Large-scale inversions between human reference assemblies

Diogo Pratas
pratas@ua.pt
Raquel M. Silva
raquelsilva@ua.pt
Armando J. Pinho
ap@ua.pt

Signal Processing Lab, IEETA / DETI,
University of Aveiro,
3810–193 Aveiro, Portugal

Abstract

The detection of large-scale inversions between intra-species genomes is fundamental to understand the dynamics of chromosomal evolution, namely to identify hallmarks associated with diseases. In this paper, we explore a method that finds exact or approximate large-scale inversions, for regions larger than 500 kb, between the three most popular assembled human genomes. We identify the relative positions for each chromosome. The inversions are computed in two maps showing at a glance the associated regions.

Introduction

How genome architecture and which macroevolutionary events emerge through time are fundamental to understand the dynamics of species evolution, namely their origin and speciation patterns [3]. Several insights into chromosome evolution have been traditionally achieved by cytogenetic procedures, for example G-banding, and by molecular karyotyping approaches, such as fluorescence in situ hybridization (FISH). More recently, array-based methods became very popular [2].

Advances in sequencing technology have increased the number of digital human genomes, raising conditions towards intra-species characteristics and diversity research. Consequently, computational approaches emerged [5], bringing high resolution, accuracy and speed with less costs. However, the de novo assembly of the next generation sequencing (NGS) reads is still problematic, mainly because the alignment of the reads from these new genomes to a high quality reference genome remains a critical aspect of data interpretation. Nevertheless, the human reference assembly is the highest quality mammalian assembly available. The main reference genome assemblies are those from the Genome Reference Consortium (GRC 38) [1], the J. Craig Venter Institute (HuRef) [4] and the Washington U. School of Medicine (CHM 1.1).

In this paper, we detect the large-scale (larger than 500 Kb) inversions between these three reference genome assemblies, exploring fundaments from an unsupervised alignment-free method [9], that is based in the hability to efficiently model the repetitiveness of genomic sequences [8].

Method

The method involves the estimation of the amount of conditional exclusive information, i. e., using only information from a reference that is required to represent a certain region of the sequence or sub-sequence [6]. The process is as follows:

1. Convert symbols outside $\mathcal{A} = \{A, C, G, T, N\}$ into "N";
2. Pseudo-randomize the "N" symbols (with uniform distribution);
3. Invert the first sequence;
4. Load to the FCM [7] the sequence created in step 3;
5. Compress the second sequence using the FCM created in step 4;
6. Filter the information sequence generated in 5;
7. Use a threshold of 1.6 to segment the sub-sequence regions;
8. Paint sub-sequence regions with different colors;
9. For each sub-sequence region, repeat the process from step 3 to 8, using the second sequence as target;

The colors are calculated automatically using a HSV scheme where only V varies.

Results

For the results presented in this section we have downloaded the genomes from the NCBI[1] site. Chromosome Y from CHM genome is absent and thus we have not reported results associated with this sequence.

Fig. 1 shows the maps from the large-scale inversions between A (GRC) and B (HuRef), while Fig. 2 shows the inversions between A(GRC) and C (CHM) assemblies. In respect to A/B, there are inversions in chromosomes 1, 2, 7, 9, 10, 11, 15 and Y. Specifically to chromosome 1, the inversions are contained in the pericentric regions (around 119 Mb). This region is also inverted between human and chimpanzee species, although in a much larger density [9].

On the other hand, for A/C the inversions are present between chromosomes 1, 2, 5, 7, 8, 9, 10, 11, 14, 15, 16, 17, 22, X. From these, most of the inversions are contained in pericentric regions, a major factor of dynamism across individuals of the same species.

Conclusions

We have proposed a procedure to detect exact or approximate inversions, larger than 500 Kb, between reference assembled genomes. The inversions have been computed in a information map.

Besides specific characteristics, the reference assembly from the Washington U. School of Medicine (CHM) seems to have a higher number of inversions compared with the Genome Reference Consortium (GRC 38), than with the J. Craig Venter Institute (HuRef).

Acknowledgements

Supported by the European Fund for Regional Development (FEDER) through the Operational Program Competitiveness Factors (COMPETE) and by the Portuguese Foundation for Science and Technology (FCT), in the context of projects PEst-OE/EEI/UI0127/2014 and Incentivo/EEI/UI0127/2014. DP is supported by the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement No. 305444 "RD-Connect: An integrated platform connecting registries, biobanks and clinical bioinformatics for rare disease research". RMS is supported by the project Neuropath (CENTRO-07-ST24-FEDER-002034), co-funded by QREN "Mais Centro" program and the EU.

References

[1] D. Church, et al. Modernizing reference genome assemblies. *PLoS Biology*, 9 (7):e1001091, 2011.

[2] K. Das and P. Tan. Molecular cytogenetics: recent developments and applications in cancer. *Clinical Genetics*, 84(4):315–325, 2013.

[3] Marta Farré and Aurora Ruiz-Herrera. Role of chromosomal reorganisations in the human-chimpanzee speciation. *eLS*, 2014.

[4] S. Levy, et al. The diploid genome sequence of an individual human. *PLoS Biology*, 5:2113–2144, 2007.

[5] C. Nielsen, et al. Visualizing genomes: techniques and challenges. *Nature methods*, 7:S5–S15, 2010.

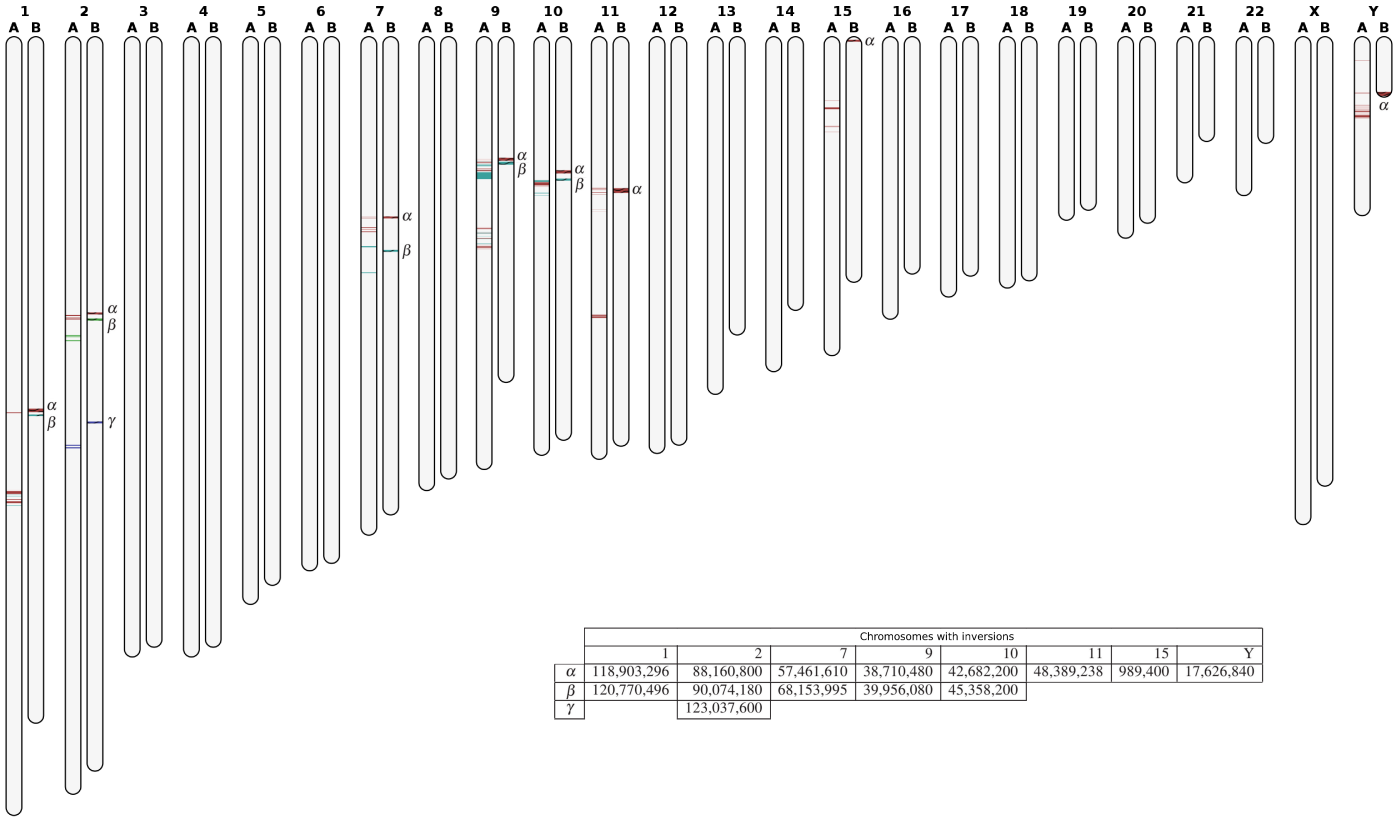
[6] A. J. Pinho, et al. Symbolic to numerical conversion of DNA sequences using finite-context models. In *Proc. of the 19th European Signal Processing Conf., EUSIPCO-2011*, Barcelona, Spain, August 2011.

[7] A. J. Pinho, et al. Complexity profiles of DNA sequences using finite-context models. In *Information Quality in e-Health*, volume 7058, pages 75–82. Springer, 2011.

[8] A. J. Pinho, et al. DNA sequences at a glance. *PLoS ONE*, 8(11):e79922, November 2013.

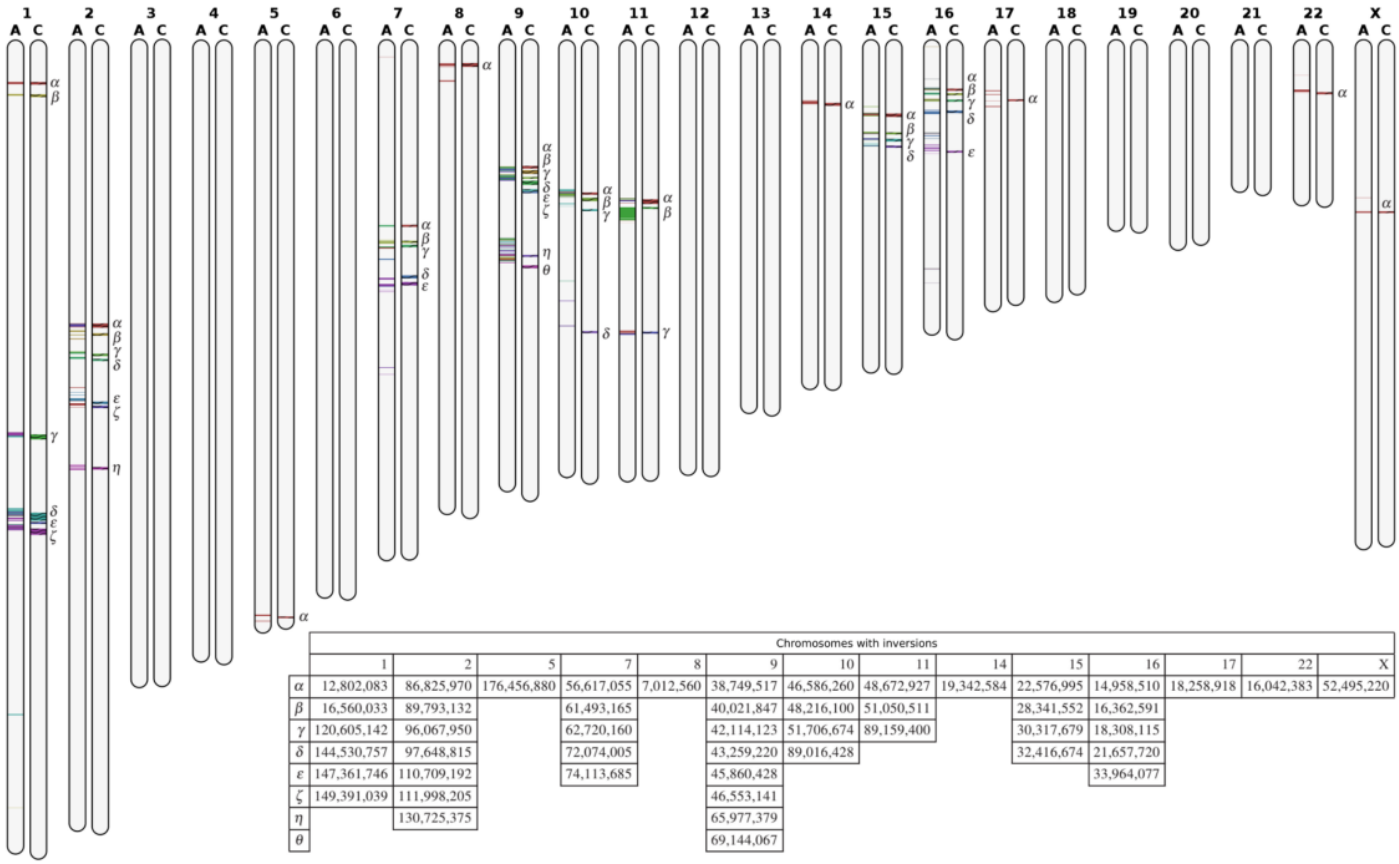
[9] D. Pratas and A. J. Pinho. A conditional compression distance that unveils insights of the genomic evolution. In *Proc. of the Data Compression Conf., DCC-2014*, Snowbird, Utah, March 2014.

[1] `ftp://ftp.ncbi.nlm.nih.gov/genomes/Homo_sapiens/Assembled_chromosomes/seq/`



	Chromosomes with inversions							
	1	2	7	9	10	11	15	Y
α	118,903,296	88,160,800	57,461,610	38,710,480	42,682,200	48,389,238	989,400	17,626,840
β	120,770,496	90,074,180	68,153,995	39,956,080	45,358,200			
γ		123,037,600						

Figure 1: Large-scale inversions between GRC (A) and HuRef (B) assemblies for each chromosome. The information maps show exact or approximate inversions with length higher than 500 kb. Each position associated with inversions, in the HuRef chromosomes, is reported in the table and marked with a greek letter according to the map.



	Chromosomes with inversions													
	1	2	5	7	8	9	10	11	14	15	16	17	22	X
α	12,802,083	86,825,970	176,456,880	56,617,055	7,012,560	38,749,517	46,586,260	48,672,927	19,342,584	22,576,995	14,958,510	18,258,918	16,042,383	52,495,220
β	16,560,033	89,793,132		61,493,165		40,021,847	48,216,100	51,050,511		28,341,552	16,362,591			
γ	120,605,142	96,067,950		62,720,160		42,114,123	51,706,674	89,159,400		30,317,679	18,308,115			
δ	144,530,757	97,648,815		72,074,005		43,259,220	89,016,428			32,416,674	21,657,720			
ε	147,361,746	110,709,192		74,113,685		45,860,428					33,964,077			
ζ	149,391,039	111,998,205				46,553,141								
η		130,725,375				65,977,379								
θ						69,144,067								

Figure 2: Large-scale inversions between GRC (A) and CHM (C) assemblies for each chromosome. The information maps show exact or approximate inversions with length higher than 500 kb. Each position associated with inversions, in the CHM chromosomes, is reported in the table and marked with a greek letter according to the map.