BIOINFORMATICS APPLICATIONS NOTE Vol. 29 no. 17 2013, pages 2199–2202
doi:10.1093/bioinformatics/btt362

Genome analysis Advance Access publication June 22, 2013

The human genome contracts again

Dmitri S. Pavlichin[1], Tsachy Weissman[2] and Golan Yona[2,*]

[1]Department of Physics and [2]Department of Electrical Engineering, Stanford University, CA 94305, USA

Associate Editor: Alfonso Valencia

ABSTRACT

Summary: The number of human genomes that have been sequenced completely for different individuals has increased rapidly in recent years. Storing and transferring complete genomes between computers for the purpose of applying various applications and analysis tools will soon become a major hurdle, hindering the analysis phase. Therefore, there is a growing need to compress these data efficiently. Here, we describe a technique to compress human genomes based on entropy coding, using a reference genome and known Single Nucleotide Polymorphisms (SNPs). Furthermore, we explore several intrinsic features of genomes and information in other genomic databases to further improve the compression attained. Using these methods, we compress James Watson's genome to 2.5 megabytes (MB), improving on recent work by 37%. Similar compression is obtained for most genomes available from the 1000 Genomes Project. Our biologically inspired techniques promise even greater gains for genomes of lower organisms and for human genomes as more genomic data become available.

Availability: Code is available at sourceforge.net/projects/genomezip/

Contact: golan.yona@stanford.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on November 29, 2012; revised on May 30, 2013; accepted on June 18, 2013

1 INTRODUCTION

With the constant advances in sequencing technologies, genome sequencing has become faster and more affordable. Although the main effort thus far has been to sequence the genomes of different organisms, the focus is gradually shifting toward sequencing different instances of the same genome (i.e. different individuals) to study the variations underlying phenotypic differences between individuals and to identify the variations that are associated with diseases and disorders. Consequently, the number of complete human genomes that have been sequenced is increasing rapidly.

As the amount of and demand for genomic data grow, the cost of storage and transmission of these data is fast becoming a bottleneck for research and future medical applications. Thus, there is a growing need for compression algorithms suited to genomic data.

Genome compression has been the subject of multiple studies in the past several years (see Supplementary Material for an overview of these related studies). The most successful methods are those that use reference genomes and code just the differences between the input genome and a reference genome (which, for humans, account for <0.2% of the genome's length). The best single-reference compression was reported in Christley *et al.* (2009), who compressed James Watson's genome to 4 MB by utilizing dbSNP (Sherry *et al.*, 2001) to represent more efficiently known SNPs in the difference map.

In this work, we report further improvements to this scheme, which result in a significant reduction of 37% in the size of the compressed genome, for only 2.5 MB. Our work is motivated by two observations. First, entropy-coding techniques can be nearly optimal in exploiting known patterns in a dataset but have not been fully optimized on human genome data. Second, in finding patterns to exploit, a wealth of biological insight remains untapped. Our improved scheme is the result of incorporating multiple sources of information on the presence of haplotypes, tag SNPs, coding and non-coding regions and other biologically motivated modifications.

2 METHODS

James Watson's genome is available (Wheeler *et al.*, 2008) as a difference map from the reference genome hg18 from the UCSC Genome Browser (Kent *et al.*, 2002). The difference map consists of three parts: (i) 3 321 840 SNPs, (ii) 156 556 deletions and (iii) 63 607 insertions.

Our overall compression scheme is similar to that of Christley *et al.* (2009). The differences lie in the way we use the statistical properties of each type of variation, and the use of other sources of information. For consistency with the starting point in Christley *et al.* (2009), we first parse the input into an 84 MB ASCII file that removes text fields redundant with the reference genome. The next steps exploit data-specific distributions to reduce the number of bits needed to code each type of variation.

2.1 SNPs

The majority of the differences between Watson's genome and the reference genome are the SNPs, and the bulk of the final file size (76%) is devoted to storing the SNPs positions. For each SNP, we need to specify the position in the genome and the nucleotide value of the polymorphism. Most of the SNPs (~2.7 million, or 82% of all SNPs) in Watson's genome were already documented in version 129 of dbSNP [For the purpose of fair comparison with Christley *et al.* (2009), we used throughout this article (unless indicated otherwise) the same version of dbSNP used in that study, which is version 129]. The remaining 0.62 million SNPs are novel with respect to that version. Representing SNPs that appear in dbSNP can be done more efficiently, as explained in section 2.1.3.

2.1.1 Novel SNP positions [space used: 1.0 MB, or 41% of the final file size]. Instead of encoding the absolute position of an SNP (which can be a large number, comparable to the length of a chromosome), we use an approach similar to Christley *et al.* (2009) and encode the relative distance from the last SNP. Storing the distances between

*To whom correspondence should be addressed.

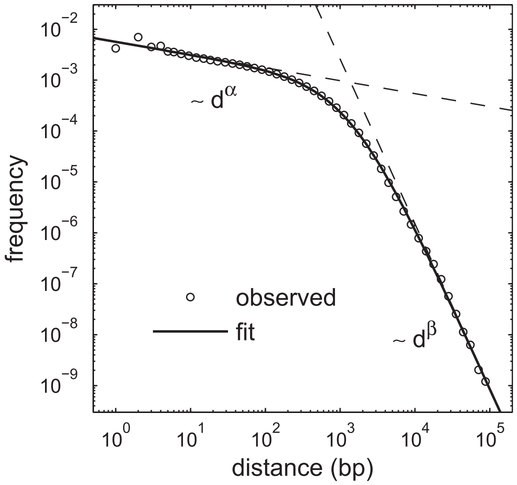

Fig. 1. The empirical SNP-to-SNP distance distribution for James Watson's 3.3M SNPs follows a double power-law distribution as in Equation (1). Circles denote the empirical distribution, and the solid line corresponds to the fit used in constructing our entropy code. Dashed lines indicate limiting behavior of the fit to the left and right of the kink. Approximate parameter values in the notation of Equation (1) are $\alpha = -0.25$, $\beta = -3.2$, $d_0 = 780$ and $\gamma = -2.7$

consecutive SNPs is sufficient to recover their absolute positions in the genome. Unlike Christley *et al.* (2009), we encode the distances by constructing an entropy code for the empirical SNP-to-SNP distance distribution. This distribution can be approximated with a mix of two power-law distributions (Fig. 1). The functional form of the double power-law distribution that we fit is (up to normalization):

$$p(d) \sim \left(d^{\frac{\alpha}{\gamma}} + (d/d_0)^{\frac{\beta}{\gamma}}\right)^{\gamma} \qquad (1)$$

Where d is the distance in base pairs (bp) between consecutive SNPs, the parameters α and β give the slope to the left and right of the kink, d_0 is the position in bp of the kink, and γ controls the curvature of the function near the kink (for $\beta < \alpha < 0$, $\gamma < 0$). Thus, we need to store only the four numerical coefficients of the functional form, saving the need to store a (~0.15 MB) code table. Having a functional form further permits us to impute distance frequencies for other human genomes. For more details see Supplementary Material.

2.1.2 Novel SNP values [space used: 100 kilobytes (KB), 4% of final file size]. The nucleotide values of SNPs conditioned on the nucleotide values in the reference sequence are not uniformly distributed. We use an arithmetic code for this conditional distribution to compress the list of novel SNP values by 16%.

2.1.3 SNPs in dbSNP [space used: 920 KB, 35% of final file size]. We use the set of all SNPs from the NCBI dbSNP database [Starting from version 132, multiple subsets (tracks) of SNPs were made available through the UCSC Genome Browser, including the following: (i) all SNPs and (ii) common SNPs (SNPs that appear in at least 1% of the population). Before version 132, dbSNP was available as a single set of (all) SNPs]. Track 129 for hg18 from the UCSC table browser contains ~9.6M SNPs, 2.7M of which appear in Watson's genome. Instead of using a vector representation of dbSNP indicating which SNPs exist in Watson's genome (as in Christley *et al.*, 2009), we use again the SNP-to-SNP distance distribution to encode known SNPs. The distances between the *indices* of Watson's SNPs in dbSNP are distributed approximately as the double power law in Equation (1), but with different parameters (see Section 2.4 of the Supplementary Material). We use an entropy code for this distribution to compress the list of Watson's SNPs in dbSNP by 22%.

It should be noted that dbSNP contains also common deletions and insertions; however, there were few matches with Watson's genome to yield a benefit in compression; therefore, we ignored non-SNP dbSNP entries.

2.2 Deletions and insertions

2.2.1 Deletions and insertions positions [space used: 370 KB, 14% of final file size]. Instead of encoding the positions separately, we merge the 157K deletion and 64K insertion positions with the SNP positions to generate a single distance distribution, resulting in shorter distances between consecutive changes (and therefore smaller code words). The combined distribution follows a double power-law distribution, as in Equation (1), from which we derive an entropy code for the distances. We use an arithmetic code to encode a list that labels each change as an SNP, a deletion or an insertion.

2.2.2 Deletion lengths [space used: 66 KB, 3% of final file size]. We construct a Huffman code for the deletion lengths, incurring a cost of 3.6 bits per deletion on average.

2.2.3 Insertion sequences [space used: 87 KB, 3% of final file size]. We separately encode the insertion lengths and the inserted sequences. We construct a Huffman code for the insertion lengths, incurring a cost of 2.9 bits per insertion on average. To encode the inserted sequences, we concatenate them into one sequence. The distribution of nucleotide values in the set of insertion sequences is not uniform, with A and T occurring roughly twice as often as G and C. We use an arithmetic code for this distribution to compress the concatenated insertion sequence by 5%.

2.3 Biologically motivated improvements

We exploit several observations about genomes to further improve the compression:

2.3.1 Conditioning on haplotype Adjacent SNPs often co-occur in blocks called haplotypes. In most cases, a block can be uniquely characterized using a subset of the SNPs in the haplotype, called tag SNPs, with all the other SNPs determined from the tag SNPs values. Therefore, we can achieve compression of the set of non-novel SNPs (those already in dbSNP) by storing only the tag SNPs for each haplotype, sufficient to identify an individual's genome's haplotypes. If the tag SNPs do not perfectly predict the remaining SNPs (which may happen if the genome's haplotypes differ from those SNPs recorded in dbSNP), then we must store a (compressible) list of wrong predictions. As long as the tag SNPs have any predictive power at all, this will result in a smaller file than storing a list with one entry per SNP in dbSNP.

In The International HapMap Consortium *et al.* (2007), it is estimated that 300 000 to 600 000 tag SNPs are sufficient to predict most of the 9.6M common SNPs in dbSNP. Thus, using tag SNPs potentially reduces the file size of non-novel SNPs by more than a factor of 10 and the file size for the entirety of Watson's genome by about a third. However, information on tag SNPs is sparse. In this study, we use the Illumina Bead Array HumanMap300k dataset, containing about 300K tag SNPs (Pe'er *et al.*, 2006). This dataset includes only 2.46M of the 9.6M SNPs in dbSNP. Only 1.56M of Watson's SNPs are associated with a tag SNP in the Illumina set (of which 88% are consistent with the predicted SNPs). This yields only ~46 KB in savings, but is expected to improve once more data on haplotypes become available, and for genomes more consistent with the predicted SNPs. For details, see Supplementary Material.

2.3.2 Conditioning on coding versus non-coding regions We use the UCSC genes track for hg18 from the UCSC table browser and find that a nucleotide in an exon is ~74% as likely to be an SNP as a nucleotide outside an exon. We can achieve some compression by constructing separate entropy codes for the positions of SNPs inside and outside exons. Given that exons make up ~2.5% of the human genome, this amounts to only ~0.01% savings in bits per SNP, but the savings would be larger for genomes that have more equal fractions of coding and non-coding elements, such as some bacterial genomes.

2.3.3 Conditioning on context in sequence We observe that the probability of a nucleotide being an SNP depends on the local sequence that includes the nucleotide (the context). In the most extreme example, the sequence CGC (and complement) is the most SNP-prone, 12 times as likely to contain an SNP in the middle nucleotide compared with AAA (and complement), the least SNP-prone. This is consistent with the observed correlation of the chromatin structure with the composition of the DNA sequence. It has been shown that the chromatin is mostly open in GC-rich regions (Dekker, 2007), suggesting that mutations are more likely in such regions. The context-based method yields ~0.33 bits per SNP in savings (or ~30 KB, 3% of the space devoted to storing novel SNPs). We find that conditioning on subsequences of length 5 maximizes savings after accounting for the overhead of storing a frequency table for each subsequence.

3 RESULTS

We applied our compression tool to each of the 1092 human genomes available from the 1000 Genomes Project (1000 Genomes Project Consortium. *et al.*, 2012), as of October 2012. The results are presented in Figure 2, showing the uncompressed file sizes (converted to the same uncompressed format as Watson's genome), and the compression ratios, labeled by population. These results confirm that the success of the algorithm in compressing genomes is not unique to Watson's genome. However, we notice a significant variation by population in both the uncompressed file size and the compression ratio achieved by our approach. Further analysis is required to understand the functional significance of the underlying differences.

It should be noted that the 1000 Genomes Project provides allele-specific information on variations, while the publicly available difference map for Watson's genome combines the variations into a single list, without specifying which allele(s) they occur at. Therefore, we modified our algorithm to account for diploids, by introducing a vector indicating homo/heterozygosity (and which allele, in case of heterozygosity). This is done by using two bits per variation, which can take the value 11 (homozygosity, both alleles have the variation), 01 or 10. Coding the allele indicator vector takes ~500 KB per genome, using a Huffman code for a Markov chain fit to the indicator vector (described in greater detail in the Supplementary Material).

Another difference between the 1000 genomes and Watson's genome is that the 1000 Genomes Project lists indels in addition to SNPs, deletions and insertions. The number of indels per genome is small (on average, there are only 179 indels out of ~3.8M variations per genome in the 1000 genomes data), and we represent each as a pair of an insertion and a deletion.

4 DISCUSSION

We have applied several entropy-coding techniques to compress the difference map between James Watson's genome and a reference genome. Compared with the software of Christley *et al.* (2009), our software uses about twice as much time and memory (for details see Supplementary Material). However, it reduces the final file size to 2.5 MB, a new lower bound by 37% over previous work. Even better results were obtained when the analysis was repeated with a more recent version of dbSNP (version 130, which included more of Watson's SNPs), with the final file size dropping from 2.54 MB to 1.96 MB, or 23% improvement overall. (More recent versions of dbSNP use assembly hg19 of the human genome, whereas Watson's genome was given with reference to hg18. Therefore, we could not test Watson's genome with more recent versions of dbSNP.) Our tests on all genomes from the 1000 Genomes Project are consistent with these results and trends, with newer versions of dbSNP improving the compression ratio, as expected. Our approach requires the use of a reference genome and dbSNP, but the cost of storing these databases is amortized for other human genomes. We have implemented several biologically motivated techniques that offer a path to future savings in compressing human and non-human genomes. Our scheme becomes more efficient, as the external databases it uses are updated to include newly observed variations and tag SNPs, effectively letting us exploit a growing standardized collection of reference genomes.

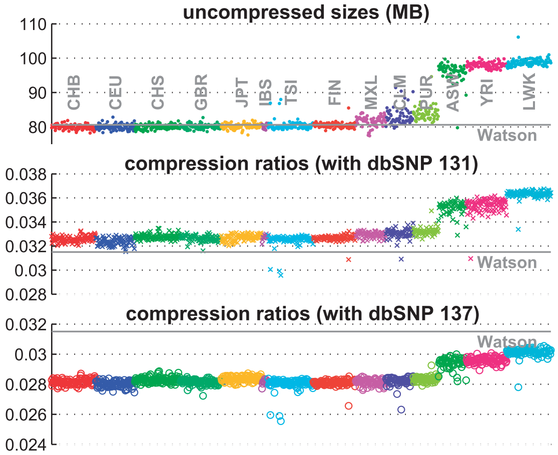

Fig. 2. Compression results for the 1000 Genomes Project. For each of the 1092 genomes, we plot the uncompressed size (top) and the compression ratio (bottom). Genomes are sorted by the population they belong to (see population codes below). Because the 1000 genomes' SNPs have been submitted to dbSNP starting with version 132, we compare the compression results for two versions of dbSNP: version 131, which is the latest before any of the new SNPs were recorded, and version 137 (common SNPs), which is the latest at the time of writing, and yields better compression results than version 137 (all SNPs). We also include results for Watson's genome on the same plot as a rough reference, although Watson's genome was compressed with respect to a different reference genome (hg18) and version of dbSNP (129) than the 1000 genomes data. Population codes (from http://www.1000genomes.org/about): **ASW:** African Ancestry in Southwest US, **CEPH:** Utah residents with Northern and Western European ancestry (CEU), **CHB:** Han Chinese in Beijing, China, **CHS:** Han Chinese South, **CLM:** Colombian in Medellin, Colombia, **FIN:** Finnish from Finland, **GBR:** British from England and Scotland, **IBS:** Iberian populations in Spain, **JPT:** Japanese in Toyko, Japan, **LWK:** Luhya in Webuye, Kenya, **MXL:** Mexican Ancestry in Los Angeles, CA, **PUR:** Puerto Rican in Puerto Rico, **TSI:** Toscani in Italy, **TRI:** Yoruba in Ibadan, Nigeria

ACKNOWLEDGEMENTS

We thank John Hamann, Nikolas Tezak and Michael Armen for helpful discussions.

Funding: NSF Center for Science of Information (NSF grant number 4101-38047). Stanford Graduate Fellowship (to D.P.).

Conflict of Interest: none declared.

REFERENCES

1000 Genomes Project Consortium *et al.* (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature*, **491**, 56–65.

Christley,S. *et al.* (2009) Human genomes as email attachments. *Bioinformatics*, **25**, 274–275.

Dekker,J. (2007) GC- and AT-rich chromatin domains differ in conformation and histone modification status and are differentially modulated by Rpd3p. *Genome Biol.*, **8**, R116.

Kent,W.J. *et al.* (2002) The human genome browser at UCSC. *Genome Res.*, **12**, 996–1006.

Pe'er,I. *et al.* (2006) Evaluating and improving power in whole-genome association studies using fixed marker sets. *Nat. Genet.*, **38**, 663–667.

Sherry,S.T. *et al.* (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acid Res.*, **29**, 308–311.

The International HapMap Consortium *et al.* (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature*, **449**, 851–861.

Wheeler,D.A. *et al.* (2008) The complete genome of an individual by massively parallel DNA sequencing. *Nature*, **452**, 872–876.