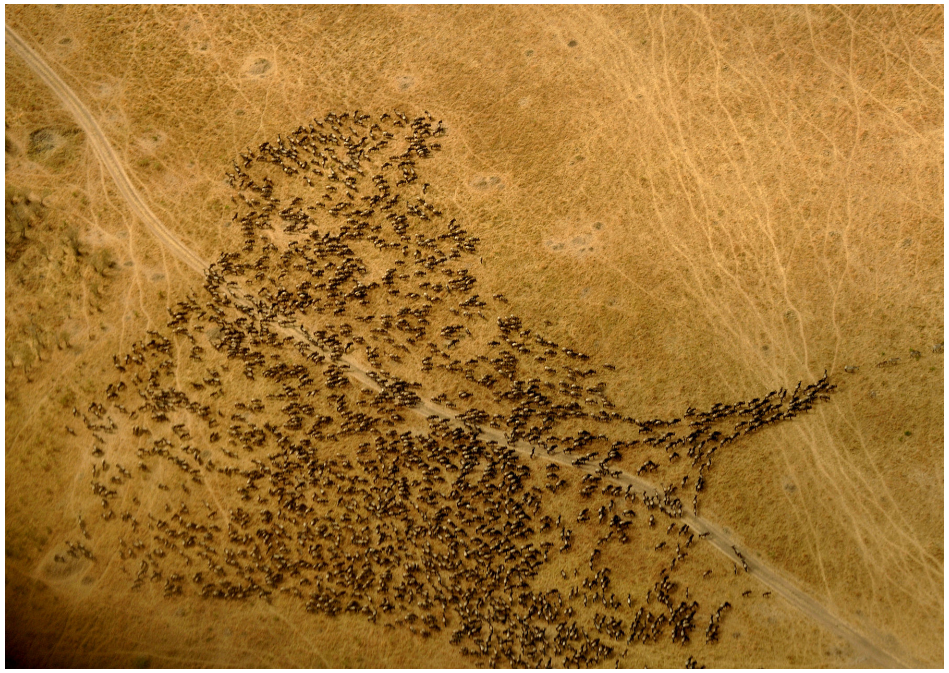
UBC Bioinformatics Class

Topic 7: Genome scans and signatures of demography and adaptation

Ultimate aims

As biologists, we are often interested in understanding demography and selection

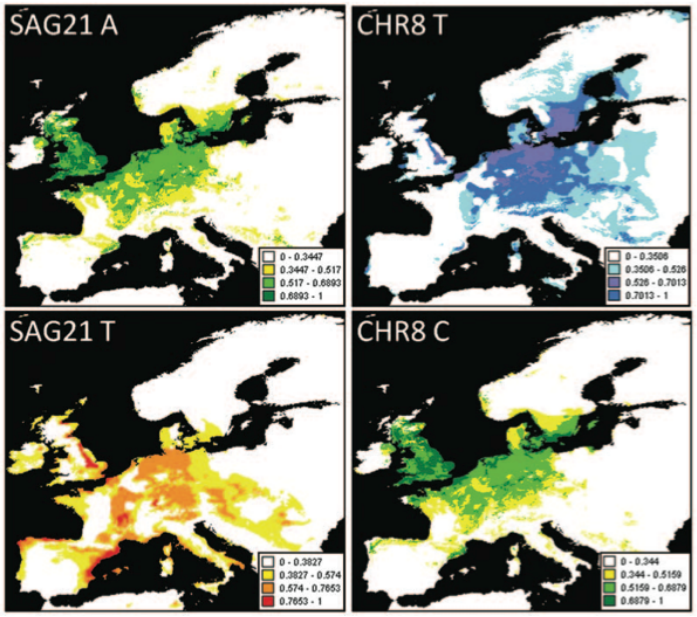


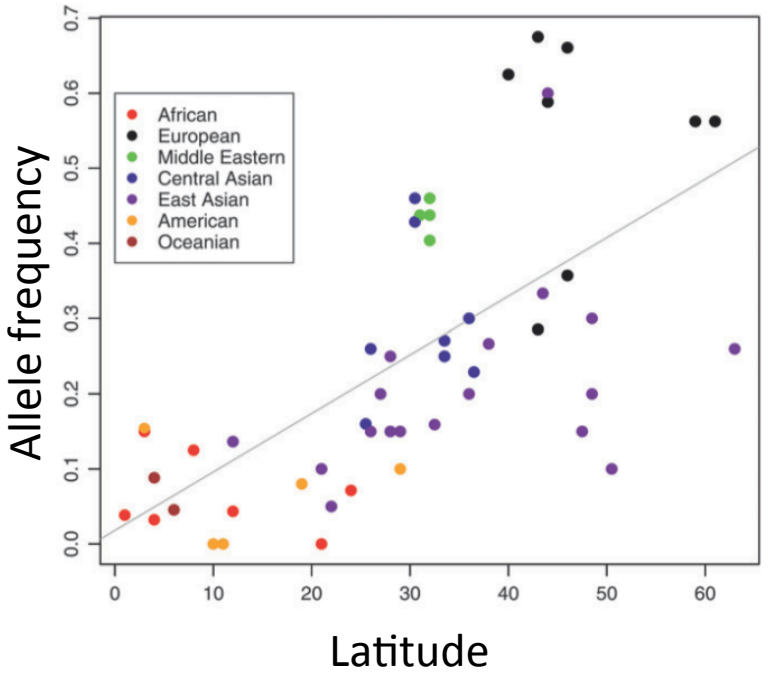
Nachman et al. (2003; PNAS)

Ultimate aims

Experimental approaches or direct observations of these phenomena are not often practical (due to sample size, logistics, etc.) or not possible (historical events)

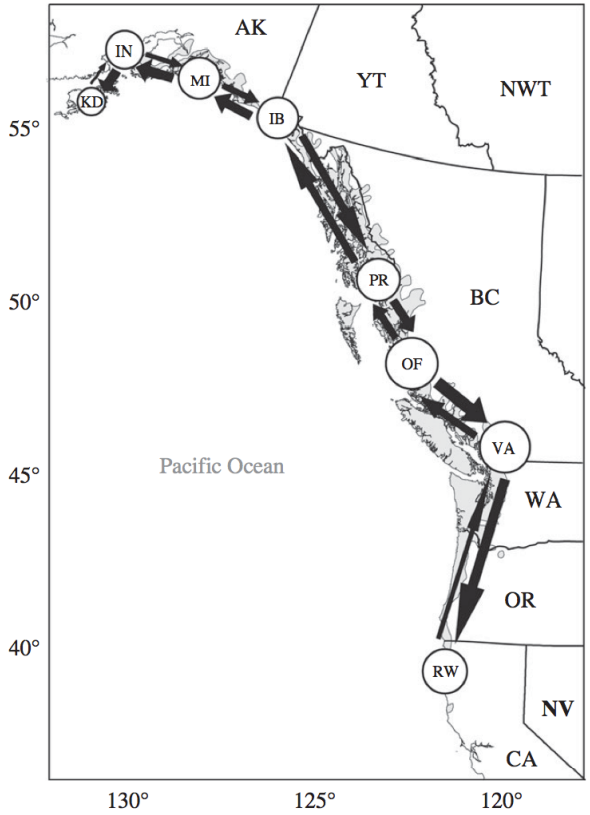
Indirect evidence through population genomic signatures

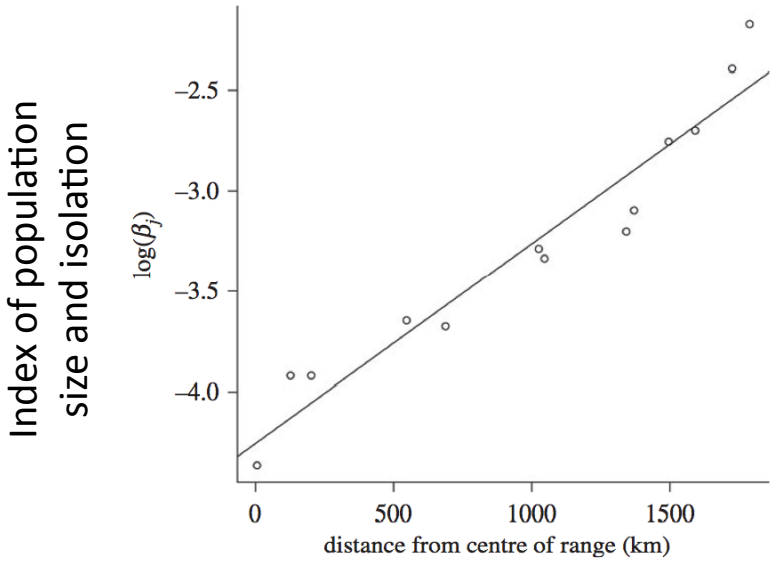




Statistics averaged across the genome

Example: how much gene flow between populations?





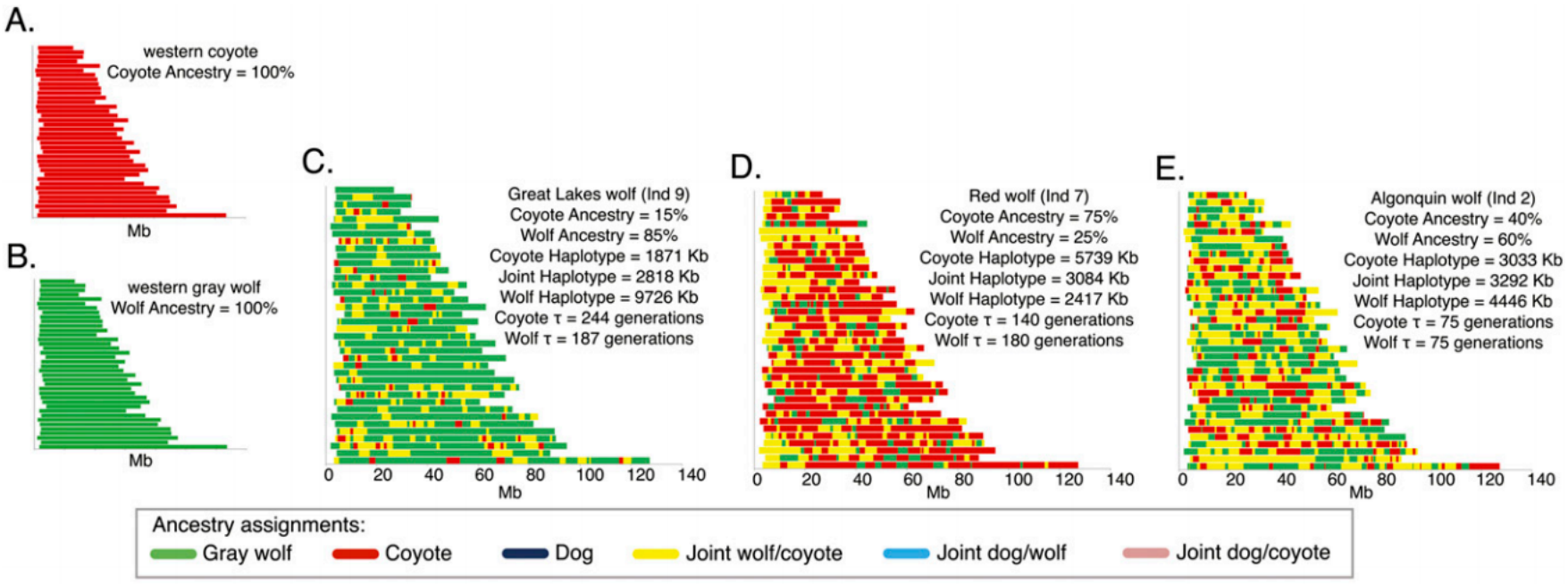
Run BayesFST on 339 SNPs to estimate migration rate and population isolation…now possible to expand this to thousands or millions of SNPs

Genomic data provides greater power due to a large number of loci

Holliday et al. (2011; PRSLB)

Patterns of variation across the genome

Example: Is a hybrid zone porous and genomically heterogeneous?

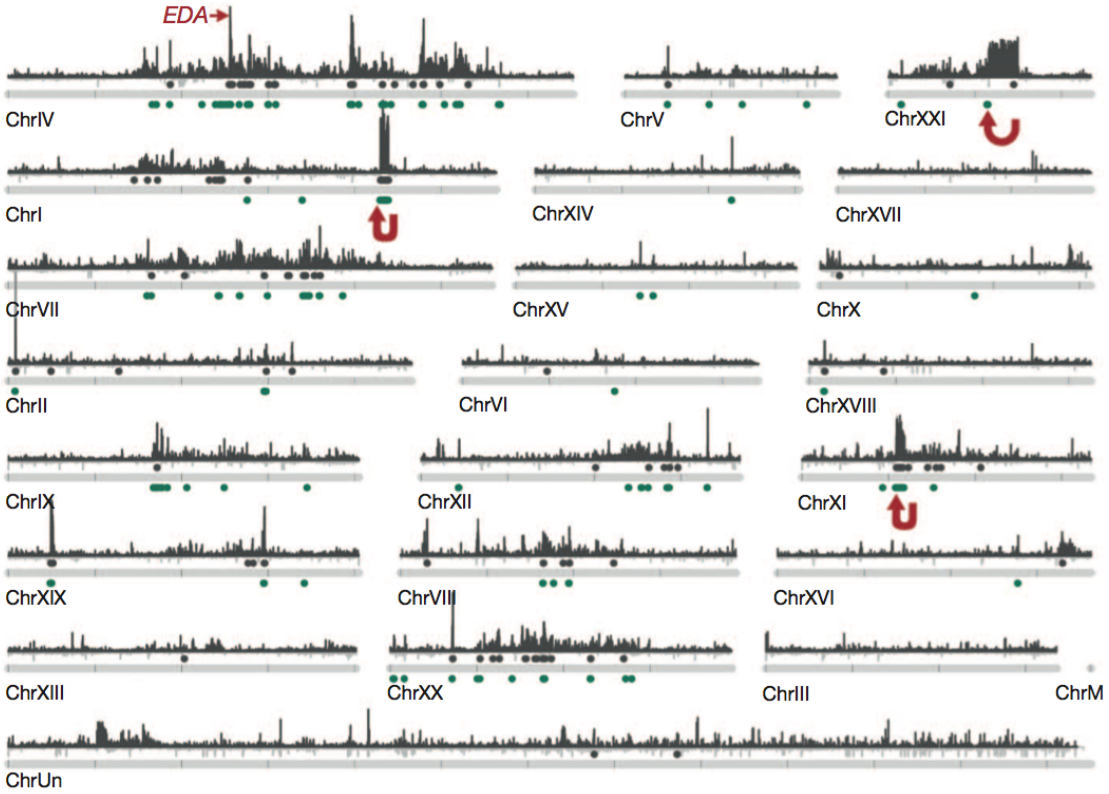


Genomic data gives greater resolution to evaluate and illustrate patterns

vonHoldt et al. (2011; Genome Research)

Genome scans and outlier analyses

Example: Which loci are contributing to local adaptation?

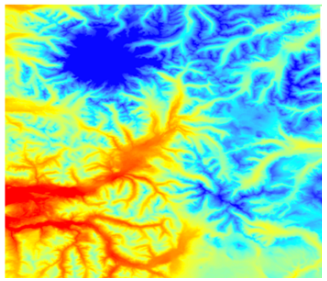


The combination of resolution and power with genomic data enables new statistical approaches that are impossible with "small data"

Jones et al. (2012; Nature)

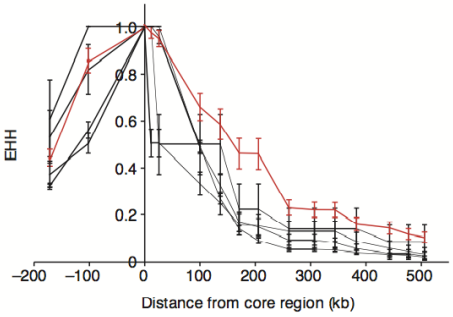
Outlier tests and signatures of selection

- **Landscape level:** differences in allele frequency among populations or environments



- **Phenotypic level**: associations between alleles and locally adapted phenotypes

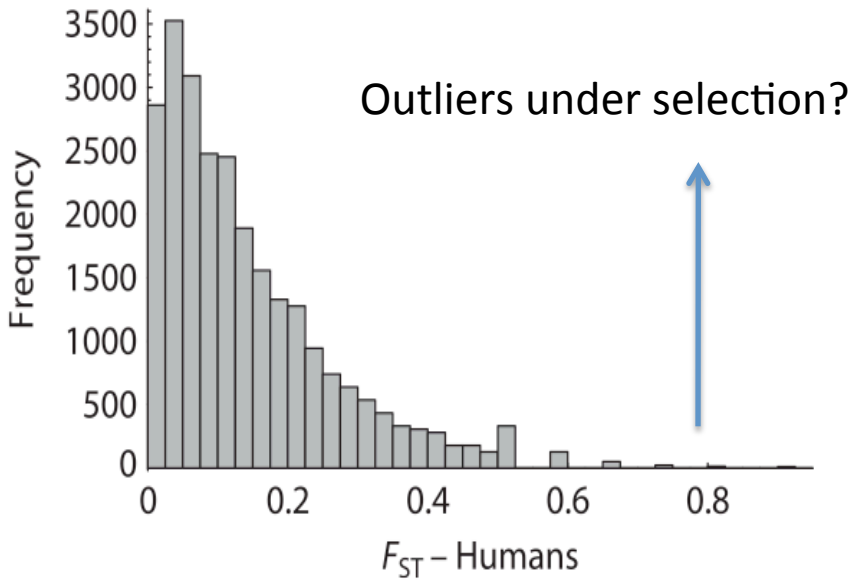
- **Sequence level**: changes in allelic diversity along a chromosome



Diversity, differentiation, and associations with phenotypes and factors affecting fitness

Landscape-level signatures of selection

F_{ST}-outlier tests



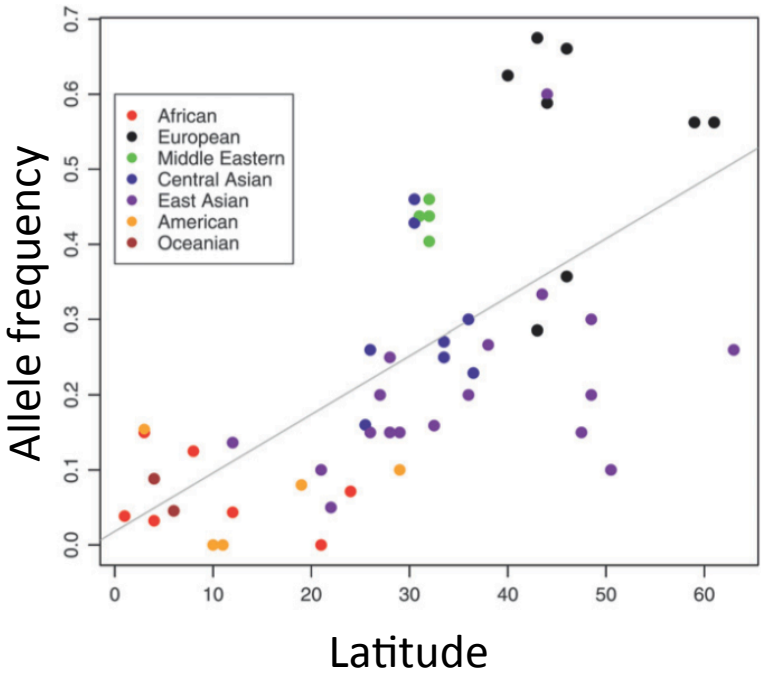
Is the variance among populations in allele frequency more than you would expect by chance?

- Compare the outliers to some background distribution expected under a null model to evaluate significance
- **BayeScan:** uses a Dirichlet distribution, akin to assuming no migration nor mutation since common ancestry
- **FDIST:** assumes an island model, uses IM to estimate demography
- **OutFLANK:** Trims the outliers, estimates X^2 distribution

Courtesy of K. Lotterhos

Landscape-level signatures of selection

Environment-Allele Associations (EAA's)



Is a SNP more highly correlated to an environment than background genomic pattern?

- Test correlation between SNP and environment after controlling for population structure
- **Bayenv:** estimates covariance matrix representing population structure from separate set of "neutral" loci
- **LFMM:** Latent-factor mixed model; estimates the population structure from SNPs in test panel
- **Beware tools don't control for population structure**

Coop et al. (2010)

Phenotypic-signatures of adaptation

SNP-phenotype associations (GWAS): one allele at a time

- Regression of phenotype on SNP (or the reverse)
- Use PCA or STRUCTURE as a covariate in a linear model and/ or a kinship matrix of relatedness in a mixed effects model
- Yields an estimate of the association between SNP and phenotype beyond what would be expected due to population structure
- Many analysis tools and improvements (used in breeding)

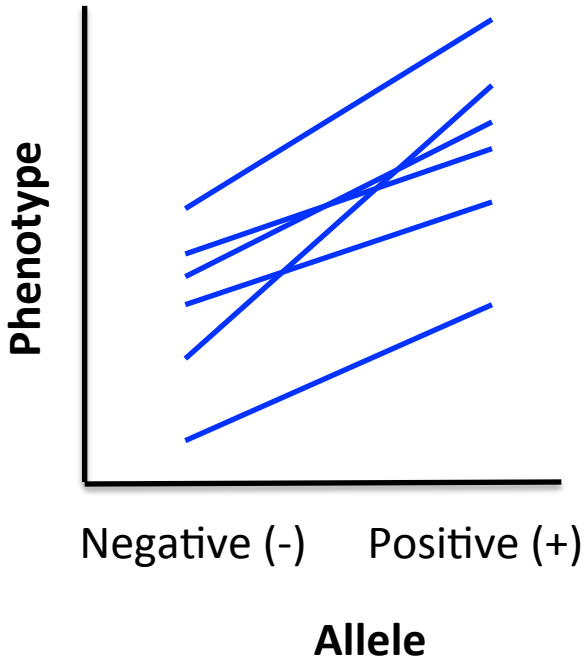
Genomic selection/prediction: multiple alleles at once without emphasis on the individual importance of each allele

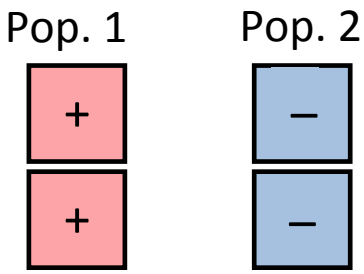
Combined approaches

GWAS-informed estimates of alleles contributing to adaptation

Berg + Coop (2014):

Get the slope of the relationship between SNP and phenotype from GWAS:





Then test whether the covariance between alleles across all phenotype-affecting loci is greater than expected based the underlying covariance among neutral loci

Does + at one locus co-occur with + at another?

Selection sweeps and diversity statistics

Population genetic metrics:

- dn/ds
- McDonald-Kreitman
- HKA

Rate of evolution in coding vs. noncoding sites (usually between species)

- Tajima's d
- Fey and Wu's H
- Extended haplotype homozygosity
- Nucleotide diversity

Diversity statistics

Application to population genomic data

Statistical analysis depends upon the structure of the underlying sequence data and reference…

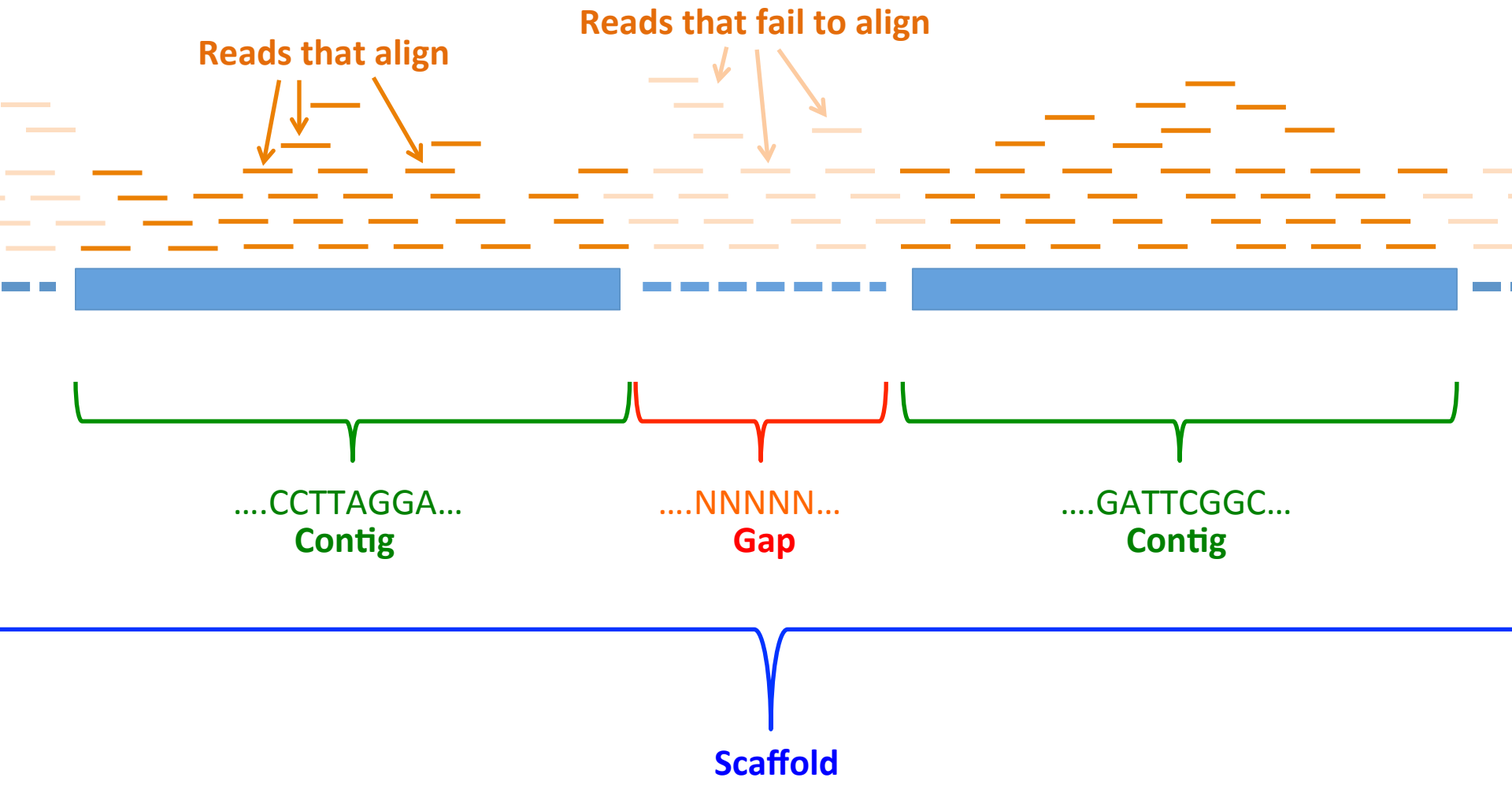
- Whole genome sequence
- Targeted capture
- RNAseq
- GBS/RAD

…and on the scale at which we are looking:

- Average across the genome
- Variance across the genome
- Identification of specific outlier regions

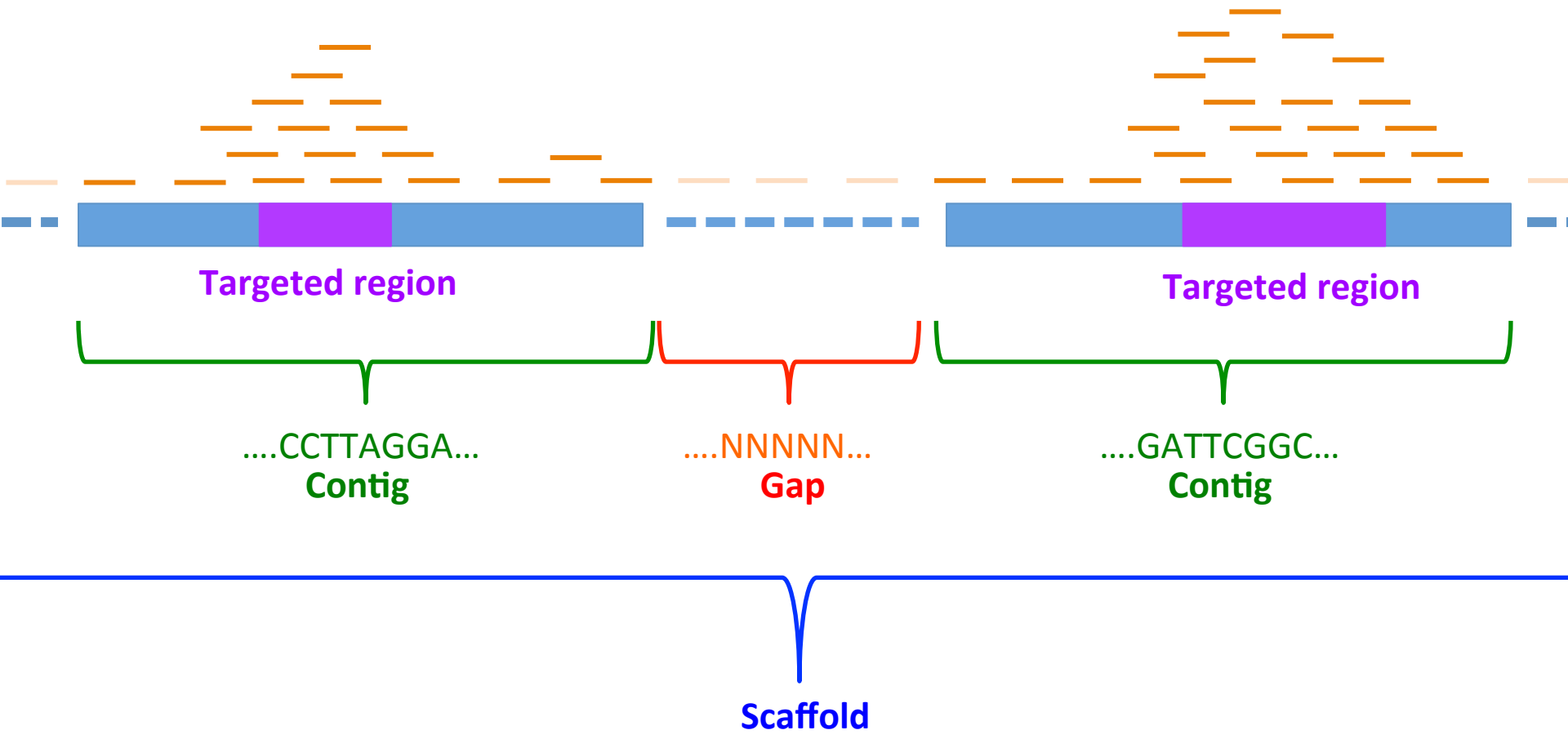
Distribution of sequence data: Whole genome sequence

(WGS) has relatively even coverage, but may miss regions of low complexity

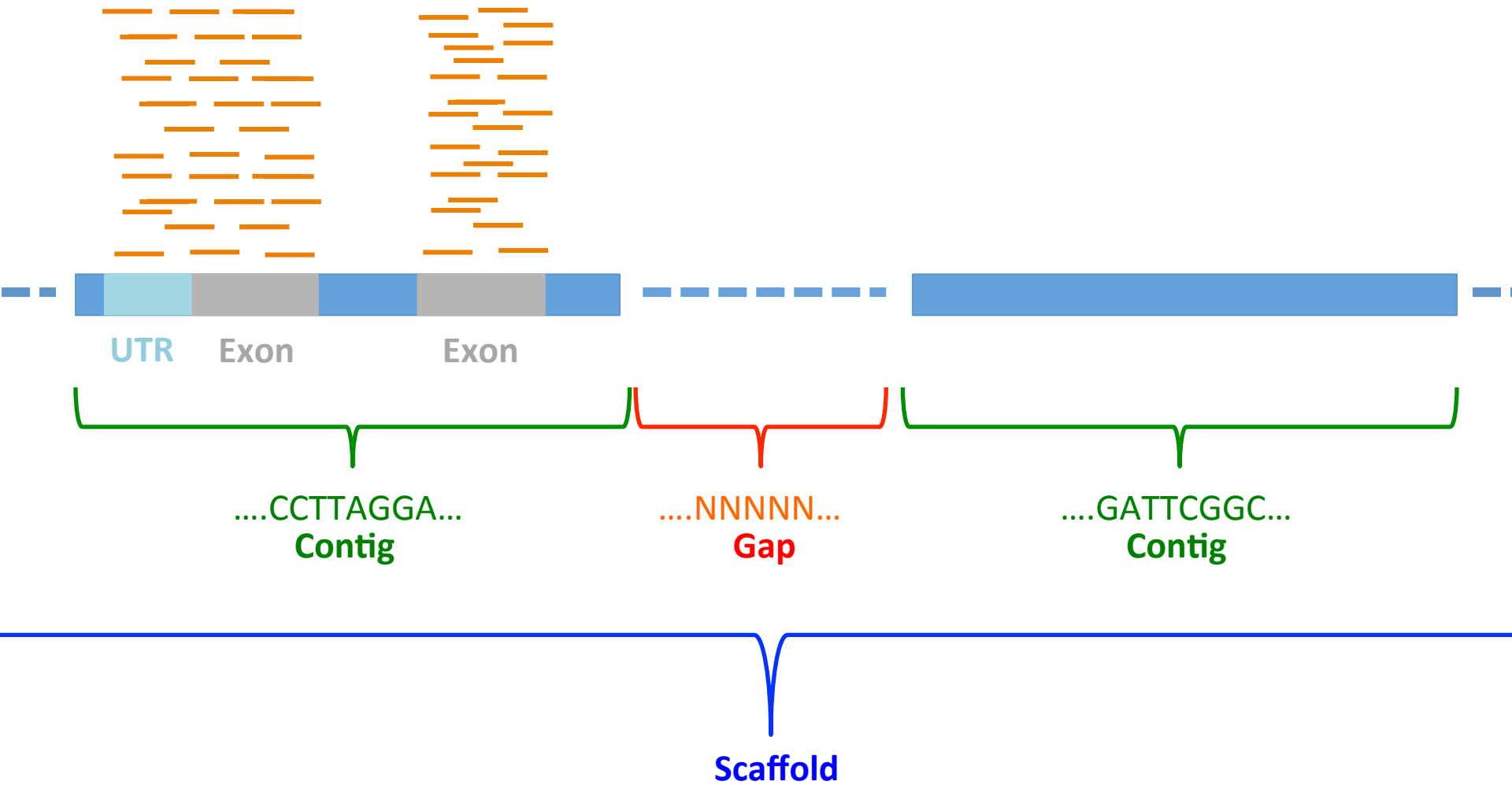


Distribution of sequence data: Targeted sequence capture

(exon capture, etc.) enriches certain regions but still gets some sequence from the rest of the genome

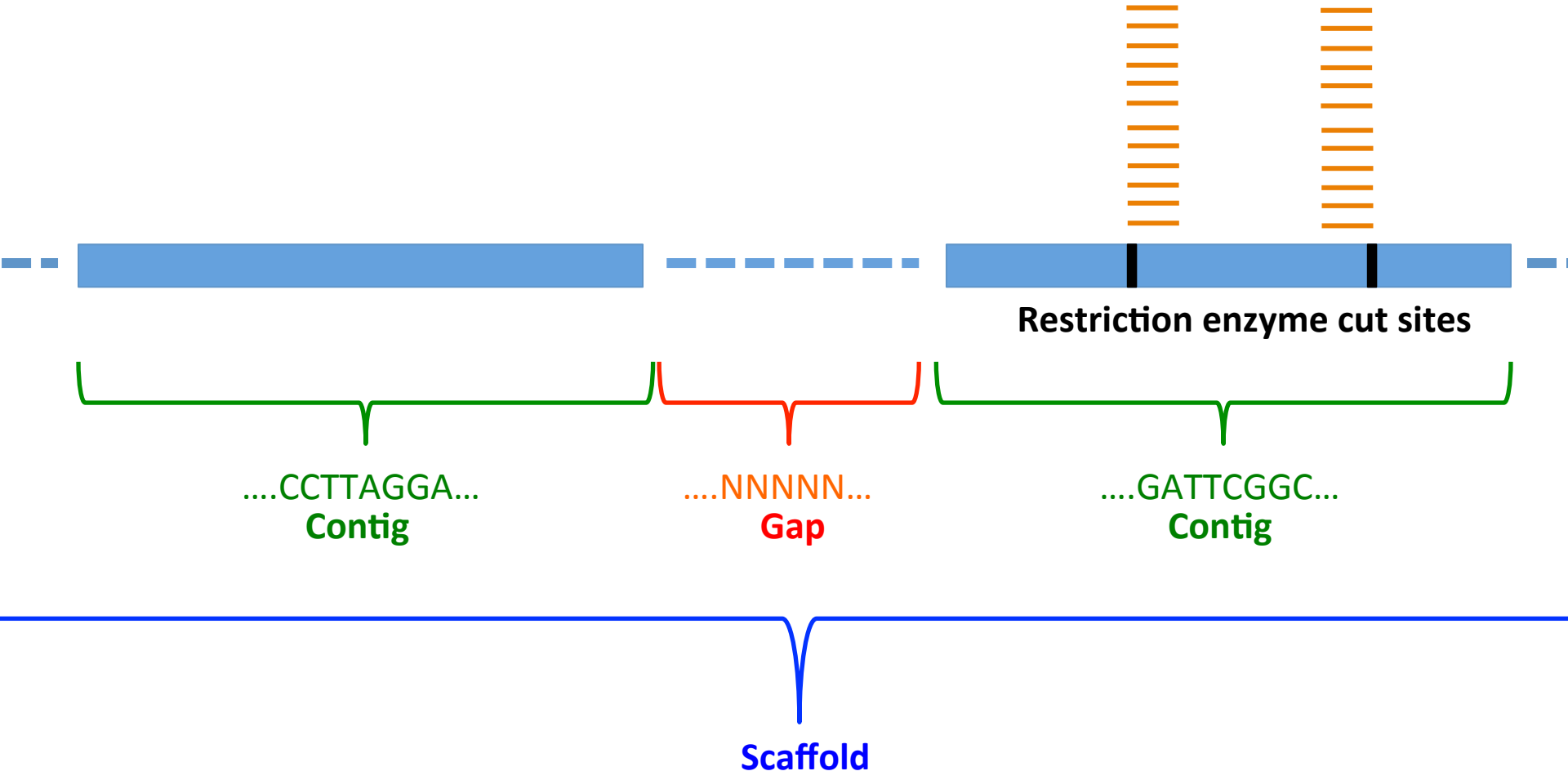


Distribution of sequence data: **RNAseq** gives very high coverage of genespace, no coverage outside of UTRs; some lowly expressed genes may not get sequenced



Distribution of sequence data: **GBS/RAD** yield small

(~100bp) segments adjacent to an enzyme cut site or a random shearing breakage; should be randomly distributed throughout the genome or near genes if methylation-sensitive

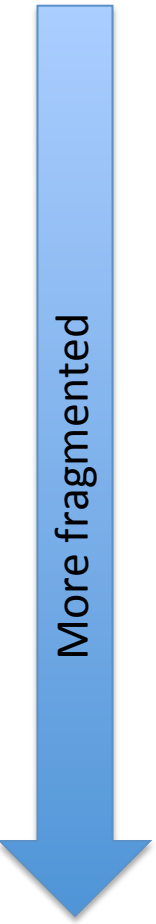


Overview: Structure of underlying sequence data

- **Whole genome sequence:** Coverage should be approximately equal over all sites, but low coverage in low complexity regions (poor sequencer performance)
- **Targeted capture:** only certain regions are targeted, but some background sequence at offtarget regions
- **RNAseq:** very high coverage of genespace, no coverage outside of UTRs; some lowly expressed genes may not get sequenced
- **GBS/RAD:** sequencing of small (~100bp) segments adjacent to an enzyme cut site or a random shearing breakage; should be randomly distributed throughout the genome or near genes if methylation-sensitive

Choice of sequencing method greatly affects the statistical approaches that you can use to assess patterns in the data

Overview: Reference genome assembly

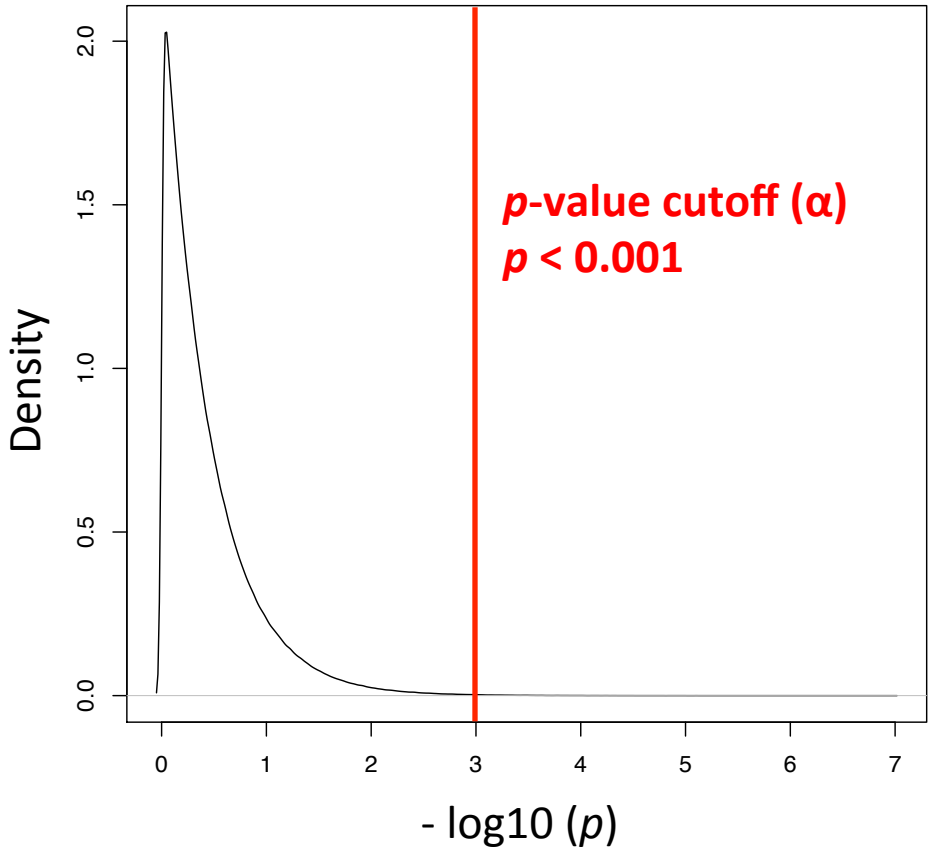


- **Archive-quality assembly:** Scaffolds are assembled into chromosomes or linkage groups
- **Scaffolded assembly:** smaller contigs are ordered into larger scaffolds which may span large regions of the genome (>1Mbp)
- **Fragmented assembly:** early-stage output from an assembler with no scaffolding of contigs. Most contigs are small (<100kbp)
- **GBS/RAD:** consensus "assembly" built using STACKS or similar approach

State of reference assembly affects how linkage information and sliding window analysis can be applied

Genome scans and outlier analyses

How to identify outliers? Most analytical approaches will generate a distribution of measurements (e.g. F_{ST}) or measures of significance (*p*-values or bayesfactors)



What now? More outliers than expected?

p-value cutoff = 0.001

Number of SNPs = 433,249

Expected # outliers: 432

Observed # outliers: 980

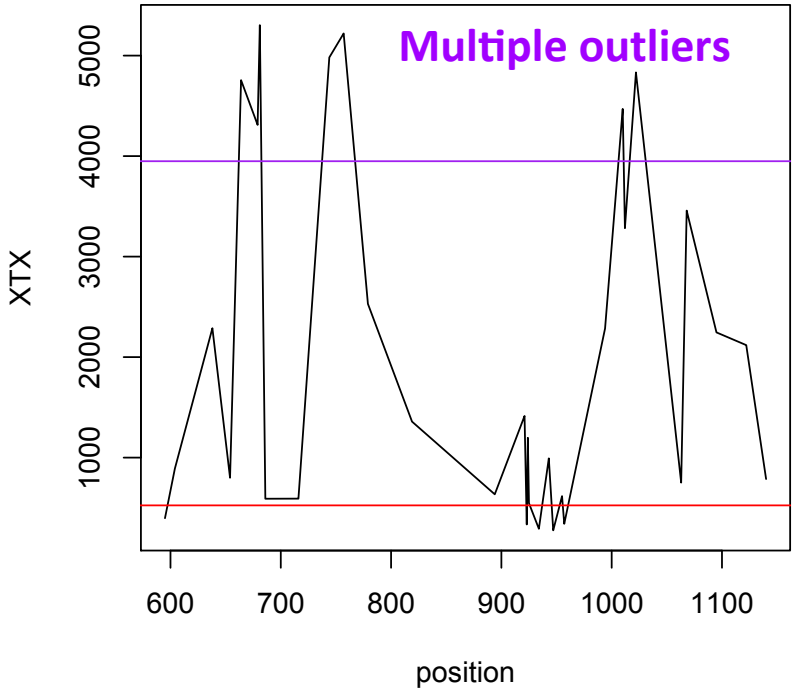
Alternatively: False discovery rate

For all tests with *q*-values < α, a proportion α will be false positives

Genome scans and outlier analyses

Problem #1 – Linkage: Does this mean that we actually have X loci that could be contributing to adaptation?

Caveat #1: Adjacent loci are not statistically independent



Solution (?): average over statistics across a distance greater than typical LD

More on this later…

Physically linked loci tend to be co-inherited and therefore statistically associated (i.e. **Linkage Disequilibrium**)

Genome scans and outlier analyses

Problem #2 – Relative significance: What can we do with measurements or bayesfactors, where we cannot evaluate the significance as we do with p-values?

P(α≠0)	Bayes Factor (BF)	log10(BF)	Jeffreys' interpretation
0.50 → 0.76	1 → 3	0 → 0.5	Barely worth mentioning
0.76 → 0.91	3 → 10	0.5 → 1	Substantial
0.91 → 0.97	10 → 32	1 → 1.5	Strong
0.97 → 0.99	32 → 100	1.5 → 2	Very strong
0.99 → 1.00	100 → ∞	2 → ∞	Decisive

From: BayeScan documentation

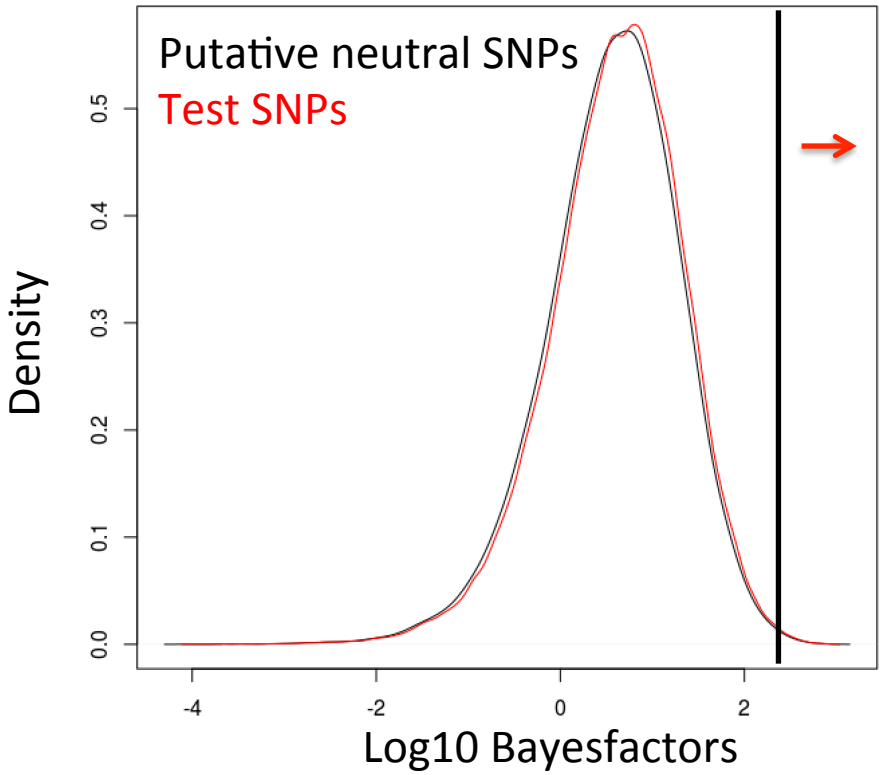
Genome scans and outlier analyses

Problem #2 – Relative significance: What can we do with measurements or bayesfactors, where we cannot evaluate the significance as we do with p-values?

One solution…

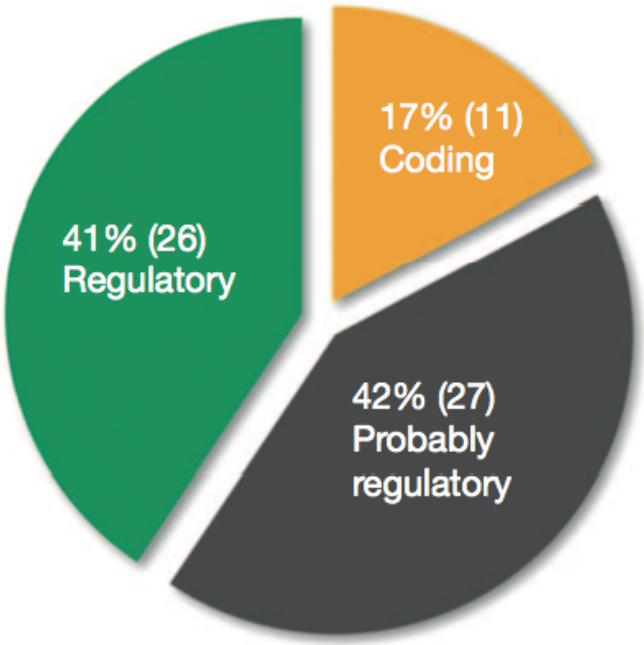
Calculate an empirical *p*-value based on distribution of loci presumed to be neutral (in non-coding regions)

Use this distribution as a set of "empirical p-values" and evaluate outliers in your test set that fall outside of this distribution



Genome scans and outlier analyses

Problem #2 – Relative significance: What can we do with measurements or bayesfactors, where we cannot evaluate the significance as we do with p-values?

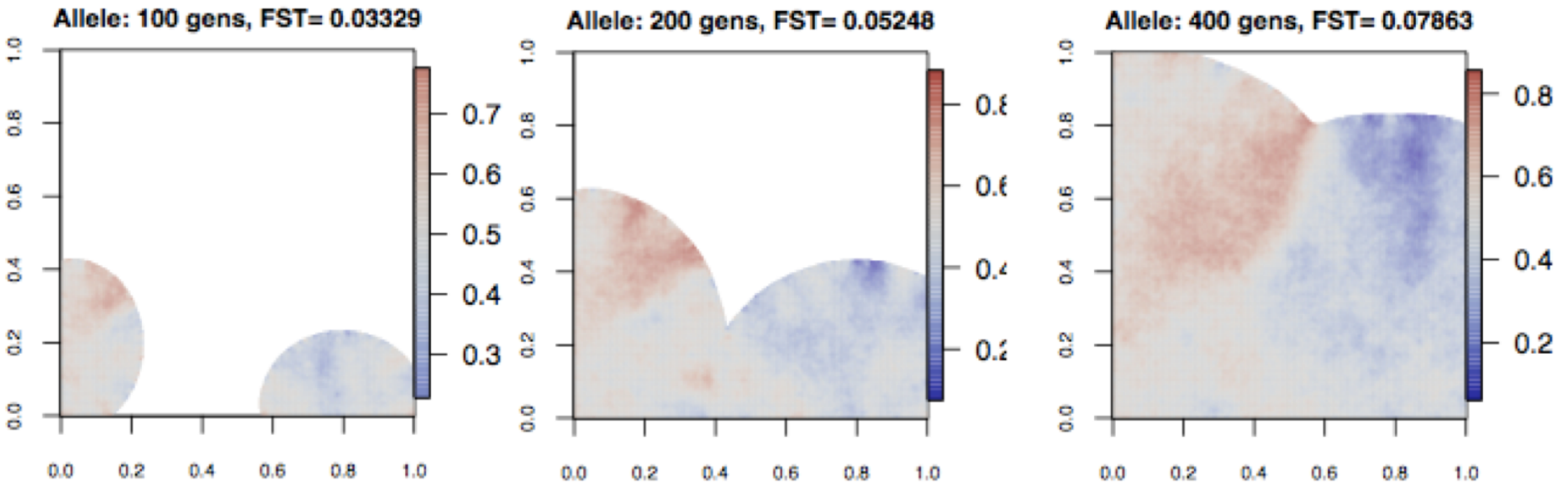


Outliers found in coding, regulatory and presumed regulatory regions in stickleback (Jones et al. 2012; Nature)

Caveat #2: Many studies have found strong signatures of selection in the non-coding gene space. For the *p*-value approach to work, the ratio of the **{putative neutral space}:{target space}** must be much larger than the ratio of true positives in each of these spaces

Genome scans and outlier analyses

Problem #3 – False Positives: Demography can yield genomic signatures that are similar to natural selection

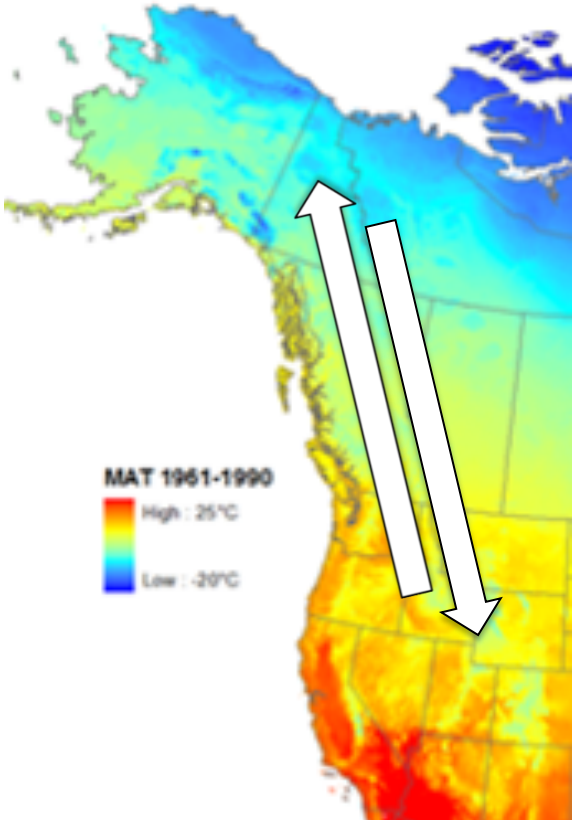


F_{ST} increases as a result of the expansion process

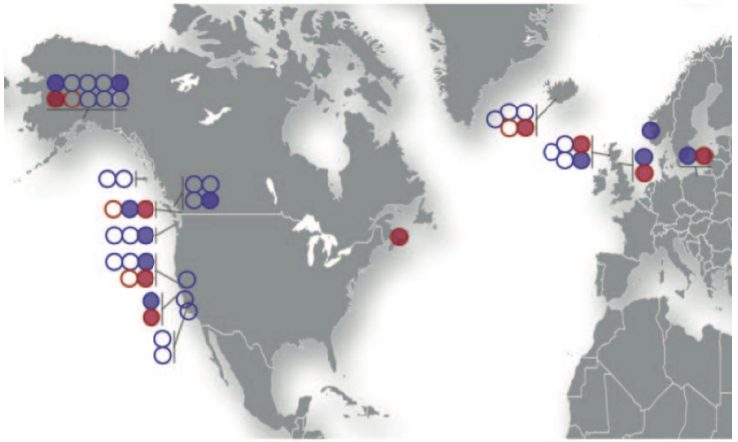
Many programs (FDIST, Bayescan, etc.) make assumptions about the underlying demography. If assumptions are violated, there may be false positives

Genome scans and outlier analyses

Problem #3 – False Positives: Demography can yield genomic signatures that are similar to natural selection



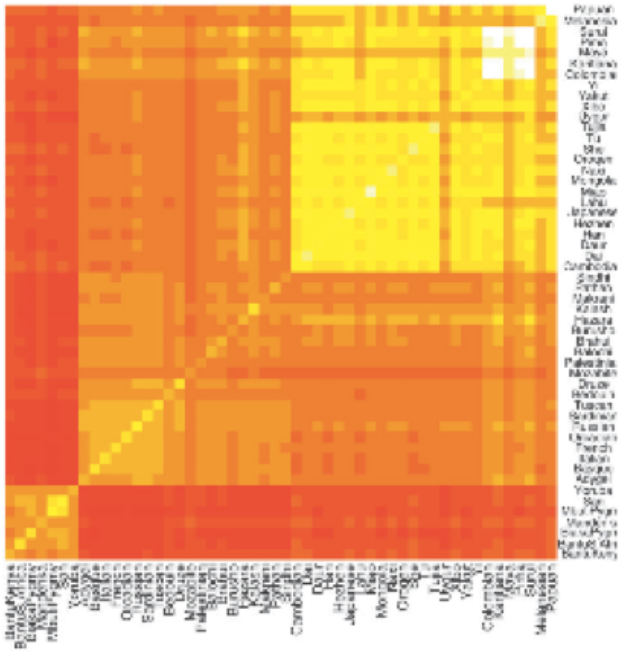
Isolation by distance can cause neutral loci to have similar structure to selected loci



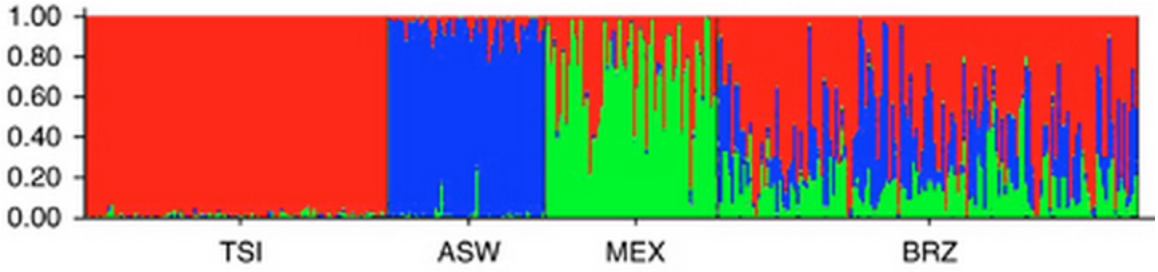
Multiple pairs of populations spanning different environments give greater power especially in "trans-gradient" direction

Genome scans and outlier analyses

Problem #3 – False Positives: Programs that use population structure estimated directly from the data (e.g. Bayenv, TASSEL, etc.) should be more resilient to these issues:



Genetic covariance matrix from Bayenv



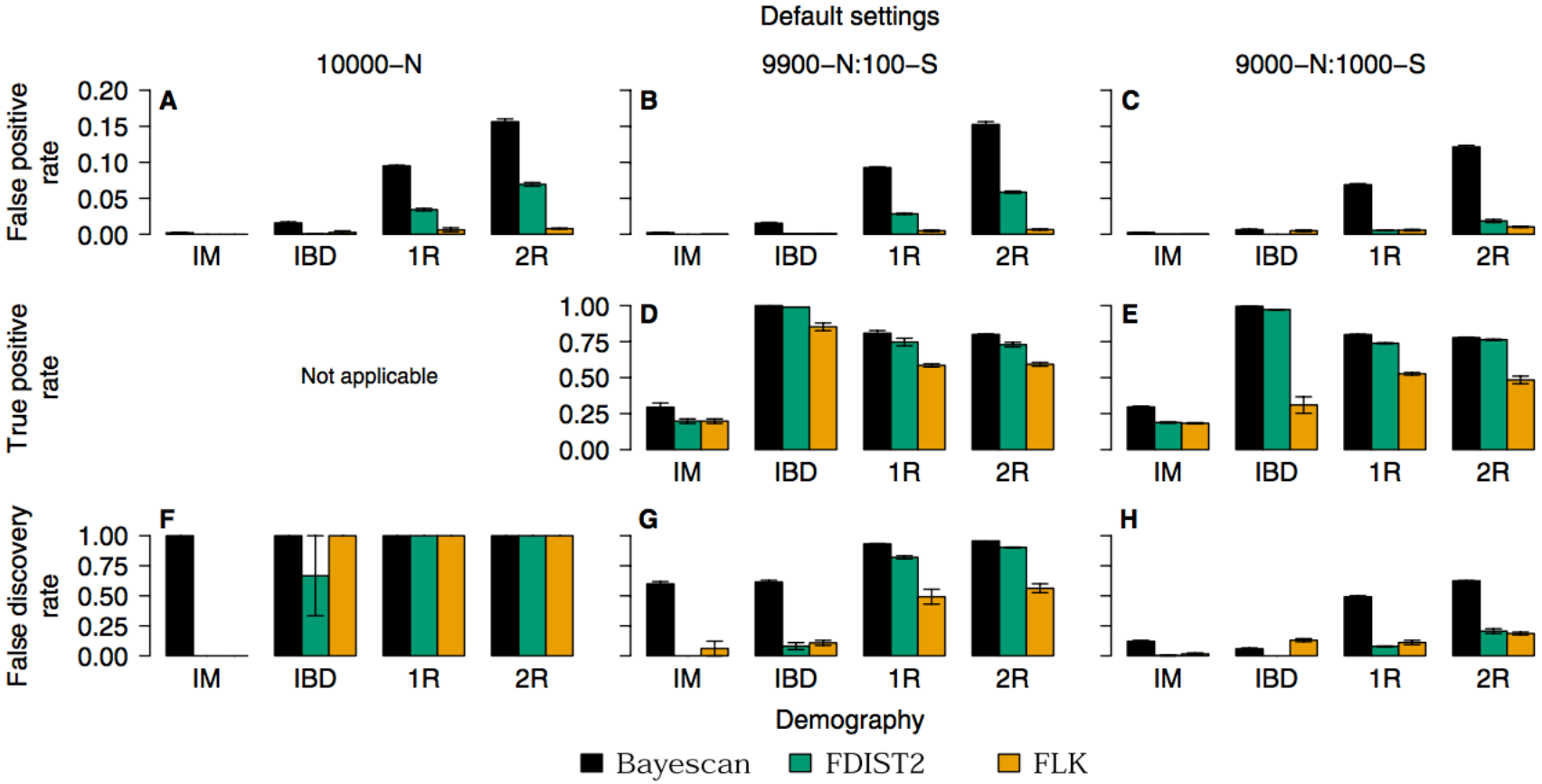
Population structure clusters from STRUCTURE (used as covariate)

Courtesy of K. Lotterhos

Giolo et al. (2012)

Genome scans and outlier analyses

Problem #3 – False Positives: Worse performance with complex patterns of population structure



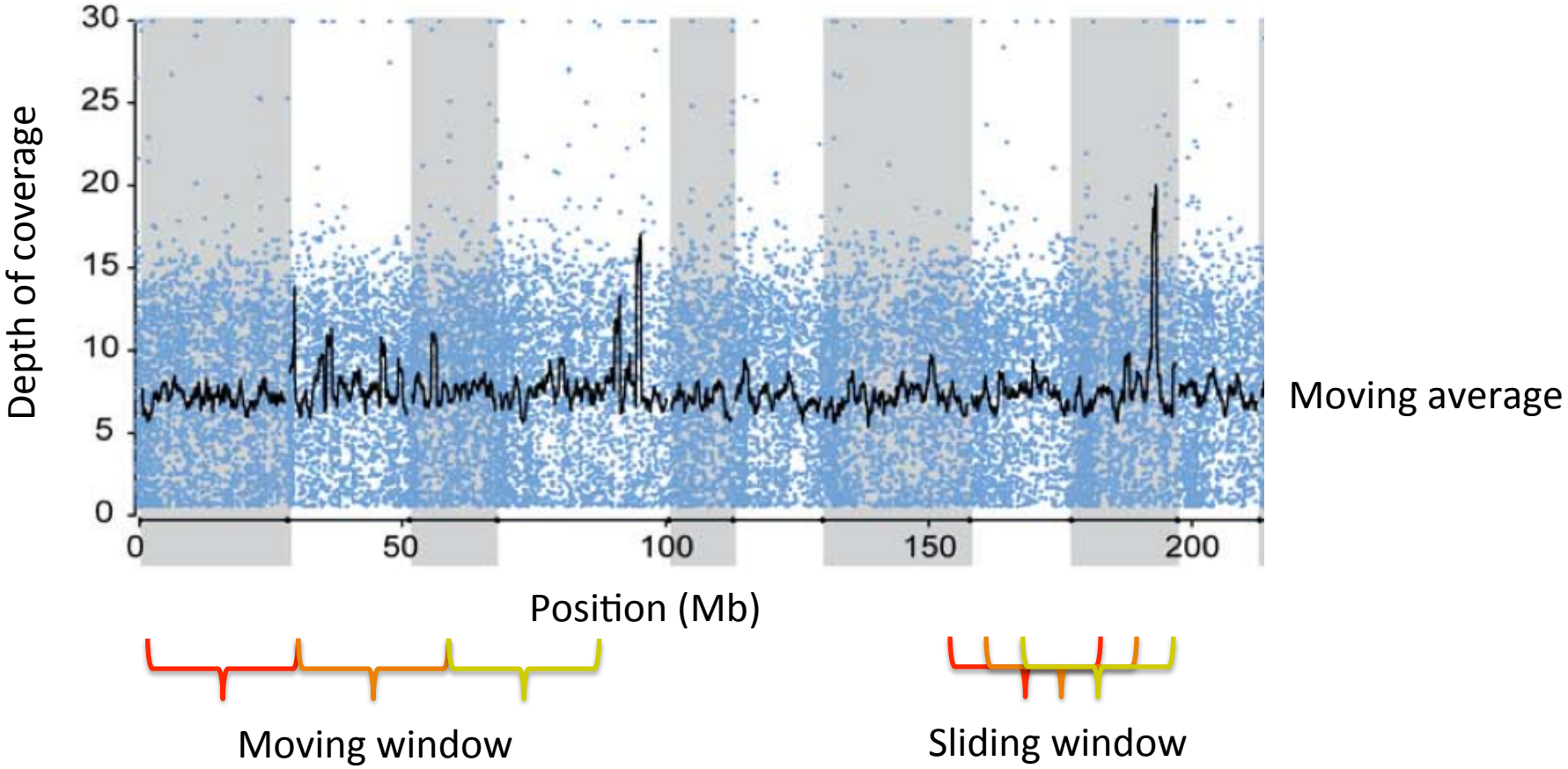
IM = Island model IBD = Isolation by Distance 1R = 1 Refugium 2R = 2 Refugia

False positives, non-independence of tests, estimating significance…

Many issues to consider…what are the most robust strategies?

Moving / Sliding window analysis

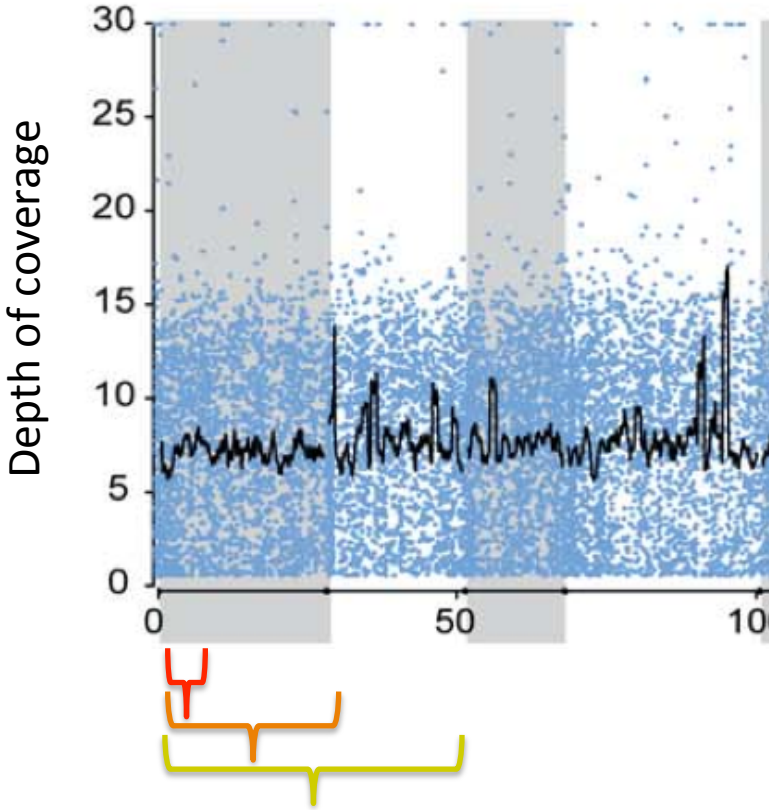
Reduce the problem of statistical non-independence by averaging across results within a given window of the chromosome



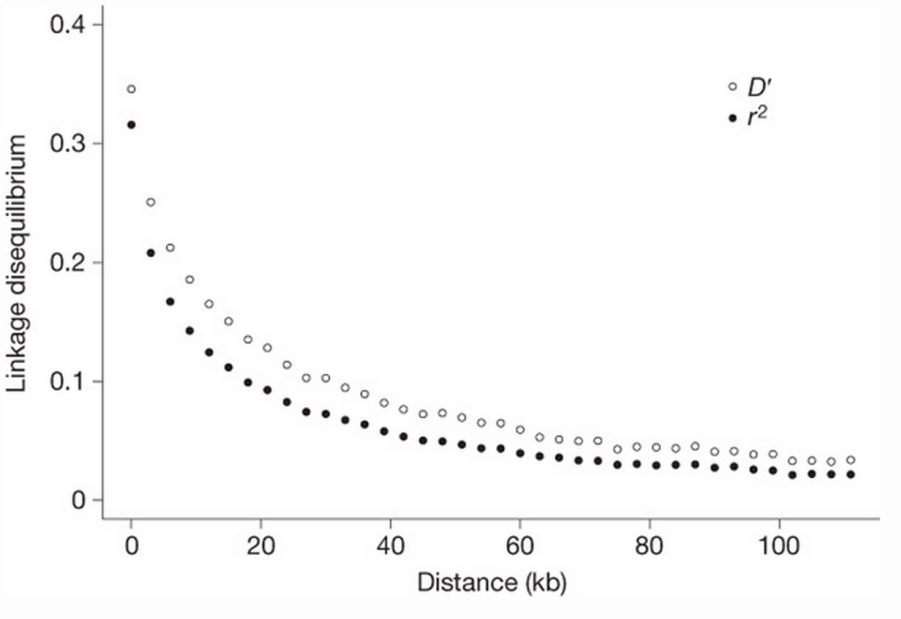
Hohenlohe et al. (2011; PLoS Genetics)

Moving / Sliding window analysis

How big should the window be? Can use information from analysis of linkage disequilibrium



Window size?



BUT: in cases of strong selection, you expect extreme LD and correlated response at many neutral loci across a wider range of the chromosome.

Rule of thumb: distance (cM) ~ s

Hohenlohe et al. (2011; PLoS Genetics)

Schacherer et al. (2009; Nature)

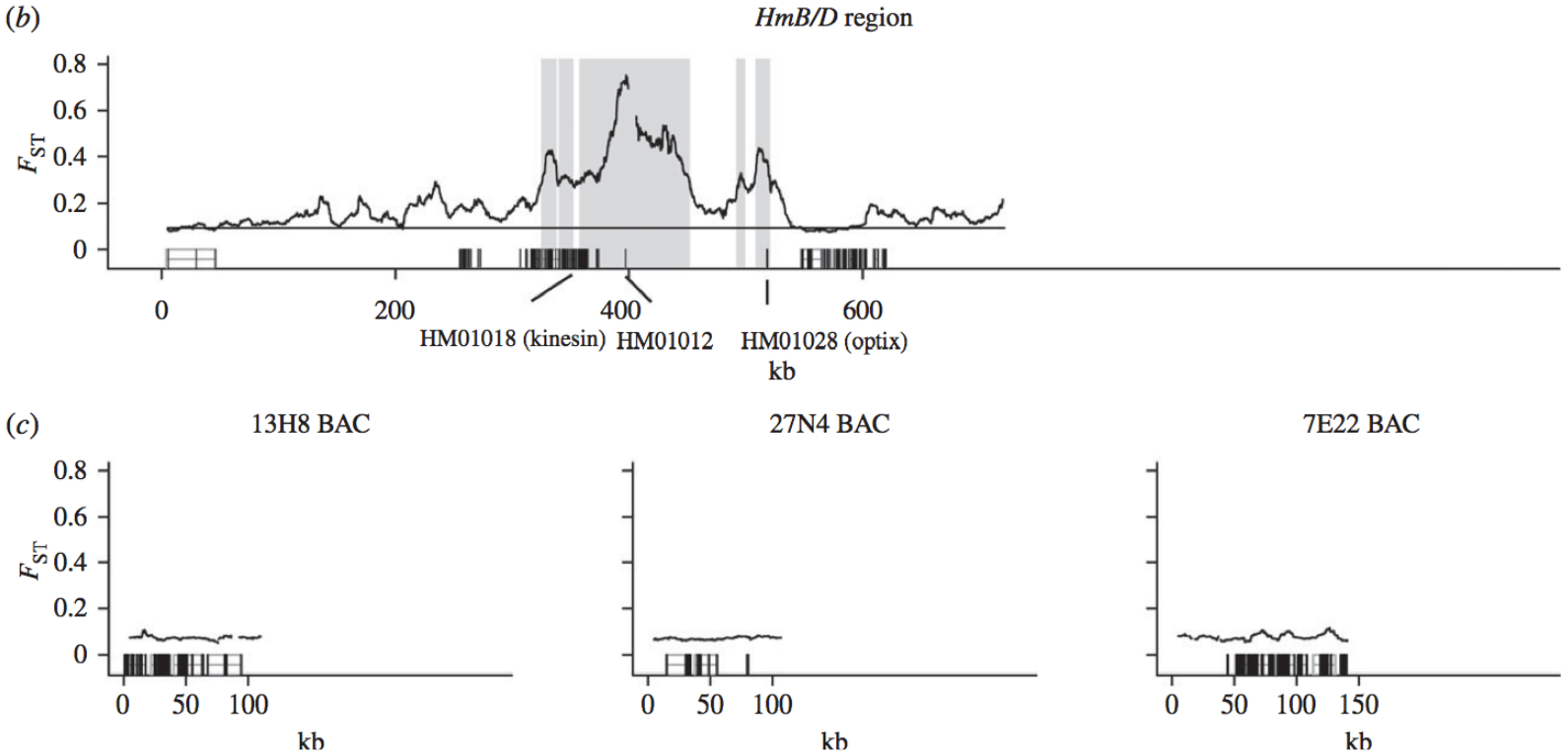
Moving / Sliding window analysis

Reduce the problem of statistical non-independence by summarizing results within a window of the chromosome

- Several approaches to calculate a summary statistic: mean, maximum, presence/absence of extreme outliers, # outliers
- Be mindful of the statistical properties of the quantity you are averaging (be careful with ratios, *p*-values, bayesfactors, etc.)
- Rare alleles may bias your scan, typically remove MAF < 0.05
- Windows may be measured in physical distance (e.g. 10kb), genetic distance (0.1 cM), or # of SNPs: each introduces a different kind of bias!
- Requires well-scaffolded reference genome or dense linkage map

Moving / Sliding window analysis

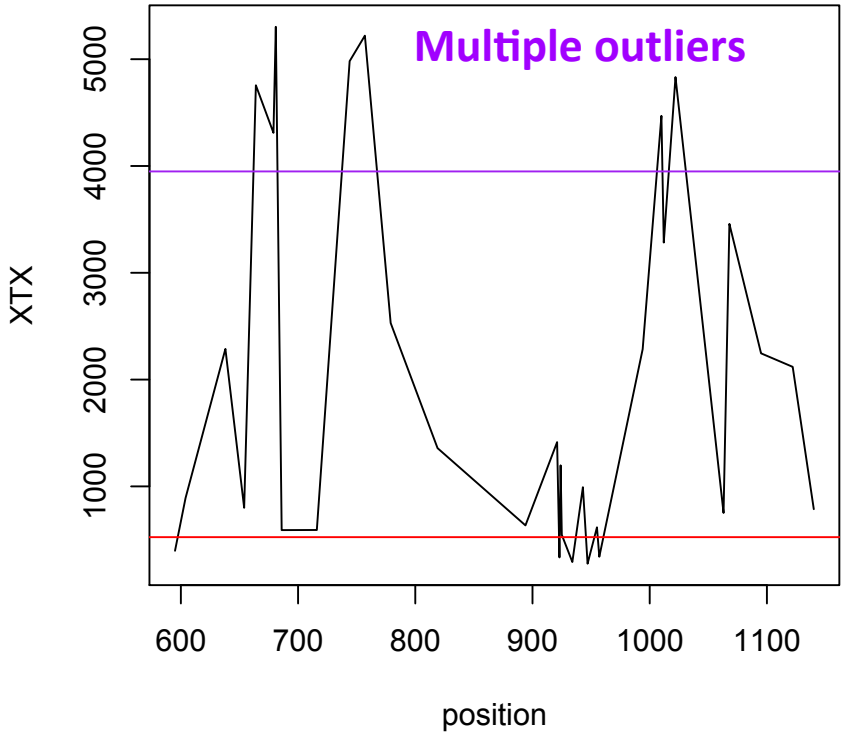
Alternative approach: Focus explicitly on trying to find signals of elevated linkage disequilibrium or large clusters of outliers ("genomic islands")

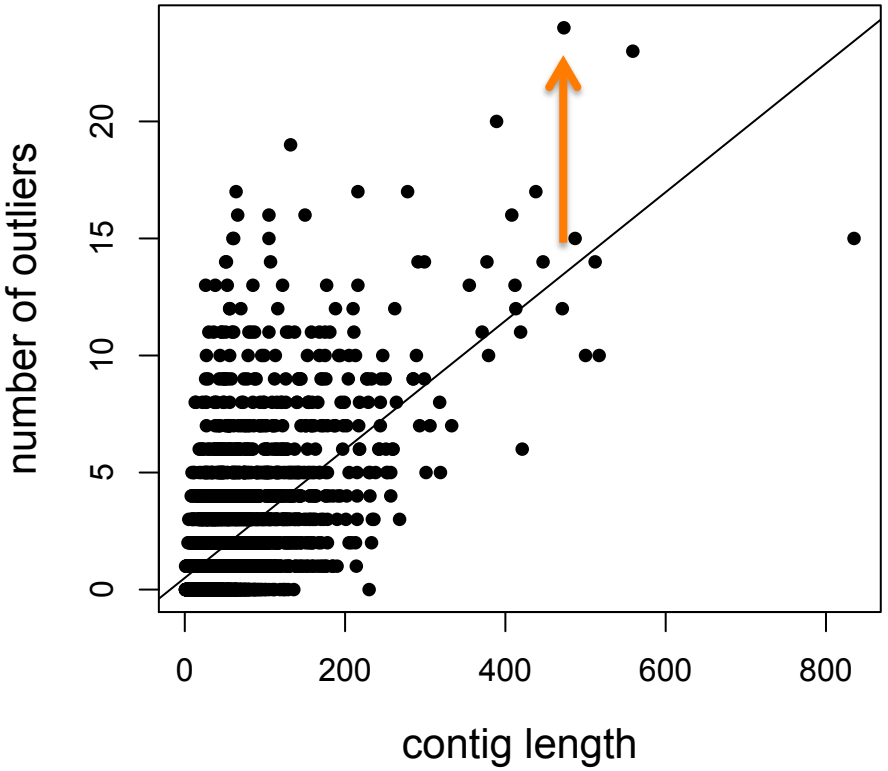


Extended window of multiple outliers is suggestive of selection (compare to 3 randomly chosen BACs below)

Genome scans with fragmented genomes

Small contigs may prevent implementation of typical sliding-window approaches. Can calculate contig-level statistics instead:

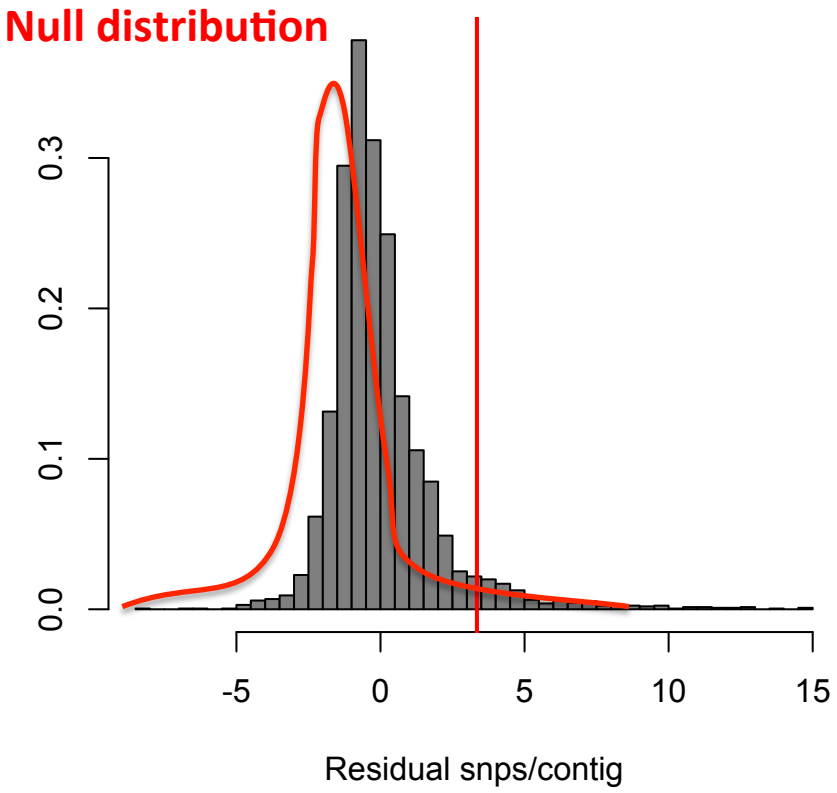




Use residuals from linear model as representation of the contig-level enrichment

Genome scans with fragmented genomes

An example of testing significance with null distribution

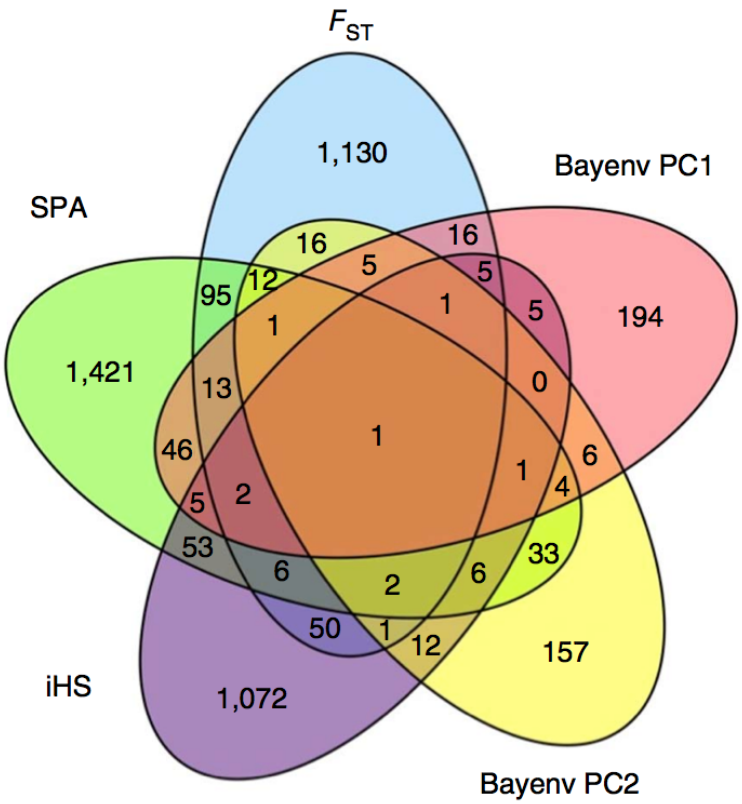


Analysis per gene or per contig:

- Randomly sample x SNPs from the overall set of SNPs, where x is the total number of outliers
- Calculate how many SNPs per contig across all contigs with hits
- Build null distribution by repeating 1000's of times
- Compare observed distribution to null distribution

Such approaches only test whether the data departs from the assumptions represented in the construction of the null distribution

Combining multiple types of genome scan

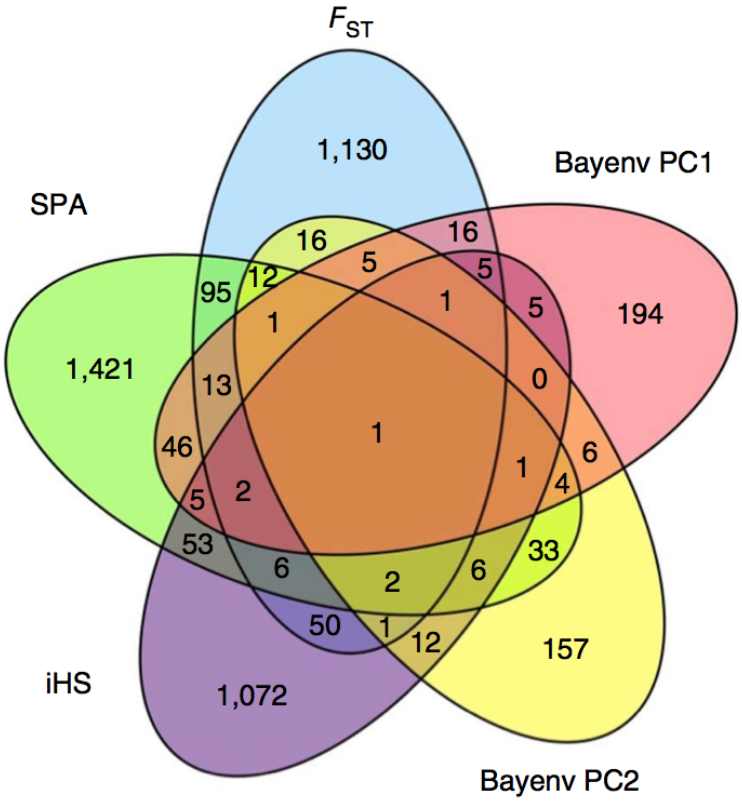


Number of 1-kb windows in the top 1% for each scan type

If SNPs are found by multiple tests, then they are more likely to be biologically important

How to test whether observed overlap is more than expected by chance?

Combining multiple types of genome scan



- Randomly sample x windows from the genome, repeating this for each of the 5 tests (with x = # of significant windows from each test)
- Calculate how times the same contigs are chosen
- Build null distribution by repeating 1000's of times
- Compare observed distribution to null distribution

This assumes that the 5 tests are independent, which is likely violated (this may be why this type of significance testing wasn't done in this paper)

Other factors to consider

- **Gene density:** background selection may be stronger near genes (reducing background heterozygosity)
- **Recombination hotspots/coldspots**: diversity tends to be reduced in recombination coldspots, due to background selection
- **Mutation hotspots/coldspots:** also affect overall diversity
- **Sequencing coverage:** affects the ability to call genotypes and should be considered when evaluating #SNPs or #outliers/region
- **Technical sources of bias:** diverse, difficult to predict, and platform specific

Practical implementation

- "Packaged" methods are best for the per-locus test statistics (Bayenv, F_{ST}, etc.) where coding your own would be inefficient and more prone to bugs
- Some packaged methods exist for calculating sliding windows and enrichment analyses, but you will have more power to do customize this to your data with your own code in R
- Using large computing clusters may be necessary, especially for MCMC based methods applied to SNPs
- It is nearly impossible to keep up with the growth of new analysis tools unless that is your primary interest…but we should still try!

The future…

- Critical need for simulating and testing how these programs work in complex demographic scenarios and highly polygenic traits
- Haplotype-level analysis will be common once long-read technology improves. There is more information in phased data, but phasing is hard, especially where selection is operating
- Experimental approaches and well-planned sampling schemes can go a long way
- Field is evolving very rapidly!

For the exercise:

- wget http://www.zoology.ubc.ca/~yeaman/bioinformatics/genomescan.tbz
- Attempt to load "`snps.seqcap.vcf.reduce`" into R, as the exercise says
- If you can't get it to load, use "`snps.gbs.vcf`" instead. It should work mostly the same, but you will have to modify the code where I've used certain contig names