Individual Genome of the Russian Male: SNP Calling and a *de novo* Assembly of Unmapped Reads

N. N. Chekanov[2], E. S. Boulygina[1], A. V. Beletskiy[2], E. B. Prokhortchouk[1,2#*], K. G. Skryabin[1,2#]
[1] Russian Scientific Centre Kurchatov Institute
[2] Bioengineering Center of the Russian Academy of Sciences
[#]Authors contributed equally to this work (in alphabetical order)
* E-mail: prokhortchouk@biengi.ac.ru
Received 03.09. 2010

ABSTRACT A somatic cell genome was recently resequenced for a patient with renal cancer. The data were submitted to the NCBI Sequence Read Archive under the accession number SRA012240. Here, we have performed SNP calling for the genome and compared it with several published genomes. We have found 2, 921, 724 SNPs, including 1, 472, 679 newly described ones. Among them, 63, 462 SNPs have been mapped to the Y chromosome and, based on 18 markers, the genome has been ascribed to the R1a1a haplogroup predominant in Russian males. The mitochondrial haplogroup has been determined as U5a, which is also common in the European part of Russia. Short reads unmapped to the human genome were used for the *de novo* assembly of DNA sequences. This resulted in genome-specific contigs (more than 100 bp in length) with an overall length of 154 kbp (for GAII) and 4.7 kbp (for SOLiD).
KEYWORDS human genome, sequencing platform, single-nucleotide polymorphism, bioinformatics
ABBREVIATIONS SNP – single-nucleotide polymorphism, RCS – reconstructed consensus sequence

INTRODUCTION

The implementation of modern sequencing platforms has allowed widely accessible sequencing of individual genomes. In August 2010, the 1000 Genomes project [1] published (at http://www.1000genomes.org/) preliminary data on the resequencing of 2,500 individual genomes from various ethnic groups. A detailed report is expected. The general purpose of these studies is to identify frequent (with a frequency of more than 1% of the population) genome variations in human populations. Apart from fundamental problems of population genetics, the medical aspect of these studies is obvious. For example, at the end of 2009, the International Cancer Genome Consortium (ICGC) was established to investigate tumor-cell genomes [2]. Russia is affiliated with this consortium through the Russian Research Centre Kurchatov Institute, the Bioengineering Center of the Russian Academy of Sciences, and the Blokhin Cancer Research Center of the Russian Academy of Medical Sciences, which are involved in studies on renal cancer-cell genomes. The first successful resequencing of the human genome in Russia was done at the end of 2009 [3]. Libraries of short DNA reads were obtained from the genome of patient N, a Russian man suffering from renal cancer, using two sequencing platforms (SOLiD and GAII). Thus, the first genome from the Slavic population, which was never been present in the population sampling of the 1000 Genomes project, was resequenced. On the other hand, it was the first step within the framework of the renal cancer-cell genome sequencing project.

In this study we have performed a bioinformatics analysis of the data on patient N's genome resequencing directed at SNP calling. In addition, we have assembled long DNA contigs specific to patient N.

MATERIALS AND METHODS

SNP calling

Short DNA sequences that had been read on a GAII sequencer were mapped using a SOAPaligner/soap2 v.2.20 alignment program [4] with default parameters; except for the paired-end reads' insert size. The acceptable insert size range was specified as 100–700 nucleotides, based on previous data [3]. Then, SNPs were identified using the SOAPsnp v.1.02 resequencing utility [5] with default parameters. The short DNA sequences that were read on a SOLiD sequencer were mapped using a Bowtie build 0.12.5 short-read aligner [6] in a quality-aware colorspace, specifying the max mismatches in the seed as two. The acceptable insert size range was specified as 600–1,400 nucleotides, which is also in accordance with the previous data [3]. SNP calling was carried out with a SAMtools 0.1.7 package [7] using only the uniquely mapped reads.

Determination of mitochondrial and Y-chromosomal haplogroups

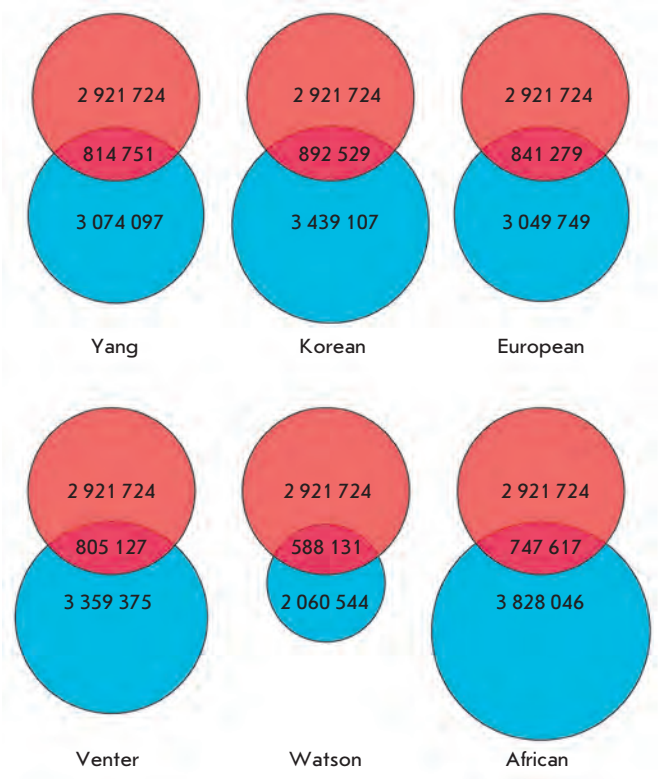
To determine the mitochondrial haplogroup, we used reads that were obtained using the SOLiD sequencer and processed with a Corona Lite package [3]. The list of mitochondrial genome SNPs, with coordinates and allele values, was acquired from the PhyloTree database (updated in August, 2010; http://www.phylotree.org/). In ascent to the mitochondrial haplogroup phylogenetic tree taken from it, we determined the allele of each distinct SNP as follows: (1) we found an allele by the specified coordinates in the RCS of mitochondrial genome, and (2) we verified these coordinates by comparing flanking sequences (no less than 10 bp from each end).

The haplogroup of the Y chromosome was determined from the reads obtained on both the GAII and SOLiD platforms and processed using the Illumina Genome Analyzer Pipeline and Corona Lite program packages, respectively [3]. The SNP list for the Y chromosome was acquired from the website http://isogg.org/ (updated in August 2010), excluding the markers that were absent in dbSNP. In ascent to the Y chromosome haplogroup tree, which was also taken from the site mentioned above, we determined the allele of each distinct SNP as follows: (1) We identified the allele in mapped nucleotide sequences from the GAII library by the coordinates of that SNP in the hg18 reference genome specified in dbSNP and verified these coordinates by comparing flanking sequences (no less than 10 bp from each end or no less than 20 bp from one end). (2) For the data from SOLiD, the allele in the Y chromosome RCS was identified by a comparison with the SNP flanking sequences acquired from dbSNP, if their size was no less than 100 bp, and RCS coverage by reads had no more than 50% gaps. The ancestral status of alleles was determined by SNP description in dbSNP.

De novo reconstruction of genome texts

We chose those reads primarily from both platforms which were not mapped to the human genome (hg18, excluding unmapped sites). The number of these sequences was 291.57 and 628.86 million for GAII and SOLiD, respectively. They were used as input data for the ABySS v.1.1.0 short read assembler [8], which offers a distributed implementation of the de Bruijn graph for the search for overlaps between k-mers (sequences whose length is k). ABySS was started several times for the optimization of the k-mer length. The optimum length of k-mer providing the longest contigs (≥200 bp) was 23 for the data from GAII and 16 from SOLiD.

Then, the sequences obtained *de novo* were mapped to the reference human genomes GRCh37 (hg19), Celera, and HuRef using the NCBI BLAST v.2.2.23 [9] with the megablast search algorithm and with enabled filtering of repeats (simple and human-specific). Sequences that were not found in any of these three reference genomes were mapped again, both to the same reference genomes and to the genomes of primates, using the discontigous megablast search algorithm.



Unique and common SNPs in different individual genomes (blue circles) compared with those of patient N (red circles).

RESULTS AND DISCUSSION

Identification of SNPs in patient N's genome

The data of patient N's genome resequencing, which was obtained using SOLiD and GAII sequencing platforms, are presented as a set of reads at the site of the National Center for Biotechnology Information (NCBI), Acc. No. SRA012240. The data had been statistically processed earlier [3]. Another immediate task of this study was to identify SNP coordinates by comparing all readings mapped to a distinct genome region (SNP calling). SNP calling was carried out separately for GAII and SOLiD data. The allele number was 1, 824, 006 and 410, 383 SNPs, respectively. The data from SOLiD were converted from the colorspace to FASTQ and combined with those from GAII, followed by the repetition of SNP calling. The total number of SNPs (2, 921, 724) exceeds the sum of SNPs identified in separate analyses

Table 1. Comparative numbers of SNPs found in different individual human genomes and the genome of patient N.

		Venter	Watson	Yang	Korean	European	African
Total SNP number		3359375	2060544	3074097	3439107	3049749	3828046
SNPs in Russian genome		1824006					
	Common SNPs	510444	365955	518294	570937	532194	479420
	One allele is the same	427096	285913	425024	457469	431977	384934
	Both alleles are the same	81957	79797	92752	113042	99967	89402
SNPs in Russian genome, SOLiD		410383					
	Common SNPs	179948	141703	187675	204235	192773	178744
	One allele is the same	116376	73735	119837	130518	125589	111031
	Both alleles are the same	27202	57292	30423	34023	33756	32133
SNPs in Russian genome, SOLiD+GAII		2921724					
	Common SNPs	805127	588131	814751	892529	841279	747617
	One allele is the same	508066	411521	486809	513621	481542	424153
	Both alleles are the same	276881	171052	307802	357562	341765	301925

Note: The data were obtained using two sequencing platforms separately and in combination

Table 2. Allele values of patient N's mitochondrial DNA known polymorphisms characterizing his affiliation with the haplogroup U5a

Haplogroup	Position	Reference allele (H2)	Diagnostic allele	SOLiD allele
L3	3594	C	C	C
N	10398	A	A	A
N	10400	C	C	C
N	10873	T	T	T
R	12705	C	C	C
UK	12308	A	G	G
U	11467	A	G	G
U5	9477	A	A	A
U5	16270	C	T	T
U5-sub	16399	A	G	G
U5a	14793	A	G	G
U5a	16256	C	T	T

of the data from each platform. This is indicative of the mutual supplementation of these two datasets in the coverage of genome regions. A comparison of allele coordinates and values was performed with the following genomes: Craig Venter [10], James Watson [11], and Huanming Yang [12], as well as genomes of a Korean [13], an African [14], and a European (CEU Trio Father NA12891 from the 1000 genomes project). The data are shown in Table 1. A comprehensive datasheet of coordinates and allele values of SNPs is shown on the site http://www.russiangenome.ru/. The figure summarizes the number of common and unique SNPs found in patient N's genome and the genomes of other individuals. We found no correlation between the resemblance of one or two equal SNP alleles (see Table. 1, rows "one allele is the same" and "both alleles are the same") and the distance between the nominal habitat of the corresponding person and Moscow, which is taken as the nominal habitat of Russians (Venter and Watson are considered Western Europeans). However, the Principal component analysis arranged individuals in accordance with the distance between their birthplaces (data not shown). The correlation is 0.89 at $p\text{-}value = 10^{-5}$.

Determination of mitochondrial and Y-chromosomal haplogroups of patient N

The identified coordinates and allele values of SNPs have made it possible to determine the mitochondrial and Y-chromosomal haplogroups of patient N's genome. Initially, we collected all reads obtained from

Table 3. Allele values of patient N's Y-chromosomal SNPs characterizing his affiliation with the haplogroup R1a1a

Haplogroup	SNP	GA allele	SOLiD allele	Ancestral allele
R	rs2032658	N/A	G	A
R	**rs17307398**	T	T	C
R	rs4481791	C	N/A	G
R	rs9786261	N/A	A	G
R	**rs891407**	G	G	C
R1	rs17307070	N/A	T	G
R1	**rs9786232**	G	G	T
R1	rs9785959	G	N/A	C
R1	rs9786197	N/A	C	T
R1	rs7067478	A	N/A	G
R1a	rs17222573	N/A	G	A
R1a	rs17307677	N/A	C	T
R1a	rs17306692	A	N/A	C
R1a1	rs17222202	N/A	A	T
R1a1	rs17316227	N/A	G	A
R1a1	rs2534636	N/A	T	T*
R1a1a	rs17222146	N/A	T	C
R1a1a	**rs17315926**	T	T	C
R1a1a	rs17221601	N/A	A	T

Note: The markers found using both sequencing platforms are drawn in bold. *rs2534636 is the back mutation for the haplogroup R1a1.

Table 4. Summary of the *de novo* reconstructed contigs that were unequivocally attributed to one of three human reference genomes.

	Not found		Found in unplaced genomic contig		Found in unlocalized genomic contig on known chromosome		Found	
	GA	SOLiD	GA	SOLiD	GA	SOLiD	GA	SOLiD
hg19	292	3	31	6	0	15	154	1
Celera	147	10	47	4	0	3	307	0
HuRef	125	9	69	8	0	0	300	0

Table 5. General statistics on *de novo* assembled contigs specific for patient N. The length of the contigs in kilobases is given in parentheses.

	GA	SOLiD
Univocally found in hg19	146 (44.7)	1 (0.3)
Simultaneously found in less than three human referense genomes	93 (27.4)	3 (0.7)
Not found in any human genome	72 (21.3)	0 (0)
Found in genomes of primates	51 (15.4)	2 (0.5)
Of them with homology > 95%	22 (6)	1 (0.2)
Total number of contigs	495 (154)	17 (4.7)

SOLiD and mapped them to the reference mitochondrial DNA (revised Cambridge Reference Sequence (rCRS); Acc. No. in GenBank: NC_012920) [15]. On the basis of these reads, an RCS was constructed and published at http://www.russiangenome.ru/. The mean coverage of the mitochondrial genome was 291. A comparison of this RCS with the reference one has shown that the mitochondrial genome of patient N belongs to the U5a haplogroup (Table. 2), one of the most common in European Russia.

The Y-chromosomal haplogroup was determined as R1a1a by four markers identified using both SOLiD and GAII and 19 markers coinciding with the data of one of two sequencing platforms (Table. 3). The coincidence of the SNP allele rs2534636 of patient N with the ancestral allele confirms the haplogroup R1a1, because this polymorphism is considered to be a result of back mutation. Since the Y chromosome is not recombinant, we can expect a high nonequilibrium coupling degree of its genetic markers. Therefore, all 63 462 SNPs identified in this work as belonging to the Y chromosome can implicitly characterize the haplotype of most men born in European Russia because of the prevalence of the R1a1a haplogroup in this region. The datasheet of all Y-chromosomal SNPs is also available at the site of the project.

De novo reconstruction of genome texts specific to patient N

The certain possibility of reconstructing a complete individual genome makes it possible to identify specific sites for a given individual. Despite the current inaccessibility of these data in the framework of the 1000 Genomes project, studies conducted by a group led by Prof. Huanming Yang at the Beijing Genomics Institute have shown that his own genome contains about 7,200 unique contigs covering about 5 million bp [16]. We have reconstructed *de novo* the unique texts of patient N's genome. All collected contigs exceeding 100 nucleotides were divided into two groups: those giving an unequivocal search result in the BLAST program (Table. 4) and those requiring additional analysis (see general statistics in Table. 5). The nucleotide sequences obtained using the SOLiD platform were insignificant both in amount and summary length. In all likelihood, this is because of the impropriety of short 25-nucleotide sequences for the reconstruction of complex genomic texts. Among the contigs collected using the GAII sequencer, the most interesting are the regions with no homology with reference human genomes, as well as those strikingly similar to genomes of primates (which have a slight difference). We can (with some degree of probability) attribute the first group of sequences to possible errors in assembling *de novo* by ABySS; however, the second group of sequences apparently cannot be the assembling errors and are characteristic of patient N. The search for open reading frames in these contigs has not revealed long (more than 30 aminoacids) coding sequences. All contigs assembled *de novo* are available at the website of the project. The difference in the number and length of contigs in the genomes of patient N and Huanming Yang can be explained by the different genome coverage (7 and 30, respectively).

Here we characterize patient N's genome compared with the reported data on other human genomes. To estimate the significance of the polymorphous and unique differences in (1) the formation of ethnic diversity and (2) the predisposition of patient N to various diseases, we need additional data on individual genomes from various ethnic groups, as well as the data obtained in associative studies using both high-density DNA chips and pangenomic sequencing. ●

This study was supported by the Federal Program Development of Russian Nanoindustry Infrastructure for 2008-2012. The authors are grateful to Prof. M.V. Kovalchuk for general assistance and for paying close attention to this work.

REFERENCES

1. Siva N. // Nat. Biotechnol. 2008. V. 26(3). P. 256.
2. Hudson T.J., Anderson W., Artez A., et al. // Nature. 2010. V. 464(7291). P. 993–998.
3. Skryabin K.G., Prokhorchuk E.B., Mazur A.M., et al. // Acta Naturae. 2009. V. 1. №3. P. 102-107.
4. Li R., Yu C., Li Y., et al. // Bioinformatics. 2009. V. 25(15). P. 1966–1967.
5. Li R., Li Y., Fang X., et al. // Genome Res. 2009. V. 19(6). P. 1124–1132.
6. Langmead B., Trapnell C., Pop M., Salzberg S.L. // Genome Biol. 2009. V. 10(3). P. R25.
7. Li H., Handsaker B., Wysoker A., et al. // Bioinformatics. 2009. V. 25(16). P. 2078–2079.
8. Simpson J.T., Wong K., Jackman S.D., et al. // Genome Res. 2009. V. 19(6). P. 1117–1123.
9. Altschul S.F., Gish W., Miller W., et al. // J. Mol. Biol. 1990. V. 215(3). P. 403–410.
10. Levy S., Sutton G., Ng P.C., et al. // PLoS Biol. 2007. V. 5(10). P. e254.
11. Wheeler D.A., Srinivasan M., Egholm M., et al. // Nature. 2008. V. 452(7189). P. 872–876.
12. Wang J., Wang W., Li R., et al. // Nature. 2008. V. 456(7218). P. 60–65.
13. Kim J.I., Ju Y.S., Park H., et al. // Nature. 2009. V. 460(7258). P. 1011–1015.
14. Bentley D.R., Balasubramanian S., Swerdlow H.P., et al. // Nature. 2008. V. 456(7218). P. 53–59.
15. Andrews R.M., Kubacka I., Chinnery P.F., et al. // Nat. Genet. 1999. V. 23(2). P. 147.
16. Li R., Li Y., Zheng H., et al. // Nat. Biotechnol. 2010. V. 28(1). P. 57–63.