Human Admixture Pipeline

Maarja Lepamets

Curriculum: Computer Science 2014/15
Institute of Computer Science
Faculty of Mathematics and Computer Science
Tartu University

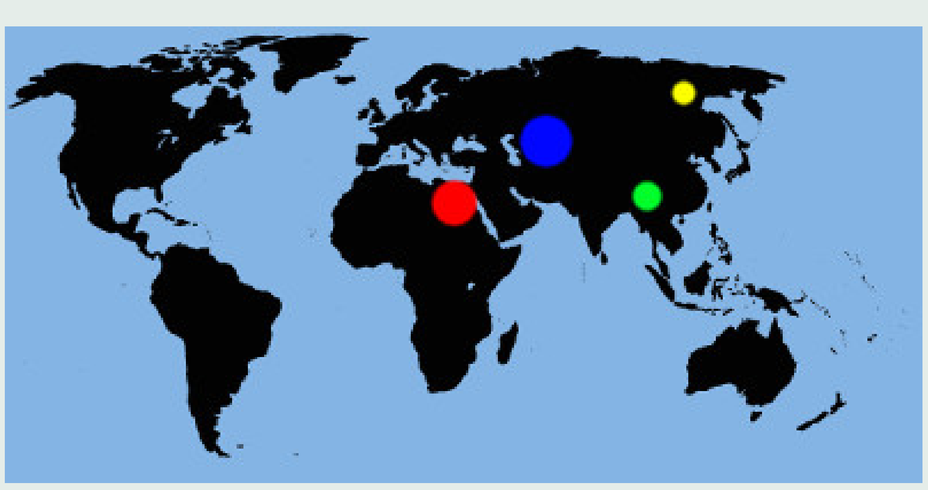
Project repository: *https://github.com/maarjalepamets/human-admixture*

Background and motivation

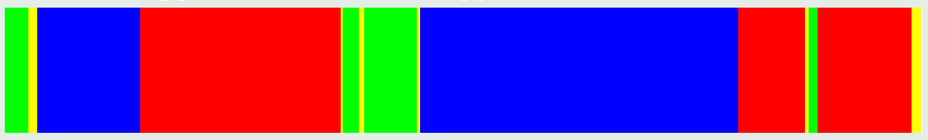
Human populations have been in interaction with each other throughout history. This suggests a steady exchange of genetic information in which new populations have developed by admixture of existing populations. It is difficult to determine the linkage and ancestry of populations but an approach has been made by Hellenthal *et al.* (2014) by using just genetic data. In particular, they have tried to identify similarities in chromosome haplotypes (groups of successive SNPs) of individuals in different populations. Numerous tools (Beagle from Browning (2007); Chromopainter and Globetrotter from Lawson *et al.* (2012)) have been used in the process of analyzing the genetic data. This project has focused on pipelining the use of such tools in order to create a single tool with which to study human admixture.

Illustrative example

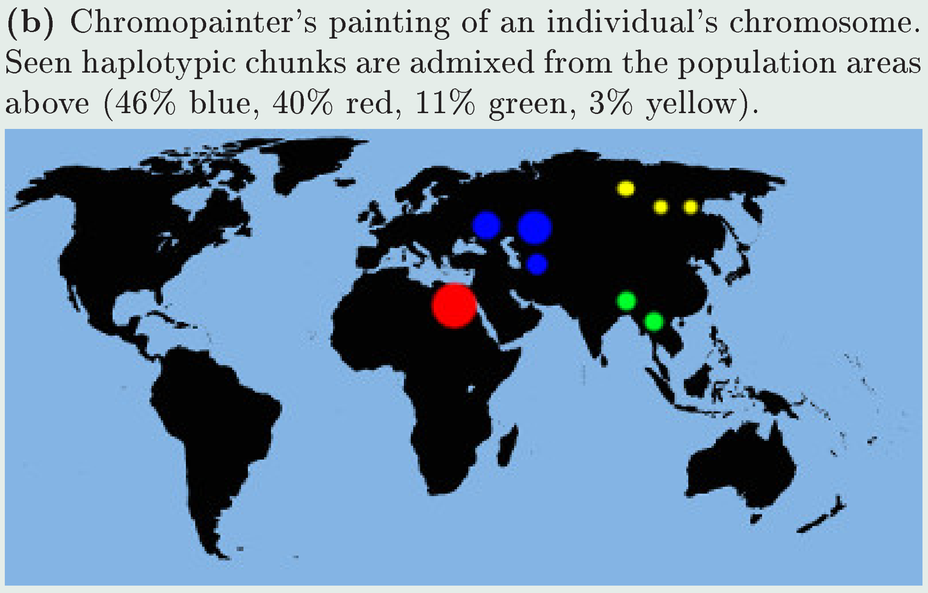
Below is a small fictitious illustration about what can be analyzed using the pipeline described on the right. Unfortunately, we did not have the permission to show the results on real data.



(a) True sources (populations). Colors indicate different donors: bigger sources have bigger markers.



(b) Chromopainter's painting of an individual's chromosome. Seen haplotypic chunks are admixed from the population areas above (46% blue, 40% red, 11% green, 3% yellow).



(c) Globetrotter's inference about which areas (populations) have been involved in the admixture.

Figure 1: (a) True sources (populations); (b) Chromopainter's painting of chunks of haplotypes; (c) Globetrotter's inference about admixture sources.

Pipeline

The pipeline assumes .fam and .bed files which are output files of whole genome association analysis toolset PLINK and consist of unphased haplotype data. It also assumes sources (populations) which are used as donors and recipients to perform the analysis. The pipeline is as follows:

1. Convert .bed to .ped, a file format used in PLINK
2. Convert .ped to .vcf, a file format used in BEAGLE v4 (modified code from The PyPedia Project (2012, 2013) is used here)
3. BEAGLE v4: use .vcf-file to phase haplotypes and output .phased.vcf. This may take a considerable amount of time, therefore phasing is done in parallel:
   (a) Phase haplotypes in different chromosomes in parallel
   (b) Combine haplotype data back together
   (c) Output phased haplotype data in file .phased.vcf
4. Use CHROMOPAINTER v2 to paint the haplotypic chunks i.e. assing chunks to given donor sources (populations). Create the necessary input files:
   (a) .haplotypes – genetic variation information of donors and recipients, convert from .vcf
   (b) .recomrates – SNP positions and genetic distances, convert from .vcf
   (c) .poplist – population file created from given donor and recipient sources
   (d) .idfile – population and identifier labels for individuals, convert from .fam
5. Use GLOBETROTTER to analyze admixture events and dates. Feed in the necessary input files:
   (a) .chunklenghts.out – Total length of DNA that individuals copy from donors, Chromopainter's output file
   (b) .samples.out – Donor haplotypes copied at each SNP of recipient haplotypes, Chromopainter's output file

References

Browning S R and Browning B L (2007) Rapid and accurate haplotype phasing and missing data inference for whole genome association studies by use of localized haplotype clustering. Am J Hum Genet 81:1084-97. doi:10.1086/521987

Hellenthal, G., Busby, G.B.J., Band, G., Wilson, J.F., Capelli, C., Falush, D. and Myers, S. (2014) Genetic Atlas of Human Admixture History Science 343:747-751

Lawson, D., Hellenthal, G., Myers, S., and Falush, D (2012) Inference of population structure using dense haplotype data PLoS Genet 8(1):e1002453

The PyPedia Project 2012, 2013, http://www.pypedia.com