Journal of Human Genetics (2013) 58, 622–626

www.nature.com/jhg

ORIGINAL ARTICLE

Complete genome sequencing and variant analysis of a Pakistani individual

Muhammad Kamran Azim[1,3], Chuanchun Yang[2,3], Zhixiang Yan[2,3], Muhammad Iqbal Choudhary[1], Asifullah Khan[1], Xiao Sun[2], Ran Li[2], Huma Asif[1], Sana Sharif[1] and Yong Zhang[2]

We sequenced the genome of a Pakistani male at 25.5x coverage using massively parallel sequencing technology. More than 90% of the sequence reads were mapped to the human reference genome. In subsequent analysis, we identified 3 224 311 single-nucleotide polymorphisms (SNPs), of which 388 532 (12% of the total SNPs) had not been previously recorded in single nucleotide polymorphism database (dbSNP) or the 1000 Genomes Project database. The 5991 non-synonymous coding variants were screened for deleterious or disease-associated SNPs. Analysis of genes with deleterious SNPs identified 'retinoic acid signaling' and 'regulation of transcription' as the enriched Gene Ontology terms. Scanning of non-synonymous SNPs against the OMIM revealed several disease and phenotype-associated variants in Pakistani genome. Comparative analysis with Indian genome sequence revealed >1.8 million shared SNPs; 32% of which were annotated in ~14 000 genes. Gene Ontology (GO) terms analysis of these genes identified 'response to jasmonic acid stimulus', 'aminoglycoside antibiotic metabolic process' and 'glycoside metabolic process' with considerable enrichment. A total of 59 558 of small indels (1–5 bp) and 16 063 large structural variations were found; 54% of which was novel. Substantial number of novel structural variations discovered in Pakistani genome enforced previous inferences that (a) structural variations are major type of variation in the genome and (b) compared with SNPs, they putatively exhibit equivalent or superior functional roles. This genome sequence information will be an important reference for population-wide genomics studies of ethnically diverse South Asian subcontinent.
Journal of Human Genetics (2013) **58,** 622–626; doi:10.1038/jhg.2013.72; published online 11 July 2013

Keywords: human genome; Pakistan; variant analysis; whole-genome sequencing

INTRODUCTION

South Asia is the home of over 1.5 billion humans, representing almost one-quarter of the world population. Early migration to this region from Africa occurred ~50 000–70 000 years before present. In recent years, genomic markers have used to study the migration patterns and relationships among different Asian ethnic groups. These efforts provided clues for two major waves of migration to South Asia from the Middle East. One wave followed a southern coastal route, around the rim of Indian subcontinent, and continued across Malaysia, Indonesia and the Philipines, whereas a distinct wave of immigrants traveled east across the Euroasian plains and turned south through the Asian mainland.[1] A recent comprehensive study carried out by HUGO Pan-Asian single-nucleotide polymorphism (SNP) Consortium[2] concluded that the southern route made a more important contribution to East and Southeast Asian populations than the northern route. Several subsequent migrations and invasions, mainly from the west, resulted in the considerable genetic diversity observed in modern South Asian populations.[1] Pakistan constitutes the north-western part of South Asia and is situated at the crossroads of Indian Subcontinent, Central Asia and the Middle East. Thus, Pakistan is located along the southern migration route.[3]

With an ethnically and linguistically diverse population of >170 million (2011 estimate: http://www.census.gov.pk), Pakistan is the sixth largest country in the world. Most of the Pakistani population has an ancestral north Indian (ANI) origin, genetically close to Middle Easterners, Central Asians and Europeans.[4] During the last decade, DNA variation among different Pakistani ethnic groups have analyzed and represented in the Human Genome Diversity Project.[1,5–10] Y-chromosomal lineage analyses and related studies have linked the Hazara and Pathan ethnic groups of Pakistan to Genghis Khan or his male ancestors[11] and Europeans[5] respectively.

Next generation DNA sequencing (NGS) technologies represent a practical way to identify and evaluate rare and previously unidentified genetic variants.[12–14] These technologies have made it

[1]Dr Panjwani Center for Molecular Medicine and Drug Research, International Center for Chemical and Biological Sciences, University of Karachi, Karachi, Pakistan and [2]BGI-Shenzhen, Shenzhen, China
[3]These authors contributed equally to this work.
Correspondence: Dr MK Azim, International Center for Chemical and Biological Sciences, University of Karachi, Karachi 75270, Pakistan.
E-mail: kamran.azim@iccs.edu
or Dr Y Zhang, BGI-Shenzhen, Beishan Road, Yantian District, Shenzhen 518083, China.
E-mail: zhangy@genomics.org.cn
Received 17 January 2013; revised 6 June 2013; accepted 7 June 2013; published online 11 July 2013

possible to develop a comprehensive catalog of genetic variation in human population samples, thereby creating a foundation for understanding human ancestry and evolution.[15,16] Large-scale studies aimed at cataloging variation, such as the 1000 Genomes Project, are currently underway. The number of genomes sequenced has grown dramatically over the last few years.[17,18] The discoveries of millions of SNPs and insertion–deletion (indel) polymorphisms indicated the necessity of whole-genome sequencing from diverse global populations to build a truly comprehensive catalog of human variation. Human genetic variation contributes a substantial fraction of disease susceptibility. The characterization of both universal and population-specific genome variation will contribute to the development of personalized medicine in the near future.[12]

Here, we report the first complete genome sequence of a Pakistani individual (designated as PK1) generated using NGS technology. Pakistan has so far been underrepresented in genome-wide surveys of human variation. The 1000 Genomes Project plans to sequence the genomes of Pakistani individuals at 2–4x coverage[17] (http://www.1000genomes.org). We sequenced the PK1 genome at >25x coverage, that is, significantly more deeply than the coverage planned by the 1000 Genomes Project. The resulting genome sequence information represents an important contribution to our knowledge of the genetic diversity of South Asia.

MATERIALS AND METHODS

Study subject and ethical statement

The study subject (designated as PK1) was a 69-year-old Pakistani male living in Karachi, Pakistan. The subject PK1 gave written informed consent to publicly disclose entire content of his genome. The Institutional Review Board of BGI approved this project after obtaining consent from the donor. Genomic DNA was isolated from peripheral blood sample using Genomic DNA isolation kit (Fermentas). Quality of the DNA was checked using 2100 Bioanalyzer (Agilent Technologies Inc. USA) and agarose gel electrophoresis. Concentration of DNA was measured with NanoDrop spectrophotometer (Thermo Inc. USA) and Qubit Fluorimeter (Life Technologies Inc. USA).

Genomic DNA library construction and genome sequencing

DNA Library preparation was carried out according to the manufacturer's instructions for sequencing on HiSeq2000 (Illumina Inc., San Diego, USA). Five microgram of genomic DNA was used for library preparation. Consequently, two paired-end libraries with insert sizes of 750 and 700 base pairs were generated for deep sequencing of PK1 genome using HiSeq 2000 (Illumina Inc.).

Data processing and read alignment

The fluorescence images were processed into sequences using the Illumina base-calling pipeline (SolexaPipeline-0.2.2.6). The human reference genome (hg19), together with the annotation of genes and repeats, was downloaded from the UCSC database (http://genome.ucsc.edu/). The SNP set of the Indian genome[19] was downloaded from web address http://krishna.gs.washington.edu/indianGenome/, and the SNP set of the 1000 Genomes Project was downloaded from website http://www.1000genomes.org/. We used SOAP (SOAPaligner version 2.20[20]) to align all short reads onto the human reference genome (hg19). To avoid misalignment, PE clusters with $\leqslant 4$ pairs were discarded.

SNPs calling. We used SOAPsnp with a statistical model based on Bayesian theory to call SNPs and the Illumina quality system to calculate the value of each possible genotype at every site. The genotype of each site was assigned as the allele types that had the highest value. The final consensus values were transformed to quality scores in Phred scale by the Illumina quality system. We used six steps to filter out unreliable portions of the consensus sequence: (1) we used a Q20 quality cutoff; (2) we required at least four reads; (3) the overall depth, including randomly placed repetitive hits, had to be <100; (4) the approximate copy number of flanking sequences had to be <2; (5) there had to be at least one paired-end read; and (6) the SNPs had to be at least 5 bp away from each other.

Indel calling. The principle of the indel calling method we used is close to alternative splice calling in transcriptome analysis (such as Tophat; http://tophat.cbcb.umd.edu/). Firstly, we selected the unmapped 75 bp reads to get the head 30 bp and the tail 30 bp to generate the PE30 reads. Then, we aligned those PE30 reads to the human reference genome (hg19) with no mismatch and no gap tolerance. If the coordinate distances of the PE reads were 1–5 bp larger or smaller than the insert size (in our case is 15 bp, 75-30-30), we considered those 75 bp reads as unmapped due to an indel. We filtered the final results by at least four hits (4 unmapped 75 bp reads with an indel).

Structural variant calling. Our two libraries were constructed with insert sizes of 750 bp and 700 bp, respectively. If the paired-end reads mapped the hg19 with the coordinate distances 3 times larger the insert size s.d. (here is 14) than the average insert size, these reads are abnormal. We grouped these reads into diagnostic paired-end (PE) clusters. To avoid misalignment, PE clusters with $\leqslant 5$ pairs were discarded. Structural variations including deletions, translocations, duplications and inversions were examined and summarized into alignment models. Reads were assembled to verify the specific coordinates of structural variation elements.

Comparative genomics and annotation of genomic variation

The entire set of genomic variation found in PK1 genome was compared with Single Nucleotide Polymorphism database (dbSNP), 1 K genome data set, OMIM and DGV (database of genomic variations). The gene annotation and genomic loci were derived from RefSeq mappings on hg19 version. Number of identified variation in gene regions was classified into (a) exonic and intronic; (b) homozygous and heterozygous variations. The exonic variants were further analyzed for potential functional effects.

Analysis of potential functional consequences of genomic variations

The non-synonymous SNPs (nsSNPs) found in PK1 genome were screened for predicting damaging effects of missense mutations using SIFT (Sorting Intolerant From Tolerant) program.[21] SIFT is a popular sequence-based amino-acid substitution prediction method available at: http://blocks.fhcrc.org/sift/SIFT.html. This program uses sequence-based predictive features to determine whether amino-acid exchanges are likely to be damaging or not. GO term enrichment analysis was carried out using GOrilla program.[22] GOrilla identifies enriched GO terms in lists of genes. It employs a flexible threshold statistical approach to discover GO terms that are significantly enriched at the top of a ranked gene list and computes an exact *P*-value according to the mHG or HG model.[22] For GO term analysis, gene lists were submitted as inputs to GOrilla server at: http://cbl-gorilla.cs.technion.ac.il with default running parameters.

RESULTS

Genome sequencing and mapping to reference genome

The individual whose genome is described in this report is Prof. Atta-ur-Rahman, who is a 69-year-old Pakistani male. The donor has no apparent genetic disorders, and his family lives in Karachi, Pakistan. Genomic DNA was subjected to sequencing using a HiSeq 2000 Genome Analyzer (Illumina Inc. San Diego, USA). Two paired-end libraries were constructed with insert sizes of 750 and 700 base pairs. A total of 78.98 Gb sequence data were generated from the two libraries (Supplementary Table 1). Using the SOAPaligner software,[20] 74.4 Gb (94%) of sequence data were aligned to the human reference genome (NCBI37/hg19). This resulted in complete coverage of the human reference genome with 25.5× sequence depth, covering about 99.5% of the human reference genome with at least one read. The individual chromosome sequence depth is shown in Supplementary Figure 1.

Identification of SNPs and analysis of variants

During SNP detection, we applied the Bayesian inference to calculate the probability and accuracy of genotypes. At each locus, the genotype with highest probability was selected as the PK1 genotype, and a quality score value was assigned as a measure of SNP call accuracy. Those loci in the PK1 consensus sequences that are polymorphic relative to the NCBI reference genome (hg19) were selected and filtered under specific criteria: quality value $\geqslant 20$, and support of the polymorphic site by at least four reads. Using the SOAPsnp software,[23] a total of 3 224 311 SNPs were detected in PK1 at an average density of 0.1%. Of these, 1,266,738 (39.2%) SNPs were identified as homozygous, whereas the remaining 1 957 573 (60.7%) were heterozygous. The chromosomal distribution of these SNPs is shown in Supplementary Table 2. Approximately one-third of all SNPs (1,031,979) were located within genes; of those, 12 896 SNPs were located in coding exon sequences. We found 876,370 SNPs in introns, of those, 153 SNPs in splice-sites. Of the 12 896 SNPs in coding exons, 6905 were synonymous and 5991 were non-synonymous substitutions; 8123 were heterozygous and the remaining 4773 were homozygous. Among the 5991 (0.18% of total SNPs) nsSNPs, 463 were novel coding variants (that is, not present in dbSNP or the 1000 Genomes Project data set). The PK1 genome has similar fraction of nsSNPs compared with Chinese, 7062 (0.23%), Watson, 7319 (0.20%) and Ventor, 6889, (0.22%).

The nsSNPs found in PK1 genome were screened for predicting damaging effects of missense mutations using SIFT program.[21] nsSNPs that lead to an amino-acid change in the protein product are of interest due to their role in protein structure–function relationship. Among the 5991 nsSNPs, 917 (15.3%) were potentially deleterious coding variants; 655 of these were heterozygous and 174 were homozygous. Examination of genes with deleterious SNPs using the GOrilla program[22] identified 'retinoic acid signaling pathway' and 'regulation of transcription' as the GO terms with enrichment among this gene set (corrected $P = 3.22 \times 10^{-4}$ and 4.27×10^{-3}, respectively). Scanning of 5991 nsSNPs against the OMIM database[24] identified 117 (1.9%) disease-associated coding variants in PK1 genome. GOrilla identified 'humoral immune response' as marginally enriched GO term in the disease-associated gene list (corrected $P = 1$). The genes involved were mannan-binding lectin serine peptidase, NOTCH2, C8A and C8B (complement component 8, α and β polypeptides).

We identified 388 532 SNPs (12% of the total PK1 SNPs) that are novel, that is, not present in dbSNP or the 1000 Genomes Project data. These novel SNPs were distributed across all chromosomes (Table 1). Of these novel SNPs, 277 859 (71.5%) were heterozygous, whereas 110 673 (28.5%) were homozygous. Further analyses revealed that 100 298 ($\sim$26%) of the novel SNPs were located in gene regions, including 1706 (0.44%) in coding exons. Among these 1706 novel coding SNPs, 731 were synonymous and 975 were non-synonymous; 1402 were heterozygous and the remaining 304 were homozygous. However, Gorilla[22] analysis of 975 novel nsSNPs containing genes did not show significant enrichment.

Table 1 The novel SNPs in PK1 genome. Chromosomal and gene region distribution of PK1 SNPs not present in dbSNP database and 1K genome data set

		SNPs in	*SNPs in*	*SNPs in coding regions*		*SNPs in*
Chromosome	*SNP*	*5-UTR*	*introns*	*Synonymous*	*Non-synonymous*	*3-UTR*
1	28 338	1123	6619	73	134	188
2	27 345	806	6433	56	74	130
3	19 643	817	5137	36	47	118
4	21 319	555	3435	13	39	67
5	17 283	478	3773	29	27	95
6	19 740	495	7200	29	37	103
7	20 733	403	5673	42	46	94
8	13 857	438	3381	13	21	89
9	17 532	374	3061	27	44	75
10	16 310	386	4347	30	34	95
11	13 972	508	3212	40	61	58
12	13 251	701	3359	37	46	100
13	8854	173	1964	15	18	25
14	9479	281	1893	19	30	38
15	10 992	246	2064	22	23	46
16	13 337	390	2198	19	27	69
17	10 644	445	2867	45	63	135
18	7926	154	1652	8	16	40
19	7668	438	1934	57	70	105
20	8646	221	1469	11	15	50
21	7 650	131	830	9	6	953
22	5635	194	1049	13	13	50
X	65 866	1778	10 360	88	83	358
Y	2512	7	48	0	1	11
Total	388 532	11 542	83 958	731	975	3092

Abbreviations: dbSNP, single-nucleotide polymorphism database; SNP, single-nucleotide polymorphism; UTR, untranslated region.

SNPs shared between individuals of Pakistani and Indian origin

Kitzman *et al.*[19] recently reported the haplotype-resolved genome sequence of a Gujarati-Indian individual. The state of Gujarat is located at the north-western part of India bordering Pakistan. Like Pakistani population, Gujaratis also have ancestral north Indian (ANI) origin. Several recent studies have examined the effects and causes of positive selection in the human genome. The accessibility of a number of entirely sequenced human genomes provided an opportunity to explore features contributing to positive selection in unprecedented detail. Therefore, a comparative analysis of PK1 and Gujarati genome sequences was carried out. Comparison between the genomes of that individual and PK1 revealed 1 825 213 shared SNPs (56% of total PK1 SNPs), of which 586 700 (32% of Pak-Indian shared SNPs) were annotated by refGene database in 14 007 gene regions. Of the shared SNPs, 101 803 were not present in the 1000 Genomes Project data. Of those novel SNPs, 24 524 SNPs are annotated in the refGene database, and 166 are non-synonymous (Tables 2 and 3).

Examination of 14 007 genes containing PK1-Indian shared SNPs using GOrilla program[22] revealed interesting correlations. Results identified seven GO terms with corrected *P*-values in the range of 10^{-3}–10^{-8} representing an array of biochemical and cellular processes (Table 4). Among these GO terms, 'response to jasmonic acid stimulus', 'aminoglycoside antibiotic metabolic process' and 'glycoside metabolic process' were identified with the strongest enrichment among this gene set (corrected $P = 1.02 \times 10^{-8}$, 2.3×10^{-6} and 4.18×10^{-6}, respectively). The next significantly enriched GO terms were 'steroid metabolic process', 'dimethylallyl diphosphate metabolic process', 'isoprenoid metabolic process' and so on. (corrected $P = 4.88 \times 10^{-4}$, 6.09×10^{-4} and 4.76×10^{-3} respectively). Interestingly, four genes of aldo-keto reductase family enzymes (that is, *AKR1C1, AKR1C2, AKR1C3* and *AKR1C4*) were involved in all of these GO terms. Two isopentenyl-diphosphate delta isomerase genes (*IDI1* and *IDI2*) were also found to be involved in some of these processes.

Table 2 Chromosomal and gene regions distribution of shared SNPs between PK1 and Gujarati-Indian genome sequences[19]

		SNPs in	*SNPs in*	*SNPs in coding regions*		*SNPs in*
Chromosome	*SNPs*	*5-UTRs*	*intron*	*Synonymous*	*Non-synonymous*	*3-UTRs*
1	5369	246	1200	14	13	28
2	4385	117	1007	8	8	16
3	3276	162	720	6	3	23
4	4520	82	588	0	0	4
5	3151	93	572	5	2	11
6	5818	87	2798	6	9	26
7	3625	64	922	4	7	5
8	2252	95	554	3	2	7
9	2439	67	472	4	3	3
10	2773	57	776	5	1	17
11	3097	125	590	10	12	10
12	2353	74	561	6	9	18
13	1651	32	261	1	2	6
14	1629	45	316	5	6	3
15	2293	39	375	4	1	4
16	1988	59	405	4	3	14
17	2665	173	601	12	17	25
18	1308	27	284	0	4	5
19	1979	104	569	10	12	27
20	1370	32	243	0	2	5
21	1175	18	153	2	1	122
22	765	20	168	1	2	8
X	41 922	1124	6,492	54	47	238
Y	0	0	0	0	0	0
Total	101 803	2942	20 627	164	166	625

Abbreviations: SNP, single-nucleotide polymorphism; UTR, untranslated region.

Moreover, 'cellular response to hydrogen peroxide' (14 genes), 'detection of chemical stimulus involved in sensory perception of smell' (four olfactory receptor genes *OR6A2, OR51B2, OR51B5* and *OR51E2*), glycolysis (5 genes) and 'regulation of insulin signal' (5 genes) were identified as enriched GO terms. Inference of these observations at population level requires more genome level studies from this region. As several of the enriched GO terms mentioned above are involved in drug metabolism, further studies may help to identify potential pharmacogenetic incompatibilities of certain drugs.

Identification of short indels

During the process of indel identification, we considered gapped alignments containing insertions or deletions of 1–5 bp. Short indels were confirmed when they were identified in both strands with a minimum of four reads. From this analysis, we identified a total of 59 558 indels in PK1 sequences, of which 32 890 were deletions and 26 668 were insertions. According to functional classification, approximately one-third of these indels (33.15%, that is, 19,746) was located within gene regions; of those, 16 609 indels were found in introns, 12 were in splice-sites. Thirty-seven indels were in coding exons and homozygous (Supplementary Table 3).

Identification of structural variants

Structural variations include deletions, insertions, inversions and other DNA sequence rearrangements. Paired-end sequencing is important for identification of large structural variants (SVs) in individual genomes relative to a reference.[25,26] We identified a total of 16 063 SVs in PK1, ranging in size from 0.1–100 kbp with an average length of 2 kb. The sum of the length of all SVs was >20 mbp (that is 20 111 213 bp) and length of majority of SVs (90%) were in the range of 500–1500 bp (Supplementary Figure 2). Of these, 8572 SVs (53% of total SVs) were not present in the DGV database (http://projects.

Table 3 Statistics of SNPs shared between PK1 and Indian[19] genome sequences

	Total	*SNPs in gene regions*	*SNPs in coding exons*	*Non-synonymous SNPs (nsSNPs)*
All SNPs	3 224 311	1 031 979	12 896	5991
Novel PK1 SNPs	388 532	100 298	1706	975
	12% of all SNPs	9.7% of SNPs in gene regions	13.2% of SNPs in coding exons	16.2% of nsSNPs
PK1-Indian shared SNPs	1 825 213	586 700	6964	3165
	56% of all SNPs	56% of SNPs in gene regions	54% of SNPs in coding exons	53% of nsSNPs
Novel PK1-Indian shared SNPs	101 803	24 524	330	166
	26.2% of novel PK1 SNPs	24.2% of novel PK1 SNPs	19.3% of novel PK1 SNPs	17% of novel PK1 SNPs

Abbreviation: SNP, single-nucleotide polymorphism.

Table 4 Gene ontology (GO) term enrichment of genes ($n = 14{,}007$) with SNVs found in both PK1 and Indian genomes. The genes AKR1C1, AKR1C2, AKR1C3 and AKR1C4 are attributed to all of the GO terms in the table

GO term	*Description*	P-*value*	*FDR* q-*value (corrected* P)	*Enrichment (N, B, n, b)*
GO:0009753	Response to jasmonic acid stimulus	1.94E-12	1.02E-8	760.27 (11404,4,15,4)
GO:0030647	Aminoglycoside antibiotic metabolic process	6.53E-10	2.3E-6	380.13 (11404,8,15,4)
GO:0016137	Glycoside metabolic process	2.38E-9	4.18E-6	304.11 (11404,10,15,4)
GO:0008202	Steroid metabolic process	3.23E-7	4.88E-4	26.37 (11404,173,15,6)
GO:0050993	Dimethylallyl diphosphate metabolic process	4.61E-7	6.09E-4	1,900.67 (11404,2,6,2)
GO:0006720	Isoprenoid metabolic process	4.96E-6	4.76E-3	60.34 (11404,54,14,4)
GO:0008207	C21-steroid hormone metabolic process	5.33E-6	4.69E-3	152.73 (11404,16,14,3)

Abbreviation: FDR, false discovery rate.
'P-value' is the enrichment *P*-value computed according to the mHG or HG model.[22] 'FDR *q*-value' is the correction of the above *P*-value for multiple testing. Namely, for the ith term (ranked according to *P*-value), the FDR *q*-value is (*P*-value * number of GO terms)/i.
Enrichment (N, B, n, b) is defined as follows: N-is the total number of genes; B-is the total number of genes associated with a specific GO term; n-is the number of genes in the top of the user's input list or in the target set when appropriate; b-is the number of genes in the intersection. Enrichment = (b/n)/(B/N)

tcag.ca/variation/), indicating the presence of novel SVs in the PK1 genome. Of the 16 063 SVs, 98.4% were large insertions or deletions. The remaining of SVs included tandem duplications (0.67%), inversions (0.062%), dispersed duplication (0.21%), and complex structures (0.60%) (Supplementary Table 4). A total of 5312 SVs were located in 2938 genes (refGene database http://hgdownload.cse.ucsc.edu/goldenPath/hg19/database/); of these, 478 were in coding exons (both in coding exons and overlapped with exons). Out of these SVs in coding exons, 17 were classified as tandem duplication; remaining were large insertions and deletions. We selected 11 SV regions with length of <1 kb for PCR validation. The fragment sizes for 8 SV regions were validated (three of them were inconclusive as the experiments were not successful) (Supplementary Figure 3).

DISCUSSION

We identified 3.22 million SNPs in PK1 genome, out of which over 0.38 million (12%) SNPs were found to be novel. Commonality of one-fourth novel SNPs in PK1 and Indian genomes indicated close relationship between these individuals (388 532 novel SNPs in PK1 versus 101 803 novel and PK1-Indian shared SNPs). Using the Markovian coalescent model applied to Chinese, European, Korean and Yoruban genome sequences, Li and Durbin[16] inferred that European and Chinese populations experienced a severe bottleneck 10 000–60 000 years before present, whereas African populations experienced a milder bottleneck from which they recovered earlier. Moreover, analyses of genome-wide SNP data sets from the CEPH Human Genome Diversity Panel samples and International HapMap Project classified the population groups studied into three genetic groups; namely Africans, Eurasians (Europeans, Middle Easterns and Central Asians including present-day Pakistan) and East Asians (also includes Americans and Oceanians).[27] The amount of variation (that is, number of SNPs and heterozygosity) we found in PK1 genome is comparable to European genome (CEU). Therefore, the present data is consistent with previous observation[27] that PK1 genome has experienced similar bottleneck like Europeans.

The study subject is of old age ($\sim$70 years) and apparently in good health. Therefore, the novel coding variants identified in this study can be linked to health status and phenotypes over the whole lifetime. As some of the PK1 coding alleles have been reported to be associated with disease, the current results may help to re-evaluate those previous reports. Moreover, our analysis showed that SVs are major type of variation in the genome. The large number of SVs identified during this study putatively having equivalent or superior functional roles than SNPs.

CONCLUSIONS

We carried out whole-genome sequencing of the Pakistani individual with 25X coverage. The present genomic data would be an important reference to add into the current deep sequenced genomes from different ethnic groups. Our analysis revealed sizeable number of unreported SNVs, short indels and structural variations. As expected deleterious non-synonymous mutations have a lower frequency than neutral variations probably due to negative selection. Human genomics can identify unknown variations associated with complex diseases widespread in South Asian subcontinent such as diabetes and cardiovascular disorders.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

Author contributions: Conceived and designed the experiments: MKA, MIC, YZ. Performed the experiments: XS, RL, HA. Analyzed the data: MKA, CY, ZY, AK. Wrote the paper: MKA, CY, ZY, AK, YZ.

Additional Information: Accession codes: SRA057506.

1 Ayub, Q. & Tyler-Smith, C. Genetic variation in South Asia: assessing the influences of geography, language and ethnicity for understanding history and disease risk. *Brief. Funct. Genomic. Proteomic.* **8,** 395–404 (2009).
2 The HUGO Pan-Asian SNP ConsortiumAbdulla, M. A., Ahmed, I., Assawamakin, A., Bhak, J., Brahmachari, S. K., Calacal, G. C. *et al.* Mapping Human Genetic Diversity in Asia. *Science* **326,** 1541–1545 (2009).
3 Hussain, J. *A history of the peoples of pakistan towards independence* (Oxford University Press, Karachi, Pakistan, 1997).
4 Reich, D., Thangaraj, K., Patterson, N., Price, A. L. & Singh, L. Reconstructing Indian Population History. *Nature* **461,** 489–494 (2009).
5 Firasat, S., Khaliq, S., Mohyuddin, A., Papaioannou, M., Tyler-Smith, C., Underhill, P. A. *et al.* Y-chromosomal evidence for a limited Greek contribution to the Pathan population of Pakistan. *Eur. J. Hum. Genet.* **15,** 121–126 (2007).
6 Siddiqi, S., Mansoor, A., Usman, S., Nasir, M., Khan, K. M. & Qamar, R. Characterization of Y-chromosomal short tandem repeat markers in Pakistani populations. *Genet. Test. Mol. Biomarkers* **15,** 165–172 (2011).
7 Mohyuddin, A., Ayub, Q., Underhill, P. A., Tyler-Smith, C. & Mehdi, S. Q. Detection of novel Y SNPs provides further insights into Y chromosomal variation in Pakistan. *J. Hum. Genet.* **51,** 375–378 (2006).
8 Mansoor, A., Mazhar, K., Khaliq, S., Hameed, A., Rehman, S., Siddiqi, S. *et al.* Investigation of the Greek ancestry of populations from northern Pakistan. *Hum. Genet.* **114,** 484–490 (2004).
9 Mohyuddin, A., Williams, F., Mansoor, A., Mehdi, S. Q. & Middleton, D. Distribution of HLA-A alleles in eight ethnic groups from Pakistan. *Tissue Antigens* **61,** 286–291 (2003).
10 Qamar, R., Ayub, Q., Mohyuddin, A., Helgason, A., Mazhar, K., Mansoor, A. *et al.* Y-chromosomal DNA variation in Pakistan. *Am. J. Hum. Genet.* **70,** 1107–1124 (2002).
11 Zerjal, T., Xue, Y., Bertorelle, G., Wells, R. S., Bao, W., Zhu, S. *et al.* The genetic legacy of the Mongols. *Am. J. Hum. Genet.* **72,** 717–721 (2003).
12 Tucker, T., Marra, M. & Friedman, J. M. Massively parallel sequencing: the next big thing in genetic medicine. *Am. J. Hum. Genet.* **85,** 142–154 (2009).
13 Li, B. & Leal, S. M. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am. J. Hum. Genet* **83,** 311–321 (2008).
14 Frazer, K. A., Murray, S. S., Schork, N. J. & Topol, E. J. Human genetic variation and its contribution to complex traits. *Nat. Rev. Genet.* **10,** 241–251 (2009).
15 DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43,** 491–498 (2011).
16 Li, H. & Durbin, R. Inference of human population history from individual whole-genome sequences. *Nature* **475,** 493–496 (2011).
17 Durbin, R. M., Abecasis, G. R., Altshuler, D. L., Auton, A., Brooks, L. D., Gibbs, R. A. *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467,** 1061–1073 (2010).
18 Drmanac, R. The advent of personal genome sequencing. *Genet. Med.* **13,** 188–190 (2011).
19 Kitzman, J. O., Mackenzie, A. P., Adey, A., Hiatt, J. B., Patwardhan, R. P., Sudmant, P. H. *et al.* Haplotype-resolved genome sequencing of a Gujarati Indian individual. *Nat. Biotechnol.* **29,** 59–63 (2011).
20 Li, R., Li, Y., Kristiansen, K. & Wang, J. SOAP: short oligonucleotide alignment program. *Bioinformatics* **24,** 713–714 (2008).
21 Kumar, P., Henikoff, S. & Ng, P. C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.* **4,** 1073–1081 (2009).
22 Eden, E., Navon, R., Steinfeld, I., Lipson, D. & Yakhini, Z. GOrilla: a tool for discovery and visualization of enriched GO terms. *BMC Bioinformatics* **10,** 48 (2009).
23 Li, R., Li, Y., Fang, X., Yang, H., Wang, J., Kristiansen, K. *et al.* SNP detection for massively parallel whole-genome resequencing. *Genome Res.* **19,** 1124–1132 (2009).
24 Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A. & McKusick, V. A. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* **33,** D514–D517 (2005).
25 Campbell, P. J., Stephens, P. J., Pleasance, E. D., O'Meara, S., Li, H., Santarius, T. *et al.* Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nat. Genet.* **40,** 722–729 (2008).
26 Wang, J., Wang, W., Li, R., Li, Y., Tian, G., Goodman, L. *et al.* The diploid genome sequence of an Asian individual. *Nature* **456,** 60–65 (2008).
27 Coop, G., Pickrell, J. K., Novembre, J., Kudaravalli, S., Li, J., Absher, D. *et al.* The role of geography in human adaptation. *PLoS. Genet.* **5,** e1000500 (2009).

Supplementary Information accompanies the paper on Journal of Human Genetics website (http://www.nature.com/jhg)