; SA, South Asian. ‡Sampled to n = 500 subjects.

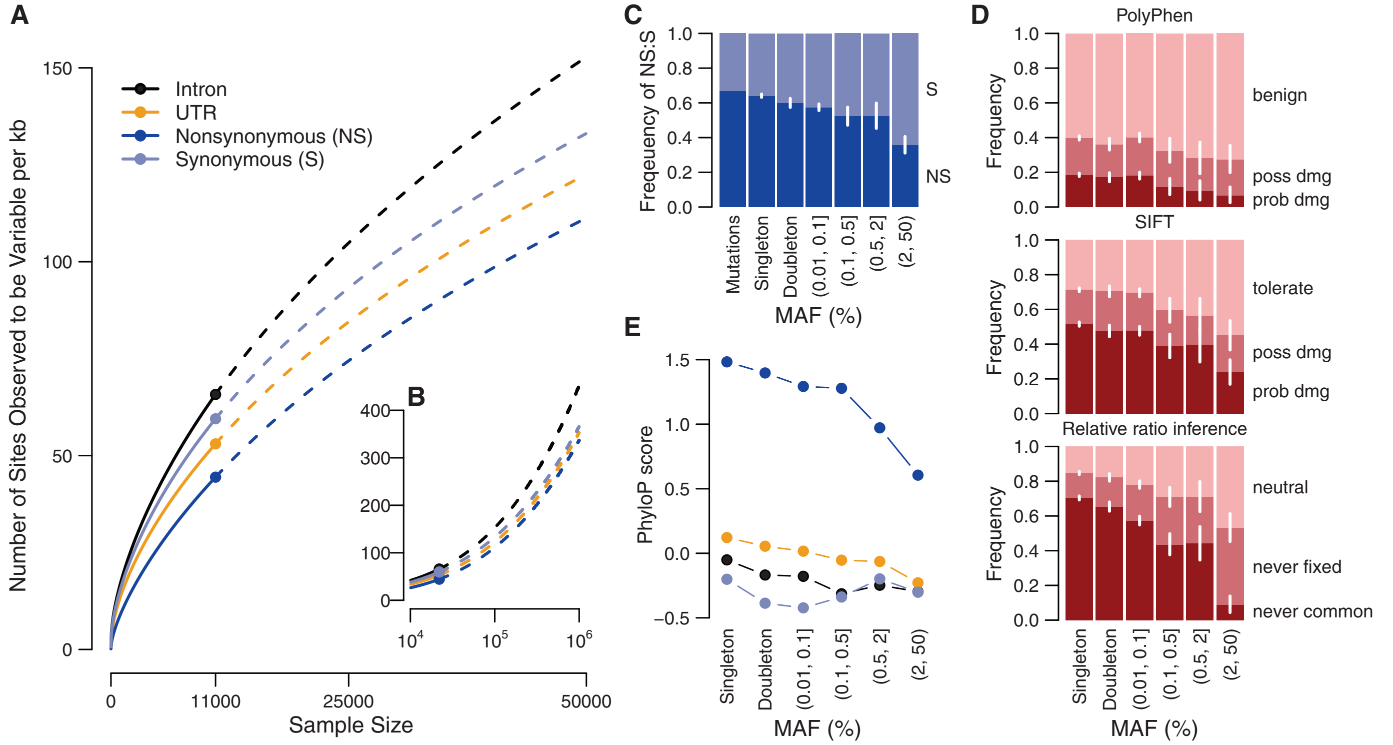

Fig. 2. (**A** and **B**) Number of variants per kilobase of intronic, UTR, NS, or S sequence with sample size increasing to 50,000 (A) and 1 million (B) Europeans. Observed numbers are given as a dot, and solid and dashed lines indicate hypergeometric expectations and jackknife projections, respectively. (**C**) Expected ratios of NS to S variants in the absence of selection and observed ratios for different MAF bins. (**D**) The proportion of NS variants predicted to be benign, possibly damaging, or probably damaging by use of PolyPhen or SIFT and the proportion of NS variants that is neutral, deleterious so that they will never become common (MAF > 5%), or never be fixed in Europeans as predicted by the relative ratios of NS:S variant abundances observed at different MAF (*2*). In (C) and (D), 95% CIs are represented by white lines. (**E**) phyloP score for intronic, UTR, NS, and S variants for different MAF bins.

carried at least one mutation reported to be a dominant cause of disease (table S6) (*13*). We also identified variants at 0.5% < MAF ≤ 2%, the so-called goldilocks variants (*20*), in that they would be common enough to be detected in large population samples and rare enough to be enriched for variants under purifying selection (Fig. 2, C to E). In the European sample, we observed 105 amino acid–changing variants in 73 genes falling within this frequency range. Half of these were predicted to be functionally damaging, relative to 31% of more common coding SNVs (>2%) and 65% of singletons. By comparison, we found 210 goldilocks variants in African Americans and 132 in South Asians, supporting the value of non-European samples for the genetic analysis of complex traits (*21*).

Rare variants can be tested in aggregate for an association with disease (*6*), in which the power of the test is strongly correlated with the cumulative MAF (cMAF) of potentially deleterious SNVs within each gene (Fig. 1, E and F, and figs. S4 and S5). Thirty-seven percent of genes had cMAFs > 0.5% of rare alleles predicted to be deleterious. We tested associations of common variants individually and rare coding variants in aggregate with the diseases represented in this study (*13*). When possible, we matched controls with cases using genome-wide genetic similarity. Nevertheless, type 1 error rate inflation consistent with effects of population stratification was observed (table S7 and fig. S6) and was worse for rare variant tests. There were no statistically significant rare variant associations and thus no compelling evidence connecting any genes with the studied diseases. Of 13 more closely examined genes reported to be associated with six of the diseases investigated (table S8) (*22*), only the association of rare variants in *IL6* with multiple sclerosis was noteworthy (OR = 12, $P = 0.007$) (table S9).

Because rare variants are typically the result of recent mutations, they are expected to be geographically clustered or even private to specific populations. Using a measure of variant sharing between two samples (*7*), we found that for common variants, any two European populations appear to be panmictic, whereas for rare variants, European populations show lower levels of sharing (fig. S7). In general, the level of sharing depends on geographic distance, with the dependence increasing substantially with decreasing allele frequency (fig. S8). The Finnish population shows substantially lower levels of sharing with other European populations than predicted by geographic distance, which is consistent with hypotheses of a historical Finnish demographic bottleneck (*23*). Levels of rare variant sharing are even lower when comparing populations from distinct continents. Thus, catalogs of rare variants will need to be generated locally across the globe (*7*, *24*).

We found substantial variation in the total abundance of variants across populations, even within Europe (Fig. 3 and fig. S7D), which is likely due to demographic history. In particular, we observed a north-south gradient in the abundance of rare variants across Europe, with increased numbers of rare variants in Southern Europe and a very small number of variants among Finns, who had about one third as many variants as southern Europeans. The gradient is consistent with observed gradients in haplotype diversity (*25*) and a Finnish ancestral bottleneck (*23*). Association mapping approaches based on rare variant diversity levels will be more susceptible to subtle effects of population stratification (*26*) and more likely to result in false-positive disease associations.

To evaluate our conclusions relative to the rest of the genome, we compared the NS:S variant ratios of the sequenced genes with the entire coding genome within the low-coverage CEU 1000 Genomes Project data. The average per subject NS:S ratio from our 202 genes was 0.54, whereas all other genes had an average ratio of 0.94 ($P < 10^{-15}$) (fig. S9). By comparison, genes found in Online Mendelian Inheritance in Man (OMIM) and the genome-wide association studies catalog (*22*) had average ratios of 0.75 and 0.78, respectively. This implies that the genes in this study are under stronger purifying selection, which is consistent with their choice as drug targets and importance to human health. Hence, our results cannot be simply extrapolated to the whole exome. Instead, it is likely that our results underestimate the average genetic diversity that will be found in more typical human gene-coding regions, primarily regarding the amount of NS variation.

This large-scale resequencing study provides a unique description of variation for 202 drug target genes and insight into the very rare spectrum of variation. Although sequencing error might be a concern, we show that the error rates in this study are low (table S3). Another caveat is that our inference of demographic parameters and mutation rates ignores the effects of background selection on synonymous variants. Despite these caveats, the results show there is an abundance of rare variation in human populations and that surveys of common variants are only observing a small fraction of the genetic diversity in any gene. Further, much of the rare variation in coding regions appears to be functional and may be crucial for yielding insights into the genetic basis of human disease. Because the genes studied are related to drug discovery, development, or repositioning efforts, this work has potential to help investigate drug target biology and drug response.

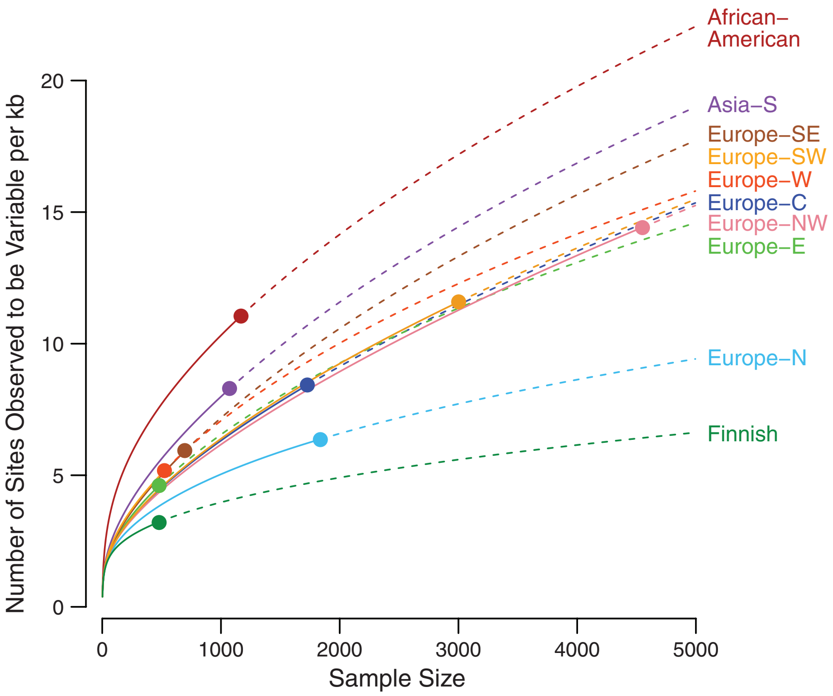

Fig. 3. Number of variants per kilobase of sequence with sample sizes increasing to 5000 people for multiple populations. Observed numbers are given as a dot, and solid and dashed lines indicate hypergeometric expectations and jackknife projections, respectively.

References and Notes

1. J. K. Pritchard, *Am. J. Hum. Genet.* **69**, 124 (2001).
2. G. V. Kryukov, L. A. Pennacchio, S. R. Sunyaev, *Am. J. Hum. Genet.* **80**, 727 (2007).
3. G. T. Marth *et al.*, 1000 Genomes Project, *Genome Biol.* **12**, R84 (2011).
4. T. A. Manolio *et al.*, *Nature* **461**, 747 (2009).
5. E. E. Eichler *et al.*, *Nat. Rev. Genet.* **11**, 446 (2010).
6. J. Asimit, E. Zeggini, *Annu. Rev. Genet.* **44**, 293 (2010).
7. S. Gravel *et al.*, 1000 Genomes Project, *Proc. Natl. Acad. Sci. U.S.A.* **108**, 11983 (2011).
8. A. Coventry *et al.*, *Nat. Commun.* **1**, 131 (2010).
9. T. Ohta, *Nature* **246**, 96 (1973).
10. S. H. Williamson *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **102**, 7882 (2005).
11. H. J. Muller, *Am. J. Hum. Genet.* **2**, 111 (1950).

12. A. P. Russ, S. Lampel, *Drug Discov. Today* **10**, 1607 (2005).
13. Materials and methods are available as supplementary materials on *Science* Online.
14. M. A. Jobling, M. Hurles, C. Tyler-Smith, *Human Evolutionary Genetics: Origins, Peoples and Disease* (Garland Science, 2003).
15. M. Livi-Bacci, *A Concise History of World Population* (Wiley-Blackwell, ed. 2, 2007), pp. 1 to 250.
16. J. Wakeley, T. Takahashi, *Mol. Biol. Evol.* **20**, 208 (2003).
17. P. Awadalla *et al.*, *Am. J. Hum. Genet.* **87**, 316 (2010).
18. D. F. Conrad *et al.*; 1000 Genomes Project, *Nat. Genet.* **43**, 712 (2011).
19. P. W. Messer, *Genetics* **182**, 1219 (2009).
20. A. L. Price *et al.*, *Am. J. Hum. Genet.* **86**, 832 (2010).
21. I. K. Kotowski *et al.*, *Am. J. Hum. Genet.* **78**, 410 (2006).
22. L. A. Hindorff *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **106**, 9362 (2009).
23. E. Salmela *et al.*, *PLoS ONE* **3**, e3519 (2008).
24. C. D. Bustamante, E. G. Burchard, F. M. De la Vega, *Nature* **475**, 163 (2011).
25. O. Lao *et al.*, *Curr. Biol.* **18**, 1241 (2008).
26. E. S. Lander, N. J. Schork, *Science* **265**, 2037 (1994).
27. J. M. Akey *et al.*, *PLoS Biol.* **2**, e286 (2004).
28. SeattleSNPs, http://pga.gs.washington.edu (2012).
29. N. Ahituv *et al.*, *Am. J. Hum. Genet.* **80**, 779 (2007).

Acknowledgments: We thank GSK colleagues who advised on the selection of genes and collections, especially W. Anderson, L. Condreay, P. Agarwal, A. Hughes, J. Rubio, C. Spraggs, and D. Waterworth; the sample preparation team, especially J. Charnecki, M. E. Volk, D. Duran, D. Briley, and K. King for data preparation; A. Slater for subject selection and preparation of genome-wide genotype data; E. Woldu for capillary sequencing; A. Nelsen, S. Buhta-Halburnt, L. Amos, and J. Forte for consent review; M. Lawson for assistance in running the association analyses; J. Brown for discussions about gene feature analyses; S. Ghosh for providing reviews of the manuscript; and G. Tian, H. Jiang, Z. Su, X. Sun, L. Yang, and X. Zhang at BGI for sequencing. We acknowledge the work of collaborating clinicians and researchers who contributed to recruiting and characterizing subjects (*13*). J.N. and D.W. were supported by a Searle Scholars Program award to J.N. D.K. is supported by a NIH Genome Analysis training grant. All variants described in this study have been submitted to dbSNP; accession nos. are included in database S2. Subject-level sequence data for CoLaus and LOLIPOP studies are available in dbGaP. Additional subject-level sequence data can be made available upon request from the authors under a Data Transfer Agreement for the purpose to understand, assess, or extend the conclusions of this paper.

Supplementary Materials
www.sciencemag.org/cgi/content/full/science.1217876/DC1
Materials and Methods
Supplementary Text
Figs. S1 to S15
Tables S1 to S17
References (*30–85*)
Databases S1 to S3

14 December 2011; accepted 3 May 2012
Published online 17 May 2012;
10.1126/science.1217876

Recurrent Hemizygous Deletions in Cancers May Optimize Proliferative Potential

Nicole L. Solimini,[1] Qikai Xu,[1] Craig H. Mermel,[2,3] Anthony C. Liang,[1] Michael R. Schlabach,[1]* Ji Luo,[1]† Anna E. Burrows,[1] Anthony N. Anselmo,[1] Andrea L. Bredemeyer,[1] Mamie Z. Li,[1] Rameen Beroukhim,[2,3,4] Matthew Meyerson,[2,3] Stephen J. Elledge[1]‡

Tumors exhibit numerous recurrent hemizygous focal deletions that contain no known tumor suppressors and are poorly understood. To investigate whether these regions contribute to tumorigenesis, we searched genetically for genes with cancer-relevant properties within these hemizygous deletions. We identified STOP and GO genes, which negatively and positively regulate proliferation, respectively. STOP genes include many known tumor suppressors, whereas GO genes are enriched for essential genes. Analysis of their chromosomal distribution revealed that recurring deletions preferentially overrepresent STOP genes and underrepresent GO genes. We propose a hypothesis called the cancer gene island model, whereby gene islands encompassing high densities of STOP genes and low densities of GO genes are hemizygously deleted to maximize proliferative fitness through cumulative haploinsufficiencies. Because hundreds to thousands of genes are hemizygously deleted per tumor, this mechanism may help to drive tumorigenesis across many cancer types.

Cancer progression is directed by alterations in oncogenes and tumor suppressor genes (TSGs) that provide a competitive advantage to increase proliferation, survival, and metastasis (*1–3*). The cancer genome is riddled with amplifications, deletions, rearrangements, point mutations, loss of heterozygosity (LOH), and epigenetic changes that collectively result in tumorigenesis (*4–7*). How these changes contribute to the disease is a central question in cancer biology. In his "two-hit hypothesis," Knudson proposed that two mutations in the same gene are required for tumorigenesis, indicating a recessive disease (*8*). In addition, there are now several examples of haploinsufficient TSGs (*9–11*). Current models do not explain the recent observation that hemizygous recurrent deletions are found in most tumors (*12*, *13*). Whether multiple genes within such regions contribute to the tumorigenic phenotype remains to be elucidated.

Recent analysis of 3131 tumors revealed 82 regions of recurrent focal deletion (*13*), averaging six deletions per tumor and 24 genes per deletion (Fig. 1C, fig. S1A, and table S1) (*14*). Breast, gastric, bladder, pancreatic, and ovarian cancers average ≥10 deletions/tumor (Fig. 1A). Several possible explanations exist for the roles of these deletions in tumorigenesis. First, they may contain a recessive TSG where mutation or epigenetic silencing of the second allele is necessary for tumorigenesis. Second, they may recur because they mark unstable genomic regions, such as fragile sites (*12*). Finally, it is possible that single-copy loss may provide a selective advantage irrespective of changes in the remaining allele.

To address the possibility that recurrent deletions are enriched for recessive TSGs, we analyzed these regions for the presence of known or putative recessive TSGs. For this purpose we used a list from the Cancer Gene Census (*15*) and a list of putative TSGs that we identified with homozygous loss-of-function (termination codon or frameshift) mutations from whole-genome sequencing of 526 tumors in the Catalogue of Somatic Mutations in Cancer (COSMIC) (Fig. 1B and tables S2 and S3) (*16*). Only 14 of 82 recurrent deletions contained a known TSG, and only 10 had a mutant or putative TSG, 6 of which were in a region with a known TSG (Fig. 1C and fig. S1). Thus, only 18 of 82 deletions can be explained by known or putative recessive TSGs. This number may increase if gene silencing is as prevalent as point mutation for gene inactivation, but this remains to be determined across all cancers. These data suggest that in addition to the two-hit mechanism, an alternative mechanism may function to provide a selective advantage to these deletions.

Of the many altered processes promoting tumorigenesis, proliferation is likely to encompass the most genes, as it is integrated into all developmental decisions. Cancer evolution relies on alterations that provide incremental increases in cell number—a function of cell duplication frequency coupled with cell survival efficiency. The average fitness increase of a single alteration in tumors is estimated to be 0.4% (*17*). Because subtle changes in proliferation rates can have profound effects on tumor fitness and clonal selection, we examined whether recurrent deletions affect regulators of cell proliferation. We

[1]Department of Genetics, Harvard University Medical School, and Division of Genetics, Howard Hughes Medical Institute, Brigham and Women's Hospital, Boston, MA 02115, USA. [2]Departments of Medical Oncology and Cancer Biology and Center for Cancer Genome Discovery, Dana-Farber Cancer Institute, Boston, MA 02215, USA. [3]Cancer Program, Broad Institute of MIT and Harvard, Cambridge, MA 02141, USA. [4]Departments of Medicine, Harvard University Medical School and Brigham and Women's Hospital, Boston, MA 02115, USA.

*Present address: Novartis Institute for Biomedical Research, Cambridge, MA 02139, USA.

†Present address: National Cancer Institute, Bethesda, MD 20892, USA.

‡To whom correspondence should be addressed. E-mail: selledge@genetics.med.harvard.edu