Global analysis of disease-related DNA sequence variation in 10 healthy individuals: Implications for whole genome-based clinical diagnostics

Barry Moore, MS[1], Hao Hu, BS[1], Marc Singleton, MS[1], Francisco M. De La Vega, DSc[2], Martin G. Reese, PhD[3], and Mark Yandell, PhD[1]

Background: Understanding how sequence variants within healthy genomes are distributed with respect to ethnicity and disease-implicated genes is an essential first step toward establishing baselines for personalized genomic medicine. **Methods:** In this study, we present an analysis of 10 genomes from healthy individuals of various ethnicities, produced using six different sequencing technologies. In total, these genomes contain more than 34 million single-nucleotide variants. **Results:** We have analyzed these variants from a clinical perspective, assaying the influence of sequencing technology and ethnicity on prognosis. We have also examined the utility of OMIM and the disease-gene literature for determining the impact of rare, personal variants on an individual's health. **Conclusions:** Our analyses demonstrate that clinical prognoses are complicated by sequencing platform-specific errors and ethnicity. We show that disease-causing alleles are globally distributed along ethnic lines, with alleles known to be disease causing in Eurasians being significantly more likely to be homozygous in Africans. ***Genet Med*** **2011:13(3):210–217.**

Key Words: *personal genomes, genome analysis, personalized genomics*

What does it mean to have a "healthy" genome? Neither J. Craig Venter's nor James Watson's genomes contain any alleles likely to cause or strongly predispose them to genetic illness.[1,2] They are also not heterozygous for any alleles raising serious reproductive issues. Given these facts, some have expressed skepticism regarding the prognostic value of personal genome sequences.[3,4] To date, the standard reply to the skeptic has been that healthy adults have healthy genomes. Although reasonable, this rebuttal presumes that we know what a healthy genome is. No doubt, a clean bill of genomic health will be the most common clinical scenario in genomic medicine. However, just what does a healthy genome look like? What is the impact of sequencing technology on prognostic accuracy? What role will ethnicity play in prognosis? Finally, how useful will existing resources, such as OMIM, be for categorizing personal genome variants as deleterious? The answers to these questions are of immediate importance for the future of genomic medicine.

To answer these questions, we have assembled a standardized dataset of single-nucleotide variants (SNVs), also called single nucleotide polymorphisms, from 10 personal genome sequences.[1,2,5–11] These genomes represent both sexes and a variety of ethnic backgrounds: three Africans, two Asians, and five whites. This second feature of the data set has allowed us to assay the impact of ethnicity on variant locations. Six different sequencing technologies are also represented. One genome was sequenced with Sanger technology,[12] one using the Roche 454 platform (Life Sciences, Branford, CT), three with ABI SOLiD (Applied Biosystems/Life Technologies, Carlsbad, CA), three with Illumina Genome Analyzer (GA) (Illumina, San Diego, CA), one with Helicos Biosciences (Cambridge, MA), and one genome was sequenced by Complete Genomics (Mountain View, CA). Moreover, one individual's genome is represented twice, sequenced on both the Illumina GA and ABI SOLiD platforms.

The diversity of ethnic backgrounds and sequencing technologies used to produce this dataset make it ideal for investigations of the impact of sequencing platform on SNV ascertainment and of ethnicity on SNV distributions. To address the clinical relevance of these factors, we have carried out two additional large-scale analyses. First, we analyzed these genomes' 34 million variants with respect to their intersection with OMIM variants, which we have systematically mapped forward to the current genome assembly; this has made it possible to globally assay the distributions of OMIM alleles in different ethnic backgrounds. Additional clinical perspective is obtained through a second, complementary analysis using a disease-gene classification system based on MeSH[13] and *Harrison's Internal Medicine.*[14] This classification system is a first step toward comprehensive disease-gene ontology, and it has allowed us to examine individual genomic variation associated with specific areas of the clinical disease-gene literature, such as oncogenesis, neonatal development, and cardiovascular disease.

Taken together, these analyses illustrate some of the challenges of whole genome variation analysis and provide a first glimpse of trends and distributions indicative of healthy genomes that will also be important for clinical prognosis of personal genome sequences.

MATERIALS AND METHODS

Genome sources

Genomes were downloaded from sites made available by the publishing authors or were obtained by personal communication with the authors. The groups originally publishing the genome are described in Table 1.

From the [1]Department of Human Genetics, Eccles Institute of Human Genetics, University of Utah School of Medicine, Salt Lake City, Utah; [2]Research and Development, Genetic Systems Division, Life Technologies, Foster City, California; and [3]Omicia, Inc., Emeryville, California.

Mark Yandell, PhD, Department of Human Genetics, Eccles Institute of Human Genetics, University of Utah School of Medicine, Salt Lake City, UT. E-mail: myandell@genetics.utah.edu.

Disclosure: Barry Moore and Mark Yandell have worked as paid consultants for Omicia Inc. Francisco M. De La Vega was an employee and held stock options of Life Technologies when this research was conducted. The other authors declare no conflict of interest.

Supplemental digital content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's Web site (www.geneticsinmedicine.org).

Submitted for publication November 9, 2010.

Accepted for publication January 6, 2011.

Published online ahead of print February 15, 2011.

DOI: 10.1097/GIM.0b013e31820ed321

Table 1 The 10 genomes dataset

Genome	Individual	Ethnicity	Platform	References	SNV count	Ti/Tv	Het/Hom	Nonsynon	OMIM loci (hom/het)	Healthy variant
0	NA19240	African	ABI SOLiD	De la Vega et al.[11]	4,082,270	2.03	1.6	11,906	48/56	25
1	NA18507	African	Illumina	Bentley et al.[5]	4,192,779	2.06	1.8	12,268	53/59	23
2	NA18507	African	ABI SOLiD	McKernan et al.[6]	4,241,042	2.02	1.7	12,620	54/59	25
3	Chinese	Asian	Illumina	Wang et al.[7]	3,074,097	2.01	1.3	9,868	46/58	27
4	Korean	Asian	Illumina	Ahn et al.[8]	3,439,107	2.07	1.4	10,622	39/71	19
5	Venter	White	Sanger	Levy et al.[1]	3,075,858	2.04	1.1	9,728	40/68	27
6	Watson	White	Roche 454	Wheeler et al.[2]	3,261,381	2.01	1.3	9,996	36/53	23
7	NA07022	White	CGenomics	Drmanac et al.[9]	3,161,103	2.16	1.7	10,681	41/73	30
8	NA12878	White	ABI SOLiD	De la Vega et al.[11]	3,274,358	1.97	1.1	10,001	33/70	31
9	Quake	White	Helicos	Pushkarev et al.[10]	2,794,408	2.03	1.4	9,968	32/81	24

For ease of discussion, genomes are referred to throughout by a single numeric identifier (0–9). Columns: 1, ID; 2, individual; 3, ethnicity; 4, platform; 5, references; 6, number of SNVs; 7, transition/transversion ratio; 8, heterozygous/homozygous ratio; 9, number of nonsynonymous SNVs; 10, number of OMIM alleles (homozygous positions/heterozygous positions); 11, number of positions where the personal genome contains a healthy allele, but the reference has the OMIM disease allele.

GVF conversion

We converted each genomes' original variant file to GVF format[15] to facilitate our analyses. Complete documentation for the GVF format can be found at http://www.sequenceontology.org/resources/gvf.html. Each file is available for download at http://www.sequenceontology.org/resources/10Gen.html.[17] In every case, we used all variants in the original source variant file, as this proved the most logical starting point for comparison, allowing each individual provider to identify variants using their own procedures. Thus, our comparisons of accuracy reflect both the sequencing platform and the (then) current read mapping and variant calling pipeline used in the original publications of these genomes.

Intergenome distance calculations

Figure 1 is based on the intergenome distance between each of the variant files. We used the following distance metric for this calculation: $D_{ij} = (N_s - [N_s \cap N_L])/N_s$, where D_{ij} is the total intergenome distance, N_s is the total number of variants in the genome (i or j) having the fewest number of variants; N_L is the total number of variants in the genome (i or j) having the greatest number of variants; and $N_s \cap N_L$ denotes the number of variants in the intersection of the two genomes. One important feature of this distance is that it is based on locations of variants, not the nucleotide of the variant; another is that D_{ij} is robust with respect to the depth of coverage. This fact controls for depth of coverage differences in our dataset.

Tree creation

We calculated a distance matrix, D for every D_{ij} pair in our dataset on a chromosome-by-chromosome basis. We restricted this calculation to the 22 autosomes to avoid complications associated with comparisons of XX and XY genomes. Phylip[18] together with the resulting 22 distance matrices was used to produce a neighbor-joining tree (default options were used) for each of the 22 chromosomes. Phylip Consense was then used to produce the consensus tree shown in Figure 1. Numbers attached to the nodes refer to the number of autosomes (N out of 22) for which that node was seen. Figure 2 was produced using the same procedures; only this time, the input data were restricted to only those variants for each genome corresponding to known OMIM alleles. The values labeling the nodes in Figure 2 are bootstraps[19] based on 100 replicates using Phylip's bootstrapping procedures.[20]

Classification of variants

We used a CGL-based script[21] to classify each SNV with respect to genomic location (integenic, intron, CDS, and untranslated region) using a nonredundant set of the gene annotations from the University of California Santa Cruz (UCSC) genome browser,[22] knownGene, and RefSeq[23] tables for build hg18 (NCBI36.3). Variants intersecting CDS regions were further classified into one of six categories on the basis of the change to the protein: synonymous, conservative substitution, nonconservative substitution, and stop gained. Designation as conservative versus nonconservative was based on the BLOSUM 62 matrix[24]; changes with a score ≥ 0 were considered conservative and those <0, nonconservative. In total, of the 34 million variants in our dataset, 38% were genic. Of these, 36% mapped to introns and 2% to exons. Among variants mapping to exons, 33% were protein coding, that is, they mapped to the CDS portion of exons; 51% of these coding variants were synonymous, 28% produced conservative substitutions, and 20% nonconservative substitutions; 0.5% produced stop-gained substitutions.

OMIM mapping

Disease-causing alleles and sequence variants implicated in disease were parsed from OMIM flat file documents.[16] OMIM indexes its coding variants according to the codon or amino acid they alter on the messenger RNA or the protein sequence reported in the original publication, rather than the currently annotated sequence. To overcome this, we used a previously described mapping process[25] to move the OMIM alleles forward to the current annotations. By far, the most common change in gene annotation over the years has been the addition of additional 5′ exons to previously annotated transcripts,[26] resulting in amino-terminal extensions in the encoded protein.

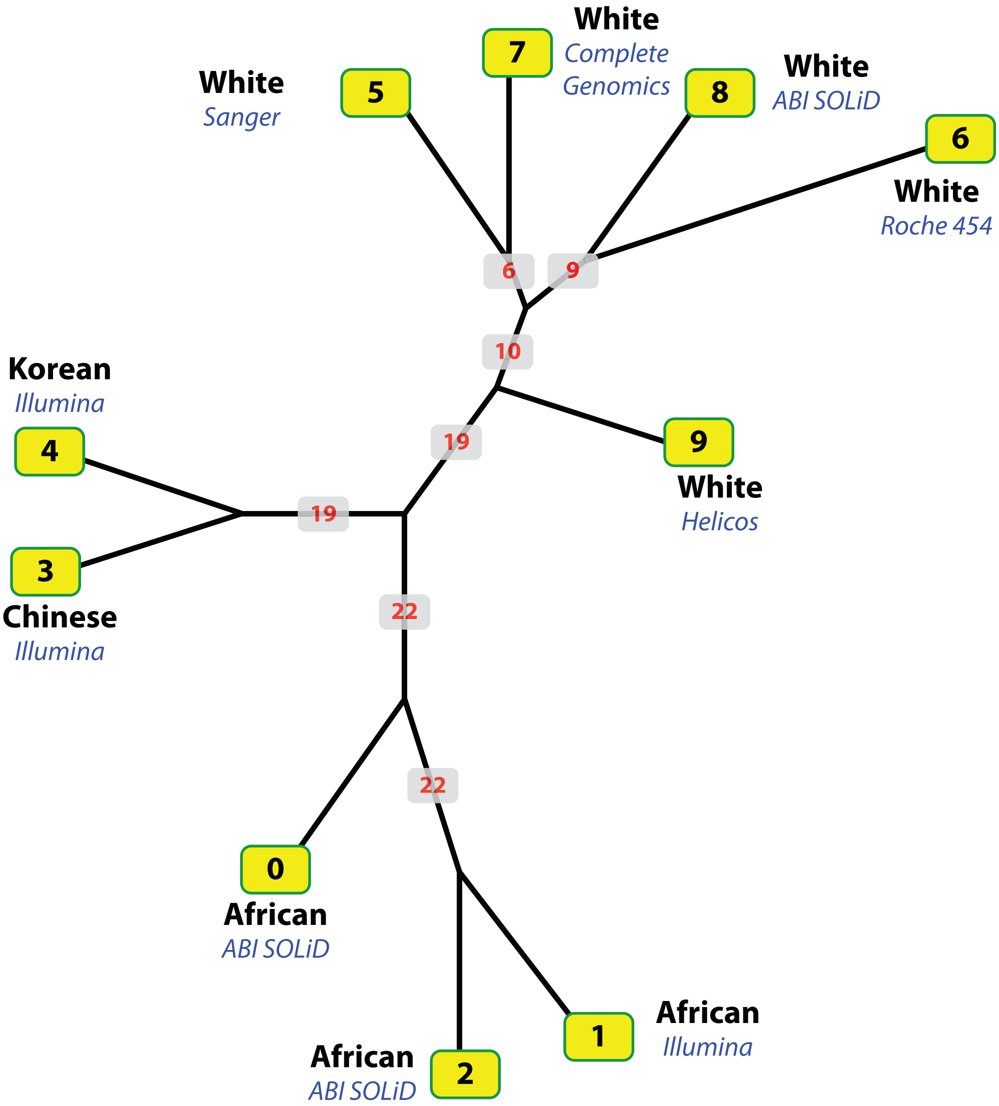

Fig. 1. Consensus neighbor-joining tree based on intergenome similarities among the 34 million SNV locations. Leaves are labeled with each genome's ID, ethnicity, and sequencing platform used to produce the sequence (Table 1). Numbers in red denote the number of autosomes supporting that node. See "Materials and Methods" for details of similarity calculations.

The mapping procedure exploits this fact and asks whether there was a single offset that will map each OMIM coding variant in a gene to a corresponding amino acid in the currently annotated protein. Assuming that two or more alleles are available, the probability that this would occur by chance would be approximately 1 of 20,[2] neglecting amino acid frequency biases. This process allowed us to map forward 9247 (88.7%) of OMIM variants located within the CDS regions of genes.

Ontology creation

We used a set of 3626 hand-curated human disease genes (the "Omicia disease gene set"), which have been documented in the literature as playing a causative or predisposing role in one or more human diseases. This list of genes includes and extends a human disease-gene set previously published by Jimenez-Sanchez et al.[27] containing 923 genes. We used a semiautomated natural language processing-based approach,[28] followed by manual review and curation to identify and assign an additional 2703 genes to our disease-gene ontology. To do so, we used the HGMD database[29] and the literature links associated with each of its curated disease-causing and predisposing variants to identify Medline abstracts related to those genes and their associated mutations. Next, we extracted the MeSH descriptors associated with those abstracts and used these to connect 1024 genes in HGMD and their associated mutations with MeSH disease terms. Then we back populated the MeSH ontology with genes that were listed in the original abstract; a round of manual review followed this. Next, we manufactured a simplified, more clinically focused disease-gene ontology by assigning MeSH disease IDs to a list of 12 high-level disease categories (Fig. 3) based on *Harrison's Internal Medicine.*[14] We also assigned additional genes to this list by hand based on peer-reviewed literature. The resulting ontology is stored in OBO format.[30] At the time of the analyses reported herein, we

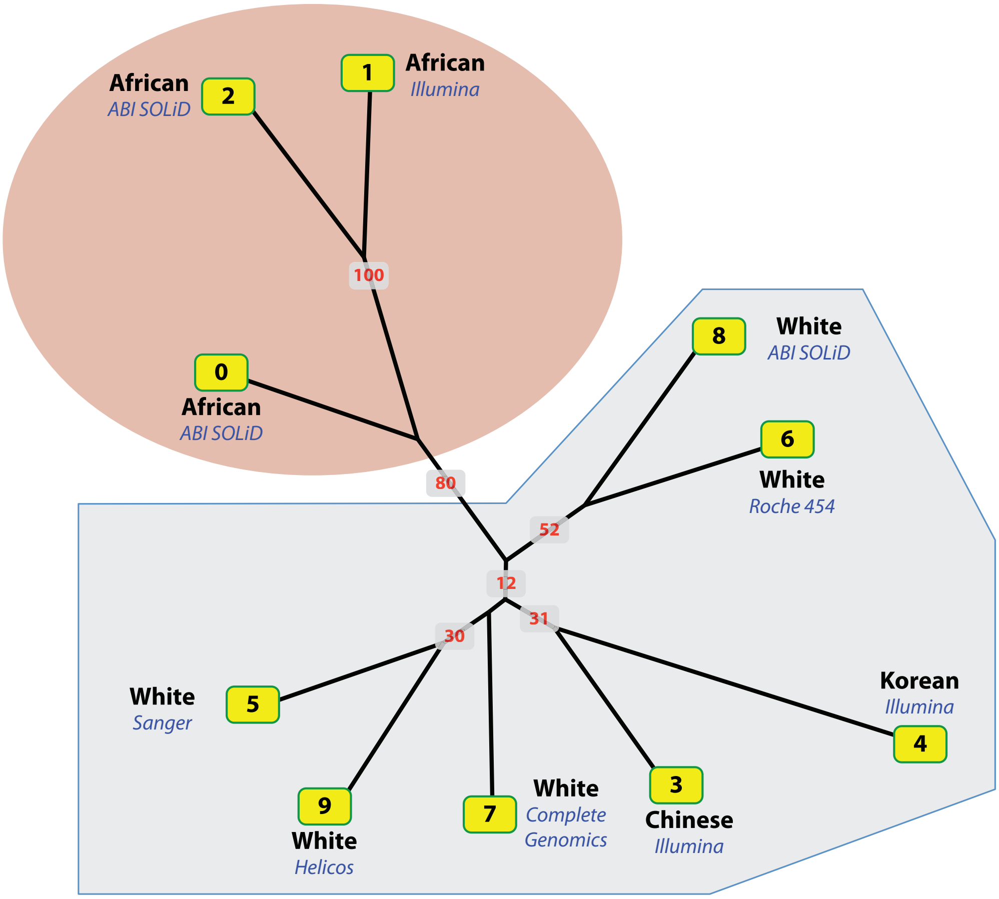

Fig. 2. Consensus neighbor-joining tree based on shared OMIM alleles. Tree constructed using same procedure used to construct Figure 1. This time, however, restricting the data to those locations where either the personal genome or reference genome contained an OMIM allele. The tree was bootstrapped 100 times, and labels on nodes are the resulting bootstraps. The tree is not as well resolved as the one shown in Figure 1 due to the many fewer OMIM alleles, when compared with all SNVs; nevertheless, the same trends with respect to ethnicity seen in Figure 1 are still present.

had manually reviewed 2703 assignments, extending the original list of Jimenez-Sanchez et al. to 3626 genes.

Expectation calculation of ontology categories

We normalized the counts of coding variants for each disease category (Figure, Supplemental Digital Content 1, http://links.lww.com/GIM/A160) in our ontology, to control for differences in gene numbers, CDS lengths, and codon usage. We used the following formula to do so. Expected counts = $T*C_w*C_a$, where T is the total number of coding SNVs genome wide for a given individual. C_w = class-specific CDS footprint/CDS footprint for the entire genome, where the term footprint refers to the sum of the nonredundant nucleotide lengths of all CDSs either within a disease category or genome wide. The third term, C_a, is a constant, equal to 0.77; this value is the fraction of nucleotide positions within CDSs, wherein a nucleotide change can produce a nonsynonymous change. Its value is based on the genetic code and amino acid frequencies for the proteome as a whole. Disease class-specific variance in this value is negligible (data not shown). Note that although this correction controls for aspects of gene number, amino acid frequency, and category size, it does not correct for selection and population-history effects—this, however, is precisely the point: our goal in producing Figure 3 was to assay the extent to which natural selection, disease prevalence, and population history have acted to perturb the numbers of coding variants within a disease category relative to this basic expectation. The numbers plotted in Figure 3 are the percent difference in the observed and expected counts for each category.

RESULTS

Dataset

We collected variant files for individual whole genomes and converted their contents to the Genome Variant Format (GVF)[15] to facilitate analysis. See "Materials and Methods" section for complete description of the dataset. In total, the dataset contains 34.6 million SNVs relative to the reference human genome. Table 1 provides a summary of the dataset.

The total number of SNVs per genome ranges from 2.8 to 4.2 M. We have characterized these variants with respect to their intersections with gene annotations (untranslated region, coding sequence [CDS], intron, and intergenic) and impact on translation (termination codons, synonymous, and nonsynonymous amino acid changes)—all with reference to the currently annotated 25,118

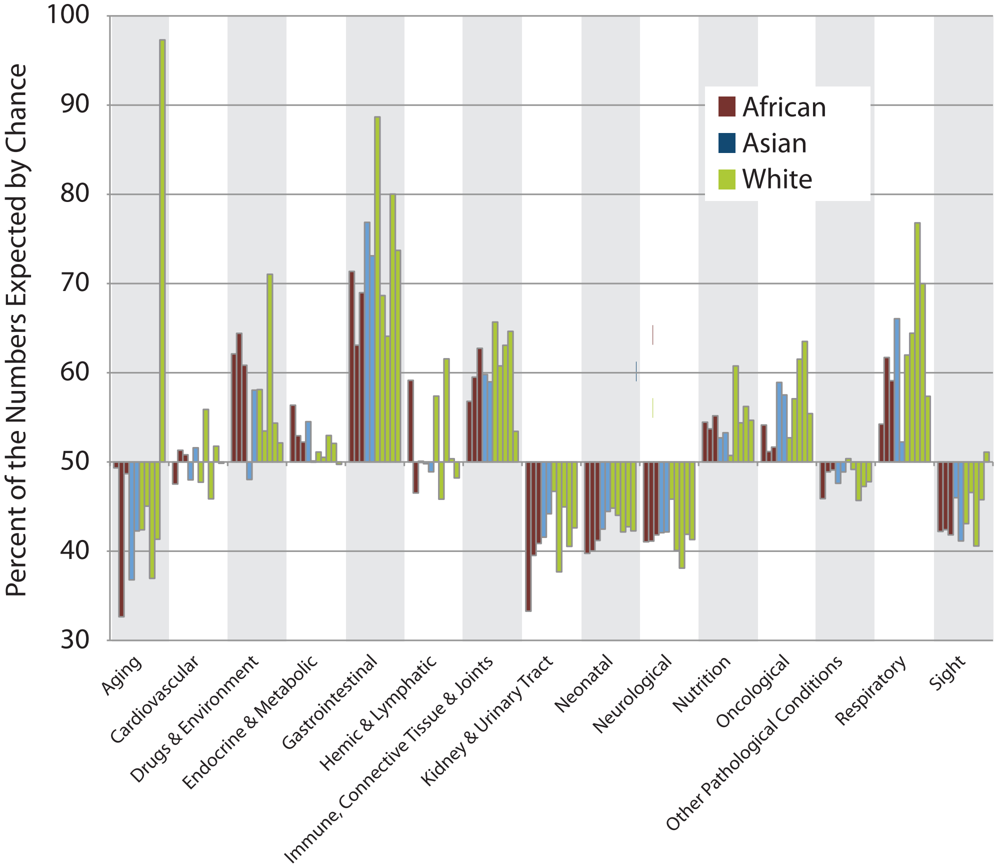

Fig. 3. Genomic load differs significantly within and between disease categories. *x* axis: disease category and *y* axis: percent deviation from expected genetic load (nonsynonymous variants) for each genome by and by category. Overall, genes in disease categories on average contained 50% fewer variants than expected by chance, thus the *y* axis is set to cross at 50% to clearly show the differences between categories. Data for individual genomes are plotted according to their listed order in Table 1 and color coded by ethnicity.

human protein-coding genes, assembled from 265,019 exons (See "Classification of Variants" in "Materials and Methods" section**)**. Variants were further classified as being present in a heterozygous or homozygous state; 1.96% of all variants are located in exons, with 0.64% in protein-coding sequences. For the individuals studied herein, an individual carries between 19,691 and 25,894 protein-coding SNVs in their genome with respect to the reference genome. Of these, approximately 49% are amino acid changing (Table 1). On a gene-by-gene basis, protein-coding sequences are altered in a total of 1,328–1,508 genes. Ten percent of these are entirely novel personal variants—still not included in dbSNP by release 132. We also examined the rate of novel variant discovery. To do so, we determined the number of variants found in each genome not present in dbSNP 126, the last version of dbSNP before the publication of any personal genomes and the current release, dbSNP 132 (Fig. 4). For variants not present in dbSNP 126, 66% were unique to a single genome, 22% of these were common to two, 7% to three, and only 815 variants were common to all 10 genomes. By dbSNP release 132, most of these variants were no longer novel due to inclusion of variants from personal genome sequences and 1000Genomes Project data in dbSNP.

A high-level perspective on variants in the 10 genomes

To obtain a high level perspective of the dataset, we constructed a neighbor-joining tree based on intergenome similarities among the 34 million SNV locations (Fig. 1). The more similar two genomes are with respect to the locations of their variants, the closer they lie to one another in the tree. See "Materials and Methods" section for details of tree construction. The topology of tree shown in Figure 1 makes clear several important trends in the data. First, the variant data are highly structured: some of the genomes are much more similar to one another than others. Globally, genomes 3 and 4, for example, share 46.5% of their variant locations in common, whereas the percentage for genomes 1 and 6 is approximately 23.9%. We also examined how these similarities varied from autosome to autosome. To do so, we constructed subtrees for each autosome and asked how often trees based on a single autosome supported nodes in the tree shown in Figure 1; these are the numbers shown in red in Figure 1. Genomes 1 and 2, for example, are always each other's closest neighbor, regardless of which of the 22 autosomes are compared. This tendency is less pronounced for other genome pairs. Genomes 3 and 4 are each other's closest neighbor for 19 of 22 autosomes. Thus, a given individual's closest neighbor in variant space is usually—but not always—the same from autosome to autosome.

Mapping ethnicity onto the tree in Figure 1 suggests that the dominant trend structuring SNV locations is ethnicity. All three African sequences are closely neighbored, as are the two Asians and five whites. Interestingly, although the whites form a well-

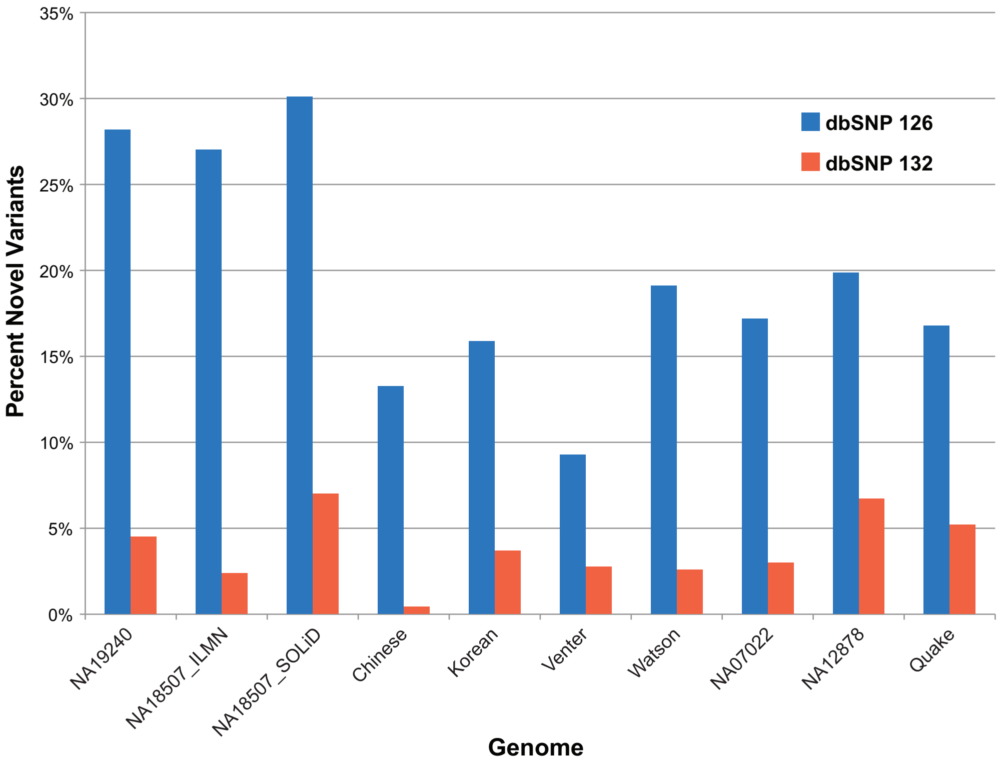

Fig. 4. Novel variant discovery. The *y* axis shows the percentage of each genome's variants (see Table 1 for total variant count) not present in dbSNP 126 (blue) and 132 (red). The lower percentages for dbSNP 132 reflect the inclusion of variants from personal genomes and 1000Genomes Project data. dbSNP 126 was released before the publication of any personal genomes.

supported clade for 19 of 22 autosomes, they are diverse, suggesting the possible existence of further substructure with this group.

Genomes 1 and 2 represent the same individual, sequenced on the Illumina and ABI SOLiD platforms, respectively. This fact makes possible some global comparisons of these two platforms as regards the accuracy of variant discovery and the impact of platform on potential diagnostic reliability. In total, 76.9% of variant locations of genomes 1 and 2 are shared in common relative to the union of both sets, a fact with implications for diagnostics. As can be seen in Figure 1, both genomes are each other's closest neighbor. By comparison, the next closest pair, the two Asian genomes (3 and 4) both sequenced on the same platform share only 46.5% of their locations in common. As we show later, however, the agreement is better at positions within genes and at genome positions associated with OMIM alleles.

OMIM analyses

We next examined the intersection of each genome's variants with known disease-causing/predisposing alleles as cataloged in OMIM.[16] OMIM indexes alleles relative to their positions in the DNA, messenger RNA, or protein sequence reported in the original publication, not relative to the current human reference genome assembly. This fact introduces many problems for analyses, because in many cases, the human reference sequence or genome annotation for the corresponding gene has changed considerably since the original publication, obscuring the relationship between the allele as reported in OMIM and its position on the current version of the reference human genome. Thus, we developed a procedure to systemically map the contents of OMIM forward to the reference human genome sequence (see "Materials and Methods"). In total, we are able to map forward almost 10,000 OMIM variants, including 88.7% of all amino acid changing alleles. See "Materials and Methods" section for additional details.

On average, each genome is heterozygous for 65 OMIM alleles and homozygous for 42 OMIM alleles (Table 1). Once again, ethnic trends are seen in these results. Calculating the ratio of homozygous/heterozygous alleles reveals that the African genomes contain a 1.6-fold excess of homozygous OMIM alleles relative to the Eurasian genomes. None of the genomes are homozygous for any severe disease-causing alleles nor are any heterozygous for alleles with serious reproductive consequences. Interestingly, in a substantial fraction of the cases, the personal genome sequence is homozygous for the healthy allele, with the reference genome containing a disease-causing allele (Table 1: "Healthy Variant").

We also examined the influence of ethnicity on shared OMIM alleles (Fig. 2). To do so, we once again constructed a neighbor-joining tree using the same procedure used to construct Figure 1. This time, however, restricting the data to those locations where either the personal genome or reference genome contained an OMIM allele. The resulting tree shows the same trends with respect to ethnicity as Figure 1; in other words, the global distribution of known disease-causing and predisposing variants within every genome in our dataset is influenced by the individual's ethnicity. Interestingly, the agreement between the Illumina and ABI SOLiD versions of genome NA18507 is much better for OMIM alleles. In total, 118 alleles were identified by the two platforms, 106 in common.

Genomic load in disease genes

We assigned 3626 human genes to 14 high-level disease categories based on peer-reviewed publications, wherein variants within the gene were shown to cause or predispose to a disease. Categories are not mutually exclusive. See "Materials and Methods" section for additional details. This disease-gene classification system allowed us to recover clinically meaningful variant sets, e.g., all variants in genes implicated in cardiovascular disease, e.g., the cardiovascular genomic load. This in turn made it possible (1) to measure the average genomic load of variants within genes specific to different disease categories; (2) to carry out cross-category comparisons; and (3) to systematically recover and analyze novel variants in disease genes. Figure, Supplemental Digital Content 1, http://links.lww.com/GIM/A160, plots the raw counts of all personal variants resulting in nonsynonymous changes to a gene's protein sequence by top-level category of the disease-gene classification system. The individuals in our dataset contained, for example, between 187 and 260 nonsynonymous variants within cardiovascular disease genes; and between 157 and 205 nonsynonymous variants within genes shown to play a role in oncogenesis and/or neoplasm progression. These raw counts, however, are not adjusted for the genomic footprint of the category. By genomic footprint, we mean the total number of coding nucleotides within all genes in a category, wherein changes can result in nonsynonymous changes to a gene's protein sequence. Figure 3 is corrected for this factor, showing the total genomic load of nonsynonymous variants corrected for genomic footprint and by each ethnic group. All genomes have about half as many variants in the genes classified as disease related as would be expected by chance, but these differences from expected are not consistent across classes. Figure 3 highlights these intercategory differences. As can be seen, an individual's genomic load varies by disease category. For example, every genome has approximately 45% of the expected number of disease-related nonsynonymous changes in genes involved in neonatal development but 75% of the expected number of nonsynonymous variants in genes involved in gastrointestinal disease. In general, trends are uniform within a category, e.g., if an excess of changes is observed in one genome, the same trend is observed in the remainder and vice versa. The aging category (generally genes involved in apoptosis, senescence, regeneration, and cell division), however, contains an exception to this trend, with genome 9 having almost 100% of the expected load.

DISCUSSION

Several recent studies have reported global statistics regarding the numbers and locations of SNVs[31–33] within both healthy and disease genomes. The work reported in this study extends these studies to the domain of clinically relevant trends in SNVs within healthy genomes, focusing on the impact of sequencing platform and ethnicity for genome-based prognosis. As Figure 1 makes clear, ethnicity has a powerful impact on the distributions of SNVs within personal genome sequences. It also provides a means to assess the global impact of sequencing technology on variant identification. Despite the six different sequencing platforms used to produce the dataset, in no case is the basic trend of ethnic similarity disrupted. For example, although five different platforms were used to produce the white genomes in our dataset, all five of these genomes form a clade. This indicates that accuracy of every platform is high enough to reveal ethnic relationships. Thus, differences in base-calling accuracy among the different platforms do not invalidate cross-platform genome comparisons for purposes of basic anthropological investigations. Moreover, we find that although the ABI SOLiD and Illumina versions of the NA18507 genome have 575,099 and 526,836 unique positions genome wide relative to each other sharing a total of 77% of their variants in common relative to the union of their sets that at locations corresponding to known disease-causing alleles, agreement is much better: 90% of variants are shared. These results make it clear that cross-platform analyses requiring great accuracy will remain problematic until sound probability models are available for base calling in every platform or at the very least until a standard set of equivalences is established between the quality values produced by the different sequencing platforms. It also has to be noted that these genomes include the first genomes to be published on all platforms, and most genomes were sequenced to relatively low coverage. Although much progress is being made in this area,[34,35] our results show that clinical prognoses cannot yet be made in a platform neutral fashion. However, because our knowledge about disease-causing mutations is heavily biased toward loci in coding regions and because these same regions tend to produce higher quality variant calls relative to the genome as a whole, the current technologies are clearly sufficient for a wide array of nondiagnostic cross-platform analysis.

We find that, similar to SNVs in general, disease-causing alleles are also distributed along ethnic lines, with Africans almost twice as likely to be homozygous for disease-causing or predisposing alleles as Eurasians. One likely explanation for this trend is background effects,[36,37] i.e., alleles with deleterious consequences in one ethnic background may well prove harmless in another. This unequal distribution of variants relative to ethnicity is likely compounded by an ascertainment bias in existing databases of variation. This bias exists due to the overrepresentation of Eurasian populations in current studies of disease. The approach taken herein of looking at the variation across broad classes of genes provides a key insight into our view of human genetic variation—the forces affecting mutational load seem to be largely constant across human populations. In contrast, we see a strong signal separating ethnicities when viewing these genomes in light of known, clinically relevant variation as cataloged in OMIM. These findings indicate that failure to adequately control for ethnicity will jeopardize the prognostic accuracy of sequence-based diagnoses; unfortunately, the impact of ethnicity on the penetrance and phenotypic severity of many disease alleles is still unknown. The need to extend studies in other clinical areas to include women and diverse ethnic populations is well established.[38] Our data demonstrate that similarly inclusive studies of personal genome sequences will be needed to assure equitable prognostic accuracy across ethnicities. Interestingly, given a larger set of personal genomes, the analysis techniques used herein would provide a means to identify and quantify background effects. Our results also suggest that further substructure in personal genome SNV distributions still awaits analysis. The structure of the white clade in Figure 1, for example, is suggestive of deeper population substructure within the whites.

In this study, we introduce the concept of personal genomic load by disease. Our disease-gene classification system has made it possible to measure the genomic load of variants within different disease categories and to carry out cross category comparisons. Genomic load varies between disease categories (Fig. 3). Furthermore, as genes are assigned to categories independent of their genomic location, these deviations are unlikely to be due to shared haplotypes acting to modulate variant numbers within a category in a coordinated fashion, especially in light of the diverse ethnicities in our dataset. One possible explanation for this observation then is a kind of ascertainment bias due to the fact that all the genomes in our dataset are from healthy adults. Simply surviving to adulthood may be correlated with restricted variation for some disease categories but less for others.

Still unknown, however, is the relationship between the magnitude of an individual's intercategory deviation from the average observed load within that category and the prognostic impact of that deviation. More genomes will be required to establish these baselines for personal prognosis. The answer to this question should be forthcoming, however, as additional genomes will allow significance thresholds to be applied to each category and for ethnicity. One genome in our dataset, however, does show a very large deviation. Genome 9 has a much increased genomic load within the aging category due to stop codons in the *CDC27* gene. On further inspection (data not shown), we believe that this is a false positive. Several *CDC27* pseudogenes exist, and it seems that for this genome, there may have been a systemic error during the read-alignment phase of the variant-calling procedure. That such events occur is itself an important point—not every gene or region of a genome will be equally accessible for prognosis by genome resequencing. These regions will need to be cataloged for accurate prognoses. Despite these facts, that this individual so stands out does demonstrate the utility of high-level summaries of genomic load made possible by an ontology-based approach, indicating that tools such as our disease-gene classification system will play a useful role in the future of whole-genome data management, analyses, and clinical prognosis and diagnostics.

ACKNOWLEDGMENTS

This work was supported by NIH SBIR Grant 1R4HG003667 (to M.G.R. and M.Y.) and NIH ARRA GO Grant 1RC2HG005619-01 (to M.Y. and M.G.R.).

REFERENCES

1. Levy S, Sutton G, Ng PC, et al. The diploid genome sequence of an individual human. *PLoS Biol* 2007;5:e254.
2. Wheeler DA, Srinivasan M, Egholm M, et al. The complete genome of an individual by massively parallel DNA sequencing. *Nature* 2008;452:872–876.
3. Bonetta L. Getting up close and personal with your genome. *Cell* 2008;133:753–756.
4. Wolinsky H. The thousand-dollar genome. Genetic brinkmanship or personalized medicine? *EMBO Rep* 2007;8:900–903.
5. Bentley DR, Balasubramanian S, Swerdlow HP, et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 2008;456:53–59.
6. McKernan KJ, Peckham HE, Costa GL, et al. Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Res* 2009;19:1527–1541.
7. Wang J, Wang W, Li R, et al. The diploid genome sequence of an Asian individual. *Nature* 2008;456:60–65.
8. Ahn SM, Kim TH, Lee S, et al. The first Korean genome sequence and analysis: full genome sequencing for a socio-ethnic group. *Genome Res* 2009;19:1622–1629.
9. Drmanac R, Sparks AB, Callow MJ, et al. Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* 2010;327:78–81.
10. Pushkarev D, Neff NF, Quake SR. Single-molecule sequencing of an individual human genome. *Nat Biotechnol* 2009;27:847–852.
11. De La Vega F, Hyland F, McLaughlin S, et al. Functional analysis of the genetic variation within the genomes of three HapMap individuals obtained by whole-genome, second-generation sequencing, 2009. Available at: http://www3.appliedbiosystems.com/cms/groups/mcb_marketing/documents/generaldocuments/cms_065553.pdf. Accessed February 1, 2011.
12. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci USA* 1977;74:5463–5467.
13. MeSH Browser, 2010. Available at: http://www.nlm.nih.gov/mesh/MBrowser.html. Accessed March 1, 2008.
14. Fauci AS, Braunwald E, Kasper DL, et al. Harrison's principles of internal medicine. New York: McGraw-Hill Medical, 2008.
15. Reese MG, Moore B, Batchelor C, et al. A standard variation file format for human genome sequences. *Genome Biol* 2010;11:R88.
16. Online Mendelian Inheritance in Man, OMIM (TM). McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD) and National Center for Biotechnology Information, National Library of Medicine (Bethesda, MD). Available at: http://www.ncbi.nlm.nih.gov/omim. Accessed March 1, 2008.
17. The 10Gen Data Set. Available at: http://www.sequenceontology.org/resources/10Gen.html. Accessed October 28, 2010.
18. Felsenstein J. An alternating least squares approach to inferring phylogenies from pairwise distances. *Syst Biol* 1997;46:101–111.
19. Efron B, Halloran E, Holmes S. Bootstrap confidence levels for phylogenetic trees. *Proc Natl Acad Sci USA* 1996;93:13429–13434.
20. Felsenstein J. Estimating effective population size from samples of sequences: a bootstrap Monte Carlo integration method. *Genet Res* 1992;60:209–220.
21. Yandell M, Mungall CJ, Smith C, et al. Large-scale trends in the evolution of gene structures within 11 animal genomes. *PLoS Comput Biol* 2006;2:e15.
22. Kent WJ, Sugnet CW, Furey TS, et al. The human genome browser at UCSC. *Genome Res* 2002;12:996–1006.
23. Pruitt KD, Tatusova T, Maglott DR. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* 2007;35(database issue):D61–D65.
24. Henikoff S, Henikoff J. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci USA* 1992;89:10915–10919.
25. Yandell M, Moore B, Salas F, et al. Genome-wide analysis of human disease alleles reveals that their locations are correlated in paralogous proteins. *PLoS Comput Biol* 2008;4:e1000218.
26. Eilbeck K, Moore B, Holt C, Yandell M. Quantitative measures for the management and comparison of annotated genomes. *BMC Bioinformatics* 2009;10:67.
27. Jimenez-Sanchez G, Childs B, Valle D. Human disease genes. *Nature* 2001;409:853–855.
28. Yandell MD, Majoros WH. Genomics and natural language processing. *Nat Rev Genet* 2002;3:601–610.
29. Cooper DN, Ball EV, Krawczak M. The human gene mutation database. *Nucleic Acids Res* 1998;26:285–287.
30. The OBO Flat File Format Specification, version 1.2. Available at: http://www.geneontology.org/GO.format.obo-1_2.shtml. Accessed February 1, 2011.
31. Pelak K, Shianna KV, Ge D, et al. The characterization of twenty sequenced human genomes. *PLoS Genet* 2010;6:e1001111.
32. Cirulli ET, Singh A, Shianna KV, et al. Screening the human exome: a comparison of whole genome and whole transcriptome sequencing. *Genome Biol* 2010;11:R57.
33. Harismendy O, Ng PC, Strausberg RL, et al. Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biol* 2009;10:R32.
34. Hoppman-Chaney N, Peterson LM, Klee EW, Middha S, Courteau LK, Ferber MJ. Evaluation of oligonucleotide sequence capture arrays and comparison of next-generation sequencing platforms for use in molecular diagnostics. *Clin Chem* 2010;56:1297–1306.
35. Mane SP, Modise T, Sobral BW. Analysis of high-throughput sequencing data. *Methods Mol Biol* 2011;678:1–11.
36. Linder CC. The influence of genetic background on spontaneous and genetically engineered mouse models of complex diseases. *Lab Anim (NY)* 2001;30:34–39.
37. Coleman DL, Hummel KP. The influence of genetic background on the expression of the obese (Ob) gene in the mouse. *Diabetologia* 1973;9:287–293.
38. Friedman MA. FDAMA—women and minorities guidance requirements under the Food and Drug Administration Modernization Act of 1997 (FDAMA) Sec. 115 Clinical Investigations. (b) Women and minorities—Section 505(b)(1) 21 U.S.C. 1998;355(b)(1): FDA, 1998.