Relative Impact of Nucleotide and Copy Number Variation on Gene Expression Phenotypes

Barbara E. Stranger, *et al.*

Science **315**, 848 (2007);
DOI: 10.1126/science.1136678

References and Notes

1. J. K. Rose, M. A. Whitt, in *Fields' Virology*, D. M. Knipe, P. M. Howley, Eds. (Lippincott, Williams & Wilkins, Philadelphia, ed. 4, 2001), pp. 1221–1244.
2. B. L. Rao *et al.*, *Lancet* **364**, 869 (2004).
3. I. Le Blanc *et al.*, *Nat. Cell Biol.* **7**, 653 (2005).
4. L. H. Luo, Y. Li, R. M. Snyder, R. R. Wagner, *Virology* **163**, 341 (1988).
5. A. Benmansour *et al.*, *J. Virol.* **65**, 4198 (1991).
6. S. B. Vandepol, L. Lefrancois, J. J. Holland, *Virology* **148**, 312 (1986).
7. Y. Gaudin, C. Tuffereau, D. Segretain, M. Knossow, A. Flamand, *J. Virol.* **65**, 4853 (1991).
8. R. W. Doms, D. S. Keller, A. Helenius, W. E. Balch, *J. Cell Biol.* **105**, 1957 (1987).
9. Y. Gaudin, *Subcell. Biochem.* **34**, 379 (2000).
10. Y. Gaudin, R. W. Ruigrok, M. Knossow, A. Flamand, *J. Virol.* **67**, 1365 (1993).
11. P. Durrer, Y. Gaudin, R. W. Ruigrok, R. Graf, J. Brunner, *J. Biol. Chem.* **270**, 17575 (1995).
12. B. L. Fredericksen, M. A. Whitt, *Virology* **217**, 49 (1996).
13. C. C. Pak, A. Puri, R. Blumenthal, *Biochemistry* **36**, 8890 (1997).
14. F. A. Carneiro, A. S. Ferradosa, A. T. Da Poian, *J. Biol. Chem.* **276**, 62 (2001).
15. S. Roche, Y. Gaudin, *Virology* **297**, 128 (2002).
16. Y. Gaudin, C. Tuffereau, P. Durrer, A. Flamand, R. W. Ruigrok, *J. Virol.* **69**, 5528 (1995).
17. S. Roche, S. Bressanelli, F. A. Rey, Y. Gaudin, *Science* **313**, 187 (2006).
18. M. Kielian, F. A. Rey, *Nat. Rev. Microbiol.* **4**, 67 (2006).
19. P. A. Bullough, F. M. Hughson, J. J. Skehel, D. C. Wiley, *Nature* **371**, 37 (1994).
20. J. J. Skehel, D. C. Wiley, *Cell* **95**, 871 (1998).
21. H. S. Yin, R. G. Paterson, X. Wen, R. A. Lamb, T. S. Jardetzky, *Proc. Natl. Acad. Sci. U.S.A.* **102**, 9288 (2005).
22. Y. Modis, S. Ogata, D. Clements, S. C. Harrison, *Nature* **427**, 313 (2004).
23. S. Bressanelli *et al.*, *EMBO J.* **23**, 728 (2004).
24. D. L. Gibbons *et al.*, *Nature* **427**, 320 (2004).
25. E. E. Heldwein *et al.*, *Science* **313**, 217 (2006).
26. Y. Gaudin, R. W. Ruigrok, C. Tuffereau, M. Knossow, A. Flamand, *Virology* **187**, 627 (1992).
27. Single-letter abbreviations for the amino acid residues are as follows: A, Ala; C, Cys; D, Asp; E, Glu; F, Phe; G, Gly; H, His; I, Ile; K, Lys; L, Leu; M, Met; N, Asn; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; and Y, Tyr.
28. F. Forster, O. Medalia, N. Zauberman, W. Baumeister, D. Fass, *Proc. Natl. Acad. Sci. U.S.A.* **102**, 4729 (2005).
29. P. Zhu *et al.*, *Nature* **441**, 847 (2006).
30. H. S. Yin, X. Wen, R. G. Paterson, R. A. Lamb, T. S. Jardetzky, *Nature* **439**, 38 (2006).
31. Y. Gaudin, H. Raux, A. Flamand, R. W. Ruigrok, *J. Virol.* **70**, 7371 (1996).
32. E. Krissinel, K. Henrick, in *CompLife 2005*, M. R. Berthold, R. Glen, K. Diederichs, O. Kohlbacher, I. Fischer, Eds., *Lecture Notes in Bioinformatics*, vol. 3695 (Springer-Verlag, Berlin, 2005), pp. 163–174.
33. F. Lafay, A. Benmansour, K. Chebli, A. Flamand, *J. Gen. Virol.* **77**, 339 (1996).
34. Y. Gaudin, *J. Virol.* **71**, 3742 (1997).
35. C. Tuffereau, J. Benejean, D. Blondel, B. Kieffer, A. Flamand, *EMBO J.* **17**, 7250 (1998).
36. L. V. Chernomordik, M. M. Kozlov, *Cell* **123**, 375 (2005).
37. C. M. Carr, C. Chaudhry, P. S. Kim, *Proc. Natl. Acad. Sci. U.S.A.* **94**, 14306 (1997).
38. W. L. Delano, The PyMOL Molecular Graphics System (DeLano Scientific, San Carlos, CA, 2002), available at www.pymol.org.

39. We thank A. Flamand for constant support on this project; J. Lepault, R. Ruigrok, M. Knossow, A. Benmansour, C. Tuffereau, and D. Blondel for helpful discussions at different stages of this work; and C. Maheu for virus purification. Data collections were performed at the Swiss Light Source (SLS), Paul Scherrer Institut, Villigen, Switzerland, and at the European Synchrotron Radiation Facility (ESRF), Grenoble, France. We acknowledge the help of T. Tomizaki (beamline X06SA, SLS), G. Leonard and D. Bourgeois (beamlines ID29 and ID23-2, ESRF), and S. Duquerroy and G. Squires in data collection. We acknowledge support from the CNRS and INRA, the CNRS program "Physique et Chimie du Vivant," the INRA Animal health department program "Les virus des animaux et leurs interactions avec la cellule," the Ministère de l'éducation nationale, de la recherche et de la technologie program "Action Concertée Incitative blanche," and the Agence Nationale de la Recherche program. S.R. was the recipient of an Agence Nationale de Recherche sur le Sida fellowship during part of this project. Coordinates and structure factors have been deposited with the Protein Data Bank under accession code 2j6j.

Supporting Online Material

www.sciencemag.org/cgi/content/full/315/5813/843/DC1
Materials and Methods
Figs. S1 to S5
Table S1
References
Movie S1

29 September 2006; accepted 3 January 2007
10.1126/science.1135710

Relative Impact of Nucleotide and Copy Number Variation on Gene Expression Phenotypes

Barbara E. Stranger,[1] Matthew S. Forrest,[1] Mark Dunning,[2] Catherine E. Ingle,[1] Claude Beazley,[1] Natalie Thorne,[2] Richard Redon,[1] Christine P. Bird,[1] Anna de Grassi,[3] Charles Lee,[4,5] Chris Tyler-Smith,[1] Nigel Carter,[1] Stephen W. Scherer,[6,7] Simon Tavaré,[2,8] Panagiotis Deloukas,[1] Matthew E. Hurles,[1]* Emmanouil T. Dermitzakis[1]*

Extensive studies are currently being performed to associate disease susceptibility with one form of genetic variation, namely, single-nucleotide polymorphisms (SNPs). In recent years, another type of common genetic variation has been characterized, namely, structural variation, including copy number variants (CNVs). To determine the overall contribution of CNVs to complex phenotypes, we have performed association analyses of expression levels of 14,925 transcripts with SNPs and CNVs in individuals who are part of the International HapMap project. SNPs and CNVs captured 83.6% and 17.7% of the total detected genetic variation in gene expression, respectively, but the signals from the two types of variation had little overlap. Interrogation of the genome for both types of variants may be an effective way to elucidate the causes of complex phenotypes and disease in humans.

Understanding the genetic basis of phenotypic variation in human populations is currently one of the major goals in human genetics. Gene expression (the transcription of DNA into mRNA) has been interrogated in a variety of species and experimental scenarios in order to investigate the genetic basis of variation in gene regulation (*1*–*8*), as well as to tease apart regulatory networks (*9*, *10*). In some respects, a comprehensive survey of gene expression phenotypes (steady-state levels of mRNA) serves as a proxy for the breadth and nature of phenotypic variation in human populations (*11*). Much of the observed variation in mRNA transcript levels may be compensated at higher stages of regulatory networks, but an understanding of the nature of genetic variants that affect gene expression will provide an essential framework and model for elucidating the causes of other types of phenotypic variation. Single-nucleotide polymorphisms (SNPs) have long been known to be associated with phenotypic variation either through direct causal effects or by serving as proxies for other causal variants with which they are highly correlated (i.e., in linkage disequilibrium) (*1*, *2*, *12*). An understanding of this association has been facilitated by the validation of millions of SNPs by the International HapMap project (*13*). However, during the last few years, structural variants, such as copy number variants (CNVs)—defined as DNA segments that are 1 kb or larger in size present at variable copy number in comparison with a reference genome (*14*)—have attracted much attention (*2*). It has become apparent that they are quite common in the human genome (*15*–*19*) and can have dramatic phenotypic consequences as a result of altering gene dosage, disrupting coding se-

[1]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, UK. [2]Department of Oncology, University of Cambridge, Cancer Research UK Cambridge Research Institute, Li Ka Shing Centre, Robinson Way, Cambridge CB2 0RE, UK. [3]Istituto di Tecnologie Biomediche-Sezione di Bari, Consiglio Nazionale della Ricerche (CNR), 70126 Bari, Italy. [4]Department of Pathology, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115, USA. [5]Broad Institute of Harvard and Massachusetts Institute of Technology, Cambridge, MA 02142, USA. [6]The Centre for Applied Genomics and Program in Genetics and Genomic Biology, The Hospital for Sick Children, MaRS Centre, Toronto, Ontario, M5G 1L7, Canada. [7]Department of Molecular and Medical Genetics, University of Toronto, Toronto, Ontario, Canada. [8]Program in Molecular and Computational Biology, University of Southern California, Los Angeles, CA 90089–2910, USA.

*To whom correspondence should be addressed. E-mail: md4@sanger.ac.uk (E.T.D.); meh@sanger.ac.uk (M.E.H.)

quences, or perturbing long-range gene regulation (*20*, *21*). Evidence has been presented that increased copy number can be positively (*18*, *22*) or negatively (*23*) correlated with gene expression levels (for example, deletion of a transcriptional repressor could serve to elevate gene expression) but the relative contribution of such large genetic variants (i.e., CNVs) and smaller variants (i.e., SNPs) to phenotypic variation has not been evaluated. It is also still unknown whether SNPs can serve as proxies to CNVs (*24*, *25*) and whether the complex nature of some CNVs requires that they be surveyed directly (*26*). We have used the phase I HapMap SNPs (*13*) and the recently described CNV data ascertained in the same HapMap populations (*26*) for correlation with genomewide gene expression variation in the same individuals.

Gene expression was interrogated in lymphoblastoid cell lines of all 210 unrelated HapMap individuals (*13*) from four populations (CEU: 60 Utah residents with ancestry from northern and western Europe; CHB: 45 Han Chinese in Beijing; JPT: 45 Japanese in Tokyo; YRI: 60 Yoruba in Ibadan, Nigeria) in four technical replicates (see Methods). Out of the 47,294 transcripts that were interrogated, the normalized values for 14,925 transcripts (14,072 genes) were included in the analysis [see Methods and (*27*)]. The SNP genotypes from phase I HapMap (*28*) were used in the analysis (see Methods). CNV data were represented by $\log_2$ ratios from comparative genomic hybridization (CGH) of each HapMap individual against a common reference individual on an array comprising 26,574 large-insert clones covering 93.7% of the euchromatic portion of the genome (*26*, *29*). Log_2 ratios from two sets of clones were analyzed: the whole set of 24,963 autosomal clones (CGH clones) and the 1322 autosomal clones corresponding to CNVs present in at least two HapMap individuals (CNV clones) (*26*). We excluded genes on sex chromosomes because of their imbalance in males and females. We performed linear regression (on each of the four populations separately) between normalized quantitative gene expression values and SNP genotypes or clone $\log_2$ ratios that were near the gene (SNP position or clone midpoint within 1 Mb and 2 Mb, respectively, of the probe midpoint position). We used different window sizes for SNPs and clones because clones are large (median size of ~170 kb) and structural variants can exert long-range effects (*21*), so a 2-Mb window is more appropriate. Statistical significance was evaluated through the use of permutations (*30*), as previously described (*1*), and a corrected P value threshold of 0.001 was applied (see Methods). Repeated permutation exercises showed that our permutation thresholds were very stable (see table S1). We tested a large number of genes so an additional correction was required. This could be done either by adjusting the threshold to a new corrected threshold above which all genes are expected to be significant (e.g., Bonferroni correction) or by setting the threshold to a value that generates a satisfactory false-discovery rate (FDR). We have used the second, and we have estimated the FDR on the basis of the number of genes tested and have required that, in all cases, at least 80% of the genes called significant are estimated to be truly significant. Given that there are 14,072 genes that lie within 1 Mb of SNPs and within 2 Mb of the full set of CGH clones, and ~7150 genes that lie within 2 Mb from the CNV clones (from 7135 to 7191 depending on the population, owing to missing data), we expect this analysis to generate false-positive association signals for approximately 14 and 7 genes, respectively, in each population.

Of the 14,072 genes tested, we detected significant associations with at least one SNP for 323, 348, 370, and 411 genes for CEU, CHB, JPT, and YRI, respectively (e.g., Table 1 and table S1). These comprise a total of 888 nonredundant genes of which 331 (37%) were replicated at the same significance level in at least one other population, and of those, 67 (8%) were significant in all four populations (Table 2 and table S2). As expected, we have limited power to detect weak effects because of the small sample sizes: The minimum detected squared regression coefficient (r^2)—which reflects the proportion of expression variance accounted for by the linear association with allele counts—was 0.27. However, some very strong effects were detected that, in some cases, had an r^2 close to 1 (Fig. 1 and fig. S1). We detected a strong preference for associated SNPs to be close to their respective genes, most of which were within 100 kb of the interrogated expression probe (Fig. 1, A and C). In summary, we detected a large number of regions that appear to carry genetic variation affecting gene expression. To evaluate the effect of experimental variation and, hence, the robustness of our associations, we compared the list of gene expression associations from our previous study (*1*) in which we detected 63 expression associations significant at the 0.05 permutation threshold in the CEU population. Of those 63 expression phenotypes, 47 went into the current analysis, of which 43 (91.5%) were called significant at the same permutation threshold (0.05) in the same population. The previous study was performed with different batches of cells, by using RNA extracted in a different laboratory, with RNA levels quantified on a different type of array (custom versus genomewide array), so the high degree of experimental and statistical replication strongly suggests that the signals we detected are robust and stable to experimental variation in expression measurements.

Of the 14,072 genes tested, we detected significant associations with at least one of the 24,962 autosomal CGH clones in 85, 44, 58, and 96 genes in CEU, CHB, JPT, and YRI, respectively (238 nonredundant genes), of which 28 (12%) were replicated at the same significance level in at least one other population, and of those, 5 (2%) were significant in all four populations (Fig. 2, Table 1, and table S3 and figs. S2 and S3). Not all associated clones were within CNVs defined using the stringent criteria of (*26*) [119 out of 303 (39%) associated clones were previously defined as CNVs], and it is likely that some of these clones encompass smaller CNVs that are detectable though asso-

Table 1. Numbers of genes with significant associations to SNPs (SNP-probe distance < 1 Mb), all CGH clones (clone-probe distance < 2 Mb), or CNV clones (clone-probe distance < 2 Mb) as assessed by permutations, together with the numbers of overlaps between SNP-associated genes and CGH or CNV clone-associated genes (probe-variant distance < 1 Mb for both SNPs and clones) (see table S4).

Gene population	CNV (2 Mb)		SNP	CNV (1 Mb) + SNP overlap	
	CGH clones	CNV clones		CGH clones	CNV clones
			Permutation threshold 0.01		
CEU	362	138	643	14	15
CHB	221	110	673	10	9
JPT	319	134	752	13	14
YRI	481	166	815	14	11
Nonredundant	1246	451	1886	28	16
			Permutation threshold 0.001		
CEU	85	40	323	9	8
CHB	44	32	348	5	6
JPT	58	40	370	8	6
YRI	96	42	411	7	6
Nonredundant	238	99	888	15	12
			Permutation threshold 0.0001		
CEU	32	18	198	5	6
CHB	14	19	204	4	4
JPT	23	20	217	6	5
YRI	27	16	251	2	2
Nonredundant	69	39	526	8	8

ciations of $\log_2$ ratios across a population, but cannot be detected as extreme outliers in their $\log_2$ ratios in any one individual [as is required for classification as a CNV in (*26*)]—(see example below). For 36 common (minor allele frequency > 0.05) CNVs (encompassing 99 CGH clones), accurate CNV genotypes were available. We used these genotypes to validate the statistical power of performing association analysis using $\log_2$ ratios directly rather than genotypes. There was strong correlation between r^2 values or *P* values generated using the $\log_2$ ratio signals or the CNV genotypes (Pearson correlation coefficients > 0.9), indicating that $\log_2$ ratios can be used directly.

Little prior data exists on CNV-expression associations against which to compare and demonstrate the robustness of our associations. One recent study (*18*) demonstrated three associations between common deletions and gene expression in a subset of the CEU. Two of these deletions are covered by our CGH data. The reported expression association caused by the largest of these two deletions is also captured in our analysis (influencing *UGT2B17*), and we extend this observation to show that this deletion also affects the expression of three other nearby genes (*UGT2B7*, *UGT2B10*, and *UGT2B11*) and that these associations replicate across all four populations. The smaller deletion of only 18 kb, reported previously (*18*) as affecting expression of *GSTM1*, is below the expected resolution of the CGH data. Nonetheless, we observe an association that, although it does not pass our stringent permutation threshold (0.001), has significant nominal *P* values in all four populations (P_{CEU} = 0.0292; P_{YRI} = 0.0018; P_{JPT} = 0.0408; P_{CHB} = 0.0185). This suggests that effects of CNVs far smaller than genomic regions that met our criteria to be called a CNV within the CGH platform can be detected and replicated in multiple populations with our analysis.

Having investigated the potential contribution of CNV to variation in gene expression by using data from all CGH clones, we interrogated the nature of CNV effects on gene expression in finer detail by performing association tests of 1322 clones within high confidence CNVs (see above) with expression of the 14,072 genes, in order to generate a set of high stringency associations for which the presence of an underlying CNV has already been validated. Significant associations with at least one of the 1322 CNV clones were detected for 40, 32, 40, and 42 genes in CEU, CHB, JPT, and YRI, respectively (99 nonredundant genes) (table S4). Thirty-four of the 99 genes (34%) associated with CNV clones have a significant signal in at least two populations (Table 2), of which 7 (7%) were associated in all populations. Some CNV clones were associated with more than one gene in the same population; a notable example was a single CNV clone associated with expression of four genes in all populations (*UGT2B* genes, see above). CNVs detected by CGH can be classified into five classes: deletion, duplication, deletion and duplication at the same locus, multiallelic, and complex (*26*); we find all classes of CNV represented among the significant associations. Despite the clear preference for genes to lie close to their associated CNVs (Fig. 1, B and D), 53% of the expression probes associated with a CGH clone were located outside the CNVs encompassing that clone (*26*). This suggests that rather than altering gene dosage, about half the CNV effects are caused by disruption of the gene (some parts of the gene, but not the probe, are within in the CNV) or affect regulatory regions and other functional regions that have an impact on gene expression. When we extended our analysis to consider associations between genes and CNVs up to 6 Mb apart, we detected a few significant long-distance associations beyond 2 Mb (table S5). These types of long-range effects are becoming more apparent through recent studies looking in detail at specific genomic regions (*20*, *31*). A small minority (5 to 15%) of the significant CNV-expression associations have a negative correlation between copy number and gene expression, which suggests that not all the detected effects are of the conventional type, wherein gene expression levels increase with gene copy number (table S3). Almost all (32 out of 34) of the associations that are shared between populations also exhibit the same direction of correlation in all populations. The two exceptions could result from the CNVs being in linkage disequilibrium with different regulatory variants in different populations or because of SNP × CNV interactions. However, the strong bias toward positive correlations between copy number and expression levels implies that the vast majority of these associations are attributable to the CNV itself, and not to a linked variant.

We next determined whether the same associations were also captured by SNPs (Fig. 2 and figs. S3 and S4). We only considered those CGH clones or CNVs within 1 Mb of the probe so that the analysis is comparable to that of the SNPs (total of 188 and 84 genes for CGH clones and CNVs, respectively). We expect some of the CNVs to be correlated with SNPs via common genealogical history (linkage disequilibrium) and therefore their effect on gene expression would also be captured by SNP associations. Fewer than 20% (in all populations) of the detected CGH clone associations overlapped with SNP associations (Table 1), even when we included CGH and SNP associations with the same gene but in different populations [28 out of 188 (14%) genes with significant CGH clone associations also had a SNP association in any population]. The same is true of CNV clone associations: Only 15 of 84 genes (18%) with CNV clone associations within 1 Mb also had a SNP association in any population, and if we required the association in the same population, only 12 (14%) of genes had a SNP association. On the basis of previous work characterizing the patterns of linkage disequilibrium around CNVs (*26*), we considered that this low overlap between CNV or CGH clone associations with SNP associations might be due in part either to a low density of successfully genotyped SNPs around some CNVs or to the suppression of apparent LD by recurrent mutation at some CNVs. Segmental duplications (SDs) are the primary cause of low

Table 2. Sharing of associations between populations.

	CGH clone (2 Mb)	CNV clone (2 Mb)	SNP (1 Mb)
CEU-CHB-JPT-YRI	5	7	67
CEU-CHB-JPT	2	4	48
CEU-CHB-YRI	1	0	11
CEU-JPT-YRI	1	0	12
CHB-JPT-YRI	3	3	28
CEU-CHB	1	3	18
CEU-JPT	2	0	15
CEU-YRI	6	6	36
CHB-JPT	4	5	51
CHB-YRI	1	3	18
JPT-YRI	2	3	27
CEU only	67	20	116
CHB only	27	7	107
JPT only	39	18	122
YRI only	77	20	212
Sum	238	99	888
Gene associations in at least two populations	28	34	331
Percentage of total	0.12	0.34	0.37
Gene associations in single populations	210	65	557
Percentage of total	0.88	0.66	0.63

SNP densities in HapMap Phase I because of the difficulties in developing robust SNP genotyping assays within them (*13*). We did not observe enrichment of segmentally duplicated sequences within the CGH and CNV clones that did not share signals with SNPs relative to those CGH and CNV clones that did share signals with SNPs. However, we observe a 2.5-fold excess of compound CNVs [CNVs with more than one mutation event, on the basis of classification of the CNVs in (*26*)] in associations that are not shared with SNPs relative to those that are shared (Fisher's exact test: $P < 0.001$). Thus our analysis suggests that recurrent mutation is a likely factor reducing overlap between CNV and SNP associations.

CNV associations that were also detected with SNPs were clearly biased toward large effect sizes (tables S1 and S3). Of the 12 genes with both SNP and CNV associations in the same population, 8 shared the association in two or more populations (giving a redundant total across the four populations of 26 shared CNV and SNP associations). The ratio of 8 out of 12 (67%) population shared associations is larger than that observed in all CNV associations (34 out of 99 = 34%) potentially suggesting that associations with higher frequency, older CNVs are more likely to be captured by SNPs. For the 26 associations (representing 12 genes; see above) captured both by CNVs and SNPs in the same population, we observed that SNPs and CNVs were themselves highly correlated for 23 out of 26 SNP-CNV pairs (Pearson correlation, $P < 0.001$) suggesting that for these cases the CNV and SNP captured the same effect, and that only a small fraction of the associations captured both by SNPs and CNVs occurs by chance. In summary, 87 out of 99 (87%) of genes with a significant CNV association are not associated with SNPs.

The large-scale (typically > 100 kb) copy number variation analyzed here appears to be associated with about 10 to 25% as many gene expression phenotypes as captured by ~700,000 SNPs, and the majority of these effects cannot be explained by altered dosage of the entire gene, but by gene disruption and its impact on the regulatory landscape of the region where these CNVs occur. When we restrict the analysis to within 1 Mb of the probe of the expressed gene, we detected 1061 genes associated with CGH clones or SNPs, 17.7% of which are associated with CGH clones, 83.6% with SNPs, and 1.3% with both. Of the 972 genes associated with CNV clones or SNPs, 8.75% are associated with CNV clones, 92.5% with SNPs, and 1.25% with both. Whereas the phase I HapMap SNPs likely capture a large fraction of the SNP effects in the genome (*13*), only a small minority of the CNVs in the genome were considered here: CNVs < 100 kb in length are far more numerous than CNVs >100 kb in length (*19*). As a consequence, 8.75 to 17.7% is a minimal estimate of the proportion of heritable gene expression variation that is explained by copy number variation.

Our study has attempted to evaluate the relative impact of CNVs and SNPs on phenotypic variation in human populations. Within the limitations of our samples, tissue type, SNP coverage, and CNV resolution, each type of genetic variation captures a substantial number of largely mutually exclusive effects on gene expression. We also demonstrate that both CNV and

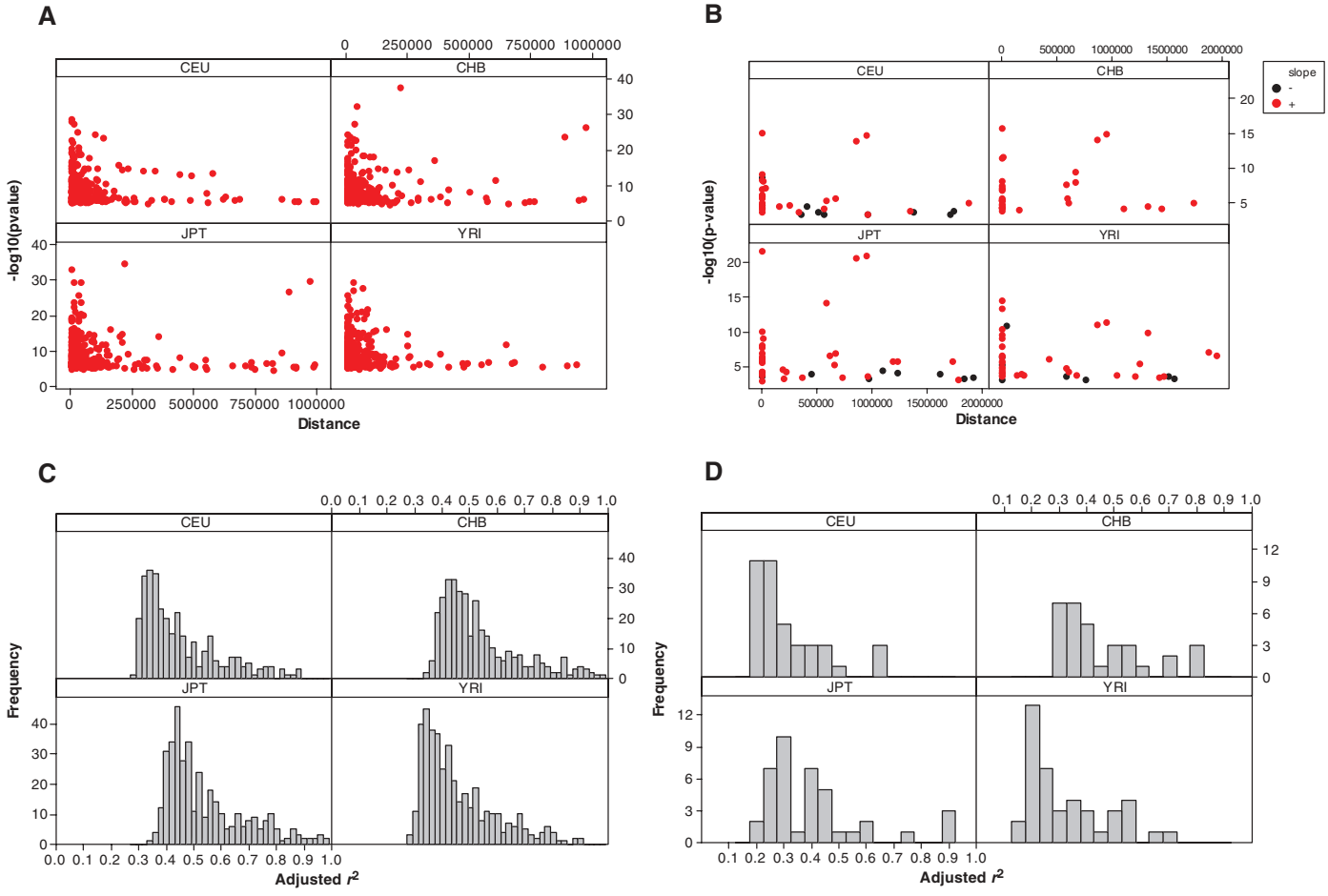

Fig. 1. Strength of association as a function of distance between (**A**) SNP and probe and (**B**) CNV and probe. Positive associations between mRNA levels and clone $\log_2$ ratios are shown in red, negative associations in black. Distance equal to zero corresponds to the probe residing within the CNV. In each population panel, only the details for the most significant association per significant gene are shown. Distribution of r^2 values for the most significant association per significant gene for (**C**) SNP-expression associations and (**D**) clone-expression associations.

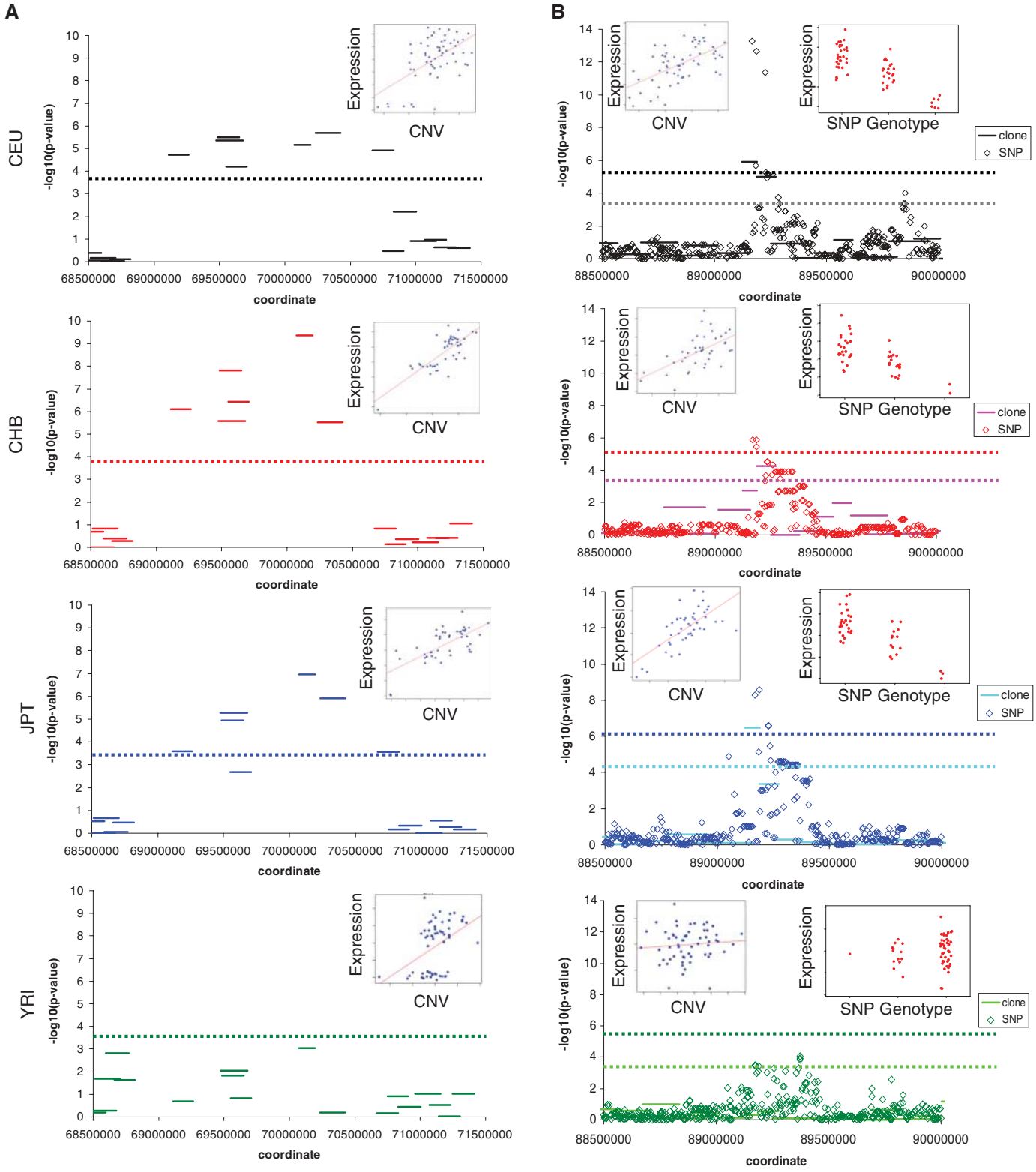

Fig. 2. Examples of SNP-expression and clone-expression associations in the four HapMap populations. (**A**) Clone-expression association for *SMN2*; chromosome 5 (chr 5). Significant associations between clones and expression are observed in CEU, CHB, and JPT, but not in YRI. (**B**) SNP-expression and clone-expression association for *GBP3*; chr 1. Both SNPs and clones are significantly associated with expression of *GBP3* in CEU, CHB, and JPT, but not in YRI. In each plot, dotted lines show the 0.001 permutation significance threshold. For clone-expression associations, all clones in the window are shown; however, the significance threshold was determined by permuting data only from those clones in CNVs where the CNV was present in at least two HapMap individuals. All coordinates shown are from Build 35 of the human genome. Inset panels show the relation between mRNA levels and SNP genotypes or clone $\log_2$ ratios, for the most significant clone or SNP in that population, which may differ across populations.

SNP associations are replicated across populations. Replication of association signals is the sine qua non of association studies, and the fact that we observe this even between diverse populations and with small sample sizes highlights the relevance and robustness of the associations we detect. Gene expression is the basis for many crucial functions in the cell, so the relative contribution of these two types of variants is an indication of the nature of the mutational and natural selection processes that contribute to phenotypic diversity and divergence. It is, therefore, essential that we interrogate both SNPs and CNVs (of all types) to perform a comprehensive exploration of genetic effects on phenotypic variation and disease. It is possible that, if a larger number of SNPs were analyzed or a higher resolution of CNVs was available, we would observe more overlap between the effects attributed to CNVs and SNPs. However, the difficulty of designing robust SNP genotyping assays in structurally dynamic regions of the genome (*26*) suggests that even with more comprehensive interrogation of SNPs and CNVs, the overlap may not be high enough for one type of variation to be sufficient for exploring the genetic causes of disease. We have also demonstrated that it is not necessary to perform such studies with CNV calls or CNV genotypes, but it is possible to use filtered CGH $\log_2$ ratios or any other type of high-quality quantitative signal that reflects underlying CNV. It has also become apparent that there are many more structural variants that contribute to phenotypic variation than our stringent criteria for what is a CNV reveal and that higher-resolution methods are necessary to elucidate their structure and function. Last, but not least, is the fact that we have only considered simple models of association in small samples, so it is very likely that if we apply more complex and realistic models (e.g., epistatic interactions) and/or larger population samples, a larger number of effects would be revealed. The results presented here reinforce the idea that the complexity of functionally relevant genetic variation ranges from single nucleotides to megabases, and the full range of the effects of all of these variants will be best captured and interpreted by complete knowledge of the sequence of many human genomes. Until this is possible we need to survey all known types of genetic variation to maximize our understanding of human evolution, diversity, and disease.

References and Notes

1. B. E. Stranger *et al.*, *PLoS Genet.* **1**, e78 (2005).
2. V. G. Cheung *et al.*, *Nature* **437**, 1365 (2005).
3. S. Doss, E. E. Schadt, T. A. Drake, A. J. Lusis, *Genome Res.* **15**, 681 (2005).
4. R. B. Brem, L. Kruglyak, *Proc. Natl. Acad. Sci. U.S.A.* **102**, 1572 (2005).
5. J. D. Storey, J. M. Akey, L. Kruglyak, *PLoS Biol.* **3**, e267 (2005).
6. M. F. Oleksiak, J. L. Roach, D. L. Crawford, *Nat. Genet.* **37**, 67 (2005).
7. S. A. Monks *et al.*, *Am. J. Hum. Genet.* **75**, 1094 (2004).
8. E. E. Schadt *et al.*, *Nature* **422**, 297 (2003).
9. E. J. Chesler *et al.*, *Nat. Genet.* **37**, 233 (2005).
10. L. Bystrykh *et al.*, *Nat. Genet.* **37**, 225 (2005).
11. E. T. Dermitzakis, B. E. Stranger, *Mamm. Genome* **17**, 503 (2006).
12. T. Pastinen, T. J. Hudson, *Science* **306**, 647 (2004).
13. International HapMap Consortium, *Nature* **437**, 1299 (2005).
14. L. Feuk, C. R. Marshall, R. F. Wintle, S. W. Scherer, *Hum. Mol. Genet.* **15** (suppl. 1), R57 (2006).
15. A. J. Iafrate *et al.*, *Nat. Genet.* **36**, 949 (2004).
16. J. Sebat *et al.*, *Science* **305**, 525 (2004).
17. E. Tuzun *et al.*, *Nat. Genet.* **37**, 727 (2005).
18. S. A. McCarroll *et al.*, *Nat. Genet.* **38**, 86 (2006).
19. D. F. Conrad, T. D. Andrews, N. P. Carter, M. E. Hurles, J. K. Pritchard, *Nat. Genet.* **38**, 75 (2006).
20. P. Stankiewicz, in *Genomic Disorders: The Genomic Basis of Disease*, J. R. Lupski, P. Stankiewicz, Eds. (Humana Press, Totowa, NJ, 2006), pp. 357–369.
21. D. A. Kleinjan, V. van Heyningen, *Am. J. Hum. Genet.* **76**, 8 (2005).
22. M. J. Somerville *et al.*, *N. Engl. J. Med.* **353**, 1694 (2005).
23. J. A. Lee *et al.*, *Ann. Neurol.* **59**, 398 (2006).
24. D. P. Locke *et al.*, *Am. J. Hum. Genet.* **79**, 275 (2006).
25. D. A. Hinds, A. P. Kloek, M. Jen, X. Chen, K. A. Frazer, *Nat. Genet.* **38**, 82 (2006).
26. R. Redon *et al.*, *Nature* **444**, 444 (2006).
27. GENEVAR—GENe Expression VARiation, www.sanger.ac.uk/genevar.
28. International HapMap Project, www.hapmap.org; release 16c.1.
29. The Copy Number Variation (CNV) Project Data Index, www.sanger.ac.uk/humgen/cnv/data.
30. R. W. Doerge, G. A. Churchill, *Genetics* **142**, 285 (1996).
31. G. Merla *et al.*, *Am. J. Hum. Genet.* **79**, 332 (2006).

32. We thank A. Clark and J. Pritchard for comments on earlier versions of the manuscript; M. Smith for assistance with software development; and M. Gibbs, J. Orwick, and C. Gerringer for technical support. Funding was provided by the Wellcome Trust to E.T.D., M.E.H., P.D., C.T.S., and N.C.; NIH to E.T.D. and S.T.; Cancer Research U.K. to S.T. and N.T.; the Leukemia and Lymphoma Society and the Brigham and Women's Hospital Department of Pathology to C.L.; and the U.K. Medical Research Council (MRC) to M.D. S.T. is a Royal Society Wolfson Research Merit Award holder. S.W.S. is supported by grants from Genome Canada/Ontario Genomics Institute and is a Scholar of the Canadian Institutes of Health Research and the Howard Hughes Medical Institute.

Supporting Online Material

www.sciencemag.org/cgi/content/full/315/5813/848/DC1
Materials and Methods
Figs. S1 to S3
Tables S1 to S5
References and Notes

24 October 2006; accepted 5 January 2007
10.1126/science.1136678

Evidence That Focal Adhesion Complexes Power Bacterial Gliding Motility

Tâm Mignot,[1]* Joshua W. Shaevitz,[2] Patricia L. Hartzell,[3] David R. Zusman[1]*

The bacterium *Myxococcus xanthus* has two motility systems: S motility, which is powered by type IV pilus retraction, and A motility, which is powered by unknown mechanism(s). We found that A motility involved transient adhesion complexes that remained at fixed positions relative to the substratum as cells moved forward. Complexes assembled at leading cell poles and dispersed at the rear of the cells. When cells reversed direction, the A-motility clusters relocalized to the new leading poles together with S-motility proteins. The Frz chemosensory system coordinated the two motility systems. The dynamics of protein cluster localization suggest that intracellular motors and force transmission by dynamic focal adhesions can power bacterial motility.

During the exhibition of gliding motility, bacteria move across solid surfaces without the use of flagella (*1*). Gliding motility is important for biofilm formation and bacterial virulence. Motility in *Myxococcus xanthus*, a Gram-negative rod-shaped bacterium, relies on two separate but coordinated motility engines. S motility is powered by type IV pili that are assembled at the leading cell pole; movement is produced as the pili bind to surface exopolysaccharides and are retracted, thereby pulling the cell forward (*2*). A motility, on the other hand, is not associated with pili or other obvious structures and is not well understood.

To investigate the A-motility system, we studied AglZ, a protein that is essential for A motility but dispensable for S motility (fig. S1, A and B) (*3*). AglZ is similar to FrzS, an S-motility protein that oscillates from one cell pole to the other when cells reverse direction (*4*) (fig. S1A). To track the localization of AglZ in moving cells, we constructed an *M. xanthus* strain containing a chimeric *aglZ-yfp* gene in place of the endogenous *aglZ* gene (fig. S2A). This chimeric gene encodes an AglZ–yellow fluorescent protein (YFP) fusion protein that was stable and functional (fig. S2, B and C). We followed AglZ-YFP localization using time-lapse video microscopy: In fully motile cells, AglZ-YFP was localized in

[1]Department of Molecular and Cell Biology, University of California, Berkeley, CA 94720, USA. [2]Department of Integrative Biology, University of California, Berkeley, CA 94720, USA. [3]Department of Microbiology, Molecular Biology, and Biochemistry, University of Idaho, Moscow, ID 83844, USA.

*To whom correspondence should be addressed. E-mail: tmignot@berkeley.edu (T.M.); zusman@berkeley.edu (D.R.Z.)