For reprint orders, please contact:
reprints@futuremedicine.com

Ancestry-related differences in gene expression: findings may enhance understanding of health disparities between populations

**Wei Zhang¹ &
M Eileen Dolan^{1,2,3†}**

[†]Author for correspondence

¹University of Chicago,
Section of
Hematology/Oncology,
Department of Medicine,
5841 S Maryland Ave.
MC 2115, Chicago,
IL 60637, USA

Tel.: +1 773 702 4441;
Fax: +1 773 702 0963;
E-mail: edolan@
medicine.bsd.uchicago.edu

²University of Chicago,
Committee on Clinical
Pharmacology and
Pharmacogenomics, Chicago,
IL 60637, USA

³University of Chicago,
Cancer Research Center,
Chicago, IL 60637, USA

'Analysis of variation in gene expression between human populations will enhance our understanding of the contribution of genetics to health disparities.'

The unveiling of the first human genome draft sequences almost 7 years ago [1,2] marked the beginning of the postgenomic era. However, because of the existence of huge genetic variations, such as those in the form of SNPs [3] and copy-number variations (CNVs) [4], there is still much to be discovered. In recent years, functional studies have revealed that natural genetic variation plays a role in complex human traits, such as risks of common diseases [5], for example, asthma, stroke, heart attack, diabetes and cancer, and variability in drug response [6–8]. The importance of genetic variation in the human genome has increasingly attracted researchers' efforts in the past few years, leading to its selection as the Breakthrough of the Year in 2007 by *Science* [9].

To explore the complexity of the human genome, parallel efforts to the Human Genome Project have included the International HapMap Project [10,11]. This project has been invaluable in understanding the complexity of human genetic variation not only among individuals of the same ancestry [12], but also between different human populations [13–16]. The goal of the HapMap Project is to develop a haplotype map of the human genome and to describe the common patterns of DNA sequence variation using 270 lymphoblastoid cell lines (LCLs) derived from individuals of African (YRI: 30 parents–child trios from the Yoruba people in Ibadan, Nigeria), Asian (CHB: 45 unrelated Han Chinese from Beijing, China; JPT: 45 unrelated Japanese from Tokyo, Japan) and European ancestry (CEU: 30 parents–child trios from the Centre d'Etude du Polymorphisme Humain samples collected from UT, USA). The newly released Phase II component of the HapMap genotypic data

contains over 3.1 million human SNPs [3] that reveal common haplotype patterns and differences in genetic variation among these four geographically distant populations. As HapMap LCL samples are commercially available, researchers can obtain relevant phenotypes (e.g., drug response and gene expression) on those cell lines and perform population-based whole-genome associations with the publicly available genotypic data to identify responsible genetic determinants or SNP markers linked to the causal elements [17].

Among various phenotypes, gene expression acts as an intermediate phenotype situated between variation in DNA sequence and other more complex cellular, tissue, organ or whole-body phenotypes. Quantitative variation in gene-expression level (e.g., mRNA transcript abundance) as a complex trait is heritable and has been mapped to the human genome as expression quantitative trait loci (eQTLs), which represent genomic regions for the genetic control of gene expression [18–20].

'Quantitative variation in gene-expression level as a complex trait is heritable and has been mapped to the human genome as expression quantitative trait loci...'

Significant gene-expression variation has previously been observed within the same human population [18–20]. More recently, using the LCL samples from the four HapMap populations, the ancestry-related or population differences in gene expression has been described by investigators utilizing various whole-genome expression microarrays and statistical approaches [17]. Comparing the unrelated 60 CEU samples and 82 Asian samples (41 CHB and 41 JPT), Spielman *et al.* showed that common genetic variants account for differences in gene expression among ethnic groups [14]. They found that approximately 25% of genes analyzed (out of approximately 4200 deemed expressed in LCLs) using the Affymetrix (CA, USA) Human Focus array were differentially expressed between the European-derived and Asian-derived populations,

while there were few differences between the two Asian populations [14]. Using the same microarray platform, Storey *et al.* found that approximately 17% of genes (out of ~5200 deemed expressed in LCLs) are differentially expressed between eight CEU samples and eight YRI samples [15]. They also found that genes differentially expressed between the CEU and YRI samples were strongly enriched in inflammatory pathways even after a strict Bonferroni correction for multiple-hypothesis tests. Included in this set of genes were several cytokines and chemokine receptors that have been implicated in numerous cardiovascular, infectious and immune-related diseases [15]. Using the Affymetrix Human Exon array [21] (containing ~1.4 million exon-level known and predicted probesets), we recently performed a comprehensive comparison of the expression of approximately 9200 transcript clusters (gene-level), which were deemed reliably expressed out of approximately 18,000 transcript clusters with RefSeq-supported [22] annotation, between 87 CEU samples and 89 YRI samples [13]. Approximately 4.5% (383) of the tested genes were found to be differentially expressed between the CEU and YRI populations. While no differential genes appeared to be over-represented on certain chromosomes, biological processes including ribosomal biogenesis, tRNA processing and antimicrobial humoral response were found to be enriched in these differential genes, suggesting their possible roles in contributing to the population differences at a higher level than that of mRNA expression and in response to environmental information [13]. Impressively, our finding of the enrichment of immune response-related genes (e.g., *CCR7* and *CXCR3*) was in agreement with the Storey *et al.* smaller-scale analysis [15]. Our findings also suggest that the frequencies of common genetic variants contribute to a substantial fraction of gene-expression variation between human populations. Previous studies have focused on *cis*-acting elements [14,15], but our results suggest that distant or *trans*-acting elements can also contribute to the population differences in gene expression. Using the Illumina (CA, USA) whole-genome expression array, Stranger *et al.* tested the population differences in gene expression among all the available HapMap populations (CEU, YRI and CHB/JPT), with the major aim to dissect genetic regulators of gene expression [16]. They estimated that the fraction of genes with significant gene-expression

variation between any two populations is between 17 and 29%. Their results also support an abundance of *cis*-regulatory variation in the human genome with the existence of limited *trans* effects [16].

'...gene expression differs among human populations and common genetic variants account for this variation in *cis* and in *trans* mode.'

Collectively, these studies demonstrate that gene expression differs among human populations and common genetic variants account for this variation in *cis* and in *trans* mode [13–16]; however, some differences exist in these findings. The most apparent discrepancy is the proportion of differential genes between two populations, ranging from 4.5 to 29% [13–16]. This is primarily due to differences in sample size, statistical approaches, level of cutoff for significance, data preprocessing/normalization procedures and microarray platforms utilized for the studies. The two most comprehensive studies [13,16] used different statistical approaches to accommodate the trio structures of the HapMap CEU and YRI samples. Our analysis incorporated both a permutation-based approach and a general linear model with Toeplitz structure for modeling parents–child covariance, while Stranger *et al.* defined a cutoff based on the assumption that parents and CEU or YRI children represent the null distribution of differences in median expression values [13,16]. The other two studies [14,15] used unrelated samples and therefore did not have to consider relatedness. There were also other technical differences in the analysis. First, we made an attempt to address the potential confounding factor of SNPs in probes [23], while other studies did not take this potential problem into account in their analyses. Second, although, generally speaking, the microarray data are consistent with other experimental quantitative techniques and across microarray platforms [24,25], the Affymetrix GeneChips® and Illumina BeadArrays used in these studies have differences in probe-set design, gene coverage and probe-set replication. Furthermore, some limitations or confounding factors (both measurable or unmeasurable) could influence the accuracy of these results [26]. A re-analysis by Akey *et al.* of the data from one expression study [14] suggested that a large proportion of the differential genes between CEU and CHB/JPT samples could be due to

systematic and uncorrectable bias [27]. This confounding factor is probably due to the age (in culture) of the CEU samples [28] and time at which analysis was performed (batch effect) relative to the other recently established CHB/JPT and YRI samples. To address the issue of age in culture, we tried a reduced quantitative transmission disequilibrium test [29] model to evaluate the contribution of genetic and nongenetic factors for the differential genes associated with SNPs [13]. The majority of our differential genes were not explained by the population identity alone, suggesting the substantial contribution of genetic factors.

‘...ancestry is a significant determinant of both susceptibility to some diseases and response to therapeutic treatments.’

Gene expression can affect susceptibility to diseases and drug response. A widely appreciated example is the multidrug resistance exhibited by tumor cells through overexpression of *MDR* genes [30]. Although disease risk and drug response are likely to be contributed by the interplay between both genetic and nongenetic factors, gene-expression variation plays a significant role, probably through upregulating or downregulating genes within physiological pathways. For example, Caucasians and African-Americans differ significantly in *EGFR* expression in prostate cancer [31], suggesting potential differences in response to *EGFR* inhibitors between these two populations. Therefore, analysis of variation in gene expression between human populations will enhance our understanding of the contribution of genetics to health disparities.

Clinically, studies have shown that ancestry is a significant determinant of both susceptibility to some diseases and response to therapeutic treatments. For example, African-American, Hispanic, Asian and Native American women have a lower incidence of breast cancer but higher mortality compared with non-Hispanic white women [32]. Differences in response to anticancer agents docetaxel and carboplatin have been observed between Asian and Caucasian advanced non-small-cell lung cancer patients [33]. Findings of population differences in expression are critical for us to investigate these health disparities between human populations. Interestingly, the enrichment of immune response-related genes in both ours [13] and a previous smaller study [15] suggests that individuals of

African and European ancestry may have different susceptibility to infectious diseases. In fact, it has been observed that African-American adults may be more susceptible to infection by certain microbes such as *Porphyrromonas gingivalis*, which causes periodontitis [34].

Obviously, the recent progress demonstrated the usefulness of the HapMap resource, which is comprised of comprehensive genotypic and gene-expression data, in identifying genetic regulators for gene-expression variation among populations. However, as the current HapMap samples include only four populations (three if combining the two Asian populations), we are still far from understanding the differences in gene expression involving other major human populations (African-Americans, Mexicans and Pacific Islanders, to name a few). To extend these kinds of studies to other populations would present more challenging problems; for example, the African-American population is believed to be an admixture of African, European and Native American descent, thus a relatively clear cut of ancestry may not be as evident as the current HapMap populations. Another limitation of the current studies is that only mRNA-level expression has been compared between populations; a natural question is does this equate to gene- or pathway-level differences in protein expression? Progress in areas such as proteomics and metabolomics will facilitate future studies of population differences in gene expression. Therefore, current success in evaluating global differences in expression is just the beginning of our understanding of how differences in genetic variation and expression variation explain differences in disease susceptibility or drug response.

Financial & competing interests disclosure

Some of the research described in this editorial was funded through the Pharmacogenetics of Anticancer Agents Research (PAAR) Group (www.pharmacogenetics.org) by the NIH/NIGMS grant U01GM61393. Affymetrix provided chips, reagents and other technical support for our exon array data described in this editorial. ME Dolan is on the Scientific Advisory Board to the NIGMS Human Genetic Cell Repository at the Coriell Institute for Medical Research, NJ, USA. The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

No writing assistance was utilized in the production of this manuscript.

Bibliography

1. Lander ES, Linton LM, Birren B *et al.*: Initial sequencing and analysis of the human genome. *Nature* 409(6822), 860–921 (2001).
2. Venter JC, Adams MD, Myers EW *et al.*: The sequence of the human genome. *Science* 291(5507), 1304–1351 (2001).
3. Frazer KA, Ballinger DG, Cox DR *et al.*: A second generation human haplotype map of over 3.1 million SNPs. *Nature* 449(7164), 851–861 (2007).
4. Redon R, Ishikawa S, Fitch KR *et al.*: Global variation in copy number in the human genome. *Nature* 444(7118), 444–454 (2006).
5. Kryukov GV, Pennacchio LA, Sunyaev SR: Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. *Am. J. Hum. Genet.* 80(4), 727–739 (2007).
6. Hartford CM, Dolan ME: Identifying genetic variants that contribute to chemotherapy-induced cytotoxicity. *Pharmacogenomics* 8(9), 1159–1168 (2007).
7. Huang RS, Duan S, Bleibel WK *et al.*: A genome-wide approach to identify genetic variants that contribute to etoposide-induced cytotoxicity. *Proc. Natl Acad. Sci. USA* 104(23), 9758–9763 (2007).
8. Huang RS, Duan S, Shukla SJ *et al.*: Identification of genetic variants contributing to cisplatin-induced cytotoxicity by use of a genomewide approach. *Am. J. Hum. Genet.* 81(3), 427–437 (2007).
9. Pennisi E: Breakthrough of the year. Human genetic variation. *Science* 318(5858), 1842–1843 (2007).
10. International HapMap Consortium: The International HapMap Project. *Nature* 426(6968), 789–796 (2003).
11. International HapMap Consortium: A haplotype map of the human genome. *Nature* 437(7063), 1299–1320 (2005).
12. Cheung VG, Spielman RS, Ewens KG *et al.*: Mapping determinants of human gene expression by regional and genome-wide association. *Nature* 437(7063), 1365–1369 (2005).
13. Zhang W, Duan S, Kistner EO *et al.*: Evaluation of genetic variation contributing to differences in gene expression between populations. *Am. J. Hum. Genet.* 82(3), 631–640 (2008).
14. Spielman RS, Bastone LA, Burdick JT *et al.*: Common genetic variants account for differences in gene expression among ethnic groups. *Nat. Genet.* 39(2), 226–231 (2007).
15. Storey JD, Madeoy J, Strout JL *et al.*: Gene-expression variation within and among human populations. *Am. J. Hum. Genet.* 80(3), 502–509 (2007).
16. Stranger BE, Nica AC, Forrest MS *et al.*: Population genomics of human gene expression. *Nat. Genet.* 39(10), 1217–1224 (2007).
17. Zhang W, Ratain MJ, Dolan ME: The HapMap resource is providing new insights into ourselves and its application to pharmacogenomics. *Bioinform. Biol. Insights* 2, 15–23 (2008).
18. Cheung VG, Conlin LK, Weber TM *et al.*: Natural variation in human gene expression assessed in lymphoblastoid cells. *Nat. Genet.* 33(3), 422–425 (2003).
19. Stranger BE, Forrest MS, Clark AG *et al.*: Genome-wide associations of gene expression variation in humans. *PLoS Genet.* 1(6), E78 (2005).
20. Morley M, Molony CM, Weber TM *et al.*: Genetic analysis of genome-wide variation in human gene expression. *Nature* 430(7001), 743–747 (2004).
21. Affymetrix: Exon probeseq annotations and transcript cluster groupings. *Affymetrix GeneChip Exon Array White Paper Collection* (2005).
22. Pruitt KD, Tatusova T, Maglott DR: NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* 35(Database Issue), D61–D65 (2007).
23. Alberts R, Terpsstra P, Li Y *et al.*: Sequence polymorphisms cause many false *cis*eQTLs. *PLoS ONE* 2(7), E622 (2007).
24. Canales RD, Luo Y, Willey JC *et al.*: Evaluation of DNA microarray results with quantitative gene expression platforms. *Nat. Biotechnol.* 24(9), 1115–1122 (2006).
25. Barnes M, Freudenberg J, Thompson S, Aronow B, Pavlidis P: Experimental comparison and cross-validation of the Affymetrix and Illumina gene expression analysis platforms. *Nucleic Acids Res.* 33(18), 5914–5923 (2005).
26. Zhang W, Dolan ME: On the challenges of the HapMap resource. *Bioinformatics* 2(6), 238–239 (2008).
27. Akey JM, Biswas S, Leek JT, Storey JD: On the design and analysis of gene expression studies in human populations. *Nat. Genet.* 39(7), 807–808; author reply 808–809 (2007).
28. Dausset J, Cann H, Cohen D *et al.*: Centre d'étude du polymorphisme humain (CEPH): collaborative genetic mapping of the human genome. *Genomics* 6(3), 575–577 (1990).
29. Abecasis GR, Cardon LR, Cookson WO: A general test of association for quantitative traits in nuclear families. *Am. J. Hum. Genet.* 66(1), 279–292 (2000).
30. Perez-Tomas R: Multidrug resistance: retrospect and prospects in anti-cancer drug treatment. *Curr. Med. Chem.* 13(16), 1859–1876 (2006).
31. Shuch B, Mikhail M, Satagopan J *et al.*: Racial disparity of epidermal growth factor receptor expression in prostate cancer. *J. Clin. Oncol.* 22(23), 4725–4729 (2004).
32. Fejerman L, Ziv E: Population differences in breast cancer severity. *Pharmacogenomics* 9(3), 323–333 (2008).
33. Millward MJ, Boyer MJ, Lehnert M *et al.*: Docetaxel and carboplatin is an active regimen in advanced non-small-cell lung cancer: a Phase II study in Caucasian and Asian patients. *Ann. Oncol.* 14(3), 449–454 (2003).
34. Schenkein HA, Burmeister JA, Koertge TE *et al.*: The influence of race and gender on periodontal microflora. *J. Periodontol.* 64(4), 292–296 (1993).