Detecting selection: methods (I)

Matteo Fumagalli

Bioinformatics for Adaptation

Adaptation & Phenotypes

Relevance to human health

Why do African Americans suffer from a higher incidence of hypertension?

OPEN ACCESS Freely available online PLoS GENETICS

Differential Susceptibility to Hypertension Is Due to Selection during the Out-of-Africa Expansion

J. Hunter Young[1*], Yen-Pei C. Chang[1], James Dae-Ok Kim[1], Jean-Paul Chretien[1], Michael J. Klag[1], Michael A. Levine[2], Christopher B. Ruff[1], Nae-Yuh Wang[1], Aravinda Chakravarti[1]

Hypertension: variants that allowed water and sodium retention and increased vascular reactivity now cause hypertension.

Relevance to human health

Why is there any increase of obesity-related disorders is some countries?

Obesity and insulin resistance: thrifty variants became unfavorable with a shift in diet.

OPEN ACCESS Freely available online PLoS GENETICS

Adaptations to Climate in Candidate Genes for Common Metabolic Disorders

Angela M. Hancock[1], David B. Witonsky[1], Adam S. Gordon[1], Gidon Eshel[2¤], Jonathan K. Pritchard[1], Graham Coop[1], Anna Di Rienzo[1*]

Relevance to human health

Why do Northern Europeans experience a dramatic susceptibility to autoimmune conditions?

Published May 25, 2009

JEM

Parasites represent a major selective force for interleukin genes and shape the genetic predisposition to autoimmune conditions

Matteo Fumagalli,[1,2] Uberto Pozzoli,[1] Rachele Cagliani,[1] Giacomo P. Comi,[3] Stefania Riva,[1] Mario Clerici,[4,5] Nereo Bresolin,[1,3] and Manuela Sironi[1]

Hygiene hypothesis: lack of exposure to pathogens in early life determines immune imbalances and predisposes to atopic and autoimmune diseases.

Relevance to human health

Evolutionary medicine

OPEN ACCESS Freely available online

Differential Susceptibility to Hyperte
Is Due to Selection during the Out-
Expansion

J. Hunter Young[1*], Yen-Pei C. Chang[1], James Dae-Ok Kim[1], Jean-Paul Chretien[1], Michael J
Christopher B. Ruff[1], Nae-Yuh Wang[1], Aravinda Chakravarti[1]

ariants that allowed water
etention and increased
y now cause hypertension.

Obesity and insulin res
thrifty variants became un
with a shift in diet.

PLoS GENETICS

Climate in Candidate Genes
etabolic Disorders

, Adam S. Gordon[1], Gidon Eshel[2¶], Jonathan K. Pritchard[1], Graham Coop[1],

Published May 25, 2009

JEM

Parasites represent a major selec
for interleukin genes and shape
predisposition to autoimmune co

Matteo Fumagalli,[1,2] Uberto Pozzoli,[1] Rachele Caglia
Giacomo P. Comi,[3] Stefania Riva,[1] Mario Clerici,[4,5] N
and Manuela Sironi[1]

esis: lack of exposure to
rly life determines immune
predisposes to atopic and
eases.

The genetic basis of phenotype variation

- Genome-wide association studies (GWAS)

- QTL mapping

Length

Escape tunnel

Weber *et al.*, 2013

Natural selection

Heritable traits that increase the fitness of the become more common.

1)Mutations arise randomly and evolve according to their effect on the fitness of the carrier

Natural selection

Heritable traits that increase the fitness of the become more common.

1)Mutations arise randomly and evolve according to their effect on the fitness of the carrier

Natural selection

Heritable traits that increase the fitness of the become more common.

1)Mutations arise randomly and evolve according to their effect on the fitness of the carrier

Natural selection

Heritable traits that increase the fitness of the become more common.

1)Mutations arise randomly and evolve according to their effect on the fitness of the carrier

Natural selection

Heritable traits that increase the fitness of the become more common.

1) Mutations arise randomly and evolve according to their effect on the fitness of the carrier

2) Sites targeted by natural selection are likely to harbour **functionality**

Outline

- Brief introduction to natural selection
- Modes of selection
- Inferring selection at the intra-species level
  - Genetic differentiation
  - Haplotype variation
  - Model-based approaches
  - Testing for significance
- Inferring selection at the inter-species level
- Detecting selection from low-depth sequencing data

Methods to infer selection

- within-species:

  Micro-evolutionary events between populations, local adaptation

Methods to infer recent selection

Sabeti et al. 2006 Science

Methods to infer recent selection

Sabeti et al. 2006 Science

Allele frequency differentiation

Allele frequency differentiation

Allele frequency differentiation

Allele frequency differentiation

Allele frequency differentiation

F_{ST}

Common measure for <u>quantifying</u> population subdivision.

$$F_{ST} = H_B / (H_W + H_B)$$

H_B: between populations

H_W: average within populations

- ❑ if H_W=0 then F_{ST}=1
- ❑ if H_B=0 then F_{ST}=0

F_{ST}

F_{ST} can be either considered a statistic or a parameter

- Method-of-moments estimator

(depends on sample allele frequencies and sample size)

Genetic variance **between** populations

$$F_{ST} = \frac{a}{a+b}$$

Genetic variance **within** populations

$$a = \frac{4n_1(\hat{p}_{1A} - \hat{p}_A)^2 + 4n_2(\hat{p}_{2A} - \hat{p}_A)^2 - b}{2\left(\frac{2n_1 n_2}{n_1 + n_2}\right)}$$

$$b = \frac{n_1 \alpha_1 + n_2 \alpha_2}{n_1 + n_2 - 1}$$

Reynolds et al 1983

- Parameter of a distribution (e.g. a Beta-Binomial)

Haplotype-based F_{ST}

F_{ST} based on haplotype differentiation between populations

$$F_{ST} = 1 - (H_W / H_B)$$

Within populations

Between populations

What is the variation within populations?

e.g. A vs B

The differ by 1 site

As defined by Eq. 3 in Hudson, Slatkin and Maddison 1992

Haplotype-based F_{ST}

$$F_{ST} = 1 - (H_W / H_B)$$

Within populations

Between populations

A
B
C

D
E
F

What is the variation within populations?

A	B	
A	C	
B	C	

Mean=?

D	E	
D	F	
E	F	

Mean=?

H_W is the average within-populations: ?

Haplotype-based F_{ST}

$$F_{ST} = 1 - (H_W / H_B)$$

Within populations Between populations

A
B
C

D
E
F

What is the variation within populations?

A	B	1
A	C	2
B	C	1

Mean=4/3

D	E	0
D	F	0
E	F	0

Mean=0/3

H_W is the average within-populations: (4/3+0/3)/2=2/3

Haplotype-based F_{ST}

$$F_{ST} = 1 - (H_W / H_B)$$

Within populations — Between populations

A
B
C

D
E
F

What is the variation between populations?

A	D	0
A	E	0
A	F	0
B	D	1
B	E	1
B	F	1
C	D	2
C	E	2
C	F	2

Mean=9/9

H_B is the average between-populations: 9/9=1

Haplotype-based F_{ST}

F_{ST} based on haplotype differentiation between populations

$$F_{ST} = 1 - (H_W / H_B) = 1 - ((2/3)/1) = 1/3 \sim 0.33$$

Haplotype-based F_{ST}

F_{ST} based on haplotype differentiation between populations

$F_{ST} = 1 - (H_W / H_B) = 1 - ((2/3)/1) = 1/3 \sim 0.33$

$F_{ST} = 1 - (?/?) = ?$

Haplotype-based F_{ST}

F_{ST} based on haplotype differentiation between populations

$F_{ST} = 1 - (H_W / H_B) = 1 - ((2/3)/1) = 1/3 \sim 0.33$

$F_{ST} = 0.75$

Sample size matters!

Allele frequency differentiation

From standing variation

Allele frequency differentiation

With migration

Migration rates

Two populations may be separated but occasionally **migrants** may move and exchange genetic material.

m = probability that an individual from one population is replaced with an individual from the other (per individual, per generation)

Allele frequency differentiation

With recent divergence

The case of Tibetans

Adaptation to high altitude

Selection in Tibetans

T

C

F_{ST}(T-C)

Selection in Tibetans

F_{ST}(T-C)~T(T-A-C)

Selection in Tibetans

F_{ST}(T-C)~T(T-A-C)

Selection in Tibetans

F_{ST}(T-C)~T(T-A-C)

Selection in Tibetans

$T(T\text{-}A\text{-}C) = -\log(1 - F_{ST}(T\text{-}C))$

T(T-A)?

Population Branch Statistic

Population Branch Statistic (PBS) = $(T_{12} + T_{13} - T_{23})/2$

Yi et al. Science 2010

Population Branch Statistic

Methods to infer recent selection

Sabeti et al. 2006 Science

Positive selection: effect on haplotypes

$t<T_{sel}$

Positive selection: effect on haplotypes

$t<T_{sel}$

$t=T_{sel}$

Positive selection: effect on haplotypes

$t<T_{sel}$

$t=T_{sel}$

$t>T_{sel}$

Positive selection: effect on haplotypes

$t<T_{sel}$

$t=T_{sel}$

$t>T_{sel}$

$t>>T_{sel}$

Selective sweep

Genetic hitchhiking

Positive selection

$t<T_{sel}$

$t>>T_{sel}$

Positive selection

$t<T_{sel}$

$t>>T_{sel}$

- Reduction of polymorphisms levels (e.g. from 7 to 5 SNPs)

Positive selection

- Reduction of polymorphisms levels
(e.g. from 7 to 5 SNPs)

Nucleotide diversity index: Watterson's Theta (1975)
with K SNPs and n chromosomes

$$\theta_W = \frac{K}{a_n}$$

$$a_n = \sum_{i=1}^{n-1} \frac{1}{i}$$

Positive selection

- Reduction of polymorphisms levels (Theta)
- ?

Positive selection

- Reduction of polymorphisms levels (Theta)
- Excess of low-frequency variants

Nucleotide diversity index: average pairwise nucleotide differences (Pi)
with $k_{i,j}$ equal to the number of nucleotide differences between sequences *i* and *j*

$$\pi = \frac{\sum_{i=1}^{n-1} \sum_{j=+1}^{n} k_{i,j}}{\binom{n}{2}}$$

Positive selection

- Reduction of polymorphisms levels (Theta)
- Excess of low-frequency variants (Pi)

Under neutrality, Theta and Pi are expected to be the same.
Tajima's D measures their difference.

$$D = \frac{\pi - \theta_W}{\sqrt{\hat{V}(\pi - \theta_W)}}$$

D<0 is suggestive of an excess of low-frequency variants

The importance of being…
The Site Frequency Spectrum

2D-SFS

Chinese

Tibetans

Yi et al. Science 2010

The importance of being…
The Site Frequency Spectrum

The importance of being…
The Site Frequency Spectrum

Tajima's D?

$$D = \frac{\pi - \theta_W}{\sqrt{\hat{V}(\pi - \theta_W)}}$$

~0.33=1/3

$$\theta_W = \frac{K}{a_n}$$

$$a_n = \sum_{i=1}^{n-1} \frac{1}{i}$$

$$\pi = \frac{\sum_{i=1}^{n-1} \sum_{j=+1}^{n} k_{i,j}}{\binom{n}{2}}$$

= 10, the number of comparisons you need to make

The importance of being…
The Site Frequency Spectrum

K=5

a_n=1/1 + 1/2 + 1/3 + 1/4=(12+6+4+3)/12=25/12

Theta=5/(25/12)=12/5

Pi=(5+5+5+5+0+0+0+0+0+0)/10=20/10=2

sd(D)=1/3

D=(2-12/5)/(1/3)=((10-12)/5)*3=-6/5=-1.2

D<0 → How to distinguish between positive and negative selection?

Positive vs. negative selection

- Excess of low-frequency variants (Tajima's D<0, Fu and Li's D/F<0)
- Extended Linkage Disequilibrium
- Excess of high-frequency derived alleles (Fay and Wu's H)

Demography matters?

n=20; L=500kbp; no selection

Demography matters?

n=20; L=500kbp; no selection

n=20; L=500kbp; no selection

Demography matters!

n=20; L=500kbp; no selection

n=20; L=500kbp; no selection

The genetic effects of population size changes

Demography matters!

n=20; L=500kbp; no selection

Minor Allele Frequency

n=20; L=500kbp; no selection

Minor Allele Frequency

- Excess of segregating sites
- Excess of low-frequency variants
- SFS-derived summary statistics may fail to distinguish between the effects of demography and selection

Demography matters?

n=20; L=500kbp; no selection

n=20; L=500kbp; no selection

Demography matters!

n=20; L=500kbp; no selection

n=20; L=500kbp; no selection

The genetic effects of population size changes

Bottleneck event: drastic reduction in size followed by recovery:

- Depletion of rare variants
- Excess of common variants

Demography matters!

n=20; L=500kbp; no selection

Minor Allele Frequency

n=20; L=500kbp; no selection

Minor Allele Frequency

- Depletion of segregating sites
- Excess of intermediate-frequency variants
- SFS-derived summary statistics may fail to distinguish between the effects of demography and selection → ?

Balancing selection

The process whereby genetic variability is maintained in the population due to selection

Features:
- High nucleotide diversity
- Excess of intermediate frequency alleles
- Excess of polymorphisms compared to interspecies divergence

More difficult to detect (shorter genomic extent)

- **Overdominance** (the heterozygote has an advantage)
- **Frequency dependence** (an advantage is conferred by a rare feature)
- **Environmental** adaptation (different alleles are advantageous in different environments)

Balancing selection

Common haplotypes separated by deep branches
(deep time to the most common recent ancestor)

Confounding factor: ancient population structure

Experimental design matters?

The effect of ascertainment bias

Deficiency of low-frequency variants

HapMap Consortium. Nature 2005

How to take neutral confounding factors into account?

Under constant population size:

How to take neutral confounding factors into account?

Under expanding population size:

How to take neutral confounding factors into account?

Under expanding population size and positive selection:

- Demography affects all loci equally, while selection changes local patterns

Outlier approach

Assign empirical *p*-values
(ranked percentiles)

Outlier approach

Assign empirical *p*-values
(ranked percentiles)

How to take neutral confounding factors into account?

Under expanding population size and positive selection:

- Demography affects all loci equally, while selection changes local patterns
  What should we do if we don't have genome-wide data?

How to take neutral confounding factors into account?

Under expanding population size and positive selection:

- Demography affects all loci equally, while selection changes local patterns
  What should we do if we don't have genome-wide data?

Simulations-based approach

Simulations-based approach

Assign *p*-values
(based on ranked percentile of observed value)

Model-based approaches

Expected distribution is derived from analytical solutions or fitted from observed data

Example: SweepFinder
The spatial patterns of fixed mutations and segregating sites with their frequencies is modeled using theoretical expectations. A (composite) likelihood ratio test is performed to test for selection.

Example: BayeScan
Based on differences in allele frequencies between populations. Departure from neutrality is assessed by measuring the locus-specific component.

Positive selection

$t<T_{sel}$

$t>>T_{sel}$

- Reduction of polymorphisms levels (Theta)
- Excess of low-frequency variants (Pi, Tajima's D, SFS)
- ?

Positive selection

- Reduction of polymorphisms levels (Theta)
- Excess of low-frequency variants (Pi, Tajima's D, SFS)
- Extended haplotype homozygosity / Extended LD

Extended Haplotype Homozygosity

$t >> T_{sel}$

…

$t >>> T_{sel}$

…

Extended Haplotype Homozygosity

$t >> T_{sel}$

…

$t >>> T_{sel}$

…

$t >>> T_{sel}$

Extended Haplotype Homozygosity

Core haplotype is 1
(Biallelic: 0 is ancestral, 1 is derived allele)

$$EHH_c(x_i) = \sum_{h \in H_c(x_i)} \frac{\binom{n_h}{2}}{\binom{n_c}{2}}$$

Core SNP

Extended Haplotype Homozygosity

Core haplotype is 1
(Biallelic: 0 is ancestral, 1 is derived allele)

$$EHH_c(x_i) = \sum_{h \in H_c(x_i)} \frac{\binom{n_h}{2}}{\binom{n_c}{2}}$$

Until marker x_i
(starting from x_0)

Extended Haplotype Homozygosity

Core haplotype is 1
(Biallelic: 0 is ancestral, 1 is derived allele)

$$EHH_c(x_i) = \sum_{h \in H_c(x_i)} \frac{\binom{n_h}{2}}{\binom{n_c}{2}}$$

Sum across all unique haplotypes
carrying the core SNP

Extended Haplotype Homozygosity

Core haplotype is 1
(Biallelic: 0 is ancestral, 1 is derived allele)

$$EHH_c(x_i) = \sum_{h \in H_c(x_i)} \frac{\binom{n_h}{2}}{\binom{n_c}{2}}$$

n_h is haplotype frequency of h

n_h is haplotype frequency of the core SNP

Sum across all unique haplotypes carrying the core SNP

Extended Haplotype Homozygosity

Core haplotype is 1
(Biallelic: 0 is ancestral, 1 is derived allele)

$$EHH_c(x_i) = \sum_{h \in H_c(x_i)} \frac{\binom{n_h}{2}}{\binom{n_c}{2}}$$

n_h is haplotype frequency of h

n_h is haplotype frequency of the core SNP

Sum across all unique haplotypes carrying the core SNP

$$EHH_c(x_i = 0) = ?$$

Extended Haplotype Homozygosity

Core haplotype is 1
(Biallelic: 0 is ancestral, 1 is derived allele)

$$EHH_c(x_i) = \sum_{h \in H_c(x_i)} \frac{\binom{n_h}{2}}{\binom{n_c}{2}}$$

n_h is haplotype frequency of h

n_c is haplotype frequency of the core SNP

Sum across all unique haplotypes carrying the core SNP

$$EHH_c(x_i = 0) = \frac{\binom{5}{2}}{\binom{5}{2}} = 1$$

Extended Haplotype Homozygosity

$$EHH_c(x_i) = \sum_{h \in H_c(x_i)} \frac{\binom{n_h}{2}}{\binom{n_c}{2}}$$

$$EHH_c(x_i = +1) = ?$$

How many unique haplotypes carrying the core SNP?
What is their frequency?

Extended Haplotype Homozygosity

$$EHH_c(x_i) = \sum_{h \in H_c(x_i)} \frac{\binom{n_h}{2}}{\binom{n_c}{2}}$$

$$EHH_c(x_i = +1) = \frac{\binom{4}{2} + \binom{1}{2}}{\binom{5}{2}} = \frac{6+0}{10} = 0.60$$

Extended Haplotype Homozygosity

$$EHH_c(x_i) = \sum_{h \in H_c(x_i)} \frac{\binom{n_h}{2}}{\binom{n_c}{2}}$$

$$EHH_c(x_i = +2) = ?$$

Extended Haplotype Homozygosity

$$EHH_c(x_i) = \sum_{h \in H_c(x_i)} \frac{\binom{n_h}{2}}{\binom{n_c}{2}}$$

$$EHH_c(x_i = +2) = EHH_c(x_i = +1) = 0.60$$

Extended Haplotype Homozygosity

$$EHH_c(x_i) = \sum_{h \in H_c(x_i)} \frac{\binom{n_h}{2}}{\binom{n_c}{2}}$$

How many unique haplotypes carrying the core SNP?
What is their frequency?

Extended Haplotype Homozygosity

$$EHH_c(x_i) = \sum_{h \in H_c(x_i)} \frac{\binom{n_h}{2}}{\binom{n_c}{2}}$$

How many unique haplotypes carrying the core SNP?
What is their frequency?
1111 with freq=2
1110 with freq=2
1000 with freq=1

$$EHH_c(x_i = +3) = ?$$

Extended Haplotype Homozygosity

$$EHH_c(x_i) = \sum_{h \in H_c(x_i)} \frac{\binom{n_h}{2}}{\binom{n_c}{2}}$$

How many unique haplotypes carrying the core SNP?
What is their frequency?
1111 with freq=2
1110 with freq=2
1000 with freq=1

$$EHH_c(x_i = +3) = \frac{\binom{2}{2} + \binom{2}{2} + \binom{1}{2}}{\binom{5}{2}} = \frac{1+1+0}{10} = 0.20$$

Extended Haplotype Homozygosity

$$EHH_c(x_i) = \sum_{h \in H_c(x_i)} \frac{\binom{n_h}{2}}{\binom{n_c}{2}}$$

n	*n choose 2*
1	0
2	1
3	3
4	6
5	10
6	15

$$EHH_c(x_i = -1) = ?$$

$$EHH_c(x_i = -2) = ?$$

Comment on differences (if any) between *EHH(x=+2)* and *EHH(x=-2)*.

Extended Haplotype Homozygosity

$$EHH_c(x_i) = \sum_{h \in H_c(x_i)} \frac{\binom{n_h}{2}}{\binom{n_c}{2}}$$

n	*n choose 2*
1	0
2	1
3	3
4	6
5	10
6	15

$$EHH_c(x_i = -1) = \frac{\binom{3}{2} + \binom{2}{2}}{\binom{5}{2}} = \frac{3+1}{10} = 0.4$$

$$EHH_c(x_i = -2) = \frac{\binom{2}{2} + \binom{1}{2} + \binom{1}{2} + \binom{1}{2}}{\binom{5}{2}} = \frac{1+0+0+0}{10} = 0.1$$

Comment on differences (if any) between *EHH(x=+2)* and *EHH(x=-2)*?

Integrated Haplotype Score

For the derived allele

For the ancestral allele

Integrated Haplotype Score

Integrated haplotype homozygosity (*iHH*)

Integrated Haplotype Score

Integrated haplotype homozygosity (*iHH*)

Integrated haplotype score:

$iHs = ln(iHH_1/iHH_0)$

↓

Genome-wide normalization in frequency bins
(to mean=0 and sd=1)

nSL

Integration with respect to physical (not genetic) map

$nSL = ln(SL_1/SL_0)$

Genome-wide normalization in frequency bins (to mean=0 and sd=1)

Cross-population Extended Haplotype Homozygosity

Integrated haplotype homozygosity (*iHH*) for **populations** A and B

Integrated haplotype score:

$XP\text{-}EHH = ln(iHH_A/iHH_B)$

↓

Genome-wide normalization in frequency bins (to mean=0 and sd=1)

Outline

- Brief introduction to natural selection
- Modes of selection
- Inferring selection at the intra-species level
  - Genetic differentiation
  - Haplotype variation
  - Model-based approaches
  - Testing for significance
- Inferring selection at the inter-species level
- Detecting selection from low-depth sequencing data

Software available

DnaSP (http://www.ub.edu/dnasp/)

Arlequin (http://cmpg.unibe.ch/software/arlequin35/)

BayeScan (http://cmpg.unibe.ch/software/BayeScan/)

libsequence (http://www.molpopgen.org/software.html)

sweep (http://www.broadinstitute.org/mpg/sweep/)

iHS (http://coruscant.itmat.upenn.edu/software.html)

nSL (http://cteg.berkeley.edu/~nielsen/resources/software/)

Pre-computed values (USCS genome browser tables)

Homemade scripts

…

Summary

- Methods to detect signatures of selection are grouped based on:
  - time of selection
  - summary statistics used

- Assessing statistical significance through empirical or expected distributions

- Investigating complex models of selection and adaptation