## 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 leukemia (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 158 humans**.

The genotyping calls for 10,003 small INDELs are listed for the 165 arrays ( 158 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 158 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 a control (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 Table 17. dbSNP accession numbers for the INDELs identified in this study.**

All INDELs identified in this study were deposited to dbSNP (build 129). The ss numbers that were assigned by dbSNP are listed.

**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.