nature genetics

Extensive genomic and transcriptional diversity identified through massively parallel DNA and RNA sequencing of eighteen Korean individuals

Young Seok Ju[1,2,9], Jong-Il Kim[1,3–5,9], Sheehyun Kim[1,2,9], Dongwan Hong[1,8], Hansoo Park[1,6], Jong-Yeon Shin[1,5], Seungbok Lee[1,4], Won-Chul Lee[1,4], Sujung Kim[5], Saet-Byeol Yu[5], Sung-Soo Park[5], Seung-Hyun Seo[5], Ji-Young Yun[5], Hyun-Jin Kim[1,4], Dong-Sung Lee[1,4], Maryam Yavartanoo[1,4], Hyunseok Peter Kang[1], Omer Gokcumen[6], Diddahally R Govindaraju[6], Jung Hee Jung[2], Hyonyong Chong[2,7], Kap-Seok Yang[2], Hyungtae Kim[2], Charles Lee[6] &
Q1 Jeong-Sun Seo[1–5,7]

Massively parallel sequencing technologies have identified a broad spectrum of human genome diversity. Here, we deep sequenced and correlated 18 genomes and 17 transcriptomes of unrelated Korean individuals. This has allowed us to construct a genome-wide map of common and rare variants and also identify variants formed during DNA-RNA transcription. We identified 9.56 million genomic variants, 23.2% of which appear to be previously unidentified. From transcriptome sequencing, we discovered 4,414 transcripts not previously annotated. Finally, we revealed 1,809 sites of transcriptional base modification, where the transcriptional landscape is different from the corresponding genomic sequences, and 580 sites of allele-specific expression. Our findings suggest that a considerable number of unexplored genomic variants still remain to be identified in the human genome, and that the integrated analysis of genome and transcriptome sequencing is powerful for understanding the diversity and functional aspects of human genomic variants.

Massively parallel sequencing technologies have revolutionized our understanding of human genome architecture. A diverse array of high-throughput sequencing platforms and strategies have now been developed and applied to the analyses of the genomes of both healthy and clinically affected individuals[1–7]. Over the past few years, a broad range of genetic variants including SNPs, copy number variants (CNVs) and other structural genomic variants[1–4,8–13] have been discovered using deep sequencing. In 2010, a public database of human variants (dbSNP 131) catalogued approximately 20.1 million SNPs, and the database of genomic variants (DGV) listed 89,427 structural genetic variants from 38 projects[14]. Next-generation sequencing technologies have also allowed gene expression analysis to provide transcriptome maps at nucleotide resolution[15,16]. However, only a limited number of studies have analyzed human genome and transcriptome sequences for the same individuals on a genome-wide scale[6,17].

A primary goal of human genetics is to understand the relationship between genomic variants and phenotypic traits. Although genome-wide association studies (GWAS) have revealed associations between hundreds of specific variants with complex traits and diseases, only a small proportion of the heritability of these genetic traits has actually been explained[18], presumably because, in part, of unknown and rare functional variants. Comprehensive discovery of the impact of unknown and rare genomic variants on missing heritability requires unbiased whole-genome sequencing. The pilot phase of the 1000 Genomes Project has been published[19], but to gain a more detailed understanding of the genomic landscape of common and rare variants in humans, more individuals from different populations should be whole-genome sequenced at high-depth coverage. Furthermore, comparisons between genomic variants and their corresponding transcriptional profiles from the same individuals need to be performed to help understand the functional aspects of these variants.

Here, we have analyzed both genomic and transcriptomic sequences from healthy Korean individuals. These include high-coverage whole-genome sequencing of ten individuals and whole-exome sequencing of an additional eight individuals. We also generated complete transcriptome sequence data from 17 of these individuals to explore the relationship between identified genomic variants and the corresponding transcriptome variants. Our analyses integrate genome and transcriptome data across multiple individuals and reveal extensive variation at both levels.

[1]Genomic Medicine Institute (GMI), Medical Research Center, Seoul National University, Seoul, Korea. [2]Macrogen Inc., Seoul, Korea. [3]Department of Biochemistry, Seoul National University College of Medicine, Seoul, Korea. [4]Department of Biomedical Sciences, Seoul National University Graduate School, Seoul, Korea. [5]Psoma Therapeutics Inc., Seoul, Korea. [6]Department of Pathology, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts, USA. [7]Axeq Technologies, Rockville, Maryland, USA. [8]Present address: Division of Convergence Technology, Functional Genomics Branch, National Cancer Center, Goyang-si, Korea. [9]These authors contributed equally to this work. Correspondence should be addressed to J.-S.S. (jeongsun@snu.ac.kr).

Received 17 December 2010; accepted 3 June 2011; published online XX XX 2011; doi:10.1038/ng.XXXX

RESULTS

SNP and short indel identification

We extracted genomic DNA from 18 unrelated individuals (11 males and 7 females) from venous blood (**Supplementary Fig. 1** and Online Methods). All the individuals have Altaic Korean origins and have no known genetic diseases. From these 18 individuals, 10 were whole-genome sequenced to an average of 26.1-fold coverage (**Table 1**). The majority of the short-reads were sequenced from both ends with 76–151 bp read lengths (**Supplementary Table 1**). We aligned the short reads to the NCBI human reference genome build 36.3 (hg18) using GSNAP[20] and Bioscope (Life Technologies).

After applying a set of bioinformatic filter conditions that were established from previous conservative training processes[3], we identified 3.45 million to 3.73 million SNPs from each whole genome sequenced (**Table 1**). The ratio between heterozygous and homozygous SNPs was 1.50 on average, which agrees well with other previously annotated genomes[21]. The ratio from whole-genome sequences obtained using earlier sequencing platforms (1.67 for individual AK1 and 1.33 for individual AK2) slightly deviated from the average, potentially reflecting underlying differences of sequencing technologies in variant detection. Comparing sequencing-derived SNPs with data obtained from the Illumina 610K genotyping array revealed a 99.94% positive predictive value for SNP detection (**Supplementary Fig. 2** and **Supplementary Table 2**). We further validated these results using PCR amplification and Sanger sequencing of 33 randomly chosen heterozygous SNPs found as singletons among ten individuals studied, which showed a 100% success rate (**Supplementary Table 3**).

The SNPs identified among the ten whole-genome–sequenced individuals clustered into a non-redundant set of 8.37 million SNPs, 21.9% (1.83 million) of which were considered to be new when compared to dbSNP131. As such, this study has increased the number of identified SNPs by approximately 8% over what was previously known, including variants discovered by the pilot phase of the 1000 Genomes Project[19]. Specifically, of these 1.83 million new SNPs, we found 73.9% (1.36 million) as singletons among the ten individuals studied, suggesting that a large proportion of these new variants are rare (**Fig. 1a**). Each individual genome has approximately 130,000 of these putative rare SNPs, consistent with the understanding that many rare variants still remain to be identified. The remaining 26.1% (0.48 million) of the new SNPs are thought to be common among Koreans (preliminary allele frequency ≥10%) but were not identified by previous genome studies, suggesting that they could be population-specific variants. A genome-wide map summarizing the location and frequency for each identified SNP is provided in our database (see URLs)[22].

We also detected 1,191,599 indels (≤30 bp) in the ten individual whole-genome sequences (**Table 1** and **Supplementary Table 4**). Compared to previously sequenced personal genomes[2,3,6], we found approximately twice as many indels from each individual. The excess may be attributed to more accurate read alignments from the increased read lengths and an increased proportion of available paired-end sequencing reads. We validated 35 randomly chosen indels that appeared in the heterozygous state and found as singletons using PCR and Sanger sequencing, and we successfully validated all the regions (**Supplementary Table 3**). A large proportion of the indels identified (32.5%) were not found in dbSNP 131 and are therefore considered new, indicating that a large fraction of indels among human populations also remain to be discovered.

Out of the total 8.37 million SNPs, 23,025 were non-synonymous (nsSNPs). On average, each individual genome contains 8,431 nsSNPs in ~4,700 genes. Considering that approximately 20,000 genes lead to mRNA transcripts, this suggests that ~25% of mRNA coding genes could lead to variable functions between individuals.

To more comprehensively investigate the variants within human genes, we captured and sequenced entire exonic regions representing 17,133 genes were from an additional eight Korean individuals (six males and two females) (**Supplementary Table 5** and Online Methods). On average, we obtained 63.9× read depth per individual for the targeted regions. From the exome-capture sequencing of these eight individuals, we found a total of 35,740 SNPs and 15,697 indels. Of these, 17,833 (49.9%) of the SNPs were non-synonymous.

To obtain a global view of genomic variants influencing protein sequences among the 18 individuals, we merged the nsSNPs identified from the ten whole-genome and the eight whole-exome sequences into a non-redundant set, which included 28,179 variants, 8,130 (28.6%) of which are new based on their absence in dbSNP 131 (**Supplementary Table 6**). A subset of the nsSNPs showed remarkably high allele frequencies among the Koreans studied compared to other populations, including Europeans and west Africans represented in the HapMap project[23]. For example, rs4961 in *ADD1*, rs17822931 in *ABCC11* and rs3827760 in *EDAR* are known to be associated with common salt-sensitive hypertension[24], dry type ear wax[25] and thick hair morphology[26], respectively. We also identified new nsSNPs with high frequency among Koreans, for example, a polymorphism that alters isoleucine to methionine on the third exon of *PAWR* that is thought to influence apoptosis and tumor suppression[27]. We annotated the putative clinical implications of the nsSNPs identified using the Trait-o-matic algorithm[3] (**Supplementary Table 7** and see URLs).

We found a subset of genes to be highly enriched for nsSNPs, here called super nsSNP genes (**Supplementary Table 8** and **Supplementary Note**). For example, *ZNF717* and *CDC27* showed ~100 times increased density of nsSNPs compared to other genes (**Table 2**). Among the 86 super nsSNP genes, 49 (57.0%) are associated with sensory and immunological function, such as olfactory receptors and *HLA*-related genes. The enrichment of nsSNPs in some of these genes, such as *PRIM2*, may be explained by hidden duplications that are yet to be represented in the human genome, such as *HYDIN* (**Supplementary Figs. 3,4**)[28,29]. Notably, a substantial number of these genes overlap with copy number variant regions (84%) that are archived in the DGV[14] as well as segmental duplications (37%) found in the human

Table 1 Summary statistics of whole-genome sequencing

Individuals	Gender	Aligned read depth	No. SNPs	Het./hom.	No. new SNPs	No. nsSNPs	No. nsSNP genes	No. indels
AK1[a]	Male	27.8×	3,453,606	1.67	339,067	10,310	5,560	170,202
AK3	Male	25.8×	3,510,326	1.49	259,958	7,515	4,296	373,687
AK5	Male	26.1×	3,595,936	1.56	285,766	8,279	4,606	367,651
AK7	Male	30.6×	3,573,029	1.56	283,133	7,516	4,275	414,397
AK9	Male	27.0×	3,700,128	1.45	294,005	7,895	4,363	507,624
AK2[b]	Female	29.3×	3,588,328	1.33	336,200	9,742	5,459	213,719
AK4	Female	23.1×	3,630,428	1.47	259,792	8,132	4,577	429,259
AK6	Female	22.3×	3,558,703	1.46	250,126	7,689	4,346	413,950
AK14	Female	26.1×	3,726,964	1.48	276,253	8,571	4,720	478,072
AK20	Female	22.7×	3,686,270	1.55	277,118	8,656	4,796	414,045
Total		260.8×	8,367,302		1,833,102	23,025	9,315	1,191,599

No., number; het./hom., ratio of heterozygous to homozygous SNPs.
[a]We considerably included single-end and shorter (2 × 36 bp) reads[3]. [b]We used Life Technologies SOLiD.

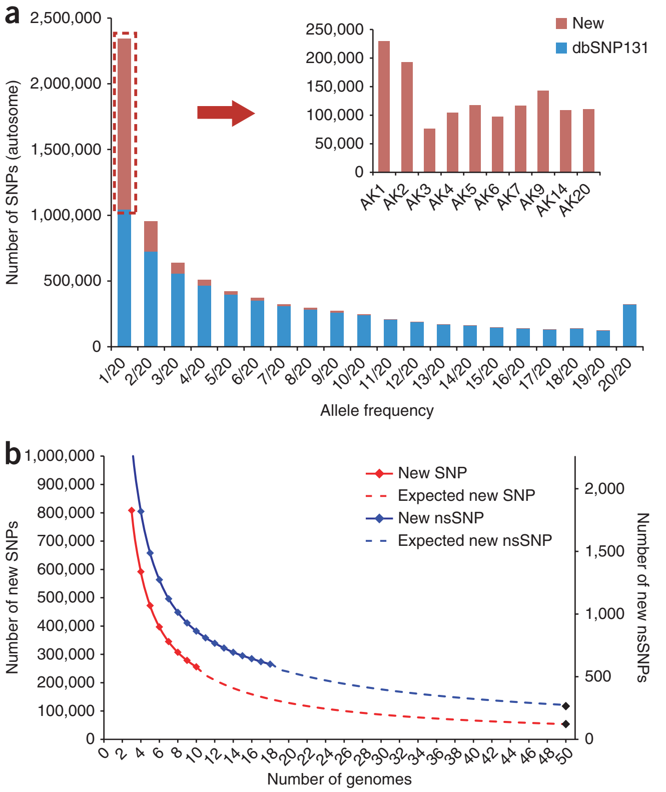

Figure 1 New and rare SNPs in individual genomes. (**a**) The allele frequencies of autosomal SNPs from whole-genome sequence data. The number of singleton and new SNPs in each individual genome is also shown. (**b**) The number of new SNPs and non-synonymous SNPs as the number of personal genomes increased through the simulation study.

reference genome, suggesting that super nsSNP genes are clustered near common structural variants in the human genome.

Rare and population-specific variants

To obtain a reasonable estimate of the number of rare variants in a given individual's genome, we simulated the number of variants that would be detected as 'new' as the number of personal genomes obtained increases (**Supplementary Note**). Our simulations suggested that, among the ~3.5 million SNPs in an individual genome, ~54,000 (1.5%) and ~28,000 (0.8%) have an allele frequency of <1% and <0.5%, respectively (**Fig. 1b**). This number should be interpreted cautiously, as approximately 2,000 SNPs in an individual genome are considered to be false positives when the positive predictive value of SNP detection is 99.94%, as it was in our study. The nsSNPs are estimated to account for a larger proportion of rare SNPs (~3% and ~1.9% of nsSNPs in an individual have of <1% and <0.5% allele frequency, respectively) in this analysis, suggesting that nsSNPs have greater diversity among individuals. This enrichment of nsSNPs among rare variants may be caused by negative selection, which removes deleterious alleles from the human population and thus drives those alleles to low allele frequency.

The deep sequencing of Korean genomes also allowed us to assess the relationship between common Korean nsSNPs (allele frequency more than ~10%) and previously known SNPs. For example, a common Korean nsSNPs (archived in dbSNP131 as rs76418769 with an estimated allele frequency of ~20% among Koreans), which alters glycine to arginine on the fourth exon of *IL23R* (a gene known to be associated with Crohn's disease and psoriasis) showed low linkage disequilibrium (LD) with nearby (<20 kb) known SNPs (**Fig. 2a**; maximum $r^2 < 0.10$). Notably, a tagging SNP (rs6687620) was noninformative for this region in Koreans (with an estimated allele frequency of 1). Indeed, 53.4% of new but common Korean nsSNPs were in insufficient LD (r^2 <0.8) with currently known SNP variants within a distance of 20 kb (**Fig. 2b** and **Supplementary Table 9**). This finding suggests that current GWAS based on common tagging SNPs observed in other ethnic groups (for example, Europeans) may have fundamental limitations in other populations for detecting common nsSNPs[30].

Table 2 Selected super nsSNP genes (top 20) from whole-genome sequencing data

				nsSNP		
Gene	Chr.	Position (Mb)	Exon length (bp)	Average number	Average density (per kb)	Increase of read depth[a]
ZNF717	3	75.869	2,749	75.0	26.28	+
OR4C3	11	48.303	991	16.5	16.65	+
CDC27	17	42.553	2,494	40.6	16.28	+
FRG2C	3	75.796	853	13.5	15.83	+
OR4C45	11	48.323	921	13.7	14.88	
OR9G9	11	56.224	919	12.9	14.04	+
FAM104B	X	55.189	342	4.6	13.45	
PRIM2	6	57.291	1,543	17.7	11.47	+
HLA-DRB1	6	32.655	807	8.5	10.53	
HLA-DPA1	6	33.144	787	7.8	9.91	
CTBP2	10	126.668	1,347	12.6	9.35	
SEC22B	1	143.808	653	6.0	9.19	+
HLA-DQB1	6	32.736	791	7.1	8.98	
OR13C5	9	106.401	958	8.4	8.77	
KCNJ12	17	21.259	1,303	11.2	8.60	
MUC4	3	196.960	3,401	28.4	8.35	
TAS2R31	12	110.743	931	7.3	7.84	+
HLA-DQA1	6	32.713	772	6.0	7.77	
OR51Q1	11	5.400	955	7.4	7.75	
HLA-A	6	30.018	1,106	8.4	7.59	

Chr., chromosome.
[a]Read depth for the corresponding gene increased by >30%. This may suggest hidden duplication of the gene in genomes of individuals sequenced.

Large deletions with breakpoints

We identified deletions greater than 1 kb among the eight individuals that were whole-genome sequenced. We excluded two of the individuals (AK1 and AK2) from deletion analyses because their genomes were sequenced using earlier sequencing platforms (for example, single-end and short-length reads (≤36 bp), making it difficult to apply current analytic strategies for detection of large deletions). All of these approaches were based on, and improved from, previous genome sequencing studies[3] (**Fig. 3a** and Online Methods). We identified a total of 5,496 large deletion segments from these eight individuals (1,348 regions when compressed using a >50% reciprocal overlap criteria). The size of these deletions ranged from 1 kb to 46 kb (**Supplementary Table 10**). We validated the deletions with an ultra-high–resolution genome-wide

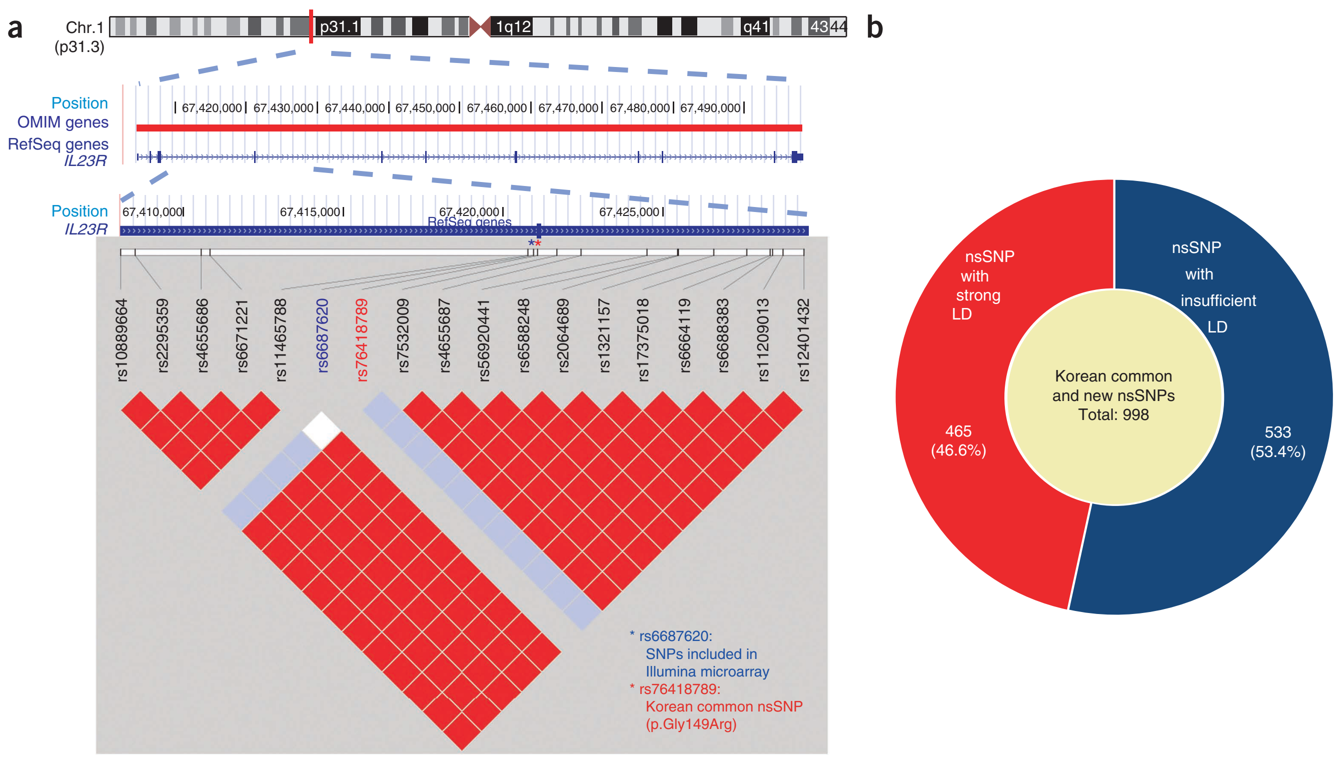

Figure 2 Linkage disequilibrium between new non-synonymous SNPs and known SNPs. (**a**) An example of a new non-synonymous SNP on *IL23R* genes (with an estimated allele frequency of ~25% in the Korean population) showed low LD ($r^2 < 0.1$) with known SNPs nearby. (**b**) We did not identify significant linkage disequilibrium for more than 50% of the new non-synonymous SNPs detected. The criteria for strong LD was $r^2 \geq 0.8$.

comparative genomic hybridization (CGH) array set comprising 24 million probes[31], and 83.8% of the tested regions were consistent with the array CGH data (**Supplementary Table 11**). Of these, 1,171 segments (407 compressed regions) were not previously catalogued in the DGV[14] using a > 50% reciprocal overlap criterion and are therefore considered to be new.

On average, we identified ~690 large deletions in each individual. These deletions overlapped ~200 genes. The coding sequences of some of the genes, such as *LCE3B*, *LCE3C*, *LILRA3*, *HYAL1* and *UBE2E3*, were frequently spanned by the deletions.

To identify nucleotide-resolution breakpoints for the large deletions, we analyzed the 'split reads' that aligned with discontinuous mapping across the junction of each deletion using modified methods of previous reports[32,33] (**Supplementary Fig. 5** and **Supplementary Note**). Our breakpoint-detection system was initially trained on NA10851, a HapMap sample of European origin that is well characterized for CNVs[8] and often used as a CGH array reference source. We successfully identified 48.4% of the breakpoints in this individual (**Supplementary Table 12**). Applying this system in the eight Korean individuals, we mapped 3,346 (60.9%) of the 5,496 large deletions identified to nucleotide resolution. These breakpoints were compressed to 1,392 non-redundant ones. Of these, 1,068 (76.7%) do not overlap with any previously identified breakpoints[33,34] (**Fig. 3b**), suggesting that a large fraction of the large deletions may be population specific.

We investigated the molecular mechanisms underlying the deletions that we identified (**Fig. 3c**). Approximately 18% of the deletions consisted of simple repetitive sequences (variable number tandem repeats) and 13% were consistent with the generation by non-allelic homologous recombination events. Generally, these inferred mechanisms are in good agreement with previous genome-wide estimation[32], except that our study found a slightly lower fraction of large deletions by transposable element insertions because of the fact that we targeted relatively large-size deletions (>1 kb). The predominant mechanism for deletion formation appeared to be that of non-homologous end joining (NHEJ; ~66%). We carried out a search for microhomology sequences for NHEJ-mediated large deletions using MEME Tools[35] and identified a significant enrichment of the 'CTCAGCCTCC' motif (MEME $E = 1.2 \times 10^{-881}$) (**Supplementary Table 13**). Out of the 1,022 non-redundant NHEJ-mediated large deletions, 511 showed the motif sequence near the estimated boundaries (<400 bp).

Transcriptome sequencing analysis

To investigate the transcriptional impact of the genome, we sequenced total RNA from 17 out of the 18 Korean individuals. We extracted the RNA samples from lymphoblastoid cell lines established by Epstein-Barr virus transformation of peripheral blood mononuclear cells (**Supplementary Table 14**). We sequenced an experimental set, which included 15 samples, using Illumina Genome Analyzer IIx. Additionally, we used another set of transcriptome sequences, including from the remaining two samples (AK1 and AK2), for validation purposes.

To prevent misalignment of RNA short reads bearing splicing junctions, which increases error rate in RNA sequencing (**Supplementary Fig. 6** and **Supplementary Table 15**), we aligned the short reads onto a set of complementary DNA (cDNA) sequences generated using ~160,000 mRNA sequences obtained from the RefSeq, UCSC and Ensembl databases (**Fig. 4a** and Online Methods).

We generated the expression map for all currently known RefSeq genes and found evidence for active transcription (expression level ≥1 rpkm[36]) of 11,101 genes in the lymphoblastoid cells used in the study (**Supplementary Table 16**).

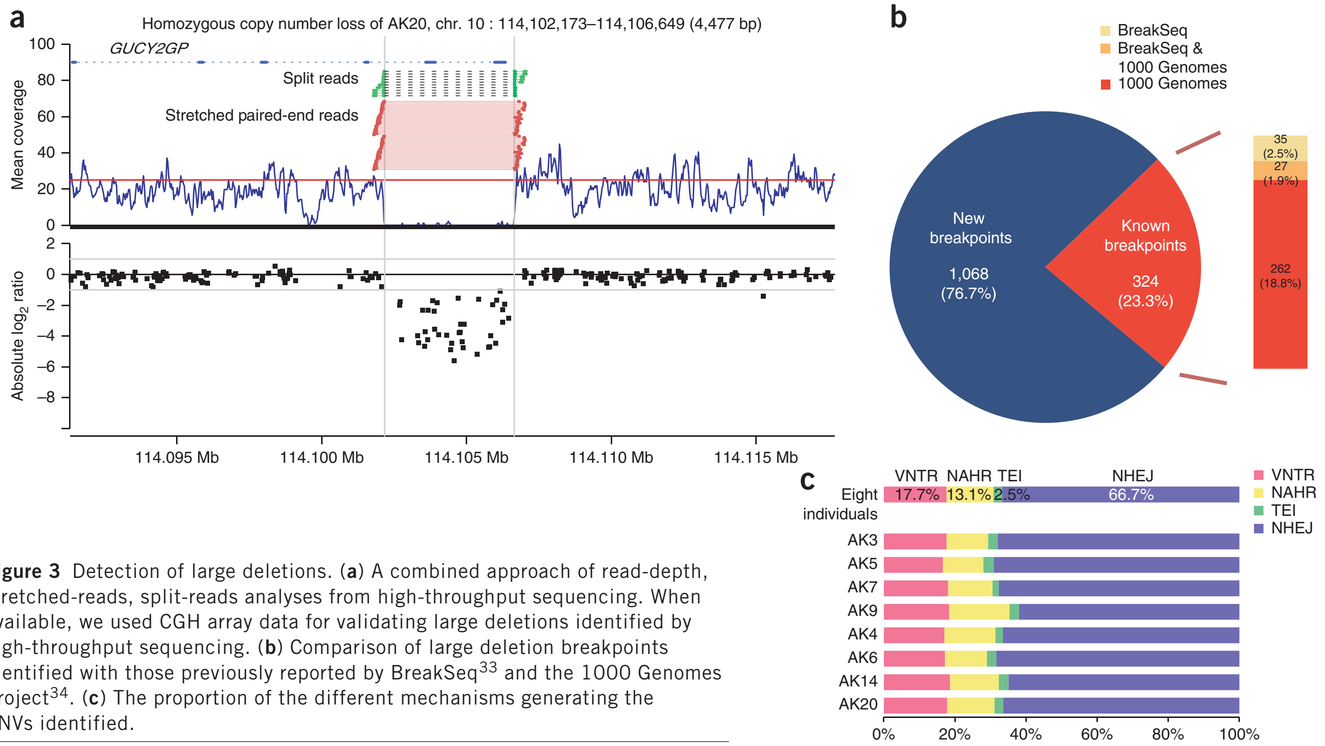

Figure 3 Detection of large deletions. (**a**) A combined approach of read-depth, stretched-reads, split-reads analyses from high-throughput sequencing. When available, we used CGH array data for validating large deletions identified by high-throughput sequencing. (**b**) Comparison of large deletion breakpoints identified with those previously reported by BreakSeq[33] and the 1000 Genomes Project[34]. (**c**) The proportion of the different mechanisms generating the CNVs identified.

To identify transcripts expressed on genomic loci previously not annotated as genes (unknown transcripts), we analyzed RNA short reads that could not be aligned to the cDNA sequences[15,37]. We aligned these unmapped reads to human genome sequences excluding known genic regions in the RefSeq, UCSC, Ensembl and GenBank databases (**Supplementary Note**). Furthermore, we removed short reads overlapping known expression sequence tags and transcripts detected in only one individual. We finally selected 4,414 regions as final unknown transcripts, which do not overlap any known genes or pseudogenes. We identified all the 4,414 unknown transcripts from at least two individuals and detected 111 unknown transcripts from all the 15 samples. We performed PCR validation on four randomly chosen unknown transcripts, and all were successful (**Supplementary Fig. 7**). These transcripts cover 2.74 Mb of the human genome and are located 4.7 kb (median distance) from the nearest gene, which is slightly longer than the median length of introns (2 kb) (**Supplementary Fig. 8** and **Supplementary Table 17**). The median length of these 4,414 unknown transcripts was 360 bp (**Fig. 4b**), which is slightly longer than that of currently known exons (130 bp). To identify the potential function of the unknown transcripts, we compared them with all known protein sequences from different species using BLASTX (see URLs). Of the 4,414 new RNA-derived genic regions, 19.5% (862) showed homology (≥20 amino acids long with ≥80% identity) with known protein sequences.

We then examined gender differences in expression level of X-chromosomal genes to explore X-inactivation profiles. We found 23 genes, including *XIST*, *PNPLA4*, *HDHD1A*, *NGFRAP1* and *GPR34*, that showed higher levels of gene expression in females compared to males (**Fig. 4c**, **Supplementary Table 18** and **Supplementary Note**). Notably, five genes (including *NGFRAP1* and *GPR34*) that appeared to escape X inactivation had not been previously shown to escape X inactivation in human fibroblasts[38]. Although our data should be interpreted cautiously because of our small sample sizes, and we cannot distinguish whether differences in expression between the sexes is caused by escape from X inactivation or differential activation in the X chromosome, the results suggest that high-throughput sequencing may be used to enable comprehensive analysis of human X-inactivation profiles.

Comparison of DNA and RNA sequence

To investigate nucleotide change during transcriptional processes, we looked for evidence of RNA variants that did not correspond exactly with their genomic sequence of origin[17,39–41]. Applying conservative filter criteria to genome and transcriptome sequences of the same individuals, we identified 1,809 sites of such transcriptional base modification (TBM; **Supplementary Note**). We identified each of the TBMs from at least two individuals. These TBMs are thought to be a result of nucleotide modification occurring during or after gene transcription (**Supplementary Table 19**). We performed PCR and Sanger sequencing on cDNA and genomic DNA from 16 loci, which validated 15 of the 16 sites (93.7%) (**Supplementary Fig. 9**). On average, the lymphoblast cell lines of each individual showed ~500 sites of TBM. Of these 1,809 TBMs, 74.1% (N = 1,341) were nucleotide transition. Of the transitions, 81.7% (N = 1,096) were A to G (N = 985) or C to T (N = 111) modifications on the coding strand, which could be explained by previously known molecular mechanisms of A to I and C to U RNA editing[39,41] (**Fig. 5a**). Of the A to G and C to T modifications, 30.2% overlapped with computationally predicted RNA editing sites[42]. Approximately 90% of the TBMs (N = 1,621) were located in untranslated regions, suggesting potential functional roles in modifying mRNA stability[43]. However, 188 of the TBMs were located in coding sequences, 128 of which resulted in different amino acids and therefore were capable of influencing protein structure or function. For example, the T to G modification in *PHB2*, found in 14 individuals, changes a stop codon (TGA) to glycine (GGA).

Next, to investigate allele-specific expression[44,45] or allelic expression imbalances, we compared the allele-specific read counts from genome and transcriptome sequence data of individuals heterozygous in a tested SNP (**Supplementary Note**). From our dataset,

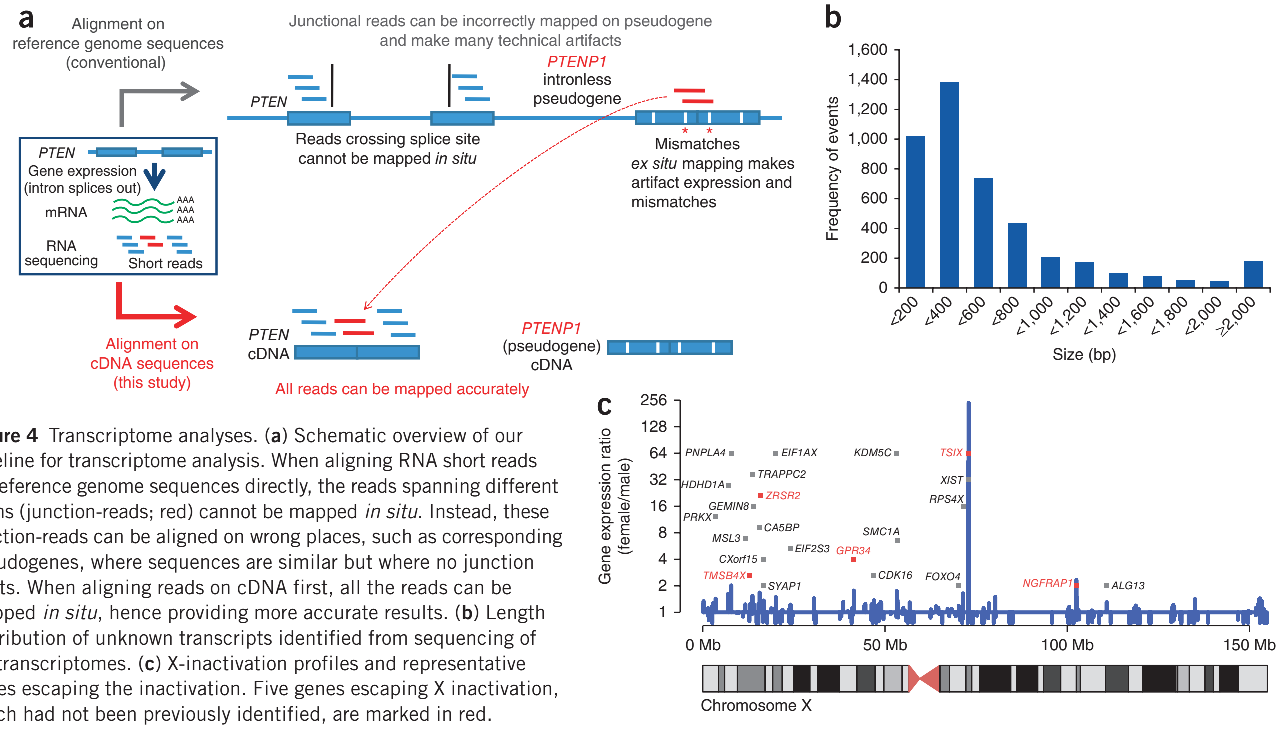

Figure 4 Transcriptome analyses. (**a**) Schematic overview of our pipeline for transcriptome analysis. When aligning RNA short reads on reference genome sequences directly, the reads spanning different exons (junction-reads; red) cannot be mapped *in situ*. Instead, these junction-reads can be aligned on wrong places, such as corresponding pseudogenes, where sequences are similar but where no junction exists. When aligning reads on cDNA first, all the reads can be mapped *in situ*, hence providing more accurate results. (**b**) Length distribution of unknown transcripts identified from sequencing of 15 transcriptomes. (**c**) X-inactivation profiles and representative genes escaping the inactivation. Five genes escaping X inactivation, which had not been previously identified, are marked in red.

we found 580 nsSNP sites that showed preferential expression of one allele compared to the other (**Fig. 5b** and **Supplementary Table 20**). For example, the wild type (G allele) of rs2549782, which generates lysine in *ERAP2*, was preferentially transcribed in the lymphoblastoid cell lines examined. Of these 580 sites, 18 SNPs signal a termination of translation. Our data suggest that allele-specific expression is widespread and may have functional importance in human gene expression variation.

New sequences from *de novo* assembly

Recent human genome studies have reported numerous new sequence variants that are not found in the human reference genome[2,46,47]. To discover new human genomic sequences among individuals used in our study, we assembled ~180 million paired-end short reads that were unmapped to the human reference genome using the *de novo* assembler Abyss[48] (Online Methods). We generated more than 12 million contigs. The maximum and N50 (where N50 is a weighted median statistic such that 50% of the entire assembly is contained in contigs equal to or larger than this value) length of the contigs were 66,717 bases and 118 bases, respectively, and we filtered for 1,937 contigs that were greater than 1 kb in length (**Supplementary Table 21**). We then compared the remaining contigs to the HuRef genome sequence[49] and human reference genome build 37.1 and finally obtained 947 contigs that could not be mapped on these genome sequences. We successfully aligned 19 of the contigs to the chimpanzee reference genome (*Pan troglodytes 2*) with 99% identity. One of these contigs showed evidence of transcription through our analyses of the transcriptomes of the eight individuals in this study. We found

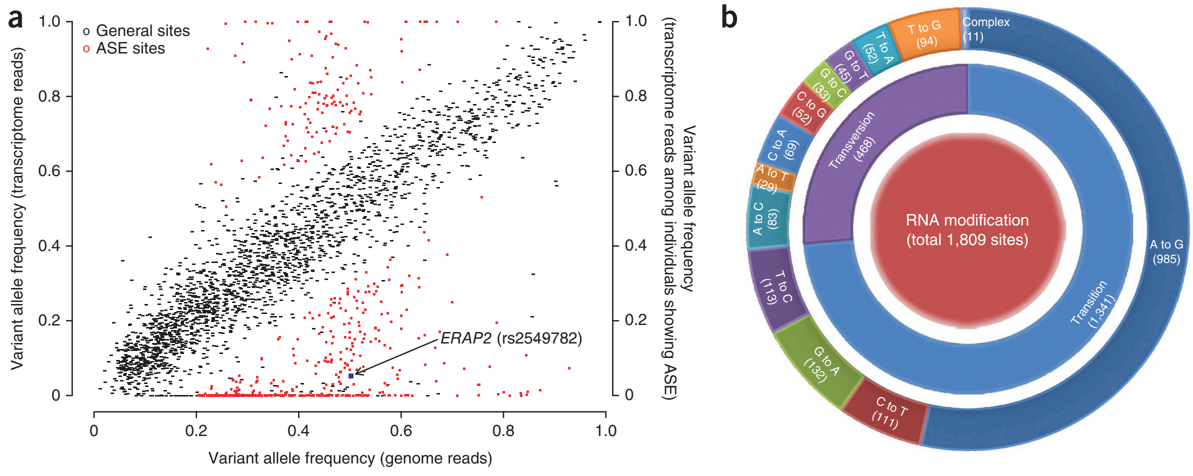

Figure 5 Comparison of genome and transcriptome sequences. (**a**) Relative contributions of each pattern of TBMs. (**b**) Genes showing allele-specific expression. We compared variant allele frequencies of genome and transcriptome sequence reads at SNP sites tested. Both the allele frequencies are balanced at the majority of the loci (black). Five hundred eighty genomic loci showing allele-specific expression are shown in red.

nine contigs in all eight individuals, and moreover, we found three contigs in the eight individuals as well as seven previously sequenced individuals (NA10851 (ref. 8) and NA12878 (ref. 19) with European ancestry, NA18507 (ref. 1), NA19240 (ref. 19), ABT (ref. 12) and KB1 (ref. 12) with African ancestry and YH (ref. 2) with Asian ancestry and Palaeo-Eskimo for ancient human[10]), suggesting that some of these contigs may be common among different human populations (**Supplementary Table 22**).

We examined the sequences of the remaining 928 contigs that were not mapped to the chimpanzee reference genome using Tandem Repeat Finder (see URLs). Out of these 928 contigs, only 30 (3.28%) were comprised of low-complexity sequences or had repetitive sequences comprising more than 50% of their length.

DISCUSSION

In order to gain a comprehensive understanding of human genomic variation as well as of susceptibility to complex diseases, many issues need to be addressed. These include (i) an understanding of the relative abundance of common and rare variants in the human genome, (ii) the amount of genomic difference between ethnic groups, (iii) the extent of LD between common and rare variants in an individual genome and (iv) the discovery of functional variants that influence complex human phenotypes and diseases. Through the genome sequencing of unrelated Koreans, we identified a considerable number of new variants and found that many rare putative functional variants likely remain to be identified despite major efforts in recent years to catalog human genomic variation. Indeed, our findings suggest that a substantial number of Korean common functional variants may not be tagged well by neighboring ‘tagging’ SNPs on microarrays. These results suggest that many association studies may have fundamental limitations, especially for populations that were not included in the initial LD assessments on the human genome[23].

As discussed in previous reports, increased lengths of sequencing reads are highly advantageous for short indel detection[3,9]. In addition to the read length, the read depth is also a critical factor. Low-coverage sequencing may miss short indels despite longer read length[13], especially for heterozygous indels[9]. Compared to SNPs, there appears to be a lack of consensus on the number of indels present in an individual’s genome[1,4]. More personal genome sequencing with longer read length and higher read depth will provide a more accurate understanding of the prevalence and locations of indels in the human genome.

Additionally, we have also identified 1,348 large deletions with nucleotide resolution. We conclude that copy number deletions can be detected accurately using whole-genome sequencing, especially if data from multiple genomes can be compared. Increased read length of short reads improved efficiency for identification of exact breakpoints in large deletions compared to previous studies[32,33]. Because the sensitivity for structural genomic variant detection by whole-genome sequencing is still incomplete, most of the structural variants we found are deletions, are large in size (>1 kb) and are located at relatively ‘easily accessible regions’ of the human genome. More personal genome sequences with high-coverage and longer DNA reads will enable detection of human structural variants more accurately, such as smaller deletions (50–500 bp), copy number gains, DNA insertions and inversions. This will provide a more complete understanding of the unbiased characteristics and functional impact of human structural genomic variation.

From transcriptome sequencing, we identified 4,414 genomic regions that are active in transcription, although they are not currently annotated as human genes. Large-scale transcriptome sequencing of various types of human cells will likely aid a more comprehensive understanding of the functional roles of these unknown transcripts.

Finally, we integrated genome and transcriptome sequence in the same individuals. This identified widespread TBMs, or nucleotide changes occurring during RNA transcription. Some of the TBMs are explained by known A to I and C to U RNA editing mechanisms. However, other patterns of base conversion, including nucleotide transversion (for example, G to A or T to G), have not been studied extensively. Notably, those patterns have recently been reported in *Arabidopsis thaliana*[50]. The TBMs may affect the susceptibility of complex diseases because they are likely to modify mRNA stability and to change amino acids of protein sequence. A combination of deeper genome and transcriptome sequencing of a variety of tissues from more individuals, including those clinically affected, will be necessary to assess the complete profile and the functional impacts of TBMs.

URLs. TIARA database, http://tiara.gmi.ac.kr/; Trait-o-matic, http://snp.med.harvard.edu/; CCDS database, http://www.ncbi.nlm.nih.gov/CCDS/CcdsBrowse.cgi; R statistics, http://www.R-project.org/; NCBI short read archive, http://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?; genome browser in GMI database, http://tiara.gmi.ac.kr/; Tandem Repeat Finder, http://tandem.bu.edu/trf/trf.html; BLASTX, http://blast.ncbi.nlm.nih.gov/Blast.cgi.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturegenetics/.

Accession codes. Whole-genome sequence data are freely available from the NCBI short read archive, with the following accession codes: SRA008370, SRA010321, SRA023746, SRA023747, SRA023748, SRA023749, SRA023750, SRA023751, SRA023752, SRA023753. SNPs and indels are deposited in the dbSNP database under the handle GMI.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

Q2

We acknowledge the anonymous Korean individuals who participated in this study. We thank Y.J. Yoo and T. Bleazard at Seoul National University for their personal comments regarding this manuscript. We are indebted to the scientists who contributed to this work but were not included in the author list. This work has been supported in part by Green Cross Therapeutics (grant # 0411-20080023 to J.-S.S.), the Korean Ministry of Knowledge Economy (grant # 10037410 to J.-I.K.), the Korean Ministry of Education, Science and Technology (grant # M10305030000 to J.-S.S; grant # 2010-0013662 to J.-I.K.) and the National Human Genome Research Institute of the US National Institutes of Health (grant # P41HG4221 to C.L.).

AUTHOR CONTRIBUTIONS

J.-S.S. and C.L. conceived of the project. J.-S.S. planned and managed the project. Y.S.J., J.-I.K., Sheehyun Kim, D.H., W.-C.L., Sujung Kim and S.-B.Y. analyzed sequencing data. D.H. and S.-S.P. developed the genome browser. J.-Y.S., S.-H.S., J.-Y.Y., H.C., K.-S.Y. and H.K. constructed libraries and executed sequencing. J.H.J. analyzed genotyping microarray experiments. H.P., S.L., H.-J.K., H.P.K. and O.G. assisted in the data analysis. Y.S.J., S.L., D.-S.L. and M.Y. performed validation analyses. J.-S.S., C.L., Y.S.J., J.-I.K., Sheehyun Kim, D.H., O.G. and D.R.G. wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/naturegenetics/.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Bentley, D.R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53–59 (2008).
2. Wang, J. *et al.* The diploid genome sequence of an Asian individual. *Nature* **456**, 60–65 (2008).

3. Kim, J.I. *et al.* A highly annotated whole-genome sequence of a Korean individual. *Nature* **460**, 1011–1015 (2009).
4. Drmanac, R. *et al.* Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* **327**, 78–81 (2010).
5. Pushkarev, D., Neff, N.F. & Quake, S.R. Single-molecule sequencing of an individual human genome. *Nat. Biotechnol.* **27**, 847–850 (2009).
6. Baranzini, S.E. *et al.* Genome, epigenome and RNA sequences of monozygotic twins discordant for multiple sclerosis. *Nature* **464**, 1351–1356 (2010).
7. Lupski, J.R. *et al.* Whole-genome sequencing in a patient with Charcot-Marie-Tooth neuropathy. *N. Engl. J. Med.* **362**, 1181–1191 (2010).
8. Ju, Y.S. *et al.* Reference-unbiased copy number variant analysis using CGH microarrays. *Nucleic Acids Res.* **38**, e190 (2010).
9. Levy, S. *et al.* The diploid genome sequence of an individual human. *PLoS Biol.* **5**, e254 (2007).
10. Rasmussen, M. *et al.* Ancient human genome sequence of an extinct Palaeo-Eskimo. *Nature* **463**, 757–762 (2010).
11. Roach, J.C. *et al.* Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science* **328**, 636–639 (2010).
12. Schuster, S.C. *et al.* Complete Khoisan and Bantu genomes from southern Africa. *Nature* **463**, 943–947 (2010).
13. Wheeler, D.A. *et al.* The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452**, 872–876 (2008).
14. Iafrate, A.J. *et al.* Detection of large-scale variation in the human genome. *Nat. Genet.* **36**, 949–951 (2004).
15. Pickrell, J.K. *et al.* Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464**, 768–772 (2010).
16. Montgomery, S.B. *et al.* Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* **464**, 773–777 (2010).
17. Li, J.B. *et al.* Genome-wide identification of human RNA editing sites by parallel DNA capturing and sequencing. *Science* **324**, 1210–1213 (2009).
18. Manolio, T.A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747–753 (2009).
19. Durbin, R.M. *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
20. Wu, T.D. & Nacu, S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* **26**, 873–881 (2010).
21. Ju, Y.S., Yoo, Y.J., Kim, J.I. & Seo, J.S. The first Irish genome and ways of improving sequence accuracy. *Genome Biol.* **11**, 132 (2010).
22. Hong, D. *et al.* TIARA: a database for accurate analysis of multiple personal genomes based on cross-technology. *Nucleic Acids Res.* **39**, D883–D888 (2010).
23. Altshuler, D.M. *et al.* Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52–58 (2010).
24. Cusi, D. *et al.* Polymorphisms of alpha-adducin and salt sensitivity in patients with essential hypertension. *Lancet* **349**, 1353–1357 (1997).
25. Yoshiura, K. *et al.* A SNP in the *ABCC11* gene is the determinant of human earwax type. *Nat. Genet.* **38**, 324–330 (2006).
26. Fujimoto, A. *et al.* A scan for genetic determinants of human hair morphology: EDAR is associated with Asian hair thickness. *Hum. Mol. Genet.* **17**, 835–843 (2008).
27. Zhao, Y. *et al.* Cancer resistance in transgenic mice expressing the SAC module of Par-4. *Cancer Res.* **67**, 9276–9285 (2007).
28. Kim, J.I., Ju, Y.S., Kim, S., Hong, D. & Seo, J.S. Detection of hydin gene duplication in personal genome sequence data. *Genomics Inform.* **7**, 159–162 (2009).
29. Alkan, C., Sajjadian, S. & Eichler, E.E. Limitations of next-generation genome sequence assembly. *Nat. Methods* **8**, 61–65 (2011).
30. McClellan, J. & King, M.C. Genetic heterogeneity in human disease. *Cell* **141**, 210–217 (2010).
31. Park, H. *et al.* Discovery of common Asian copy number variants using integrated high-resolution array CGH and massively parallel DNA sequencing. *Nat. Genet.* **42**, 400–405 (2010).
32. Conrad, D.F. *et al.* Mutation spectrum revealed by breakpoint sequencing of human germline CNVs. *Nat. Genet.* **42**, 385–391 (2010).
33. Lam, H.Y. *et al.* Nucleotide-resolution analysis of structural variants using BreakSeq and a breakpoint library. *Nat. Biotechnol.* **28**, 47–55 (2010).
34. Mills, R.E. *et al.* Mapping copy number variation by population-scale genome sequencing. *Nature* **470**, 59–65 (2011).
35. Bailey, T.L. & Elkan, C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **2**, 28–36 (1994).
36. Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* **5**, 621–628 (2008).
37. Toung, J.M., Morley, M., Li, M. & Cheung, V.G. RNA-sequence analysis of human B-cells. *Genome Res.* **21**, 991–998 (2011).
38. Carrel, L. & Willard, H.F. X-inactivation profile reveals extensive variability in X-linked gene expression in females. *Nature* **434**, 400–404 (2005).
39. Wulff, B.E., Sakurai, M. & Nishikura, K. Elucidating the inosinome: global approaches to adenosine-to-inosine RNA editing. *Nat. Rev. Genet.* **12**, 81–85 (2011).
40. Levanon, E.Y. *et al.* Systematic identification of abundant A-to-I editing sites in the human transcriptome. *Nat. Biotechnol.* **22**, 1001–1005 (2004).
41. Conticello, S.G. The AID/APOBEC family of nucleic acid mutators. *Genome Biol.* **9**, 229 (2008).
42. Kiran, A. & Baranov, P.V. DARNED: a DAtabase of RNa EDiting in humans. *Bioinformatics* **26**, 1772–1776 (2010).
43. Rosenberg, B.R., Hamilton, C.E., Mwangi, M.M., Dewell, S. & Papavasiliou, F.N. Transcriptome-wide sequencing reveals numerous *APOBEC1* mRNA-editing targets in transcript 3′ UTRs. *Nat. Struct. Mol. Biol.* **18**, 230–236 (2011).
44. Knight, J.C. Allele-specific gene expression uncovered. *Trends Genet.* **20**, 113–116 (2004).
45. Pastinen, T. Genome-wide allele-specific analysis: insights into regulatory variation. *Nat. Rev. Genet.* **11**, 533–538 (2010).
46. Khaja, R. *et al.* Genome assembly comparison identifies structural variants in the human genome. *Nat. Genet.* **38**, 1413–1418 (2006).
47. Li, R. *et al.* *De novo* assembly of human genomes with massively parallel short read sequencing. *Genome Res.* **20**, 265–272 (2010).
48. Simpson, J.T. *et al.* ABySS: a parallel assembler for short read sequence data. *Genome Res.* **19**, 1117–1123 (2009).
49. Venter, J.C. *et al.* The sequence of the human genome. *Science* **291**, 1304–1351 (2001).
50. Meng, Y. *et al.* RNA editing of nuclear transcripts in *Arabidopsis thaliana. BMC Genomics* **11** (Suppl 4), S12 (2010).

ONLINE METHODS

Sample collection. All protocols of this study were approved by the institutional review board of Seoul National University Hospital (C-0806-023-246). Informed consent was obtained from all the individuals who participated in this study. Genomic DNA samples were obtained from venous blood of 18 anonymous and apparently healthy Altaic Korean individuals with at least three of generations of Korean ancestors using a Gentra Puregene Blood Kit (QIAGEN). For whole-exome sequencing of eight individuals, exon regions comprising ~180,000 exons, 700 miRNAs and 300 noncoding RNAs (with a total length of 38 Mb) were captured using the SureSelect Human All Exon Kit (ver. 1.0.1, Agilent Inc.) using the manufacturer's standard protocols. The exons are defined in the consensus coding sequence (CCDS) database (see URLs).

Genome sequencing and sequence alignment. Genomic DNA was sequenced using Illumina Genome Analyzer IIx instruments and Life Technologies SOLiD instruments (for AK2) following the manufacturer's standard protocols. For whole-genome sequencing, at least two DNA libraries were constructed to minimize the short-read redundancy of PCR duplicates, which could bias the read depth of sequencing coverage. Most Illumina sequencing libraries were constructed for paired-end sequencing (with an insert size of ~500 bp); the exception was AK1, which was reported previously[3]. Paired-end sequencing was performed for the whole genome and the whole exome by Illumina GAII sequencing, mostly as 2 × 76 bp and 2 × 101 bp runs (up to 2 × 151 bp). Short reads were aligned to the NCBI reference human genome assembly (build 36.3 and build 37.1) using the GSNAP[20] alignment program, with allowance for 5% mismatches. A single position was randomly chosen when short reads had multiple positions with identical highest alignment scores. SOLiD instruments (ver. 2 and 3) were run for whole-genome sequencing of AK2, and pipeline software (Bioscope tools) was used for sequence alignment and data analysis.

Variant calling and gene annotation. SNPs and indels were identified from the GSNAP[20] results of sequence alignment. Mismatches, insertions and deletions were counted at all positions of the human reference. To accurately identify variants, we applied a calling filter for both SNPs and indels as described previously[3]. SNPs were defined based on satisfaction of the following three conditions: (i) the number of uniquely mapped reads at the position was ≥4; (ii) the average base quality for the position was ≥20; and (iii) the allele ratio at the position was ≥20% for heterozygous SNPs and ≥90% for homozygous SNPs. Indels were obtained by the same procedure, except a homozygous indel was called if the allele ratio was ≥60%.

We created our own annotation pipeline to assign each variant to a coding sequence, untranslated region, intron or intergenic regions based on the coordinates of RefSeq genes. SNPs that altered an amino acid were considered to be non-synonymous SNPs. In addition, promoter regions were defined as the 1-kb region upstream of the start site of each gene.

Detection of large deletions. To find large deletions, we used read-depth (RD) paired-end read and split-reads information of each personal genomes and performed pairwise comparison between genomes. First, we calculated the normalized (to 25.0×) personal whole-genome read depth of coverage in a 30-bp window size as follows:

$$\text{normalized}\,\text{RD}_{\text{30-bp window, person}} = \text{RD}_{\text{30-bp window, person}} \times \frac{25.0\times}{\text{average whole-genome}\,\text{RD}_{\text{person}}}$$

Then we compared the 30-bp window coverage between two individuals as:

$$(\text{RD deviation} = \text{normalized}\,\text{RD}_{\text{30-bp window, person A}} / \text{normalized}\,\text{RD}_{\text{30-bp window, person B}})$$

using all available combinations (N = C(8,2) = 28). To be defined as a deletion candidate, the read-depth deviation should be increased (4/3 = 1.33) or decreased (3/4 = 0.75) for more than 33 windows in a row (>1 kb long). Likewise, to identify regions of homozygous deletion for all individuals considered, we also counted regions with in which the read depth for all individuals was less than 5× for more than 33 windows in a row.

For the next step, we investigated stretched paired-end reads, aligning each end onto each flanking region of large deletion candidates (defined as <1 kb from the estimated junction). We regarded deletion candidate regions with ≥2 stretched paired-end reads as suggestive.

Thereafter, we checked read-depth changes and paired-end reads near deletion candidate regions for all individuals in parallel. Regions with nearby unstable read depths, which make it difficult to determine deletion, were removed. If an individual region showed a remarkable read-depth decline compared to flanking regions or if read depths were variable among individuals, it was considered a large deletion. If any individuals showed fitting of stretched reads to the large deletion regions, all individuals who showed a clear decline in read depth for the regions were regarded as carrying the deletion in their genome (see the **Supplementary Note** for more details).

Sample preparation for RNA sequencing. For RNA extraction, immortalized lymphoblastoid cell lines were established from 17 individuals by Epstein-Barr virus transformation of mononuclear cells (Seoul Clinical Laboratories Inc.). Lymphoblastoid cell lines were cultured in RPMI 1640 media containing 15% FBS at 37 °C in a humidified 5% CO_2 environment. RNA samples were extracted using an RNAiso Plus Kit (Takara Bio Inc.). cDNA was synthesized from total RNA according to the standard protocol of Illumina Inc. for high-throughput sequencing.

Sequence alignment for transcriptome. Using GSNAP[20], we aligned short reads from transcriptome sequencing to a set of constructed mRNA sequences instead of the reference human genome to avoid mapping errors resulting from mRNA splicing (see the **Supplementary Note** for more details). We generated the mRNA sequences set using information about exons from the RefSeq, UCSC and Ensembl gene databases. All information was downloaded from the UCSC genome browser. Exons for a total of 161,250 genes were available (33,907 from RefSeq, 65,271 from UCSC and 62,072 from Ensembl). The mRNA sequences were generated from human reference genome NCBI Build 36.3 based on their exonic positions.

After mapping the short reads from transcriptome sequencing onto the set of 161,250 mRNA sequences, the mapping information for each base (read depth and type and number of mismatches) was transformed into the genomic location from the mRNA scale. Results of transcriptome sequencing, such as expression level and variants information, were obtained from this mapping information.

***De novo* assembly with unmapped reads.** To find new contigs, we conducted a *de novo* sequence assembly using read data not aligned to the reference human genome. Before merging vertices into contigs, we discarded all reads that contained any ambiguities ('N's) and those with the lowest base quality scores ('B's). *De novo* sequence assembly of the filtered read data was carried out using the ABySS[48] version 1.2.1 short-read assembler and the message passing interface protocol. To assess assembly performance of overlapping sub-string values (*k*-mer), we compared assemblies of 181.2 million paired-end reads for *k* values ranging from 25–34 bp and found the optimal size (32 bp) of *k*-mers with parameters of four coverage depths and two erode bases (see the **Supplementary Note** for more details).

Validation using PCR and Sanger sequencing. We selected 33 heterozygous SNPs and 35 indels identified in single individuals for validation by Sanger sequencing. These regions were validated using genomic DNA extracted from peripheral blood and amplified by PCR using flanking primers (**Supplementary Table 3**). For RNA modifications, we amplified genomic DNA and cDNA from 16 of these samples by PCR using flanking primers. PCR amplifications were performed in a total volume of 50 µl with 50 ng genomic DNA, 10 pmol of forward and reverse primer each and a standard volume of Ex Taq (Takara), Ex Taq buffer (Takara) and dNTPs (Takara) under the following thermocycling conditions: 95 °C for 10 min then 35 cycles of 95 °C for 30 s, 60 °C for 30 s, 72 °C for 30 s and finally 72 °C for 10 min. PCR products were purified with an AccuPrep PCR Purification Kit (Bioneer Inc.). The purified products were then validated using an ABI 3730xl DNA analyzer with ABI BigDye Terminator cycle sequencing

(Applied Biosystems). We validated four unknown transcripts using PCR and gel electrophoresis (**Supplementary Fig. 4**).

Validation of deletions using CGH arrays. We compared our deletion findings to data obtained from ultra-high–resolution CGH arrays comprising 24 million genome-wide probes[31]. We performed 24 million CGH array experiments for four females (AK4, AK6, AK14 and AK20). Before comparing our deletions to the array data, we chose deletion regions that had at least five CGH array probes in both the deletion regions and their flanking regions. To apply statistical analysis on our validation, we used a Wilcoxon rank-sum test using the $\log_2$ ratio of the CGH array data. First, for all subject deletions, we classified the $\log_2$ ratio of probes in two sets based on whether the probe is from a deletion region or from its flanking region. We applied a Wilcoxon rank-sum test to these two $\log_2$ ratio sets using the 'wilcox.test' function in R and calculated the P value of the null hypothesis for each deletion. Finally, regions that had $P < 0.05$ were defined as validated deletions.

doi:10.1038/ng.872

QUERY FORM

Nature Genetics	
Manuscript ID	[Art. Id: 872]
Author	
Editor	
Publisher	

AUTHOR:

The following queries have arisen during the editing of your manuscript. Please answer queries by making the requisite corrections directly on the galley proof. It is also imperative that you include a typewritten list of all corrections and comments, as handwritten corrections sometimes cannot be read or are easily missed. Please verify receipt of proofs via e-mail

Query No.	*Nature of Query*
Q1	Please carefully check the spellings of all author names and affiliations.
Q2	Please check that all funders have been appropriately acknowledged and that all grant numbers are correct.