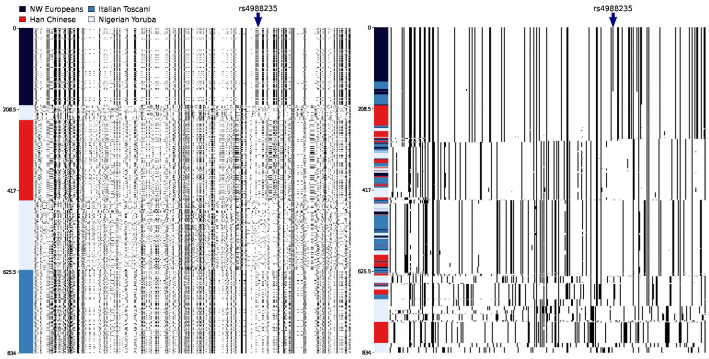
Haplostrips: revealing population structure through haplotype visualization

Marnetto D(1,2,3)[†] , HuertaSánchez E(4)

(1)Department of Molecular Biotechnology and Health Sciences, University of Turin, , Turin, Italy
(2)Department of Integrative Biology, University of California, , Berkeley, CA, USA
(3)Current Affiliation: Estonian Biocentre, Institute of Genomics, University of Tartu, Tartu, Estonia
(4) School of Natural Sciences, University of California Merced, Merced, CA, USA

[†] Email: davide.marnetto@ut.ee

Motivation
Population genetic analyses often identify polymorphic variants in regions of the genome that indicate the effect of nonneutral evolutionary processes. However, in order to obtain deeper insights into the evolutionary processes at play, we often resort to summary statistics, sacrificing the information encoded in the complexity of the original data. In addition, this simplification might represent a risk when leveraging hits Genome Wide Association Studies hits without considering the population structure at those loci. Here, we present haplostrips, a tool to visualize polymorphisms of a given region of the genome in the form of independently clustered and sorted haplotypes,
Methods
Haplostrips is a commandline tool written in Python and R, that uses variant call format files as input and generates a heatmap view. It is available at:
https://bitbucket.org/dmarnetto/haplostrips. It can already be applied in several fields and in all living systems for which a phased haplotype is available to visualize complex effects of, among others: introgression, domestication, selection, demographic events. An additional step has been implemented to apply Haplostrips to sparse genetic data where genotypes can be completely missing, e.g. ancient genomes.
Results
Haplostrips is able to sort and cluster haplotypes, using only the distance between the genetic sequences, regardless of the metainformation supplied. This turns the disorganized heatmap, of which an example is represented on the left of Fig. 1, into an informative plot that reveals hidden haplotype structures, as seen on the right of Fig. 1. This meaningful plot is delivered without losing the basic information encoded in variant sequences. We present the LCT region as a case study to underline the utility of fully considering this basic information in order to inform scientific hypotheses and decisions, leveraging both modern and ancient genomic data.


NW Europeans
Italian Toscani
Han Chinese
Nigerian Yoruba
rs4988235
0
208.5
417
625.5
834
rs4988235
0
208.5
417
625.5
834