HapRFN: a deep learning method for identifying short IBD segments

Gundula Povysil[1], Djork-Arné Clevert[2], Sepp Hochreiter[1]

[1] Institute of Bioinformatics, Johannes Kepler University Linz, Austria, [2] Bayer Pharma AG, Berlin, Germany

Abstract

A segment of DNA is called identical by descent (IBD) in two or more individuals if it is identical because it was inherited from a common ancestor. IBD segments can be used to uncover hidden familial relationships, detect the population of origin of an individual or find interbreedings between humans and ancient hominins like the Neandertal. IBD segments can also be used to find the cause of diseases via a technique called IBD mapping.

For the above applications HapFABIA was shown to be superior to other IBD detection methods by detecting short IBD segments that are tagged by rare DNA variants via biclustering. Nevertheless, HapFABIA still has several problems: (1) To decide whether individuals possess an IBD segment is often difficult because of the soft bicluster membership supplied by HapFABIA. (2) HapFABIA can only extract 10-30 IBD segments at once and therefore needs to perform multiple iterations. However, the IBD segments identified in different iterations may not be decorrelated, thus they may be redundant and overlapping or even split into smaller segments. (3) Processing very large data sets is time intensive.

We recently introduced “Rectified Factor Networks” (RFNs) as an unsupervised deep learning approach. Each code unit of the RFN represents a bicluster and therefore an IBD segments, where samples for which the code unit is active share the bicluster (IBD segment) and features (DNA variants) that have activating weights to the code unit tag the IBD segment. HapRFN overcomes the problems of HapFABIA. (1) RFNs provide sparser codes via their rectified linear units that immediately supply bicluster memberships as factors being different from zero. (2) RFNs can learn thousands of factors and therefore many IBD segments simultaneously. Therefore, all IBD segments are mutually decorrelated, thus are not redundant and do not overlap. (3) RFNs allow for much faster processing of very large data sets using techniques from deep learning like efficient matrix multiplications and implementations of networks on graphical processing units (GPUs).

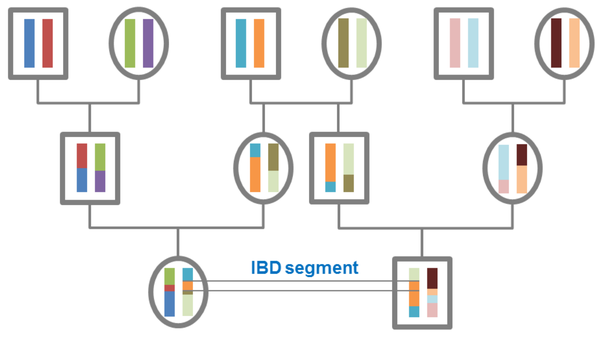
To keep feature membership vectors sparse, we introduce a Laplace prior on the parameters. Therefore, only few features contribute to activating a code unit, that is, only few features belong to a bicluster. In order to enforce more sparseness of the sample membership vectors, we introduce dropout of code units. Dropout means that during training some code units are set to zero at the same time as they get rectified. Dropout avoids co-adaptation of code units and reduces their correlations.

As a result HapRFN makes it possible to process very large data sets and to determine the size and number of IBD segments more precisely. With HapRFN we are able to accurately detect familial relationships, populations of origin, or interbreeding with ancient genomes in data sets with thousands of individuals. Furthermore, finding disease associations via IBD mapping becomes more reliable which might be the key to uncover unknown hereditary causes of multifactorial diseases.

Identity by Descent (IBD)

IBD (identity by descent):
identical DNA sequence in 2 or more individuals because it was inherited from a common ancestor

SNVs (single nucleotide variants):
small differences between DNA sequence of sample and reference, can be used to distinguish individuals, shared if sequence is IBD

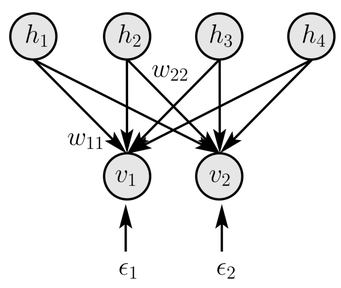


RFNs

Rectified Factor Networks:

$$\boldsymbol{v} = \boldsymbol{W}\,\boldsymbol{h} + \boldsymbol{\epsilon}$$

- $\boldsymbol{v} \in \mathbb{R}^m$: visible units/observations
- $\boldsymbol{h} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$: hidden units
- $\boldsymbol{W} \in \mathbb{R}^{m \times l}$: weight matrix
- $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Psi})$: additive noise

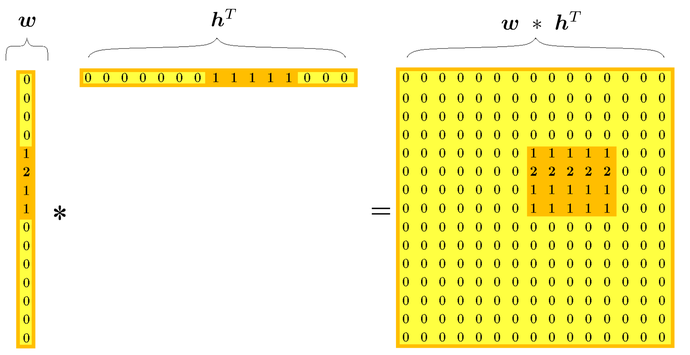


- rectified posterior mean → non-negative and exact zero code units
- dropout → sparse code units
- generalized alternating minimization algorithm
- based on posterior regularization method

Biclustering – IBD Segments

bicluster: pair of sample set and feature set for which samples are similar to each other on the features and vice versa

IBD segment = bicluster to which
- **SNVs (features)** belong if they tag the IBD segment (tagSNVs)
- **individuals (samples)** belong if they share the IBD segment



HapFABIA

- biclustering of SNV data to find short IBD segments shared by multiple individuals
- based on low-frequency and rare SNVs (minor allele frequency < 5%)

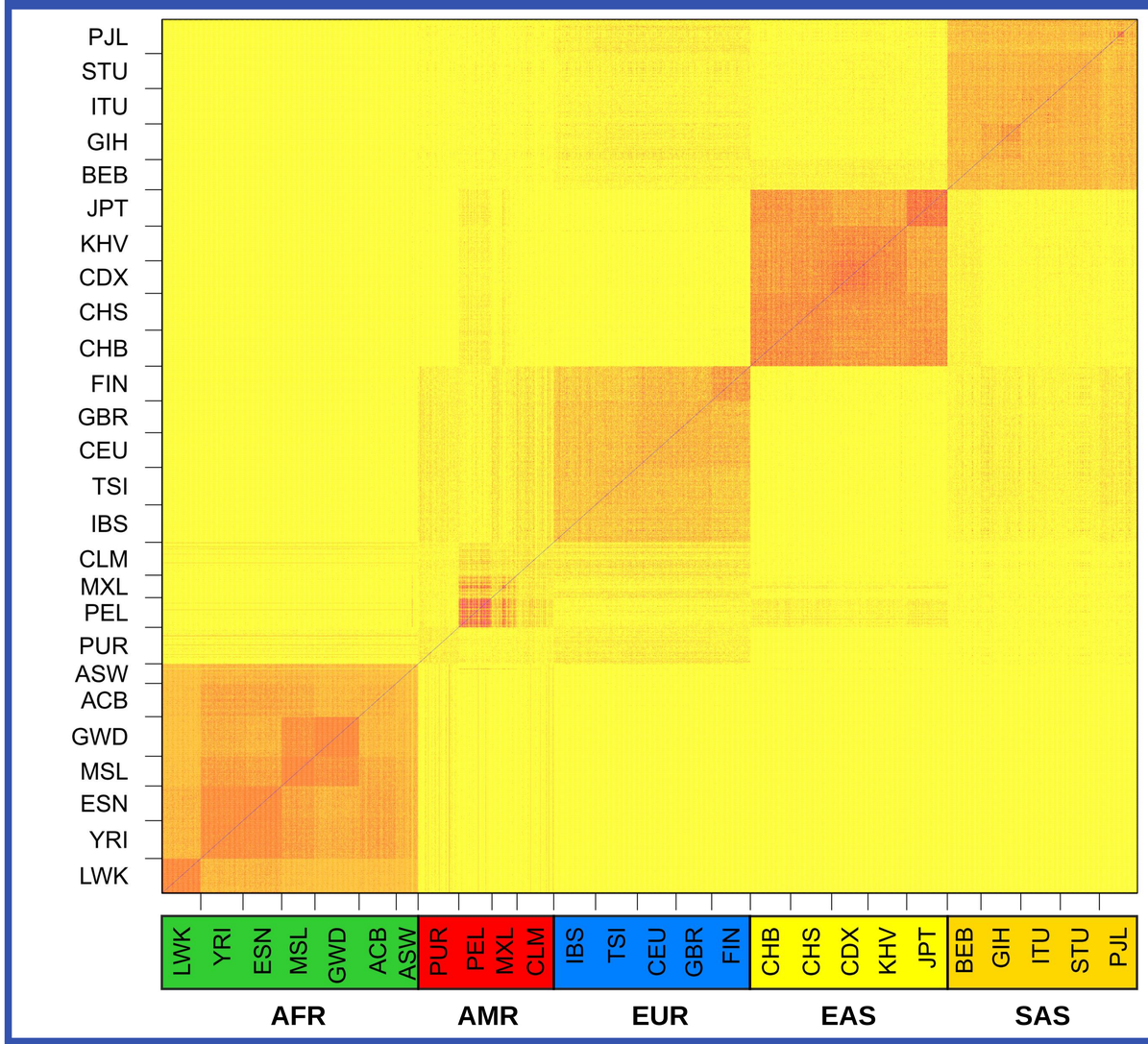
Problems:
(1) soft bicluster membership: deciding whether individuals possess IBD segment is difficult
(2) multiple iterations: HapFABIA can only extract 10-30 IBD segments at once, IBD segments identified in different iterations may not be decorrelated, may be redundant and overlapping or even split into smaller segments
(3) time intensive

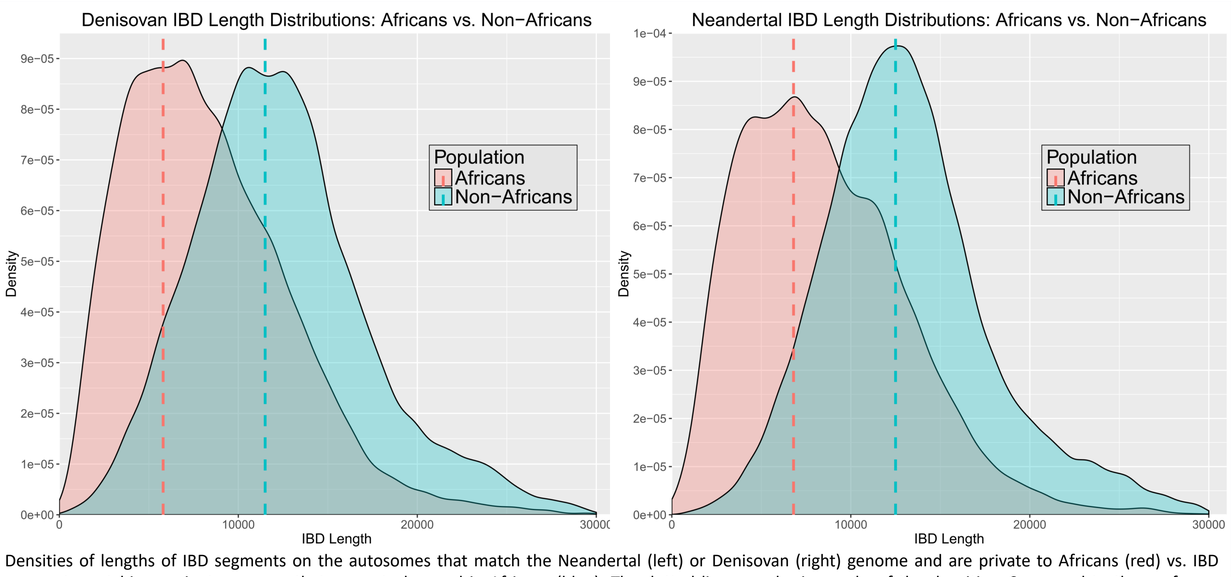
HapRFN

- each code unit of RFN represents bicluster/IBD segment
- individuals for which code unit is active share IBD segment
- features (SNVs) that have activating weights to code unit tag IBD segment

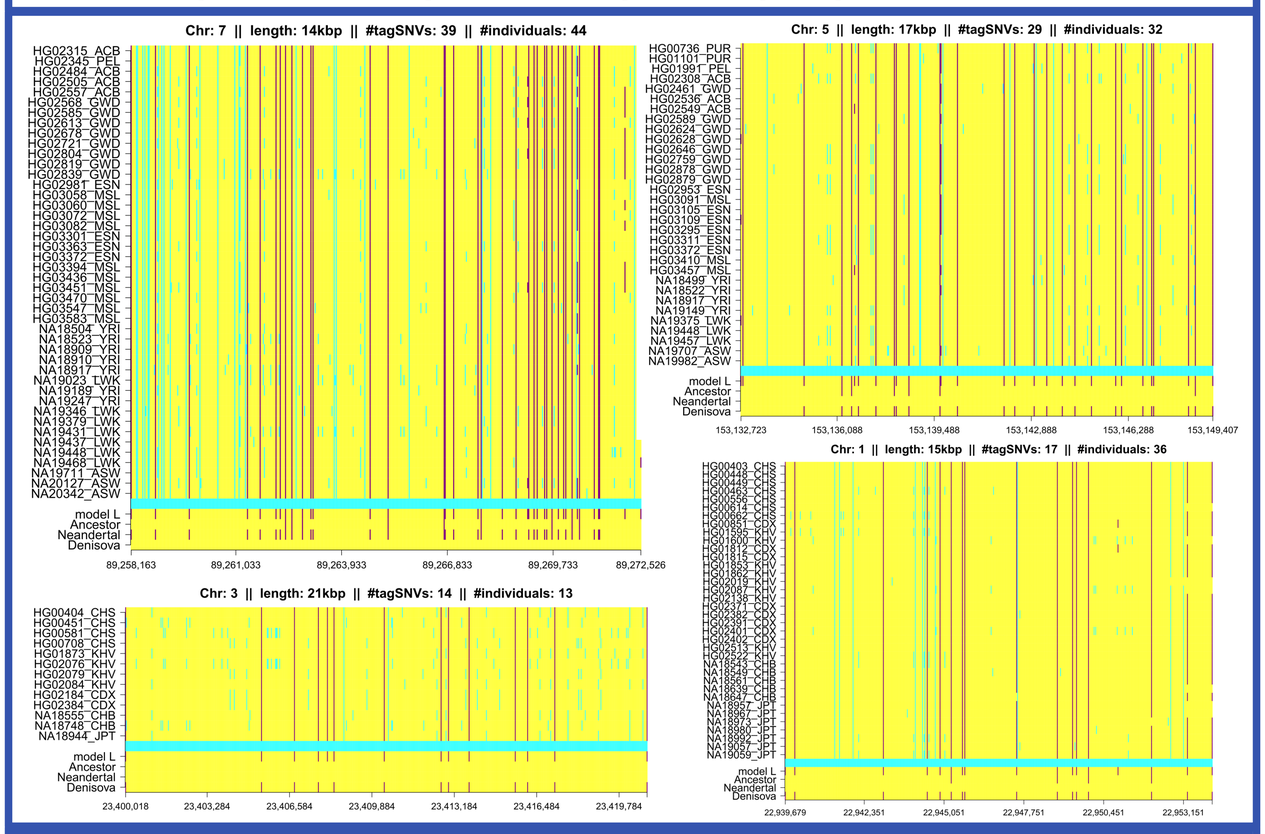
HapRFN overcomes the problems of HapFABIA:
(1) RFNs provide sparser codes via their rectified linear units that immediately supply bicluster memberships as factors being different from zero.
(2) RFNs can learn thousands of factors and therefore many IBD segments simultaneously.
(3) RFNs are much faster through techniques from deep learning like efficient matrix multiplications and implementations of networks on graphical processing units (GPUs).

Results





Densities of lengths of IBD segments on the autosomes that match the Neandertal (left) or Denisovan (right) genome and are private to Africans (red) vs. IBD segments matching ancient genomes that are not observed in Africans (blue). The dotted lines emphasize peaks of the densities. Compared to those of non-Africans, African IBD segments that match an ancient genome are enriched in regions of shorter segment lengths.



Conclusion

- rectification → easier interpretation of bicluster membership
- simultaneous detection of thousands of IBD segments
- better separation of IBD segments → less intensive post-processing
- main results unchanged but more reliable

Clevert D.-A., Mayr A., Unterthiner T., Hochreiter S. (2015) Rectified factor networks. Advances in Neural Information Processing Systems 28 (NIPS 2015), eds. Cortes C., Lawrence N. D., Lee D. D., Sugiyama M., and Garnett R. (Montreal, QC), 1846–1854.
Gunawardana A., Byrne W. (2005) Convergence theorems for generalized alternating minimization procedures. Journal of Machine Learning Research, 6:2049–2073.
Ganchev K., Gillenwater J., Taskar B. (2010) Posterior regularization for structured latent variable models. Journal of Machine Learning Research, 11:2001–2049.
Hochreiter S. (2013) HapFABIA: Identification of very short segments of identity by descent characterized by rare variants in large sequencing data. Nucleic Acids Res. 41:e202.
Povysil G., Hochreiter S. (2016) IBD Sharing between Africans, Neandertals, and Denisovans. Genome Biol Evol. http://dx.doi.org/10.1093/gbe/evw234

Acknowledgements: We thank the NVIDIA corporation for the GPU donation that facilitated our work.