Published in final edited form as:

Curr Opin Syst Biol. 2017 February ; 1: 102–108. doi:10.1016/j.coisb.2016.12.018.

Global variation in gene expression and the value of diverse sampling

Derek E. Kelly^{a,b}, Matthew E. B. Hansen^a, and Sarah A. Tishkoff^{a,b,c}

^aDepartment of Genetics, University of Pennsylvania, Philadelphia, PA 19104, USA

^bGenomics and Computational Biology Graduate Group, University of Pennsylvania, Philadelphia, PA 19104, USA

^cDepartment of Biology, University of Pennsylvania, Philadelphia, PA 19104, USA

Abstract

The genomics era has accelerated our understanding of how genetic and epigenetic factors influence both normal variable traits and disease risk in humans. However, the majority of “omics” studies have focused on individuals living in urban centers, primarily from Europe and Asia, neglecting much of the genetic and environmental variation that exists across worldwide populations. Comparative studies of gene regulation in ethnically diverse populations are informing our understanding of how evolutionary forces have shaped the genetic and molecular mechanisms underlying complex traits, and studying gene expression in different environmental contexts is enabling the dissection of disease-related pathways such as immune response. Such approaches are vital to the equitable application of genomics and medicine.

Introduction

Molecular and sub-cellular traits like gene expression, epigenetic modifications, and chromatin state form the foundation upon which tissue and organism level phenotypes are built [1]. Examination of the co-variation between genetic variation and these molecular traits, as is done in genome-wide association tests, can facilitate a mechanistic understanding of how complex traits arise, including disease risks [2–4]. A systems biology approach, incorporating information on genetic interactions and transcription networks, is useful to explore how naturally occurring genetic variants perturb these networks [5]. Equally important is understanding how molecular networks are influenced by environmental factors including diet, lifestyle, and infectious disease status [5]. By extending these approaches to diverse populations, we will learn more about the biological processes that account for population differences in disease susceptibility and some of the remarkable examples of local human adaptation. Moreover, failure to include diverse groups in studies of molecular traits will bias our understanding of gene expression mechanisms and can exacerbate

Corresponding Author: Sarah A. Tishkoff, tishkoff@mail.med.upenn.edu.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

existing health care disparities [6,7]. In this review we discuss recent studies of inter-population variation in molecular traits and conclude by discussing several promising research areas and current challenges.

Recent insights from global sampling

Impact of demography and natural selection on human diversity

The patterns of genotypic and phenotypic diversity in humans are shaped by demographic history as well as adaptation to diverse environments. Anatomically modern humans arose in Africa nearly 200 thousand years ago (kya), and within the last 100 ky migrated out of Africa and across the globe [8]. This global migration event is characterized by a series of population bottlenecks, or “founder” events, reducing levels of genetic diversity as populations migrated from West to East across Eurasia and into the Pacific Islands and the Americas [9,10]. Africans, which have maintained large population sizes relative to non-Africans, have the highest levels of genetic diversity on a global scale, as well as high levels of population substructure [11]. As humans spread across the globe, they encountered new climates, geography, food sources, and pathogens, which presented novel selective pressures [8]. In addition, early humans migrating out of Africa encountered archaic populations including Neanderthals and Denisovans, resulting in low levels of admixture and introgression of archaic genomes into modern humans genomes (representing ~1 – 6% of non-African genomes) [12]. Plant and animal domestication led to substantial shifts in diet, as populations moved from a hunting and gathering lifestyle to an agricultural lifestyle rich in grains and carbohydrates or towards a pastoralist lifestyle rich in milk and protein within the past 10 ky.

One of the most striking examples of local human adaptation is the persistence of lactase gene expression into adulthood, a trait common in populations that practice dairying. The enzyme lactase metabolizes the sugar lactose common in milk into glucose and galactase. This trait has arisen independently multiple times in human history as populations domesticated cattle and began incorporating milk into their diet [13–15]. This example highlights several key features of human trait variation. For one, despite being a relatively simple trait regulated by the expression of a single enzyme, European and African populations with the lactase persistence trait differ in their underlying causal variants. Moreover, these variants reside in an intronic region of the neighboring gene, *MCM6*, acting over a considerable genomic distance. Evidence suggests that the causal genetic variants primarily act by altering epigenetic mechanisms during development and aging [3]. Finally, the lactase enzyme is specifically expressed in enterocytes of the small intestine, and the trait is only apparent when individuals actually consume lactose, highlighting that population differences can be cell-type specific and condition dependent.

Population differences in patterns of gene expression

The most comprehensive global studies of cellular expression to date come from studies of lymphoblastoid cell lines (LCLs) derived from B cells taken from whole blood samples. LCLs have been derived for many of the worldwide genome reference panels collected over

the last few decades, including the Human Genome Diversity Panel [16], HapMap [17], and gEUVADIS and the 1000 Genomes project [18].

Comparing European and African populations from the gEUVADIS project, Lappalainen *et al.* [18] found that more distantly related populations tend to have a greater number of differentially expressed genes, and that they tend to differ because of differences in expression of different gene transcripts rather than differences in expression levels of the same transcripts. Between British (GBR) and Tuscan (TSI) populations, as little as 8% of transcriptome variation was due to transcript usage, while transcript usage accounted for 85% of the transcriptome variation between a European descent population from Utah (CEU) and the West African Yoruba (YRI) population. Using a more globally diverse sample, Martin *et al.* [16] did not observe a significant correlation between population divergence and differentially expressed transcripts in LCLs. Studies of LCL gene expression also conflict on the proportion of overall expression variance that can be attributed to ancestry, ranging from 2–3% [18,19] to 25% [16]. The discrepancies may be attributable to the diversity of genetic backgrounds sampled (see Table 1) or lower power in Martin *et al.* due to smaller sample sizes (45 individuals in [16] vs. 462 in [18]). Differences in sample storage can also have effects on LCL expression as some LCLs are decades old and have been through numerous cycles of thawing and refreezing, which are known to have effects on gene expression levels [20,21]. It should also be noted that studies of differences in gene expression across populations must carefully control for batch effects [22] and comparison of results between studies can be difficult due to differences in populations studied, data types used (array vs. sequencing), and experimental protocols.

While LCLs have provided a baseline for how much gene expression variation among populations is due to different genetic or epigenetic backgrounds, these cells are highly irregular cells and do not capture the cellular heterogeneity nor the normal expression profile of native tissue. To understand inter-population gene expression variation in a native tissue, Hughes *et al.* [23] studied placental samples from four American ethnic groups (see Table 1), estimating that 7.8% of variation in gene expression was between-groups. Genes with significant between-group expression variance were found to be enriched for functions related to immune response, cell signaling, and metabolism, which are all phenotypes that can have adaptive significance [24,25].

The genetic basis of gene expression differences between populations

Over the past decade, genome-wide association tests have been used to discover genetic variants that influence gene expression, referred to as expression quantitative trait loci (eQTL) [26]. As with all association tests that do not include every genetic variant, significantly associated variants are not necessarily functional themselves but instead may “tag” a functional variant that resides on the same haplotype. Because the causal variants underlying eQTL associations are in most cases not known and populations often differ in haplotype frequencies and patterns of linkage disequilibrium (LD), eQTL findings may not replicate across populations. In addition, many eQTL studies rely on imputed genotypes, often using the haplotypes from a reference population to make the genotype inference. However, if the study population differs from the reference population in regards to having

novel genetic variants or different patterns of LD, the imputation may suffer imputation inaccuracies. Consequently, cross-population comparison of eQTLs at imputed variants may suffer biases in the association accuracy if the reference populations underlying the imputation are not genetically similar to the populations being compared.

Few inter-population gene expression studies to date have had sufficient sample sizes to map eQTLs in each population separately. The best powered study in LCLs found at least one eQTL in at least one population for 60% of genes analyzed (FDR 5%) [18]. Cross-population comparisons have shown that the specific eQTL variants are often not genome-wide significant in multiple populations; Stranger *et al.* found that 77% of eQTLs are population specific, 23% are shared between two or more populations, and no eQTLs are shared across all populations. However, for genes with a significant association in at least two populations, the vast majority (>98%) of eQTLs do show consistent direction of effect as well as strongly correlated effect sizes between population pairs [17], suggesting that eQTLs tag shared gene regulatory variation and that the failure to reach significance in multiple populations is likely due to allele frequency differences or other factors that can alter the statistical power of an association test.

Epigenetic modifications can differ between populations

As with gene expression, LCLs have been a popular choice for studying the genetic control of epigenetic mechanisms at the population scale [27], but only a handful of studies have attempted to systematically assess epigenetic variation across populations. LCLs have been used to assess population variation in DNA methylation [19,28,29] and histone modifications [30] across ethnically diverse populations. Carja *et al.* [19] estimated that ancestry explained 6% of the variance for methylation, while Kasowski *et al.* [30] found that ancestry explained less than 20% of the variance in different histone modifications and DNA-binding protein locations (specifically CTCF and cohesin) as measured by ChIP-seq.

Fraser *et al.* [28] mapped QTLs for DNA methylation (“mQTLs”), finding that, similar to eQTLs, mQTLs are not always shared between populations, with only 11 out of 124 (8.9%) of mQTLs being found in both the YRI and CEU. The authors also compared their identified mQTLs with those previously reported for a subset of the same YRI cell lines, finding that only 34.6% of their high-confidence mQTLs replicated. This seemingly low percentage may be attributable to differences in statistical procedures, what variants were considered, or the aforementioned limitations of LCLs.

Teasing apart the respective genetic and environmental contributions to cross-population differences in epigenetic modifications discussed above is often difficult in practice because genetic ancestry and environmental conditions are often confounded variables (i.e. highly correlated). One approach is to study populations with the same genetic ancestry but that reside in different environments, which controls for the genetic background and allows for environmental influences to be isolated. A recent example of this is Idaghdour *et al.*, who studied gene expression differences in leukocytes of Moroccan individuals of Arab and Amazigh descent occupying urban and rural environments [31]. The complementary approach is to examine populations with different genetic ancestries that live in the same location, which controls for shared environmental factors and allows for genetic associations

to be identified without environmental confounding. Samples of DNA and tissue from the right balance of populations are key for either approach. Given the relative ease of blood sampling from which DNA, RNA, and epigenetic profiles can be extracted, expanded blood sampling from global population would be an important research resource. Such a global “omics” bio-bank would be useful not only for distinguishing environmental and genetic effects in an important tissue type that impacts human disease and immunity, but would also be an ideal data source for establishing analysis protocols and provide a reference point for other cell and tissue types.

A recent cross-population study in Africa examined variation in DNA methylation from three populations using both the above strategies. Fagny *et al.* [29] identified differentially methylated sites between short statured central-African rainforest hunter-gatherers (RHGs, commonly known as Pygmies) and two groups of neighboring Bantu agriculturalists. One of the Bantu groups lives an agrarian lifestyle while the other lives in the rain-forest with a similar lifestyle as the RHGs. By comparing the differentially methylated sites between these three populations, Fagny *et al.* found that the methylation differences between the rain-forest dwelling Bantu and RHGs, reflecting deep ancestral differences, are enriched near developmental genes. In contrast, methylation differences between the rain-forest dwelling Bantu and the agrarian Bantu, which captures differences due to recent shifts in lifestyle and environment, were enriched near genes involved in immunity and host/pathogen interactions. Studies like this exemplify the need for careful cross-population study design to tease apart environmental from genetic influences on epigenetic modifications.

Differential environmental exposures reveal context dependent expression

Differences in steady-state gene expression levels can influence phenotypic variation, but some key biological pathways, such as the immune response to pathogens, are only active when the triggering environmental factor is present [32] and only within specific cell types [33]. A handful of studies have investigated population differences in immune response within specific cell populations [32]. One notable finding is that individuals with recent African ancestry show higher average expression of immune response genes during T cell activation than individuals of European or Asian ancestry [34].

Two recent studies have shed light on how evolutionary forces have shaped differences in immune response between individuals of African and European descent. Quach *et al.* [35] and Nédélec *et al.* [36] measured gene expression levels in CD14+ monocytes [35] or macrophages derived from CD14+ cells [36], before and after immune challenges with pathogens or stimulating ligands. These studies found extensive differences in the immune response due to ancestry, both in terms of gene expression levels before and after stimulation and the magnitude of response. Similar to the results of Ye *et al.* [34], African ancestry was associated with a strong inflammatory response and functional tests also showed an increased ability to clear bacterial infection [36]. Both studies also provide evidence of Neanderthal ancestry contributing to immunity differences, particularly with regards to viral infections [35]. Performing similar challenge studies in groups with different histories of pathogen exposure may reveal what genes and pathways are activated in response to specific

pathogens, as well as key genes that have been tuned by natural selection to improve resistance.

Challenges and Future Directions

Recent comparative studies in global populations have revealed how gene expression and regulation can differ among human groups in a handful of tissue types. From here, key research directions include creating a reference set of high quality whole genome sequences that includes greater representation of non-European genetic ancestries, incorporating high-throughput techniques for measuring protein abundance and alternatively spliced gene transcripts, studying more diverse cell types across many ethnic groups, and inclusion of other intermediate traits, particularly the human metabolome.

Beyond gene expression: measuring protein abundance and alternatively expressed isoforms

Over 90% of protein coding genes in humans are capable of encoding multiple transcripts and thus multiple gene isoforms by combining different exons through a process known as “alternative splicing” [37]. This process can have phenotypic effects on par with changes in gene expression levels [37,38]. While gene expression studies using RNA sequencing generally infer patterns of alternative splicing, short-read sequencing is limited in its ability to quantify expression at the transcript level [39]. The increasing availability of longer reads [39] and improved bioinformatics tools [40] will lead to a greater understanding and appreciation of the role of alternative splicing in evolution and phenotypic outcomes. A globally diverse set of transcriptomes will also be useful for mapping population specific alternative splicing patterns.

Messenger RNA (mRNA) levels have been a popular proxy for protein abundance because the technologies for quantifying RNA such as microarrays and short-read sequencing are well established [41,42]. However, it is well understood that steady-state RNA levels do not perfectly correlate with protein abundance [43,44]. Ribosome profiling quantifies RNA transcripts undergoing translation, giving a more functionally relevant measure of gene expression [45]. Newer technologies for high-throughput measurements of protein abundance [43,44] will also provide improved measures of cellular state beyond mRNA levels.

Global cell-type sampling beyond blood

Due to practical and ethical issues, studying within- or between-population variation in gene expression has been limited to accessible tissues like blood, placenta, adipose, or fibroblast cells [46]. To provide a more comprehensive view of gene regulation throughout the human body, the Genotype-Tissue Expression (GTEx) project [47] is mapping eQTLs across 54 post-mortem human tissues in hundreds of individuals. In its most recent phase, the GTEx consortium has identified an eQTL for approximately 80% of all tested genes (82.6% of expressed genes, 78.3% of all lincRNA or protein coding genes) and approximately 90% of autosomal protein coding genes (90.2% of expressed protein coding genes and 88% of all annotated protein coding genes) [48]. Fine scale, tissue specific eQTL data can be used to

refine GWAS hits [49] and suggest causal tissues underlying complex diseases [50]. While this will provide an extremely useful catalog of genetic variants and their impact on gene expression, the majority of individuals in the GTEx cohort are of European ancestry, potentially limiting the global applicability of their findings.

A promising alternative approach for measuring cell-type specific expression is the use of inducible pluripotent stem cell (iPSC) lines. These cell lines can be generated from blood and differentiated into numerous cell types [51], and importantly, inter-individual differences in cellular phenotypes are recapitulated during cell reprogramming, potentially making them suitable for the mapping of eQTLs and other molecular phenotypes in the differentiated cell-types [52,53]. We note, however, that this is an area of active research and the limits to this approach are currently not well known. Combining this approach with technologies for introducing population or region specific genetic variants *in vitro*, such as CRISPR/Cas9 technology, could be a promising route [3,4] for examining cross-population differences in gene expression regulation in diverse tissues that obviates many of the difficulties in tissue collection across ethnically diverse populations.

Global sampling of the human metabolome

In addition to gene expression differences, measuring differences in intermediate or molecular phenotypes, such as metabolite levels, is important for understanding population differences underlying complex traits. Molecular intermediates can be markers for specific biological processes central to phenotypic expression and disease etiology, and can also be markers of environmental exposures [54,55]. Profiling metabolites can help to map the human “metabolome,” and like gene expression, measuring how metabolites co-vary with genetic variation and environmental variation can give important insights into the genetic and environmental basis of complex traits.

Conclusion

The recent explosion in functional genomic tools has enabled scientists to profile genetic, epigenetic, and gene expression variation genome-wide, giving an increasingly clear picture of how a genomic blueprint (imperfectly) determines molecular processes and, ultimately, phenotypes. As these technologies mature, they will be increasingly applied to populations far from the science hubs of the Western world, allowing for a more complete understanding of human genetic and phenotypic variation, informing a better understanding of evolution, molecular biology, and medicine.

Acknowledgments

This work is funded by the National Institutes of Health [grants 1R01DK104339-01 and 1R01GM113657-01] and the National Science Foundation [grant BCS-1317217] to S.A.T., National Institutes of Health [grant T32ES019851-02] to M.E.B.H. through the Center of Excellence in Environmental Toxicology at the University of Pennsylvania, and the National Institutes of Health [grant 4T32AI007532-19] to D.E.K through the Parasitology program at the University of Pennsylvania.

References

1. Lappalainen T. Functional genomics bridges the gap between quantitative genetics and molecular biology. *Genome Res.* 2015; 25:1427–1431.doi: 10.1101/gr.190983.115 [PubMed: 26430152]
2. Civelek M, Lusis AJ. Systems genetics approaches to understand complex traits. *Nat Rev Genet.* 2014; 15:34–48.doi: 10.1038/nrg3575 [PubMed: 24296534]
3. Labrie V, Buske OJ, Oh E, Jeremian R, Ptak C, Gasiunas G, Maleckas A, Petereit R, Zvirbliene A, Adamonis K, et al. Lactase nonpersistence is directed by DNA-variation-dependent epigenetic aging. *Nat Struct Mol Biol.* 2016; 23:566–573. [PubMed: 27159559]
4. Claussnitzer M, Dankel SN, Kim K-H, Quon G, Meuleman W, Haugen C, Glunk V, Sousa IS, Beaudry JL, Puviindran V, et al. FTO Obesity Variant Circuitry and Adipocyte Browning in Humans. *N Engl J Med.* 2015; 373:895–907.doi: 10.1056/NEJMoa1502214 [PubMed: 26287746]
5. Zhu J, Zhang B, Smith EN, Drees B, Brem RB, Kruglyak L, Bumgarner RE, Schadt EE. Integrating Large-Scale Functional Genomic Data to Dissect the Complexity of Yeast Regulatory Networks. *Nat Genet.* 2008; 40:854–861.doi: 10.1038/ng.167 [PubMed: 18552845]
6. Petrovski S, Goldstein DB. Unequal representation of genetic variation across ancestry groups creates healthcare inequality in the application of precision medicine. *Genome Biol.* 2016; 17:157.doi: 10.1186/s13059-016-1016-y [PubMed: 27418169]
7. Manrai AK, Funke BH, Rehm HL, Olesen MS, Maron BA, Szolovits P, Margulies DM, Loscalzo J, Kohane IS. Genetic Misdiagnoses and the Potential for Health Disparities. *N Engl J Med.* 2016; 375:655–665.doi: 10.1056/NEJMSa1507092 [PubMed: 27532831]
8. Beltrame MH, Rubel MA, Tishkoff SA. Inferences of African evolutionary history from genomic data. *Genet Hum Orig.* 2016; 41:159–166.doi: 10.1016/j.gde.2016.10.002
9. Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, Feldman MW. Genetic Structure of Human Populations. *Science.* 2002; 298:2381–2385.doi: 10.1126/science.1078311 [PubMed: 12493913]
10. Ramachandran S, Deshpande O, Roseman CC, Rosenberg NA, Feldman MW, Cavalli-Sforza LL. Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proc Natl Acad Sci U S A.* 2005; 102:15942–15947.doi: 10.1073/pnas.0507611102 [PubMed: 16243969]
11. Tishkoff SA, Reed FA, Friedlaender FR, Ehret C, Ranciaro A, Froment A, Hirbo JB, Awomoyi AA, Bodo J-M, Doumbo O, et al. The Genetic Structure and History of Africans and African Americans. *Science.* 2009; 324:1035–1044.doi: 10.1126/science.1172257 [PubMed: 19407144]
12. Slatkin M, Racimo F. Ancient DNA and human history. *Proc Natl Acad Sci.* 2016; 113:6380–6387.doi: 10.1073/pnas.1524306113 [PubMed: 27274045]
13. Enattah NS, Jensen TG, Nielsen M, Lewinski R, Kuokkanen M, Rasinpera H, El-Shanti H, Seo JK, Alifrangis M, Khalil IF, et al. Independent Introduction of Two Lactase-Persistence Alleles into Human Populations Reflects Different History of Adaptation to Milk Culture. *Am J Hum Genet.* 2008; 82:57–72.doi: 10.1016/j.ajhg.2007.09.012 [PubMed: 18179885]
14. Tishkoff SA, Reed FA, Ranciaro A, Voight BF, Babbitt CC, Silverman JS, Powell K, Mortensen HM, Hirbo JB, Osman M, et al. Convergent adaptation of human lactase persistence in Africa and Europe. *Nat Genet.* 2007; 39:31–40.doi: 10.1038/ng1946 [PubMed: 17159977]
15. Ranciaro A, Campbell MC, Hirbo JB, Ko W-Y, Froment A, Anagnostou P, Kotze MJ, Ibrahim M, Nyambo T, Omar SA, Tishkoff SA. Genetic Origins of Lactase Persistence and the Spread of Pastoralism in Africa. *Am J Hum Genet.* 2014; 94:496–510.doi: 10.1016/j.ajhg.2014.02.009 [PubMed: 24630847]
- 16. Martin AR, Costa HA, Lappalainen T, Henn BM, Kidd JM, Yee M-C, Grubert F, Cann HM, Snyder M, Montgomery SB, Bustamante CD. Transcriptome Sequencing from Diverse Human Populations Reveals Differentiated Regulatory Architecture. *PLoS Genet.* 2014; 10:e1004549. Using the most ethnically diverse RNA-seq dataset to date, comprising 7 non-European populations from 3 continents, the authors find that ancestry contributes to 25% of observed gene expression variance. doi: 10.1371/journal.pgen.1004549 [PubMed: 25121757]
17. Stranger BE, Montgomery SB, Dimas AS, Parts L, Stegle O, Ingle CE, Sekowska M, Smith GD, Evans D, Gutierrez-Arcelus M, et al. Patterns of Cis Regulatory Variation in Diverse Human

Populations. *PLoS Genet.* 2012; 8:e1002639.doi: 10.1371/journal.pgen.1002639 [PubMed: 22532805]

18. Lappalainen T, Sammeth M, Friedländer MR, 't Hoen PA, Monlong J, Rivas MA, González-Porta M, Kurbatova N, Griebel T, Ferreira PG, et al. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature.* 2013; 501:506–511.doi: 10.1038/nature12531 [PubMed: 24037378]
- 19. Carja O, MacIsaac JL, Mah SM, Henn BM, Kobor MS, Feldman MW, Fraser HB. Worldwide patterns of human epigenetic variation. *bioRxiv.* 2015; Using a subset of individuals from Martin *et al.* [16], this study finds that ancestry contributes more to DNA methylation variance than gene expression variance (6% vs. 2%, respectively), and that populations can be more reliably clustered using methylation measurements. doi: 10.1101/021931
20. Yuan Y, Tian L, Lu D, Xu S. Analysis of Genome-Wide RNA-Sequencing Data Suggests Age of the CEPH/Utah (CEU) Lymphoblastoid Cell Lines Systematically Biases Gene Expression Profiles. *Sci Rep.* 2015; 5:7960.doi: 10.1038/srep07960 [PubMed: 25609584]
21. Çali kan M, Pritchard JK, Ober C, Gilad Y. The Effect of Freeze-Thaw Cycles on Gene Expression Levels in Lymphoblastoid Cell Lines. *PLoS ONE.* 2014; 9:e107166.doi: 10.1371/journal.pone.0107166 [PubMed: 25192014]
22. Akey JM, Biswas S, Leek JT, Storey JD. On the design and analysis of gene expression studies in human populations. *Nat Genet.* 2007; 39:807–808.doi: 10.1038/ng0707-807 [PubMed: 17597765]
- 23. Hughes DA, Kircher M, He Z, Guo S, Fairbrother GL, Moreno CS, Khaitovich P, Stoneking M. Evaluating intra- and inter-individual variation in the human placental transcriptome. *Genome Biol.* 2015; 16:54. This study measured inter-population gene expression differences in placenta, notable for being a recent examination of a native tissue other than blood from nominally healthy cohorts. They find that ancestry contributes to expression variance on par with other studies (7.8%), but genes that are differentially expressed between groups are enriched in biological pathways related to immunity, cell signaling, and metabolism. doi: 10.1186/s13059-015-0627-z [PubMed: 25887593]
24. Karlsson EK, Kwiatkowski DP, Sabeti PC. Natural selection and infectious disease in human populations. *Nat Rev Genet.* 2014; 15:379–393.doi: 10.1038/nrg3734 [PubMed: 24776769]
25. Fan S, Hansen MEB, Lo Y, Tishkoff SA. Going global by adapting local: A review of recent human adaptation. *Science.* 2016; 354:54–59.doi: 10.1126/science.aaf5098 [PubMed: 27846491]
26. Albert FW, Kruglyak L. The role of regulatory variation in complex traits and disease. *Nat Rev Genet.* 2015; 16:197–212. [PubMed: 25707927]
27. Taudt A, Colome-Tatche M, Johannes F. Genetic sources of population epigenomic variation. *Nat Rev Genet.* 2016; 17:319–332. [PubMed: 27156976]
28. Fraser HB, Lam LL, Neumann SM, Kobor MS. Population-specificity of human DNA methylation. *Genome Biol.* 2012; 13:R8–R8.doi: 10.1186/gb-2012-13-2-r8 [PubMed: 22322129]
- 29. Fagny M, Patin E, MacIsaac JL, Rotival M, Flutre T, Jones MJ, Siddle KJ, Quach H, Harmant C, McEwen LM, et al. The epigenomic landscape of African rainforest hunter-gatherers and farmers. *Nat Commun.* 2015; 6:10047. This is the only study to measure epigenetic differences across populations in a native tissue (whole blood) and in a non-Western setting. Comparing RHG and Bantu-speaking agriculturalists and hunter-gatherers, the authors separate the contribution of recent environment (forest vs. urban) and ancestry (RHG vs. Bantu), finding that these factors alter methylation of genes related to immunity and development, respectively. doi: 10.1038/ncomms10047 [PubMed: 26616214]
30. Kasowski M, Kyriazopoulou-Panagiotopoulou S, Grubert F, Zaugg JB, Kundaje A, Liu Y, Boyle AP, Zhang QC, Zakharia F, Spacek DV, et al. Extensive Variation in Chromatin States Across Humans. *Science.* 2013; 342:750–752.doi: 10.1126/science.1242510 [PubMed: 24136358]
31. Idaghdour Y, Czika W, Shianna KV, Lee SH, Visscher PM, Martin HC, Miclaus K, Jadallah SJ, Goldstein DB, Wolfinger RD, Gibson G. Geographical genomics of human leukocyte gene expression variation in southern Morocco. *Nat Genet.* 2010; 42:62–67.doi: 10.1038/ng.495 [PubMed: 19966804]
32. Raj T, Rothamel K, Mostafavi S, Ye C, Lee MN, Replogle JM, Feng T, Lee M, Asinowski N, Frohlich I, et al. Polarization of the Effects of Autoimmune and Neurodegenerative Risk Alleles in Leukocytes. *Science.* 2014; 344:519–523.doi: 10.1126/science.1249547 [PubMed: 24786080]

33. Lee MN, Ye C, Villani A-C, Raj T, Li W, Eisenhaure TM, Imboywa SH, Chipendo PI, Ran FA, Slowikowski K, et al. Common Genetic Variants Modulate Pathogen-Sensing Responses in Human Dendritic Cells. *Science*. 2014; 343:1246980–1246980.doi: 10.1126/science.1246980 [PubMed: 24604203]
34. Ye CJ, Feng T, Kwon H-K, Raj T, Wilson M, Asinovski N, McCabe C, Lee MH, Frohlich I, Paik H, et al. Intersection of population variation and autoimmunity genetics in human T cell activation. *Science*. 2014; 345:1254665–1254665.doi: 10.1126/science.1254665 [PubMed: 25214635]
- 35. Quach H, Rotival M, Pothlichet J, Loh Y-HE, Dannemann M, Zidane N, Laval G, Patin E, Harmant C, Lopez M, et al. Genetic Adaptation and Neandertal Admixture Shaped the Immune System of Human Populations. *Cell*. n.d; 167:643–656.e17. This study, along with Nédélec *et al.* [36], is one of the only studies to incorporate immune challenges of primary cells from individuals of different ancestries with eQTL mapping and scans of natural selection. Measuring CD14⁺Rosalie Sears T cell response to bacterial and viral ligands, the authors find that Europeans show evidence of a decreased response to viral infection, and that several key immunity genes show evidence of introgression from Neanderthals. doi: 10.1016/j.cell.2016.09.024 [PubMed: 27768888]
- 36. Nédélec Y, Sanz J, Baharian G, Szpiech ZA, Pacis A, Dumaine A, Grenier J-C, Freiman A, Sams AJ, Hebert S, et al. Genetic Ancestry and Natural Selection Drive Population Differences in Immune Responses to Pathogens. *Cell*. n.d; 167:657–669.e21. Similar to Quach *et al.* [35], this study incorporates immune challenges with eQTL mapping and scans of natural selection, though in a cohort of European and African Americans. Macrophages derived *in vitro* were challenged with *Listeria* and *Salmonella*, which triggered distinct immune responses. African ancestry was associated with an overall stronger response to infection, specifically in genes related to inflammation, as well as a greater ability to clear bacterial infections. doi: 10.1016/j.cell.2016.09.025 [PubMed: 27768889]
37. Scotti MM, Swanson MS. RNA mis-splicing in disease. *Nat Rev Genet*. 2016; 17:19–32. [PubMed: 26593421]
38. Li YI, van de Geijn B, Raj A, Knowles DA, Petti AA, Golan D, Gilad Y, Pritchard JK. RNA splicing is a primary link between genetic variation and disease. *Science*. 2016; 352:600–604.doi: 10.1126/science.aad9417 [PubMed: 27126046]
39. Abdel-Ghany SE, Hamilton M, Jacobi JL, Ngam P, Devitt N, Schilkey F, Ben-Hur A, Reddy ASN. A survey of the sorghum transcriptome using single-molecule long reads. *Nat Commun*. 2016; 7:11706.doi: 10.1038/ncomms11706 [PubMed: 27339290]
40. Vaquero-Garcia J, Barrera A, Gazzara MR, González-Vallinas J, Lahens NF, Hogenesch JB, Lynch KW, Barash Y. A new view of transcriptome complexity and regulation through the lens of local splicing variations. *eLife*. 2016; 5:e11752.doi: 10.7554/eLife.11752 [PubMed: 26829591]
41. van Dijk EL, Auger H, Jaszczyszyn Y, Thermes C. Ten years of next-generation sequencing technology. *Trends Genet*. 2014; 30:418–426.doi: 10.1016/j.tig.2014.07.001 [PubMed: 25108476]
42. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, Szczekniak MW, Gaffney DJ, Elo LL, Zhang X, Mortazavi A. A survey of best practices for RNA-seq data analysis. *Genome Biol*. 2016; 17:13.doi: 10.1186/s13059-016-0881-8 [PubMed: 26813401]
43. Battle A, Khan Z, Wang SH, Mitrano A, Ford MJ, Pritchard JK, Gilad Y. Impact of regulatory variation from RNA to protein. *Science*. 2015; 347:664–667.doi: 10.1126/science.1260793 [PubMed: 25657249]
44. Cenik C, Cenik ES, Byeon GW, Grubert F, Candille SI, Spacek D, Alsallakh B, Tilgner H, Araya CL, Tang H, et al. Integrative analysis of RNA, translation, and protein levels reveals distinct regulatory variation across humans. *Genome Res*. 2015; 25:1610–1621.doi: 10.1101/gr.193342.115 [PubMed: 26297486]
45. Brar GA, Weissman JS. Ribosome profiling reveals the what, when, where and how of protein synthesis. *Nat Rev Mol Cell Biol*. 2015; 16:651–664. [PubMed: 26465719]
46. Sajuthi SP, Sharma NK, Chou JW, Palmer ND, McWilliams DR, Beal J, Comeau ME, Ma L, Calles-Escandon J, Demons J, et al. Mapping adipose and muscle tissue expression quantitative trait loci in African Americans to identify genes for type 2 diabetes and obesity. *Hum Genet*. 2016; 135:869–880.doi: 10.1007/s00439-016-1680-8 [PubMed: 27193597]

47. Ardlie KG, Deluca DS, Segrè AV, Sullivan TJ, Young TR, Gelfand ET, Trowbridge CA, Maller JB, Tukiainen T, Lek M, et al. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science*. 2015; 348:648–660.doi: 10.1126/science.1262110 [PubMed: 25954001]
48. Aguet F, Brown AA, Castel S, Davis JR, Mohammadi P, Segre AV, Zappala Z, Abell NS, Fresard L, Gamazon ER, et al. Local genetic effects on gene expression across 44 human tissues. *bioRxiv*. 2016; doi: 10.1101/074450.
49. Barbeira A, Dickinson SP, Torres JM, Torstenson ES, Zheng J, Wheeler HE, Shah KP, Edwards T, Nicolae D, Cox NJ, Im HK. Integrating tissue specific mechanisms into GWAS summary results. *bioRxiv*. 2016; doi: 10.1101/045260.
50. Ongen H, Brown AA, Delaneau O, Panousis N, Nica AC, Dermitzakis ET. Estimating the causal tissues for complex traits and diseases. *bioRxiv*. 2016; doi: 10.1101/074682.
51. Avior Y, Sagi I, Benvenisty N. Pluripotent stem cells in disease modelling and drug discovery. *Nat Rev Mol Cell Biol*. 2016; 17:170–182. [PubMed: 26818440]
52. Burrows CK, Banovich NE, Pavlovic BJ, Patterson K, Gallego Romero I, Pritchard JK, Gilad Y. Genetic Variation, Not Cell Type of Origin, Underlies the Majority of Identifiable Regulatory Differences in iPSCs. *PLoS Genet*. 2016; 12:e1005793.doi: 10.1371/journal.pgen.1005793 [PubMed: 26812582]
53. Kilpinen H, Goncalves A, Leha A, Afzal V, Ashford S, Bala S, Bensaddek D, Casale FP, Culley O, Danacek P, et al. Common genetic variation drives molecular heterogeneity in human iPSCs. *bioRxiv*. 2016; doi: 10.1101/055160.
54. Johnson CH, Ivanisevic J, Siuzdak G. Metabolomics: beyond biomarkers and towards mechanisms. *Nat Rev Mol Cell Biol*. 2016; 17:451–459. [PubMed: 26979502]
55. Schlebusch CM, Gattepaille LM, Engström K, Vahter M, Jakobsson M, Broberg K. Human Adaptation to Arsenic-Rich Environments. *Mol Biol Evol*. 2015; 32:1544–1555.doi: 10.1093/molbev/msv046 [PubMed: 25739736]
56. Spielman RS, Bastone LA, Burdick JT, Morley M, Ewens WJ, Cheung VG. Common genetic variants account for differences in gene expression among ethnic groups. *Nat Genet*. 2007; 39:226–231.doi: 10.1038/ng1955 [PubMed: 17206142]
57. Storey JD, Madeoy J, Strout JL, Wurfel M, Ronald J, Akey JM. Gene-Expression Variation Within and Among Human Populations. *Am J Hum Genet*. 2007; 80:502–509. [PubMed: 17273971]
58. Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, Thorne N, Redon R, Bird CP, de Grassi A, Lee C, et al. Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science*. 2007; 315:848–853.doi: 10.1126/science.1136678 [PubMed: 17289997]
59. Kasowski M, Grubert F, Heffelfinger C, Hariharan M, Asabere A, Waszak SM, Habegger L, Rozowsky J, Shi M, Urban AE, et al. Variation in Transcription Factor Binding Among Humans. *Science*. 2010; 328:232–235.doi: 10.1126/science.1183621 [PubMed: 20299548]
60. Wu L, Candille SI, Choi Y, Xie D, Li-Pook-Than J, Tang H, Snyder M. Variation and Genetic Control of Protein Abundance in Humans. *Nature*. 2013; 499:79–82.doi: 10.1038/nature12223 [PubMed: 23676674]

- Human inter-population variation in gene expression and regulation is reviewed
- Expression, eQTLs, and epigenetic modifications can vary between groups
- Sampling greater global diversity is required to assess population specificity
- More cell type sampling beyond blood is needed to assess tissue specificity

Table 1

Cross-population studies of gene expression and gene regulation, including the sampled populations, the tissue or cell type examined, and the technology platform used.

Individuals Sampled	Tissue	Assays	Reference
60 CEU, 41 CHB, 41 JPT	LCLs	RNA microarray, genotypes from HapMap	[56]
8 CEU, 8 YRI	LCLs	RNA microarray, genotypes from HapMap	[57]
60 CEU, 45 CHB, 45 JPT, 60 YRI	LCLs	RNA BeadChip array, CNV from HapMap 1	[58]
5 European, 2 East Asian, 3 Nigerian	LCLs	ChIP-seq for TFs (NFkB and PolII)	[59]
194 individuals of Arab and Amazigh ancestry 109 CEU, 80 CHB, 82	Leukocytes	RNA BeadChip array, DNA SNP chip genotyping	[31]
GIH, 82 JPT, 82 LWK, 45 MEX, 138 MKK, 108 YRI (726 Total)	LCLs	RNA BeadChip array, Genotype data from HapMap 3	[17]
30 CEU trios, 30 YRI trios	LCLs	Methylation BeadChip array, array genotypes from HapMap	[28]
5 CEU, 7 YRI, 2 Asian, 1 Caucasian, 4 San	LCLs	RNA-seq, ChIP-seq for chromatin modifications (H3K27ac, H3K4me1, H3K4me3, H3K36me3, and H3K27me3) and DNA-binding proteins s (CTCF and SA1)	[30]
91 CEU, 95 FIN, 94 GBR, 93 TSI, 89 YRI	LCLs	mRNA and small RNA-seq, Genotypes from 1000 Genomes Phase 1	[18]
53 CEU, 33 YRI, 8 CHB, 1 JPT	LCLs	2D LC-MS/MS, Genotype data from HapMap 3	[60]
4 San, 7 Mbuti Pygmies, 7 Mozabites, 7 Pathan, 7 Cambodians, 6 Yakut, 7 Maya	LCLs	RNA-seq, whole exome sequencing, whole genome sequencing	[16]
213 European Americans (EA), 112 African Americans (AA), 74 East Asian Americans (EA)	CD4+ T lymphocytes, CD14+CD16– monocytes	mRNA GeneChip array, Exome BeadChip array	[34]
6 Yakut, 7 Cambodians, 7 Pathan, 7 Mozabite, 7 LCLs Maya.		Methylation BeadChip array, genotypes from HGDP-CEPH	[19]
112 Central African rainforest hunter-gatherers, 260 Bantu-speakers	Whole blood	Methylation microarray, DNA microarray genotyping	[29]
10 AA, 10 EU, 10 EA, 10 South Asian-American (SA)	Placenta	RNA-seq	[23]
80 AA, 94 EU	CD14+ macrophages	RNA-seq, ATAC-seq, Exome BeadChip genotyping	[36]
100 European, 100 African in Belgium	CD14+ monocytes	RNA-seq, whole-exome sequencing, BeadChip genotyping	[35]