Published online 25 January 2011

Nucleic Acids Research, 2011, Vol. 39, No. 7 ***e45***
doi:10.1093/nar/gkr009

A novel compression tool for efficient storage of genome resequencing data

Congmao Wang[1] and Dabing Zhang[1,2,*]

[1]School of Life Sciences and Biotechnology and [2]Bio-X Center, Key Laboratory of Genetics & Development and Neuropsychiatric Diseases, Ministry of Education, Shanghai Jiao Tong University, Shanghai 200240, China

Received November 4, 2010; Revised December 13, 2010; Accepted January 3, 2011

ABSTRACT

With the advent of DNA sequencing technologies, more and more reference genome sequences are available for many organisms. Analyzing sequence variation and understanding its biological importance are becoming a major research aim. However, how to store and process the huge amount of eukaryotic genome data, such as those of the human, mouse and rice, has become a challenge to biologists. Currently available bioinformatics tools used to compress genome sequence data have some limitations, such as the requirement of the reference single nucleotide polymorphisms (SNPs) map and information on deletions and insertions. Here, we present a novel compression tool for storing and analyzing Genome ReSequencing data, named GRS. GRS is able to process the genome sequence data without the use of the reference SNPs and other sequence variation information and automatically rebuild the individual genome sequence data using the reference genome sequence. When its performance was tested on the first Korean personal genome sequence data set, GRS was able to achieve ~159-fold compression, reducing the size of the data from 2986.8 to 18.8 MB. While being tested against the sequencing data from rice and *Arabidopsis thaliana*, GRS compressed the 361.0 MB rice genome data to 4.4 MB, and the *A. thaliana* genome data from 115.1 MB to 6.5 KB. This *de novo* compression tool is available at http://gmdd.shgmo.org/Computational-Biology/GRS.

INTRODUCTION

The development of new DNA sequencing technologies, such as next-generation sequencing (NGS) and single-molecule sequencing, has enabled the research of genomics and functional genomics to advance to new levels (1,2). Due to the dramatic reduction of sequencing cost and increase of sequencing efficiency, these new high-throughput sequencing technologies have become effective and routine applications in the 'resequencing' of individual genomes for detecting sequence variation between the individual and the reference genome (3). Resequencing individual genomes can facilitate the investigation of the relationship between sequence and phenotypic variations. To date, several personal human genome sequencing data have been released (2,4–6). Sequencing of individual human genomes is believed to provide molecular basis for personalized medicine. Furthermore, more resequencing data are being generated from various organisms. Individual genome resequencing in animals such as mouse and pig, and in plants such as rice, maize and soybean, has proven to be extremely powerful in investigating genome variations, such as single nucleotide polymorphisms (SNPs), deletions, insertions and rearrangements.

However, how to store and manage the huge amount of sequencing data has become a challenge to biologists. For example, the storage of one 2009 human reference genome (i.e. UCSC hg19 assembly) requires up to 905 MB with the tar.gz compression format (7). Thus, a total of 90 500 MB (nearly 88.38 GB) hard disk storage space with the tar.gz compression format would be required to store data for 100 individuals in genetic disease studies. It is noteworthy that different individuals within one species share higher consensus nucleotide sequence; for example, human has ~99.9% common genome sequence with DNA sequencing errors of 0.01% (8,9). Moreover, the electronic transfer of sequencing data is a bottleneck, even though some tools have been developed to compress the files and increase the network bandwidth.

Currently, several methods for compressing genomic sequence data have been reported (10–13). However, these compression tools can not process the genome sequence data without the reference SNPs map or information about sequence variations, such as insertions or deletions.

Here, we present a general Genome ReSequencing (GRS) tool for storing and managing the individual genome resequencing data without having to rely on

*To whom correspondence should be addressed. Tel: +86 021 34205074; Fax: +86 021 34204869; Email: zhangdb@sjtu.edu.cn

known reference SNPs maps or other information on sequence variation. We demonstrate the power of this GRS tool in processing genome resequencing data, using whole genome sequencing data sets from human and the model plants *Arabidopsis* and rice.

MATERIALS AND METHODS

Data set

Data sets used to test GRS include KOREF_20090131 and KOREF_20090224, two of the first Korean personal genome sequences (4), and two different versions of the reference genome sequence from *Arabidopsis thaliana* (TAIR8 and TAIR9) and rice (TIGR5 and TIGR6) (14–16), respectively.

Software availability

GRS is implemented in C and Shell. It will be freely available for non-profit use. The source code and its executable file are available at http://gmdd.shgmo.org/Computational-Biology/GRS.

Architecture of the GRS tool

The main modules in GRS connect the input chromosome file, the intermediate data and the final compressed file. The architecture of GRS is shown in Figure 1. When an individual genome sequence data needs to be compressed, GRS first evaluates each chromosome varied sequence percentage (δ) based on the reference chromosome. Then it filters the longest identical nucleotide sequence and extracts the different sequence ($\delta \leq 0.03$), and precodes the different sequence file that has been generated to reduce the file size. Then, GRS uses the Huffman coding strategy to compress the reduced different sequence file to the bz2 type file and generates the command file to decompress the compressed file. If $0.03 \leq \delta \leq 0.1$, GRS will cut chromosome into n pieces and calculate each different rate δ_i $(1 \leq i \leq n)$ to find the position with minimal $\sum\delta_i$, then compress each piece with the same strategy as the one used for $\delta \leq 0.03$. Individual genome sequence data that has

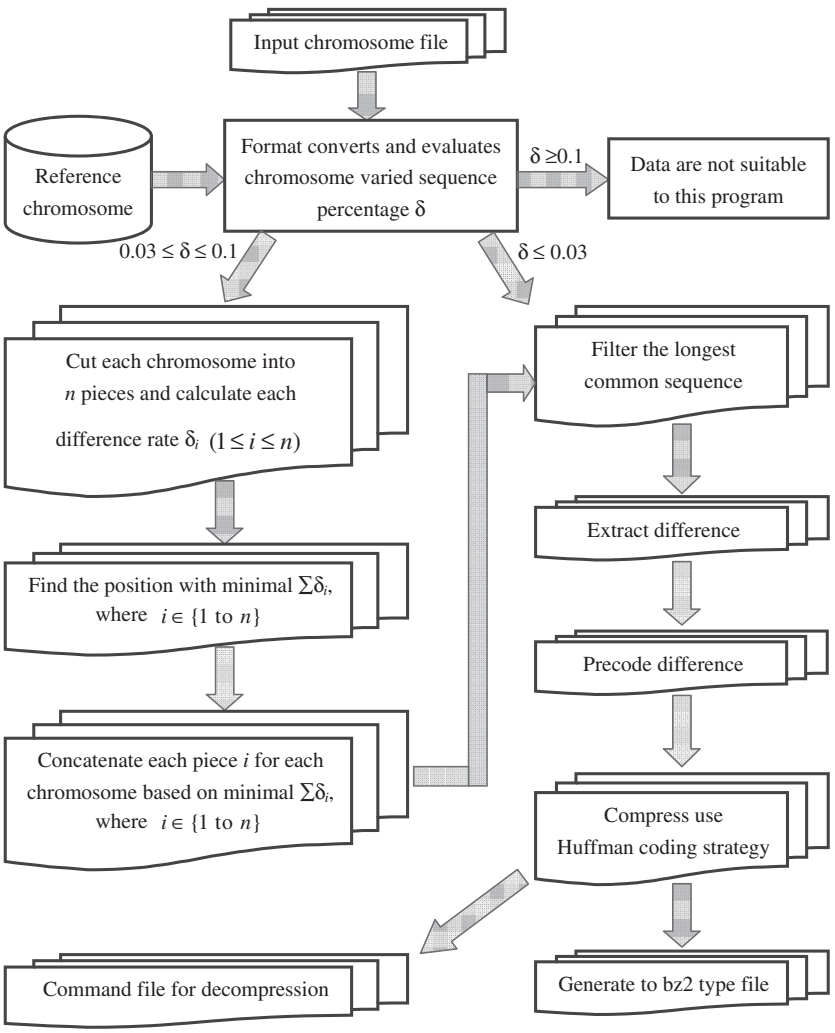

Figure 1. Architecture of the GRS Tool. The main modules in GRS connect the input chromosome file, the intermediate data and the final compressed file. Details of the processing procedure are described in the text.

been compressed using GRS can be later decoded easily with the GRS decoding tool.

Evaluation of individual genome sequence variation

Higher percentage of nucleotide sequence variation from the reference genome leads to longer time to run GRS and results in larger compressed file for an individual genome. When the individual genome sequence data needs to be compressed, GRS checks whether the users use the correct reference chromosome data, quantifying the percentage of varied nucleotide sequence in an individual genome is thus required. Here we used the following method to calculate δ. We used the formula $\delta = \sum_{i=0}^{9} c_i/t$, where c_i means the number of different nucleotide sequence between the individual and the reference; i means the type of nucleotide; including A, T, C, G, N, a, t, c, g and n; t means the total DNA base number in the reference genome sequence data.

Recording the longest common local nucleotide sequence and the changed sequence

It is reported that finding and recording the varied nucleotide sequence of two sequences equals to finding their longest common local sequence (17). Using a matrix graph, the longest common sequence can be extracted effectively. Taking two sequences 'gaNGCTA' and 'gNGTNA' as an example, their longest common sequence is 'gNGTA'. That is to say that each nucleotide of 'gaNGCTA' in x-axis is used to compare with the whole sequence of 'gNGTNA' in y-axis, and each common nucleotide in y-axis direction will be marked with a red circle, after the base by base comparison the longest common sequence will be marked in the whole matrix (Supplementary Figure S1). GRS can find the minimal changes between two genome sequences using the modified UNIX diff program (18).

RESULTS

Huffman encoding for varied nucleotide sequence information and individual genome sequence rebuilding

Figure 2a shows the module of presenting the raw information on sequence difference between the individual and the reference sequences generated based on the modified UNIX diff program. When processing the varied sequence information, the '>', '<' and the '\n' between adjacent nucleotides is removed by GRS (Figure 2b). Then 'a' (add) is converted to 'i' and 'c' (change) to 'h'. In addition, the base information below the 'd' (deletion) can be removed since the deleted sequence information can be extracted based on the sequence position such as N5 and N6 (Figure 2a and b). Also, the '—' and the highlighted bases can be removed because these sequences can be recovered by using the sequence at N9 and N10 (Figure 2a and b).

Next, ',' is changed to '\t' and '\t' is added to each side of 'i', 'd' and 'h' by GRS to make the rebuilding language more readable by the computer. If there are two numbers at the side of 'i', 'd' or 'h', the second nucleotide position

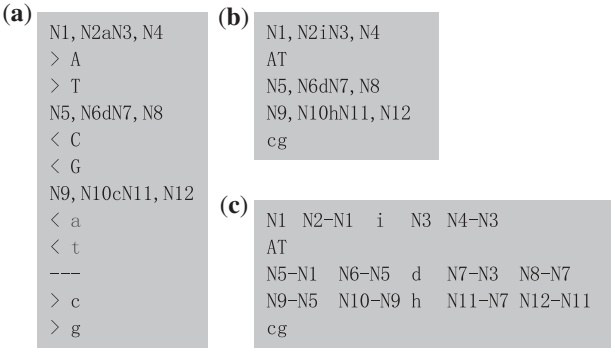

Figure 2. Processing changes file of DNA base, genome position and recover language. (**a**) Raw changes between two sequences generated based on the modified UNIX diff program. N1 to N12 indicate the nucleotide sequence position ranging from N1 to N12; 'a' is the insertion of nucleotide(s); 'd' is the deletion of nucleotide(s) and 'c' is the changed nucleotide(s). In addition, symbol ',' between N1 and N2 means positions start from N1 to N2. Symbol '>', '<' and '—' are the keywords when the whole individual genome sequence is rebuilt on basis of the reference genome sequence. (**b**) Changes file with redundant information deleted. (**c**) Changes file generated based on the subtracted number, which is more readable to the computer.

will be recorded using the subtracted number of the first nucleotide position to the second one. Therefore, at each side of 'i', 'd' and 'h', the number N5, N7, N9 and N11 is replaced by the subtracted number of their corresponding nucleotide position N1, N3, N5 and N7 to reduce the file size (Figure 2b and c). Eventually, the individual genome sequence information can be recorded as the format shown in Figure 2c using Huffman coding (19).

To encode the processed individual sequence data more effectively, each nucleotide sequence and its relevant number are recorded with the same binary value since it can be decoded uniquely with the help of 'i', 'd' and 'h'. Table 1 shows an example of the encoding strategy using the varied sequence information of *A. thaliana* chromosome 1 with TAIR8 as the reference and TAIR9 as the individual genome, showing that the larger counts of the symbol are reduced to the shorter encoding value. Then the bit file is able to be generated to the char code, for instance, the taken bits '01000001' presents the corresponding ASCII code 'A'.

Performance of GRS

Performance of the GRS tool was tested in three cases. When two Korean genome sequence data (KOREF_20090131 and KOREF_20090224) were used (4), the raw file with 2986.8 MB in size (KOREF_20090224) was reduced to a 18.8-MB compressed file, achieving a ~159-fold compression rate (Table 2). In addition, we also compressed the raw file of rice genome from 361.0 MB to 4.4 MB with the compression rate ~82 fold (Table 3), and 115.1 MB of *A. thaliana* genome to 6.5 KB with nearly 18 133 fold of compression (Table 4). Furthermore, the good performance of GRS was revealed by the calculated compression and decompression time of these three genomes (Supplementary Table S1).

Table 1. Huffman encoding for DNA base, genome position and recover language

DNA base	Relevant number	Counts	Encoding value
A	0	95	0110
T	1	155	000
C	2	132	1110
G	3	105	1100
N	4	101	1010
a	5	110	1111
t	6	98	0111
c	7	80	0010
g	8	106	1101
n	9	83	0011
d		31	101 110
h		15	101 111
i		54	10 110
\t		210	100
\n		168	010

Each DNA base and its relevant number are encoded with the same binary value based on the Huffman encoding strategy. Shown here is the encoding table for changes file generated for chromosome 1 of the *A. thaliana* genome using TAIR8 as reference and TAIR9 as the individual genome. Character d means delete sequence, h means change sequence and i means insert sequence.

Table 2. Performance of GRS in compressing the KOREF_20090224 human genome using KOREF_20090131 as the reference

Chromosome number	Varied sequence percentage (%)	Raw file size	Compressed file size	Compression rate
1	0.656 929	239.7 MB	1.3 MB	184.4
2	0.716 863	235.6 MB	1.3 MB	181.2
3	0.630 572	193.4 MB	987.4 KB	200.6
4	0.762 314	185.5 MB	1.1 MB	168.6
5	0.711 956	175.4 MB	964.9 KB	186.1
6	0.649 071	165.7 MB	884.9 KB	191.7
7	0.912 855	154.0 MB	1.0 MB	154.0
8	0.639 359	141.8 MB	746.4 KB	194.5
9	0.774 539	136.0 MB	844.0 KB	165.0
10	0.705 819	131.3 MB	750.4 KB	179.2
11	0.720 238	130.4 MB	738.0 KB	180.9
12	0.638 779	128.3 MB	685.6 KB	191.6
13	0.550 377	110.7 MB	508.4 KB	223.0
14	0.529 220	103.1 MB	473.4 KB	223.0
15	0.589 095	97.3 MB	484.6 KB	205.6
16	0.808 032	86.1 MB	554.7 KB	158.9
17	0.818 430	76.4 MB	494.1 KB	158.3
18	0.666 472	73.8 MB	399.0 KB	189.4
19	0.744 553	61.9 MB	390.4 KB	162.4
20	0.493 781	60.5 MB	276.0 KB	224.5
21	0.579 505	45.5 MB	221.2 KB	210.6
22	0.632 448	48.2 MB	256.3 KB	192.6
M	0.108 715	16.5 KB	183.0 B	94 543.8
X	3.299 049	150.2 MB	3.1 MB	48.5
Y	1.768 076	56.0 MB	578.9 KB	99.1
The whole genome	0.804 282	2986.8 MB	18.8 MB	158.9

The verified sequence percentage of each chromosome, the size of raw sequence file and compressed file, as well as the compression rate are shown.

Table 3. Performance of GRS in compressing rice genome of TIGR6 using TIGR5 as the reference

Chromosome number	Varied sequence percentage (%)	Raw file size (MB)	Compressed file size	Compression rate
1	0.757 801	42.0	1.4 MB	30.0
2	0.013 898	34.8	1.4 KB	25 453.7
3	0.168 381	35.3	46.6 KB	775.7
4	0.096 345	34.2	35.3 KB	992.1
5	0.069 046	29.0	6.0 KB	4949.3
6	0.000 000	30.3	0	
7	0.027 041	28.8	4.0 KB	7372.8
8	0.479 452	27.6	115.5 KB	244.7
9	0.000 000	22.3	0	
10	1.128 503	22.4	770.1 KB	29.8
11	0.188 992	27.6	2.3 MB	12.0
12	0.000 000	26.7	0	
The whole genome	0.244 122	361.0	4.4 MB	82.0

The verified sequence percentage of each chromosome, the size of raw sequence file and compressed file, as well as the compression rate are shown.

Table 4. Performance of GRS in compressing *A. thaliana* genome of TAIR9 using TAIR8 as the reference

Chromosome number	Varied sequence percentage (%)	Raw file size (MB)	Compressed file size	Compression rate
1	0.016 314	29.4	715.0 B	43 116.3
2	0.036 145	19.0	385.0 B	51 747.9
3	0.046 910	22.7	2.9 KB	6709.0
4	0.000 301	17.9	1.9 KB	9647.2
5	0.063 888	26.1	604.0 B	45 311.0
The whole genome	0.032 712	115.1	6.5 KB	18 132.7

The verified sequence percentage of each chromosome, the size of raw sequence file and compressed file, as well as the compression rate are shown.

DISCUSSION

With the advance of DNA sequencing technologies, more and more genome resequencing projects, such as the International HapMap Project and the 1000 Genomes Project, have been initiated (20,21). As a result, compression of the huge amount of genome sequencing data has become an important issue (10,11). Currently available tools have limitations in effectively processing the large amount of genome reseqencing data. For example, tools developed by Brandon *et al.* (10) and Christley *et al.* (11) are limited by not only the known reference SNPs map, but also the possible loss of sequence information, such as large structural variations (SVs) including sequence rearrangements and segment duplications. Even though the advent of sequencing technologies facilitates the processing of individual genome sequence such as reassembling genome

sequences using the sequencing reads on basis of the reference genome (3,4), the current sequence compression tools are not very suitable for this purpose. Moreover, comprehensive reference SNPs maps are unavailable for many organisms such as rice, *A. thaliana* and other species, making it hard to compress these genome resequencing data using the available tools. In this study, we show that GRS is a *de novo* genome compression approach for compressing resequencing data, which is applicable to the genome data management of many species.

Varied sequence percentage plays a critical role in compressing the genome resequencing data. GRS employs a novel approach to deal with the resequensing data, especially for those data sets with higher variation between the reference genome and the individual genome. The key point of GRS is to splice the reference chromosome and the input chromosome into the same intervals, and then calculate each corresponding pair of the varied sequence percentage δ_i based on each nucleotide frequency. Subsequently, concatenating piece i to make the modified reference chromosome and modified input chromosome creates the minimum varied sequence percentage based on the δ_i value (Figure 3). Then the minimum change file can be compressed using GRS and the chromosome piece with a higher value of varied sequence percentage can be compressed using the general and routine file compression method such as 7-Zip. When the chromosome size is too big or the computer memory capability is limited, it is useful to splice the reference chromosome and the input chromosome data. In this study, GRS grouped each chromosome sequencing data of the Korean genome (4), into 50, 25 or 10 million per piece, respectively. Similar compression capabilities were obtained (i.e. file size is ~19 MB), demonstrating the flexibility and reliability of GRS.

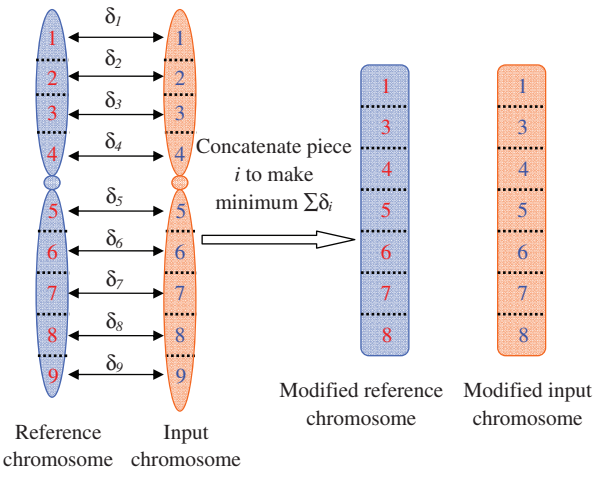

Figure 3. Method to resolve the minimum varied sequence percentage between the reference and input chromosomes and assemble the pieces together. Here is an example showing that two chromosomes are spliced into nine parts with relevant δ_i. Part 1, 3, 4, 5, 6, 7 and 8 are put together because δ_2 and δ_9 with a higher value than others based on the threshold of varied sequence percentage.

CONCLUSIONS

In this article, we designed and implemented a generic tool, GRS, for *de novo* compression of genome resequencing data. GRS is simple to use and does not need the reference SNPs map, thus can be widely used for many genomes, especially those without reference SNPs. Case studies using the sequencing data of human, rice and *A. thaliana* genomes have demonstrated the good performance of GRS in sequencing data compression.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge Jianping Hu for editing this manuscript and Xiaobin Wu and Xiaoyi Guo for helpful discussions.

FUNDING

National Key Basic Research Developments Program, Ministry of Science and Technology, P. R. China (2009CB941500, 2006CB101700); National '863' High-Tech Project (2006AA10A102); National Natural Science Foundation of China (30725022 and 30600347); Shanghai Leading Academic Discipline Project (B205). Funding for open access charge: Special Funding for Transgenic Organisms (2008ZX08012-002).

Conflict of interest statement. None declared.

REFERENCES

1. Horner,D.S., Pavesi,G., Castrignano,T., Meo,P.D.D., Liuni,S., Sammeth,M., Picardi,E. and Pesole,G. (2009) Bioinformatics approaches for genomics and post genomics applications of next-generation sequencing. *Brief. Bioinform.*, **11**, 181–197.
2. Pushkarev,D., Neff,N.F. and Quake,S.R. (2009) Single-molecule sequencing of an individual human genome. *Nat. Biotechnol.*, **27**, 847–852.
3. Service,R.F. (2006) The race for the $1000 genome. *Science*, **311**, 1544–1546.
4. Ahn,S.M., Kim,T.H., Lee,S., Kim,D., Ghang,H., Kim,D.S., Kim,B.C., Kim,S.Y., Kim,W.Y., Kim,C. *et al.* (2009) The first Korean genome sequence and analysis: Full genome sequencing for asocio-ethnic group. *Genome Res.*, **19**, 1622–1629.
5. Wheeler,D.A., Srinivasan,M., Egholm,M., Shen,Y., Chen,L., McGuire,A., He,W., Chen,Y., Makhijani,V., Roth,G.T. *et al.* (2008) The complete genome of an individual by massively parallel DNA sequencing. *Nature*, **452**, 872–876.
6. Wang,J., Wang,W., Li,R., Li,Y., Tian,G., Goodman,L., Fan,W., Zhang,J., Li,J., Zhang,J. *et al.* (2008) The diploid genome sequence of an Asian individual. *Nature*, **456**, 60–66.
7. Rhead,B., Karolchik,D., Kuhn,R.M., Hinrichs,A.S., Zweig,A.S., Fujita,P., Diekhans,M., Smith,K.E., Rosenbloom,K.R., Raney,B.J. *et al.* (2010) The UCSC Genome Browser database: update. *Nucleic Acids Res.*, **38**, D613–D619.
8. Lander,E.S., Linton,L.M., Birren,B., Nusbaum,C., Zody,M.C., Baldwin,J., Devon,K., Dewar,K., Doyle,M., FitzHugh,W. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.

9. Snyder,M., Du,J. and Gerstein,M. (2010) Personal genome sequencing: current approaches and challenges. *Genes Dev.*, **24**, 423–431.
10. Brandon,M.C., Wallace,D.C. and Baldi,P. (2009) Data structures and compression algorithms for genomic sequence data. *Bioinformatics*, **25**, 1731–1738.
11. Christley,S., Lu,Y., Li,C. and Xie,X. (2009) Human genomes as email attachments. *Bioinformatics*, **25**, 274–275.
12. Tembe,W., Lowey,J. and Suh,E. (2010) G-SQZ: compact encoding of genomic sequence and quality data. *Bioinformatics*, **26**, 2192–2194.
13. Soliman,T.H., Gharib,T.F., Abo-Alian,A. and Sharkawy,M.E. (2009) A Lossless Compression Algorithm for DNA sequences. *Int. J. Bioinform. Res. Appl.*, **5**, 593–602.
14. Huala,E., Dickerman,A., Garcia-Hernandez,M., Weems,D., Reiser,L., LaFond,F., Hanley,D., Kiphart,D., Zhuang,J., Huang,W. *et al.* (2001) The Arabidopsis Information Resource (TAIR): A comprehensive database and web-based information retrieval, analysis, and visualization system for a model plant. *Nucleic Acids Res.*, **29**, 102–105.
15. Ouyang,S., Zhu,W., Hamilton,J., Lin,H., Campbell,M., Childs,K., Thibaud-Nissen,F., Malek,R.L., Lee,Y., Zheng,L. *et al.* (2007) The TIGR Rice Genome Annotation Resource: improvements and new features. *Nucleic Acids Res.*, **35**, D883–D887.
16. Rhee,S.Y., Beavis,W., Berardini,T.Z., Chen,G., Dixon,D., Doyle,A., Garcia-Hernandez,M., Huala,E., Lander,G., Montoya,M. *et al.* (2003) The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community. *Nucleic Acids Res.*, **31**, 224–228.
17. Myers,E.W. (1986) An O(ND) Difference Algorithm and Its Variations. *Algorithmica*, **1**, 251–266.
18. Miller,W. and Myers,E.W. (1985) A File Comparison Program. *Software-Pract. Exper.*, **15**, 1025–1040.
19. Huffman,D. (1952) A method for the construction of minimum redundancy codes. *Proc. IRE*, **40**, 1098–1101.
20. Frazer,K.A., Ballinger,D.G., Cox,D.R., Hinds,D.A., Stuve,L.L., Gibbs,R.A., Belmont,J.W., Boudreau,A., Hardenbol,P., Leal,S.M. *et al.* (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature*, **449**, 851–862.
21. Kaiser,J. (2008) A plan to capture human diversity in 1000 genomes. *Science*, **319**, 395.