

Identifying footprints of selection in genome-wide SNP data from haplotype structure REHH

Mathieu Gautier

UMR INRA/CIRAD/IRD/SupAgro CBGP

18th October 2016

REHH Overview

Comparisons of local haplotype diversity

- **iHS** : Within-Population ("outliers")
- **rSB** and **rSB** : Between populations (pairwise)
- Empirical (model-free) approach : calibration of the statistics is based on the observed distribution

Haplotype Data

- Need a prior phasing of the data (with e.g., FASTPHASE, BEAGLE, SHAPE-IT) and optionnally ((but recommended) **imputation** (with e.g., FASTPHASE, IMPUTE)
⇒ 3 data format allowed in REHH
- Map information (genetic > physical distances)
- For **iHS** only : ancestral/derived allele status of SNPs

General principles

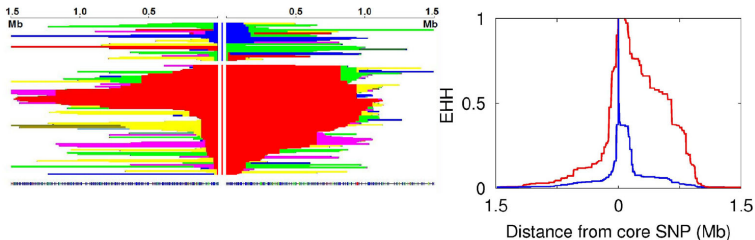
- Selection (when recent) tends to reduce haplotype diversity
(\Leftrightarrow) increase Linkage Disequilibrium
- Extended Haplotype Homozygosity (Sabeti *et al.*, 2002)
 - EHH : An estimator of Haplotype Variability
 - Probability that two randomly chosen chromosomes carrying the core allele of interest are IBD for the entire interval from the core region to a distance x (assuming SNP homozygosity=IBS \Rightarrow IBD)
 - EHH detects the transmission of an extended haplotype without recombination

$$\bullet \quad EHH_i(x) = \frac{\sum_{k=1}^{h_i} \binom{n_k^{h_i}}{2}}{\binom{n_i}{2}} = \sum_{k=1}^{h_i} \frac{n_k^{h_i} (n_k^{h_i} - 1)}{n_i (n_i - 1)} \quad \left(\simeq \sum_{k=1}^{h_i} \left(\frac{n_k^{h_i}}{n_i} \right)^2 = \sum_{k=1}^{h_i} (f_k^{h_i})^2 \right)$$

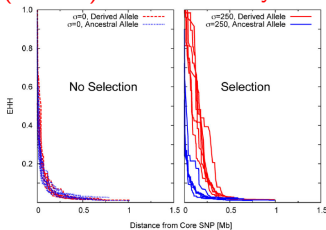
- h_i : Nb. of unique haplotypes carrying the core allele
- $n_k^{h_i}$: Nb. of each unique haplotypes ($f_i^{h_i}$ = observed freq)
- $n_i = \sum_{k=1}^{h_i} n_k^{h_i}$: Nb. of haplotypes carrying the core allele

EHH as a function of the distance to core allele

By construction EHH at a core allele decreases with distance



(Recent) selection decays the decrease



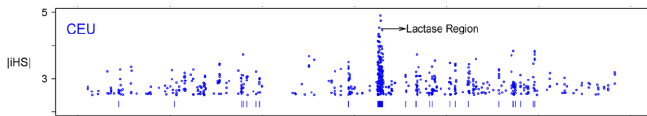
Genome wide scan

Limitations of EHH

At the genome scale, decrease depends on recombination rate ("cold spot" might mimic selection) \Rightarrow Difficult to compare across positions

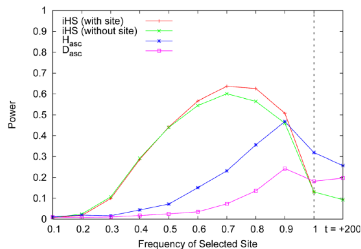
A new measure iHS (Voight *et al.*, 2006)

- iHH_A et iHH_D area under EHH curves for the Ancestral and Derived core allele
- $iHH = \frac{iHH_A}{iHH_D}$ and $\widetilde{iHH} = \ln(iHH)$ (\Rightarrow correct for recombination)
- Standardization : $iHS = \frac{(\widetilde{iHH}_i - \mu_{\widetilde{iHH}})^2}{\sigma_{\widetilde{iHH}}}$ (\Rightarrow per-class of ancestral allele freq. \simeq age)
- Ex. : $|iHS| > 2.58 \Rightarrow$ outlier at the 0.1% (nominal) threshold



Comparisons across populations (Tang et al., 2007)

iHS-based test lacks power when the variant is close to fixation



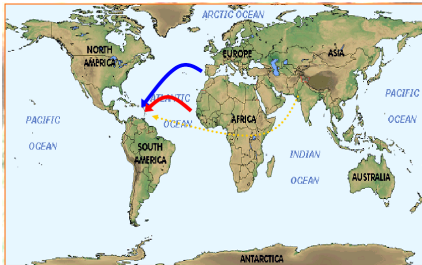
Comparisons across populations : R_{SB} (Tang et al., 2007) or XP-EHH (Sabeti et al., 2007)

- within population : $EHS \simeq \frac{p_A^2 EHH_A + p_D^2 EHH_D}{p_A^2 + p_D^2} \Rightarrow iES$ (area under the EHS curve)

- across population : $\widetilde{R_{SB}} = \log \left(\frac{iES_1}{iES_2} \right) \Rightarrow \text{standardization : } R_{SB} = \frac{\left(\widetilde{R_{SB}}_i - \text{med} \widetilde{R_{SB}} \right)^2}{\sigma_{\widetilde{R_{SB}}}^2}$

- Same definition for XP-EHH (except iES is not computed the same way)

The Creole cattle from Guadeloupe : 3 origins ?



- **European Taurine** (Iberia) with the first spanish colonizer (1493) et more recently (French breeds)
- **African Taurines** similar routes than the slave trade ones between the 16th and 18th centuries
- **Zebu** From India, introduced in America (South and US) during the 19th century

The study (Gautier and Naves, 2011)

Design

- 140 CGU individuals combined with data on pop. from European (275 ind. from 11 breeds), African (6 breeds) and Zebu (n=3 breeds) ancestry
- Genome Scan (29 autosomes) with a 50K assay (44,507 SNPs after QC) :
 - Phasing done with FASTPHASE
 - Allele polarization for iHS according to ADMIXTURE cluster freq. weighted by average CGU ancestry prop. (anc. allele=the one with the highest weighted allele frequencies)

Example of whole genome scan

- iHS scan within CGU (n=280 haplotypes)
- CGU vs. EUT (n=530 combined haplos.) comparisons using rSB and XP-EHH

The data

Haplotype Files (default format)

- Example with file EHH_10_CGU.hap.bz2
- 240 CGU haplos for Chromosome 10 (BTA10) with 1,877 SNPs

```
1 2 2 1 2 2 2 1 2 2 1 2 2 2 1 2 1 2 2 2 1 2 2 2 2 2 1 1 1 2 2 2 2 1 ... [1,877 col.]
2 1 2 1 2 2 1 1 2 2 1 2 2 1 2 1 1 1 1 2 2 1 2 2 2 1 1 2 1 1 2 1 1 1 1 ... [1,877 col.]
3 2 1 1 2 1 2 2 2 1 2 1 1 1 1 1 2 2 1 2 2 2 2 1 1 2 2 2 2 1 1 1 1 1 ... [1,877 col.]
4 2 1 1 2 2 1 1 2 2 1 2 1 1 1 1 1 2 2 2 2 2 2 2 2 2 1 1 1 2 1 2 1 1 1 ... [1,877 col.]
5 1 2 2 2 1 1 2 2 2 1 2 2 2 1 1 2 2 2 1 2 1 2 2 2 1 1 2 2 2 2 1 1 1 1 1 ... [1,877 col.]
.....
[280 rows in total]
```

Map File

- 44,507 SNPs from 29 chromosomes (file map.all.inp)

```
F0100190 1 113642 1 2
F0100220 1 244699 1 2
F0100250 1 369419 1 2
F0100270 1 447278 1 2
.....
[44,045 rows in total]
```

Loading the data and Computing iHH's (and iES's) on CGU

```
cnt=0
for(i in 10:15){#Example for Five chromosomes only
  cnt=cnt+1
  #File Name
  tmp.hapfile=bzfile(paste0("EHH_",code.chr[i],"_CGU.hap.bz2"))
  #Read data
  tmp.hap=data2haplohh(hap_file=tmp.hapfile,
                       map_file="map.all.inp",
                       chr.name=i)

  #Compute stats
  tmp.scan=scan_hh(tmp.hap,threads=4)
  if(cnt==1){
    wgscan.cgu=tmp.scan
  }else{
    wgscan.cgu=rbind(wgscan.cgu,tmp.scan)
  }
}
head(wgscan.cgu)
```

Computing and plotting iHS

```
ihs.cgu=ihh2ihs(wgscan.cgu,minmaf=0.05,freqbin=0.025)  
  
head(ihs.cgu$iHS,25)  
ihs.cgu$frequency.class  
distribplot(ihs.cgu$iHS$iHS)  
  
ihspplot(ihs.cgu)
```

Comparing CGU vs. EUT with rSB and XP-EHH

Computing iES for European Taurines

```
cnt=0
for(i in 10:15){
  cnt=cnt+1
  tmp.hap=data2haplohh(hap_file=bzfile(paste0("EHH_",code.chr[i],"_EUT.hap.bz2")),
                      map_file="map.all.inp",chr.name=i)
  tmp.scan=scan_hh(tmp.hap,threads=4)
  if(cnt==1){wgscan.eut=tmp.scan}else{wgscan.eut=rbind(wgscan.eut,tmp.scan)}
}
```

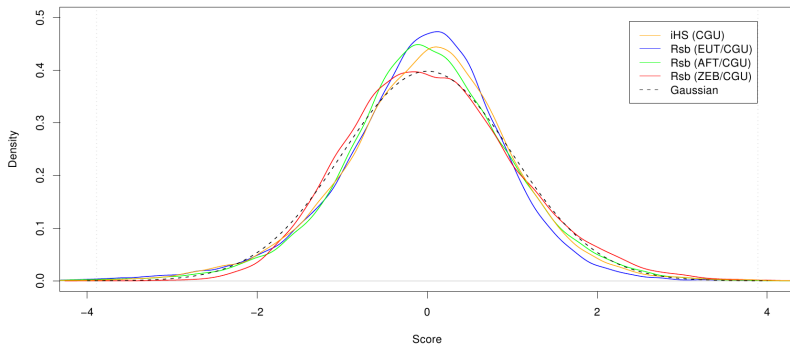
Computing Rsb and XpEHH (CGU vs. EUT)

```
rsb=ies2rsb(wgscan.cgu,wgscan.eut)
xpehh=ies2xpehh(wgscan.cgu,wgscan.eut)

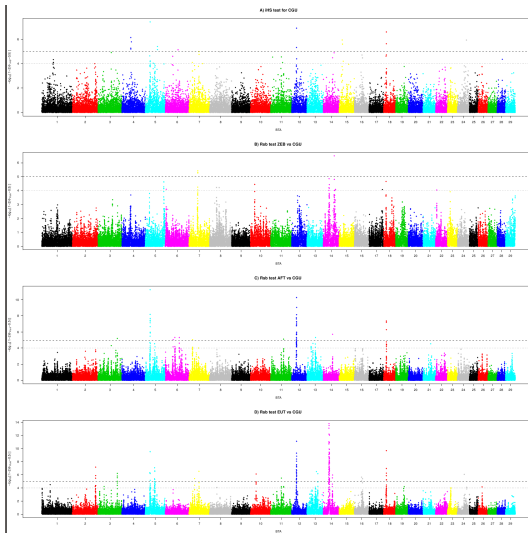
rsbplot(rsb)
xpehhplot(xpehh)
```

Computing iHS within CGU and R_{SB} CGU versus each of the 3 ancestries

A) Score Distributions



Footprints of Selection



Footprints of selection

Variant origin

- Most signals are common to several tests (including admixture based tests)
- 14/15 significant regions with an excess of ZEB ancestry (better adaptation ?)

Candidate genes and biological functions

- Reproduction : RXFP2, PMEPA1, IGFBP3 et SLCA5
- Metabolism : CYP46A1, PDE1B
- Disease resistance : CENTD3

Vizualising haplotype structure around RXFP2 SNP

```
i=12
tmp.hap=data2haplohh(
  hap_file=bzfile(paste0("EHH_",code.chr[i],"_CGU.hap.bz2")),
  map_file="map.all.inp",chr.name=i)
layout(matrix(1:2,2,1))

bifurcation.diagram(tmp.hap,mrk_foc=456,all_foc=1,nmrk_l=20,nmrk_r=20,
  main="Bifurcation diagram (RXFP2 SNP on BTA12): Ancestral Allele")

bifurcation.diagram(tmp.hap,mrk_foc=456,all_foc=2,nmrk_l=20,nmrk_r=20,
  main="Bifurcation diagram (RXFP2 SNP on BTA12): Derived Allele")
```


Key features of REHH

- Identify footprints of selection using haplotype structure (LD information)
 - Within population (iHS) [ancestral allele information needed in theory]
 - Across pairs of population (rSB or XP-EHH) (polarization of the signal)
 - If $n_{pop} > 2$, combining signals?
- Computationally efficient (e.g., parallel computing)
- Indirect approach (Need complementary approaches to understand/validate the origin of the signal)
- For more details, see the REHH vignette
<https://cran.r-project.org/web/packages/rehh/vignettes/rehh.pdf>