Haplostrips: revealing population structure through haplotype visualization

Davide Marnetto[1,2] Emilia Huerta-Sánchez[3]

January 25, 2017

[1] Dept. of Molecular Biotechnology and Health Sciences, University of Turin, Italy

[2] Dept. of Integrative Biology, University of California Berkeley, CA, USA

[3] School of Natural Sciences, University of California Merced, CA, USA

Abstract

1. Population genetic analyses often identify polymorphic variants in regions of the genome that indicate the effect of non-neutral 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.

2. 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. *Haplostrips* is a command-line tool written in Python and R, that uses VCF files as input and generates a heatmap view.

3. *Haplostrips* is available at: https://bitbucket.org/dmarnetto/haplostrips. It can 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.

4. *Haplostrips* can reveal hidden patterns of genetic variation without losing the basic information encoded in variant sequences.

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/2041-210X.12747

1 Introduction

A haplotype is an arrangement of specific alleles occurring in the same chromosome within a given genetic segment. Often genetic variation is studied through summaries of single nucleotide polymorphisms (e.g. frequency of mutations). However, haplotypes provide more information because we can see the combination of alleles that are present on a single chromosome. Access to tools to examine the complete patterns of genetic variation in these regions, and not just summary statistics, will help further elucidate the underlying evolutionary processes.

To our knowledge, the first tool to fulfill this need is *inPHAP* (Jäger *et al.*, 2014). It features a graphical interface to show sequences of alleles and to aggregate them interactively and according to meta information. Although it is useful for basic haplotype visualization and analysis, it requires manually supplying groupings to observe summaries, and does not provide allele polarization, dataset merging, and a command-line environment. Therefore, *inPHAP* is less suitable for automated hypothesis generation in population genetics.

Other methods to visualize haplotype structure provide insights on the abundance of several haplotypes in a population (Paradis, 2010), or on the linkage disequilibrium between variants (Barrett *et al.*, 2005). However, these methods generate a data summary, and do not show the full sequence of variants directly. Here we present a tool to assist researchers in visualizing the polymorphisms of a given region of the genome. In particular, this tool provides the user a few options to reveal hidden haplotype structure that may not be apparent when the haplotypes are plotted in a random order. Therefore, beyond being a visualization software, *Haplostrips* can be used to gain information about the evolutionary processes responsible for the observed haplotype patterns (e.g. positive selection, introgression etc.).

2 Description

Haplostrips is a command-line tool written in Python and R. It takes advantage of the preexisting Python package *Pandas* (McKinney, 2011) and the R package *gplots* (Warnes *et al.*, 2015) to manage input data and draw the plot, respectively. The software handles variation data, selecting the window of interest, extracting the haplotype data from the phased genotypes, polarizing variant sites and filtering them for mapping and genotype qualities. It keeps the samples belonging to populations of interest and chooses only the most informative sites, eliminating variations with very low frequency in all the populations to be plotted. Finally it produces a heatmap plot that displays the haplotypes in rows while each column represents a SNP within a region of interest. Haplotypes are labeled with a color defined by metadata, e.g. populations, from a file supplied by the user. Derived alleles are represented as black spots and ancestral alleles are represented as white spots (see Figure 1).

A key feature of *Haplostrips* is being able to sort and cluster haplotypes using only the distance between the genetic sequences, regardless of the meta information supplied. This turns the disorganized heatmap, of which an example is represented in Figure 1 A, into an informative plot that reveals hidden haplotype structures, as seen in Figure 1 B. Bringing together similar haplotypes and ordering them with respect to a reference has been proven productive in previous work (Huerta-Sánchez *et al.*, 2014) where visualization of the data led to the observation that the haplotype at high frequency in Tibetans originated from another population, a conclusion that was not evident from statistical summaries of the data. Also, an early version of this tool has been applied in a recent project (Racimo *et al.*, 2017, 2016), to substantiate Denisovan and Neanderthal introgressions in modern human populations.

2.1 Input

The user can supply as input a VCF genotype file, similar to those produced by the 1000 genome project (The 1000 Genomes Project Consortium, 2015). This format has become a standard for genetic variation data, and this makes our tool portable, versatile and simple to use. In addition, where Tabix indexes are available, *Haplostrips* uses the *pysam* package (Li *et al.*, 2009) to perform a fast retrieval of the region of interest. More VCF files can be supplied to the tool, which is capable of merging them using the reference allele of variants present in one VCF to infer missing data in others, or simply working on the intersection. *Haplostrips* can run iteratively over many windows of interest supplied with a file that contains the genomic intervals and populations to be plotted: this can be useful to visualize windows resulting from genome wide scans, e.g. GWAS. Another accepted input is the format generated by *ms* (Hudson, 2002), a widely used software to generate samples under a variety of neutral models. This feature allows one to observe the direct effects of particular demographic histories without any further parsing or coding.

2.2 Clustering, sorting and other options

The clustering is optional and is performed hierarchically via the single agglomerative method based on Manhattan distances, using the *stats* library in R (R Core Team, 2013). The Manhattan distance is simply the number of SNPs with different alleles in two sequences. The clustering brings together similar haplotypes, generating a thicker row in the plot for those that are more abundant, and a thicker row in the label column for the populations where they are more represented. The resulting dendrogram, which can be visualized as well, is re-ordered by decreasing similarity with a reference haplotype or a consensus sequence of a defined population. The reordering is performed using the minimum distance method, to ensure that the closest haplotype to the reference is always shown at the top. Available ordering options also include (1) performing the

clustering after the population grouping, (2) sorting the haplotypes for increasing differences from the reference one or (3) keeping the input order. The first option can be used to investigate population specific Linkage Disequilibrium or other effects, while the second one allows the comparison the heatmap with a plot showing the increasing number of differences to the reference haplotype, also reported in a separate file. This method lets the user deal with a simpler quantity, avoiding the clustering step. The heatmap and the distances to the reference haplotype can be optionally output to tab-delimited files.

The user can define the populations or groups of interest to be plotted, which are associated to the samples by another input file. Alleles at variant sites can be polarized for ancestral/derived status, using the ancestral allele provided in the INFO field of the VCF file. Knowing the ancestral or derived state of the allele is important for understanding the time of arrival of the mutation in the human lineage, and informs what the correct evolutionary models need to be applied in analysing a dataset. Sites can be filtered for genotype and mapping qualities or for having a low intra-population minor allele frequency in all populations plotted, as cited above. This last filter is particularly important because usually only a small portion of the polymorphic sites is informative, while most of them have frequency so low that would result in nearly uniform columns (white or black) in the plot. As an example only 344 out of 2463 polymorphic sites were plotted in Figure 1, while all sites with a maximum within-population MAF below 0.05 were removed.

2.3 Choice of the region to be plotted

It is worthwhile to point out that *Haplostrips* is useful for inspecting local patterns. Selecting a region that is too long can make the interpretation and haplotype clustering difficult, to the point where the plot loses its meaning. Consequently *Haplostrips* is not optimized for large regions and the RAM usage is dependent on their dimension. Windows of around 1000 sites before the filtering steps tend to provide good resolution, though this depends on the nature

of the region that will be plotted and on the SNP density of your dataset.

3 Usage example

We show in Figure 1 B a sequence that encompasses the genes LCT and MCM6. The plot allows us to visually distinguish a haplotype at high frequency in all Europeans populations, but at very low frequency in Africans. Interestingly, the Italian Toscani population possess a different and more variable set of haplotypes, consistent with a higher incidence of lactose intolerance in this population. We can observe that East Asians carry an almost exact copy of the Northern European haplotype at moderate frequencies. However 2 out of 3 sites where they differ are rs4988235 and rs182549, which have previously been associated with lactose intolerance (Enattah *et al.*, 2002). This suggests that the haplotype background of this variant site originated before the European-Asian split, whereas the associated allele arose more recently. Note that this insight would not be derivable from a haplotype network view, for example, as the SNPs contributing to the haplotypes are not presented, and one cannot distinguish clearly between the selected European haplotype and the very similar moderate frequency Chinese haplotype. It remains to be determined what models of positive selection can lead to the high similarity of the selected haplotype in the Europeans and the closest Asian haplotype.

4 Conclusion

Haplostrips can be used to conduct exploratory analyses, confirm hypotheses about candidate regions, or even substantiate findings in scientific publications. It can be applied in all living systems for which haploid or phased diploid genotype datasets are available to visualize complex effects of, among others: introgression, domestication, selection and demographic events. Although existing tools already address the task of visualizing haplotypes, *Haplostrips* includes

the ability of independent haplotype clustering and providing meaningful plots without sacrificing the basic information encoded in the genetic sequences. *Haplostrips* is downloadable at https://bitbucket.org/dmarnetto/haplostrips

5 Authors Contribution

E.H.S. supervised the project, D.M developed the software, D.M. and E.H.S. wrote the manuscript and documentation.

6 Acknowledgements

We thank Tyler Linderoth, Stefan Prost, Fernando Racimo, Fergal Casey and Peter Wilton for their insightful comments and software testing. This work was funded by UC Merced startup funds and NSF-DEB 1557151 to E.H.S.

7 Data Accessibility

Data used in the Usage example section were downloaded from the 1000 Genomes Project (The 1000 Genomes Project Consortium, 2015) ftp site, following the link for Phase 3 VCF data at http://www.internationalgenome.org/data.

References

Barrett, J.C., Fry, B., Maller, J. & Daly, M.J. (2005) Haploview: Analysis and visualization of LD and haplotype maps. *Bioinformatics*, **21**, 263–265.

Enattah, N.S., Sahi, T., Savilahti, E., Terwilliger, J.D., Peltonen, L. & Järvelä, I. (2002) Identification of a variant associated with adult-type hypolactasia. *Nature genetics*, **30**, 233–237.

Hudson, R. (2002) Ms a Program for Generating Samples Under Neutral Models. *Bioinformatics*, **18**, 337–338.

Huerta-Sánchez, E., Jin, X., Asan, Bianba, Z., Peter, B.M., Vinckenbosch, N., Liang, Y., Yi, X., He, M., Somel, M., Ni, P., Wang, B., Ou, X., Huasang, Luosang, J., Cuo, Z.X.P., Li, K., Gao, G., Yin, Y., Wang, W., Zhang, X., Xu, X., Yang, H., Li, Y., Wang, J.J., Wang, J.J. & Nielsen, R. (2014) Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. *Nature*, **512**, 194–197.

Jäger, G., Peltzer, A., Nieselt, K., Jager, G., Peltzer, A. & Nieselt, K. (2014) inPHAP : Interactive visualization of genotype and phased haplotype data. *BMC Bioinformatics*, **15**, 1–14.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., Data, G.P. & Sam, T. (2009) The Sequence Alignment / Map format and SAMtools. **25**, 2078–2079. URL `https://github.com/pysam-developers/pysam`.

McKinney, W. (2011) pandas: a Foundational Python Library for Data Analysis and Statistics. *Python for High Performance and Scientific Computing*, 1–9.

Paradis, E. (2010) Pegas: An R package for population genetics with an integrated-modular approach. *Bioinformatics*, **26**, 419–420.

R Core Team (2013) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL `http://www.R-project.org/`.

Racimo, F., Gokhman, D., Fumagalli, M., Hansen, T., Moltke, I., Albrechtsen, A., Carmel, L., Huerta-Sánchez, E. & Nielsen, R. (2017) Archaic adaptive introgression in TBX15/WARS2. *bioRxiv*, **In Press**.

Racimo, F., Marnetto, D. & Huerta-Sánchez, E. (2016) Signatures of archaic adaptive introgression in present-day human populations. *Molecular Biology and Evolution*, **In Press**.

The 1000 Genomes Project Consortium (2015) A global reference for human genetic variation. *Nature*, **526**, 68–74.

Warnes, G.R., Bolker, B., Bonebakker, L., Gentleman, R., Liaw, W.H.A., Lumley, T., Maechler, M., Magnusson, A., Moeller, S., Schwartz, M. & Venables, B. (2015) *gplots: Various R Programming Tools for Plotting Data*. URL `http://CRAN.R-project.org/package=gplots`, r package version 2.17.0.

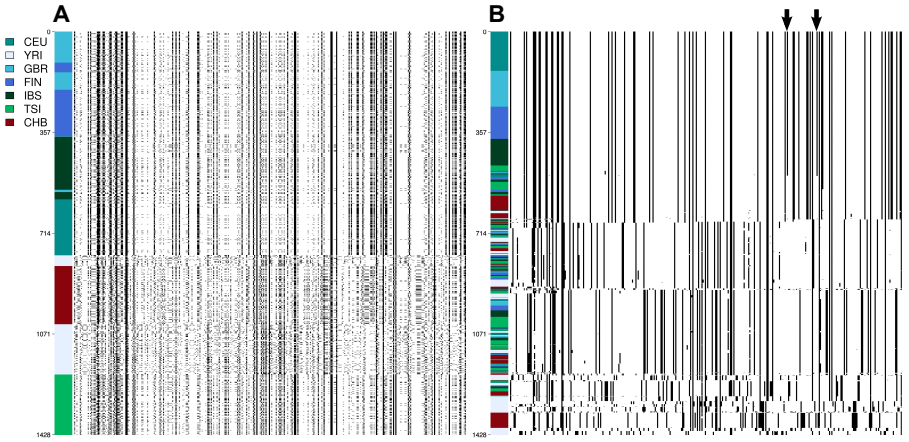

Figure 1: Haplostrips plot of LCT and MCM6: A) unsorted haplotypes, B) haplotypes clustered and sorted by increasing distance with CEU consensus. The arrows indicate the position of rs4988235 and rs182549, associated with lactase persistence. CEU = Utah residents with north-western European ancestry, GBR = British, FIN = Finnish, TSI = Toscani in Italia, IBS= Iberian in Spain, CHB = Han Chinese in Beijing, YRI = Yoruba in Ibadan, Nigeria