A Computational Analysis of Human Genetic Variation Highlights Signatures of Natural Selection in the Human Genome

Matteo Fumagalli$^{1,2}$, Uberto Pozzoli$^2$, Rasmus Nielsen$^3$, Manuela Sirioni$^2$, and Linda Pattini$^1$

1) Dipartimento di Bioingegneria, Politecnico di Milano
2) Bioinformatic Lab, Scientific Institute IRCCS “E. Medea”, Bosisio Parini (LC)
3) Department of Integrative Biology, University of California, Berkeley, USA

Abstract—As humans spread out of Africa they encountered a wide range of different environments and ecosystems, to which they may have been under selective pressure to adapt. Several studies have found immune and defense related genes to have been under positive selection in human evolution. However, little is known about the degree to which adaptation to local environment is primarily related to pathogens, or if most local adaptation is in response to other local environmental conditions.

By correlating genetic variation of more than 50 human populations with a set of environmental variables, we identified signatures of human local genetic adaptation. We showed that while majority of human variation can be explained by neutral factors, several loci show signatures of natural selection and this effect is mostly explained by adaptation to pathogen environment.

Detecting loci under natural selection owing to local adaptation may help addressing evolutionary history of modern complex diseases and optimize further designs of disease association studies to better unveil relationship between mutation and disorder susceptibility.

I. INTRODUCTION

During their evolution humans encountered a wide range of different environments and it is conceivable to think that they have been under strong selective pressures in order to adapt to them. Local genetic adaptation may be considered one of the most important factor shaping human genetic variation among different geographically distributed populations. Great phenotypic diversity across populations has a genetic basis and results from adaptive processes (Harris and Meyer, 2006) in response to changes in climate, diet or pathogen load.

With the advent of new genome-wide SNP data from many human populations, the interest and the potentiality of identifying signatures of positive natural selection and adaptation at the molecular level have greatly increased. Most of current strategies are unable to underline the environmental pressure that acted as selective pressure on the evolution of detected genes a priori. Additionally methods detecting polymorphisms whose frequencies are greatly correlated with environmental variables seem to be promising strategies to highlight signatures of human local adaptation, especially when beneficial variants have a weak phenotypic effect (Novembre and Di Rienzo, 2009).

Here we identify signatures of human environmental genetic adaptation and we attempt to disentangle the contributions of different environmental factors using methods based on correlations between population allele frequencies and environmental variables. More specifically, we quantify the main role played by pathogens in human adaptation, and identify specific pathways in the immune system mainly affected by the selective pressure imposed by different types of pathogens.

II. MATERIALS AND METHODS

A. Data retrieval

We investigated spatial allele frequencies using genotype data for 55 distinct human populations, comprising more than 1,500 individuals, by joining data from Human Genome Diversity Panel and from HapMap Phase III. A total of more than 500k SNPs were analyzed. SNPs were assigned as genic if they were located in transcribed regions or no further than 500 bp upstream transcription start sites. On the opposite, SNPs were assigned as intergenic if no genes has been mapped in a 100 kbp windows around the polymorphic site. We retrieved an ensemble of environmental variables and assigned correspondent values to each country. We chose a total of 14 environmental variables that may be good candidates to describe the overall pressure that humans had to adapt to: climatic or geographic factors (distance from the sea, mean annual temperature, mean annual precipitation rate, mean annual relative humidity, mean annual short wave radiation flux), subsistence strategies (relative amount of human activity spent in agriculture, animal husbandry, fishing, hunting, gathering) and pathogen diversity (number of different species of viruses, bacteria, protozoa, helminths).

B. Multiple regressions

We applied a Projection to Latent Structures (PLS, also known as Partial Least Squares) multiple regression to model the relationship between population allele frequencies of each SNP and a matrix describing environmental factors. For the model comprising all the environmental factors we applied an Uninformative Variable Elimination algorithm before the regression. In this way we could remarkably increase prediction accuracy by not considering predictors having very low regression coefficients. This algorithm (here named UVE-PLS) can handle highly correlated predictors and can effectively separate the weight of each predictor in the multiple correlation even in case of strong collinearity among variables (which is likely be the case for environmental factors).
For each regression we computed the cross-validated prediction accuracy ($Q^2$), estimated by a leave-one-out procedure. $Q^2$ can be considered a measure of the shift in allele frequencies among populations according to our set of environmental variables.

For each bin of the distribution of $Q^2$ we calculated the enrichment of genic and intergenic SNPs as previously proposed (Coop et al., 2009). Specifically, for each interval we computed the fraction of genic (or intergenic) SNPs divided by the total fraction of genic (or intergenic) SNPs. Confidence intervals were computed by bootstrap resampling of 200 kbp segments from the genome.

C. Partial correlations

For each gene we assessed the relationship between the locus-specific population genetic distances matrix and the distance matrix for environmental variables via partial Mantel correlations. This is a non parametric statistical test for association between two distance matrices, controlling the effect of a third matrix to remove spurious correlations. The latter independent distance matrix is calculated as the overall population genetic distances among populations computed over all loci and therefore it accounts for the non independence of populations and corrects for neutral demographic events. Specifically for each correlation we calculated the improvement of explained variance (here called $I(R^2)$) adding the environmental matrix to the correlation between the first and third matrix.

Statistical significance was assessed by permuting rows (populations) within the same stratus (continent) for the dependent matrix and recalculating our statistic for permuted data. To reduce computational time we computed approximated $p$ values using an asymptotic method (Knijnenburg et al., 2009). Briefly, $p$ values were computed approximating the upper tail of the distribution of permuted statistics by a Generalized Pareto distribution.

All computation were performed in the R environment.

III. RESULTS

A. Excess of genic SNPs for high values of prediction accuracy

We examined the relative abundance of genic versus intergenic SNPs in the upper tail of distribution of $Q^2$ values. Notably we found a significant enrichment of genic SNPs compared to intergenic SNPs located at high values of $Q^2$, suggesting the action of natural selection driving the differential allele frequency shift among human populations (Fig. 1). We also verified that this excess it not biased by confounding factors like ascertainment bias, background selection or population structure.

We then applied the same regression strategy to unveil the different contribution that distinct environmental factor classes had in in the excess of functional SNPs for high values of prediction accuracy. We ran multiple regressions considering separately only climate, subsistence strategies or pathogens predictors. Notably, even if we observed a higher abundance of genic compared to intergenic SNPs for high levels of $Q^2$ for each model, only including pathogens predictors leads to a striking and significant enrichment of genic SNPs.

B. Quantify the amount of selection given by environmental factors

Our next goal was to identify the relative fraction of loci whose population genetic variation is significantly correlated with specific environmental factors. These genes would be good candidate to be under selection under a particular selective pressure.

We assessed the relationship between the population genetic distance of each gene and the distance matrix for environmental variables via partial Mantel correlations. This is a non parametric statistical test for association between two distance matrices, controlling the effect of a third matrix to remove spurious correlations. The latter independent distance matrix is calculated as the overall population genetic distances among populations computed over all loci and therefore accounts for the non independence of populations and corrects for neutral demographic events.

While most of the genetic distance variance is explained by the pure population demography there is a non negligible contribution of environmental predictors. Several loci show non zero values of improvement of explained variance $I(R^2)$, suggesting a role of all environmental factors shaping human genetic variation.

Assessing the statistical significance, we observed a striking predominance of genes variation is significantly associated with pathogens distance matrix (194 loci) rather than subsistence (4 loci) or climate (no associated genes), indicating that pathogens have been the most important environmental factor and they shaped population variation at a larger number of genes. Focusing on pathogens factors, removing helminth richness from the model leded to a drastic decrease of $I(R^2)$ (Fig. 2). This may suggest that parasitic worms have had the strongest impact on human adaptation among infectious agents, while the opposite situation is observed for virus richness.
There could be several reasons to explain why we observed such a greater extent of selection given by adaptation to pathogens compared to other environmental factors. First this may effectively pinpoints pathogens as the main factor shaping human variation during evolution. Second, adaptation to diet or climate could involve weaker selection coefficients or multigenic mechanism more often. In these cases we presumably have less power to detect selection and development of advanced methods and further investigations are needed to fully understand molecular signatures of selection left by different kind of selective processes. Finally, selection target, and therefore number of loci and pathways somehow involved in a particular biological function, is more likely to be wider in the case of regulation of immune system than modulation of metabolism or heat stress response or other mechanisms regarding adaptation to climate.

V. CONCLUSION

Population genetics analyzes represent outstanding retrospective investigations on the role of evolutionary processes in shaping genetic variation and detecting putative loci subjected to natural selection. On the other hand, to clearly assess the adaptive significance of variants at a particular gene focused functional experiments are needed as direct insights into biochemical mechanisms of adaptation. Studies based on integrating population genetics and functional experiments on candidate loci provided direct evidences of the effect of selection on alternative splicing and protein structure (Cagliani et al., 2010) or gene expression. In particular evidences of the action of positive selection on gene expression regulation and the emerging of new technologies to detect expression quantitative trait loci (eQTLs) should encourage further studies in this direction.

In a clinical perspective, detecting loci that underwent a non neutral evolution owing to adaptation represents a valuable opportunity to address evolutionary history of modern complex diseases. In fact, there are evidences that variation on susceptibility genes for common diseases may be a result of changes in selective pressure during human history. Notably, findings from population genetics studies may optimize further designs of disease association studies to better unveil real relationship between allelic status and modulation of disorder susceptibility especially if the latter is controlled by genetic variation of multiple genes, each one having a modest contribution to phenotype.

These considerations are even more notable when assessing resistance to immune-related disorders. An enrichment of signatures of recent positive selection have been found investigating SNPs associated with autoimmune or infectious diseases by genome-wide association studies, pinpointing the evolutionary impact on fitness hold by immune diseases (Barreiro and Quintana-Murci, 2010). Simplest expectation is that positively selected alleles are responsible for an increased protection against infectious
diseases, and this is true for several examples (e.g. in malaria resistance). However cases of positively selected haplotypes associated with an increased risk of infection may states for a situation where those variants conferred resistance to a previous and different pathogen load. This is the principle of the hygiene hypothesis which assumed that humans have adapted to a past pathogen-rich environment that no longer exists in modern societies thanks to better environmental conditions and health care system (Sironi and Clerici 2010). Further deep analyzes of pathogen-associated genes and their effects on disease phenotypes may shed light both in the evolutionary history of modern complex diseases and in the association between alleles and disorders.

References