# Natural genetic variation caused by small insertions and deletions in the human genome

Ryan E. Mills, W. Stephen Pittard, Julienne M. Mullaney, et al.

*Genome Res.* published online April 1, 2011
Access the most recent version at doi:10.1101/gr.115907.110

# Natural genetic variation caused by small insertions and deletions in the human genome.

Ryan E. Mills[1,2], W. Stephen Pittard[3], Julienne M. Mullaney[4,5], Umar Farooq[4], Todd H. Creasy[4], Anup A. Mahurkar[4], David M. Kemeza[4], Daniel S. Strassler[4], Chris P. Ponting[6], Caleb Webber[6], and Scott E. Devine[1,4,5,7,8]

1. Department of Biochemistry, Emory University School of Medicine Atlanta, GA 30322;
2. Department of Pathology, Brigham and Women's Hospital, Boston, MA 02115;
3. Bimcore, Emory University, Atlanta GA 30322;
4. Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore MD 21201;
5. Division of Endocrinology, Diabetes, and Nutrition, Department of Medicine, University of Maryland School of Medicine, Baltimore, MD 21201;
6. MRC Functional Genomics Unit, Department of Physiology, Anatomy and Genetics, University of Oxford, Oxford OX1 3QX, UK;
7. Winship Cancer Institute, Emory University, Atlanta GA 30322;
8. Greenebaum Cancer Center, University of Maryland School of Medicine, Baltimore, MD 21201

**Corresponding Author:**
Scott E. Devine, Ph.D.
Institute for Genome Sciences
University of Maryland School of Medicine
Room 615 BioPark II
Baltimore, MD 21201
Phone: (410) 706-2343
Email: sdevine@som.umaryland.edu

Running title: small INDELs in human genomes
Key words: insertion, deletion, INDEL, genetic variation

February 14, 2011

**ABSTRACT**

Human genetic variation is expected to play a central role in personalized medicine. Yet only a fraction of the natural genetic variation that is harbored by humans has been discovered to date. Here we report almost two million small insertions and deletions (INDELs) that range from 1 bp to 10,000 bp in length in the genomes of 79 diverse humans. These variants include 819,363 small INDELs that map to human genes. Small INDELs frequently were found in the coding exons of these genes and several lines of evidence indicate that such variation is a major determinant of human biological diversity. Microarray-based genotyping experiments revealed several interesting observations regarding the population genetics of small INDEL variation. For example, we found that many of our INDELs had high levels of linkage disequilibrium (LD) with both HapMap SNPs and with high-scoring SNPs from genome-wide association studies. Overall, our study indicates that small INDEL variation is likely to be a key factor underlying inherited traits and diseases in humans.

Supplemental material is included with this manuscript. The INDEL variants reported in this study have been deposited to dbSNP and the genotyping array data has been deposited to the Gene Expression Omnibus (both under the DEVINE_LAB handle).

## INTRODUCTION

The age of personalized genomics is well underway. Human genomes are being sequenced at unprecedented rates (Levy et al. 2007; Wheeler et al. 2008; Wang et al. 2008; Bentley et al. 2008; Ley et al. 2008; Kim et al. 2009; Ahn et al. 2009; McKernan et al. 2009; Schuster et al. 2010; 1000 Genomes Project Consortium 2010) and projects are underway to sequence the genomes of at least 1,000 additional humans (Hayden, 2008; 1000 Genomes Project Consortium). The long-range goal of these studies is to "crack the code" of natural genetic variation, i.e., to understand how changes in our genetic blueprints influence human traits. The prime motivation for these studies is to improve human health by understanding how genetic variation affects health, diseases, and medical treatments. Personal genome sequences will be used to predict the future health of individuals and to develop customized medical treatments that are optimized based on the genetic variation that is detected.

A critical first step in this process is to gain a comprehensive knowledge of the genetic variation that is harbored by human populations and to develop databases of informative variants that can be used to predict human health. Several types of natural genetic variation have been identified in humans, including single nucleotide polymorphisms (SNPs; The International SNP Map Working Group, 2001; The International HapMap Consortium, 2003; The International HapMap Consortium, 2005; 1000 Genomes Project Consortium 2010), small insertions and deletions (INDELs) ranging from 1 bp to 10 kb in length (Weber et al., 2002; Bhangle et al., 2005; Mills et al. 2006; Korbel et al., 2007; Kidd et al. 2008; 1000 Genomes Project Consortium 2010), and larger structural variants ranging from 10 kb to several megabases in length (Iafrate et al. 2004; Sebat et al., 2004; Tuzan et al. 2005; McConnell et al. 2006; Hinds et al. 2006; Conrad

et al. 2006; Kidd et al. 2008; Conrad et al. 2010a; Conrad et al. 2010b; 1000 Genomes Project Consortium 2010; Mills et al. 2011). Minisatellites, microsatellites, and transposon insertions also have been identified within these variant classes (Weber et al. 2002; Mills et al. 2006; Mills et al. 2007). Although SNPs and larger structural variants have received considerable attention, small INDELs remain largely under-discovered, and methods for studying these INDELs have lagged behind methods for analyzing other forms of variation. As a consequence, we know relatively little about the impact of small INDEL variation on human biology and diseases.

In this study, we identified almost two million small INDELs in the genomes of 79 diverse humans. More than 800,000 of these small INDELs mapped to human genes, including the coding exons and promoters of these genes. Several lines of evidence indicate that coding INDELs in particular are likely to affect gene function in humans. We also developed new microarray-based INDEL genotyping technologies to study the population genetics of our small INDELs in diverse humans. Overall, our study indicates that small INDEL variation is extensive in human genomes and is likely to have a major impact on human biology and diseases.

## RESULTS

### Small INDEL discovery using DNA sequencing data generated from 79 humans.

To gain a better understanding of small INDEL variation in human populations, we examined 98 million Applied Biosystems (ABI) DNA re-sequencing traces that had been deposited into the trace archive at the National Center for Biotechnology Information (NCBI) (Supplemental Table 1). Many of these traces were generated previously by genome centers for

SNP discovery projects or for BAC sequencing projects that were not used in published genome assemblies. Traces that were generated from targeted, PCR-based re-sequencing projects were excluded from our analysis. The final trace set includes DNA sequence information from 79 diverse humans, making it an ideal resource for variation discovery (Supplemental Table 2). We confirmed that these traces provide excellent coverage of the human genome (Supplemental Fig. 1).

By comparing these 98 million traces to build hg18 of the reference human genome sequence (The International Human Genome Sequencing Consortium, 2001), we identified 1.96 million non-redundant small INDELs from an initial discovery set of 3.4 million INDELs (Table 1, Supplemental Tables Chr 1 to Chr Y; See Methods). We previously established that our INDEL discovery pipeline has a validation rate of 97.2% (Mills et al. 2006; please see Methods). We also determined that 1.36 million of the 1.96 million INDELs in this study (69.3%) were discovered in more than one independent trace or could be confirmed in the chimpanzee (The Chimpanzee Sequencing and Analysis Consortium, 2005) or Celera (Venter et al. 2001) genomes (Table 1). Thus, additional validation was achieved through these comparisons.

Our INDELs were found on all 24 human chromosomes at an average spacing of one INDEL per 1,500 bp of DNA. They ranged from 1 bp to 10,000 bp in length and follow a size distribution where the majority of INDELs were <100 bp in length (Supplemental Fig. 2; Wheeler et al. 2008). Like known SNPs, which affect ~15 million base pairs of DNA (dbSNP build 129), our INDEL variants affected 11.9 million base pairs of the human genome. Thus, the

amount of genetic variation that is caused by small INDELs, in terms of base pairs, is considerable and approaches that caused by known SNPs.

**Structural variants (SV's) and transposon insertions.**

Our INDEL discovery range (of 1 bp to 10,000 bp) overlapped the discovery ranges that have been defined for structural variants (SV's), copy number variants (CNV's), and transposon insertions. For example, the 1000 Genomes Project recently defined SV's as variants that are greater than 50 bp in length (1000 Genomes Project Consortium, 2010; Mills et al. 2011), and other groups have defined SV's as variants that are greater than 1 kb (Iafrate et al. 2004). We identified 7,245 variants that are >50 bp among our collection of 1.96 million INDELs, 957 of which are also >1,000 bp (Table 1 and Supplemental Table 4). Thus, a relatively small fraction of our 1.96 million INDELs (0.4%) are sufficiently large to overlap these other variant classes. We compared these 7,245 variants: i) to SV's in The Database of Genomic Variants (Iafrate et al. 2004), ii) to SV's that were reported recently by Conrad et al. (Conrad et al. 2010a; Conrad et al. 2010b) and iii) to SV's reported by the 1,000 Genomes Project (1000 Genomes Project Consortium, 2010; Mills et al. 2011). We found that 3,582 of our 7,245 variants (49.4%) are novel compared to these other variants (Supplemental Table 4). Many of the remaining 3,663 variants provide breakpoint resolution for known SV's at the single nucleotide level for the first time. We also suspected that some of our INDELs might have been caused by transposon insertions (Supplemental Fig. 2). Indeed, 1,150 transposon insertions were identified in our INDEL collection, including Alu, L1, and SVA insertions (Table 1 and Supplemental Table 16). A total of 170/1,150 (14.8%) of these insertions were previously identified by our laboratory using transposon-seq methodologies (Iskow et al. 2010).

**Comparisons with INDELs in personal genomes and other human populations.**

We next wished to compare our 1.96 million variants to the small INDELs that have been discovered in personal genomes and human populations. We first compared our variants to the INDELs that that have been deposited to dbSNP and found that 37% (726,871/1.96 million) of our INDELs had been deposited previously. Thus, 63% of our INDELs are novel compared to the INDELs in dbSNP (build 129). We next examined five of the personal human genomes that have been sequenced, including four "healthy" genomes (Levy et al. 2007; Wheeler et al. 2008; Wang et al. 2008; Bentley et al. 2008) and the genome of a patient with acute myelogenous leukemia (Ley et al. 2008). Twenty-two percent (432,958) of our 1.96 million INDELs were present in one or more of these genomes (Fig. 1; Supplemental Table 3). Finally, we compared our 1.96 million INDELs to the 1.48 million INDELs that were recently reported by the 1000 Genomes Project (Fig. 1C; 1000 Genomes Consortium 2010). We determined that 463,377 of our 1.96 million INDELs (23.6%) were present in the 1000 Genomes Project dataset (Fig. 1C). The relatively small overlap between the 1.96 million INDELs that were discovered in this study and INDELs from these other studies (Fig. 1, Supplemental Table 3; dbSNP) suggests that INDEL discovery is likely to be incomplete in human populations.

**Small INDELs in human genes.**

Although INDELs could, in principle, affect a number of functional elements in the human genome, of particular interest are variants that alter human genes. A large number (819,363 or 42%) of the 1.96 million INDELs in this study mapped to known human genes, and 2,123 INDELs (0.1%) affected the coding exons of these genes (Fig. 2, Supplemental Fig. 3, and Supplemental Table 5). These "coding INDELs" were identified by comparing the co-ordinates

of our INDELs to those of annotated RefSeq and Ensembl genes (build hg18). An initial set of 1,715 coding INDELs was identified that affected 990 of the 20,705 RefSeq genes that were examined (equivalent to 4.8% of the RefSeq genes). An additional 408 coding INDELs were identified in 215 Ensembl genes that were not annotated in the RefSeq collection. Thus, a combined total of 1,300 exons and 1,205 genes were affected by coding INDELs in our collection (equivalent to ~5.8% of the genes in the human genome). These data indicate that apparently healthy humans harbor a substantial genetic load of coding INDELs.

A total of 184 of our 2,123 (8.7%) coding INDELs overlapped one or more exon boundaries, and these variants generally would be expected to abolish gene function (Supplemental Fig. 3, Supplemental Table 5). The remaining coding INDELs (1,939/2,123 or 91.3%) fell entirely within coding exons. More than half of these INDELs (1,037/1939 or 53.5%) would be expected to cause frame-shifts that would introduce premature termination codons (PTC's). Some of these PTC's are predicted to occur at positions that would target the encoded mRNA for nonsense mediated decay (NMD), and such variants generally would abolish gene function. However, PTC's that occur in the final 3' exon (or in the 3' 50 nucleotides of the penultimate exon) may produce mRNAs that escape NMD; these mRNAs could, in principle, encode novel proteins that have gain-of-function or dominant-negative effects. The remaining coding INDELs (902/1,939 or 46.5% of those within exons) produced INDELs that were multiples of three nucleotides, and thus, maintained the open reading frames (ORFs) of the original proteins (Supplemental Table 5).

Many (357/1,205 or 29.6%) genes were affected independently by two or more coding INDELs (Supplemental Fig. 4; Supplemental Table 5). For example, nine in-frame exon variants of the DSPP gene (involved in the development of the dentin portion of teeth; Mastranglo et al. 2007) were identified (Supplemental Table 5). It is tempting to speculate that such polymorphisms might play a role in dental health. Moreover, fourteen coding exon variants of the EP400 tumor suppressor gene (Fuchs et al. 2001) were identified (Supplemental Table 5). A total of 52 genes that previously have been implicated in human cancers harbored coding INDELs (Supplemental Table 5). At least four gene families that previously were known to be highly polymorphic in human populations (Hughes et al. 1998; Shimomura et al. 2002; Suzuki et al. 2002; Gilead and Lancet 2003) also carried a large number of coding INDELs. In particular, 46 exon variants were identified in 23 different keratin genes; 14 exon variants were identified in seven collagen genes; 40 exon variants were identified in 32 olfactory receptor genes; and 11 exon variants were identified in two HLA genes (Supplemental Fig. 4, Supplemental Table 5). Thus, coding INDELs contribute to the high levels of genetic variation that are characteristic of these gene families. In each case, these high levels of genetic variation are known to be maintained for a biological purpose (Hughes et al. 1998; Shimomura et al. 2002; Suzuki et al. 2002; Gilead and Lancet 2003). These examples, along with the many additional examples in Supplemental Table 5, illustrate the point that coding INDELs may potentially cause a great deal of human biological diversity.

**Coding INDELs that are likely to cause human phenotypes.**

We next set out to identify INDELs that are likely to cause phenotypic consequences in humans. Although some of our coding INDELs could, in principle, cause gain-of-function

phenotypes, our analysis was focused solely on human INDELs that would disrupt gene function. In order to identify INDELs that are likely to be disruptive in humans, we identified INDELs that (*i*) reside within coding exons and cause frame-shifts or overlap exon boundaries, (*ii*) affect more than just the last exon, and (*iii*) affect all Ensembl transcripts for the overlapped gene. Filtering on these three criteria resulted in 548 INDELs that affect 394 human genes (Supplemental Table 5). To gain insight into the likely phenotypic consequences of these disruptions, we identified 101 genes from this set that are associated with targeted disruption phenotypes in the mouse (Supplemental Table 6; Eppig et al., 2005; Eppig et al. 2007). Of these 101 experimentally-disrupted genes, 84 (83%) yield an abnormal phenotype upon homozygous disruption in the mouse (Supplemental Table 6).

We also identified five additional genes for which heterozygous disruptions are known to cause human diseases. As might be expected, the human disorders that are associated with these gene disruptions are non-lethal, often are late-onset, and are associated with relatively mild or variable phenotypes that might easily go undetected. The five genes, and their associated diseases, are *CEL* [maturity-onset diabetes of the young; OMIM: 609812; Raeder et al. 2006], *IRAK3* [early onset asthma; OMIM:611064; Baluci et al. 2007], *KRT14* [ectodermal dysplasia syndromes; OMIM:161000 and 125595; Lugassy et al. 2006], *MYO6* [progressive hearing loss; OMIM: 606346; Sanggard et al. 2007; Hilgert et al. 2008], and *RP1* [retinitis pigmentosa; OMIM:180100; Pierce et al. 1999; Guillonneau et al. 1999] (Supplemental Table 6). Three additional human genes with disruptive INDELs have mouse orthologs that are being used as human disease models in the hemizygous state. These are *CACNA1F* [congenital stationary night blindness; OMIM: 300071; Mandery et al. 2005], *NBN* [Nijmegen Breakage Syndrome;

OMIM: 251260; Dumon-Jones et al. 2003; Resnick et al. 2003; Tanzanella et al. 2003] and *TCOF1* [Treacher Collins syndrome; OMIM:154500; Dixon et al. 2000; Marzalek et al. 2003; Shoo et al. 2004] (Supplemental Table 6). These analyses indicate that at least some of our INDELs are likely to have phenotypic consequences in humans, even when present in just one copy.

**Evidence for strong purifying selection on coding INDELs.**

To investigate whether purifying selection might act upon coding INDELs (particularly those that are detrimental), we next examined the relative distributions of INDELs compared to SNPs in the human genome. Our computational pipeline simultaneously detects both INDELs and SNPs from ABI trace data, and has high levels of accuracy with both classes of variation (Tsui et al. 2003; Mills et al. 2006, please see Methods). We determined that the genome-wide ratio of INDELs to SNPs was 0.19 (i.e. 1 INDEL for every 5.3 SNPs; Supplemental Table 7). Both genes and intergenic regions had ratios that were similar to this genome-wide average (0.20 vs. 0.18, respectively; Fig. 2A). Although most gene features (promoters, introns, terminators) also had similar ratios, the ratio for coding exons was well below this genome-wide average (0.046; Fig. 2A). We confirmed that this difference was caused by a reduction of INDELs. This difference is most likely explained by the fact that INDELs create major changes in coding exons (including frameshifts and amino acid changes) whereas SNPs often produce synonomous changes that have little or no impact on gene function. Thus, strong purifying selection appears to eliminate INDELs that map to coding exons much more frequently than SNPs. Despite these reductions, coding INDELs were still fairly abundant in the genomes examined, indicating that coding INDELs are nevertheless likely to have a substantial impact on humans (see also below).

**Microarrays for genotyping small INDELs in humans.**

Although microarrays have been used widely to genotype SNPs (The International HapMap Consortium, 2003; The International HapMap Consortium, 2005) and relatively large structural variants (Sebat et al. 2004; McCarroll et al. 2006; Hinds et al. 2006; Conrad et al. 2006; Conrad et al. 2010a; Mills et al. 2011), microarray-based assays for interrogating small human INDELs have received little attention. This is particularly true for INDELs in the size range of 1 bp to 100 bp, which account for ~99% of the INDELs in our study. As a consequence, we lack robust tools to study our INDELs and the small INDELs that are being discovered in personal human genomes. Thus, we set out to develop new genotyping methods to interrogate small INDELs on microarrays. In particular, we developed assays to genotype INDELs that are 1 bp to 100 bp in length. We sampled several commercial platforms and found that the most successful approach was to adapt methods that originally were developed for the Affymetrix 5.0 and 6.0 SNP arrays (The International HapMap Consortium, 2005). Like these arrays, we used reduced representation and statistical probe clustering models to develop assays for INDEL genotyping (Fig. 3). However, in contrast to Affymetrix SNP arrays, our probe sets were designed to specifically interrogate novel junctions that define small INDELs using a series of overlapping 25-mer probes (Supplemental Fig. 5). The probe design protocols, library files, and Affymetrix Power Tools (APT) software packages were modified accordingly to accommodate the unique format of our custom INDEL data. Overall, successful probe sets with good BRLLM-P clustering characteristics were developed for 10,003 INDELs using a confidence cutoff of 0.05 (Fig. 3D; Supplemental Table 8; Methods). Trace frequency data indicate that these 10,003 INDELs were representative of the larger INDEL collection (Supplemental Fig. 6). As a measure of accuracy, we also performed validation studies on a representative set of INDELs using PCR methodologies. The vast majority (268/271 or 99%) of

the genotyping calls were identical for the two sets of measurements (Fig. 3D, Supplemental Table 9). Likewise, arrays that were hybridized with the same genomic DNA probes on separate days had >99% identical calls, indicating that our methods are highly reproducible (Supplemental Table 8).

**Population genetics of small INDELs.**

Overall, 8,836 of the 10,003 INDELs on our arrays (88.4%) showed allelic variation in the panel of 161 diverse humans that was examined (Fig. 3E; Supplemental Table 8). For an additional 212 (2.1%) INDELs, we detected solely the trace (B) allele. Since the trace allele is highly abundant and also differs from the reference genome, these INDELs also were confirmed. This is a high confirmation rate (90.5%), given that we did not probe all of the original humans that were used for INDEL discovery. In many of the remaining cases (649/955 or 68%), the trace allele was confirmed in multiple traces, in the chimp genome, or in the Celera genome (or combinations of these three). Likewise, our INDEL detection pipeline has a validation rate of 97.2% (Mills et al. 2006), further supporting the conclusion that most of the remaining variants are likely to be rare but authentic INDELs.

Our microarrays allowed us to measure the allelic frequencies of our INDELs in the population that was examined (Fig. 3E; Supplemental Tables 8). A total of 7,738 (77.4%) of the 10,003 INDELs had minor allelic frequencies that were ≥5%, and the remaining 2,265 (22.7%) had minor allelic frequencies of <5%. Thus, the majority of INDELs detected in our trace experiments meet the definition of common human genetic variation (Fig. 3E; Table 2;

Supplemental Table 8). We also examined the allelic frequencies of INDELs that mapped to RefSeq genes (Table 2, Supplemental Tables 8 and 10). Although only 22.7% of the 10,003 INDELs on the array were rare (MAF < 5%), a much larger percentage of the coding INDELs on the array were rare (72/111 or 64.9%), suggesting that purifying selection was acting on these alleles. In contrast to the data in Fig. 2A, these population frequency data indicate that even the coding INDELs that we can observe are enriched for deleterious variants compared to non-coding INDELs. A two-tailed Fisher's exact test revealed that this result was highly significant ($p = 5.1 \text{ X } 10^{-21}$; Table 2). We also examined the number of triplet (in-frame) vs. non-triplet (frame-shifting) coding INDELs, and observed an enrichment of rare alleles among the non-triplet (frame-shifting) group. These data indicate that frame-shifting INDELs are under the strongest negative selection ($p = 0.00036$), which is not surprising, given that this class is expected to be the most disruptive. These data provide evidence that even the coding INDELs that have not yet been eliminated by selection are enriched for deleterious alleles, and are likely to have an impact on humans.

We next examined INDEL inheritance patterns in the 30 Yoruban (YRI) trios that were examined in our study (containing a mother, father, and child). Ninety-nine percent of the INDELs were inherited in a Mendelian fashion in the YRI trios, indicating that most of the small INDELs in our study are polymorphisms rather than *de novo* mutations (Supplemental Table 8). The remaining 1% of the INDELs had non-Mendelian transmission patterns. This class of INDELs is likely to be enriched for genotyping errors. It should be noted that since the error could be in the parent or the child, the per sample error rate is less than 1%. We also determined whether the allelic distributions were in agreement with Hardy-Weinberg predictions in each of

the three subpopulations that were genotyped (i.e., the polymorphism discovery resource (PDR), YRI, and Han Chinese (CHB) populations; Supplemental Table 11). An average of 96.8% of the INDELs were in Hardy-Weinberg equilibrium in the three populations ($p = 0.01$). Although statistical and genotyping errors alone can account for the non-HWE distributions observed, it is also possible that some of these unusual distributions were caused by sample relatedness (Fig 3F) or natural selection.

We also compared the allelic frequencies of our INDELs in two diverse human populations (YRI and CHB; Fig. 3F-H). A total of 2,106 (21.1%) of the INDELs on our arrays had variation in one of the two populations but not in the other (Figs. 3G,H and Supplemental Table 12). Interestingly, most (1,748) of these population-specific variants were YRI-specific variants (in these cases, the YRI population had both INDEL alleles but the CHB population had only one of the two alleles, Fig. 3G). An additional 554 (5.5%%) of the INDELs had allelic frequency differences of >0.5 between the two populations.

**Linkage Disequilibrium with SNPs.**

We next examined the LD of our INDELs compared to the SNPs that have been genotyped in the YRI and CHB populations by the HapMap project (The International HapMap Consortium, 2005). A total of 5,218 INDELs had perfect LD ($r^2 = 1$) with at least one SNP in the CHB HapMap panel, and 6,084 INDELs had useful LD ($r^2 > 0.8$; Supplemental Table 13). Likewise, 5,674 INDELs had useful LD with SNPs that were genotyped in the YRI HapMap panel ($r^2 > 0.8$). Most of the INDELs that had variation on our arrays (8,836 or 88.4%) were in

LD with at least one SNP in the CHB and/or YRI populations (Fig. 4A). These experiments demonstrate that thousands of small INDELs can be efficiently genotyped and integrated into the human HapMap with this approach.

Finally, we determined whether high-scoring SNPs from genome-wide association studies likewise had high levels of LD with any of our INDELs. Such associations could establish connections between our INDELs and human traits (including diseases). Over 56,000 SNPs from 118 published genome-wide association studies recently were collected into a single convenient database (Johnson and O'Donnell, 2009; Supplemental Table 14). We used this database to determine whether any of the SNPs in these studies had high levels of pairwise LD with our INDELs. Indeed, this approach was highly successful, and 2,290 SNPs were identified that had high levels of LD ($r^2 > 0.8$) with our INDELs (Supplemental Table 15). In fact, 1,193 SNPs had perfect LD ($r^2 = 1$) with INDELs on our arrays (Fig. 4B). Almost half of the INDELs (1,102 or 48%) were located in genes (Fig. 4B), and many of these INDELs could be envisioned to affect gene function (Fig. 4C). Thus, in conjunction with GWAS studies, our INDEL genotyping platform can be used to identify INDEL candidates that may influence human traits and diseases.

## DISCUSSION

We provide strong evidence that small INDEL variation is not only abundant in humans, but is also likely to contribute to human phenotypic diversity. More than 800,000 of the small

INDELs in our study mapped to human genes, including 2,123 small INDELs that mapped to the coding exons of these genes (Fig. 2, Supplemental Tables Chr1 to Chry and Supplemental Table 5). Many of these coding INDELs would be expected to alter gene function (Fig. 2 and Supplemental Tables 5 and 6). We also identified more than 39,000 INDELs in the promoter regions of genes (Supplemental Tables Chr1 to Chry). Such INDELs could be envisioned to contribute to the allele-specific gene expression differences that have been observed in humans (Yan et al. 2002; Cheung et al. 2003; Cheung and Spielman, 2009). Taken together, these two classes of INDELs alone (coding and promoter INDELs) are likely to harbor many variants that affect human biology.

Several independent lines of evidence indicate that coding INDELs in particular are likely to be detrimental in humans. First, we found that ~75% of the INDELs that arise in coding exons appear to be eliminated by strong purifying selection (Fig. 2A). Moreover, our INDEL genotyping experiments further indicated that most of the remaining coding INDELs are also under selective pressure as revealed by an over-representation of rare alleles in this class (Table 2). Taken together, these data indicate that coding INDELs in general, and frame-shifting INDELs in particular, have negative effects on human gene function. A comparison of our most detrimental coding INDELs with targeted deletions of orthologous mouse genes provided additional support for this idea (Supplemental Tables 5 and 6). Finally, many of our INDELs (including coding INDELs) were in perfect LD with high-scoring SNPs that have been identified previously in GWAS studies (Fig. 4). Therefore, it is likely that coding INDELs and any associated effects are being detected in these studies. Taken together, these observations suggest

that coding INDELs (particularly those that cause frame-shifts) are likely to be responsible for a great deal of phenotypic diversity and diseases in humans.

Coding INDELs also have been identified in at least some of the personal genomes that have been sequenced. For example, several hundred coding INDELs have been identified in the Venter (Ng et al. 2008) and Watson (Wheeler et al. 2008) genomes. Although coding INDELs have not been reported in most of the other personal genomes that have been sequenced, it is not clear whether attempts were made to identify coding INDELs. Our data suggest that coding INDELs are likely to be present all human genomes (Fig. 2; Supplemental Tables 5, 10). In further support of this conclusion, exome re-sequencing projects and the 1000 Genomes Project also have identified coding INDELs (Ng et al. 2009; Ng et al. 2010; 1000 Genomes Project Consortium 2010).

**INDEL microarrays for genotyping.**

Salathia et al previously described an approach for genotyping INDELs in the size range of 25 bp to 7260 bp in *Arabidopsis thaliana* (Salthia et al 2007). These INDELs were genotyped using 70-mer oligonucleotide probes using Comparative Genome Hybridization (CGH). In contrast, our approach more closely resembles the methods that were developed for SNP genotyping on the Affymetrix 6.0 array (The International HapMap Consortium 2005). In particular, we derived statistical clustering models from population data to call the INDEL genotypes using BRLMM-P. However, in contrast to the Affymetrix 6.0 approach, we used a series of 25-mer probes that spanned the unique junctions of INDEL alleles (Supplemental Fig. 5). Thus, our approach is unique compared to both the CGH and the SNP-based approaches that

are outlined above. The smaller (25-mer) oligonucleotide probes that were employed with our Affymetrix platform can discriminate alternative INDEL alleles that differ by as few as 1 to 25 base pairs, and these probes performed well over the entire range that was tested in our study (1 bp to 100 bp). This size range is ideal for human INDELs, because >99% of human INDELs are smaller than 100 bp, and most human INDELs are smaller than 25 bp (Supplemental Fig 2). Although our initial array contained a relatively small number of INDEL assays (10,003), this approach is directly scalable to much larger Affymetrix array formats.

In the past, genotyping arrays have been extremely valuable for validating and genotyping SNPs that were initially discovered by DNA sequencing (The International HapMap Consortium 2005). Since INDEL array technologies have lagged behind SNP arrays, it has not been possible to perform similar follow-up studies on the small INDELs that have been discovered in personal genomes and human populations. Our approach now can be used to perform these studies. The several million INDELs that have been discovered in personal genomes and human populations now can be validated with our platform. This approach will be particularly useful for genotyping INDELs that have eluded detection with Illumina-based sequencing due to the complexities of mapping short reads that involve INDELs. Thus, even as the price of whole genome sequencing continues to drop, array-based technologies are likely to play a complementary role in validating and genotyping INDELs.

Our study argues that small INDELs should be fully discovered, genotyped, and integrated into the human HapMap. INDELs that map to coding exons, promoters, and other functionally important sites could be given the highest priority. Our custom INDEL array

technology now provides a means for accomplishing this goal. Larger array formats could be designed and manufactured to genotype the several million INDELs that have been discovered by us and other groups. Once integrated into the HapMap, these small INDELs could be readily detected through imputation, thus facilitating efforts to identify high-scoring markers and causative variants in GWAS studies. As efforts turn to routine whole genome re-sequencing, small INDELs could be imputed from the SNPs that are detected using this integrated map of phased variation. Small INDELs can be technically challenging to detect with Next-generation trace mapping and this would provide an alternative way to detect and annotate small INDELs. The largest and most interesting challenge ahead will be to use medical histories, treatment successes, and other phenotypes to identify specific INDELs that affect human biology, with the overall goal of developing a comprehensive framework of predictive health.

## METHODS

### INDEL discovery pipeline.

Our INDEL discovery pipeline has been described previously (Mills et al. 2006). In this earlier study, we conducted a series of PCR-based validation studies with variants that were detected by our pipeline (Mills et al. 2006). In particular, we examined 215 variants using PCR-based methodologies and validated 209/215 (97.2%) of the variants. Importantly, these validation studies were carried out with the same 24 humans that were used for INDEL discovery (24 humans from the polymorphism discovery resource; Collins et al. 1999). Thus, we knew with certainty that any variant identified in these traces should also be found by PCR in the 24 humans. We examined a range of INDEL sizes and sequences in this validation study,

including single bp INDELs, repeat expansions, transposon insertions, coding INDELs, and SNPs. We achieved an overall validation rate of 97.2% (false discovery = 2.8%).

In our previous study, we found that a single ABI Sanger trace was sufficient to accurately identify INDELs > 1 bp in length (from 2 bp to 10,000 bp; Mills et al. 2006). However, single base pair INDELs were less accurately called by a single trace, and we implemented a double hit rule to call 1 bp INDELs. This led to a similarly high validation rate (97.3%) for single bp INDELs (Mills et al. 2006). The number of times that a given INDEL was detected in traces is indicated in Supplemental Tables Chr1-Chry. We also used the chimp and Celera genomes to evaluate whether the trace alleles could be validated by at least one additional source (Table 1; Supplemental Tables Chr1-ChrY; Mills et al. 2006).

**Identification and analysis of INDELs from NCBI trace data.**

DNA traces were obtained from the trace archive at NCBI (http://www.ncbi.nlm.nih.gov/Traces/trace.cgi). Traces were processed and compared to the hg18 build of the human genome using quality scores to guide the analysis, as described previously (Mills et al. 2006). INDELs were defined as insertions or deletions in the 1 bp to 10,000 bp size range that could be detected by comparing ABI traces to the reference genome as outlined previously (Tsui et al., 2003; Bennett et al., 2004; Mills et al, 2006). All variants were entered into a PERL dbfile hash module to identify redundancies, and then into a MySQL database, which was used to store and analyze the data. Supplemental Tables Chr1 to ChrY contain all of our variants along with coordinates and other information. INDELs were mapped to RefSeq genes as follows. First, RefSeq gene tracks were obtained from the UCSC human

genome browser (http://www.genome.ucsc.edu). The coordinates of our variants were compared to the coordinates of all RefSeq genes to identify variants that overlapped these genes (and specific features within these genes). For coding INDELs, this process was repeated with the Ensembl gene track downloaded from the Ensembl site (http://www.ensembl.org) and a non-overlapping set was developed. Additional annotation was obtained from the RefSeq database at NCBI (http://www.ncbi.nlm.nih.gov/RefSeq) and OMIM (http://www.ncbi.nlm.nih.gov/Omim).

**Affymetrix INDEL microarrays.**

Custom PERL scripts were written to identify INDELs that would be suitable for probing on Affymetrix arrays using protocols that have been developed for SNPs (http://www.affymetrix.com). For example, to adapt our INDELs to the reduced representation probing that is used for SNPs, we identified INDELs that fell within three size intervals of *Sty* I/*Nsp* I restriction fragments: 1) 0.2-0.8 kb, 2) 0.8 – 1 kb, 3) 1 0- 1.2 kb. Group 1 was the preferred size, but the other size ranges also were included, if necessary. Probes were developed as outlined in Supplemental Fig. 5. Probes were compared to the RepeatMasker track of the human genome (build hg18) and were set aside if they overlapped known repeats. In some cases, probes were designed for INDELs involving repeat expansions and transposons and these probes were allowed to contain repeats. Such probes generally did not perform well and were filtered out at later stages (see below). Probes also were compared to the SNPs and INDELs present in build 128 of dbSNP and probes that overlapped SNPs or other INDELs were not used. 1,500 SNP probe sets from the Affymetrix 6.0 SNP array were included on our arrays as positive controls and for quality control analysis. Arrays were hybridized using the Affymetrix 6.0 array

kits and protocols (www.affymetrix.com). The Affymetrix Power Tools (APT) software package (http://www.affymetrix.com/support/developer/powertools/changelog/index.html) was modified to handle INDEL cdf files and the multiple probe designs that are required for INDELs. CEL files were normalized and the quality was assessed using the APT program apt-geno-qc. High quality arrays (cqc >0.4) were analyzed using the APT program apt-probeset-genotype using BRLLM-P at a cutoff of 0.05.

The select probes feature in the Affymetrix APT software was used to identify the best probes from all of the initial probes that were included on the array. For each INDEL, between 12 and 24 probes (depending on the INDEL type) were included on the array to discriminate between the two INDEL states (Supplemental Fig. 5). The best-performing probes were identified using a combined cutoff of AIC value <325, and FLDAB >3. Both the AIC and FLDAB parameters are derived from the clustering data and provide measures of an assays' ability to discriminate the A and B states. The probes that were selected at these cutoffs were able to discriminate the two INDEL alleles very well, and had excellent clustering properties in BRLLM-P as well as high validation rates. All other probes were filtered from further consideration. Assays that had more than 60 no calls also were removed. This approach led to a final set of 10,003 assays that were used for all remaining studies.

## ACKNOWLEDGEMENTS

We thank Shari Corin and Paul Doetsch for critical reading of the manuscript. We thank Brian Cotton and Victor Felix for computational help and Stephanie Steiner for technical assistance with PCR experiments. We thank Andy Neuwald for advice on statistical analysis. We thank Brandon Young, Kurt Donner, Christofer Bertani, Eric Schell, Ali Pirani, and the Affymetrix Custom Array Program for technical assistance. We thank Jim Nemesh and Bob Handsaker for their help with assessing the LD properties of our genotyped INDEL set. We thank the centers that provided small INDEL data from personal genome sequences. This work was funded by grants from SUN Microsystems and the National Human Genome Research Institute, National Institutes of Health (F32HG004207 and R01HG002898).

## REFERENCES

1000 Genomes Project Consortium. 2010. A map of human genome variation from population-scale sequencing. *Nature* **467:** 1061-73.

Ahn, S.M., Kim, T. H., Lee, S., Kim, D., Ghang, H., Kim, D.S., Kim, B.C., Kim, S.Y., Kim, W.Y., Kim, C., *et al*. 2009. The first Korean genome sequence and analysis: Full genome sequencing for a socio-ethnic group. *Genome Res.* **19**: 1622-1629.

Balaci, L., Spada, M.C., Olla, N., Sole, G., Loddo, L., Anedda, F., Naitza, S., Zuncheddu, M.A., Maschio, A., Altea, D. *et al.* 2007. IRAK-M is involved in the pathogenesis of early-onset persistent asthma. *Am. J. Hum. Genet.* **80**: 1103-1114.

Bentley, D.R., Balasubramanian, S., Swerdlow, H.P., Smith, G.P., Milton, J., Brown, C.G., Hall,

K.P., Evers, D.J., Barnes, C.L., Bignell, H.R. *et al*. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**: 53-59.

Bhangale, T.R., Rieder, M. J., Livingston, R. J., and Nickerson, D.A. 2005. Comprehensive identification and characterization of diallelic insertion-deletion polymorphisms in 330 human candidate genes. *Hum. Mol. Genet.* **14**: 59–69.

Cheung, V.G., Conlin, L.K., Weber, T.M., Arcaro, M., Jen, K.Y., Morley, M., and Spielman, R.S. 2003. Natural variation in human gene expression assessed in lymphoblastoid cells. *Nat. Genet.* **33**: 422-425.

Cheung, V.G., and Spielman, R.S. 2009. Genetics of human gene expression: mapping DNA variants that influence gene expression. *Nat. Rev. Genet.* **10**: 595-604.

Collins, F.S., Brooks, L.D., and Chakravarti, A. 1999. A DNA polymorphism discovery resource for research on human genetic variation. *Genome Res.* **8**: 1229-1231.

Conrad, D.F., Andrews, T.D., Carter, N.P., Hurles, M.E., and Pritchard, J.K. 2006. A high-resolution survey of deletion polymorphism in the human genome. *Nat. Genet.* **38**: 75-81.

Conrad, D.F., Pinto, D., Redon, R., Feuk, L., Gokcumen, O., Zhang, Y., Aerts, J., Andrews, T.D., Barnes, C., Campbell, P., et al. 2010a. Origins and functional impact of copy number variation in the human genome. *Nature* **464**: 704-712.

Conrad, D.F., Bird, C., Blackburne, B., Lindsay, S., Mamanova, L., Lee, C., Turner, D.J., and Hurles, M.E. 2010b. Mutation spectrum revealed by breakpoint sequencing of human germline CNVs. *Nature Genet*. **42**: 385-391.

Dixon, J., Brakebusch, C., Fassler, R., and Dixon, M.J. 2000. Increased levels of apoptosis in the prefusion neural folds underlie the craniofacial disorder, Treacher Collins syndrome. *Hum. Mol. Genet.* **9**: 1473-1480.

Dumon-Jones, V., Frappart, P.O., Tong, W.M., Sajithlal, G., Hulla, W., Schmid, G., Herceg, Z., Digweed, M., and Wang, Z.Q. 2003. Nbn heterozygosity renders mice susceptible to tumor formation and ionizing radiation-induced tumorigenesis. *Cancer Res.* **63**: 7263-7269.

Eppig, J. T., Blake, J.A., Bult, C.J., Richardson, J.E., Kadin, J.A., Ringwald, M. 2007. Mouse genome informatics (MGI) resources for pathology and toxicology. *Toxicol Pathol*. **35**: 456-457.

Eppig, J. T., Bult, C.J., Kadin, J.A., Richardson, J.E., Blake, J.A., Anagnostopoulos, A., Baldarelli, R.M., Baya, M., Beal, J.S., Bello, S.M., *et al.* 2005. The Mouse Genome Database (MGD): from genes to mice--a community resource for mouse biology. *Nucleic Acids Res.* **33**: D471-5.

Fuchs, M., Gerber, J., Drapkin, R., Sif, S., Ikura, T., Ogryzko, V., Lane, W.S., Nakatani, Y., and Livingston, D.M. 2001. The p400 complex is an essential E1A transformation target. *Cell* **106**: 297-307.

Gilad, Y. and Lancet, D. 2003. Population differences in the human functional olfactory repertoire. *Mol. Biol. Evol.* **20**: 307-314.

Guillonneau, X., Piriev, N.I., Danciger, M., Kozak, C.A., Cideciyan, A.V., Jacobson, S.G., and Farber, D.B. 1999. A nonsense mutation in a novel gene is associated with retinitis pigmentosa in a family linked to the RP1 locus. *Hum. Mol. Genet.* **8**: 1541-1546.

Hayden, E. C. 2008. International genome project launched. *Nature* **451**: 378-379.

Hilgert, N., Topsakal, V., van Dinther, J., Offeciers, E., Van de Heyning, P., and Van Camp, G., 2008. A splice-site mutation and overexpression of MYO6 cause a similar phenotype in two families with autosomal dominant hearing loss. *Eur. J. Hum. Genet.* **16**: 593-602.

Hinds, D.A., Kloek, A.P., Jen, M., Chen, X. and Frazer, K.A. 2006. Common deletions and SNPs are in linkage disequilibrium in the human genome. *Nat. Genet.* **38**: 82–85.

Hughes, A. L. and Yeager, M. 1998. Natural selection at major histocompatibility complex loci of vertebrates. *Annu. Rev. Genet.* **32**: 415-435.

Iafrate, A.J., Feuk, L., Rivera, M.N., Listewnik, M.L., Donohue, P.K., Qi, Y., Scherer, S.W., and Lee, C. 2004. Detection of large-scale variation in the human genome. *Nat. Genet.* **36**: 949-951.

International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. *Nature* **409**: 860-921.

Iskow, R.C., McCabe, M.T., Mills, R.E., Torene, S., Pittard, W.S., Neuwald, A.F., Van Meir, E.G, Vertino, P.W., and Devine, S.E. 2010. Natural mutagenesis of human genomes by endogenous retrotransposons. *Cell* **141**: 1253-1261.

Johnson, A.D., and O'Donnell, C.J. 2009. An open access database of genome-wide association studies. *BMC Med. Genet.* **10**:6.

Kidd, J.M., Cooper, G.M., Donahue, W.F., Hayden, H.S., Sampas, N., Graves, T., Hansen, N., Teague, B., Alkan, C., Anonacci, F., et al. 2008. Mapping and sequencing of structural variation from eight human genomes. *Nature* **453**: 56-64.

Kim, J.I., Ju, Y. S., Park, H., Kim, S., Lee, S., Yi, J.H., Mudge, J., Miller, N.A., Hong, D., Bell, C.J. *et al*. 2009. A highly annotated whole-genome sequence of a Korean individual. *Nature* **460**: 1011-1016.

Korbel, J. O., Urban, A.E., Affourtit, J.P., Godwin, B., Grubert, F., Simons, J.F., Kim, P.M., Palejev, D., Carriero, N.J., Du, L., *et al*. 2007. Paired-end mapping reveals extensive structural variation in the human genome. *Science* **318**: 420-426.

Levy, S., Sutton, G., Ng., P.C., Feuk, L., Halpern, A.L., Walenz, B.P., Axelrod, N., Huang, J., Kirkness, E.F., Denisov, G., *et al.* 2007. The diploid genome sequence of an individual human. *PLoS Biol*. **5**: e254.

Ley, T.J., Mardis, E.R., Ding, L., Fulton, B., McLellan, M.D., Chen, K., Dooling, D., Dunford-Shore, B.H., McGrath, S., Hickenbotham, M., *et al.* 2008. DNA sequencing of a cyogenetically normal acute myeloid leukaemia genome. *Nature* **456**: 66-72.

Lugassy, J., Itin, P., Ishida-Yamamoto, A., Holland, K., Huson, S., Geiger, D., Hennies, H.C., Indelman, M., Bercovich, D., Uitto, J., *et al*. 2006. Naegeli-Franceschetti-Jadassohn syndrome and dermatopathia pigmentosa reicularis: two allelic ectodermal dysplasias caused by dominant mutations in DRT14. *Am J. Hum. Genet.* **79**: 724-730.

Mansergh, F., Orton, N.C., Vessey, J.P., Lalonde, M.R., Stell, W.K., Tremblay, F., Barnes, S., Rancourt, D.E., and Bech-Hansen, N.T. 2005. Mutation of the calcium channel gene Cacna1f disrupts calcium signalling, synaptic transmission and cellular organization in mouse retina. *Hum. Mol. Genet.* **14**: 3035-3046.

Marzalek, B., Wisniewski, S.A., Wojcicki, P., Kobus, K., and Trzeciak, W.H. 2003. Novel mutation in the 5' splice site of exon 4 of the TCOF1 gene in the patent with Treacher Collins syndrome. *Am. J. Med. Genet.* **123A**: 169-171.

Mastrangelo, F., Scioletti, A.P., Tranasi, M., Tecco, S., Sberna, M.T., Rinci, R., Grilli, A., Stuppia, L., Gherlone, E., Tete, S., *et al.* 2007. Dentin sialophosphoprotein expression during human matrix development. *J. Biol. Regul. Homeost. Agents* **21**: 33-39.

McCarroll, S.A., Hadnott, T.N., Perry, G.H., Sabeti, P.C., Zody, M.C., Barrett, J.C., Dallaire, S., Gabriel, S.B., Lee, C., Daly, M.J., *et al*. 2006. Common deletion polymorphisms in the human genome. *Nat. Genet.* **38**: 86-92.

McKernan, K., J., Peckham, H.E., Costa, G.L., McLaughlin, S.F., Fu, Y., Tsung, E.F., Clouser, C.R., Duncan, C., Ichikawa, J.K., Lee, C.C., *et al.* 2009. Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using tow base encoding. *Genome Res.* **19**: 1527-1541.

Mills, R.E., Luttig, C.T., Larkins, C.E., Beauchamp, A., Tsui, C., Pittard, W.S., and Devine, S.E. 2006. An initial map of insertion and deletion (INDEL) variation in the human genome. *Genome Res.* **16**: 1182-1190.

Mills, R.E., Bennett, E.A., Iskow, R.C., and Devine, S.E. 2007. Which transposable elements are active in the human genome? *Trends Genet.* **23**: 183-191.

Mills, R.E., Walter, K., Stewart, C., Handsaker, R.E., Chen, K., Alkan, C., Abyzov, A., Yoon, S.C., Ye, K., Cheetham, R.K., et al. (2011) Mapping copy number variation by population-scale genome sequencing. *Nature* **470**: 59-65.

Ng, P.C., Levy, S., Huang, J., Stockwell, T.B., Walenz, B.P., Li, K., Axelrod, N., Busam, D.A., Strausberg, R.L., and Venter, J.C. 2008. Genetic variation in an individual human exome. *PLoS Genet*. **4**: e10000160.

Ng, S.B., Turner, E.H., Robertson, P.D., Flygare, S.D., Bigham, A.W., Lee, C., Shaffer, T., Wong, M., Bhattacharjee, A., Eichler, E.E., et al. 2009. Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* **461**: 272-276.

Ng, S.B., Bigham, A.W., Buckingham, K.J., Hannibal, M.C., McMillin, M.J., Gildersleeve, H.I., Beck, A.E., Tabor, H.K., Cooper, G.M., Mefford, H.C. et al. 2010. Exome sequencing identifies *MLL2* mutations as a cause of Kabuki syndrome. *Nature Genet.* **42**: 790-793.

Pierce, E.A., Quinn, T., Meehan, T., McGee, T.L., Bersen, E.L., and Dryja, T.P. 1999. Mutations in a gene encoding a new oxygen-regulated photoreceptor protein cause dominant retinitis pigmentosa. *Nat. Genet.* **22**: 248-254.

Raeder, H., Johansson, S., Holm, P.I., Haldorsen, I.S., Mas, E., Sbarra, V., Nermoen, I., Eide, S.A., Grevle, L., Bjorkhaug, L., *et al*. 2006. Mutations in the CEL VNTR cause a syndrome of diabetes and pancreatic exocrine dysfunction. *Nat. Genet.* **38**: 54-62.

Resnick, I.B., Kondratenko, I., Pashanov, E. Maschan, A.A., Karachunsky, A., Togoev, O., Timakov, A., Polyakov, A., Tverskaya, S., Evgrafov, O., and Roumiantsev, A.G. 2003. 657del5 mutation in the gene for Nijmegen breakage syndrome (NBS1) in a cohort of Russian children with lymphoid tissue malignancies and controls. *Am. J. Med. Genet.* **120**: 174-179.

Salathia, N., Lee, H.N., Sangster, T.A., Morneau, K., Landry, C.R., Schellenberg, K., Behere, S., Gunderson, K.L., Cavalieri, D., Jander, G., and Queitsch, C. 2007. Indel arrays: an affordable alternative for genotyping. *The Plant Journal* **51**: 727-737.

Sanggaard, K.M., Kjaer, K.W., Eiberg, H., Nurnberg, G. Nurnberg, P., Hoffman, K., Jensen, H., Sorum, C., Rendtorff, N.D., and Tranebjaerg, L. 2008. A novel nonsense mutation in MYO6 is associated with progressive nonsyndromic hearing loss in a Danish DFNA22 family. *Am. J. Med. Genet.* **46**: 1017-1025.

Schuster, S.C., Miller, W., Ratan, A., Tomsho, L.P., Giardine, B., Kasson, L.R., Harris, R.S., Peterson, D.C., Zhao, F., Qi, J. Alkan, C. *et al*. 2010. Complete Khoisan and Bantu genomes from southern Africa. *Nature* **463**: 943-947.

Sebat, J., Lakshmi, B., Troge, J., Alexander, J., Young, J., Lundin, P., Maner, S., Massa, H., Walker, M., Chi, M., *et al*. 2004. Large-scale copy number polymorphism in the human genome. *Science* **305**: 525–528.

Shimomura, Y., Aoki, N., Schweizer, J., Langbein, L., Rogers, M.A., Winter, H., and Ito, M. 2002. Polymorphisms in the human high sulfer hair keratin-associated protein 1, KAP1, gene family. *J. Biol. Chem.* **277**: 45493-45501.

Shoo, B.A., McPherson, E., and Jabs, E.W. 2004. Mosaicism of a TCOF1 mutation in an individual clinically unaffected with Treacher Collins syndrome. *Am. J. Med. Genet.* **126A**: 84-88.

Sullivan, L.S., Heckenlively, J.R., Bowne, S.J., Zuo, J., Hide, W.A., Gal, A., Denton, M., Inglehearn, C.F., Blanton, S.H., and Daiger, S.P. 1999. Mutations in a novel retina-specific gene cause autosomal dominant retinitis pigmentosa. *Nat. Genet.* **22**: 255-259.

Suzuki, O. T., Sertie, A.L., DerKaloustian, V.M., Kok, F., Carpenter, M., Murray, J., Czeizel, A.E., Kliemann, S.E., Rosemberg, S., Monteiro, M., *et al.* 2002. Molecular analysis of collagen XVIII reveals novel mutations, presence of a third isoform, and possible genetic heterogeneity in Knobloch syndrome. *Am. J. Hum. Genet.* **71**: 1320-1329.

Tanzanella, C., Antoccia, A., Spadoni, E., diMasi, A., Pecile, V., Demori, E., Varon, R., Barseglia, G.L., Tiepolo, L., and Maraschio, P. 2003. Chromosome instability and nibrin protein variants in NBS heterozygotes. *Eur. J. Hum. Genet.* **11**: 297-303.

Taylor, M.S., Ponting, C.P., and Copley, R.R. 2004. Occurrence and consequences of coding sequence insertions and deletions in Mammalian genomes. *Genome Res.* **14**: 555-566.

The Chimpanzee Sequencing and Analysis Consortium. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437**: 69-87.

The International HapMap Consortium. 2003. The International HapMap Project. *Nature* **426**: 789-796.

The International HapMap Consortium. 2005. A haplotype map of the human genome. *Nature* **437**: 1299-1320.

The International SNP Map Working Group. 2001. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**: 928-933.

Tsui, C., Coleman, L.E., Griffith, J.L., Bennett, E.A., Goodson, S.G., Scott, J.D., Pittard, W.S, and Devine, S.E. 2003. Single nucleotide polymorphisms (SNPs) that map to gaps in the human SNP map. *Nucleic Acids Res.* **31**: 4910-4916.

Tuzun, E., Sharp, A.J., Bailey, J.A., Kaul, R., Morrison, V.A., Pertz, L.M., Haugen, E., Hayden, H., Albertson, D., Pinkel, D., *et al*. 2005. Fine-scale structural variation of the human genome. *Nat. Genet*. **27**: 727-732.

Venter, J. C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Holt, R.A., Gocayne, J.D., Amanatides, P. *et al.* 2001. The sequence of the human genome. *Science* **291**: 1304-1351.

Wang, J., Wang, W., Li, R., Li, Y., Tian, G., Goodman, L., Fan, W., Zhang, J., Li, J., Zhang, J., *et al*. 2008. The diploid genome sequence of an Asian individual. *Nature* **456**: 60-65.

Weber, J. L., David, D., Heil, J., Fan, Y., Zhao, C., and Marth, G. 2002. Human diallelic insertion/deletion polymorphisms. *Am. J. Hum. Genet.* **71**: 854–862.

Wheeler, D.A., Srinivasan, M., Egholm, M., Shen, Y., Chen, L., McGuire, A., He, W., Chen, Y., Makhijani, V., Roth, G.T., *et al.* 2008. The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452**: 872-876 .

Yan, H., Yuan, W., Velculescu, V.E., Vogelstein, B. and Kinzler, K.W. 2002. Allelic variation in human gene expression. *Science* **297**: 1143.

## FIGURE LEGENDS

**Figure 1. Comparisons of our data with small INDELs identified from other projects.**

**A-B**. Diagrams comparing the 1.96 million INDELs discovered in this study with the small INDELs that were identified in four personal genomes. **A**. Comparison of our 1.96 million INDELs (light blue, top) with Venter (Levy et al. 2007), and Watson (Wheeler et al. 2008) INDELs, **B**. Comparison of our 1.96 million INDELs (light blue, top) with Han Chinese (Wang et al. 2008), and Yoruban (Bentley et al. 2008) INDELs, **C**. Comparison of our 1.96 million INDELs (light blue, top) with the 1.48 million INDELs identified by the 1000 Genomes Project (1000GP; 1000 Genomes Project Consortium 2010).

**Figure 2. Distribution of coding exon variants in the human genome.**

**A.** The figure depicts a typical RefSeq gene and its features. 819,363 small INDELs from our study mapped to RefSeq genes. The INDEL to SNP ratios for each genomic compartment are indicated. **B.** The 1,205 genes that were affected by 2,123 coding exon variants (Supplemental Table 5) are indicated on the map of human chromosomes (colored marks to the left of the chromosomes indicate an affected gene). Each mark is indexed by color

to indicate gene function (and is cross referenced to the pie chart below). A red mark to the right of each chromosome indicates that the affected gene previously was linked to a known disease. The pie chart shows the functional breakdown of the coding variants.

**Figure 3. Affymetrix INDEL genotyping arrays.**

**A-C.** A region of a custom Affymetrix INDEL microarray is shown following hybridization and scanning using protocols established for the Affymetrix 6.0 array. Section **C** contains 1,500 Affymetrix SNPs that were developed for the HapMap project and are also present on the SNP 6.0 array. These were included as positive controls. The average cqc for our arrays after excluding arrays with scores below 0.4 was 2.2, with a range of 0.53 – 3.67. The call rate was 96.2%. Section **B** contains a manufacturing control. Section **A** represents the remainder of the array, which contains INDEL probes. **D.** Plot of signal intensities for a typical set of INDEL probes following BRLLM-P analysis. Note that three distinct clusters were obtained for the three INDEL states (AA, AB, BB). PCR validation studies were conducted in parallel to evaluate the accuracy of the calls (Supplemental Table 9). A typical result is shown for INDEL 210917. The 24 individuals from the polymorphism discovery resource (PDR; Collins et al. 1999) that were sampled by PCR are shown in red (the calls were 100% concordant between the arrays and the PCRs). The overall validation rate with 12 representative INDEL assays in 24 individuals was 99% (Supplemental Table 9). **E.** Allelic frequencies. The allelic frequencies are plotted for the 10,003 INDELs that were examined on the INDEL microarrays (Supplemental Table 8). Although the majority of variation meets the definition of common genetic variation (where the minor allele has a frequency of ≥5%), rare INDELs also were identified. **F.** Structure Plot of INDEL data. The INDEL genotypes from our arrays were

analyzed for population substructure. The PDR panel, which was designed to capture global diversity, has a large degree of substructure (as indicated by the colored peaks—right). The Yoruban (YRI) and CHB populations also have some residual substructure. **G.** Population-specific INDEL variation. INDELs were identified where both INDEL alleles were present in one population but only one allele was present in the other. An example of a YRI-specific INDEL is shown. Note that both A and B alleles are present in the YRI population (and all three genotypes are present), whereas only the B allele (and one genotype) is present in the CHB population. The INDEL shown (1384822) is a 3 bp coding INDEL. **H.** An example of a CHB-specific INDEL.

**Figure 4. Linkage disequilibrium between SNPs and INDELs.**

**A.** The $r^2$ value was calculated for each SNP within a 1Mbp window of a given INDEL using the SNP genotypes that have been reported for HapMap3 (http://hapmap.ncbi.nlm.nih.gov/) and our INDEL genotyping data from the same samples. For each population (YRI, CHB), the SNP with the maximum $r^2$ value was identified (Supplemental Table 13). INDELs in perfect LD with a SNP have an $r^2$ of 1.0. **B.** LD was also examined for high scoring SNPs (with p values < 0.001) that were identified in 118 GWAS studies (Supplemental Table 14; Johnson and O'Donnell, 2009). GWAS SNPs that have high levels of LD ($r^2 > 0.8$) with INDELs are summarized. **C.** Examples of INDELs from "**B**" that map to functional regions of genes. The 16 examples were taken from a larger collection of 1,102 INDELs that have high levels of LD with GWAS SNPs and also map to genes (Supplemental Table 15).

**Table 1. Summary of small INDELs identified from 98 million re-sequencing traces**.

98 million human traces were obtained from the NCBI trace archive and were compared to the reference human genome to identify potential variants. All of the data was assembled into a MySQL database and deposited into dbSNP (Supplemental Tables Chr1 – ChrY). INDEL-positive traces were compared to a) other INDEL-positive traces, b) the chimp genome (The Chimpanzee Sequencing and Analysis Consortium, 2005), and c) the Celera genome (Venter et al. 2001) to determine whether the allele was identified in other genomes and could be classified as a "double hit" allele. INDEL-positive traces from independent projects and centers were tracked to provide independent confirmation (Supplemental Tables Chr1 – ChrY). INDELs were mapped to the promoters, 5' and 3' UTRs, splice sites, and introns of genes. This and other annotation was tracked in a MySQL database (and is listed in Supplemental Tables Chr1–ChrY). INDELs that were caused by Alu, L1, and SVA retrotransposon insertions are listed in Supplemental Table 16.

**Table 2. Analysis of rare INDEL variants in RefSeq genes.**

For each class of INDELs examined, the number of rare INDELs (MAF <5%) vs. the total number of INDELs in the class is listed. Two-tailed Fisher's exact tests were performed between the overall set of 10,003 INDELs that were genotyped on the array and each of the other classes that are listed. The data indicate that rare INDELs are over-represented in coding exons compared to the 10,003 INDELs. None of the remaining classes had such enrichment, suggesting that coding exons are the most sensitive to INDELs.

## Supplemental Figures and Tables

**Supplemental Figure 1. Trace coverage of the human genome.**

The 98 million human traces that were obtained from the NCBI trace archive were mapped to the human genome using our pipeline and then placed into 200 Kb bins to examine the genomic coverage. The traces provide fairly uniform coverage of the sequenced portions of the human genome. Areas that lack traces mostly correspond to regions of the reference genome that were not sequenced. In order to manage possible ascertainment biases, we implemented a maximum cutoff of 3500 traces per bin. The RefSeq gene distribution is depicted below the trace map.

**Supplemental Figure 2. Size vs. frequency distribution of our INDELs.**

The figure displays a size vs. frequency plot of our INDELs and indicates that smaller INDELs (<100 bp) are the most abundant in our data set. A peak is shown (arrow) that corresponds to Alu insertion polymorphisms.

**Supplemental Figure 3. Three types of exon alterations caused by coding INDELs.**

The top panel depicts deletions that include one or more exons. The middle panel depicts "boundary deletions," in which part of an exon is deleted along with flanking non-exon sequences. The bottom panel depicts insertions and deletions that occur entirely within exons (this is the most abundant class). The complete set of coding INDELs is listed in Supplemental Table 5.

**Supplemental Figure 4. Genes that were affected by multiple coding INDELs.**

A sample of 50 of the 357 genes that were affected by more than one coding exon variant is shown. The name of the gene, function, and number of alleles discovered are listed. A red square next to the gene indicates that the gene previously has been linked to a known disease. The color scheme is consistent with that in Figure 2 and Supplemental Table 5. The full set of 357 genes is listed in Supplemental Table 5.

**Supplemental Figure 5. 25-mer probe design for our Affymetrix INDEL arrays**.

Six distinct examples are depicted (3 insertions and 3 deletions) along with the probe sets that were designed to interrogate each type. For the simplest insertions and deletions (1-3 bp), two sets of probes were designed to interrogate the two dimorphic states: 1) probes that were complementary to the reference genome sequence, and 2) probes that were complementary to the trace sequence. A total of six probes were designed for each of the two states: a) a 25-mer with the interrogated base(s) at the center (position 0), b) a 25-mer that was offset 4 base pairs to the left (-4) and another that was offset 4 bp to the right (+4). Three additional probes were generated against the reverse complement sequences of the first three probes. Thus, six probes were developed for each state for a total of 12 probes to interrogate the INDEL. For INDELs that were larger than 3 bp, a similar approach was used except that probe sets were designed for each of the unique junctions (at positions 0,-4, +4 plus the reverse complement of each probe for a total of 6 for each junction). Thus, 18 probes were developed for these INDELs. For INDELs that were larger than 25 bp, three additional unique probes were generated that were complementary to the central region of the occupied state and the reverse complements of these probes (a total of 24 probes for these INDELs). The Affymetrix Power Tools (APT) command

line software that was originally developed for SNPs was adapted to our probe design strategy to score each of these probe sets.

**Supplemental Figure 6. Analysis of trace frequencies for array INDELs.**

The trace frequencies were examined for the 1.96 million set of INDELs in our collection vs. the 10,003 that were genotyped on the array. The 10,003 INDELs on the array have a similar range of trace frequencies as the entire set.

**Supplemental Table 1. Summary of INDELs identified from traces generated by 12 centers.**

The statistics are shown for the small INDELs discovered. For each center, the following data were tabulated: the center generating the trace (abbreviations as listed at the trace archive); the total number of variants discovered from the traces generated by that center, the number of non-redundant variants; the number of variants that matched the chimpanzee genome; the number of variants that matched the Celera genome; and the number of double hit variants (confirmed by at least two independent data sets).

**Supplemental Table 2. Traces used for INDEL discovery.**

For each trace, the center project and other annotation in the trace record at NCBI were examined to determine the human donor(s). The traces that were used in our study were originally generated by sequencing centers for genome-wide SNP discovery and DNA sequencing. Additional traces were derived from BAC sequences that were initially generated during the human genome project but were not included in the final genome assembly. Traces

that were generated by the ENCODE project were specifically excluded to avoid regional biases. The traces were derived from a diverse set of healthy donors and were ideal for genome-wide INDEL discovery.

**Supplemental Table 3. Overlap of our INDELs with those reported in five published personal genomes.**

The co-ordinates for small INDELs reported in diploid personal genomes were obtained from the centers and compared to our INDELs. We performed the comparisons using the precise co-ordinates reported and also repeated the analysis allowing for some differences in co-ordinates due to differences in INDEL discovery and annotation. JCV corresponds to the genome of J. Craig Venter (Levy et al. 2007), JDW to James D. Watson (Wheeler et al. 2008), JH to a Han Chinese individual (Wang et al. 2008), NA10857 to a Yoruban individual (Bentley et al. 2008), and AML to an individual with acute myeloid leukaemia (Ley et al. 2008). The numbers above the diagonal indicate the number of INDELs shared between each set for various degrees of positional accuracy. The numbers below the diagonal show the percentage of overlap as compared to the smallest set between the two.

**Supplemental Table 4. INDELs that overlap with structural variants.**

INDELs were identified from our collection of 1.96 million that were >50 bp in length. A total of 7,245 INDELs >50 bp were identified, and 957 of these INDELs were >1 kb. This set of larger INDELs was compared to the structural variants (SV's) that have been identified by the research community. In particular, our set of 7,245 INDELs was compared to variants in the Database of Genomic Variants (Iafrate et al. 2004), to SV's that were reported by Conrad et al.

2010a and 2010b, and to SV's that were identified by the 1000 Genomes Project (1000 Genomes Project Consortium). Comparisons were performed as described previously (Mills et al. 2011).

**Supplemental Table 5. Coding INDELs**.

The co-ordinates and other information for the 2,123 variants that affect coding exons are listed. A light blue color in the 'INDEL_ID' column indicates that the INDEL was included on the custom array. A yellow color in the 'Name' column indicates that this gene is associated with a known disease. The color of the 'Summary' column is consistent with the color scheme labeled in Figure 2B. The following information is provided: INDEL_ID: local unique variant identification number; LOCATION: chromosome and co-ordinates of the INDEL on build hg18 of the UCSC genome browser. For insertions in the trace, only a single co-ordinate is given (the site where the insertion occurred); LEN: size of variant; INDEL_TYPE: Whether it is an insertion or a deletion relative to the chimpanzee genome; WEBLINK (GPLINK): an active hyperlink to the UCSC genome browser at the co-ordinates where the variant was identified; DOUBLE_HIT_TYPE: If the allele qualifies as a double-hit INDEL, then the type of evidence that was used to determine double hit status is provided; TI_NUMS: The trace identifier (TI) number(s) for all traces in which an INDEL was discovered; NUM_TRACES: the number of trace files that had this variant allele; NUM_UNIQUE_SOURCES: The number of unique sources providing independent evidence for the INDEL from different centers/projects (including chimp, Celera, traces from independent centers); CENTER_NAMES: listing of center names (defined in tracedb) that produced the traces in which this variant was found; CENTER_PROJECTS: listing of center projects (defined in tracedb) associated with the traces in which the variant was found; ACCESSION: RefSeq or Ensembl gene accession number;

NAME: Common name for gene; SUMMARY: Summary annotation for the gene that was obtained from UCSC and OMIM; FUNCTION: Classification of the genes identified using known functions; TYPE_OF_OVERLAP: How the INDEL overlaps exon(s) of the gene, as described in Supplemental Figure 3; MULTIPLE OF 3? Y = yes, N = no; FRAMESHIFT? Indicates whether the INDEL causes a frame-shift (Y = yes, N = no); IN-FRAME STOP CODON, Indicates whether the INDEL retains the original frame but introduces a stop codon at the INDEL site (Y = yes, N = no); MGI_ID: The Mouse Genome Informatics (MGI) identifier given to the uniquely orthologous gene in mouse; MGIPhenosReported: The abbreviated phenotypic categories reported by MGI for a given gene obtained by experiments on the mouse ortholog. The phenotypic categories are abbreviated as follows: **at** adipose tissue, **bn** behaviour/neurological, **ca** cardiovascular system, **ce** cellular, **cr** craniofacial, **da** digestive/alimentary, **em** embryogenesis, **ee** endocrine/exocrine gland, **gs** growth/size, **hve** hearing/vestibular/ear, **he** hematopoietic system, **ho** homeostasis, **im** immune system, **lpo** lethality-postnatal, **lpr** lethality-prenatal, **la** life span-post-weaning/aging, **ldt** limbs/digits/tail, **lb** liver/billiary system, **mu** muscle, **ne** nervous system, **no** normal, **ot** other, **pi** pigmentation, **ru** renal/urinary system, **rp** reproductive system, **rs** respiratory system, **sk** skeleton, **scn** skin/coat/nails, **to** taste/olfaction, **tv** touch/vibrissae, **tu** tumourigenesis and **ve** vision/eye. SEQ: Sequence of the variant.

**Supplemental Table 6. Genes with disruptive INDEL variants whose mouse orthologs, when disrupted, have known phenotypes.**

Disruptive INDELs were defined as those INDELS that (i) either wholly reside within a coding exon where the INDEL is not a multiple of 3, or overlap an exon boundary, (ii) do not

only affect the last exon, and (iii) affect all Ensembl transcripts for the overlapped gene. The mouse phenotypic categories were obtained from the Mouse Genome Informatics resource (www.informatics.jax.org). The following information is tabulated: the gene symbol; number of likely disruptive INDEL Alleles; Known disease associations; Mouse homozygous deletion phenotype; mouse heterozygous deletion phenotype. The mean number of likely disruptive INDELs is 1.4, the median is 1 and the standard deviation is 1.0. Thus, on average, these genes do not overlap more than one likely disruptive INDEL.

**Supplemental Table 7. INDEL:SNP ratios.**

The number of INDELs and SNPs were tabulated using the traces described in Supplemental Figure 1. INDELs and SNPs were counted within all RefSeq genes (defined as -3Kb from the first exon to +0.5 Kb downstream of the last exon) and for non-genic areas. Genes were subdivided further into the following features: Promoter (the 3 kb region upstream of the first exon), exon, intron, and 3' terminator (the 0.5 kb region downstream of the last exon).

**Supplemental Table 8. INDEL microarray genotyping calls in 161 humans**.

The genotyping calls for 10,003 small INDELs are listed for the 170 arrays (161 humans) examined in this study. The parameters used with Affymetrix Power Tools (APT) and BRLLM-P are listed above, along with statistics. The identities for the 161 humans are listed on the top row and include: 24 humans from the polymorphism discovery resource (Coriell-yellow), 90 humans from the HapMap3 plate of Yoruban trios (Coriell-orange), 42 humans from the HapMap CHB collection (Coriell-green), the Venter genome (Coriell-red), and 3 controls (blue). Technical replicates are indicated (grey). The same DNA samples were hybridized on separate

days in these cases. The genotyping calls were >99% identical on these replicates. The probeset IDs for 10,003 INDELs are listed in column A. INDELs on the X and Y chromosome were not included. Following hybridization, the data were normalized, analyzed for quality control, and clustered using BRLLM-P. For each INDEL, the call was either -1 (NN), 0 (AA), 1 (AB), or 2 (BB) using a confidence cutoff of 0.05. The other columns are: len = length of INDEL; coding = whether it is a coding INDEL (Y or N); Common = whether the minor allele is ≥5%; RefSeq gene = whether the INDEL falls within a known RefSeq gene (-3 kb to +0.5 kb); gene component = subregion of gene affected; NN, AA, AB, BB = number of calls for each genotype; A freq. = frequency of A allele; B freq. = frequency of B allele.

**Supplemental Table 9. Validation of array calls by PCR**.

The genotypes of 12 representative INDELs that were genotyped by PCR in 24 humans were compared to the results obtained with the same DNAs on our INDEL microarrays. 268/271 (99%) of the calls were identical with the two methods (excluding no-calls). PCR validation was carried out using platinum Taq polymerase (Invitrogen) using the manufacturer's instructions in 50 µl reactions. The following cycling parameters were used: 4 minutes at 94°C, and then 35 cycles of 1 minute at 94°C, 1 minute at 60°C, and 2 minutes at 72°C, followed by 10 minutes at 72°C. Custom primers are available upon request. INDEL genotypes were scored using restriction endonucleases (Mills et al. 2006), or by sequencing at least 20 cloned PCR products from each individual.

**Supplemental Table 10. Coding INDEL annotation and genotypes from microarrays.**

The genotyping results for the coding INDELs on our arrays. For each INDEL, the following data are listed: INDEL_ID: local unique variant identification number; LEN: size of variant; COMMON: Whether the allele is common (Y) or rare (N); GENOTYPES: NN, AA, AB, BB; FREQ_A: Frequency of the A (reference) allele; FREQ_B: Frequency of the B (trace) allele.

**Supplemental Table 11. Hardy Weinberg calculations.**

The genotyping results from the INDEL microarrays were subjected to Hardy-Weinberg analysis. The allelic distributions were examined and subjected to a $\chi^2$ test (significance level of $p = 0.01$) to determine whether the alleles were in Hardy-Weinberg equilibrium. If the alleles were in agreement with a Hardy-Weinberg equilibrium ($\chi^2<6.64$), then a 0 was entered (the Hardy-Weinberg hypothesis cannot be rejected). If the Hardy-Weinberg hypothesis was rejected ($\chi^2>6.64$), then a 1 was entered (for Yes, the Hardy-Weinberg can be rejected at this significance level).

**Supplemental Table 12. Summary of INDEL array genotypes by population.**

The genotypes from Supplemental Table 8 were broken down by subpopulations and then used to count population-specific variation in the Yoruban and Han Chinese populations. The headings are the same as for Supplemental Table 8, except the three subpopulations are indicated by “PDR”, “YRI”, and “CHB”.

**Supplemental Table 13. LD analysis with HapMap SNPs.**

The $r^2$ value was calculated for each SNP within a 1Mbp window of a given INDEL using the SNP genotypes that have been reported for HapMap3 (http://hapmap.ncbi.nlm.nih.gov/) and our INDEL genotyping data from the same samples. For each population (YRI, CHB), the SNP with the maximum $r^2$ value was identified.

**Supplemental Table 14. GWAS database (Johnson and O'Donnell, 2009).**

Over 56,000 SNPs from 118 published genome-wide association studies recently were collected into a single convenient database (Johnson and O'Donnell, 2009). This table represents that database and was downloaded from that reference. The database was used to determine whether any of the SNPs in these studies had high levels of pairwise LD with our INDELs. The original studies are referenced within the table. The following data were added to the original table (columns AW to BD): CHB INDEL ID (CHB_id); pairwise r2 value for INDEL and the SNP in column R (CHB_rsq); if the INDEL mapped to a known gene, the gene name is listed (CHB_gene_Name); if the INDEL mapped to a known gene, the gene feature is listed (CHB_geneType). Equivalent data are listed for INDELs that were examined in the YRI population (columns BA to BD).

**Supplemental Table 15. GWAS SNPs that have high LD (>0.8 $r^2$) with INDELs.**

SNPs with LDs above $r^2$ of 0.8 with INDELs from Supplemental Table 14 are listed and compared. Headings are the same as in Supplemental Table 14.

**Supplemental Table 16. INDELs caused by new retrotransposon insertions.**

INDELs were screened to identify new retrotransposon insertions within our collection. If a retrotransposon insertion was identified, we also determined whether it was flanked by a target site duplication (TSD). The following features were listed: the local INDEL ID (INDEL_ID); chromosomal location (LOCATION); INDEL length (LEN); a link to the INDEL in the UCSC browser (WEBLINK); the trace TI number (TI_NUMS); transposon class (CLASS); whether the INDEL was in a gene (IN_GENE?); if yes, the gene is listed (GENE); if in Iskow et al. 2010 (Yes = Y, No = N); whether there was a flanking target site duplication (HAS_TSD); if yes, the length of the target site duplication (TSD_LEN); and the sequence of the target site duplication on the left (TSD_SEQ_LEFT); and right (TSD_SEQ_RIGHT); and the sequence of the transposon (SEQ).

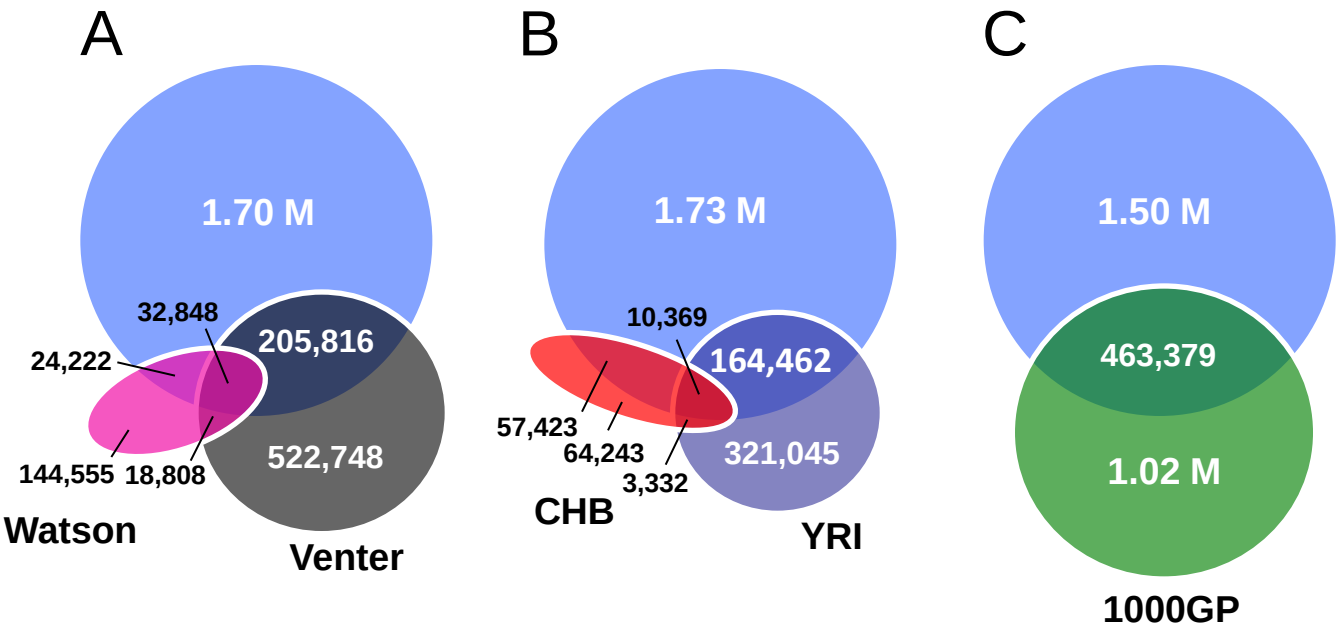
**Supplemental Tables chr1-chry.**

The complete data set of the 1,955,656 variants identified in this study is organized by chromosome into multiple Microsoft Excel files. All of these variants have been deposited into build 129 of dbSNP under the DEVINE_LAB handle. Some chromosomes have more than 64,000 variants and required multiple tabs. The coordinates refer to build hg18 of the human genome sequence. The column definitions are as follows: INDEL_ID: local unique variant identification number**;** GPLINK: an active hyperlink to the UCSC genome browser at the coordinates where the variant was identified**;** START: start coordinate of the variant**;** STOP: stop coordinate of the variant**;** IS_REVERSE: 'Y' for variants which mapped to the complementary strand**;** LEN: size of variant**;** DOUBLE_HIT: 'Y' for variants which had additional validation, defined as having another trace identifying the same variant, or the same allele identified in

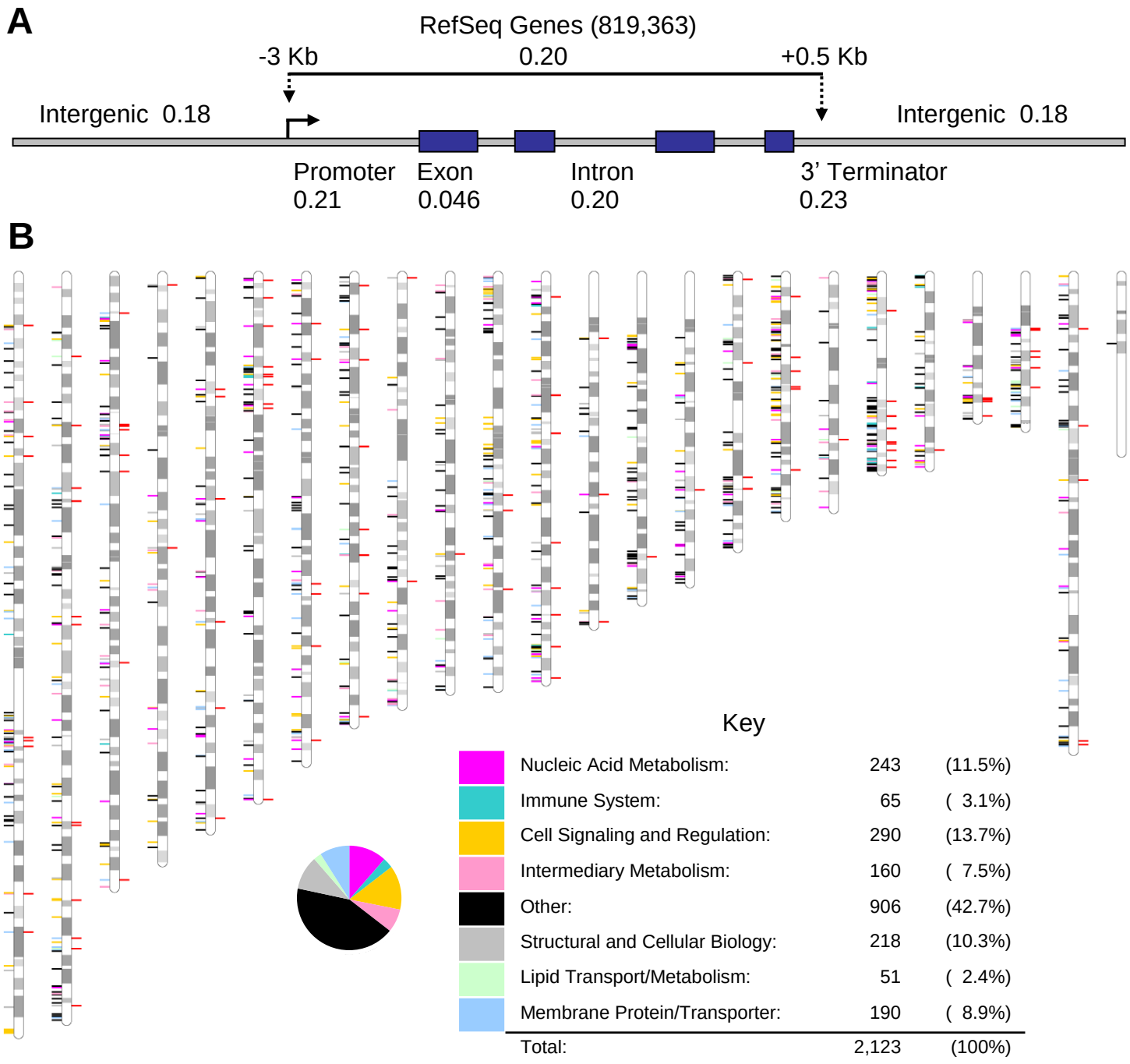
either the chimpanzee or Celera genome sequences; DOUBLE_CENTER: 'Y' for variants that were found in traces submitted by different sequencing centers**;** REF_TYPE: type of variant (INSertion or DELetion) relative to the reference sequence**;** CHIMP_TYPE: type of variant in the chimpanzee genome (panTro1). This value is 'T' for traces that could not be mapped to the chimpanzee genome and 'U' for traces that could be mapped but for which no allele matching the reference sequence or trace could be identified; CELERA_TYPE: same as chimp_type only with the Celera reference genome; CHIMP_CHR: chromosome to which the trace mapped in the chimpanzee genome. This value is 'U' if chimp_type was 'T'**;** CHIMP_START: start coordinate of the variant allele identified in the chimpanzee genome. This value is for the start coordinate of where the trace mapped if chimp_type is 'U'; CHIMP_STOP: stop coordinate of the variant allele identified in the chimpanzee genome. This value is for the stop coordinate of where the trace mapped if chimp_type is 'U'; CELERA_CHR,CELERA_START,CELERA_STOP: same as chimp only with the Celera reference genome; NUM_TRACES: the total number of trace files that had this variant allele**;** TRACENAMES: listing of the trace names in which this variant was found, delimited by commas; TIS: listing of TI numbers for the traces in which the variant was found**;** CENTER_NAMES: listing of center names (defined in tracedb) that produced the traces in which this variant was found**;** CENTER_PROJECTS: listing of center projects (defined in tracedb) associated with the traces in which the variant was found**;** PROJECT_NAMES: listing of center project names (defined in tracedb) for traces in which the variant was found; TRACE_TYPE_CODES: listing of the trace_type_code (defined in tracedb) for traces in which the variant was found**;** STRATEGIES: listing of the strategy (defined by tracedb) for each trace in which a variant was found**;** IS_GENOMIC: 'Y' for traces with the 'Genomic' field as 'G' or 'Genomic', as defined by tracedb. Variants identified only in traces with IS_GENOMIC as 'N'

are omitted; METHOD: indicates whether a variant was identified through the bl2seq or findmatch algorithms; GENE_NAME: RefSeq or Ensembl accession number for the gene to which this variant maps. This value is 'U' for instances where no gene affected; GENE_TYPE: indicates the type of overlap the variant has with the gene. Possible values are promoter (upstream 3 Kbp), terminator (downstream 500 bp), intron, exon, or 'U' for instances where no gene was affected. 'Exon' here can be coding or non-coding UTRs; IS_CODING: 'Y' for variants where the gene_type is 'exon' and which fall within the coding start and stop coordinates for a gene, as defined in RefSeq or Ensembl; CLASS: simple classification of the variant identified, using Sputnik, repeatmasker, and an in-house classification package; CLASS_TYPE: simple categorization of the variants identified. Possible values are single base, repeat expansion, transposons, and other; CONTIG: the assembly contig of the human genome reference sequence in which the variant was found (for dbSNP submission purposes)**;** SEQ: the sequence of the variant identified.

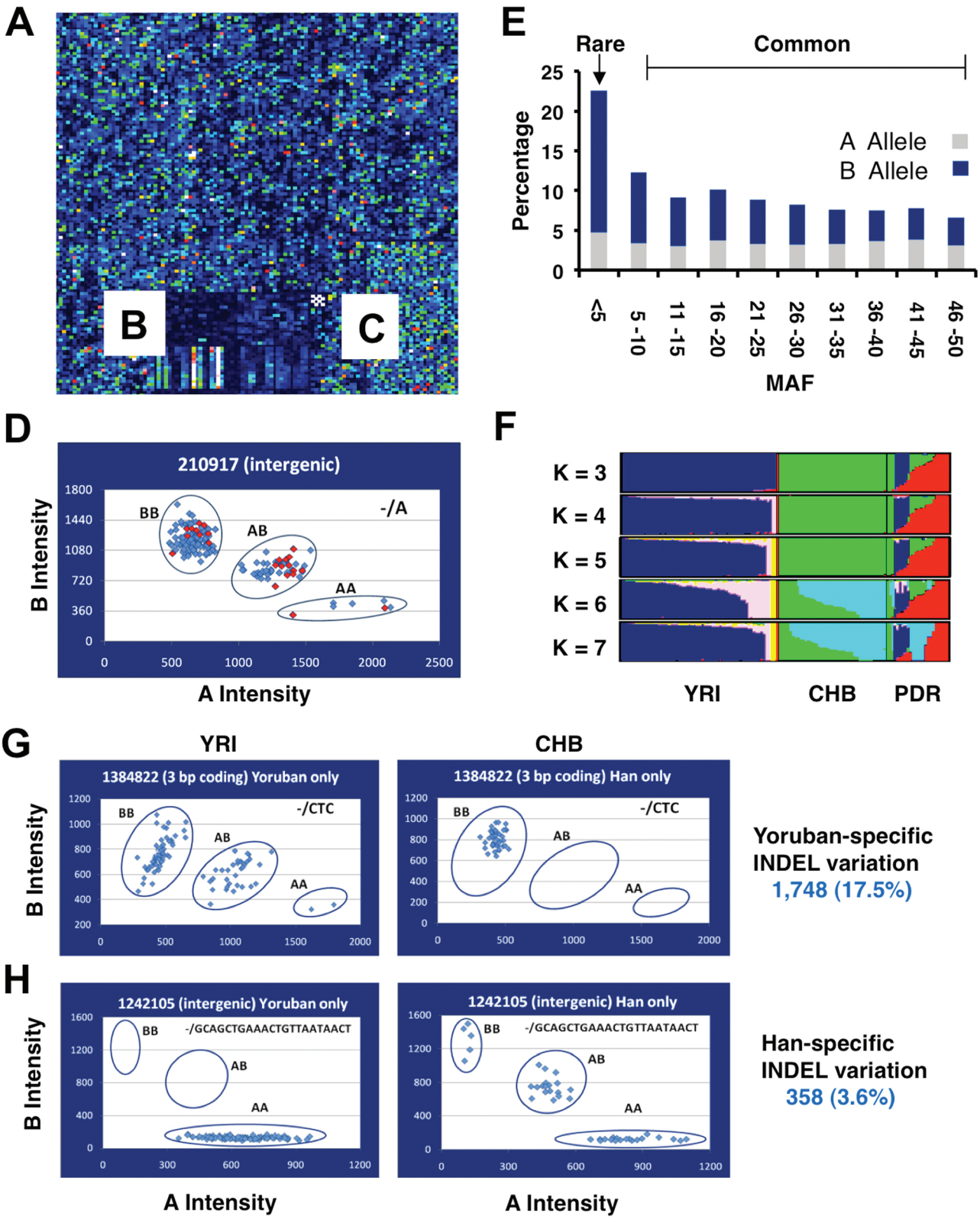
# Figure 1


A
1.70 M
32,848
205,816
24,222
144,555
18,808
522,748
Watson
Venter
B
1.73 M
10,369
164,462
57,423
64,243
3,332
321,045
CHB
YRI
C
1.50 M
463,379
1.02 M
1000GP

# Figure 2

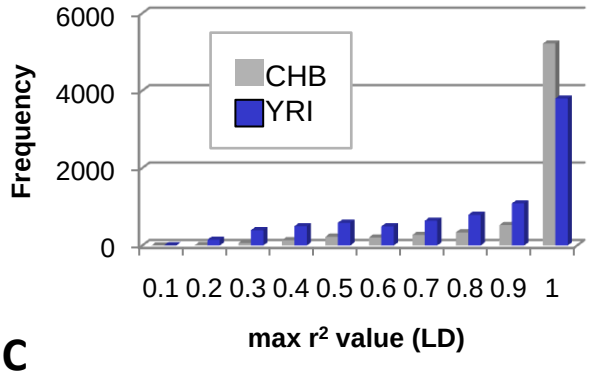


# Figure 3



# Figure 4

**A**




**B**

| Class | Number |
|---|---|
| GWAS SNPs in LD with $r^2 > 0.8$ | 2,290 |
| GWAS SNPs in perfect LD ($r^2 = 1$) | 1,193 |
| INDELs in genes (LD $r^2 > 0.8$) | 1,102 |
| exons total | 62 |
| coding INDELs | 6 |
| introns | 991 |
| promoters | 29 |
| 3' terminators | 20 |

**C**

| SNP | Disease/trait | INDEL | $r^2$ | Variant | Gene |
|---|---|---|---|---|---|
| **INDELs in coding exons** | | | | | |
| rs618145 | Coronary artery | 363513 | 1.0 | -/GTGACAGCCTGCAAC | NM_015225 |
| rs12685939 | Type I diabetes | 2235459 | 1.0 | -/AAA | NM_153267 |
| rs8067494 | Type II diabetes | 188688 | 1.0 | -/CAA | NM_000676 |
| rs8112054 | 2 hour glucose | 1192262 | 1.0 | -/ATA | NM_182515 |
| rs7940239 | Rheum. arthritis | 1286116 | 0.84 | -/GACA | NM_001004740 |
| rs700760 | Bone mass | 1122616 | 0.83 | -/TTT | NM_024617 |
| **INDELs in non-coding exons** | | | | | |
| rs11259906 | Crohn's disease | 1010795 | 1.0 | -/GT | NM_001717 |
| rs1055055 | Hypertension | 253658 | 1.0 | -/T | NM_173642 |
| rs3790043 | Parkinson's disease | 1585720 | 1.0 | -/C | NM_023112 |
| rs2657879 | Fasting glucose | 1060257 | 1.0 | -/AC | NM_207344 |
| rs6791490 | Triglyc/HDL ratio | 2619148 | 1.0 | -/AAC | NM_018010 |
| **INDELs in promoters** | | | | | |
| rs6747513 | ALS | 361854 | 1.0 | -/ATCATTATTACCACAGTG | NM_017750 |
| rs11087256 | Hypertension | 2154277 | 1.0 | -/AATA | NM_006363 |
| rs4712969 | Type I diabetes | 816478 | 1.0 | -/TGTAT | NM_001098486 |
| rs1455453 | Weight/Type 2 diab. | 1813108 | 1.0 | -/GCTACCTT | NM_021634 |
| rs10507353 | Crohn's disease | 382264 | 1.0 | -/A | NM_005977 |

## Table 1

| | |
|---|---|
| **Traces screened** | 98,350,511 |
| **Bases analyzed (trimmed)** | 42,000,606,219 |
| **Reference human genome (hg18)** | 3,107,677,273 |
| **Cumulative coverage** | 13.5X |
| | |
| **Variants identified** | 3,400,787 |
| **Unique (non-redundant) variants** | 1,955,656 |
| Insertions | 978,721 |
| Deletions | 976,935 |
| **Total bases affected** | 11,909,159 |
| **Total double hit** | 1,355,228 |
| Match chimp allele | 878,574 |
| Match celera allele | 648,941 |
| Match variant allele | 688,990 |
| **Total double center** | 477,295 |
| **Transposon insertions (with TSDs)** | 1,150 |
| Alu | 1,004 |
| L1-Ta/pre-Ta | 70 |
| SVA | 58 |
| **INDELs overlapping SV size range** (>50 bp) | 7,245 |
| (>1,000 bp) | 957 |

## Table 2

| Class | No. Rare/Total | % | P-value |
|---|---|---|---|
| All INDELs on Array | 2,265/10,003 | 22.7 | N/A |
| In RefSeq genes | 1,044/4,466 | 23.4 | 0.34 |
| Not in RefSeq genes | 1,227/5,537 | 22.2 | 0.50 |
| Exon (coding) | 72/111 | 64.9 | $5.1 \times 10^{-21}$ |
| (non-coding) | 32/122 | 26.2 | 0.33 |
| Intron | 889/4,044 | 22.0 | 0.41 |
| Promoter | 36/132 | 27.3 | 0.41 |
| Terminator | 15/57 | 26.3 | 0.53 |