Environmental Association Analysis to identify adaptive loci

Angela Hancock

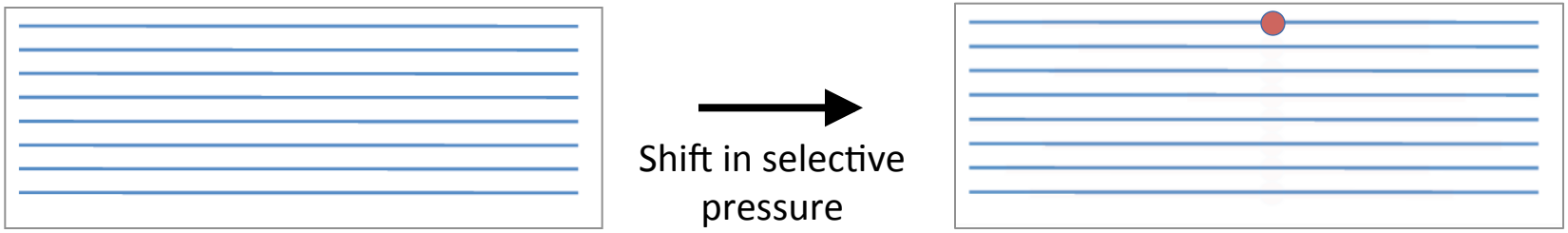
January 30, 2018

Max Planck Institute for
Plant Breeding Research

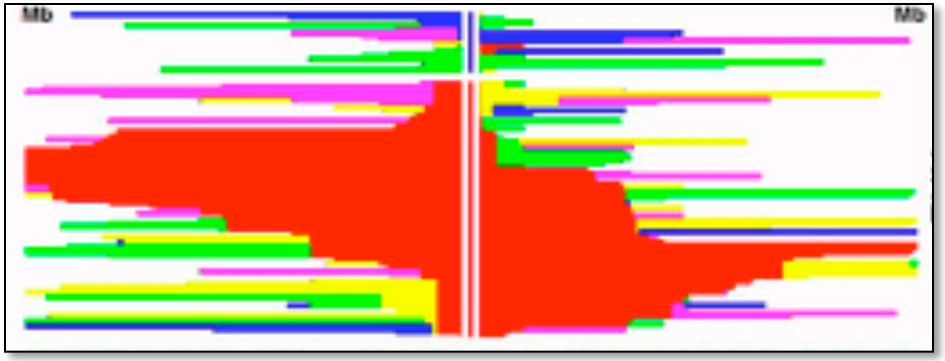
MAX-PLANCK-GESELLSCHAFT

How do populations adapt?

Hard sweep model of adaptation



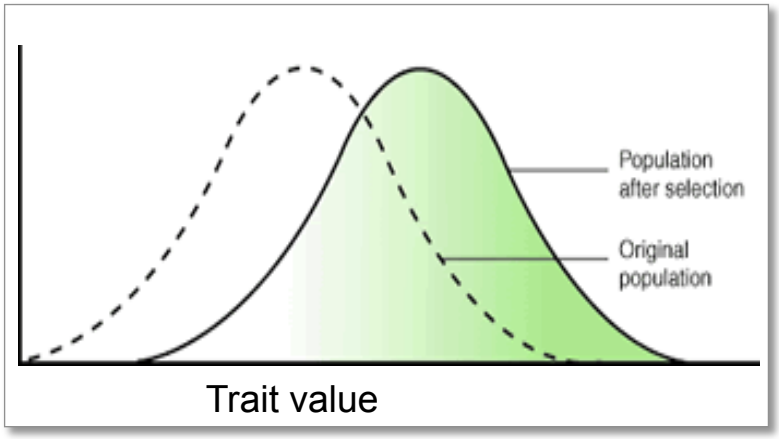
Haplotype structure can be used to identify regions implicated in *hard sweeps*



But what if adaptive traits are highly polygenic?

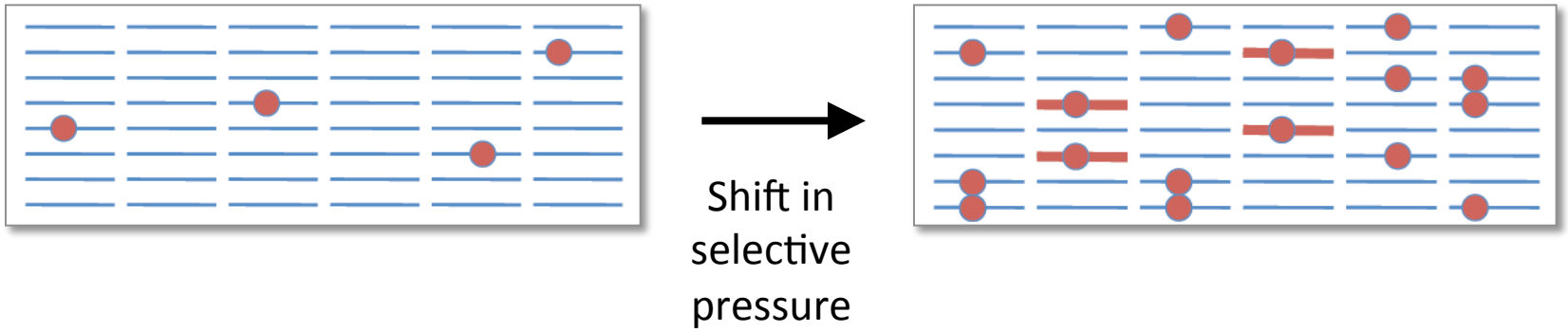
When adaptation acts mainly on variation that is already segregating in the population…

selection results in a shift in the distribution of the adaptive trait…



and shifts in the distributions of many underlying genetic variants

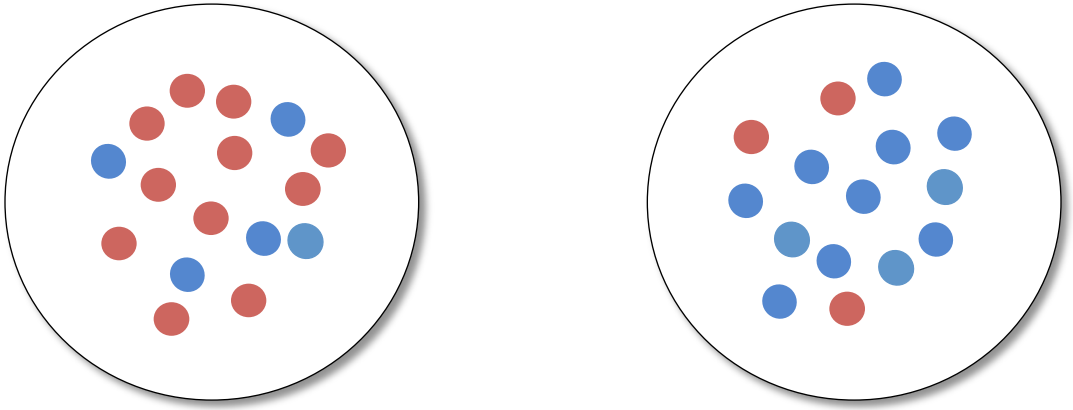
Hard sweeps may not play a major role under a polygenic model



Polygenic selection results in small shifts in frequencies at many loci, most of which were present in the population when the selection pressure arose

How can we detect adaptation from standing genetic variation?

Population differentiation

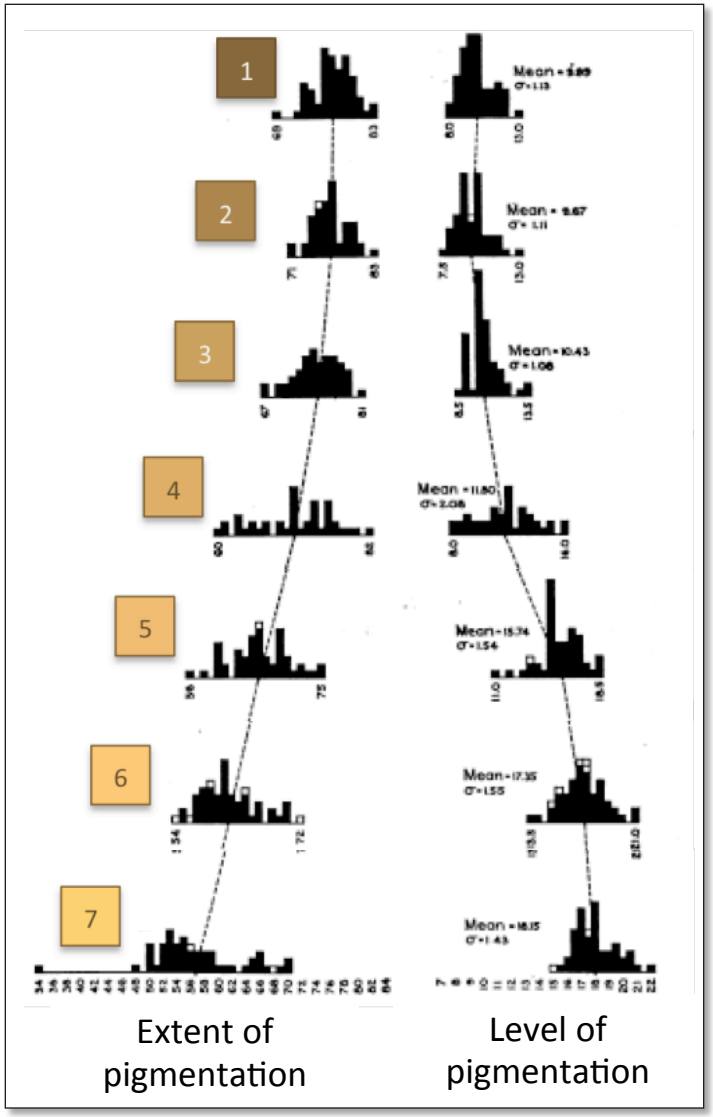


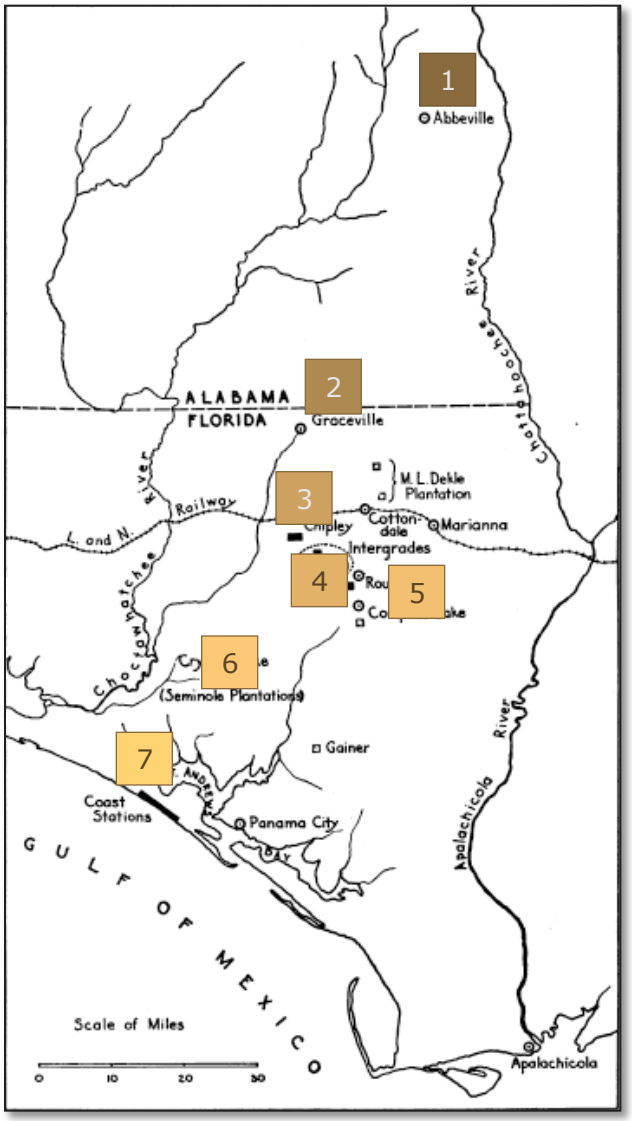
Population 1 Population 2

These methods rely on the simple assumption that the two populations differ with respect to some selection pressure

… but can we do better?

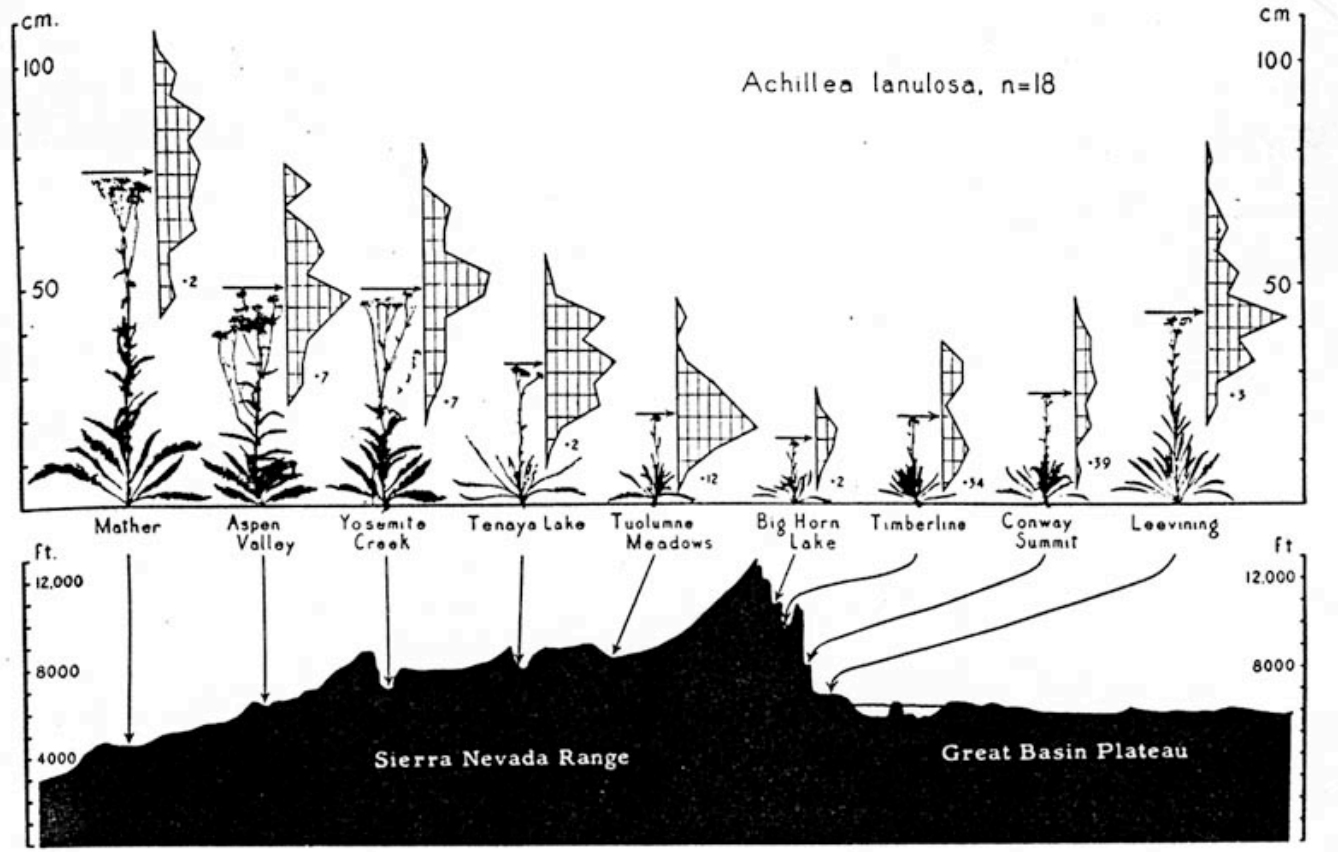
Concordance between phenotypes and environment suggests adaptation





Sumner 1929

Heights of yarrow plants vary with altitude



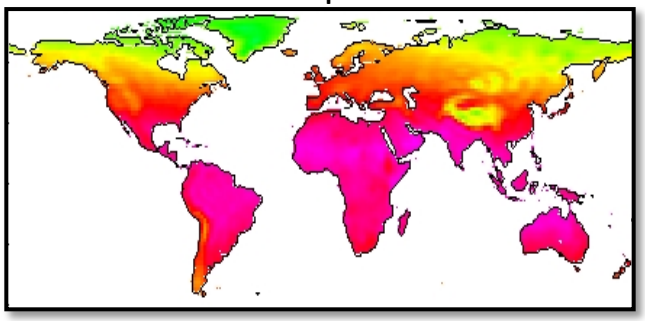
from Clausen, Keck and Heisey, 1948

Body size and stature are correlated with temperature

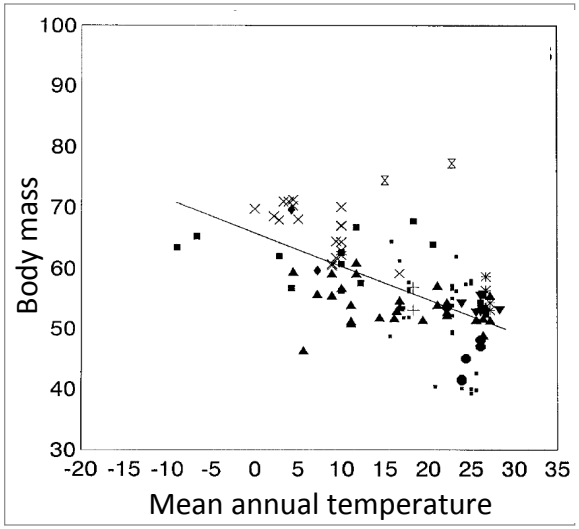
Bergmann's Rule: *Body size decreases with increasing temperature*

Allen's Rule: *Surface area relative to body size increases with increasing temperature*

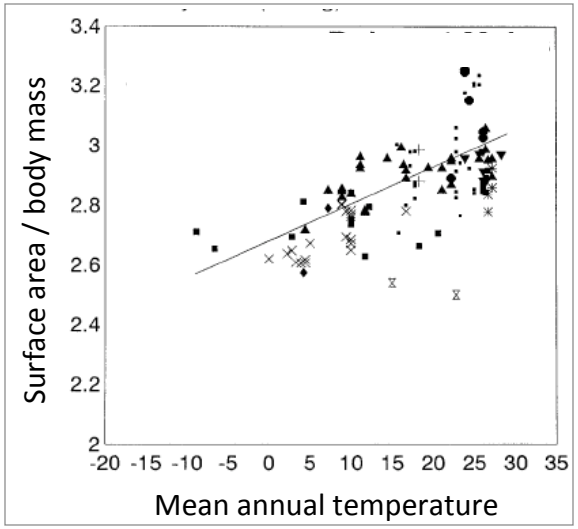
Mean temperature



Bergmann's Rule

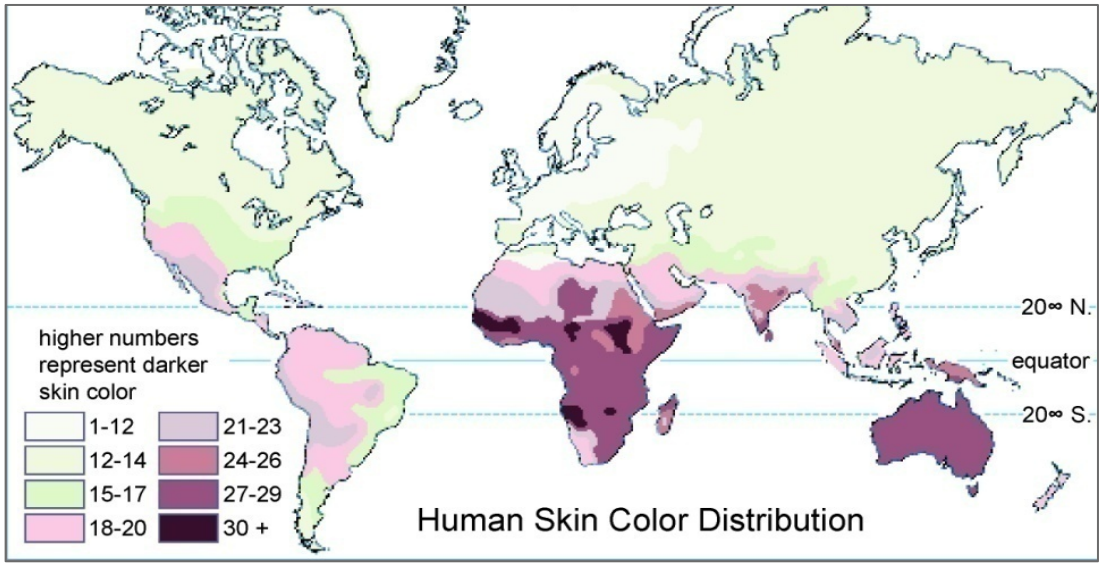


Allen's Rule

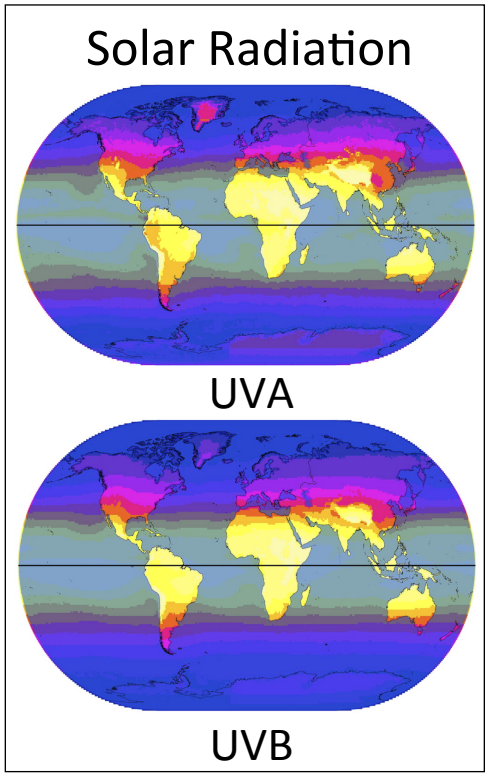


In humans pigmentation varies with latitude

Pigmentation



from Barsh PLOS Biology 2003, adapted from Biasutti 1953



Julian Huxley coined the term *cline* to describe these patterns

In 1938, Huxley coined the term *cline* to describe '*a gradation in measurable characters*'

He reviewed contemporary studies of phenotypic variation in populations and argued that this is an area that should receive greater attention

No. 3587, JULY 30, 1938 NATURE 219

Clines: an Auxiliary Taxonomic Principle

By Dr. Julian Huxley, F.R.S., Zoological Society, Regent's Park, London

MODERN taxonomy, after a phase of 'splitting', adopted the integrating principle of geographical replacement, thus uniting numerous forms previously styled species as geographical races or subspecies of a polytypic species or *Formenkreis*[1]. It seems desirable, however, to go still further in the direction of synthesis. The naming and description of subspecific forms, however necessary, is not sufficient. Further, if not supplemented by some other method, it has two actual disadvantages—it focuses undue attention on named forms as against those which remain unnamed, even when the degree of distinctness is only slightly less in the latter; it also conveys a false impression of uniformity within the named group, and thus tends to inhibit the study of intra-group regularities of variation.

Some special term seems desirable to direct attention to variation within groups, and I propose the word *cline*, meaning a gradation in measurable characters. This, being technical, seems preferable to such a term as character-gradient or phrases such as 'geographical progression of characters', used by W. F. Reinig in his recent book "Elimination und Selektion". (Naturally, when it can be shown that such characters are non-genetic in origin, they will be valueless for taxonomic purposes.) Prefixes can be used to denote clines of different types, for example, ecocline, genocline (gradient in genes), geocline (geographical cline), chronocline (paleontological trend), etc. The term could be extended if desired, for example, ontocline for regular trends in individual development.

Clines may be of inter- or intra-group nature. Inter-group clines connect the mean values of the subspecies of a polytypic species (or of the species of a geographical subgenus or *Artenkreis*[1]). Numerous regularities of this sort are known, for example, the Rules of Bergmann, Gloger, Allen, etc. Rensch[1,2] has recently summarized the subject. Good examples affecting colour or size are found in many birds represented in Britain (wrens, puffins, spotted woodpeckers, bullfinches, tits, etc.). An illuminating case is that of the wrens inhabiting Fair Isle[3]. These are not sufficiently distinct to be given a separate subspecific name, but are intermediate in character as well as in position between *T. t. troglodytes* of the mainland and the Orkneys and *T. t. zetlandicus* of the Shetlands. To subsume these facts by a cline is to direct attention to a regularity that is concealed if we restrict ourselves to specification by the naming of areal groups.

Intra-group clines concern continuous variation within a population. Relatively little work has as yet been done on this laborious subject, for example, tongue-length in bees[4]; percentage of 'spectacled' forms in guillemots; fin-rays in fish[5]; pattern in lady-beetles[6]; vertebræ in fish[7]; temperature-resistance in *Drosophila*[8], etc. Sumner[9], in a coastal subspecies of deermouse (*Peromyscus*), has shown that the adaptively cryptic colour of the pelage changes gradually as one passes inland from white sand to dark soil. Still further inland there exists a distinct and much darker subspecies on very dark soil, which also shows a colour cline, though less pronounced. Where the two meet, there is a narrow zone about three miles wide where the mean colour changes very rapidly, and the variability is much higher. Off the coast, on an isolated island of white sand, lives a much paler subspecies. Here we have, first, an inter-group cline comprising three subspecies, and also intra-group ones within the two inland groups. These run in the same direction as the inter-group cline, but are much less steep. These two geographical clines are separated by a very steep genetic cline (genocline) at the inter-breeding zone.

In plants, ecological clines appear to be the commonest type. Gregor[10], in *Plantago maritima* has shown that each ecological habitat selects out a particular assemblage of genetic types, so that a regular ecocline will run from more to less saline surroundings. It is probable that similar ecoclines are to be found among land-snails (Rensch[1]).

It is in no way intended that specification by clines should replace any of the current taxonomic methods. It would constitute a supplementary method which, it is suggested, would correct certain defects inherent in that of naming areal groups, notably in stressing continuity and regularity of variation as against mere distinctness of groups. It is important to note that clines for different characters may run in different directions (shrikes[11], fox-sparrows[12], lincoln sparrows[13] etc.).

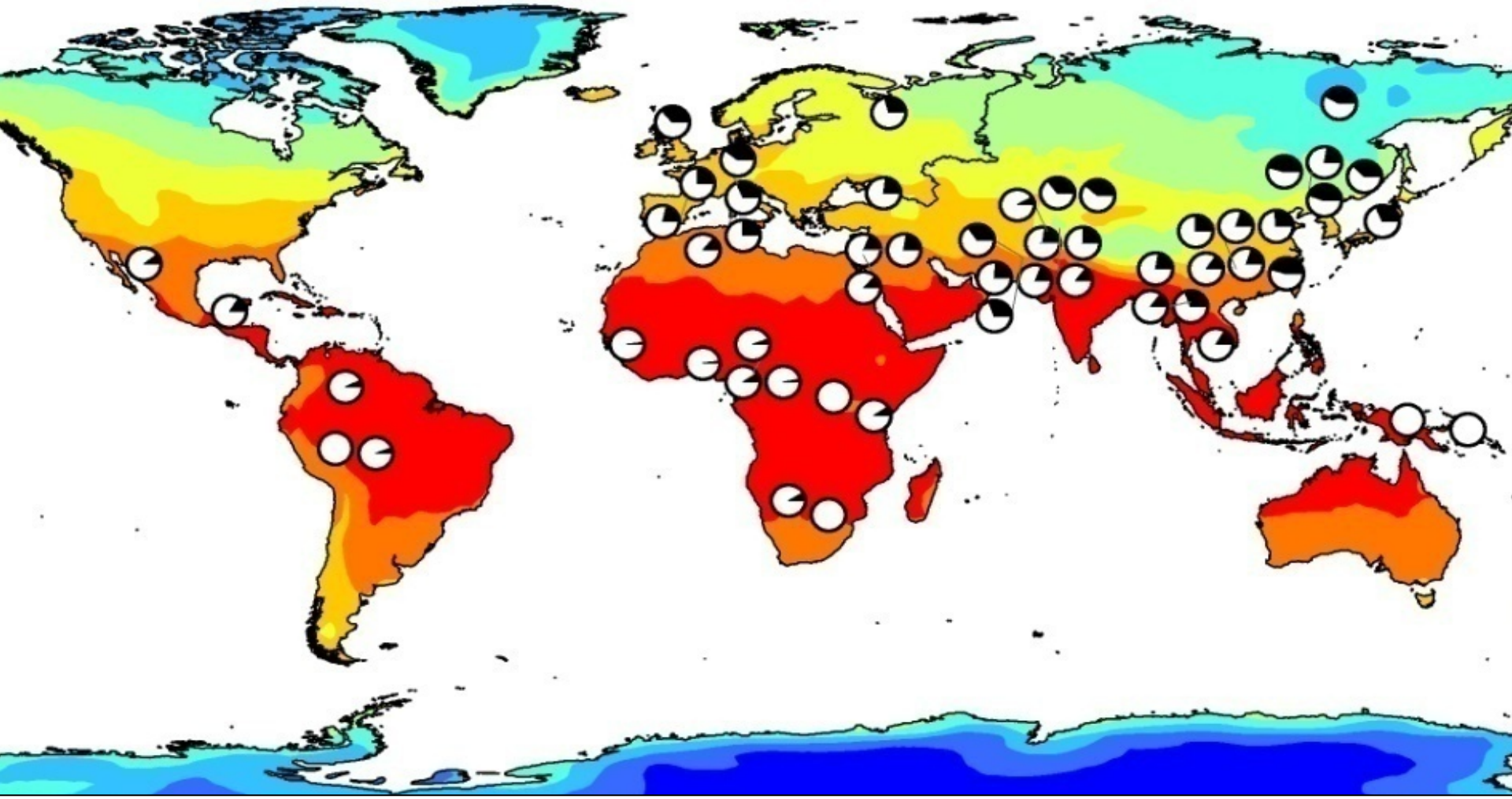
It would seem certain that, once attention is concentrated on this subject, regularities of intra-group variation will be found to be common—the rule rather than the exception. The correlation of these with environmental factors will undoubtedly often not be easy; and where the environmental factors vary in a complex way, the mere detection of regular phenoclines may be difficult, though not impossible[14]. However, if the study of such regularities is actively pursued, I would prophesy that we shall eventually gain a new picture of species. In many cases at least, the species will prove to consist of a population showing adaptive clines running in various directions: the continuous gradation will be broken up by various forms of isolation, which, by impeding interbreeding and the free flow of genes, will accentuate the mean adaptive differences between adjacent groups, as well as in some cases introducing non-adaptive differences[15]. The term cline is put forward as a step in this direction.

I have to make grateful acknowledgments to a number of systematists with whom I have discussed the subject, and who allow me to say that they believe that the use of the concept in taxonomy would be useful, notably Mr. M. A. C. Hinton, Mr. N. D. Riley and Mr. J. R. Norman of the British Museum (Natural History), Mr. J. S. L. Gilmour and Dr. W. B. Turrill of the Royal Botanic Gardens, Kew, Mr. B. W. Tucker of Oxford, the Rev. F. C. R. Jourdain, and

…if the study of such [intragroup] regularities is actively pursued, I would prophesy that we shall eventually gain a new picture of species. In many cases at least, the species will prove to consist of a population showing adaptive clines running in various directions: the continuous gradation will be broken up by various forms of isolation, which, by impeding interbreeding and the free flow of genes, will accentuate the mean adaptive differences between adjacent groups, as well as in some cases introducing nonadaptive differences. The term cline is put forward as a step in this direction.

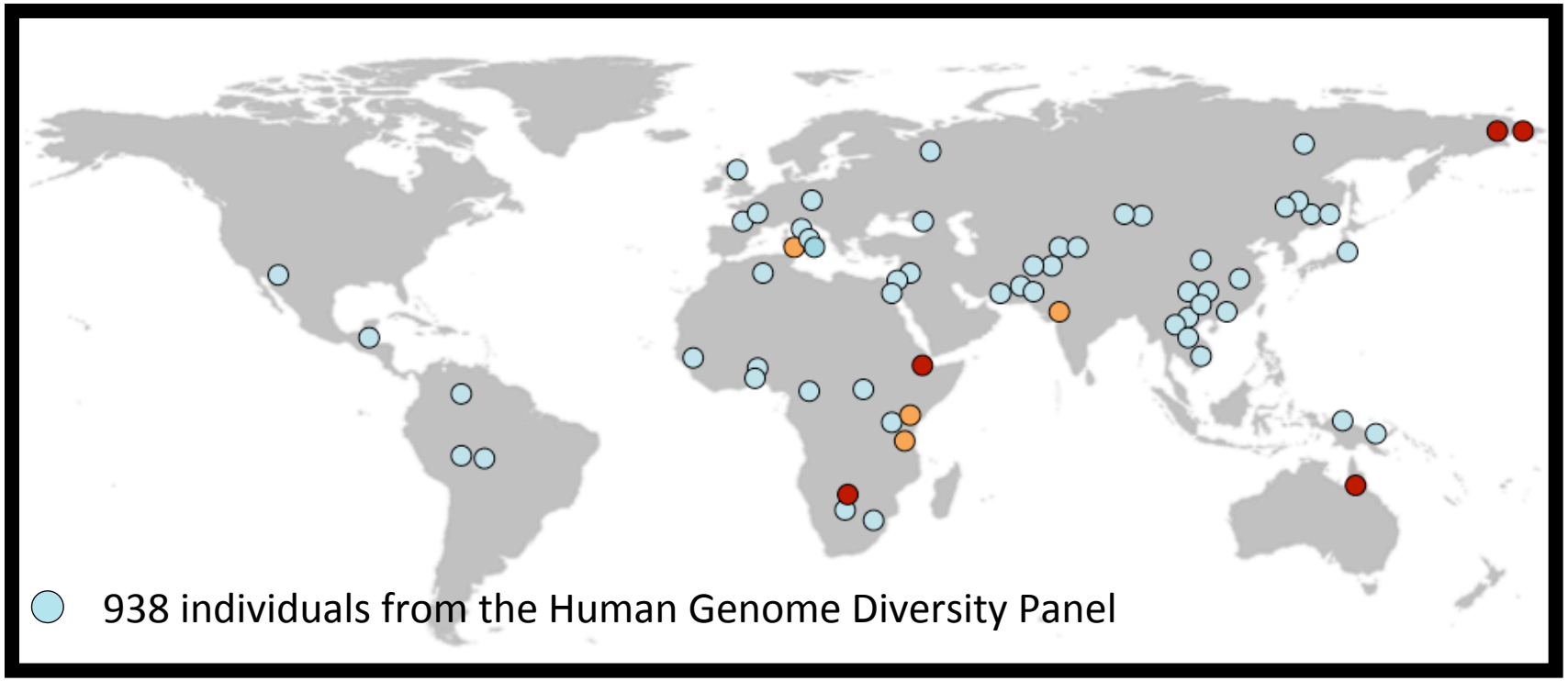
The paper was really an argument for 'lumping' vs. 'splitting'

Can we use clinal patterns to identify adaptive *genetic loci*?

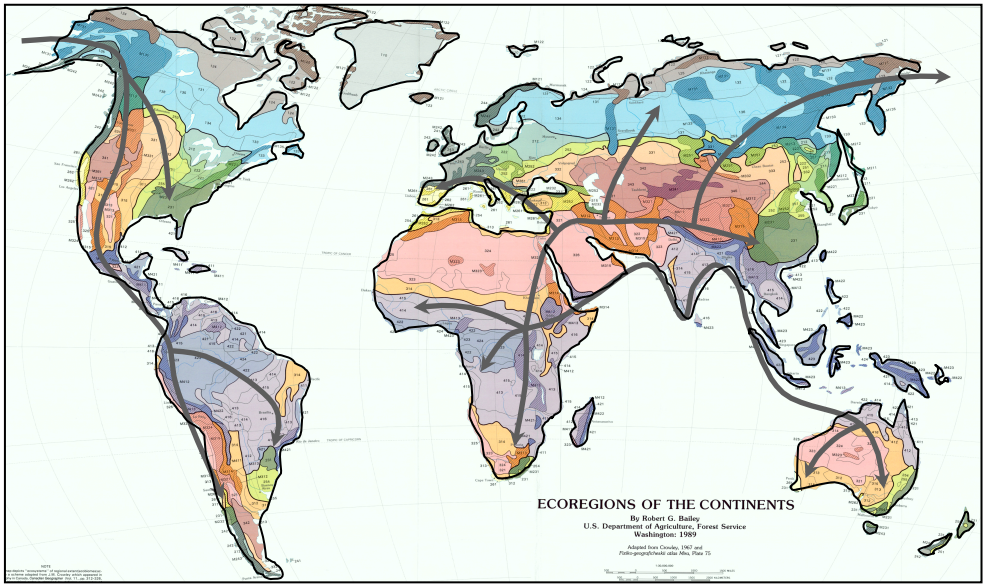


Can we use clinal patterns to identify adaptive *genetic loci*?

Candidate gene study to examine correlations in Energy metabolism genes compared to background



Population history confounds efforts to identify adaptive loci with **e**nvironmental **a**ssociation **a**nalysis (**EAA**)



- False positives result when population history is correlated with the environment
- Controlling for population structure can help expose true signals

Solution: Model population structure when assessing evidence for correlation with the environment

Bayenv compares the strength of evidence for a model with an effect of environment to the null

Null model:

$$H_0: \; y = \beta_0 + \mu + \varepsilon$$

intercept

random error

β_0: Population allele frequency

μ: Random effect due to population history

The null model contains a term for population history

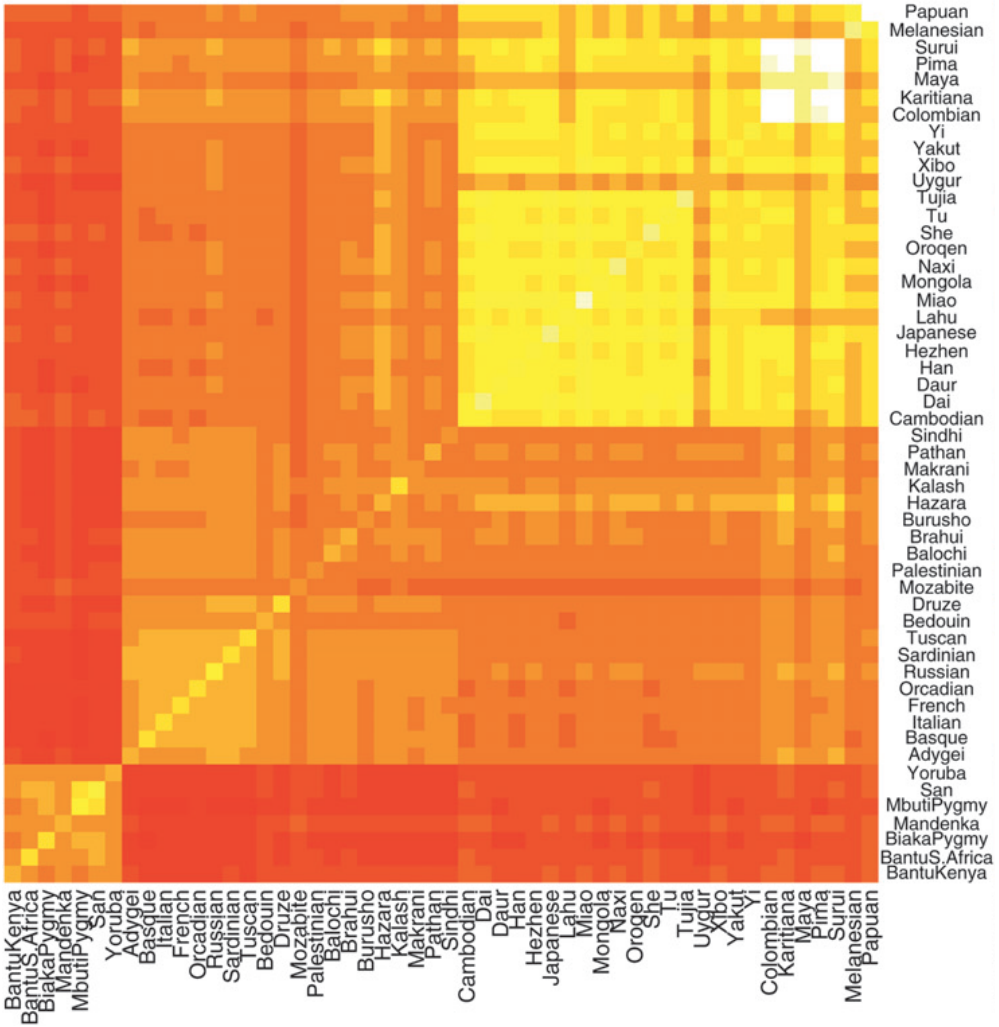
Alternative model:

$$H_1: \; y = \beta_0 + \beta_1 x + \mu + \varepsilon$$

$\beta_1 x$: Environmental effect

Coop et al., 2010

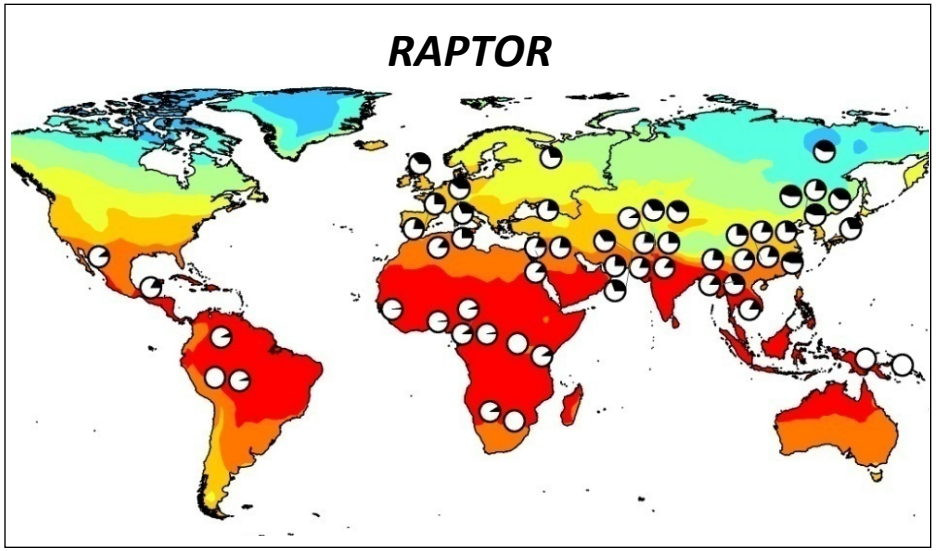
Bayenv uses the (predicted) variance/ covariance matrix to control for population structure



Energy metabolism variants are associated with winter temperature

Energy metabolism genes associated with winter climate

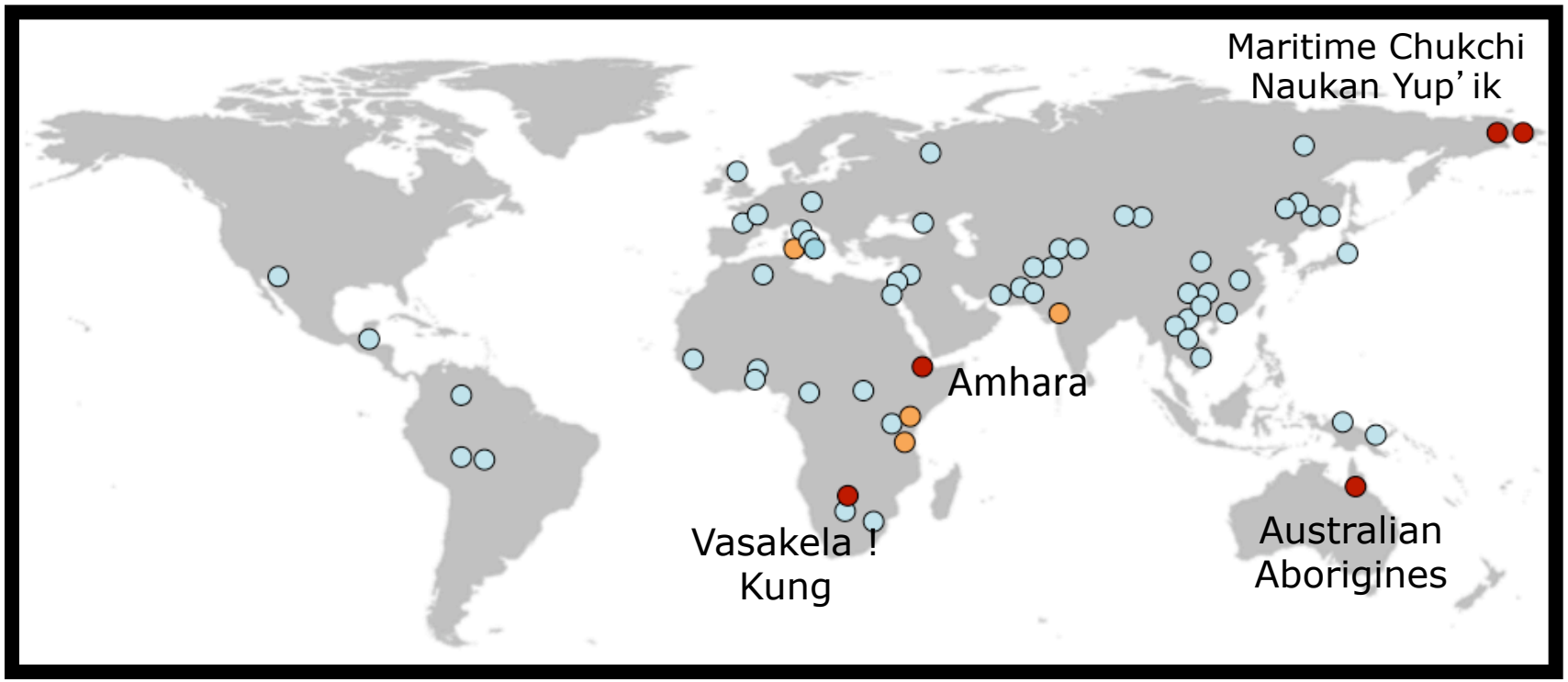
ACE	*MEF2A*
FABP2 A54T	*NUDT6*
EGFR	*PCSK1*
EPHX2 R287Q	*PPARGC1A*
FGF2	*PTK2B*
LEPR K109R	*RAPTOR*
MAPK1	*UCP3*



Hancock et al., 2008

Genetic variants that influence risk of metabolic syndrome may be involved in cold tolerance

Ilumina 650K SNP chip in 61 human population



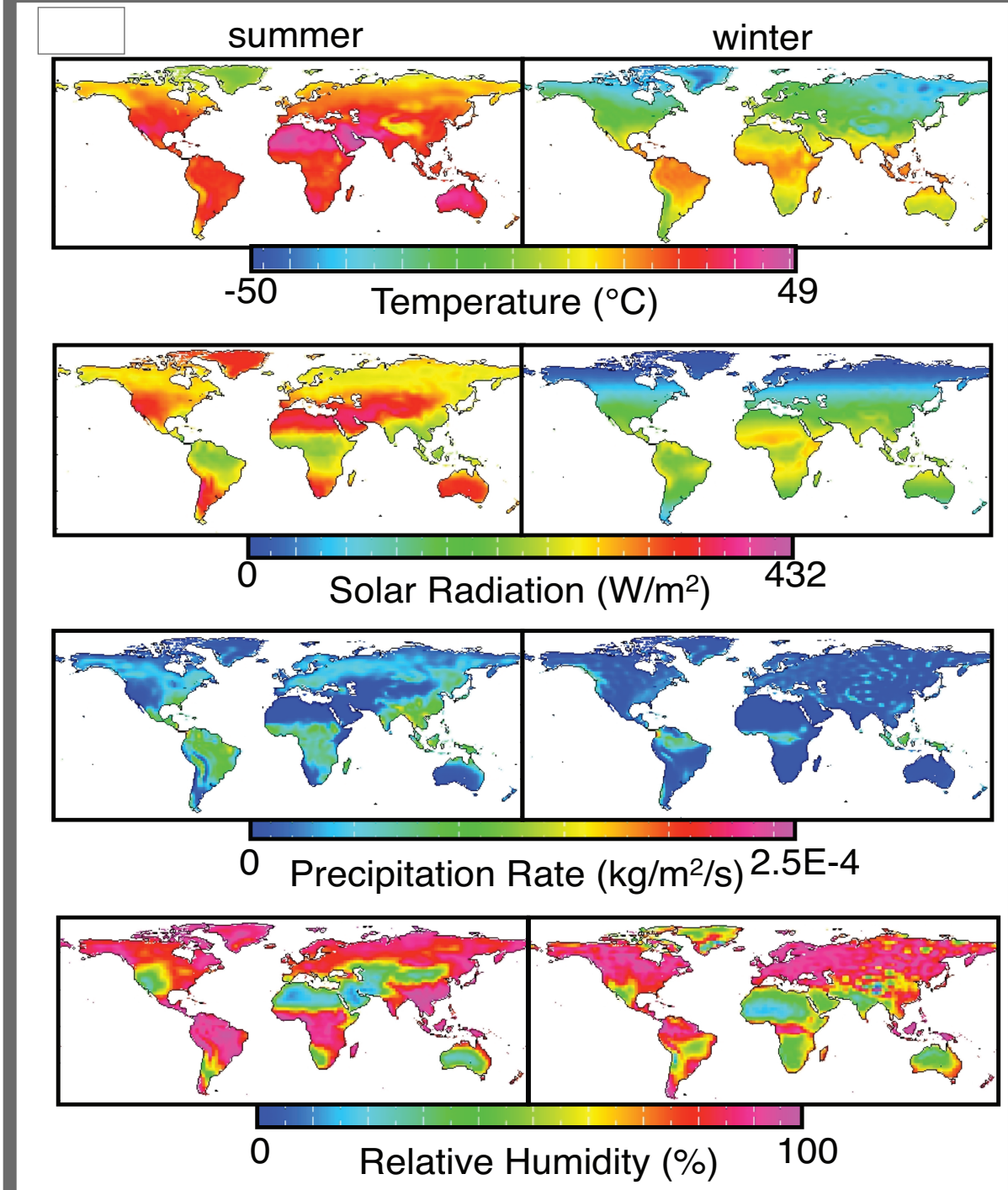
Genome-wide data from 1344 individuals from 61 populations:

- 938 individuals from the Human Genome Diversity Panel
- 288 individuals from HapMap Phase 3
- 118 individuals genotyped for these projects

Climate Variables

Climate data source:

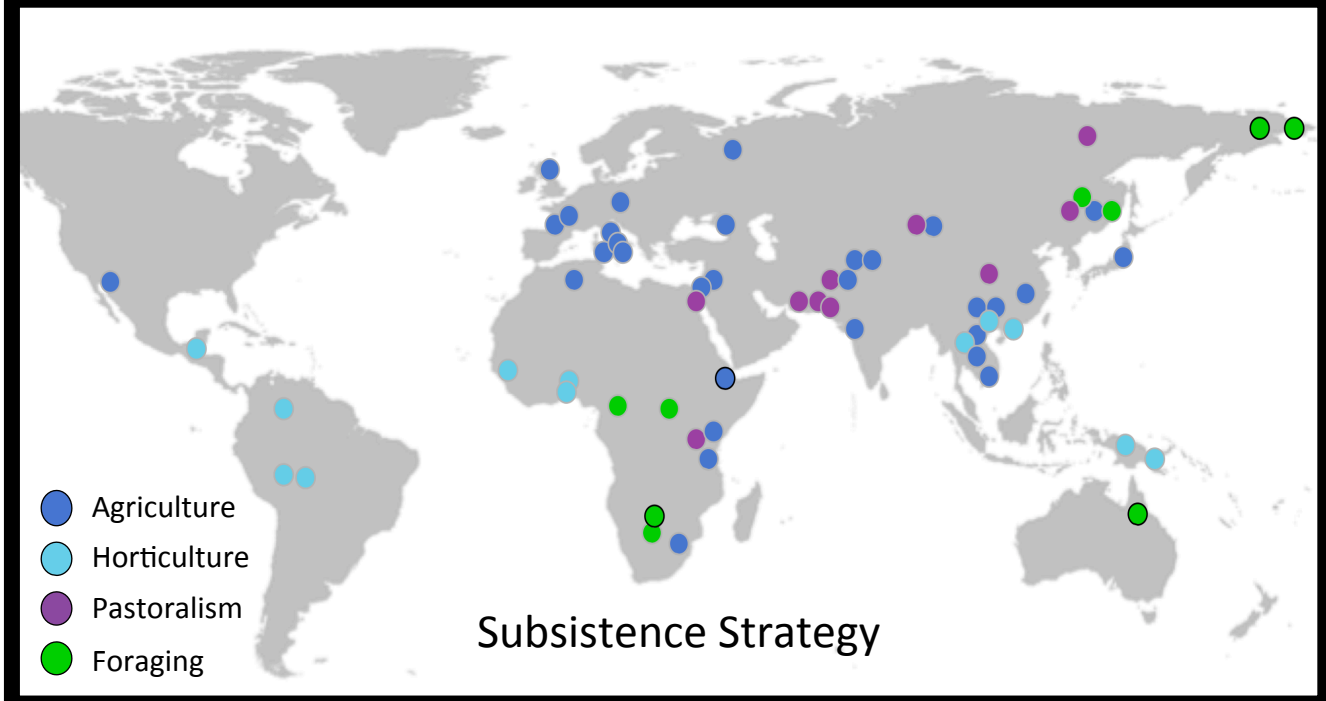
NCEP/NCAR Reanalysis Project (Kistler et al., 2001)

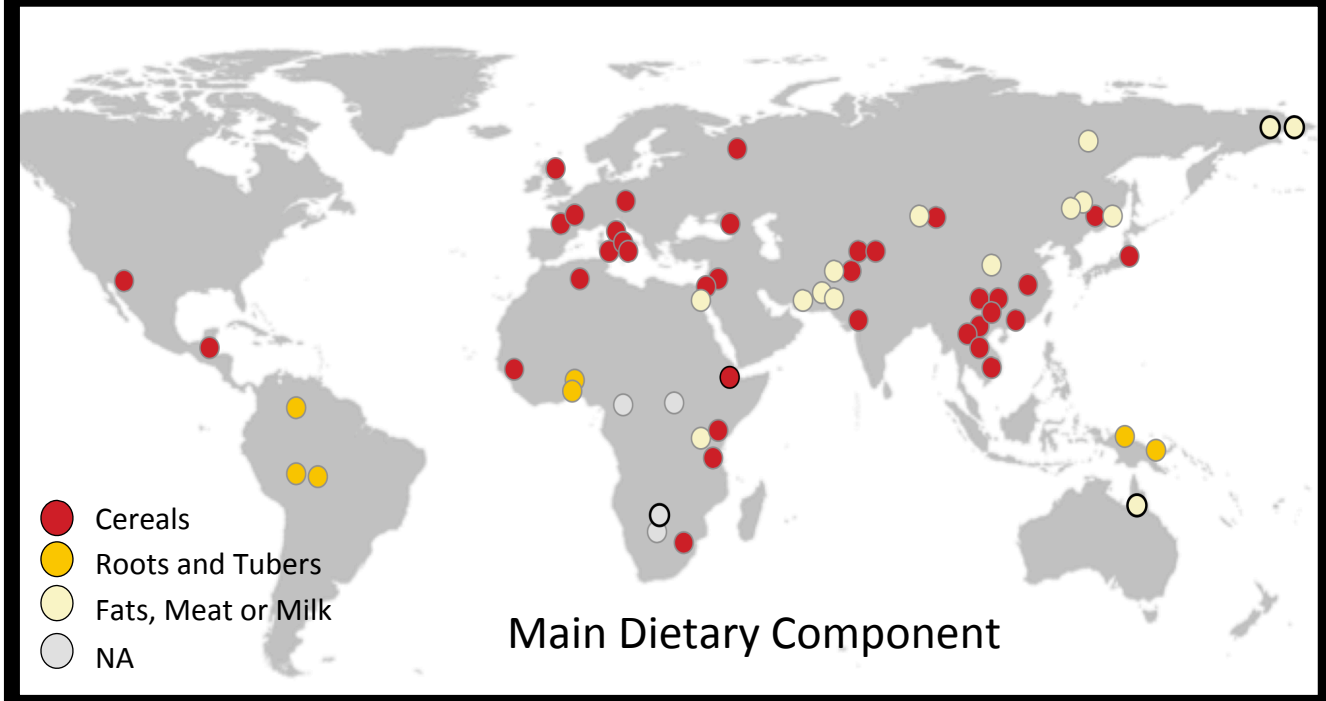


Diet and Subsistence

Data sources:

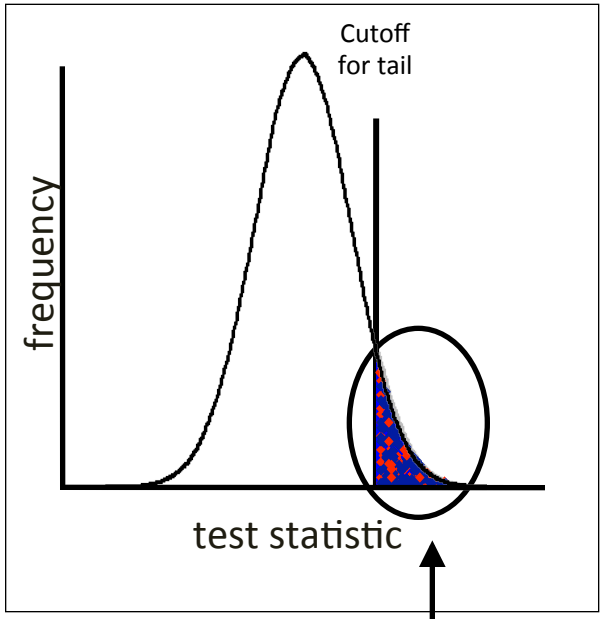
- *Ethnographic Atlas (Murdock 1967)*
- *Encyclopedia of World Cultures (Levinson 1991-97)*





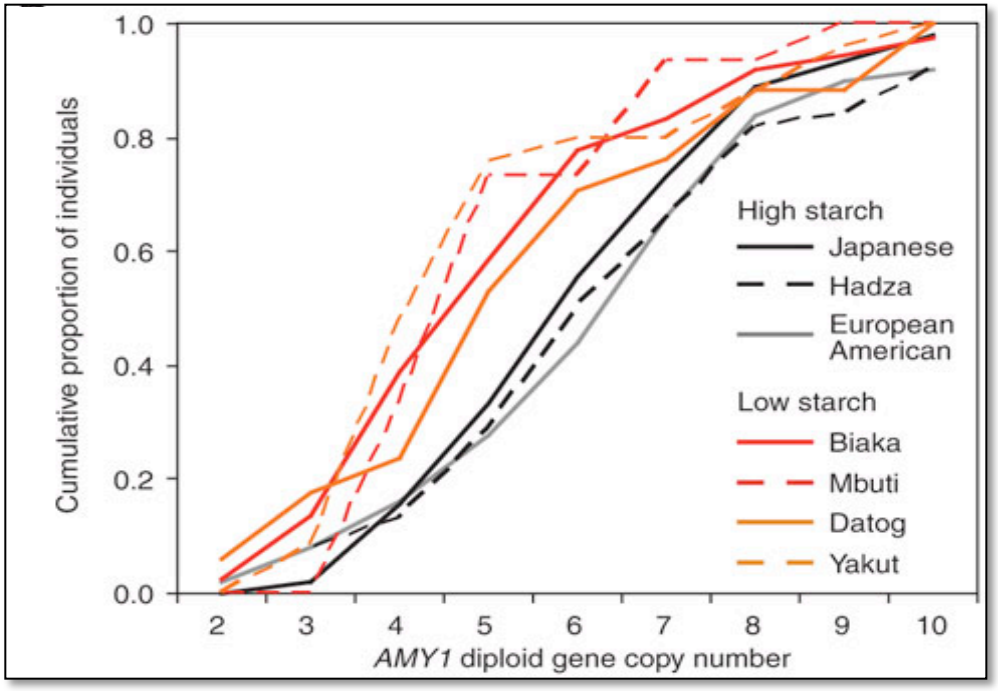
Methodology

- Calculated the correlation between each SNP and environmental variable
- Assessed evidence for enrichments of test SNPs relative to control SNPs in the tail of distribution



Is the proportion of ***test SNPs*** greater than the proportion ***control SNPs***?

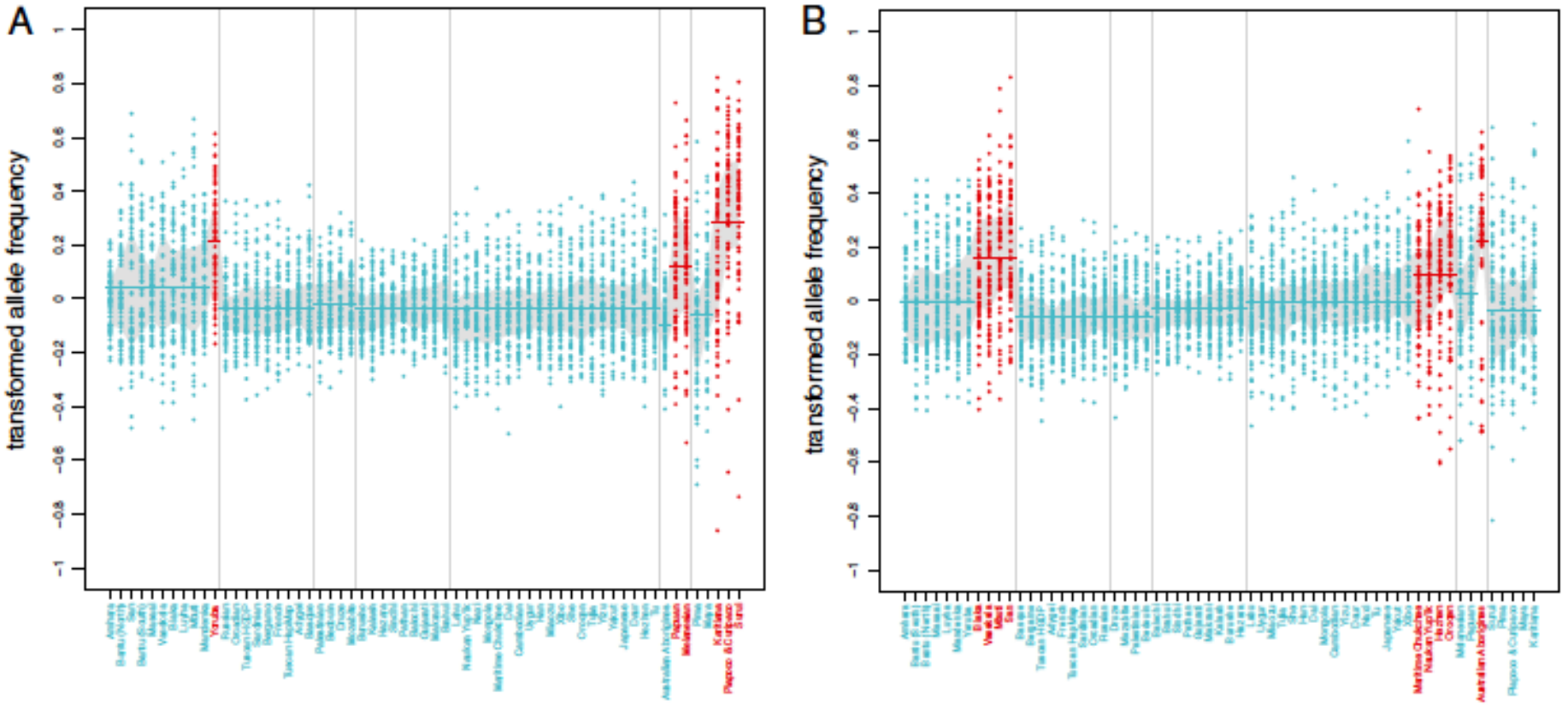
AMY1 underlies variation in starch metabolism among populations



Perry et al, 2007

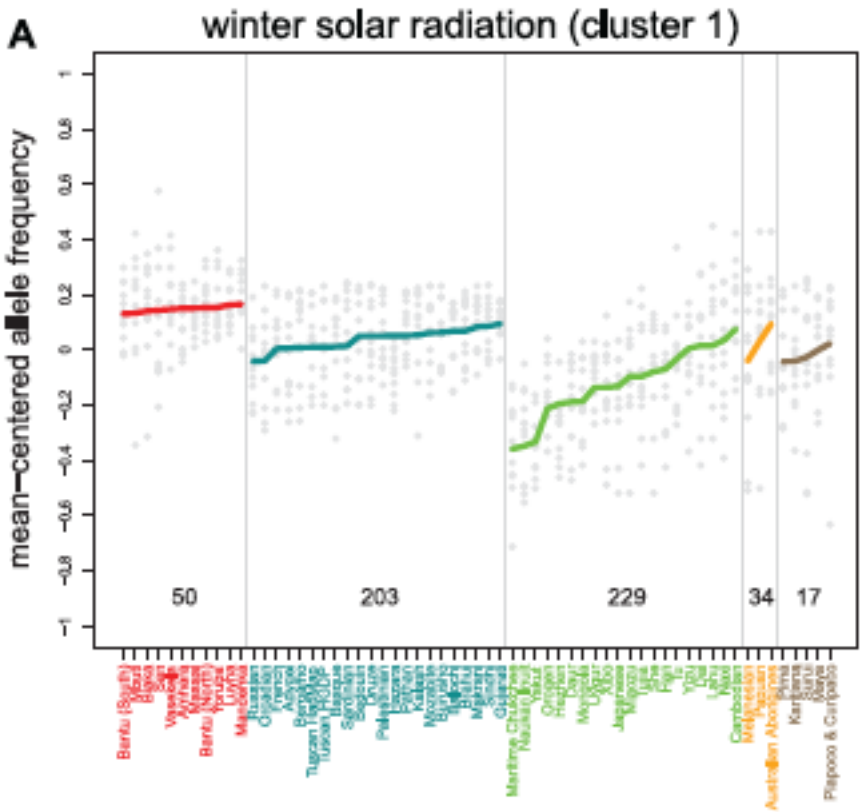
This pattern may translate to variation in disease risk under certain dietary conditions

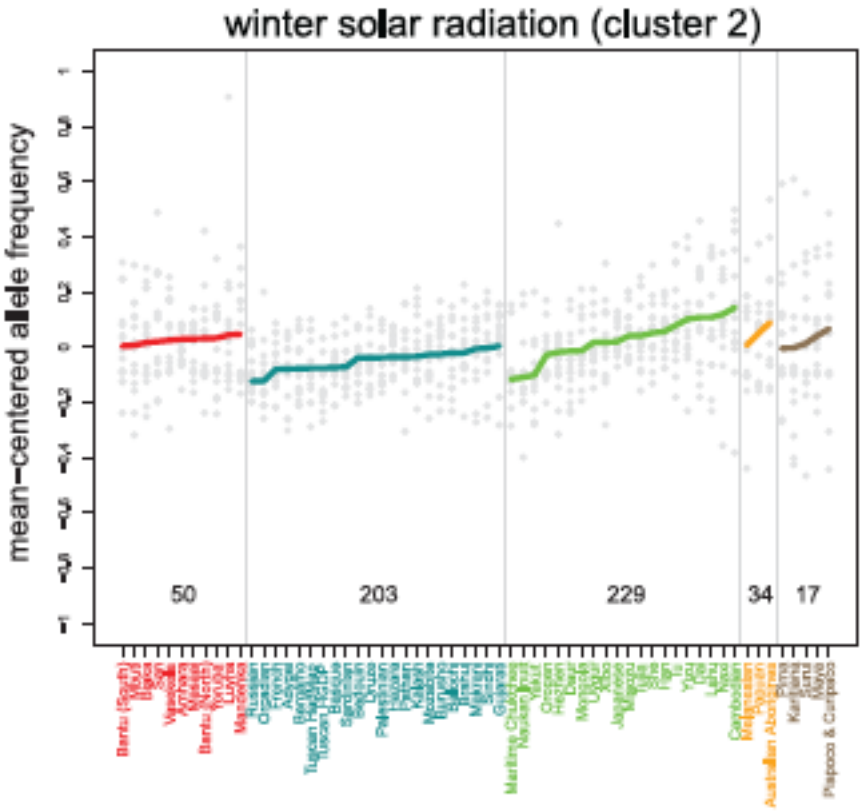
Patterns that come out of our analysis: dichotomous variables



Patterns we found were always driven by concordant shifts in allele frequencies across regions

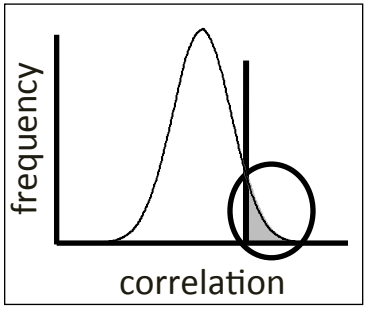
Patterns that come out of our analysis: continuous variables





Correlation are often shared between regions
but can be specific to a single region

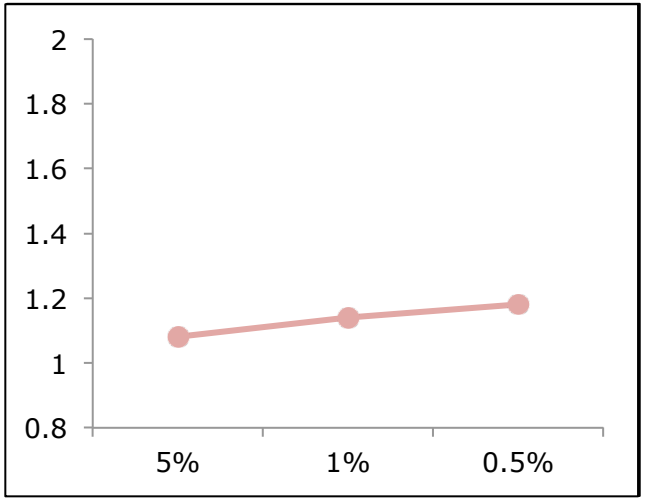
Is there evidence for adaptation to climate and subsistence *overall*?



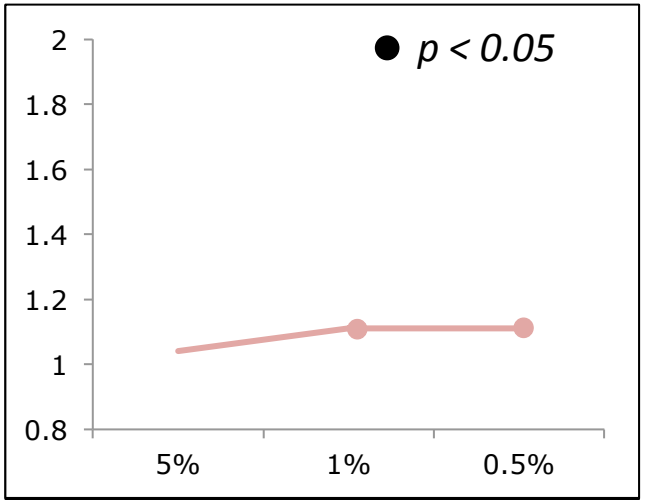
Is the proportion of ***genic SNPs*** > the proportion ***nongenic SNPs***?

Is the proportion of ***NS SNPs*** > the proportion ***nongenic SNPs***?

Climate

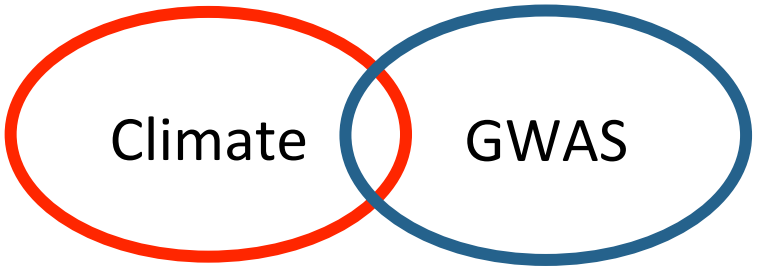


Subsistence



Hancock et al. 2010; Hancock et al. 2011

GWAS have been conducted for 1000s of traits in humans



Overlap with GWAS

SNP	Associated trait	Nearby genes
rs12913832	Hair Color (Black vs. Blond)	HERC2
	Hair Color (Black vs. Red)	
rs1667394	Eye Color (Blue vs. Green)	OCA2
	Hair Color (Blond vs. Brown)	
rs28777	Hair Color (Black vs. Blond)	MATP
	Hair Color (Black vs. Red)	
rs35391	Tanning	MATP
rs2313132	Systemic Lupus Erythematosus	PCDH18
rs4613763	Crohn's disease	PTGER4
rs6074022	Multiple sclerosis	CD40
rs10484554	Psoriasis	HLA-C
rs2187668	Systemic Lupus Erythematosus	HLA-DQA1
rs2187668	Celiac Disease	
rs9461688	IL18 protein levels	HLA-C
rs10484554	AIDS progression	HLA-C
rs11203203	Type 1 Diabetes	UBASH3A
rs20541	Psoriasis	IL13
rs185819	Height	TNXB (HLA class III)
rs6899976	Height	L3MBTL3
rs10486776	Stroke	upstream of MEOX2
rs10488360	Factor VII	upstream of SDK1
rs10490823	Bone Mineral Density (Hip)	upstream of CTNNB1
rs210138	Testicular germ cell tumor	BAK1

Pigmentation

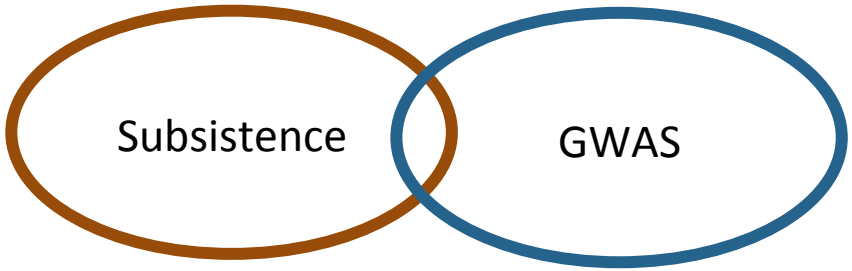
Immunity

Height

Cardiovascular Disease

Hancock et al., PLoS Gen 2011

Overlap with GWAS



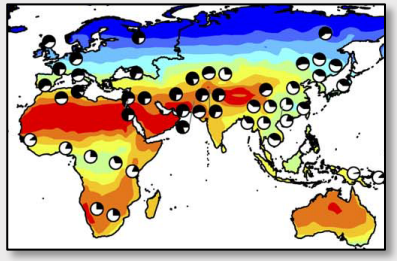
SNP	Associated Trait	Strongest signal is with the variable:	Closest Gene
rs2722425	Fasting Plasma Glucose	Roots and Tubers	ZMAT4
rs17779747	QT interval	Roots and Tubers	KCNJ2
rs2237892	Type 2 Diabetes	Cereals	KCNQ1
rs2269426	Plasma Eosinophil Count	Fat, Meat, Milk	TNXB, CREBL1 (mHC class III)
rs7395662	HDL cholesterol	Foragers	MADD, FOLH1
rs10507380	Electrocardiographic Traits	Pastoral	RPL21
rs9642880	Urinary bladder cancer	Pastoral	MYC, BC042052

Hancock et al., PNAS 2010

Variants associated with metabolic syndrome traits overlap with subsistence

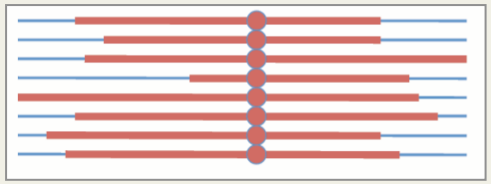
Is detection improved by using the environment?

Environmental Correlations



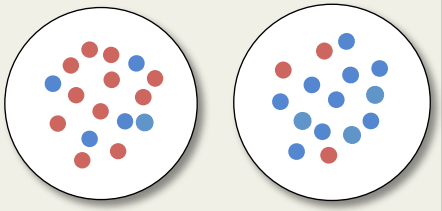
versus

Haplotype Homozygosity



and

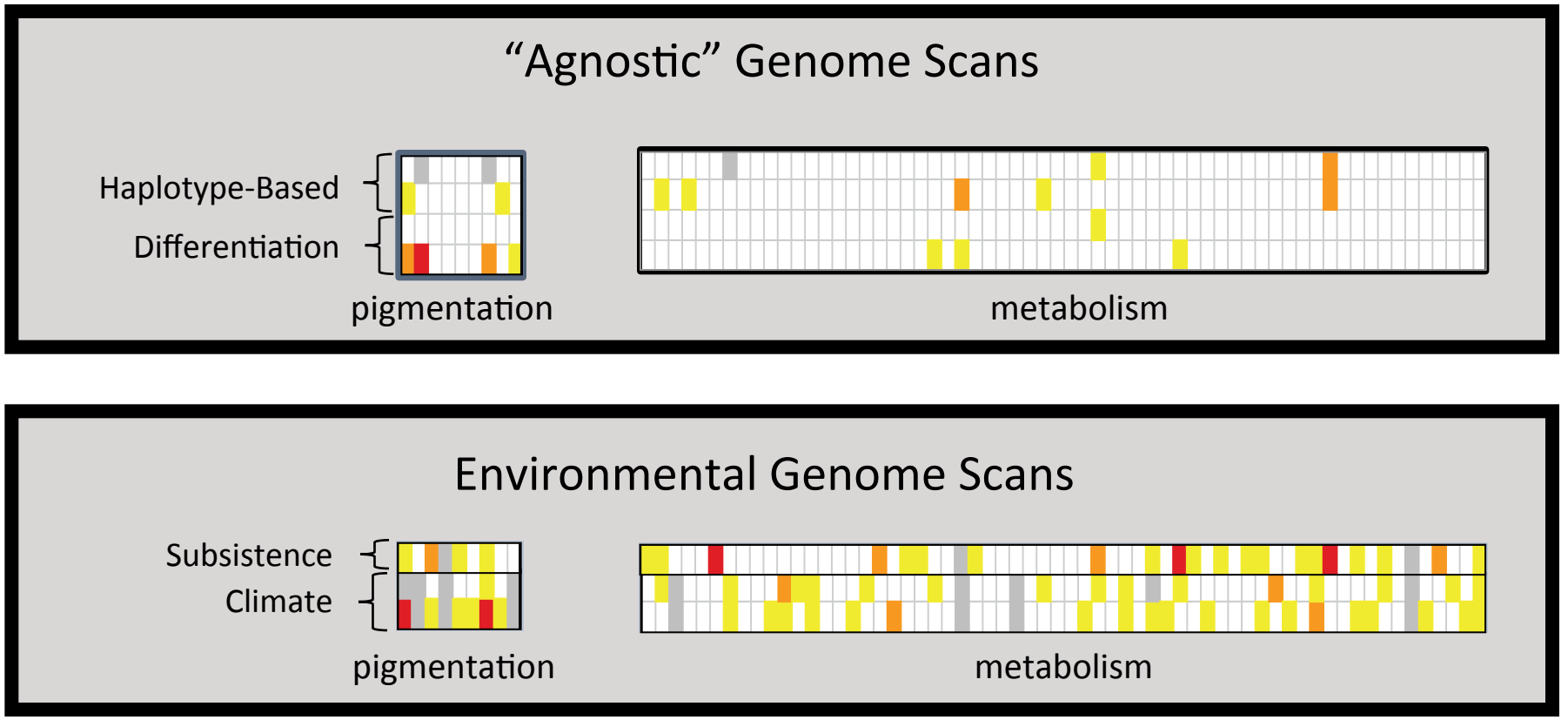
Population Differentiation



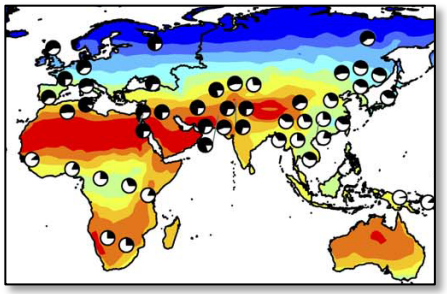
Approach:

1. Compiled sets of variants associated with two phenotypes that are hypothesized to be under spatially-varying selection
2. Compared environmental correlations to traditional methods

EAA improves power compared to traditional approaches



Hancock et al., Philos Trans Roy Soc, 2010



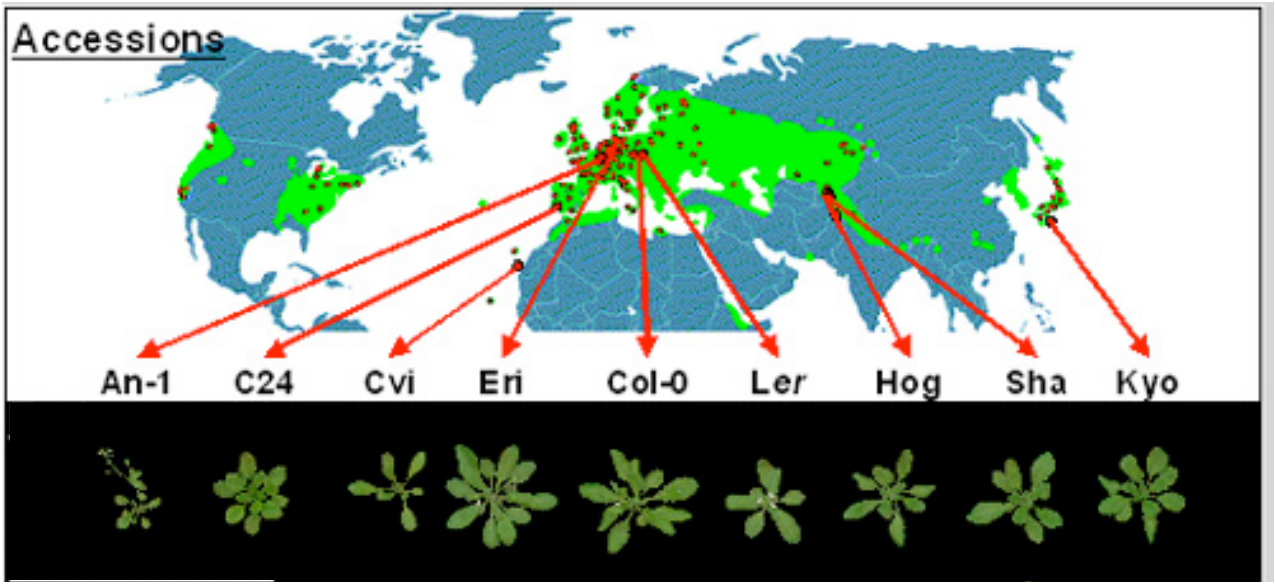
Conclusions: Clinal patterns in humans

- Subtle shifts in allele frequencies across regions and loci were important for adaptation to climate and subsistence in humans

- Environmental association analysis (**EAA**) can identify variants that will be missed using hard sweep or simple differentiation (F_{ST}) to detect positive selection

A. thaliana examples

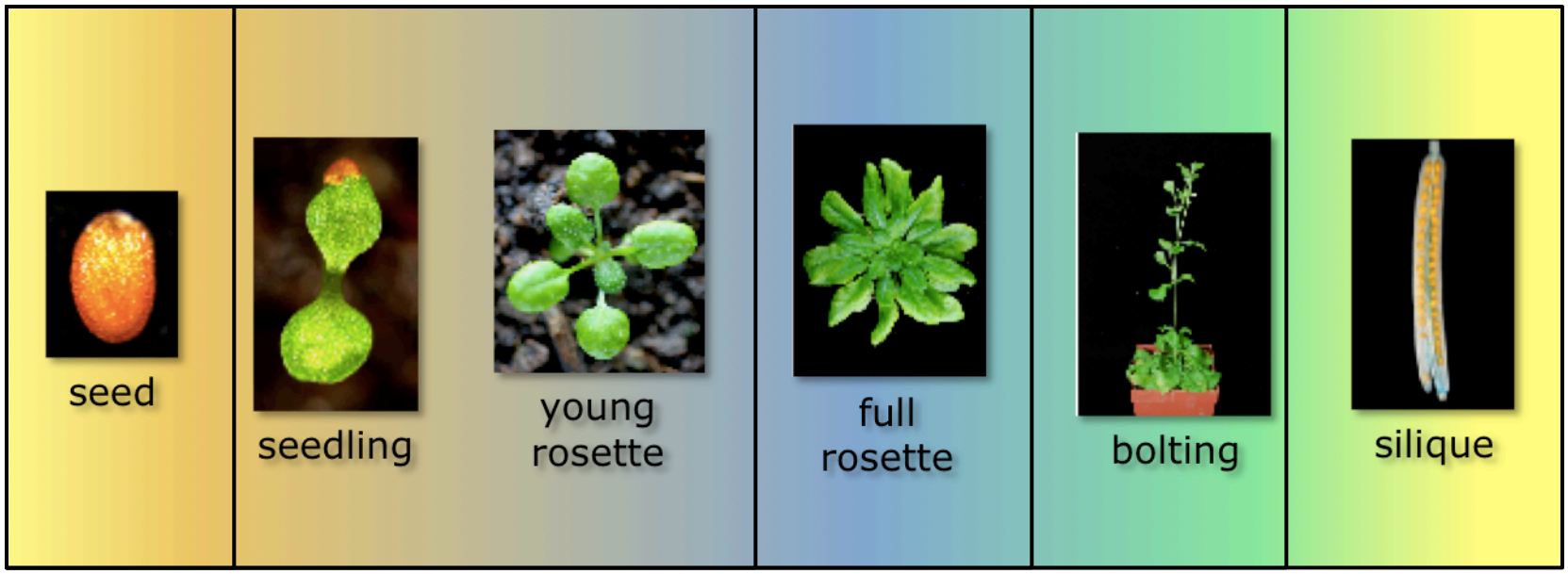
A. thaliana is a useful model for learning about adaptations to climate



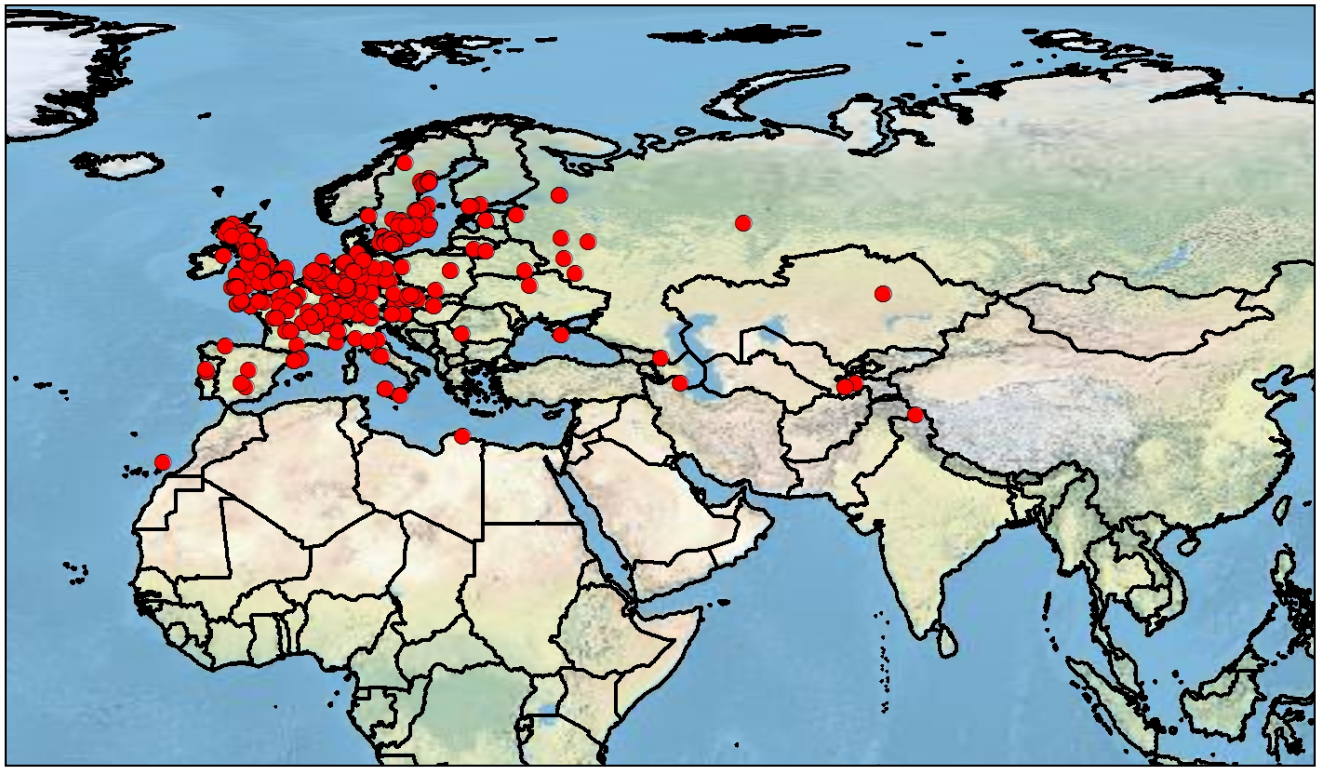
from Van Norman and Benfey 2009, adapted from Matthieu Reymond

- *A. thaliana* occupies a large range with extensive environmental variation
- Since *Arabidopsis thaliana* is largely inbreeding, the same accession can be used for genotyping, phenotyping and for follow-up functional analyses
- Genetic, genomic and molecular biology resources are well-developed for *A. thaliana* and complementary work provides valuable information about biology

A. thaliana is a winter annual



Which loci underlie adaptation to climate in *Arabidopsis thaliana*?



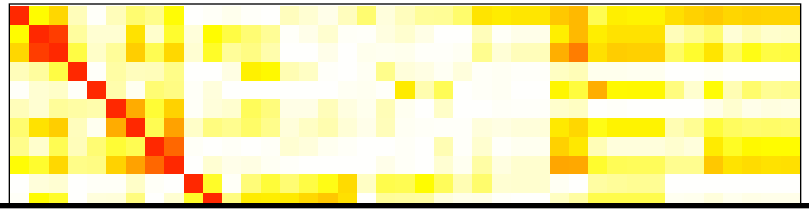
Using data for ≈215,000 SNPs from 948 accessions, I identified variation that is strongly correlated with 13 climate variables

Hancock et al., 2011

Climate variables included in the analysis

41 climate variables summarizing information about:

- Temperature
- Precipitation
- PAR
- Humidity
- Season Lengths
- Aridity

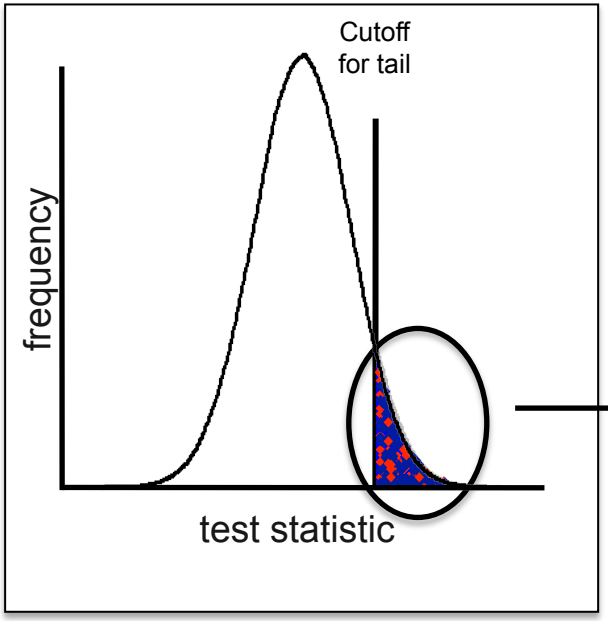


Pruned the complete set to choose 13 representative variables that fit into several general groups:

- Daylength
- Extremes in T and precipitation
- Seasonality of T and precipitation
- PAR
- Humidity
- Season lengths
- Aridity

Do the results represent true signatures of positive selection?

Are variants that are likely to have functional effects enriched in the tails of the climate correlation distributions?



Is the proportion of ***functional SNPs*** greater than the proportion of ***other SNPs***?

How often do we see a similar enrichment in 1000 permuted datasets?

Biological processes are enriched in the tails

photosynthesis
red light signaling pathway

Photosynthesis

indoleacetic acid biosynthetic process*
gynoecium development*
maintenance of root meristem identity*
positive gravitropism*
cellular response to water deprivation*
stomatal complex development
cotyledon development
cotyledon vascular tissue pattern formation*
phloem or xylem histogenesis

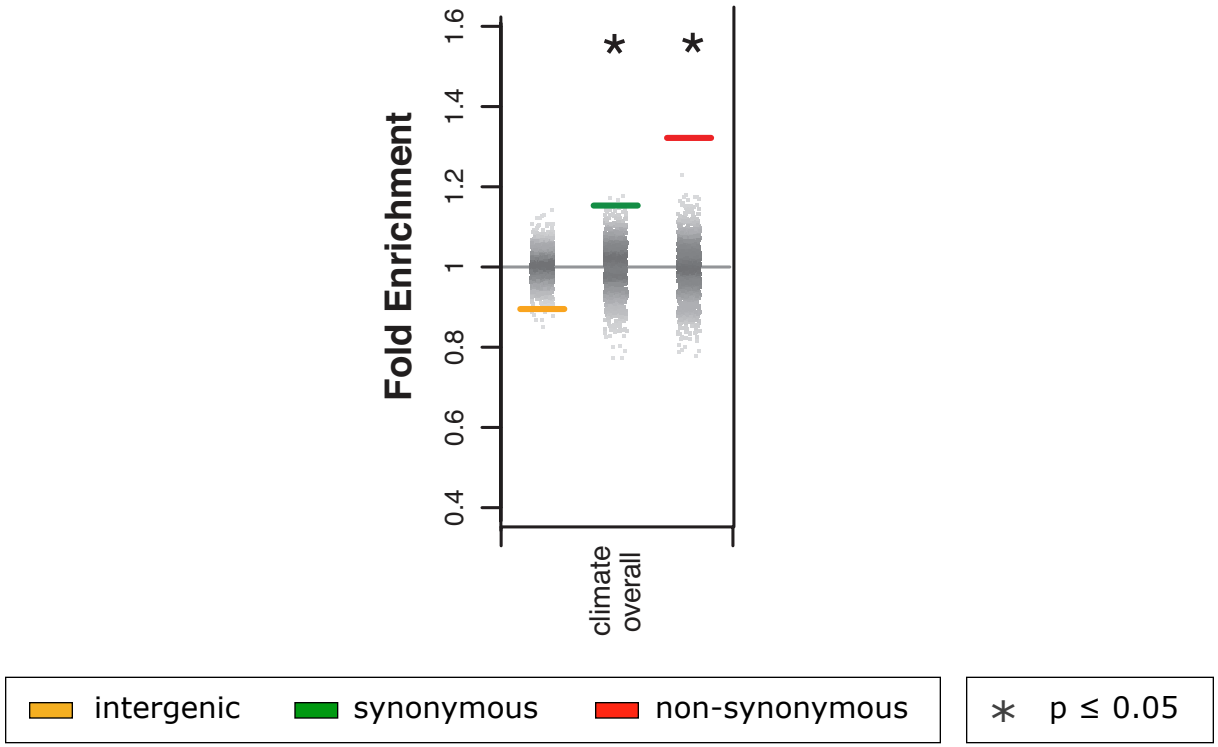
auxin-related processes

regulation of defense response
jasmonic acid mediated signaling pathway

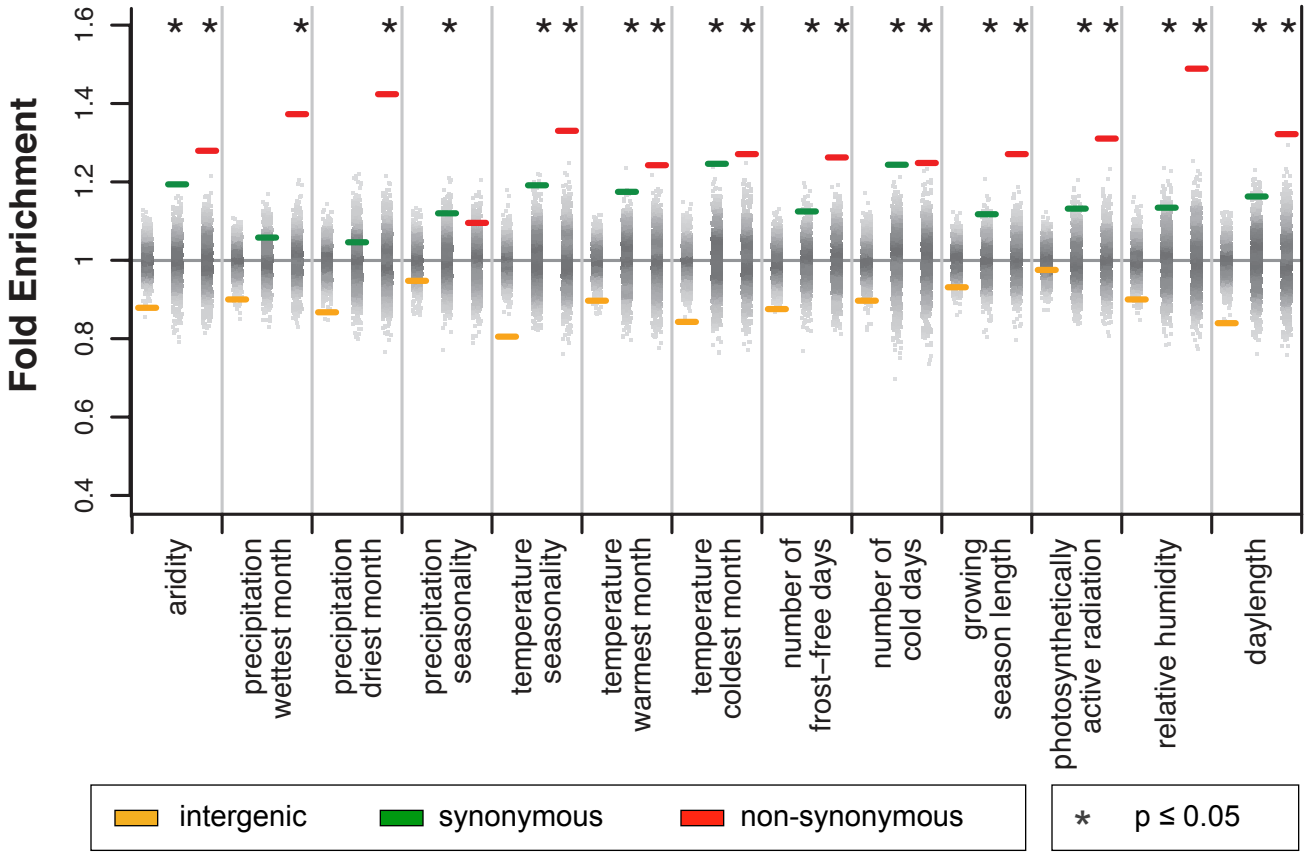
defense

* significant with $p<6.83 \times 10^{-5}$ with Bonferroni

NS SNPs are enriched among top climate associations

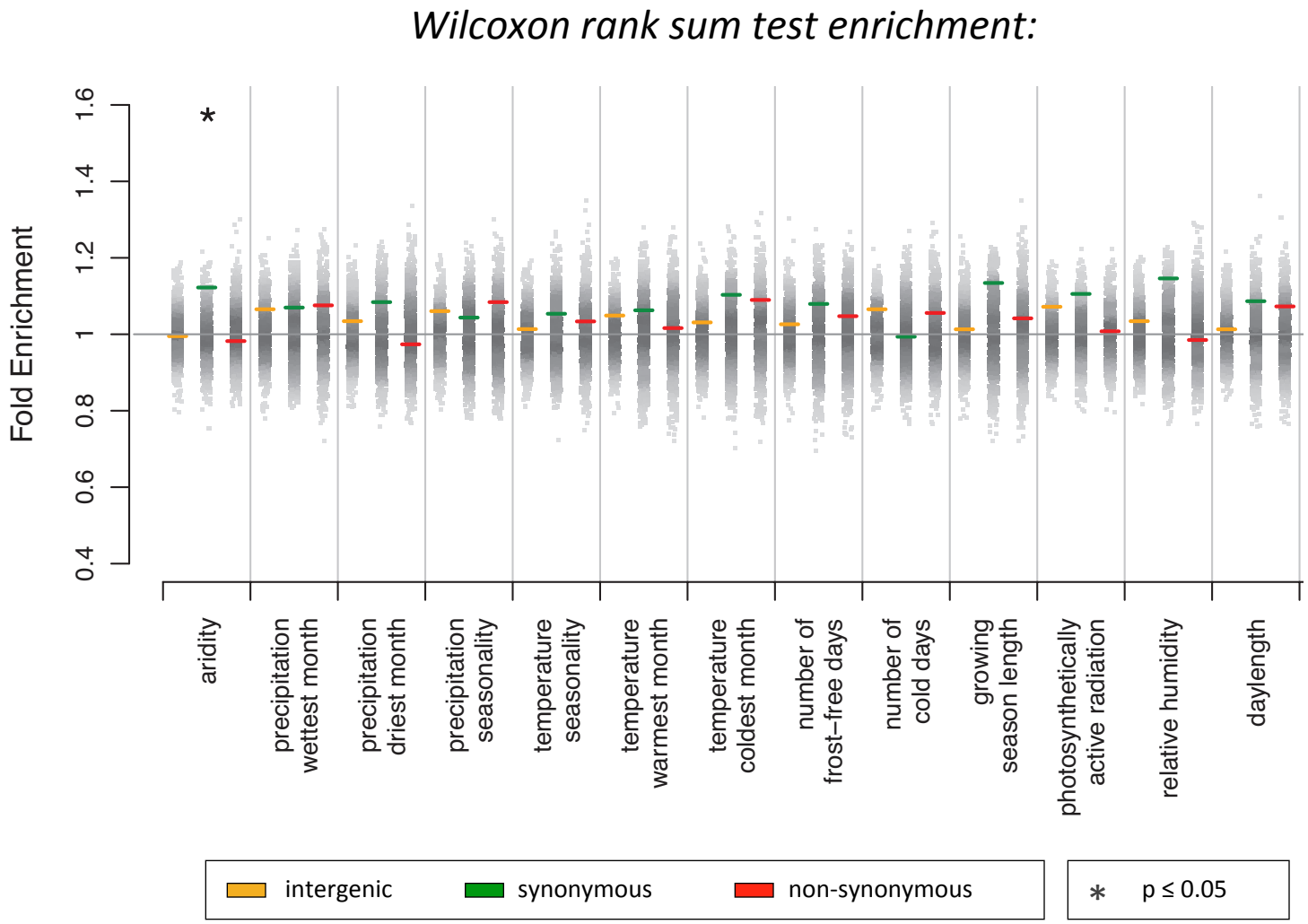


There is variation in signals among variables tested



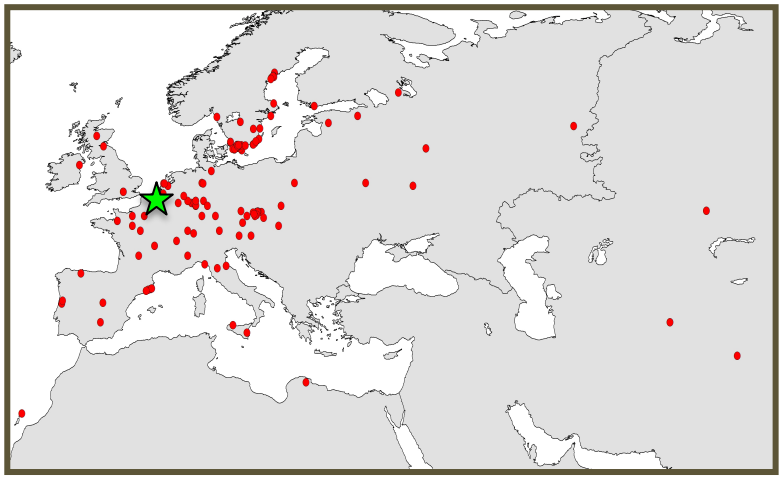
Variables related to water availability show the strongest signals

No NS enrichment using a method that doesn't account for relatedness among individuals



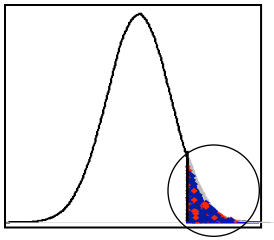
The climate correlation results predict relative fitness in a common environment

147 accessions

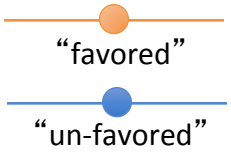


Planted in a common garden in France

Can we predict relative fitness in a particular climate using genome scan results?



SNPs from the extreme tail of the climate correlation distributions

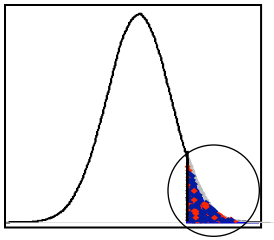


For each SNP, determine which allele is likely to be favored in the climate of interest

Fitness ~ # favorable alleles

Ask whether the number of “favorable” alleles predicts fitness

Can we predict relative fitness in Lille, France?



SNPs from the extreme tail of the climate correlation distributions

- Selected all SNPs in the 0.01% tail for any of the 13 variables
- If found for multiple, took the variable with the most extreme signal
- Removed SNPs in linkage disequilibrium

285 SNPs

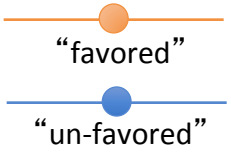
↓

254 SNPs

↓

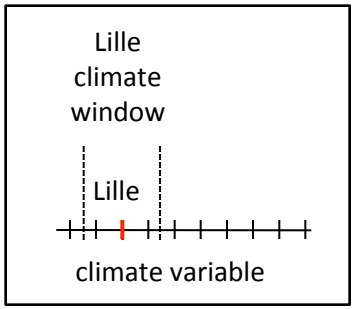
108 SNPs

Can we predict relative fitness in Lille, France?

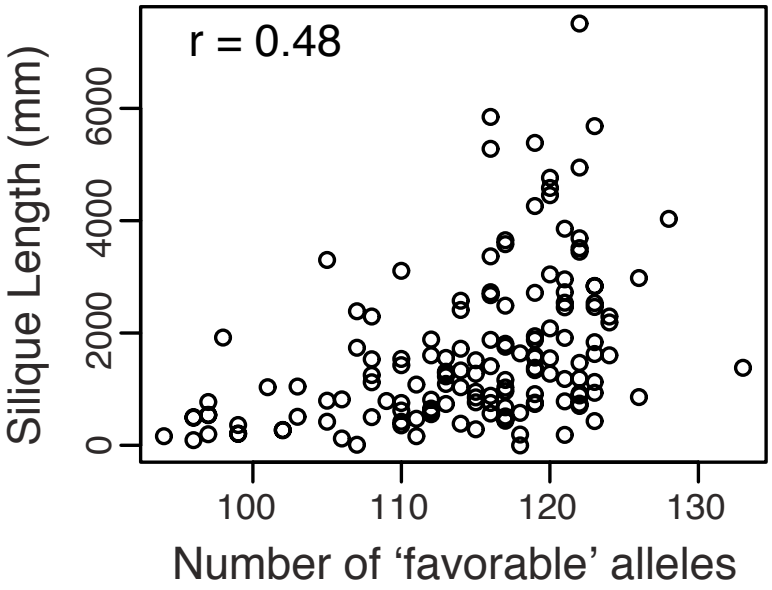


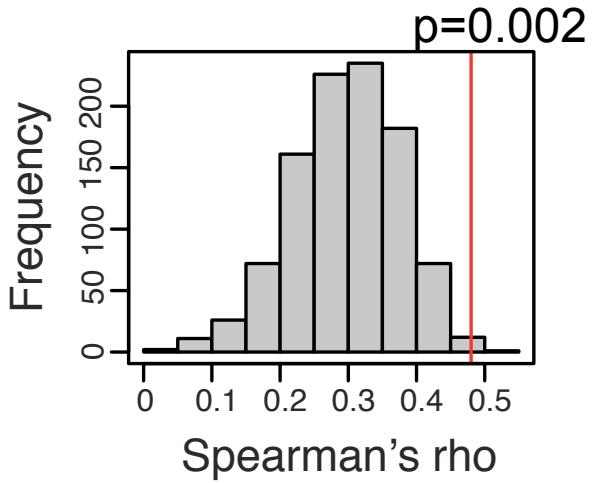
For each SNP, determine which allele is likely to be favored in the Lille climate

- Defined a "climate window" for Lille based on the distribution of climate for complete set of 948 accessions
- For each SNP, determined the "favored" allele by asking which allele was more common in the *Lille climate window*
- Calculated the the number of 'favorable' alleles for each of the 147 accessions with fitness data



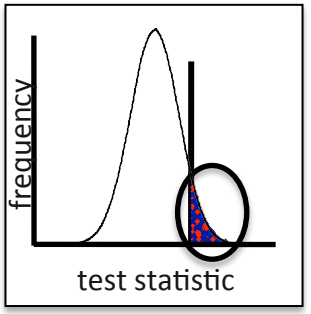
The climate correlation results predict relative fitness in a common environment





The tails of the climate correlation distributions are enriched for locally adaptive variants

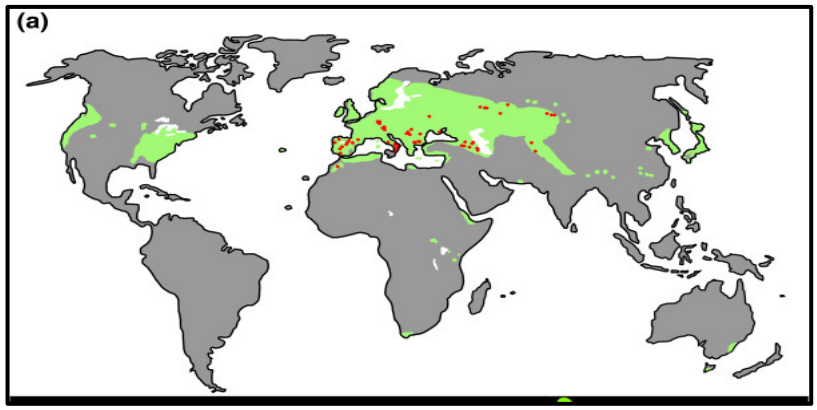
Likely functional loci are enriched in the tails of the climate correlation distributions



1. Reasonable GO categories
2. NS/S SNPs
3. Loci involved in local adaptation from a field fitness study

Evidence for local adaptation in *A. thaliana*

Evidence of climate adaptation in the ‘1001 Genomes’ Project data

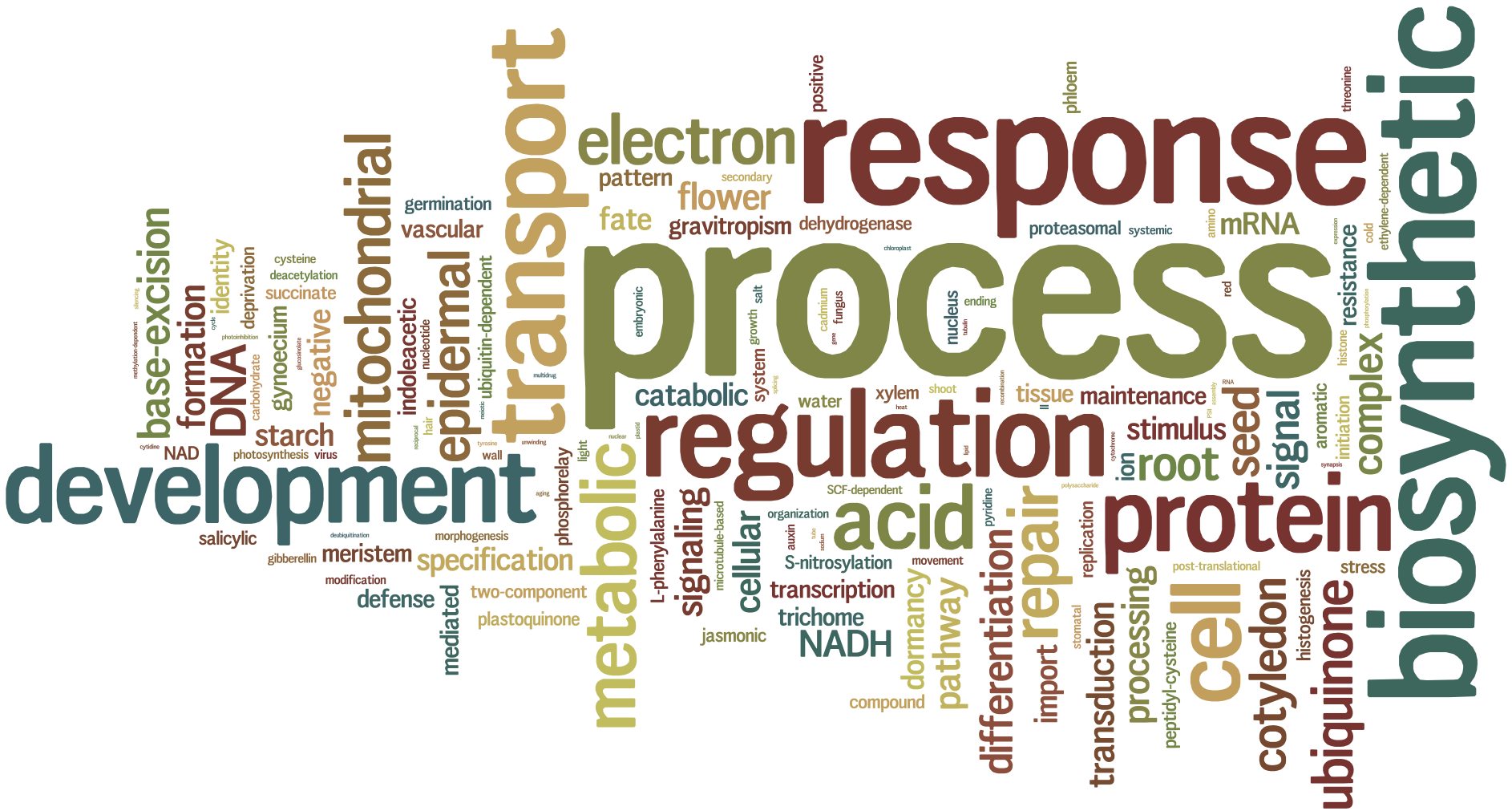


Complete genome sequence data (Illumina) from 1137 accessions collected across Eurasia and North America

1001 Genomes Consortium, 2016

Which phenotypes were important for adaptation to climate?

Which biological processes are over-represented among the strongest correlations with climate?



Biological processes enriched with precipitation-related variables

Precipitation in the driest month

- **Maintenance of root meristem identity**
- Indoleacetic acid biosynthetic process
- Mitochondrial electron transport,

Precipitation in the wettest month

- **Pyridine nucleotide biosynthesis**
- Base-excision repair
- **Root hair cell tip growth**
- **Stomatal complex morphogenesis**

Enriched biological processes provide validation and suggest novel hypotheses

1001 Genomes Consortium, 2016

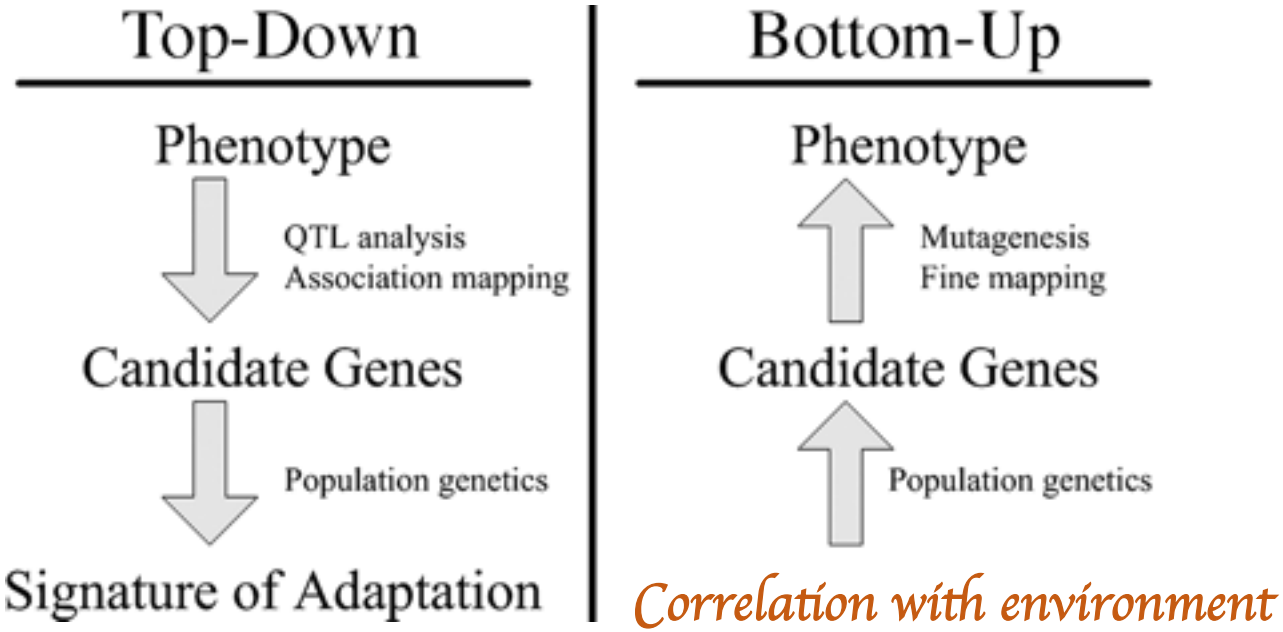
Precipitation variables have the strongest enrichment of signals

10 loci with false discovery rate < 5%

- tRNA/rRNA methyltransferase (NS)
- 2 CYP450s (NS)
- Phosphatidylserine decarboxylase 3 (NS)
- Ethylene Response Factor 1 (ERF1)
- Electron carrier, oxidation reduction
- Copia TE
- Calmodulin-binding
- Beta-glucosidase 37 (Mirosinase)
- Pectin-like super-lyase
- Mildew-resistance (ATMLO11)

Environmental association analysis is applicable to diverse species

‘Top-down’ and ‘Bottom-up’ approaches for identifying adaptive variation



Ross-Ibarra et al., 2007

Increasing number of genome-wide SNP datasets enable EAA in diverse species

Examples of additional applications of EAA approaches*

- Drosophila (Adrion et al. 2015, Machado et al. 2016)
- Medicago (Yoder et al. 2014)
- Pine (Eckert et al., 2012, Yeaman et al. 2016)
- Teosinte/Maize (Pyhäjärvi et al. 2013)
- Norway Spruce (Chen et al. 2012)
- Arabidopsis lyrata (Turner et al. 2008)
- Sheep (Kijas et al., Lv et al. 2014)
- Sticklebacks (Guo et al. 2015, Wang et al. 2014)
- Soybean (Leamy et al. 2016, Bandillo et al. 2017)
- Sorghum (Lasky et al. 2015)
- Spruce (Hornoy et al. 2015)
- Arabidopsis halleri (Fischer et al. 2013, Kubota et al. 2015)

*mainly candidate gene, GBS or exome studies