Genomic and Functional Approaches to Genetic Adaptation

Elena Carnero Montoro

TESI DOCTORAL UPF / 2013

Thesis Director

Dra. ELENA BOSCH

DEPARTAMENT DE CIÈNCIES EXPERIMENTALS I DE LA SALUT

UNIVERSITAT POMPEU FABRA

Fitxer PDF de la tesi dividit en 7 parts

Part 1 de 7 pàg 0 - 17	Introduction Cap. 1 – Cap 2
Part 2 de 7 pàg 18 - 20	Introduction Cap. 3 : 3.1, 3.2
Part 3 de 7 pàg. 20 – 25	Introduction Cap. 3 : 3.2.1
Part 4 de 7 pàg. 25 – 28	Introduction Cap. 3: 3.2.2 – 3.2.3
Part 5 de 7 pàg. 29 - 64	Introduction Cap. 4 – Cap. 6
Part 6 de 7 pàg. 65 - 183	Objectives, Results, Discussion, Concluding remarks
Part 7 de 7 pàg. 184 – 233	References, Annexes

Part 3 de 7

particularly, the recent availability of large catalogs of genetic variability in different human populations such as those provided by the HapMap or the 1000Genomes projects (Altshuler et al. 2010) (Abecasis et al. 2012) provided an unique opportunity to investigate the impact of natural selection on our genome.

Consequently, genome-wide scans of positive selection became possible and popular, and many large-scale studies have been published on different species and populations, searching for the footprints of different molecular signatures of positive selection both at specific lineage level and at population level. This genomic revolution has represented a great opportunity to overcome the limitations of candidate genes studies.

3.2.1 Population-based genome wide scans

Since Akey et al (2002) work, more than 20 genome-wide scans have been performed to detect recent or ongoing positive selection in humans, with an increase over time of the number of individuals, populations and markers included in the analysis.

The combination of results from all these studies have revealed in one or more studies more than 5000 regions of the genome with some type of signature of positive selection, encompassing more than 400Mb in the genome, and containing more than 4,000 UCSC RefSeq genes. Undoubtedly, this multiplication of candidate loci regarding the ones discovered by gene-by-gene studies, hold a promising chance to guide us to a richer understanding of where positive selection has shaped patterns of genetic variability (Akey 2009) and which traits have been important for the adaptation of each population.

It is important to state that most of the aforementioned studies have based their strategy for identifying targets on the outlier approach,

meaning, the computational construction of an empirical distribution based on specific statistical tests and on the identification of outlier regions as possible targets for positive selection (see figure 7). Applying this approach needs some strong assumptions, as for example the fact that demography is affecting the whole genome in the same way, and that positive selection is not frequent enough to produce values in the middle of the distribution. Another delicate issue is how to establish the threshold statistical significance. By default, it is set at the 99% or 95% percentile of the distribution.

Another way to assess the confidence of the results of genome-wide studies is to examine the overlap among outlier loci among studies. In a review by Akey in 2009, he only identified a 14%

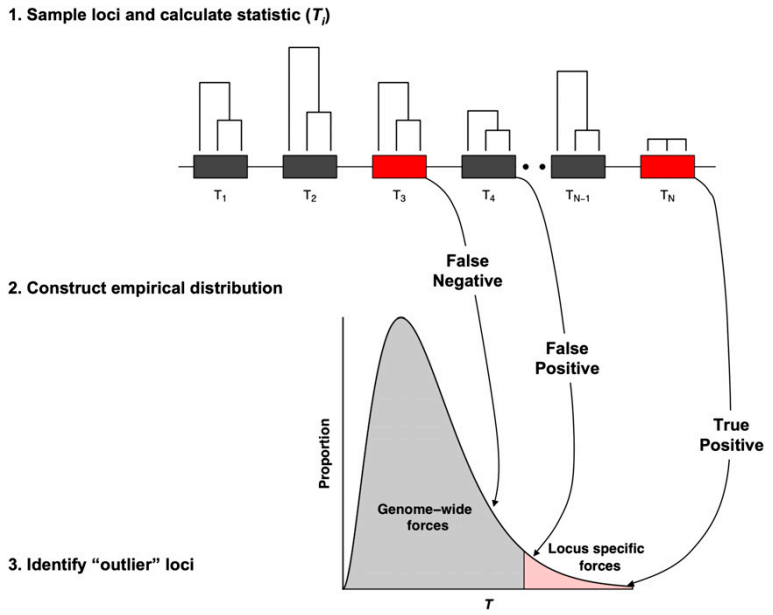

Figure 7. Population genomics study design for detecting positive selection. For each locus, a neutrality test is calculated and an empirical distribution of results is constructed. Outlier loci, affected by selection in a locus-specific mode, are identified as being targeted by positive selection. Because different genomic properties can affect coalescence processes in a way that resembles selection or obscuring the signal of selection, false positive and true positive

cases are likely to appear in the distribution (Akey 2009).

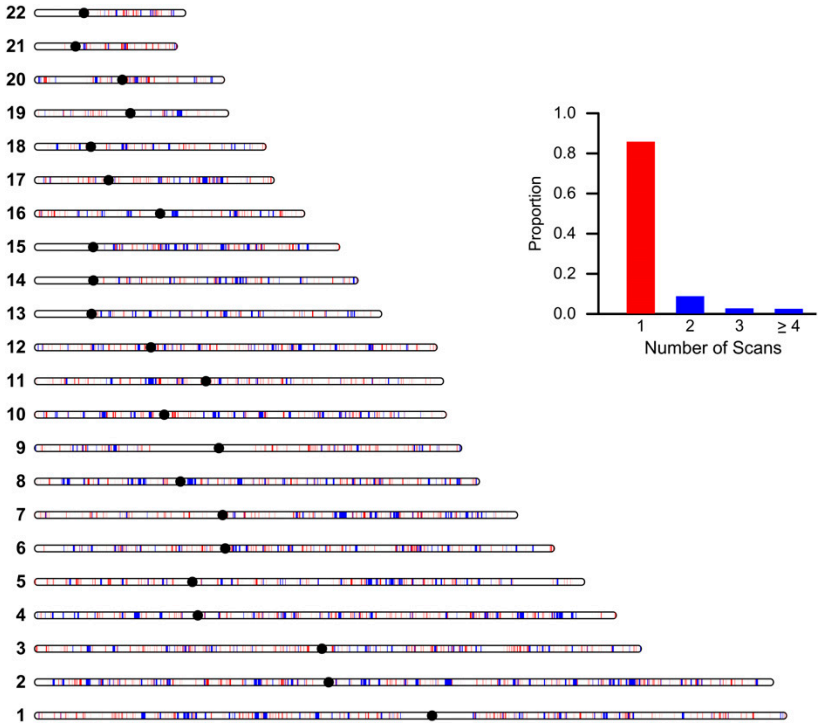

Figure 8. Integrated genomic map of positive selection. It includes results from 10 different scans. The colors represents the number of scans that have identified each target locus. As showed in the histogram, most of the loci were only identified in one scan (Akey 2009).

overlap among 10 different studies that had used the same type of data (figure 8). Furthermore, the combination of all of them did not include some of the best described examples of positive selection, such as the *G6DP* and *DARC* genes, probably due to the low variant density in the regions where they were located in these catalogs of variation (see figure 8).

Despite the small overlap, more than 700 regions have been identified in two or more studies. A number of the loci within these regions had previously been identified as candidates for adaptation,

and more interestingly, many new well-supported regions are available to be considered as shaped by positive selection.
The results of these studies show that much of the adaptation is not shared among populations, and that it only occurred recently during the human dispersal. This widespread local adaptation is not a surprise, since different populations have encountered very heterogeneous environmental conditions during their adaptation history.

Maps of positive selection hold great promise for achieving a detailed description on how natural selection has shaped our genomes and they give examples of possible new targets of positive selection. However they also represent big challenges to understand what really happened in our species involving those identified elements. Apart from the little overlapping among different studies that has already been discussed, these studies present other main limitations:

i) Loci identified as outliers may only represent the most extreme cases of positive selection, and many of those which played a role in adaptation - through lower selective coefficients- are surely presented as false negatives. On the other hand, false positives may also be prevalent in the tail of the distribution, due to different possible coalescent histories along the genome.

ii) Because most of the scans were based on genotype data, two additional problems may arise:
   a. Many of the statistical tests were developed to be applied to sequence data, not to genotype data. Since SNPs genotyped in high-throughput arrays are usually common variants, rare alleles are surely missing in the dataset, thus affecting the performance of some neutrality tests as those based on the SFS (see previous section for details on these tests).
   b. The lack of a complete variant catalog might imply that

the identification of the actual targeted variant is impossible.

iii) Regions detected as targets of selection are large and they span hundreds of kilobases. While many of the regions contain several genes and the actual target of selection is difficult to identify, other regions are intergenic or isolated without any known genetic elements which makes it impossible to speculate about their role on selection.

iv) For most genes, there is no biological evidence of which phenotype subjected to selection they might influence.

In summary, genome-wide studies are powerful resources to begin disentangling the impact of positive selection on the genome. They have allowed the identification of a large number of putative candidate genes, but they are surely not the end of the story since, in most cases, they do not provide an explanation about the adaptive processes behind them.

While the majority of selection scans where based on SNP data catalogs, basically the ones released by HapMap and Perlegen projects (Altshuler et al. 2010), the recent publication of the sequence data from the 1,000 genomes project (Abecasis et al. 2012) is allowing a new generation of scans which can use the full catalog of human variation obtained from 1,092 individuals from different worldwide populations.

The recent genome-wide study by Grossman et al. (2013), represents a big step forward to overcome the aforementioned limitations and towards the identification of putative selected variants. The progress made by them was possible due to several facts. First, they used a composite likelihood method (CMS) that combines different population genetics statistical tests in one, narrowing down from 20 to 100 times the putative candidate

region (Grossman et al. 2010). Second, they applied their method to full sequence genome variation, from the recently published phase1 data from the 1,000genome project, and thus, they were able to explore the complete allele spectrum of different populations to identify putative selected variants. Third, they took advantage of the latest publications of the Encyclopedia of DNA Element (ENCODE) Consortium (Dunham et al. 2012), as well as of the extended genome-wide association (GWA) studies database to appraise the functional relevance of the variation detected as a target of positive selection.

The ENCODE Consortium has provided a very rich catalog of functional annotated elements, identified by a variety of different high-throughput next generation sequencing technologies, as for example the are the characterization of novel non-coding elements, enhancers, patterns of methylation, expression, chromatin states, of the complete human genome of different tissues and cell lines.

This project by Grossman et al (2013) represents a decisive shift in the research of evolutionary adaptation since it is the first one that provides a rich list of positively selected candidate variants as well as a description of their possible functional predicted impact, as described in figure 11, which will be the base of generating new hypotheses about adaptation and new research lines in human evolutionary genetics.

3.2.2 Comparative genome-wide scans

As explained in the previous section, statistical tests using divergence data compare rates of potentially functional and non-functional replacements between species, and try to identify the proportion of functional substitutions that have been fixed due to positive selection. The classical debate among evolutionary biologists about what proportion of differences between different species have been fixed because they were beneficial and because