Measure of unevenness in human genomes, described as a self-affine phase transition in a 'spin-chain' model [version 1; peer review: awaiting peer review]

Sergey Feranchuk

Department of Physics, Smolensk State University, Smolensk, Russian Federation

Abstract

Background: Non-Gaussian distribution of polymorphic positions across a genome can substantially influence the results of any approach to molecular evolution based on a 'classical' probability model. The infinite dispersion of non-Gaussian perturbations is a challenge in an attempt to accept it in a probability-based model of evolution.

Methods: Here a model is proposed where non-Gaussian distribution is introduced to an exact solution of the 'Ising model'; it describes a behavior of one-dimensional chain of spins in an approaching to a phase transition. The distribution of fragments which are identical between two genomes is similar to distribution of islands of spins with the same orientation, in the model where non-integer dimension is introduced.

Results: Application of this model allows us to compare the relative contributions of non-Gaussian perturbations for pairs of human genomes from different ethnic groups. An evolution of the three human races in a most compact presentation is considered, rates of development on the separated stages of the evolution are assumed to be proportional to a value of relative unevenness between the appropriate groups of genomes. In the resolved model, the meaningful details of the separation between Asian and European races are clarified, in a period around ten thousand years ago; a particular viewpoint to the separation of the African race is also presented.

Conclusion: The proposed approximation of non-Gaussian perturbations in human genomes allows to support the statements which are otherwise missed in the scientific investigations of the early history of modern humans.

Keywords

mutation rate, self-affine development, human evolution, phase transition, genome coverage, Landau-Zener transition
Corresponding author: Sergey Feranchuk (feranchuk@gmail.com)

Author roles: Feranchuk S: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: The author(s) declared that no grants were involved in supporting this work.

Copyright: © 2021 Feranchuk S. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Feranchuk S. Measure of unevenness in human genomes, described as a self-affine phase transition in a ‘spin-chain’ model [version 1; peer review: awaiting peer review] F1000Research 2021, 10:149 https://doi.org/10.12688/f1000research.51192.1

First published: 25 Feb 2021, 10:149 https://doi.org/10.12688/f1000research.51192.1
Introduction
The issues about an unevenness of a genome arose in particular in a distribution of coverage of sequencing reads mapped to a genome. There, the deviations of reads’ coverage from a ‘classical’ Lange-Waterman model, which was constructed following a Poisson distribution for short genome fragments, reflects some features of self-affinity for most frequent genome fragments in addition to the previously observed over-dispersion of a ‘Poisson’ peak. There, the effect was observed stably in several analyzed genomes and is to be treated as a robust enough phenomenon to be discovered further and in depth.

The features of self-affinity in DNA sequences were detected at the very early days of genomics, in a classical work of Peng et al.\cite{1}; a definition of the fractal dimension for the one-dimensional series was proposed there, and DNA sequences were a model of ‘fractal’-like series.

The self-affinity features in a phenomenon imply the influence of perturbations with infinite dispersion, or the presence of a so-called ‘fat-tail’ in their probability distribution, and these features are difficult to detect and describe.

Here, an approach is proposed to detect and apply a measure to these ‘perturbations’ focusing on the phenomenon mentioned, which was observed in coverage distributions. The relevance of a proposed research project is demonstrated on a model of evolution of humans restored from some of the present-day human genomes; confusions which are accumulated in solving of this scientific problem were in fact a motif to drive out the research.

Methods
Self-affinity features are a property of ‘transitional’ period, and a description of these features is borrowed from approaches to a so-called ‘Ising model’, the model where a phase transition in a mutual orientation of spins in a crystal is explained. The heating of magnetic crystal leads to abrupt disappearance of magnetic momentum, and an approximation of this phenomenon is simulated in the Ising model. In a very simplified form, the interactions in crystal are presented there as an increase of energy if two spins in a linear chain are oriented in the same direction, and a critical temperature of phase transition ($T_c$) is derived from a strength of these interactions (‘coupling constant’).

The cooling led in turn to a sudden appearance of magnetic momentum, and a decrease of a temperature close to a critical temperature leads to accumulation of ‘islands’, long enough fragments where spins are oriented in the same order. Ordered fragments in a one-dimensional spin chain can be compared to identical fragments of genome sequences, and a distribution of these fragments allows to describe there the features of self-affinity.

In the frames of statistical physics, an expression for the probability of island of length $k$ was obtained by Dziamagra\cite{1}, as an applied case of the so-called ‘Landau-Zener transition’:

$$ p(k, \tau) \sim e^{\tau k^2} $$

This is a point where a distribution with infinite dispersion can be introduced, assuming that a power coefficient $2$ in this Gaussian-like distribution is substituted to some floating power coefficient $D < 2$; a dimension of ‘intrinsic’ self-affinity of the underlaid process. The model constructed above depends on the two flexible parameters; intrinsic dimension $D$ and a parameter $\tau$, a rate of cooling, or a rate of approaching a transitional phase. This model allows us to explain over-dispersion of the genome coverage distributions mentioned above (Figure 1) and to fit the parameters to a measure of unevenness of human genomes, trying to reconstruct their evolution most precisely. The distribution of ‘islands’ for human genomes can be obtained as distribution of lengths of completely identical fragments in the genomes; lists of polymorphic positions (SNP) from “1000 genomes” project were used as a representation of genomes. A similar distribution of fragment sizes is observed for this data (Figure 2A; Table 1). The clusters for the three races are clearly seen for both genetic distance and for tails of ‘island’ sizes in genome-genome comparison (Figure 2B). To interpret this, higher unevenness relative to same genetic distance means higher ‘equilibration rate’, higher mutation rates, and a lesser slope of a fitting line.

Results
For the two independent populations, the distance does not depend on heterogeneity in populations; a simple model of exponential development can provide a dependency of average genetic distance within population.

$$ p_x = p_0 x + Tr_x \frac{1}{N_0} \sum_n (n \alpha^n) $$

$$ k \sim e^m = 1 + m $$

$$ d_x = p_0 + Tr_x N_g + Tr_y N_g $$

The simplified model of evolution of human ethnic groups is shown in Figure 3, and for further consideration it was reduced to just clarify a separation between three human races. In this case, the model can be further reduced to a system of equations (2). The events which are assumed here as events of separation between races are (a) separation between modern Asians and modern Europeans, which happened nearly just after an expansion to America, about ten thousand years ago; and (b) expansion of modern humans to Eurasia, about fifty thousand years ago.

$$ p_a = b_a p_a e + tk_a r $$
$$ p_e = b_e p_e + tk_e r $$
$$ d_{ae} = p_{ae} + 2tr $$
$$ p_A = b_A p_0 + (T + T) k_4 r $$
$$ p_{ae} = b_{ae} p_0 + Tk_{ae} r $$
$$ d_{ae} = p_0 + 2(T + t) r $$
Table 1. Approximated averaged similarities between human races and within a race, according to the detailed chart in Figure 2.

|                    | 'unevenness' m | genetic distance p |
|--------------------|----------------|--------------------|
| Europe Europe      | 0.0138         | 0.25               |
| Europe Asia        | 0.0144         | 0.27               |
| Asia Asia          | 0.0137         | 0.22               |
| Europe Africa      | 0.0152         | 0.29               |
| Asia Africa        | 0.0153         | 0.29               |
| Africa Africa      | 0.0131         | 0.24               |

The mutation rate $10^{-9}$ per nucleotide per year can be transformed as 3 per genome per year = 3000 per genome per thousand years, 3000 mutations per 700000 SNP, so that $r$ in the equations above should be about 0.004.

Having a requirement that $p_0 \geq 0$, the $r$ should be less than 0.0029. Rates of development are assumed to be unknown, what is only known is a dependency between a rate of development and a linear coefficient $m$. Values of $b_x$, $b_y$, $p_x$, $k_y$, which are attributes of a passed history are also assumed to be unknown.

For a marginal but the most confident assumption, $p_0 = 0$, $p_x = 0.21$, $k_y = 1.65$ and $k_y = 1.81$. Unevenness in a comparison between groups is a weighted average of evolutionary paths.
from a time of separation, so that if $kA \sim mA$, $k_e \sim me$, then $k_{ae} \sim (m_A T + m_A (T + t) + m_e t) / 2(T + t)$; $k_{ae} = 0.0141$. Following a log-linear approximation, $k_a$ and $k_e$, for modern Asians and Europeans, should be about 1.74 and 1.76.

The genetic diversity of Asians in a time of separation is then estimated as 0.15, much lower than a pool of genotypes just before the separation ($p_{ae} = 0.21$). For Europeans, the pool of genotypes was wider, about 0.20.

What is known is that Eurasia is a continent with good communications, and that it was populated mostly by ancestors of present-day Asian race before the time of last separation, or ‘crash’, in better words. It is also known that ancestors of modern Europeans began to expand to Europe mostly after that ‘crash’; the founders of the latest expansion originated from tribes which developed slowly for a long time somewhere in an area of central Asia mountains. The wide expansion to Eurasia for the ancient pre-Asian race was characterized, instead, by a substantive increase of genome unevenness. Some of it now is lost, some are kept in native Americans, and, for modern Asians, the instabilities in diverged genomes were neutralized by a long enough period of stable slow development after the crash.

Conclusions
Selection of individuals was almost the same as in Schiffels and Durbin, the difference with that model is that non-Gaussian features in genomes are considered here explicitly. This has a substantial influence on a reconstructed history of the three human races. Dealing with self-affine phenomena is difficult and risky, but it by no way can be ignored in any of valuable challenges to a present-day science.

Data availability
Underlying data
Zenodo: Measure of unevenness in human genomes: supplement software utilities and intermediate data files, http://doi.org/10.5281/zenodo.4495444.

This project contains the utilities to convert, pre-process and compare genotypes.

Data are available under the terms of the Creative Commons Zero “No rights reserved” data waiver (CC0 1.0 Public domain dedication).

The International Genome Sample Resource: 1000 Genomes phase 3 release, https://www.internationalgenome.org/data-portal/data-collection/phase-3

References
1. Kuzmin D, Feranchuk SI, Sharov VV, et al.: Stepwise large genome assembly approach: a case of Siberian larch (Larix sibirica Ledeb). BMC Bioinformatics. 2019; 20(Suppl 1): 37. PubMed Abstract | Publisher Full Text | Free Full Text
2. Peng CK, Buldyrev SV, Havlin S, et al.: Mosaic organization of DNA nucleotides. Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics. 1994; 49(2): 1685-9. PubMed Abstract | Publisher Full Text
3. Dziarmaga J: Dynamics of quantum phase transition: exact solution in quantum Ising model. arxiv.org. 2005; 95(24). Publisher Full Text
4. 1000 Genomes Project Consortium, Abecasis GR, Auton A, et al.: An integrated map of genetic variation from 1,092 human genomes. Nature. 2012; 491(7422): 56–65. PubMed Abstract | Publisher Full Text | Free Full Text
5. Schiffels S, Durbin R: Inferring human population size and separation history from multiple genome sequences. Nat Genet. 2014; 46(8): 919-25. PubMed Abstract | Publisher Full Text | Free Full Text
6. Sergey F: “Measure of unevenness in human genomes”: supplement software utilities and intermediate data files (Version 1.5) (Data set). Zenodo. 2020. http://www.doi.org/10.5281/zenodo.4495444
The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com