Published online 4 November 2010 *Nucleic Acids Research, 2011, Vol. 39, Database issue* ***D883–D888***
doi:10.1093/nar/gkq1101

TIARA: a database for accurate analysis of multiple personal genomes based on cross-technology

Dongwan Hong[1], Sung-Soo Park[2], Young Seok Ju[1,3,4], Sheehyun Kim[1,4], Jong-Yeon Shin[1,2], Sujung Kim[2], Saet-Byeol Yu[2], Won-Chul Lee[2], Seungbok Lee[5], Hansoo Park[6], Jong-Il Kim[1,2,5,*] and Jeong-Sun Seo[1,2,3,4,5,*]

[1]Genomic Medicine Institute, Medical Research Center, Seoul National University, [2]Psoma Therapeutics Inc., [3]Department of Biochemistry and Molecular Biology, Seoul National University College of Medicine, Seoul 110-799, [4]Macrogen Inc., Seoul 153-801, [5]Department of Biomedical Sciences, Seoul National University Graduate School, Seoul 110-799, Korea and [6]Department of Pathology, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts 02115, USA

Received August 15, 2010; Revised October 14, 2010; Accepted October 18, 2010

ABSTRACT

High-throughput genomic technologies have been used to explore personal human genomes for the past few years. Although the integration of technologies is important for high-accuracy detection of personal genomic variations, no databases have been prepared to systematically archive genomes and to facilitate the comparison of personal genomic data sets prepared using a variety of experimental platforms. We describe here the Total Integrated Archive of Short-Read and Array (TIARA; http://tiara.gmi.ac.kr) database, which contains personal genomic information obtained from next generation sequencing (NGS) techniques and ultra-high-resolution comparative genomic hybridization (CGH) arrays. This database improves the accuracy of detecting personal genomic variations, such as SNPs, short indels and structural variants (SVs). At present, 36 individual genomes have been archived and may be displayed in the database. TIARA supports a user-friendly genome browser, which retrieves read-depths (RDs) and log2 ratios from NGS and CGH arrays, respectively. In addition, this database provides information on all genomic variants and the raw data, including short reads and feature-level CGH data, through anonymous file transfer protocol. More personal genomes will be archived as more individuals are analyzed by NGS or CGH array. TIARA provides a new approach to the accurate interpretation of personal genomes for genome research.

INTRODUCTION

Recently developed high-throughput DNA technologies have revolutionized human genomics. Massively parallel sequencing—next generation sequencing (NGS)—has been used to analyze nearly 20 personal genomes (1–10). The cost of sequencing a single genome is decreasing dramatically, and we are now approaching an era in which personal genomic sequencing will cost US$1000. The sequencing of a large number of individual genomes, possibly more than 1000, is expected to be complete within the next year (http://www.1000genomes.org). Current sequencing technologies, which provide sufficient read depth (RD), enable the detection of genome-wide SNPs and short indels with >99.9% accuracy (3,4).

Comparative genomic hybridization (CGH) arrays have been used to detect copy number variants (CNVs), a major type of structural variant (SV) in the human genome (11–16). CNVs are irregular in size and often reside in ambiguous regions (e.g. repetitive sequences) making them difficult to detect by NGS technologies alone. Although several sequencing approaches have attempted to detect CNVs (6,17–19), CGH arrays remain a standard approach to CNV detection (11–16).

Human genomic variants are believed to have important functional impacts on human biology and medicine. To evaluate the potential biological functions of the large number of variants, it is essential to develop intuitive

*To whom correspondence should be addressed. Tel: +82 2 740 8246; Fax: +82 2 741 5423; Email: jeongsun@snu.ac.kr
Correspondence may also be addressed to Jong-Il Kim. Tel: +82 2 740 8421; Fax: +82 2 741 5423; E-mail: jongil@snu.ac.kr

These authors wish it to be known that, in their opinion, the first three authors should be regarded as joint First Authors.

Table 1. Summary of massively parallel sequencing data in TIARA

Sample name	Technology	Read length (in bp)	Insert size	Number of reads	Total bases	Sequencing coverage	Aligned coverage	SNPs	Indels	CNV (region)
AK1	Illumina Genome Analyzer	1 × 36	200	519 486 218	18 701 503 848	35.9x	27.8x	3 453 653	170 202	1237 (24 193 059)
		2 × 36		1 646 543 336	59 275 560 096					
		2 × 88		123 322 768	10 852 403 584					
		2 × 106		177 416 122	18 806 108 932					
AK2	AB SOLiD	2 × 25	1500	6 371 995 780	159 299 894 500	109.6x	27.5x	3 586 271	213 718	607 (9 248 044)
		2 × 50	4700	3 390 922 334	169 546 116 700					
AK4	Illumina Genome Analyzer	2 × 76	500	444 312 562	33 767 754 712	25.7x	23.1x	3 630 428	429 258	696 (8 463 889)
		2 × 101		430 032 812	43 433 314 012					
AK6	Illumina Genome Analyzer	2 × 36	500	55 752 362	2 007 085 032	24.5x	22.3x	3 558 703	413 949	706 (11 958 848)
		2 × 76		540 079 624	41 046 051 424					
		2 × 101		301 478 526	30 449 331 126					
NA10851	Illumina Genome Analyzer	2 × 36	500	1 114 121 056	40 108 358 016	28.3x	25.0x	3 683 016	319 266	1309 (23 198 937)
		2 × 76		318 924 496	24 238 261 696					
		2 × 101		203 842 434	20 588 085 834					

methods for comparing multiple genomes using raw-level data generated by diverse technologies. Moreover, the cooperative integration of different genomic technologies is necessary for high-accuracy detection of variants, especially of CNVs (6,10,16). Although many genomic databases and browsers have been developed (20–24), the comparison and integration of genomic data sets from different platforms is not yet feasible.

We describe here a new database, the Total Integrated Archive of Short-Read and Array (TIARA; http://tiara.gmi.ac.kr) database, integrated with a genome browser. This database accumulates raw-level personal genomic data from whole genome NGS and CGH arrays. At present, it contains 36 individual genomic data sets that have been analyzed and reported by the Genomic Medicine Institute (GMI) at Seoul National University (6,10,16). To retrieve the large quantities of genomic data in real-time on TIARA, we have implemented an efficient index using Apache Lucene (http://lucene.apache.org) along with client-side Asynchronous JavaScript and XML (AJAX) scripts that reduce the volume of data exchange and processing within the web server.

MATERIALS AND METHODS

Massively parallel sequencing data

TIARA contains massively parallel sequencing data from five individuals, three of whom—[AK1 (6), AK2 (16), and NA10851 (10,16)]—have been described previously. The other two genomes deposited in TIARA, AK4 and AK6, were sequenced using the Illumina Genome Analyzer. The average RDs of the sequencing coverage for these five individuals were 27.8x, 27.5x, 25.0x, 22.3x, and 23.1x, respectively. The details of the whole genome sequencing process have been described previously (6,10,16). Briefly, short-reads from the Illumina Genome Analyzer and AB SOLiD were aligned using the GSNAP and BioScope alignment tools, respectively, with respect to the human reference genome build 36.3 (6,16,25). The RDs of sequencing coverage were obtained by adjusting the effects of GC content as described previously (10,18).

High-resolution CGH array data

CGH array data from 33 individuals (11 Koreans, 10 Chinese, 10 Japanese, 1 European and 1 West African) were obtained using a whole genome tiling CGH array comprising 24-M probes (16) (Supplementary Table S1). In addition to the usual type of CGH data, which depends on a comparison with a reference sample (NA10851), the absolute or reference-free CGH array data were also provided.

Genome variants

SNPs and indels were discovered by applying conservative filter criteria to the NGS data as described elsewhere (6). Briefly, four matches from uniquely aligned short reads with a quality score $\geq$20 were required for SNP identification. CNVs were identified in the CGH array using the ADM2 algorithm (16,26) in the Agilent Genomic Workbench Standard Edition 5.0.14. The summary statistics of each individual genome are provided in Table 1.

RESULTS

System configuration

The TIARA system mainly consists of a 'genome data repository' and a 'genome browser' (Figure 1). The genome data repository has three types of storage archive: (i) a 'Lucene index file system', (ii) a 'MySQL database' and (iii) an 'anonymous file transfer protocol (FTP) archive'. These archives were built on a virtualization file system designed to support high-performance computing clusters. The Lucene index file system includes inverted index files for real-time query processing

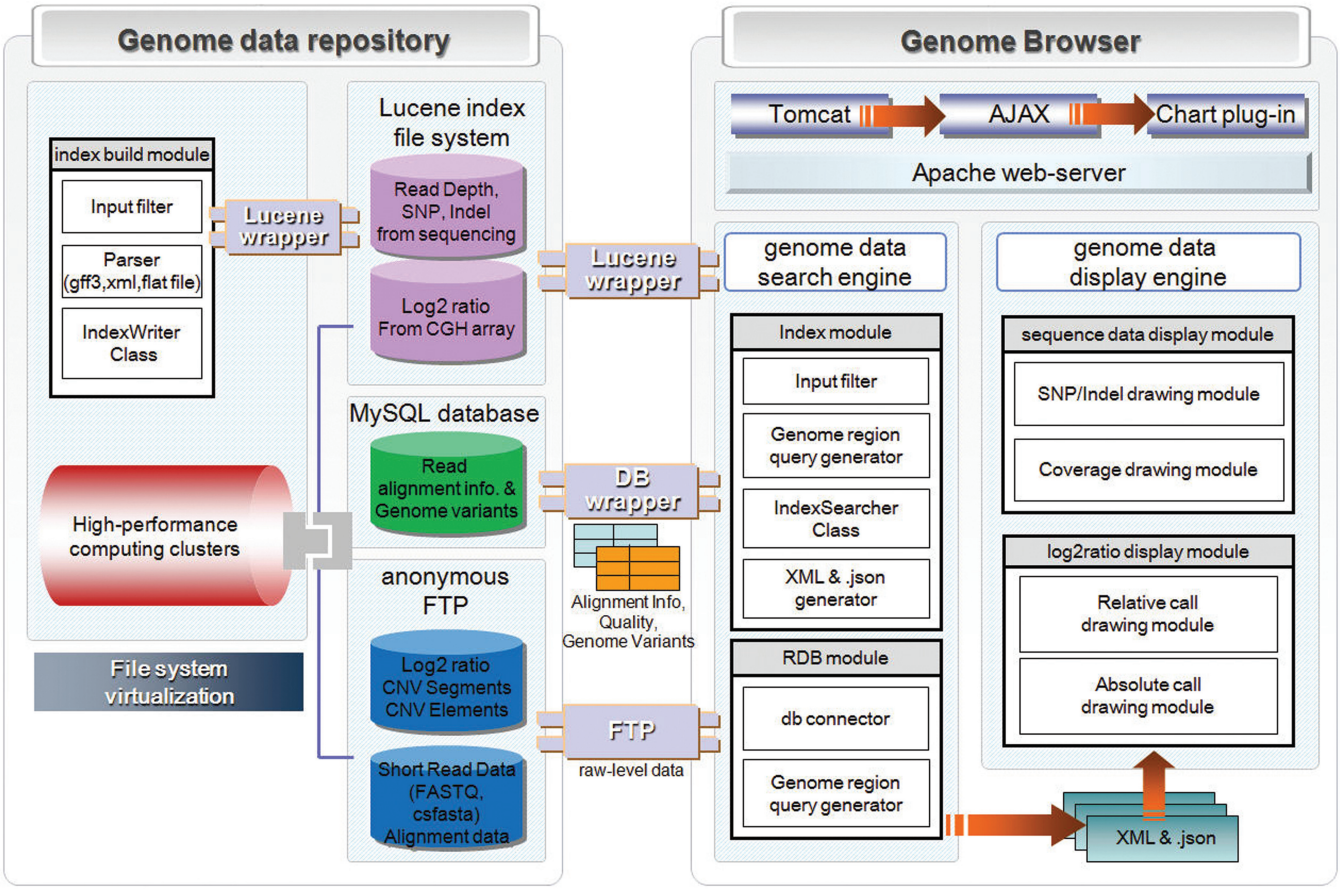

Figure 1. System configuration of TIARA.

of genomic data, including SNPs, indels, RDs and log2 ratios. Inverted index files are generated using an 'index build module'. The MySQL database stores information about the aligned short reads, such as read length, alignment position and quality. The anonymous FTP archive enables downloading of the raw CGH and short read data as well as the filtered genome variants, including SNPs, non-synonymous SNPs, indels and CNVs.

The genome browser consists of: (i) 'a genome data search engine' and (ii) 'a genome data display engine'. The genome data search engine retrieves genomic data from the genome data repository. The genome data display engine visualizes the SNPs, indels, RDs and log2 ratios obtained from the genome data search engine in the genome browser of TIARA. The properties of the genome data that are exchanged between modules are XML and JavaScript Object Notation (JSON) files. The modules in each engine are described in detail in the Supplementary Information 1.

User interface of TIARA

In this section, we describe the user interface of TIARA, the structure of which is displayed in Figure 2a. In area (A) of Figure 2a, the user can specify the genomic region and individual regions of interest for browsing. Areas (B), (C), (D) and (E) present, respectively, the RefSeq gene, SNPs, indels and RDs from the high-throughput sequencing data. Areas (F) and (G) present the CNV regions and log2 ratios from the high-resolution CGH array data, respectively. Once the user selects or deselects an individual genome data set, the personal genome data are displayed in or removed from areas (C), (D), (E) and (G). The 'GeneSearch' button allows the user to browse the genome data for a specific gene selected by the user. For example, the user can browse the TP53 gene locus (Figure 2b). The 'XMLDownload' button exports an XML document that contains structured information describing the SNPs, indels, RDs and log2 ratios visualized in the genome browser. The downloaded XML document permits analysis of the selected genomic region using other genomic browsers or custom scripts of the user's creation. A schema of each XML document is shown in the Supplementary Figures 1 and 2.

(i) High-throughput sequencing data
  (a) SNP display window (see (C) of Figure 2a)
  All SNPs detected across multiple individuals for a selected genomic region are displayed as points in the SNP display window. Homozygous and heterozygous SNPs are colored in blue and red, respectively. Users may click on one of these SNPs to receive information about the short read data. For example, Figure 2c shows the information on short reads for

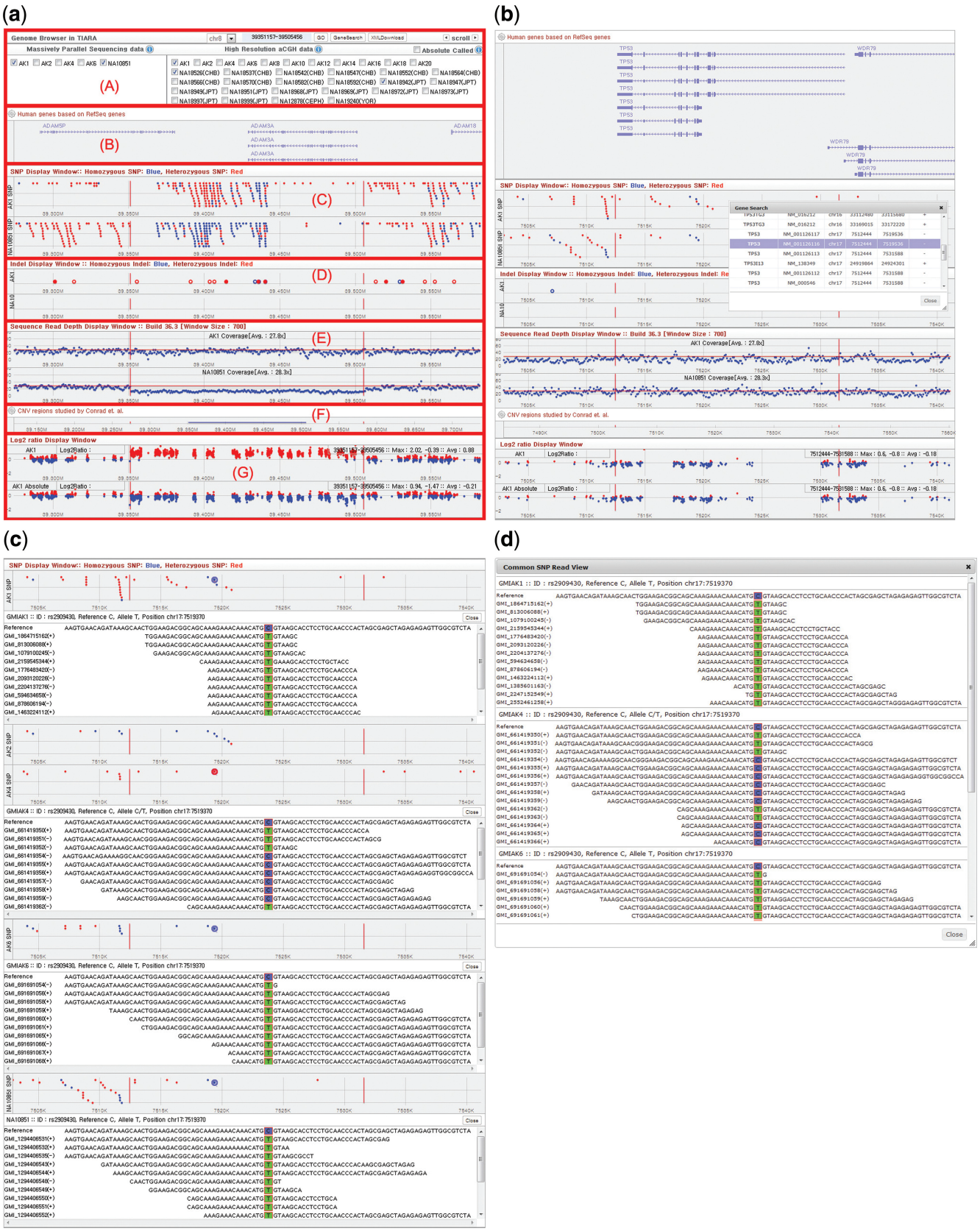

Figure 2. User interface of TIARA. (**a**) The user interface consists of areas (A) control panel, (B) Refseq gene, (C) SNPs, (D) indels, (E) RD display window from high-throughput sequencing, (F) CNV regions and (G) log2 ratio display window for high-resolution CGH array data. (**b**) Retrieval of genomic data for the TP53 gene using the gene name search function. (**c**) An example of heterozygous and homozygous SNPs for the same position in several selected individuals. These SNPs are related to colorectal and endometrial cancer. (**d**) An example of the popup window for the common SNPs displaying.

the SNP at the 74 583 581 bp position of chromosome 14. A comparison of multiple genomes provides clues as to the functional impact of each variant. For example, the SNP shown in Figure 2c appears to have a high frequency because five individuals have the SNP (homozygous SNPs for AK1 and AK6 and heterozygous SNPs for AK2, AK4 and NA10851). Although the SNP is associated with colorectal and endometrial cancers according to reports in the Online Mendelian Inheritance in Man (OMIM) database, its high allele frequency suggests that its disease effect may be limited. In addition, we provide the popup window that is available to display the common SNPs from multiple individuals in Figure 2d.

(b) Indel display window (see area (D) of Figure 2a)
The indel start position is marked with a circle. As with the SNPs, homozygous and heterozygous indels are colored in blue and red, respectively. The insertions are indicated by a filled circle and the deletions are indicated by an open circle.

(c) RD display window (see area (E) of Figure 2a)
The coverage graph for the genomic region selected by the user is drawn in the RD display window, the size of which is adjusted according to the amount of data extracted from that region of the genome.

(ii) High-resolution CGH array data

(a) CNV region display window (see area (F) of Figure 2a)
To enhance the CNV study, CNV regions reported by Conrad *et al*. (15) were browsed in area (F). A line indicates the CNV region from the start position to the stop position.

(b) Log2 ratio display window (see area (G) of Figure 2a)
The log2 ratios of the CGH array for individuals selected by the user may be visualized in the log2 ratio display window. Both the conventional log2 ratio (relative to the CGH reference DNA, NA10851) and the reference-free log2 ratio (10) are displayed if the 'Absolute called' checkbox is selected. By combining the RDs, relative log2 ratios and absolute log2 ratios, the copy number status of a selected region from individual genomes can be identified accurately. For example, as shown in Figure 2a, although the relative log2 ratio in the genomic region of AK1 appears to be a gain (G), it is apparent that AK1 has no CNVs because both the RD (E) and the absolute log2 ratio (G) indicate that not AK1, but NA10851 contains CNVs in the region.

DISCUSSION

We have described the development of the TIARA genome database, into which massively parallel sequencing data, high-resolution array CGH data and genomic variants of human whole genomes have been deposited. The TIARA genome browser is a unique visualization tool that facilitates multi-individual and cross-technology analysis of complex human genomic variations. TIARA will be upgraded to improve the efficiency of genome research by developing advanced genome browser functions and by adding more personal genomes. GMI-SNU has recently completed sequencing of the entire genomes of 10 Korean individuals using NGS and high-resolution CGH arrays. Our group plans to analyze 1000 Asian genomes and release the data through TIARA before the end of the next year. We believe that TIARA and the genomic data will prove to be an invaluable resource for human genome research.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

Korean Ministry of Knowledge Economy (grant number 0411-20100061); Korean Ministry of Education, Science and Technology (grant number 2010-0013662); Green Cross Therapeutics (0411-20080023). Funding for open access charge: Korean Ministry of Education, Science and Technology (grant number 2010-0013662).

Conflict of interest statement. None declared.

REFERENCES

1. Levy,S., Sutton,G., Ng,P.C., Feuk,L., Halpern,A.L., Walenz,B.P., Axelrod,N., Huang,J., Kirkness,E.F., Denisov,G. *et al.* (2007) The diploid genome sequence of an individual human. *PLoS Biol*., **5**, e254.
2. Wheeler,D.A., Srinivasan,M., Egholm,M., Shen,Y., Chen,L., McGuire,A., He,W., Chen,Y.J., Makhijani,V., Roth,G.T. *et al.* (2008) The complete genome of an individual by massively parallel DNA sequencing. *Nature*, **452**, 872–876.
3. Wang,J., Wang,W., Li,R., Li,Y., Tian,G., Goodman,L., Fan,W., Zhang,J., Li,J., Guo,Y. *et al.* (2008) The diploid genome sequence of an Asian individual. *Nature*, **456**, 60–65.
4. Bentley,D.R., Balasubramanian,S., Swerdlow,H.P., Smith,G.P., Milton,J., Brown,C.G., Hall,K.P., Evers,D.J., Barnes,C.L., Bignell,H.R. *et al.* (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, **456**, 53–59.
5. McKernan,K.J., Peckham,H.E., Costa,G.L., McLaughlin,S.F., Fu,Y., Tsung,E.F., Clouser,C.R., Duncan,C., Ichikawa,J.K., Lee,C.C. *et al.* (2009) Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Res*., **19**, 1527–1541.
6. Kim,J.I., Ju,Y.S., Park,H., Kim,S., Lee,S., Yi,J.H., Mudge,J., Miller,N.A., Hong,D., Bell,C.J. *et al.* (2009) A highly annotated whole-genome sequence of a Korean individual. *Nature*, **460**, 1011–1015.
7. Pushkarev,D., Neff,N.F. and Quake,S.R. (2009) Single-molecule sequencing of an individual human genome. *Nat. Biotechnol*., **27**, 847–852.
8. Drmanac,R., Sparks,A.B., Callow,M.J., Halpern,A.L., Burns,N.L., Kermani,B.G., Carnevali,P., Nazarenko,I., Nilsen,G.B., Yeung,G. *et al.* (2010) Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science*, **327**, 78–81.
9. Snyder,M., Du,J. and Gerstein,M. (2010) Personal genome sequencing: current approaches and challenges. *Genes Dev*., **24**, 423–431.
10. Ju,Y.S., Hong,D., Kim,S., Park,S.S., Lee,S., Park,H., Kim,J.I. and Seo,J.S. (2010) Reference-unbiased copy number variant

analysis using CGH microarrays. *Nucleic Acids Res.*, doi:10.1093/nar/gkq730.
11. Iafrate,A.J., Feuk,L., Rivera,M.N., Listewnik,M.L., Donahoe,P.K., Qi,Y., Scherer,S.W. and Lee,C. (2004) Detection of large-scale variation in the human genome. *Nat. Genet.*, **36**, 949–951.
12. Redon,R., Ishikawa,S., Fitch,K.R., Feuk,L., Perry,G.H., Andrews,T.D., Fiegler,H., Shapero,M.H., Carson,A.R., Chen,W. *et al.* (2006) Global variation in copy number in the human genome. *Nature*, **444**, 444–454.
13. McCarroll,S.A. and Altshuler,D.M. (2007) Copy-number variation and association studies of human disease. *Nat. Genet.*, **39**, S37–S42.
14. Perry,G.H., Ben-Dor,A., Tsalenko,A., Sampas,N., Rodriguez-Revenga,L., Tran,C.W., Scheffer,A., Steinfeld,I., Tsang,P., Yamada,N.A. *et al.* (2008) The fine-scale and complex architecture of human copy-number variation. *Am. J. Hum. Genet.*, **82**, 685–695.
15. Conrad,D.F., Pinto,D., Redon,R., Feuk,L., Gokcumen,O., Zhang,Y., Aerts,J., Andrews,T.D., Barnes,C., Campbell,P. *et al.* (2010) Origins and functional impact of copy number variation in the human genome. *Nature*, **464**, 704–712.
16. Park,H., Kim,J.I., Ju,Y.S., Gokcumen,O., Mills,R.E., Kim,S., Lee,S., Suh,D., Hong,D., Kang,H.P. *et al.* (2010) Discovery of common Asian copy number variants using integrated high-resolution array CGH and massively parallel DNA sequencing. *Nat. Genet.*, **42**, 400–405.
17. Korbel,J.O., Urban,A.E., Affourtit,J.P., Godwin,B., Grubert,F., Simons,J.F., Kim,P.M., Palejev,D., Carriero,N.J., Du,L. *et al.* (2007) Paired-end mapping reveals extensive structural variation in the human genome. *Science*, **318**, 420–426.
18. Yoon,S., Xuan,Z., Makarov,V., Ye,K. and Sebat,J. (2009) Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome Res.*, **19**, 1586–1592.
19. Chiang,D.Y., Getz,G., Jaffe,D.B., O'Kelly,M.J., Zhao,X., Carter,S.L., Russ,C., Nusbaum,C., Meyerson,M. and Lander,E.S. (2009) High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nat. Methods*, **6**, 99–103.
20. Rhead,B., Karolchik,D., Kuhn,R.M., Hinrichs,A.S., Zweig,A.S., Fujita,P.A., Diekhans,M., Smith,K.E., Rosenbloom,K.R., Raney,B.J. *et al.* (2010) The UCSC Genome Browser database: update 2010. *Nucleic Acids Res.*, **38**, D613–D619.
21. Li,G., Ma,L., Song,C., Yang,Z., Wang,X., Huang,H., Li,Y., Li,R., Zhang,X., Yang,H. *et al.* (2009) The YH database: the first Asian diploid genome database. *Nucleic Acids Res.*, **37**, D1025–D1028.
22. Axelrod,N., Lin,Y., Ng,P.C., Stockwell,T.B., Crabtree,J., Huang,J., Kirkness,E., Strausberg,R.L., Frazier,M.E., Venter,J.C. *et al.* (2009) The HuRef Browser: a web resource for individual human genomics. *Nucleic Acids Res.*, **37**, D1018–D1024.
23. Shumway,M., Cochrane,G. and Sugawara,H. (2010) Archiving next generation sequencing data. *Nucleic Acids Res.*, **38**, D870–D871.
24. Kersey,P.J., Lawson,D., Birney,E., Derwent,P.S., Haimel,M., Herrero,J., Keenan,S., Kerhornou,A., Koscielny,G., Kahari,A. *et al.* (2010) Ensembl Genomes: extending Ensembl across the taxonomic space. *Nucleic Acids Res.*, **38**, D563–D569.
25. Wu,T.D. and Nacu,S. (2010) Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics*, **26**, 873–881.
26. Lipson,D., Aumann,Y., Ben-Dor,A., Linial,N. and Yakhini,Z. (2006) Efficient calculation of interval scores for DNA copy number data analysis. *J. Comput. Biol.*, **13**, 215–228.