SNP calling from RNAseq data

Steve Palumbi Lab

Hopkins Marine Station, Stanford University

Analysis of sequence information from RNA-Seq data. (workshop) SNP-calling and overview of methods to analyze the output. **Steve Palumbi lab at Stanford/Hopkins.**

Part 1: 35 minutes

Overview: from Bam files to genotypes: Steve Palumbi [3 min]

Bowtie2 and SamTools: Bryan Barney and Nathan Truelove [10]

FreeBayes: Noah Rose and Elora López [10]

vcfTools and the 0,1,2 genotype file: Beth Sheets and Megan Morikawa [3]

PCA and FST: Megan Morikawa and Bryan Barney [5]

Part 2: 30 minutes

Mentored file manipulation workshop from fastQ files to 0,1,2 genotype matrix using demo input files

Part 3: 30 minutes

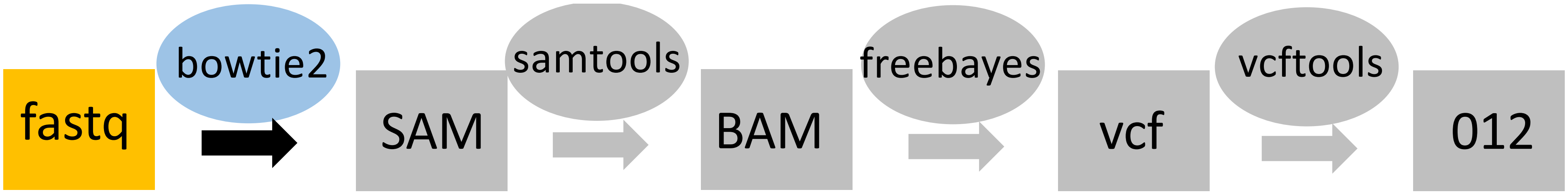
Overview: using genotype data: Steve Palumbi [1 min]

NgsAdmix and linkage: Bryan Barney [7]

Outliers and environmental correlations: Luke Thomas and Nathan Truelove [5]

Somatic mutations: Elora López [5]

dN/dS and eQTLs: Noah Rose [9]



Mapping in Bowtie2

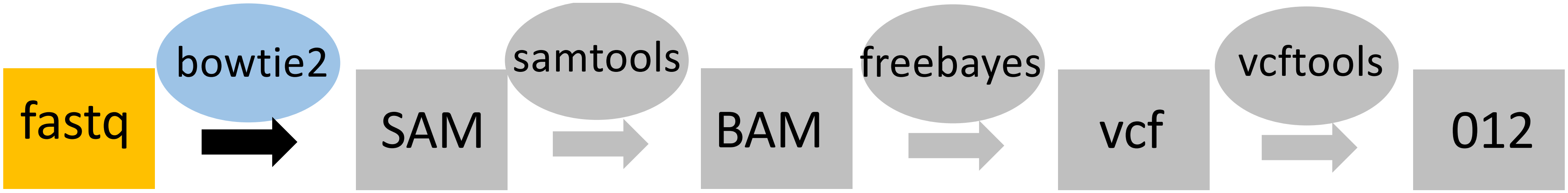
Bowtie2 lines up reads to a reference genome or transcriptome

End-to-End: Uses all the base-pairs

```
Read:      G A C T G G G C G A T C T C G A C T T C G
           | | | | |     | | | | | | | | | | | | | |
Reference: G A C T G - - C G A T C T C G A C A T C G
```

Local: Base-pairs at the ends can be discarded

```
Read:      A C G G T T G C G T T A A – T C C G C C A C G
                   | | | | | | | | |   | | | | | |
Reference: T A A C T T G C G T T A A A T C C G C C T G G
```



fastq → bowtie2 → SAM → samtools → BAM → freebayes → vcf → vcftools → 012

Alignment Score

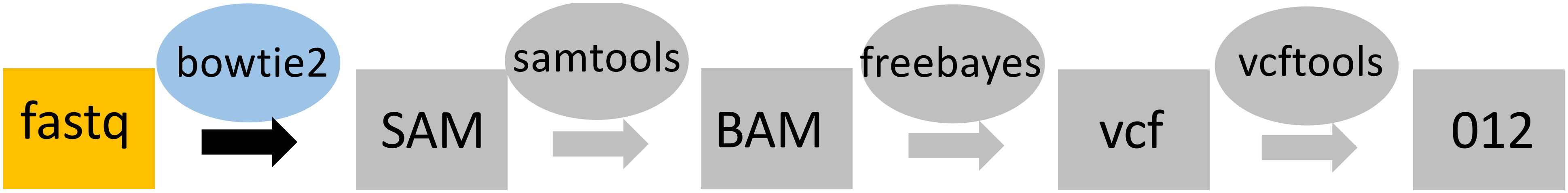
How similar the read is to the reference

End-to-End Example:

Mismatch = -6
Read Gap = -11
Best alignment score = 0

Local Example:

Mismatch = -6
Read Gap = -11
Base that matches Reference = +2
Best Alignment Score = 2 x Read length


fastq bowtie2 SAM samtools BAM freebayes vcf vcftools 012

Minimum Alignment Score

- Expressed as a Function of Read Length:

  f(x) = 0 + -0.6 * x, where x is the read length

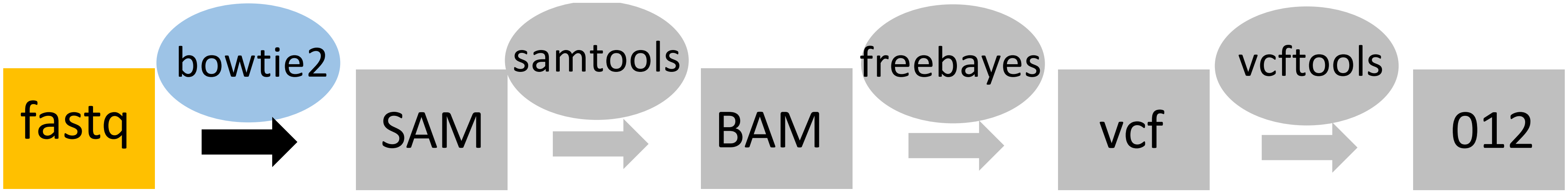
- End-to-end alignment mode default is:

  -0.6 + -0.6 * read length

- For a 50 base-pair read:

  -0.6 + -30 = -30.6

- Default: 5 mismatches/2 read gaps/Combos



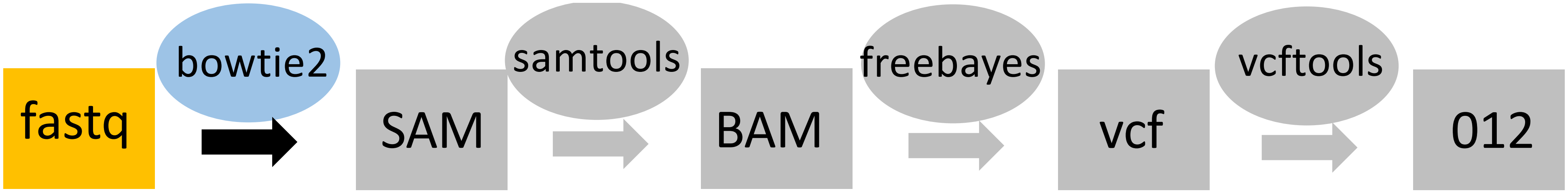
Optimize Mapping Parameters

End-to-End Example:

Mismatch = -6

Read Gap = -11

Best alignment score = 0



Optimize Mapping Parameters

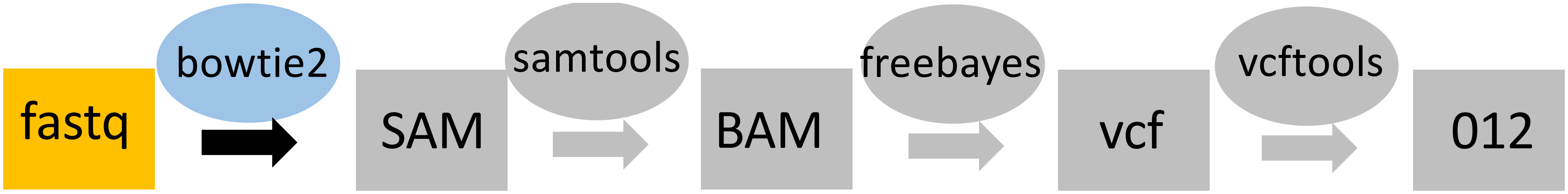
--score min

- Changes the default minimum alignment score to be considered valid.

Default: L,0,-0.6 = -30.6 for 50 bp reads

- Optimized for 2 mismatches/1 mismatch and 1 read gap

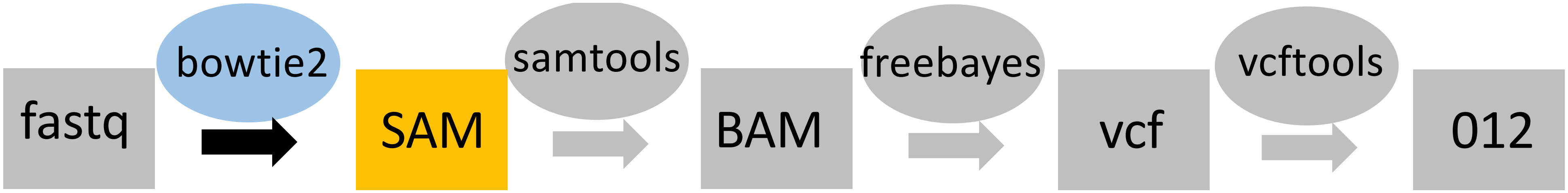
--score-min L,0,-0.36 = -18


fastq
bowtie2
SAM
samtools
BAM
freebayes
vcf
vcftools
012

Preset Mapping Parameters

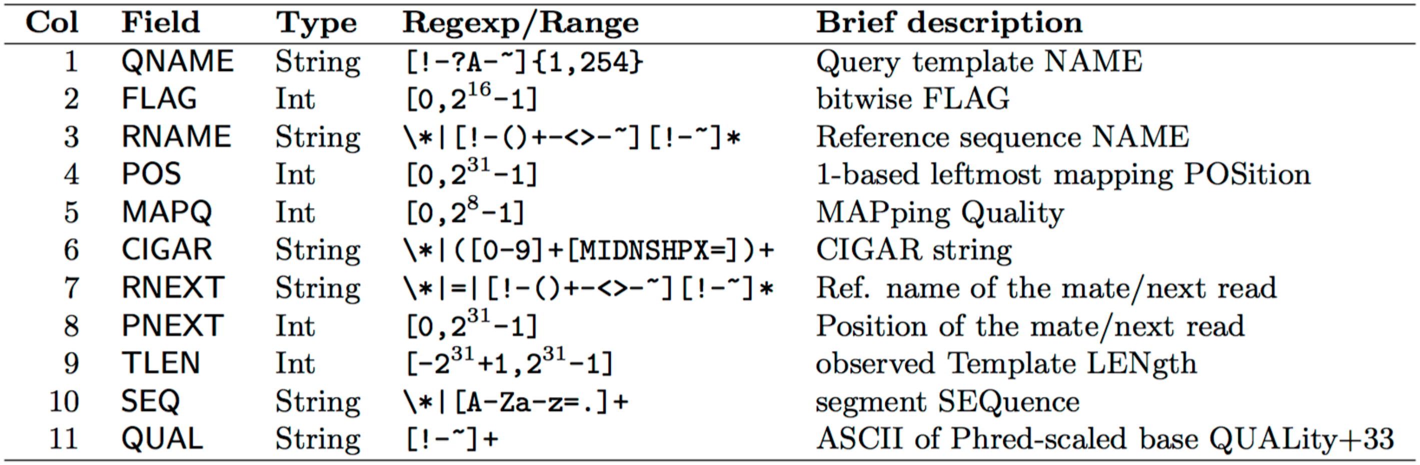
Verify that the preset meets your mapping requirements

--very-fast
--fast
--sensitive
--very-sensitive



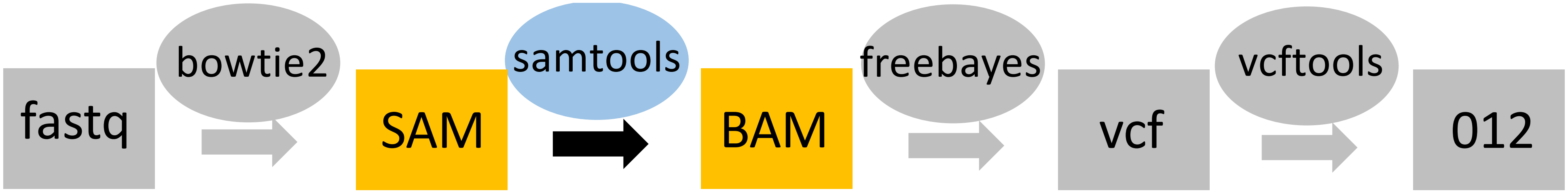
Mapping : Bowtie outputs a SAM file

SAM files contain a list of reads, each read will get a series of ‘fields’ associated with it that describe the mapping result



Col	Field	Type	Regexp/Range	Brief description
1	QNAME	String	[!-?A-~]{1,254}	Query template NAME
2	FLAG	Int	$[0,2^{16}-1]$	bitwise FLAG
3	RNAME	String	*|[!-()+-<>-~][!-~]*	Reference sequence NAME
4	POS	Int	$[0,2^{31}-1]$	1-based leftmost mapping POSition
5	MAPQ	Int	$[0,2^{8}-1]$	MAPping Quality
6	CIGAR	String	*|([0-9]+[MIDNSHPX=])+	CIGAR string
7	RNEXT	String	*|=|[!-()+-<>-~][!-~]*	Ref. name of the mate/next read
8	PNEXT	Int	$[0,2^{31}-1]$	Position of the mate/next read
9	TLEN	Int	$[-2^{31}+1,2^{31}-1]$	observed Template LENgth
10	SEQ	String	*|[A-Za-z=.]+	segment SEQuence
11	QUAL	String	[!-~]+	ASCII of Phred-scaled base QUALity+33

Different mapping utilities (bowtie2, BWA, etc) will have different ‘additional fields’ that you might use for filtering



Mapping : Samtools converts SAM to BAM, sorts, & indexes

- SAM files are human readable plain text
- BAM files are binary versions of SAM that are smaller and easier for the computer to process

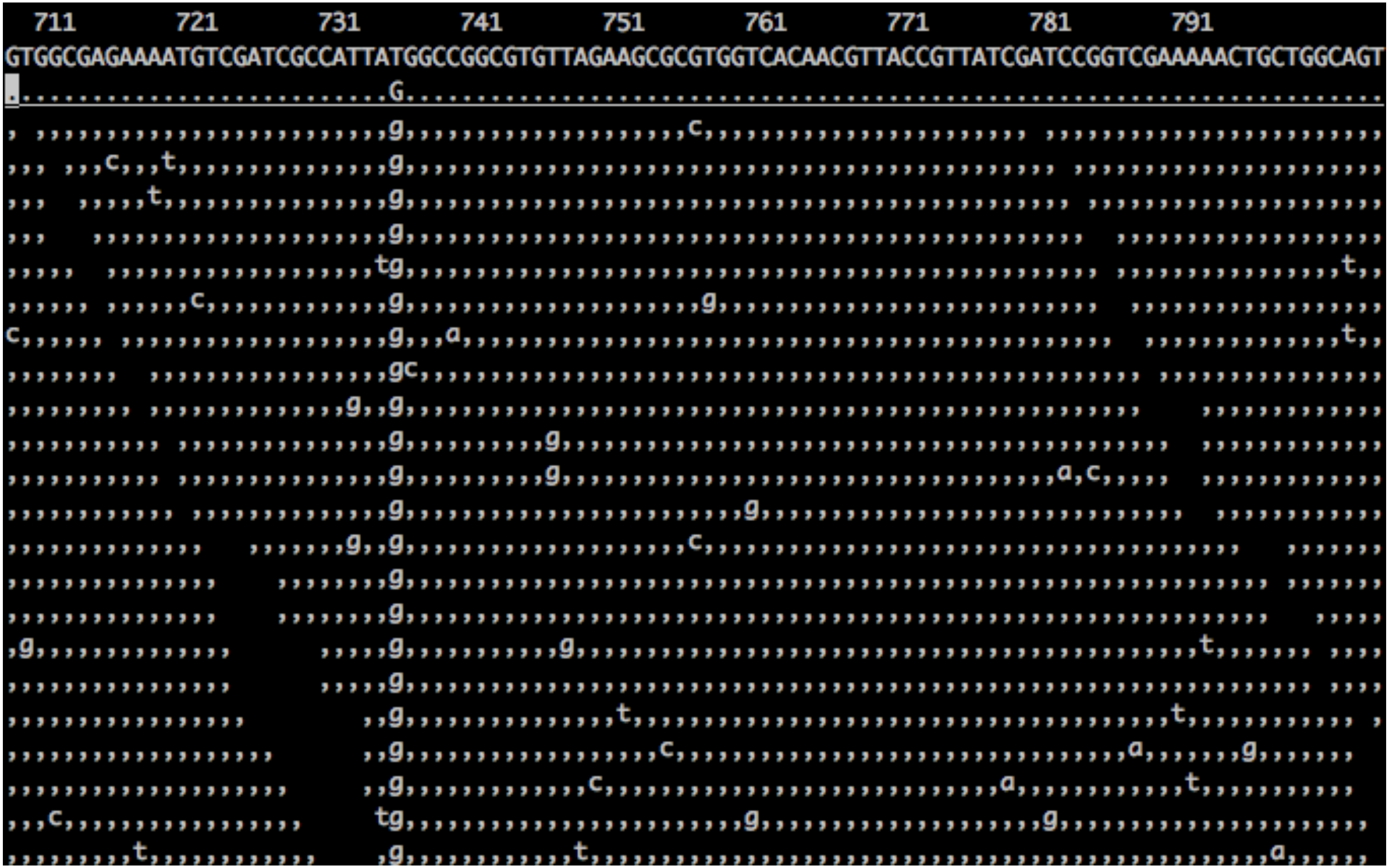
Sorting:

- Groups your reads by where in the assembly they mapped

Indexing:

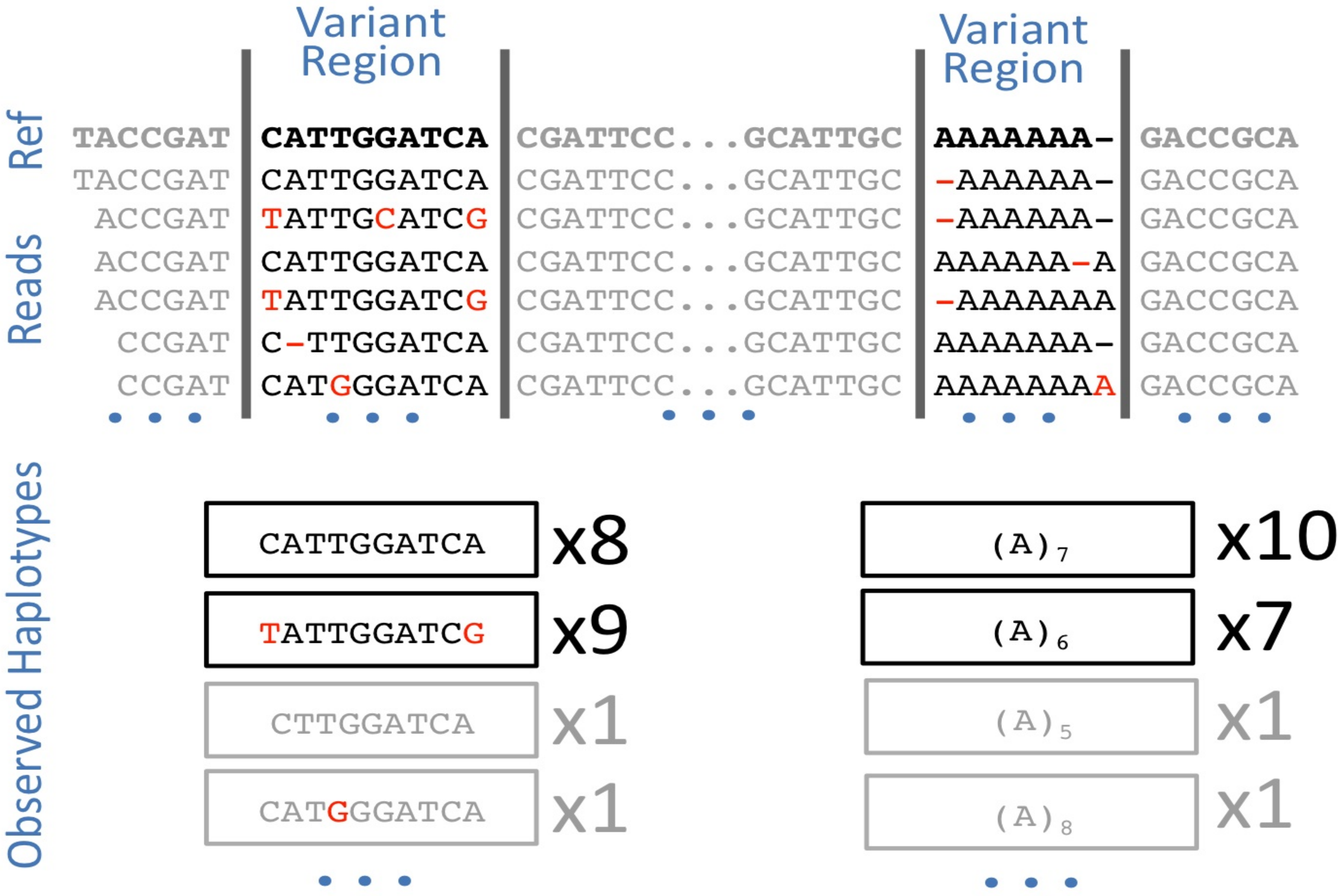
- Makes a table that contains information about:
  - How many reads mapped
  - Where they mapped
  - Reads that didn’t map

SNP Calling



```
  711       721       731       741       751       761       771       781       791
GTGGCGAGAAAATGTCGATCGCCATTATGGCCGGCGTGTTAGAAGCGCGTGGTCACAACGTTACCGTTATCGATCCGGTCGAAAAACTGCTGGCAGT
...........................G.....................................................................
, ,,,,,,,,,,,,,,,,,,,,,,,,,g,,,,,,,,,,,,,,,,,,,,c,,,,,,,,,,,,,,,,,,,,,,, ,,,,,,,,,,,,,,,,,,,,,,,,
,,, ,,,c,,,t,,,,,,,,,,,,,,,g,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,, ,,,,,,,,,,,,,,,,,,,,,,
,,,  ,,,,,t,,,,,,,,,,,,,,,,g,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,, ,,,,,,,,,,,,,,,,,,,,,
,,,   ,,,,,,,,,,,,,,,,,,,,,g,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,  ,,,,,,,,,,,,,,,,,,,
,,,,,  ,,,,,,,,,,,,,,,,,,,tg,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,, ,,,,,,,,,,,,,,,,t,,
,,,,,, ,,,,,,c,,,,,,,,,,,,,g,,,,,,,,,,,,,,,,,,,,,g,,,,,,,,,,,,,,,,,,,,,,,,,,,  ,,,,,,,,,,,,,,,,,,
c,,,,,, ,,,,,,,,,,,,,,,,,,,g,,,a,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,  ,,,,,,,,,,,,,,t,,
,,,,,,,,  ,,,,,,,,,,,,,,,,,gc,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,, ,,,,,,,,,,,,,,,,
,,,,,,,,, ,,,,,,,,,,,,,,g,,g,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,    ,,,,,,,,,,,,,
,,,,,,,,,,, ,,,,,,,,,,,,,,,g,,,,,,,,,,g,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,  ,,,,,,,,,,,,,
,,,,,,,,,,, ,,,,,,,,,,,,,,,g,,,,,,,,,,g,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,a,c,,,,,  ,,,,,,,,,,,,,
,,,,,,,,,,,, ,,,,,,,,,,,,,,g,,,,,,,,,,,,,,,,,,,,,,,,g,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,  ,,,,,,,,,,,,
,,,,,,,,,,,,,,   ,,,,,,,g,,g,,,,,,,,,,,,,,,,,,,,c,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,   ,,,,,,,
,,,,,,,,,,,,,,,    ,,,,,,,,g,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,, ,,,,,,,
,,,,,,,,,,,,,,,    ,,,,,,,,g,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,   ,,,,,
,g,,,,,,,,,,,,,,      ,,,,,g,,,,,,,,,,,g,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,t,,,,,,, ,,,,
,,,,,,,,,,,,,,,,      ,,,,,g,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,, ,,,,
,,,,,,,,,,,,,,,,,        ,,g,,,,,,,,,,,,,,,t,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,t,,,,,,,,,,,, ,
,,,,,,,,,,,,,,,,,,,      ,,g,,,,,,,,,,,,,,,,,,c,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,a,,,,,,,g,,,,,,,
,,,,,,,,,,,,,,,,,,,,     ,,g,,,,,,,,,,,,,c,,,,,,,,,,,,,,,,,,,,,,,,,,,,a,,,,,,,,,,,,t,,,,,,,,,,,
,,,c,,,,,,,,,,,,,,,,,     tg,,,,,,,,,,,,,,,,,,,,,,,,g,,,,,,,,,,,,,,,,,,,,g,,,,,,,,,,,,,,,,,,,,,,
,,,,,,,,,t,,,,,,,,,,,,    ,g,,,,,,,,,,,,t,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,a,,,,,,
```

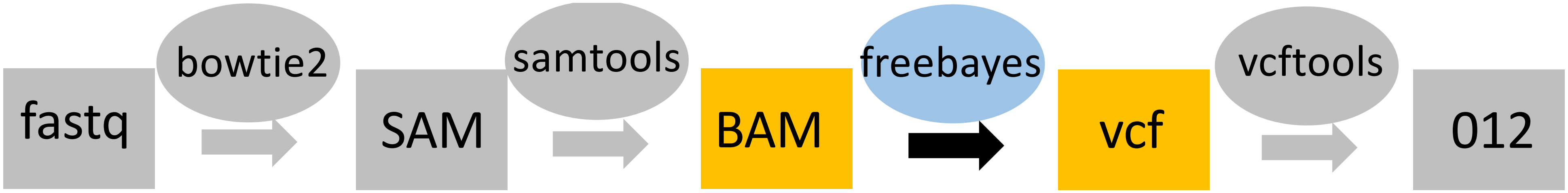
Haplotype-based approach



		Variant Region		Variant Region	
Ref	TACCGAT	CATTGGATCA	CGATTCC...GCATTGC	AAAAAAA–	GACCGCA
Reads	TACCGAT	CATTGGATCA	CGATTCC...GCATTGC	–AAAAAA–	GACCGCA
	ACCGAT	TATTGCATCG	CGATTCC...GCATTGC	–AAAAAA–	GACCGCA
	ACCGAT	CATTGGATCA	CGATTCC...GCATTGC	AAAAAA–A	GACCGCA
	ACCGAT	TATTGGATCG	CGATTCC...GCATTGC	–AAAAAAA	GACCGCA
	CCGAT	C–TTGGATCA	CGATTCC...GCATTGC	AAAAAAA–	GACCGCA
	CCGAT	CATGGGATCA	CGATTCC...GCATTGC	AAAAAAAA	GACCGCA
	• • •	• • •	• • •	• • •	• • •

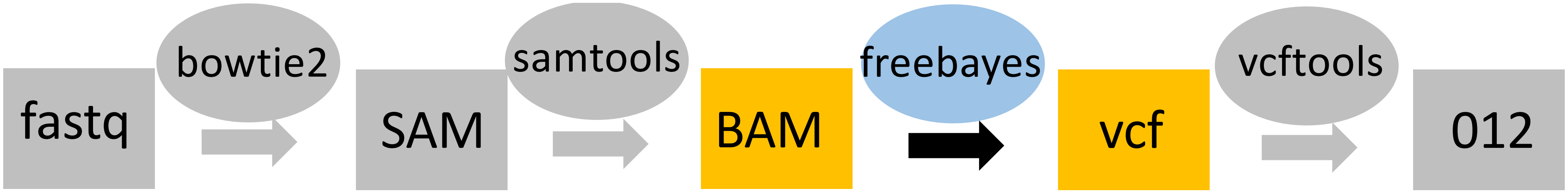
Observed Haplotypes

CATTGGATCA	x8	$(A)_7$	x10
TATTGGATCG	x9	$(A)_6$	x7
CTTGGATCA	x1	$(A)_5$	x1
CATGGGATCA	x1	$(A)_8$	x1
• • •		• • •	



Freebayes pipeline (minimal)

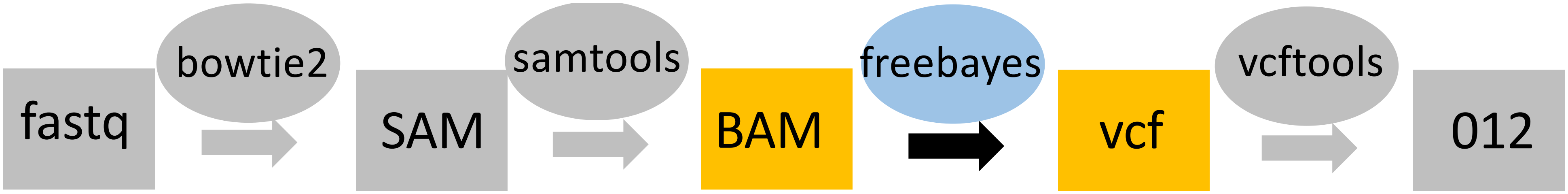
- Start with fastq reads
- Map reads (e.g. bowtie2 or hisat2)
- If paired end libraries with PCR amplification, remove PCR duplicates (Picardtools)
- Sort, compress, and index alignments (samtools)
- Call SNPs (Freebayes) Minimal call: freebayes –f ref.fa *.bam > out.vcf
- Filter SNPs (vcffilter)



fastq bowtie2 SAM samtools BAM freebayes vcf vcftools 012

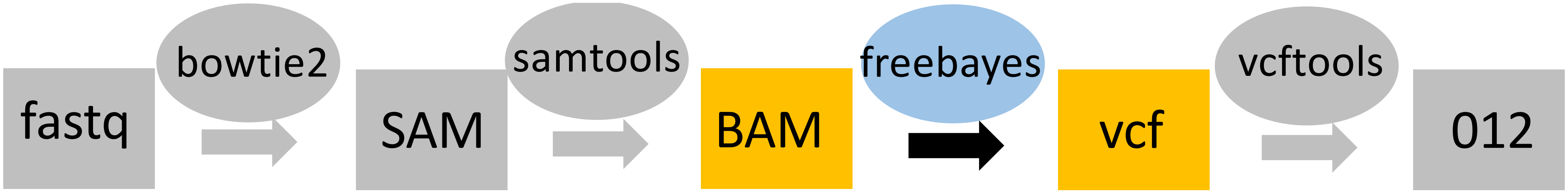
Pros and Cons of Freebayes

- Pros
  - Fast, sophisticated model
  - Easy interface, easy to customize via command line arguments
  - Good support for local multithreading (freebayes-parallel) and cluster parallelization (just split a bed file of your contigs into as many jobs as you like)


fastq → bowtie2 → SAM → samtools → BAM → freebayes → vcf → vcftools → 012

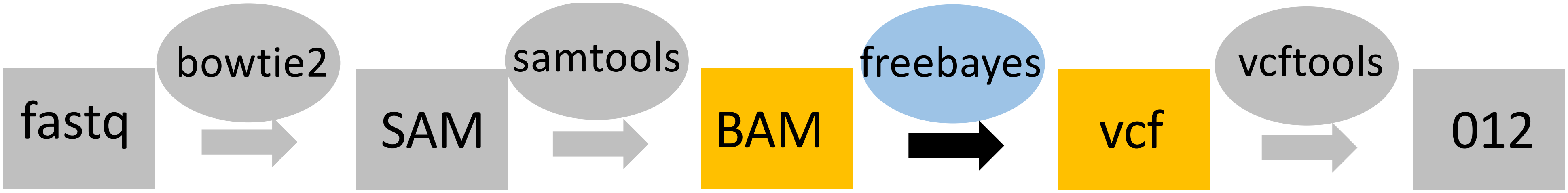
Pros and Cons of Freebayes

- Cons
  - Relentlessly haplotype based, so it can sometimes be hard to get just, like, normal biallelic SNPs (this is a feature too)
    - Utilities like vcffilter, vcfallelicprimitives, and vcfbiallelic help
  - Under rapid development, so sometimes tools change or useful features haven’t been implemented or documentation is less good



GATK’s HaplotypeCaller

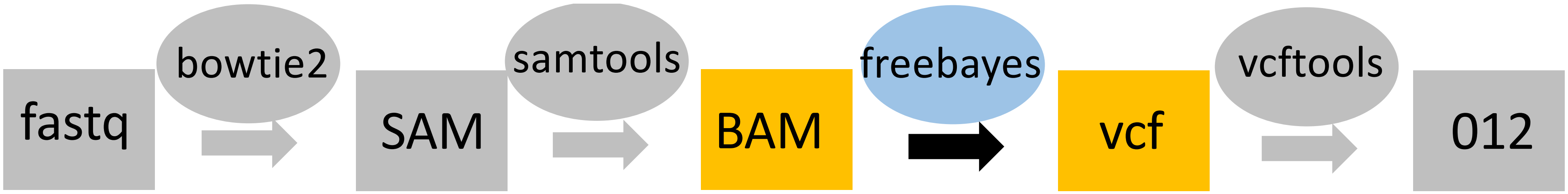
- Defines “active regions”
- Determines haplotypes by reassembling the active region
- Determines likelihoods of the haplotypes given the read data
- Assigns sample genotypes
- Outputs VCF or gVCF file



fastq → bowtie2 → SAM → samtools → BAM → freebayes → vcf → vcftools → 012

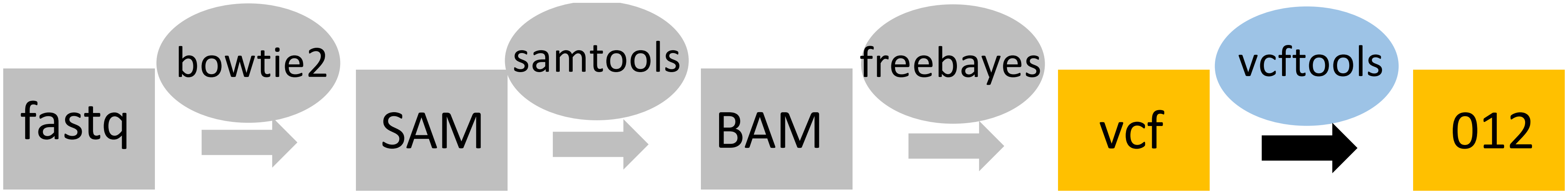
Pros and Cons of GATK

- Pros
  - Extensively documented, lots of support
  - Clear, easy-to-interpret output



Pros and Cons of GATK

- Cons
  - Not as easily customizable, not as easy interface as Freebayes
  - Slower than Freebayes



fastq → bowtie2 → SAM → samtools → BAM → freebayes → vcf → vcftools → 012

012 SNP Matrix - format

- Variant call file (.vcf) – list of alleles and their likelihoods
- Use vcftools to convert your filtered SNP file (.vcf) into a 012 matrix
- Each row is a sample, each column is a SNP
  - First column is sample number, starting at 0
- 0 : both copies of reference allele
- 1: heterozygous
- 2: both copies of alternate allele

ex:

```
0 0 1 0 0
1 0 0 0 2
2 2 0 1 0
3 2 0 0 1
```

012 SNP Matrix vs. other methods

- 012 genotype calls does not represent uncertainty about genotype
    - Ex: If we only have 2 mapped reads, both the alternate allele at the locus, this could be homozygous alternate or a heterozygote where we did not sample the other allele
- We can remove uncertainty by filtering for SNP calls that we are very confident about
- This is the strategy we are using in the pipeline today

Other methods: using genotype likelihoods

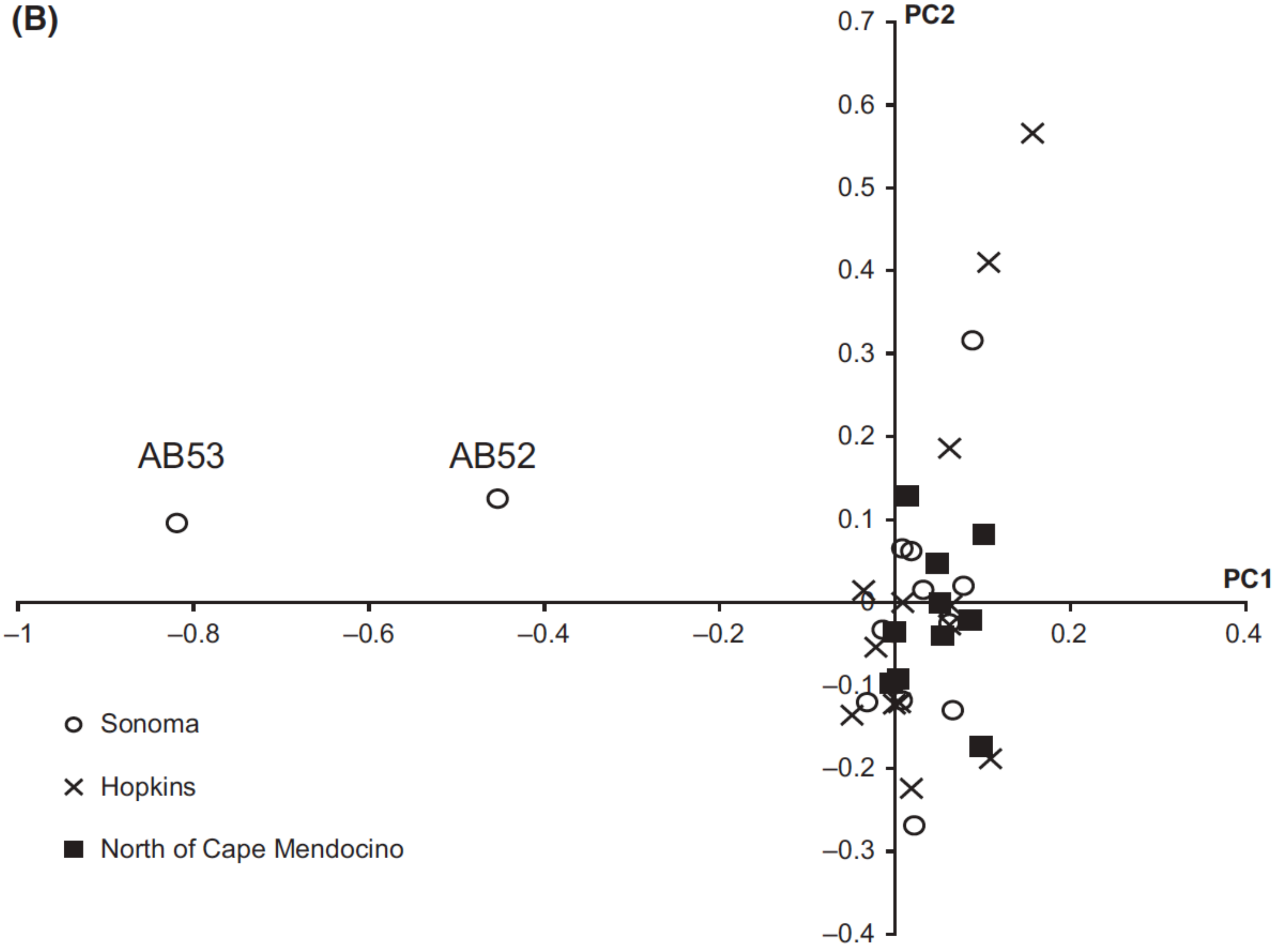
- Fewer programs use this format, examples are GPAT++, Angsd

Principle Components Analysis

Taking thousands of SNPs into account simultaneously

- Population structure
  - Detecting outliers
- An axis of variation to compare to environmental variables

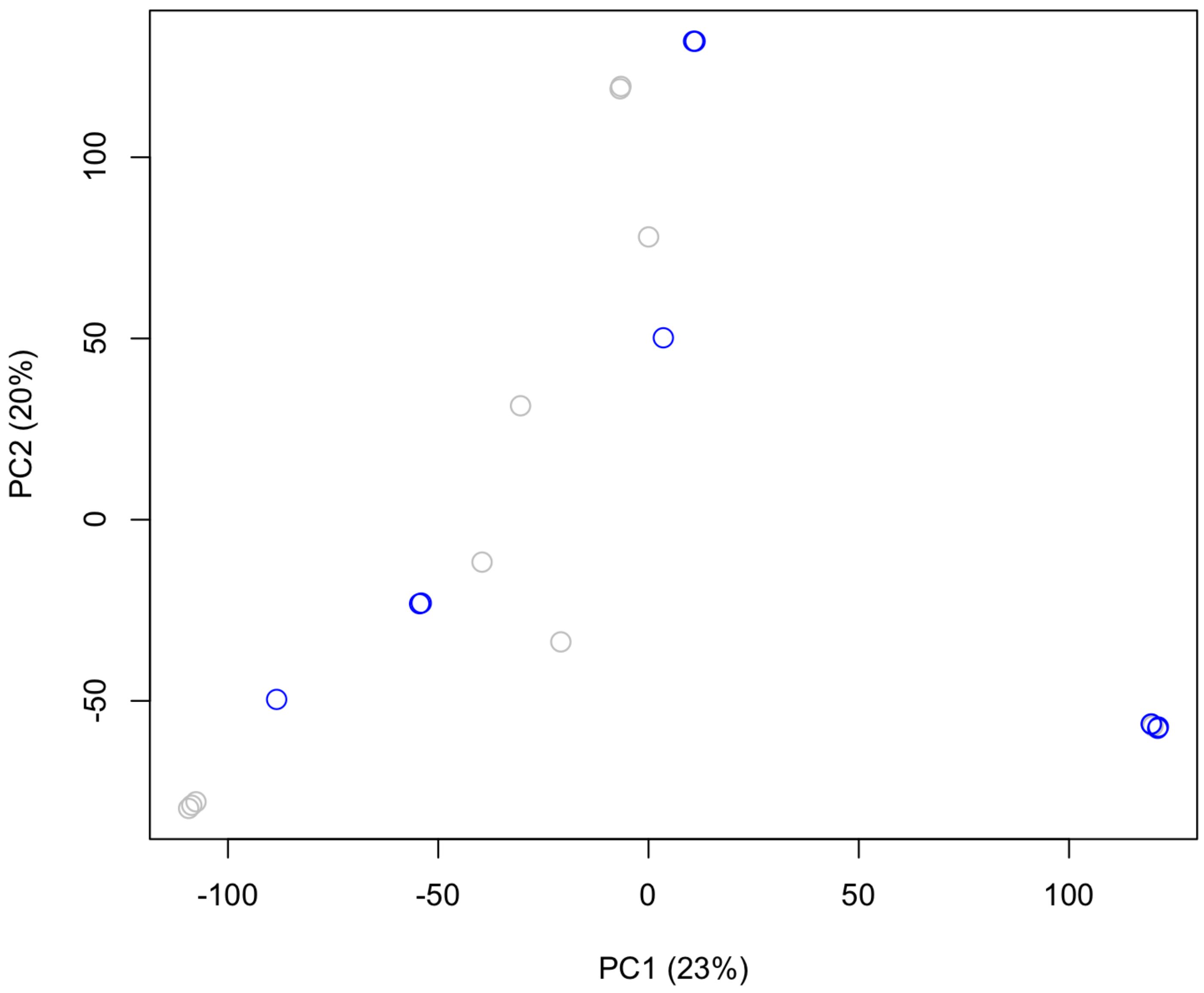
Population differentiation



Transcriptome-wide, no differentiation

de Wit & Palumbi 2013

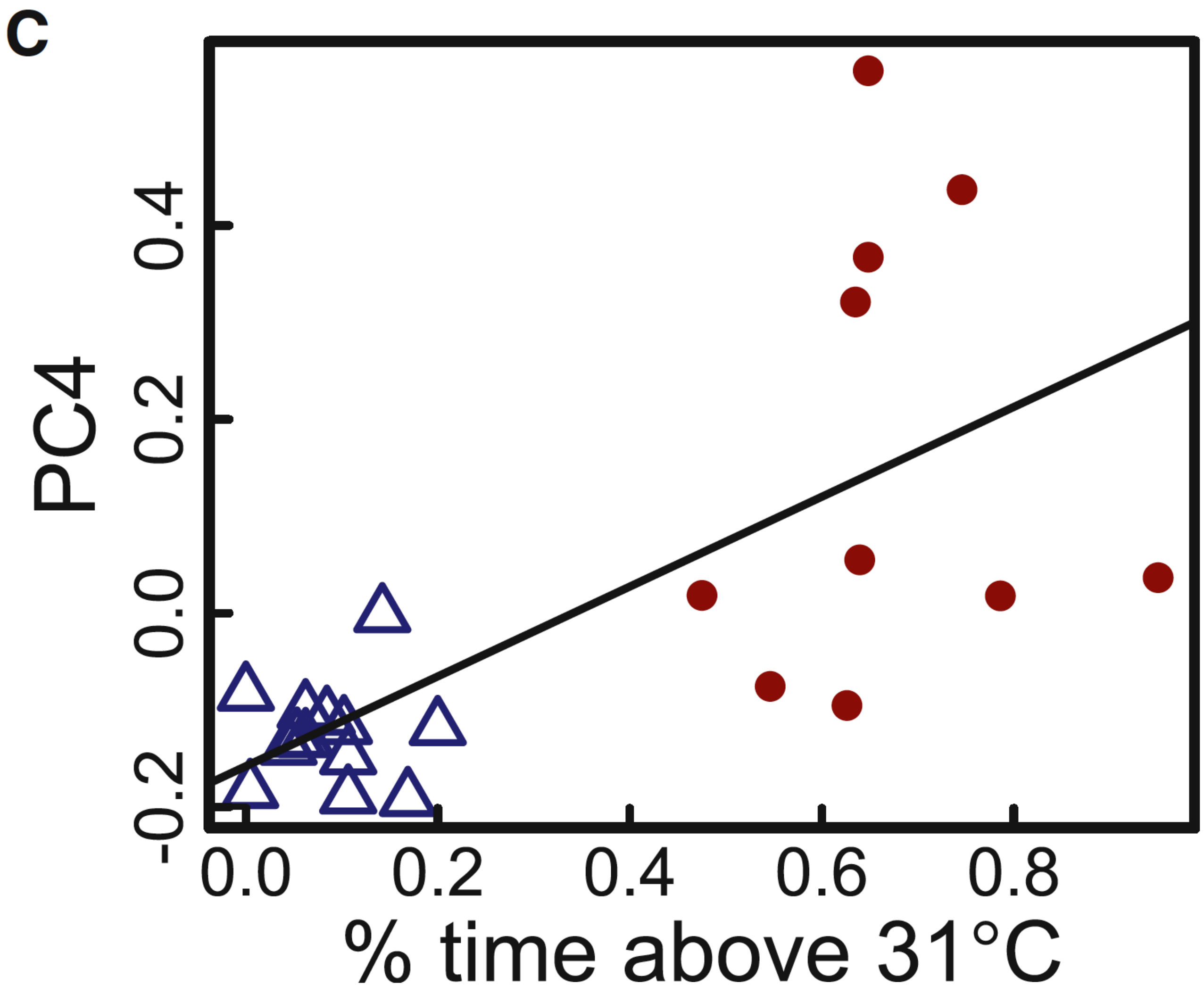
Clones in natural populations



11 points on PC1 & PC2 from 20 samples

Morikawa & Palumbi in prep

Correlation to environmental variables



PC4 and correlation to environment

Bay & Palumbi 2014

F_{ST} analysis

- F_{ST} as an index of genetic differentiation
  - Ranges from 0 (identical) to 1 (completely different)

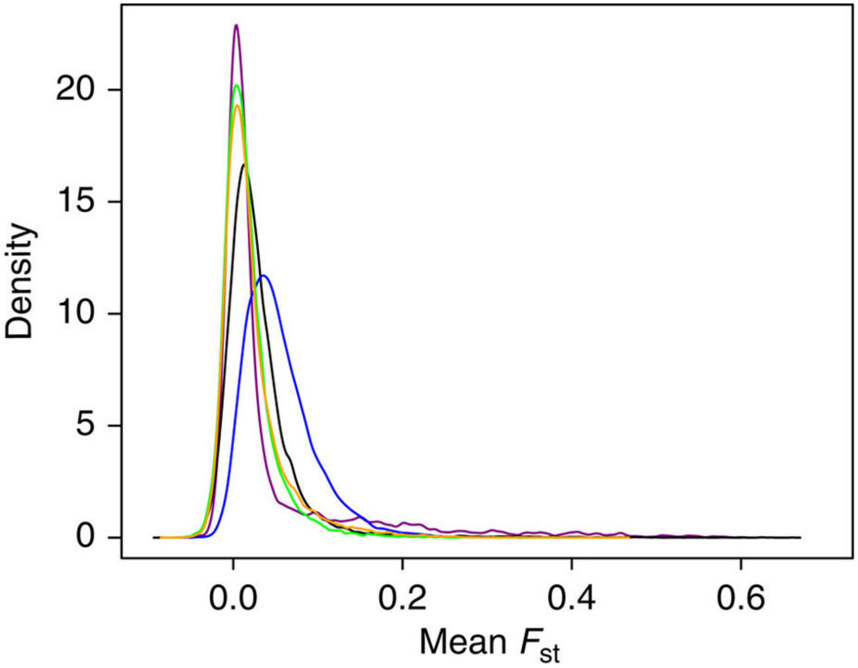
- Classic measure – can compare to MANY studies

- Wright (1953) F_{ST} vs Weir and Cockerham (1984) F_{ST}

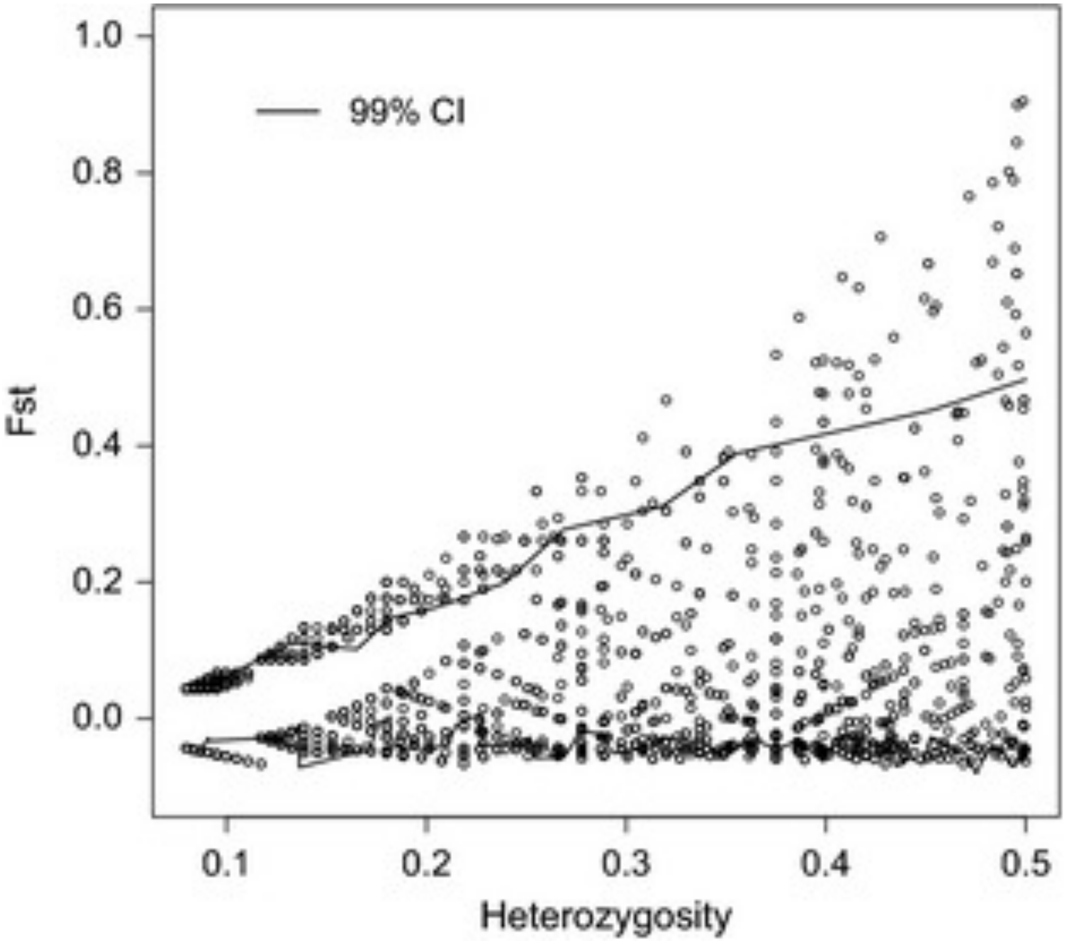
- With transcriptomic levels of data, we need to look at patterns, not necessarily individual loci

F_{ST} analysis

Compare density distributions of pairwise F_{ST} between pops



Compare F_{ST} to heterozygosity to find unusually high F_{ST} loci



Tutorial

https://github.com/bethsheets/SNPcalling_tutorial

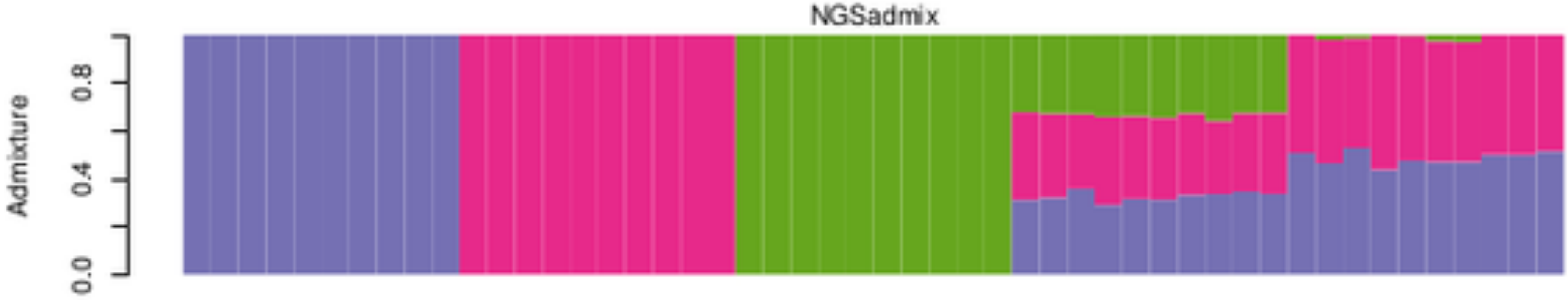
Why use transcriptomics?

- Reduced representation
- Focus on parts that matter (protein coding)
- Expression links SNPs to phenotype

Population structuring : NGSadmix

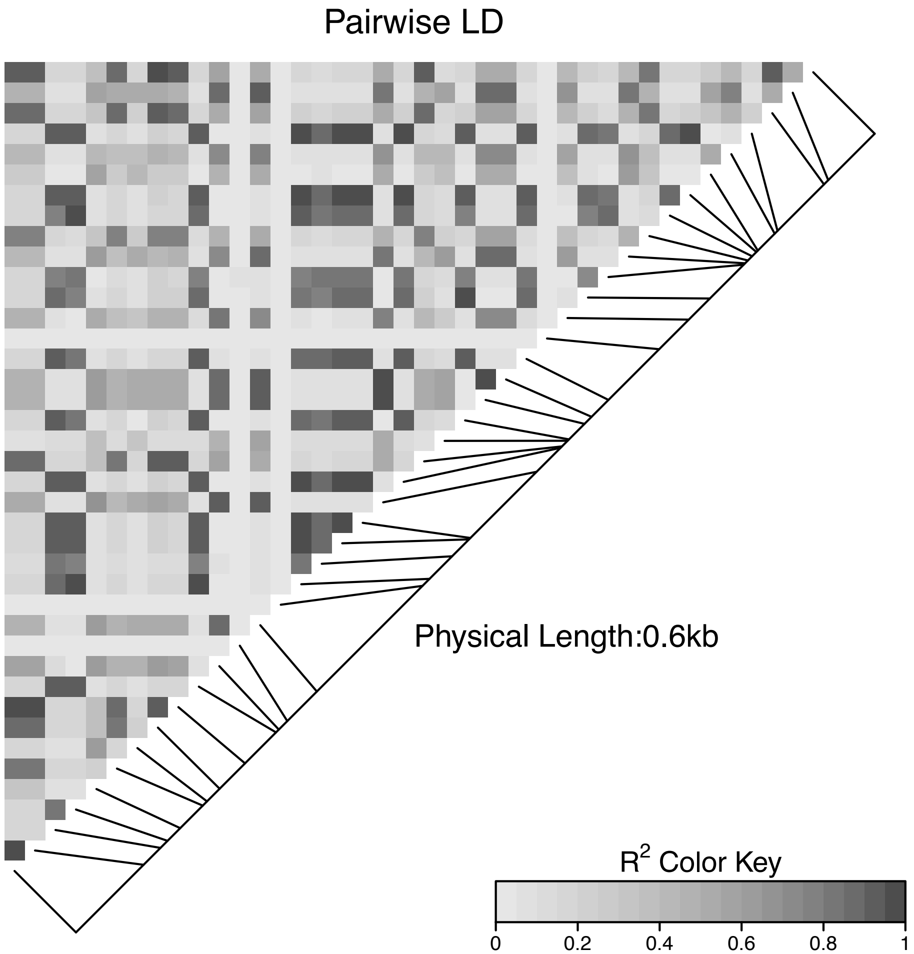
http://www.popgen.dk/software/index.php/ANGSD
http://www.popgen.dk/software/index.php/NgsAdmix

NGSadmix takes genotype likelihood files and outputs a view of population membership of individuals



Linkage disequilibrium : within contig

- Arises from a lack of recombination between SNPs
- SNPs inherited as a block, not necessarily adjacent



Barney & Palumbi in prep

BUT: LD may exist beyond the extent of individual transcripts, across multiple genes on the same chromosome, or across chromosomes!

Linkage disequilibrium : whole genome

Extra data are needed for better understanding of physical linkage at the chromosomal level:

- linkage maps from pedigreed individuals
- OR (*AND* is better!), a well-assembled genome

Pairwise LD calculations for all SNPs throughout assembly

- Generates a matrix of r2 values for each pair of SNPs

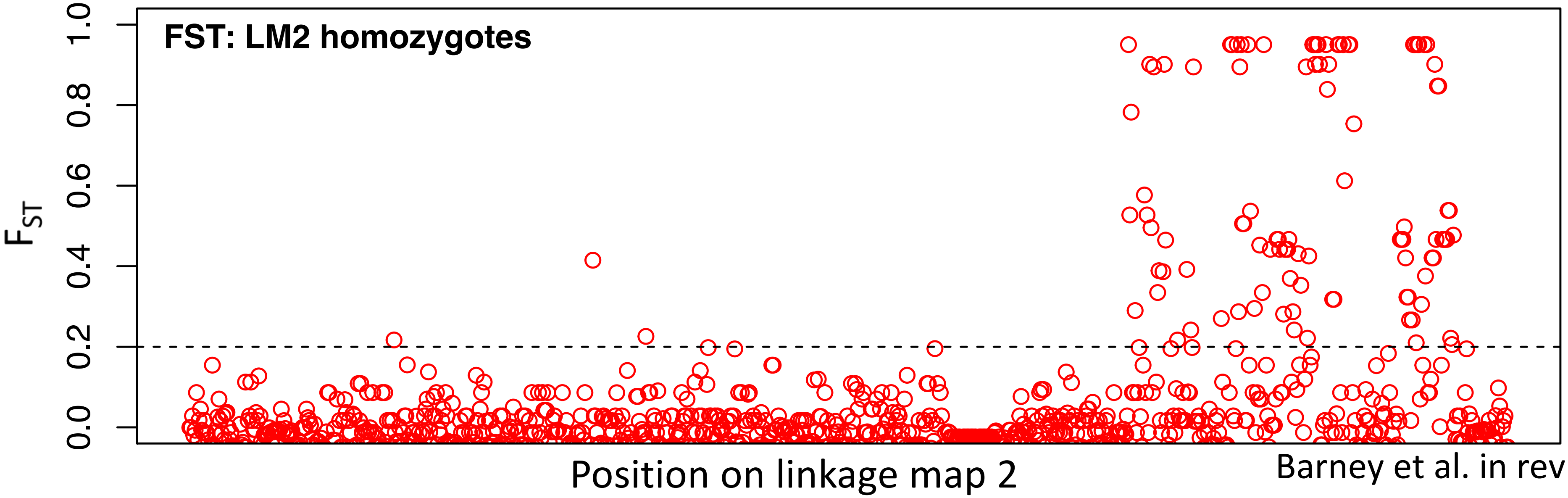
Decision: How to cluster the SNPs, and what cutoff to use?

- single-linkage clustering (A linked to B, B to C, so A,B,C in cluster
- $r^2 > 0.75$

Linkage disequilibrium : whole genome

“Islands of divergence” or supergenes?

A GO enrichment analysis of linked region may reveal overrepresentation of genes of related function, the classical definition of a supergene



Barney et al. in review

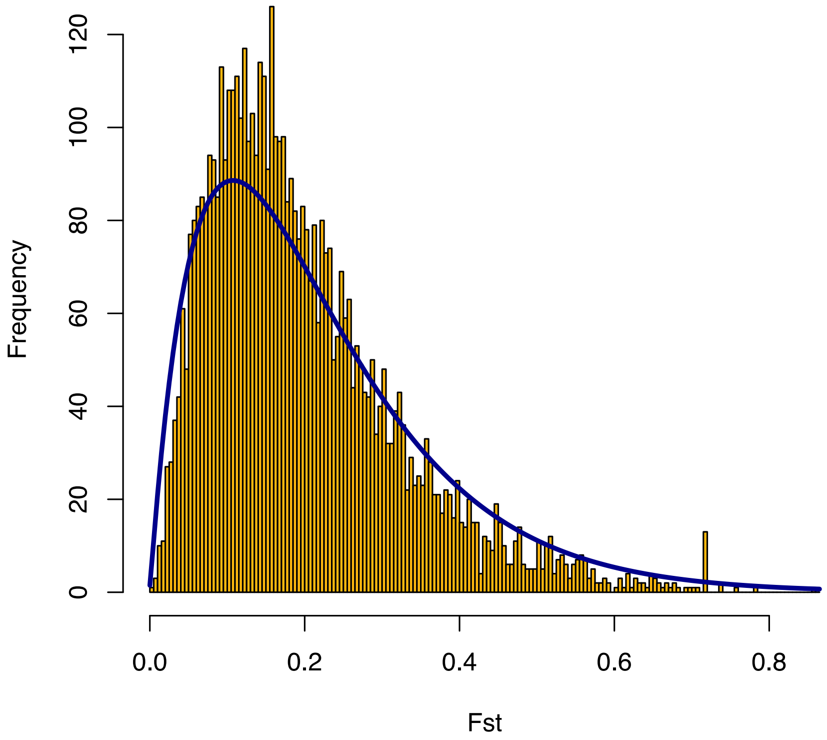
Genome wide scans for selection: identify genomic regions that exhibit signatures of diversifying selection.

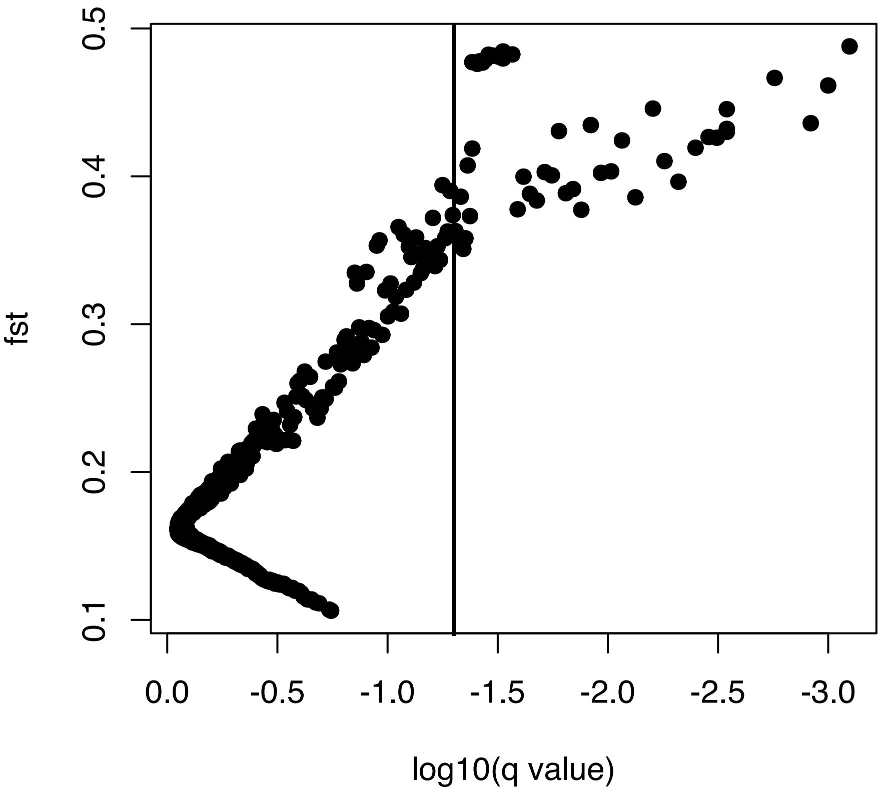
- Two main approaches:
  1. **Population differentiation** (PD) approaches

  2. **Ecological association** (EA) approaches

Population differentiation (PD) approaches

- Identifying loci that show unusual allele frequency differentiation among populations





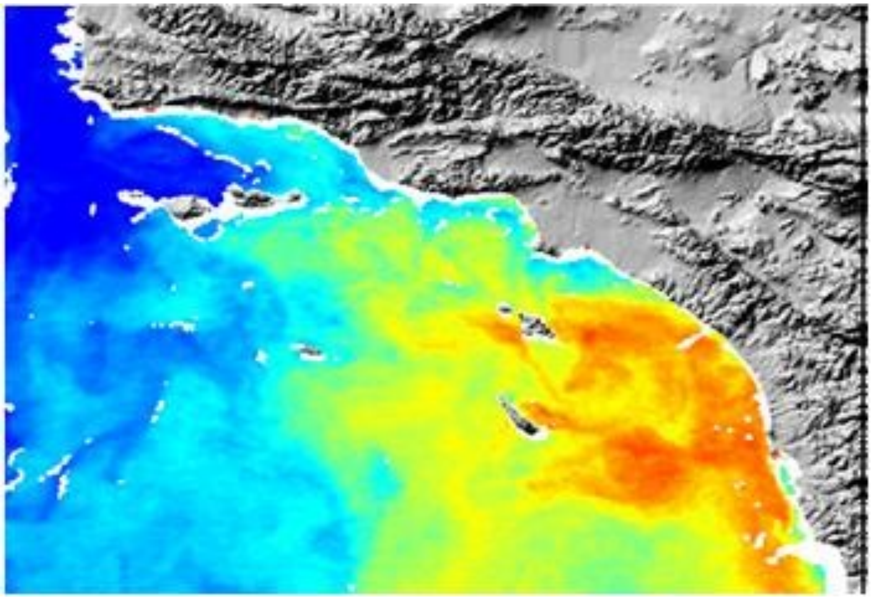
Population differentiation (PD) approaches

- Pros:
  - Can be effective at identifying genes under selection without known phenotypes
  - Does not require *a priori* information concerning the environmental forces that act as selective pressures.
  - Can screen high number of markers to identify candidate genes for further investigation.

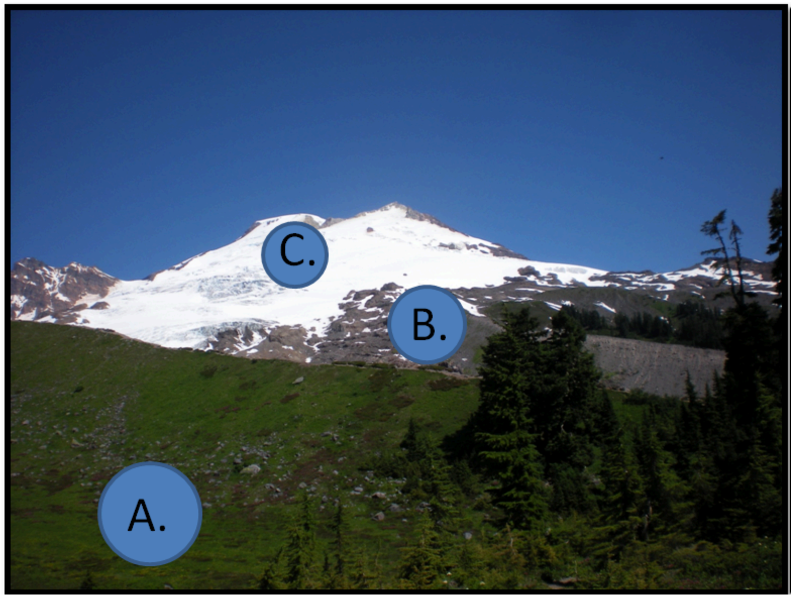
- Cons
  - Plagued by false positives and false negatives
  - Candidate outlier loci often vary in pairwise population comparisons, and therefore overall divergence (global FST) may not detect candidates that are under selection in only a portion of populations.
  - Limited power in detecting balancing selection and various forms of weak divergent selection.

Ecological association (EA) approaches

- Identifying loci with a strong association between allele frequencies and environmental variables



http://www.sccwrp.org



https://sites.google.com/site/nkooyers/

Ecological association (EA) approaches

- Pros:
  - Uncover selected loci without knowledge of selective environment (can feed the programs lots of data and see what comes out)
  - More powerful than PD approaches

- Cons
  - Requires detailed environmental data
  - If IBD then surprisingly low power (need to account for population structure)
  - High variability among runs

Genome wide scans for selection: Conclusions

- Assist in identifying loci under selection
- Be careful as can lead to numerous false-positives!
- Best used in conjunction with GWAS and linkage-mapping approaches

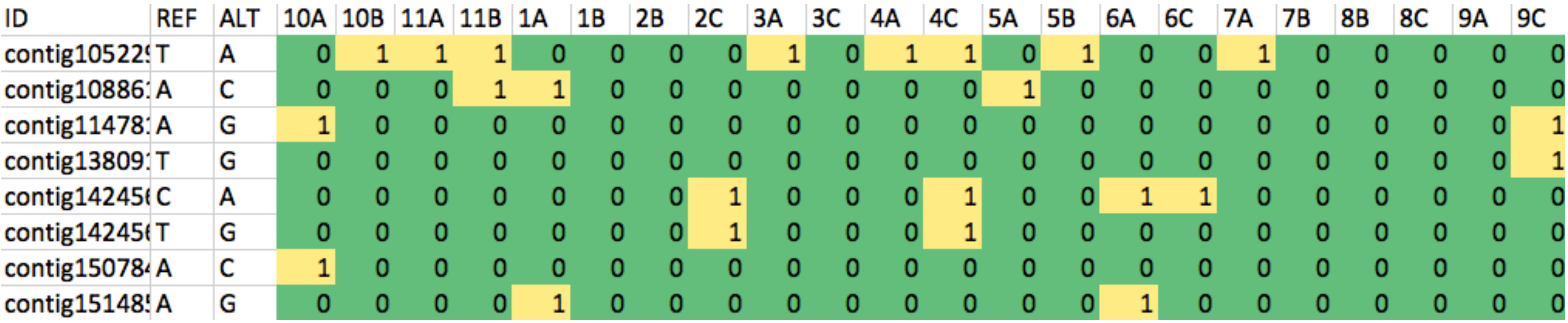
For reviews on the topic:

- **Lotterhos and Whitlock (2015) Mol. Ecol.**
- **Hoban et al. (2016) Am. Nat.**
- **Rellstab et al. (2015) Mol. Ecol.**
- **Narum and Hess (2011) Mol. Ecol. Res.**
- **Haasl and Payseur (2016) Mol. Ecol.**

- **Common programs:**
  - BayeScan
  - Arlequin
  - Lositan
  - OutFLANK
  - PCAdapt
  - BAYENV2
  - BAYESCENV

Custom Analyses of VCF files

- Identifying somatic variants



ID	REF	ALT	10A	10B	11A	11B	1A	1B	2B	2C	3A	3C	4A	4C	5A	5B	6A	6C	7A	7B	8B	8C	9A	9C
contig10522	T	A	0	1	1	1	0	0	0	0	1	0	1	1	0	1	0	0	1	0	0	0	0	0
contig10886	A	C	0	0	0	1	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
contig11478	A	G	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
contig13809	T	G	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
contig14245	C	A	0	0	0	0	0	0	0	1	0	0	0	1	0	0	1	1	0	0	0	0	0	0
contig14245	T	G	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0
contig15078	A	C	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
contig15148	A	G	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0

8 of 214 somatic variants identified in a single coral colony

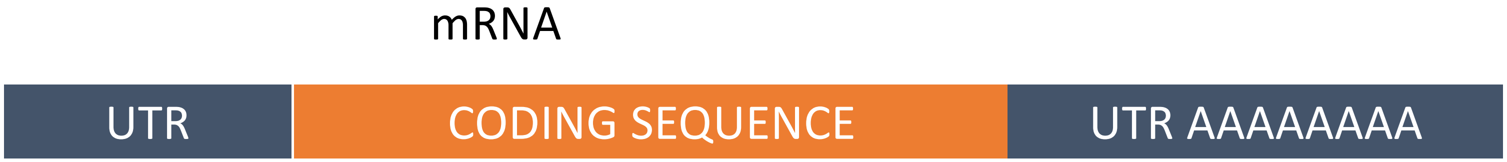
Green = homozygote

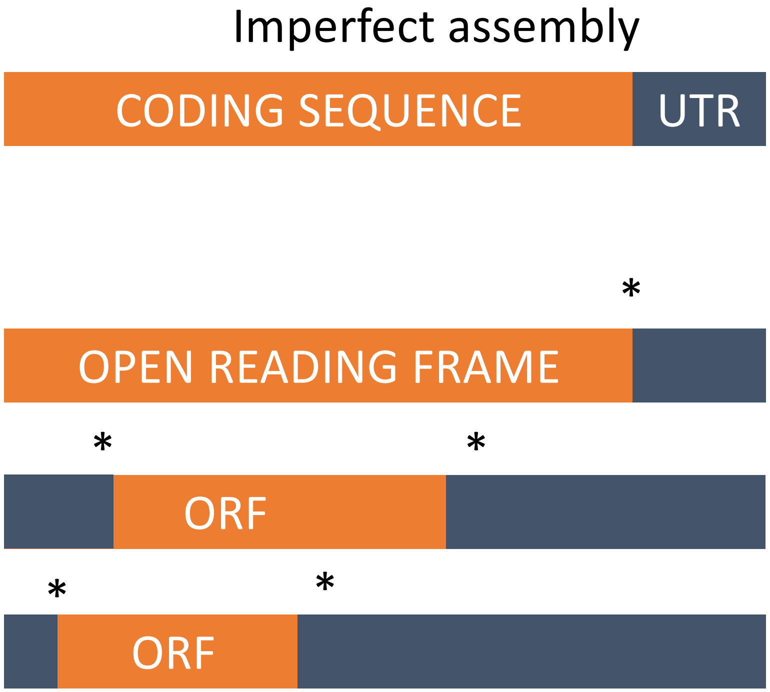
Yellow = heterozygote

How to link SNPs to phenotypes

- Protein coding
- eQTLs

Protein coding change

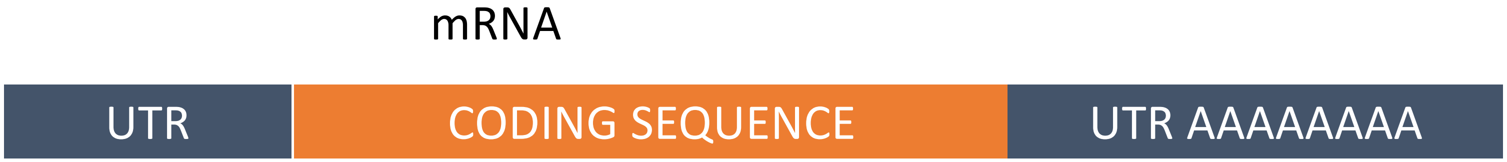




Goal: find protein coding sequences in transcripts

Tools:
OrfPredictor
Blastx
GMAP

Protein coding change

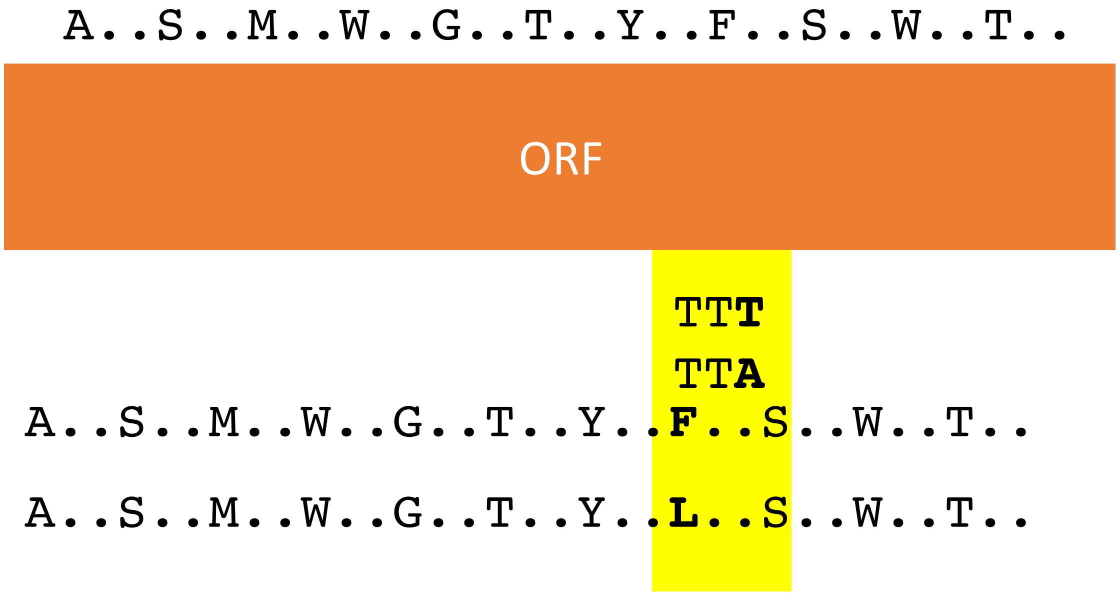


VCF says:

Contig1 24 T A

Goal: Find amino acid changes due to SNP variants

Tools:
Biopython
SnpEff



A..S..M..W..G..T..Y..F..S..W..T..

ORF

TT**T**

TT**A**

A..S..M..W..G..T..Y..**F**..S..W..T..

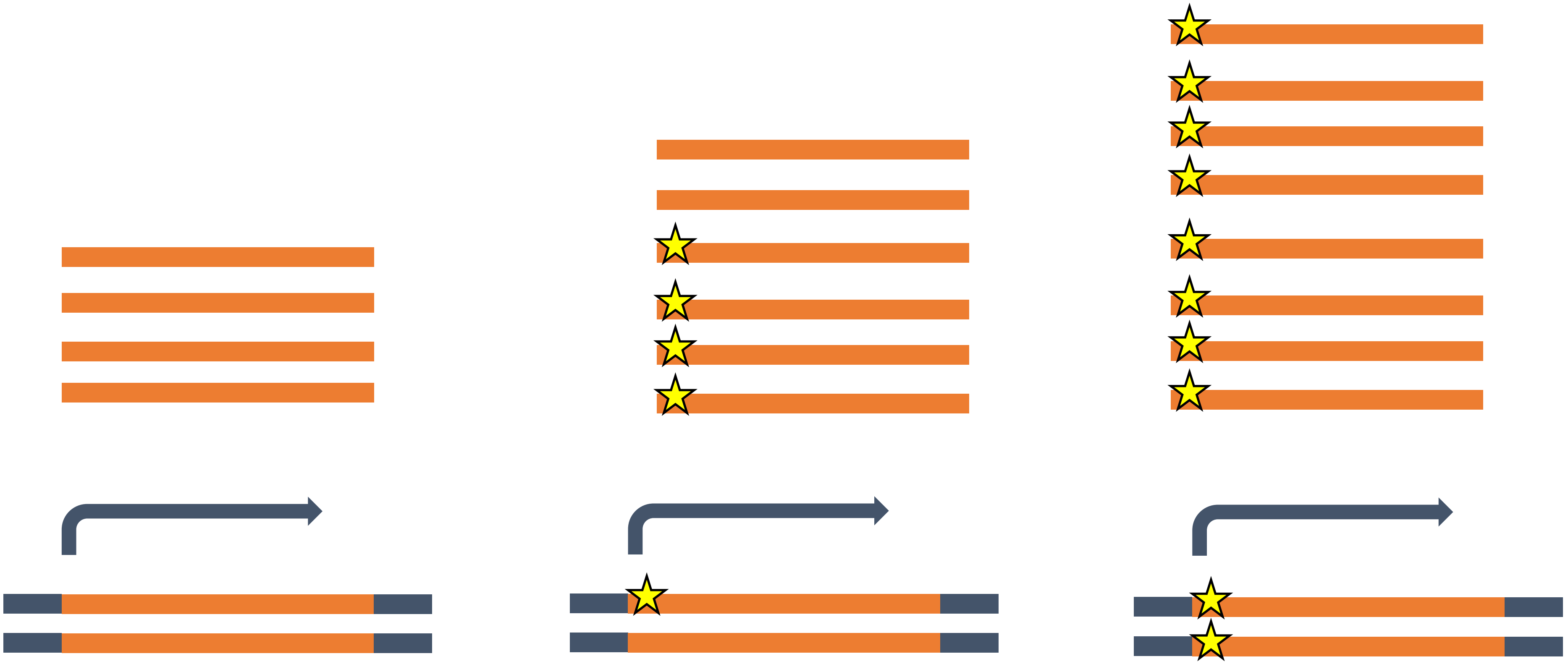
A..S..M..W..G..T..Y..**L**..S..W..T..

Questions to ask with protein coding SNPs

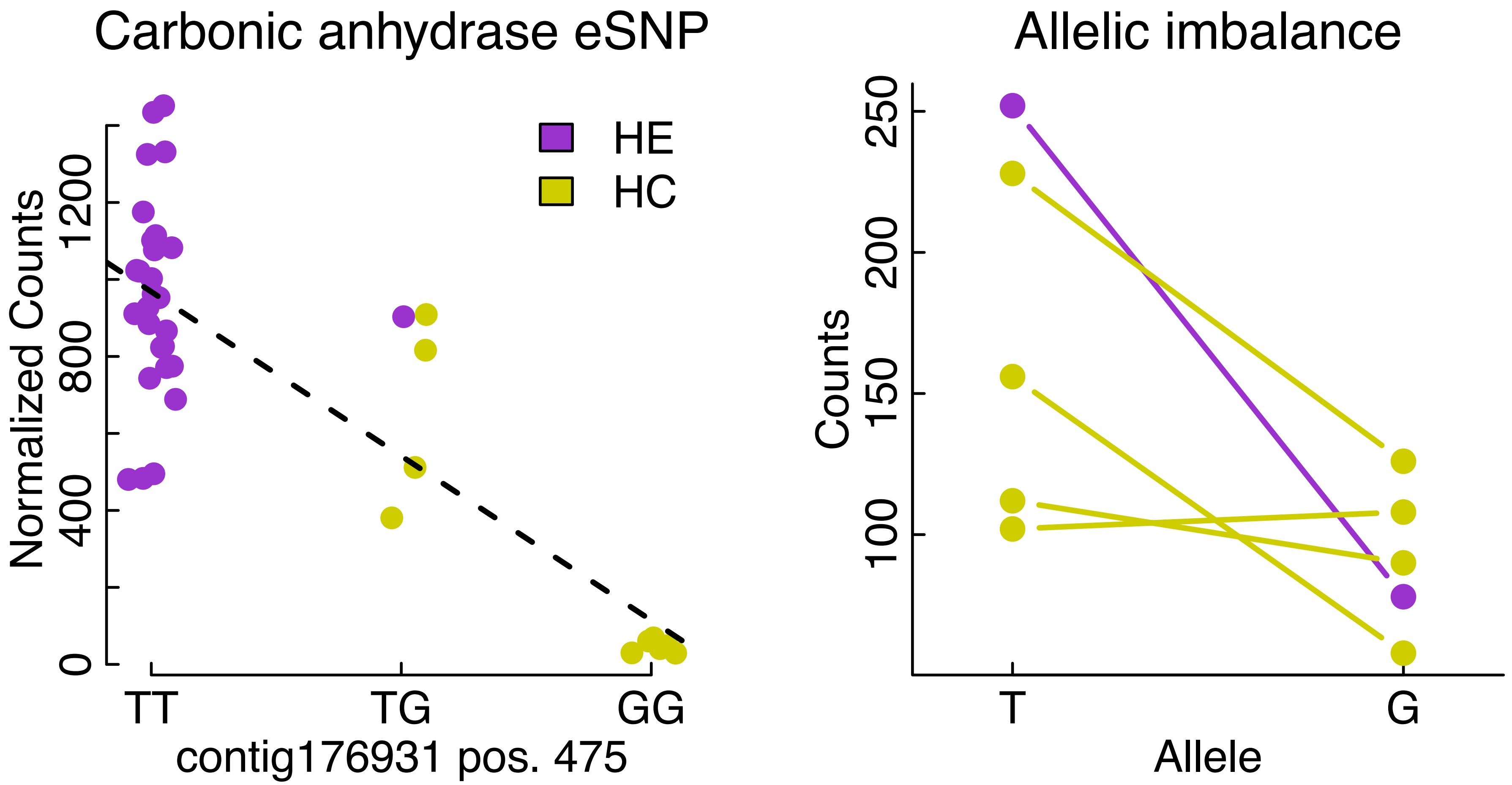
- Does a protein sequence show strong differences between populations or species?
- Do non-synonymous variants fall in residues known to be important from multi-alignment conservation or crystallography?
- Are proteins evolving to be different more quickly than you would expect from genetic drift?
  - McDonald Kreitman test for multiple samples from two species
  - dN/dS (rate of non-synonymous change over rate of synonymous change) is a general index of the rate of protein evolution

How to detect eQTLs:

SNPs that are correlated with expression



eQTLs explain variation within and between species



Questions to ask with eQTL SNPs (eSNPs)

- Do interesting genes show strong differences in expression between populations or species?
- Are these genes involved in gene networks or pathways that you are interested in?
- Do many genes in the same pathway show expression changes in the same direction?