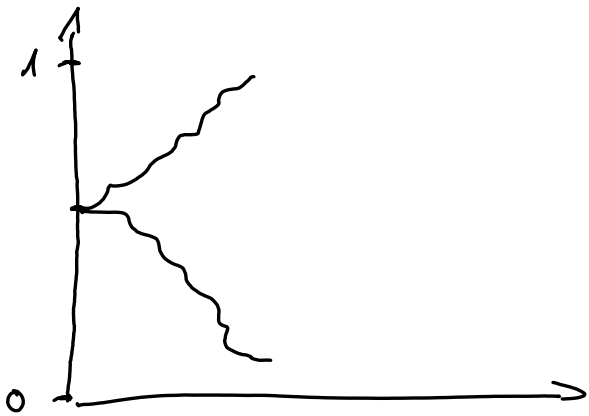
Population genetics:

HWE : $\boxed{\text{infinite pop}}$, etc.

↓

if not:



sampling
↓
random effects
↓
fixation of 1 allele
||
DRIFT

↳ if there is migration:
or there are $\boxed{\text{sub-populations}}$

Island model

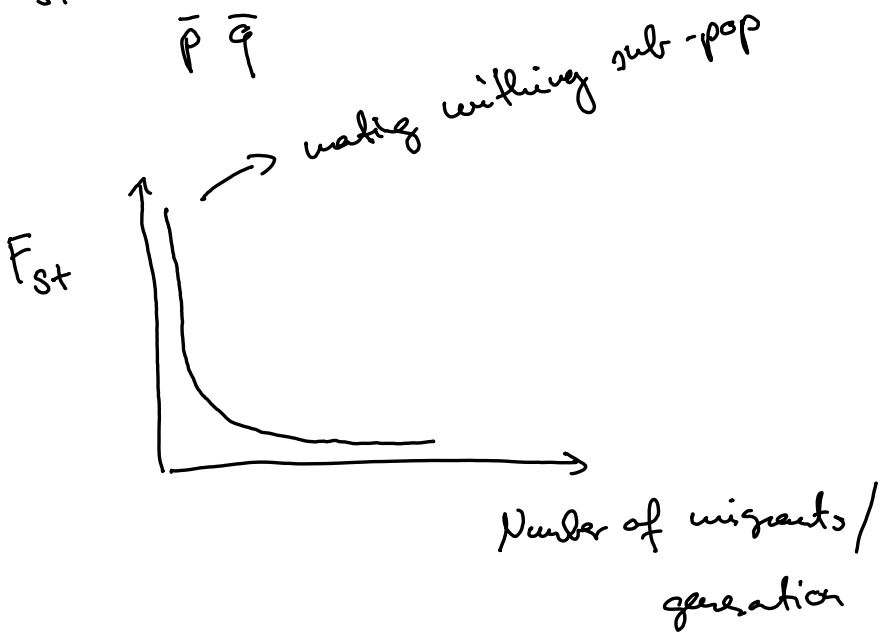
F_{st} –

fixation
sub relative to whole

still "fantasy" world:

- same size
- don't grow/extinct
- equal migration ↑

$$F_{st} = \frac{var(p)}{\bar{p}\,\bar{q}}$$

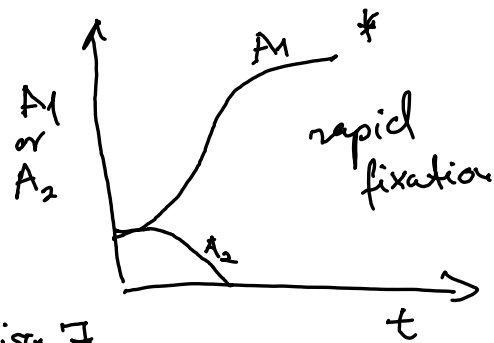


↓ what if there is selection

s = selection

A_1A_1	A_1A_2	A_2A_2
1	1 if recesive	$1-s$
	$1-hs$ if intermediate phenotype	

relative fitness (offspring)



* selective sweep → little polymorphism ↓

→ In real life lot of polymorphisms

Allozyme – variant of protein measured

↳ a lot of variation!

Molecular evolution: after DNA seq data

"History encoded in DNA"

Neutral theory of molecular evolution

A G T C T C G A T

A C T C T C G A T

polymorphism

probability of a new mutation

$$\frac{1}{2N} = \mu$$

also the prob of that mutation of becoming fixed

"wobble": 3rd bp, probs. don't change amino acid

↳ synonymous change

non-syn

synonymous

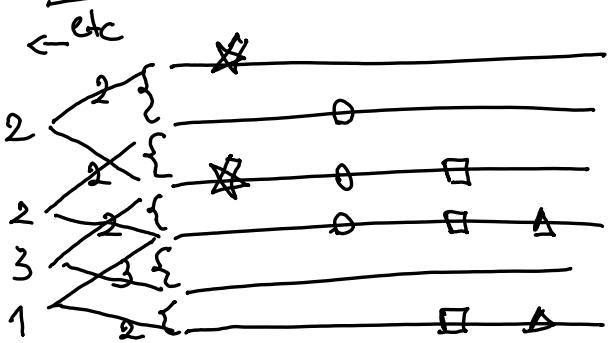
$$\frac{dN}{dS} = \frac{N \text{ mutations} / N \text{ sites}}{S \text{ synonymous mutations} / S \text{ sites}} = \omega$$

if $\omega \gg 1$ | $\omega > 2$ → signal of selection, ⊕

if $\omega = 1$ neutral

if $\omega \ll 1$ negative, purifying selection ⊖

$\boxed{\text{Tajima D}}$ relative level of diversity



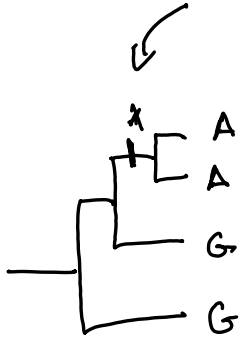
{ π nucleotid diversity }

S # of segregating sites

→ $\pi = 2.2$

$S = 4$

Greg Ray – signal of selection in non-coding conserved regions across taxa?

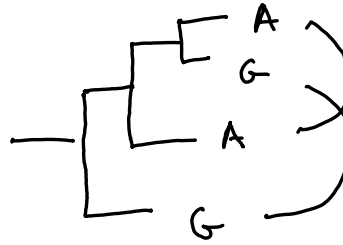


* A mutation

but neutral?

selection?

if



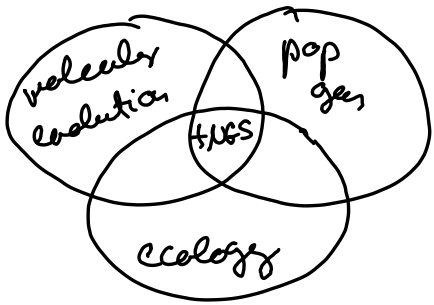
same role / env

same role / environment

↓

indicator of function

Ecological Genomics:



disentangling neutral and ⊕/⊖ selection

Monday

Life update

Discussion of paper

} sign up for 1 2 weeks prior talk to Melissa

Wednesday

Computational analysis

Library : collection of DNA / RNA prepared for sequencing

↓

Prep. method:

- WGS
- RNA-seq
- exome caption
- GBS
- Amplicon
- Bisulfite, ATAC-seq

⇓

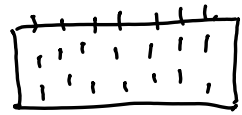
load on sequencer

- Illumina
- PacBio / ONT

Illumina - Biotech company

1. Library prep : • fragmentation
   • attach adaptor

2. Cluster generation

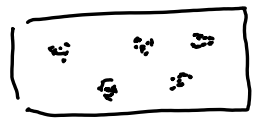


Complementary adapters → fragments align

have the primer site

↓ PCR

Bridge so that it can go both ways



clusters

3. Sequencing by synthesis + dNTPs fluorescent

recorded by machine

4. Data analysis

images generated -> analysed

<u>RNA - seq:</u>

gene expression
novel transcripts

1. Time-points
2. Conditions (treatment vs control)
3. Environments
4. Populations

✓ Transcript abundance count

✓ Allele variance

Errors: mapping back to reference
RNA low conc. or high turnover
Tissue specific variation

1. Sample collection – tissues | Shock freeze
avoid contamin.

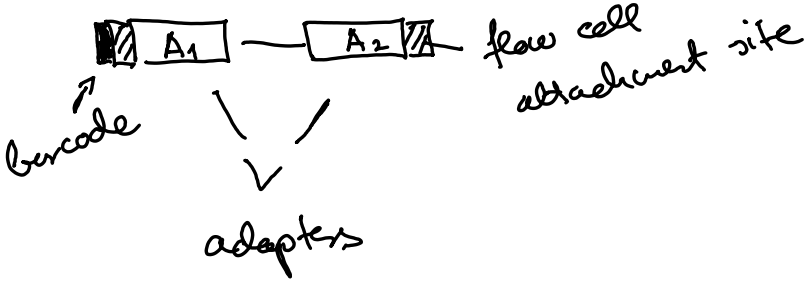
2. RNA isolation
+ quality of RNA
↳ gel electophoresis
↳ nanodrop, qubit, Bioanalyzer

3. rRNA depletion + polyA enrichment
lncRNA has polyA tail?
(she said yes)

4. Fragmentation (200-300 bp)

5. cDNA synthesis reverse transcriptase
RNA dependent polymerase

6. adding adapters
oligonucleotides
- amplification element (3' and 5' end)
- primary sequencing site (START site for sequencing)



7. Amplification - PCR

usually, 8-12 cycles before loading to flow cell

↓

ready to be sequenced

GBS - Genotyping by sequencing

to reduce amount of DNA to be sequenced

RADseq ← umbrella term

restriction site associated DNAseq

uses specific sites instead of random chop-up

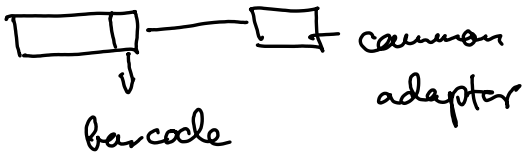
Most common: ApeKI

GCWGC
CGWCG

1. Extract genomic DNA

2. Digest with 1 restriction enzyme

3. Ligate adapters



4. multiplex (pool) individuals

5. PCR

preferentially
enriches short
sequences

→ NGS sequencing

also repetitive sites are
short seq ∈ filtered out
↳ maybe
good if
not interested

Problem: too long seq → not sequenced by
Illumina

↑

if other strand
has mutation in cut site

↓

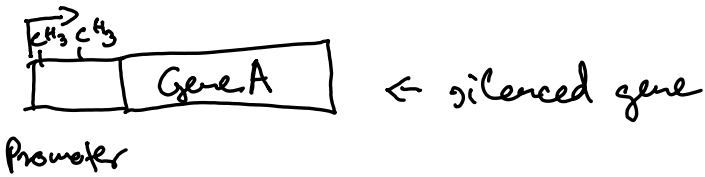
allele
drop

- if you use a restriction enzyme that only cuts methylated C → epigenetics information

- also cheap and fast (as amount of DNA is lower)

Epigenetics analysis:

A) Bisulfite – Identify meth-C usually in CG islands



← silenced gene

CG island in my promoter?

1. Isolate DNA + bisulfite

only convert met-C → U

leaves C intact

↳ in PCR A is complementary to U → compare to untreated

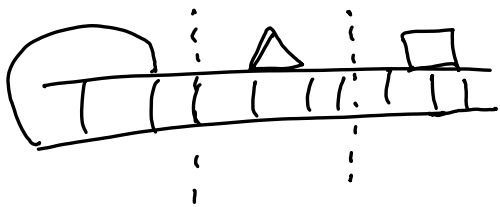
B) looking at histones instead

Chip-seq = chromatin immuno precipitation

should also be TF!

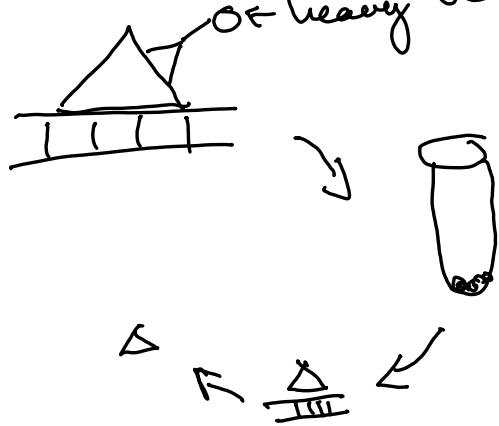
1. Fix all protein to DNA its contacting

2.



→ fragment

3. Antibody stepific to protein of interest



→ pour off top

sequence ←

C) ATAC-seq chromatin accessability

1. Introduce manifactured transposon

cut out → copy → paste elsewhere

extremely active

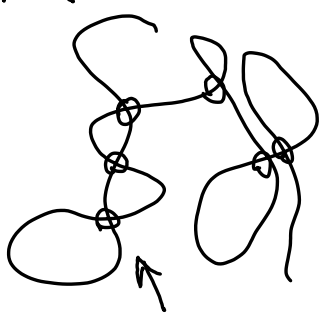
can't do it where DNA is packaged

2. DNA is fragmented as transposon does its job + adapter is ligated

↓

sequencing

! D) packaged DNA



what sequences are binding each other?

DNA-DNA interaction

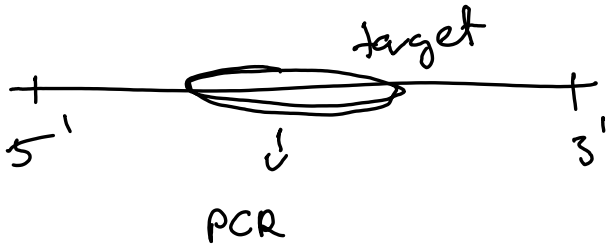
stick together, cut →

circularize + remove stick apparatus

$\boxed{\text{Hi-C seq}}$

Amplicon sequencing

↑ result of amplification



"amplicon = PCR product"

1. PCR
2. pooled
3. sequenced together

undergoes
- simplification*
- replicated genetic material
- direct repeat (head-to-tail)
- inverted repeat (head-to-head or tail-to-tail)

→ can detect low frequency variation

→ allows for multiplexing (working with multiple genes at the same time)

Uses:

- genome targeting
- detection of mut. hotspots
- SNPs

* select short, specific region of interest for sequencing

→ Metagenomics - recovery of DNA from environmental samples

↓

biodiversity

Steps:

1) Sample selection - e.g. environmental samples

- "mock communities" DNA from multiple species put together into 1

2) DNA / RNA extraction

3) Select region of interest, e.g. 16S RNA

4)

↓

Sample 1

G1

Sample 2

G1

+ add primers ← you need prior info for this

+ barcode

↓

purification - magnetic beads

↓

put samples together = pool library

↓

sequence

Bioinfo steps:

- quality control
- trimming
- demultiplexing

⋮

Exon / Exome Capture

← diseases are more likely here

but also used in ecology / evolution

e.g. detect selection, phylogenetic reconstruction, ancient DNA

1) Sample collection
   ↳ isolation of DNA
2) Fragmentation
   ↳ blunt ends
   no overhang
   ↳ adenine tails
   ↳ adapters added

have to know stuff a priori

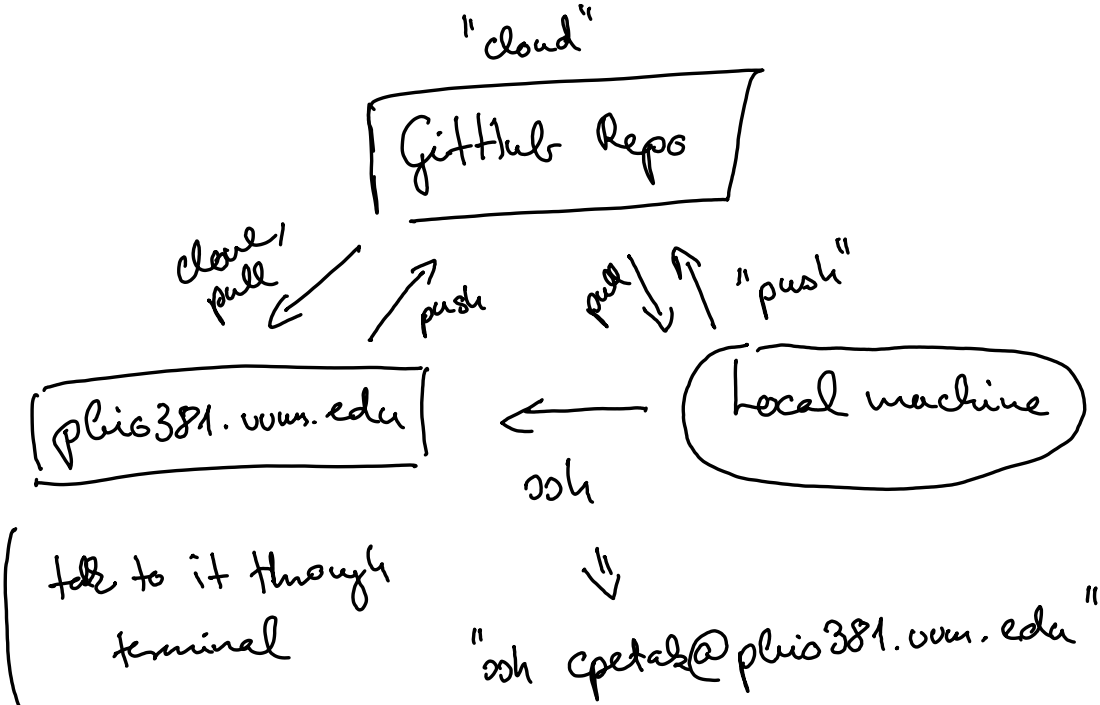
3) Library enrichment

probes = from	◆ PCR	◆ de novo	◆ divergent annotated genome
	~100 loci	~100 - 1000	~100 - 1000
need to know →	PCR primers	transcriptome / WGS	annotated genome

these probes then pull exons because of sequence pairing

4) Purify and collect target DNA

5) Quality control



talk to it through terminal

go to shared place to grab data, don't change these!

cd /data/...

in tutorial online

Population genomics

[Selection] and [drift] across the genome

env. change → abundance of species / pops Δ
→ distribution Δ
→ selection force Δ

} demographic

→ signals are very similar :-(

⇓

the challenge

Why study across the genome?

Why not pop. genetics?

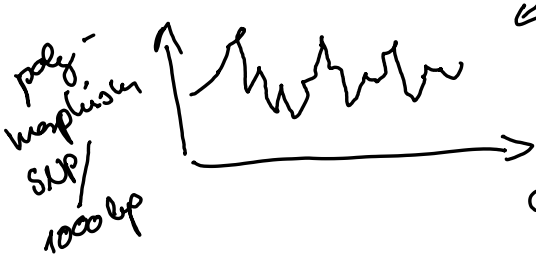
← huge variation!

due to recombination
spacial autocorrelation

↓

LD
=
linkage
↓ disequilibrium
with
↑ recombination

↓? look up



Alleles are not independent

- drift acts globally!!

vs

- selection acts locally!!

on the chromosome

"outliers in the variation by drift"

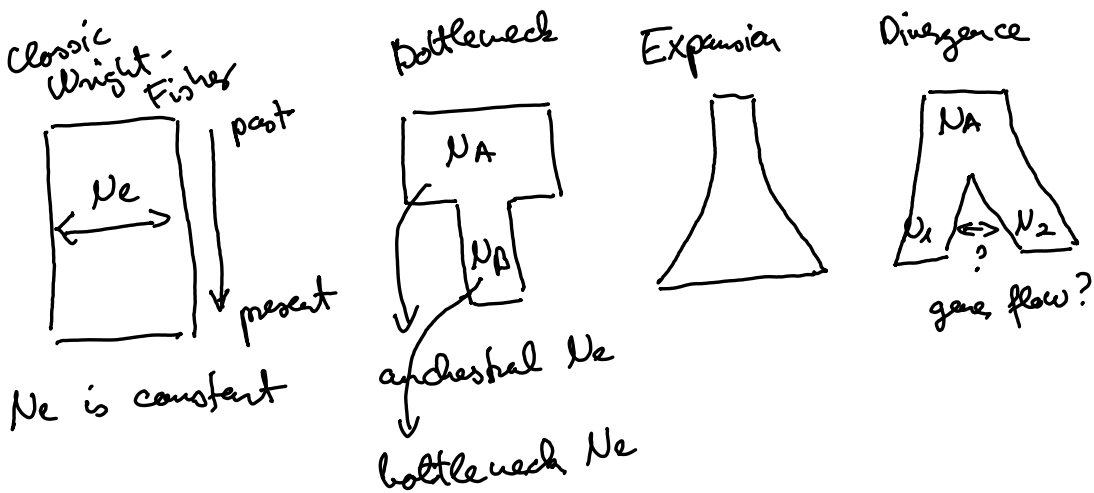
Demographic history

what is the hist. of the population?

- N_e - effective population size
  smaller → more drift effect

- Migration (i.e. gene flow) among "populations"
  ↓
  individuals exchanging genes aka demes

- Divergence times



⇓

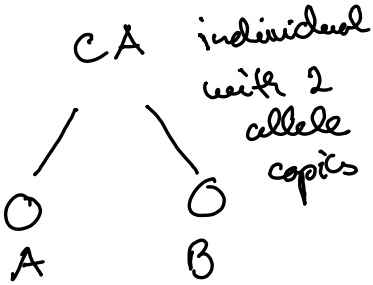
which one? → Coalescent theory

W-F model, selection, drift

coalescent theory assumes no recombination, no natural selection, and no gene flow or population structure, meaning that each variant is equally likely to have been passed from one generation to the next.

Probability that if you sample 2 allele copies they came from the same parent (1 generation away)

how gene variants sampled from a population may have originated from a common ancestor



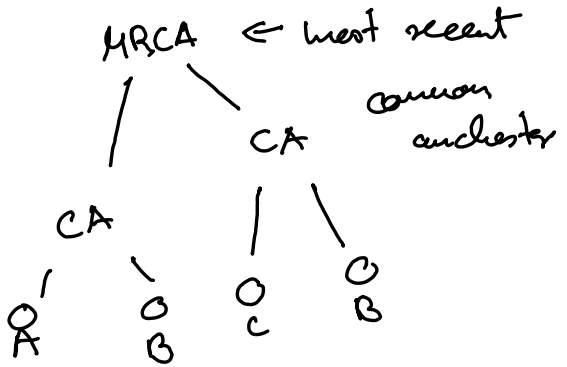
in diplaid pop:

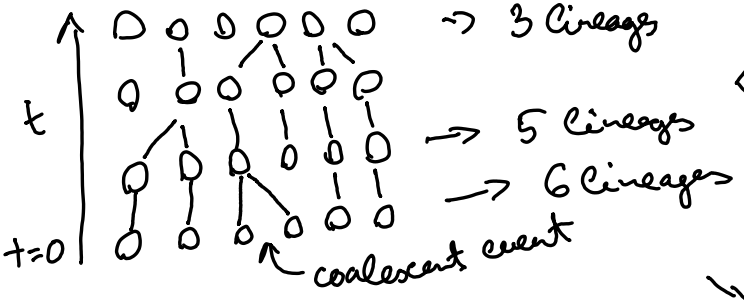
$2N_e$ = # allele copies

$$pr(CA) = \frac{1}{2N_e}$$

"phylogenetic tree for genes"

looking at 1 gene locus





← alleles copies being passed down

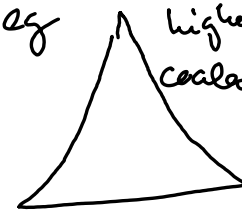
↳ you don't "see" nodes not connected to main

$2N_e = 6$ ← constant

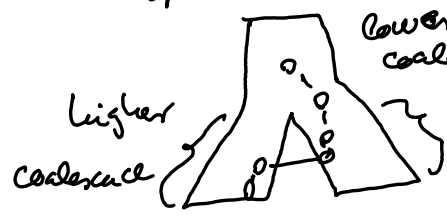
you want to do this with <u>many genes</u>, not to get "fooled"

! rate of coalescence ↑ if Ne ↓ → can estimate

eg



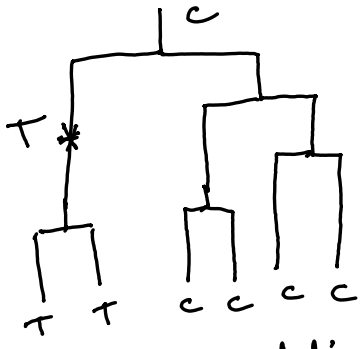
or



estimate migrations based on deeper coalescence

↓

Genealogy



now model mutation

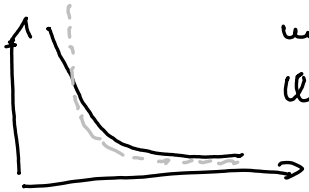
2T
4C

$\frac{2}{6} = 0.33$

how genealogy + mutations
↳ allele freq. we are seeing

→ most alleles are rare

alternate alleles



under W-F model

if pop growing more rare mut

if pop bottleneck more rare alleles

↓

an graph?

???

$\Theta = N_e \mu$

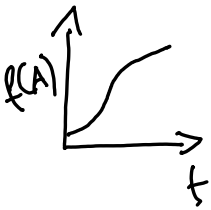
bigger pop, more mut in population

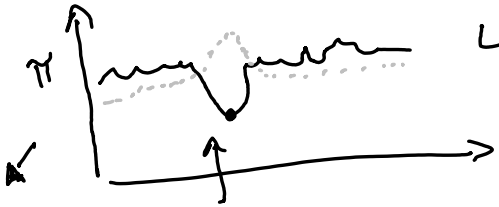
site frequency spectrum (SFS)

Selection

depends on type of selection

here we look at positive selection





selection on this allele, loci close to this are also going to decrease in π

π = nucleotide diversity

"average pair-wise difference

genealogy [|] just a stick

↖ new mut after sweep →

→ same signal as population expansion!

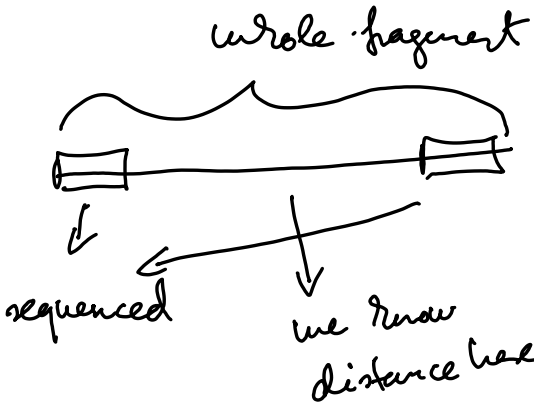
but that would be global!

↓

that is why you need

<u>multiple genealogy</u>

Pair-end sequencing



→ easier to map to reference than single-end

Pipeline

we have · fastq · gz → visualize FastQC

↓

Trimmomatic for trimming

align reads to reference genome
with BWA (best for our particular dataset)

post-processing

samtools
sambamba
(.bam)

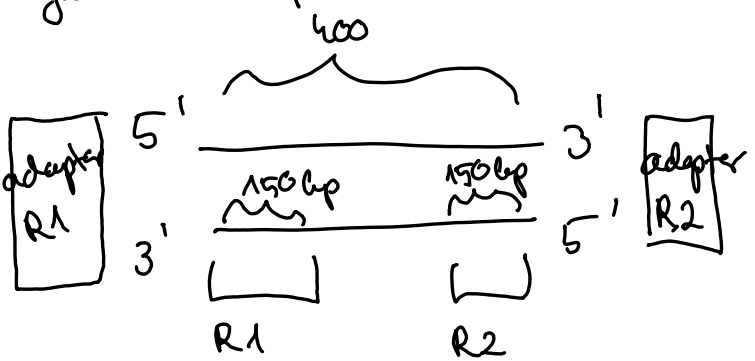
convert to binary form
easier for downstream
operations

get rid of duplicate
reads + stats
e.g. dept of coverage

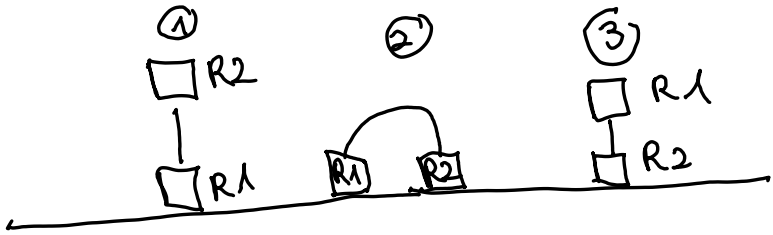
Pair-end seq.

you have forward and reverse read

R = read



sequenced at different time points



bead ↘ sequenced

→ 2 fastq files for same fragment in pair-end sequencing!

Quality scores

Probability machine messed up reading a base:

$$p = 10^{-\frac{38}{10}}$$ ← quality score in fastq matched to number (table)

$p = 0.0038$ that machine was wrong

but there can be other sources of error!
e.g. PCR error

my pop → XCV