

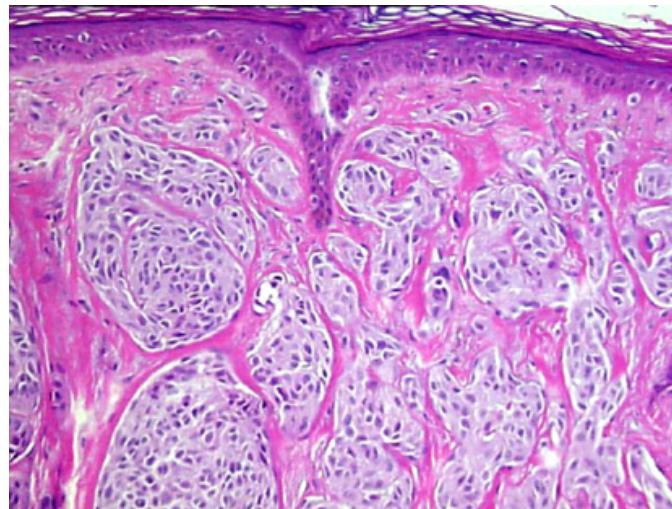
# Intrahost and pooled variation

Mark Stenglein, GDW 2018



Will mainly focus on intrahost virus populations, but other populations can be studied using similar methods and are similar in principle

Rare somatic variants in cancer  
(cancer subclones)



Population genomics using  
pools of individuals (Pool-Seq)

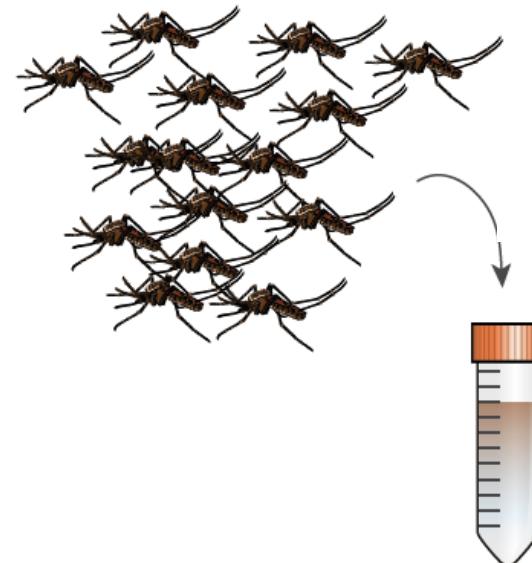
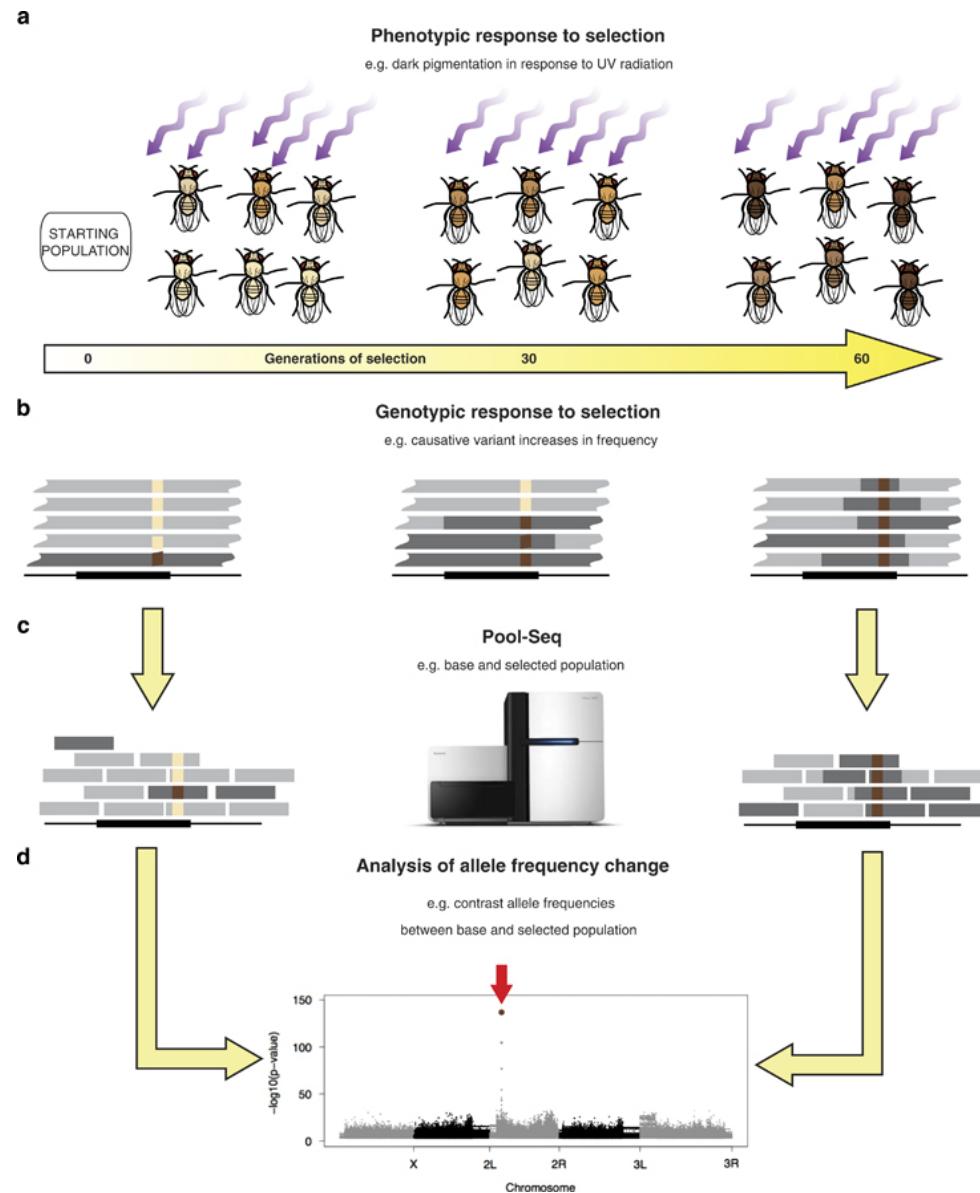


image: Magro et al (2006) Modern Path.

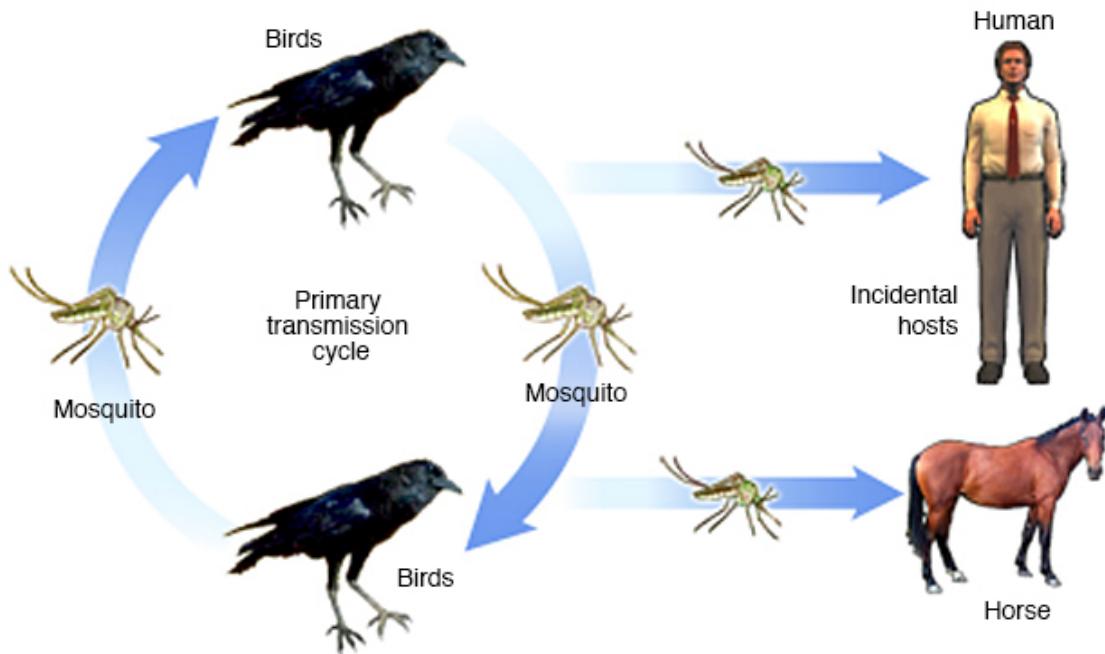
“Evolve & Resequence”



Schlötterer et al Heredity 2015

## Why study intrahost viral variation?

different population characteristics in different hosts



© MAYO FOUNDATION FOR MEDICAL EDUCATION AND RESEARCH. ALL RIGHTS RESERVED.

pre-existing variants may provide drug resistance or immune escape

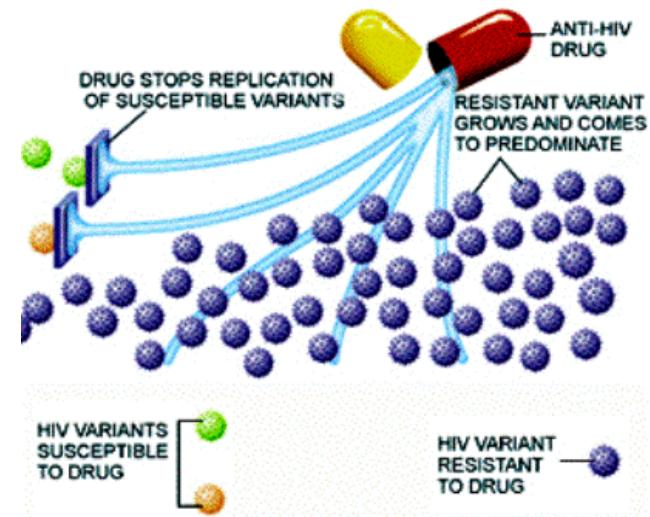
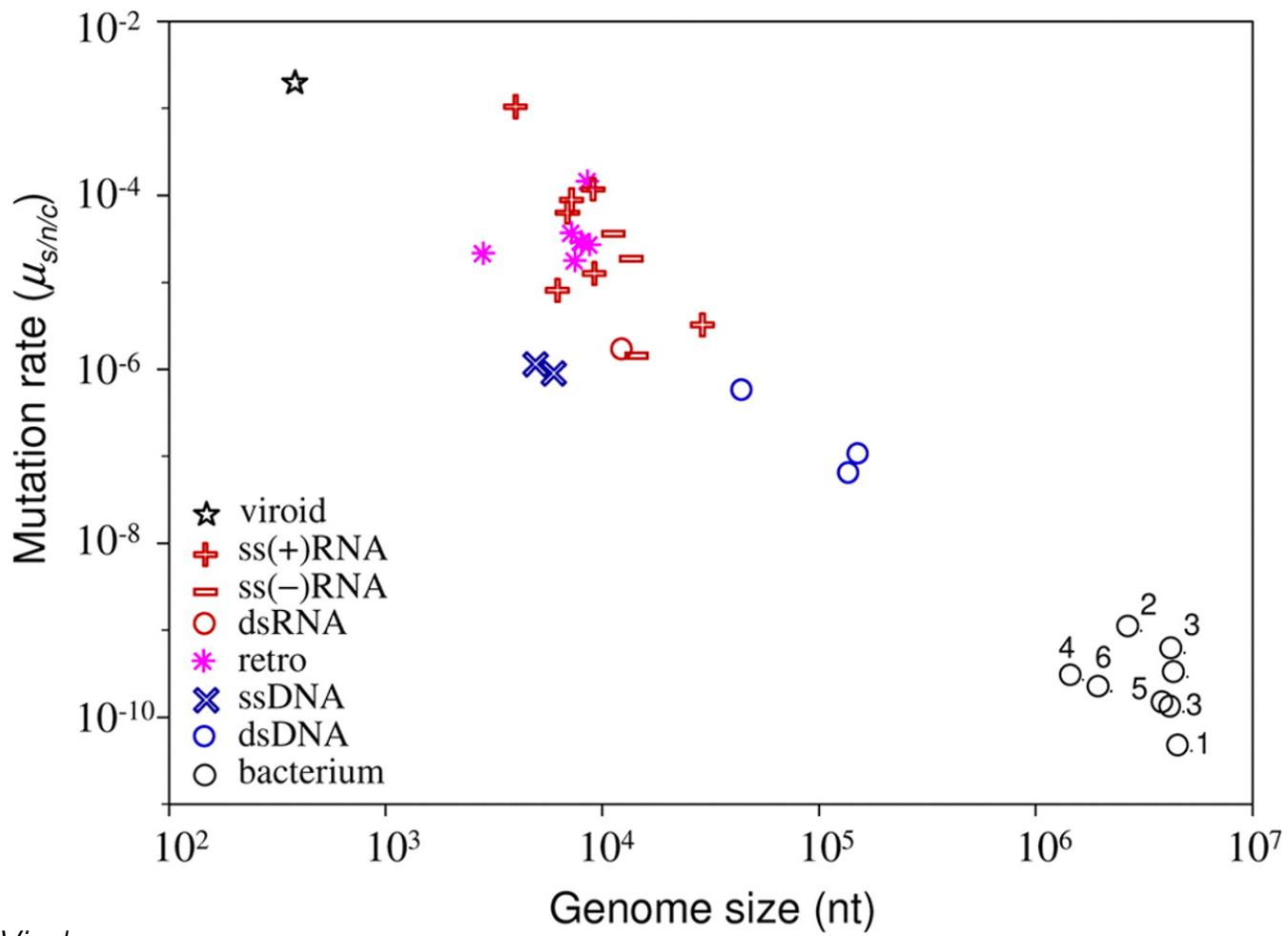
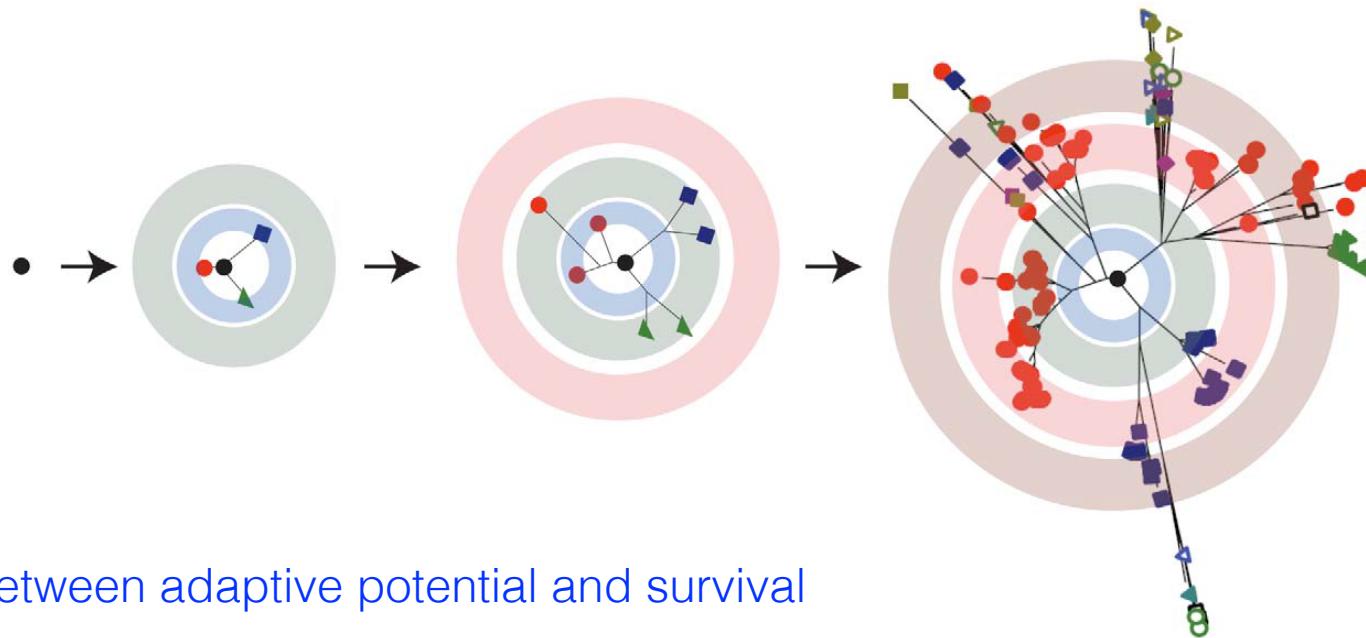


Image: Mayo clinic

(RNA) Viruses typically have error rates  $\approx 1 / \text{genome size}$



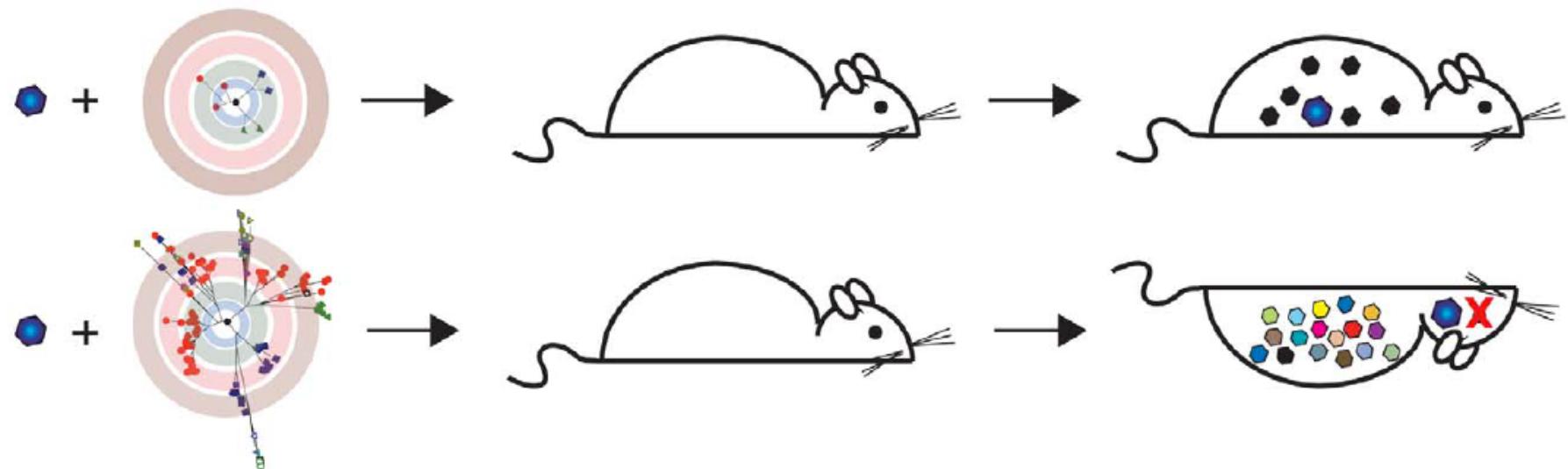
As a consequence of error-prone replication,  
intrahost virus populations can diversify rapidly



Balance between adaptive potential and survival

*Lauring and Andino (2010) PLoS Pathogens*

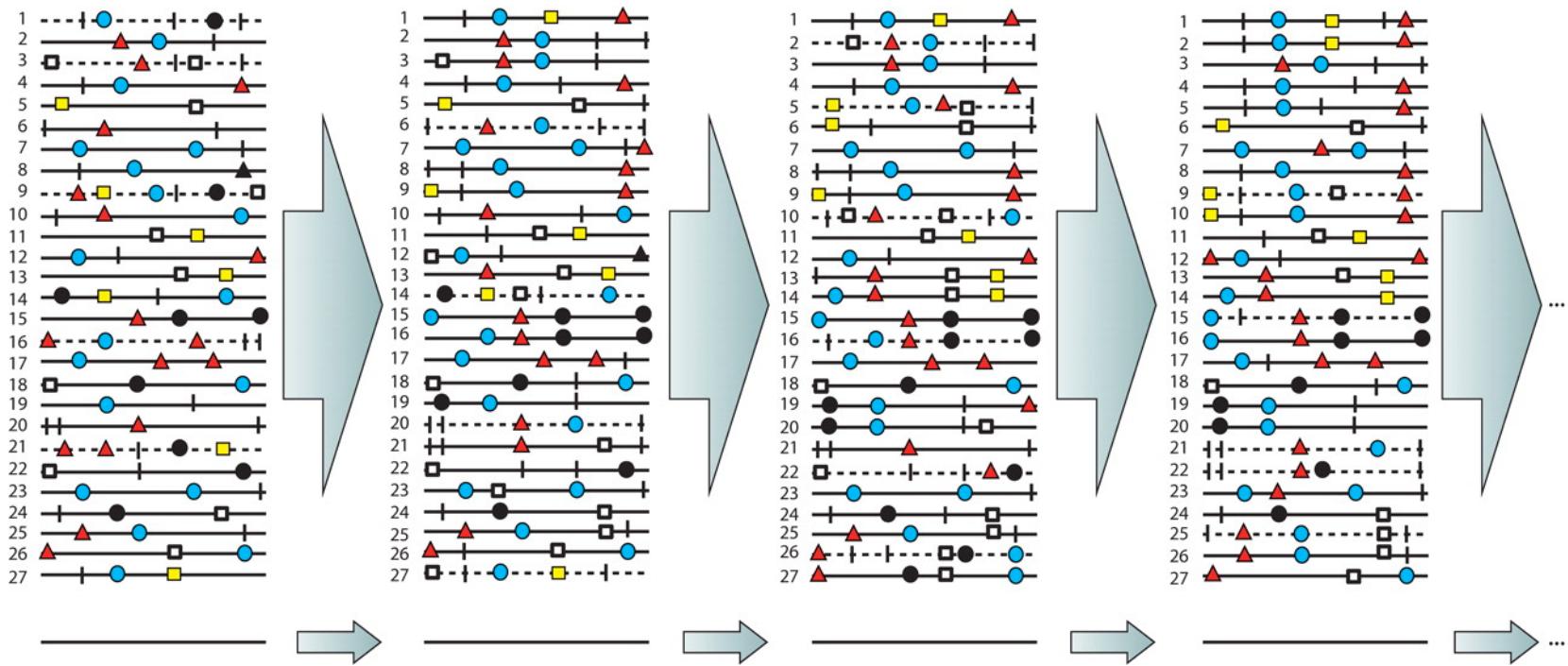
Intrahost viral population diversity can have a functional impact



*Lauring and Andino (2010) PLoS Pathogens*

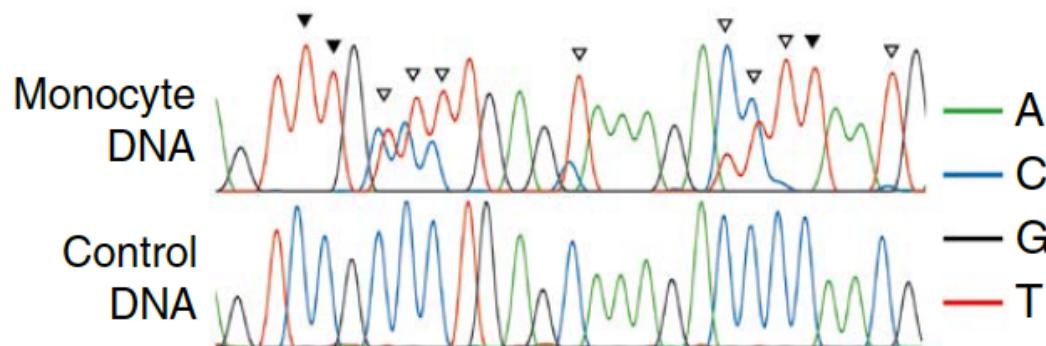
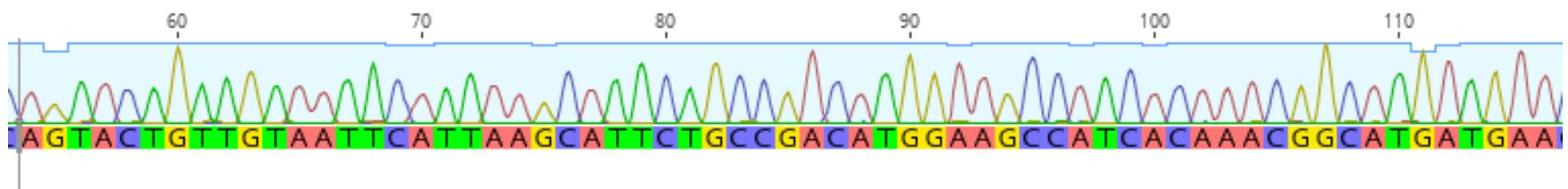
*Vignuzzi et al (2006) Nature*

The shifting ‘mutant swarm’ may not change consensus sequence



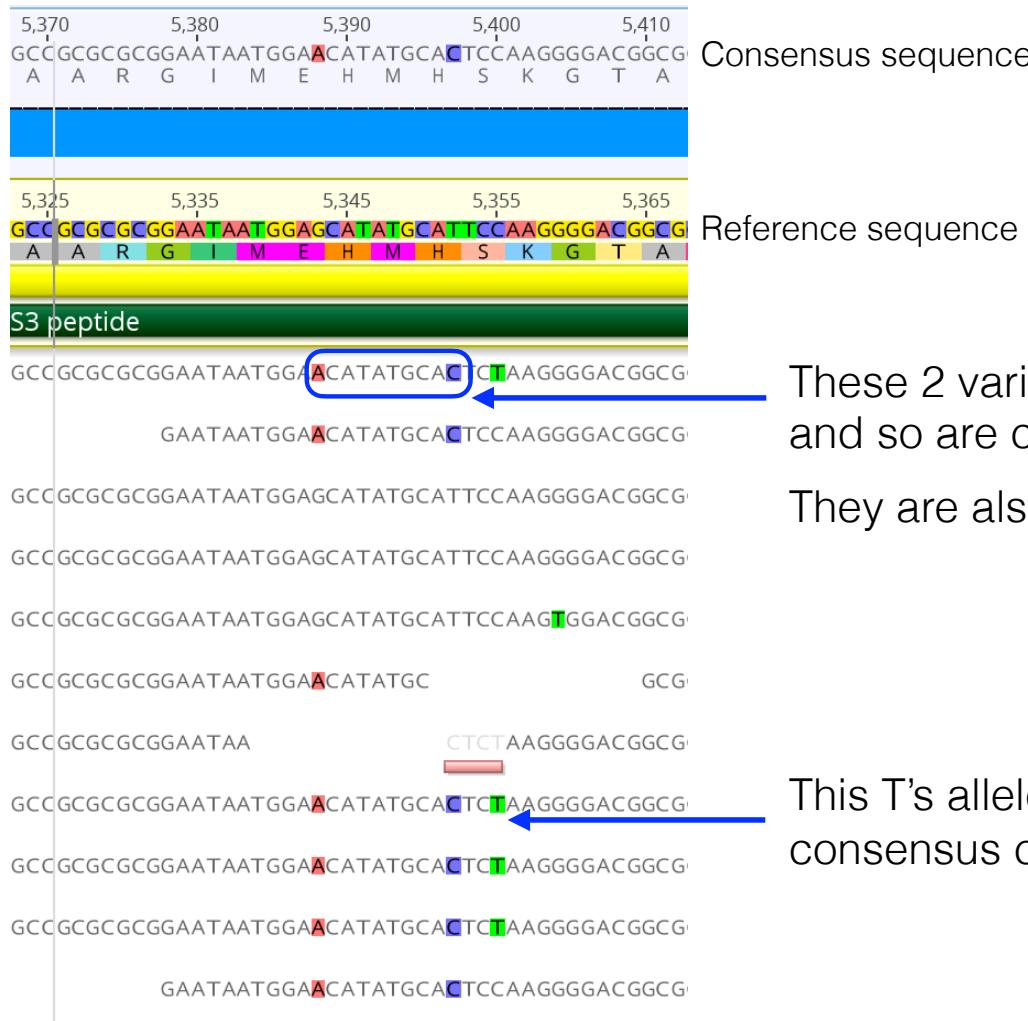
Domingo et al (2012) *Microbiol Mol Biol Reviews*

Sanger sequencing typically produces consensus sequence



It is possible to analyze chromatograms to obtain variant frequencies

Goal: identify variants, their frequencies, and potential functional impact



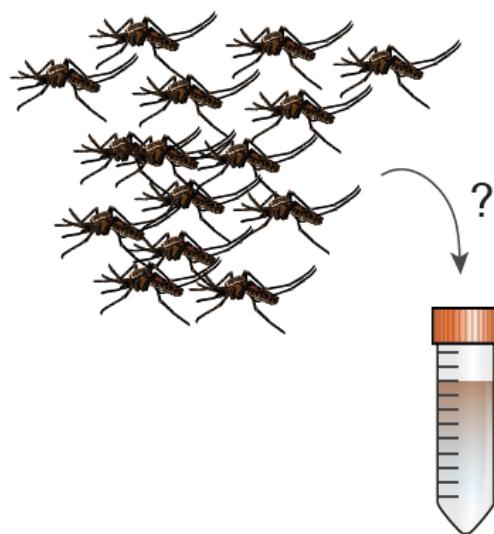
These 2 variants are at a >50% allele frequency  
and so are consensus changing variants

They are also both synonymous mutations

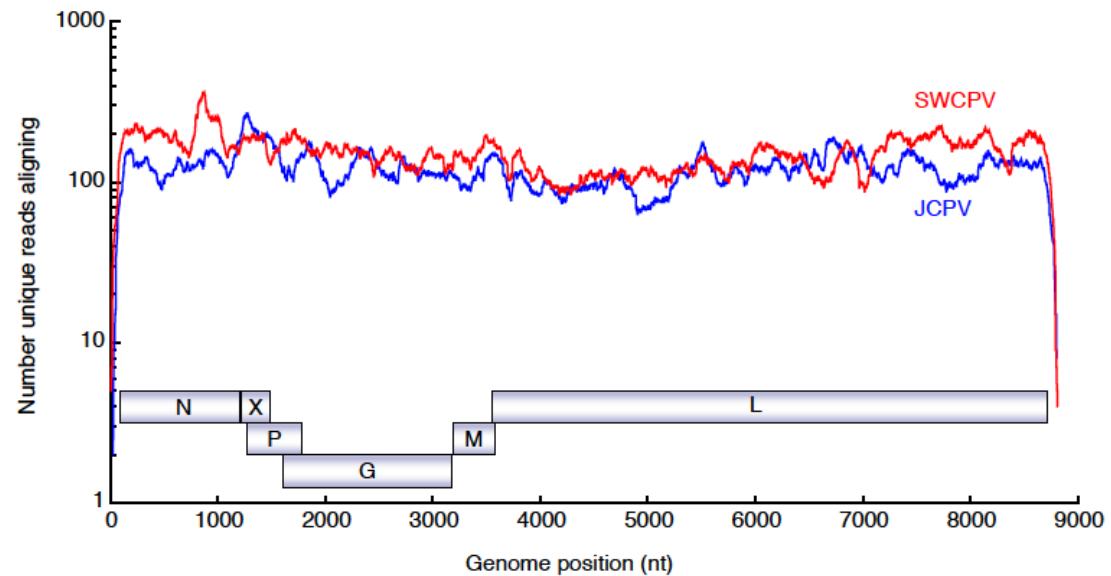
This T's allele frequency is 42% -> no  
consensus change

## Biological and technical limitations to the ability to detect rare variants

Pool size could limit the ability to detect rare variants



unlikely to observe variants with frequency < 1% in these datasets



carpet python bornaviruses



Distinguishing sequencing errors from true rare variants is a central challenge

GGAATAATGGA~~A~~CATATGCA~~C~~T~~C~~TAAAGGGGACGGCGCTCATGTAT

GAATAATGGA~~A~~CATATGCA~~C~~T~~C~~CCAAGGGGACGGCGCTCATGTAT

GGAATAATGGAGCATATGCATTCCAAGGGGACGGCGCTCATGTAT

GGAATAATGGAGCATATGCATTCCAAGGGGACGGCGCTCATGTAT

GGAATAATGGAGCATATGCATTCCAAG~~T~~GGACGGCGCTCATGTAT

GGAATAATGGA~~A~~CATATGC GCGCTCATGTAT

GGAATAA CT~~C~~TAAGGGGACGGCGCTCATGTAT  
  

GGAATAATGGA~~A~~CATATGCA~~C~~T~~C~~TAAAGGGGACGGCGCTCATGTAT

GGAATAATGGA~~A~~CATATGCA~~C~~T~~C~~TAAAGGGGACGGCGCTCATGTAT

GGAATAATGGA~~A~~CATATGCA~~C~~T~~C~~TAAAGGGGACGGCGCTCATGTAT

GAATAATGGA~~A~~CATATGCA~~C~~T~~C~~CCAAGGGGACGGCGCTCATGTAT

sequencing error, or real low frequency variant?

## Variant calling is also sensitive to mapping

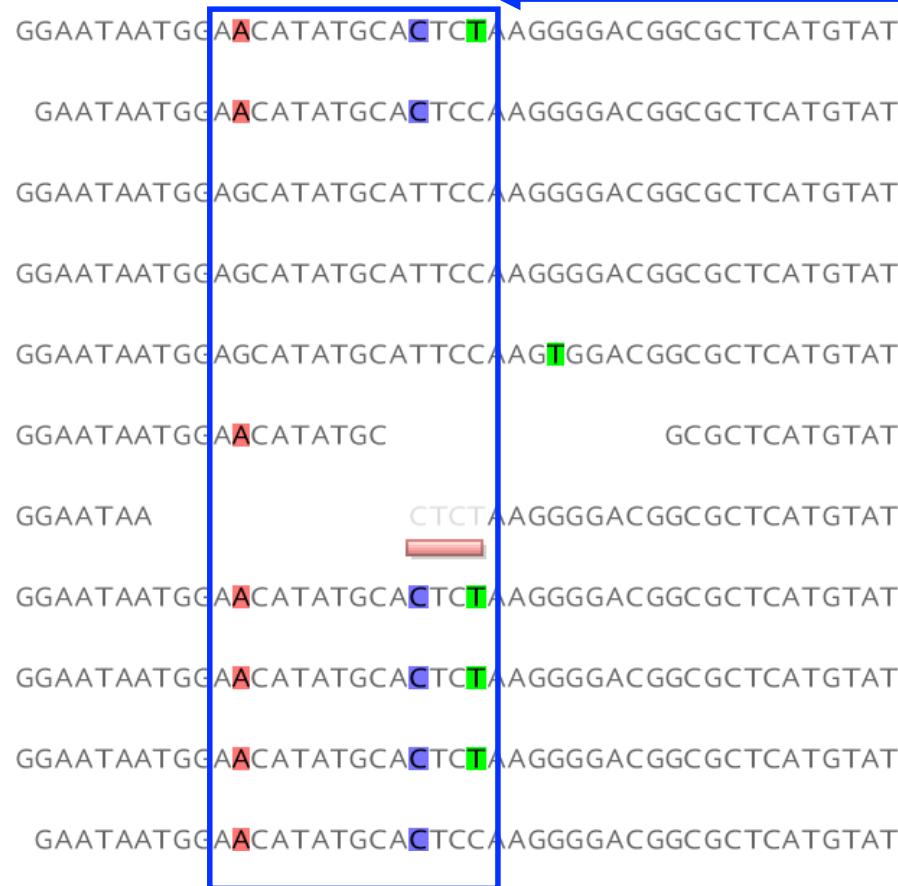
GGAATAATGGA~~A~~CATATGCA~~CT~~~~T~~AAGGGGACGGCGCTCATGTAT  
GAATAATGGA~~A~~CATATGCA~~CT~~~~T~~CCAAGGGGACGGCGCTCATGTAT  
GGAATAATGGAGCATATGCATTCCAAGGGGACGGCGCTCATGTAT  
GGAATAATGGAGCATATGCATTCCAAG~~T~~GGACGGCGCTCATGTAT  
GGAATAATGGA~~A~~CATATGC GCGCTCATGTAT  
GGAATAA  
GGAATAATGGA~~A~~CATATGCA~~CT~~~~T~~AAGGGGACGGCGCTCATGTAT  
GGAATAATGGA~~A~~CATATGCA~~CT~~~~T~~AAGGGGACGGCGCTCATGTAT  
GAATAATGGA~~A~~CATATGCA~~CT~~~~T~~CCAAGGGGACGGCGCTCATGTAT



These bases were soft-trimmed (not aligned), but they support variant basecalls

Different mapping software could well produce different results.

Another issue is linking or ‘phasing’ variants (haplotype reconstruction)



3 haplotypes evident here

**G T C** [reference sequence]

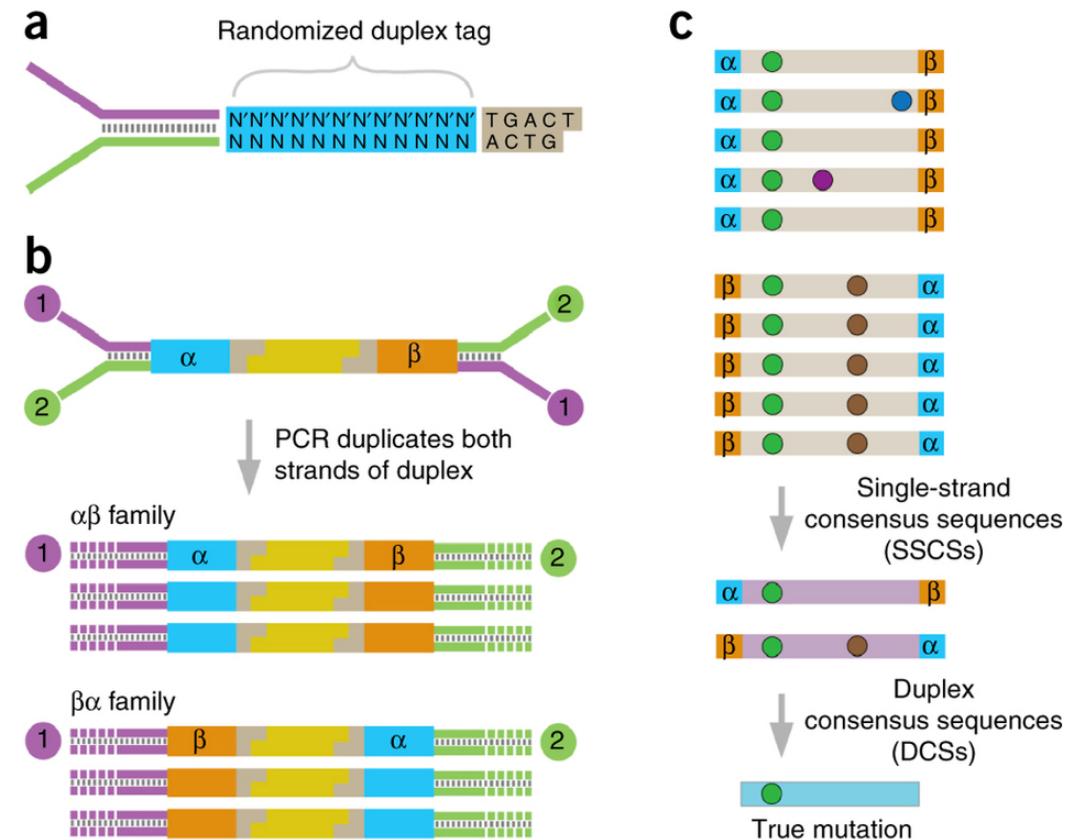
**A C C** [2 mutations]

**A C T** [3 mutations]

Much harder to link distant variants using short read data

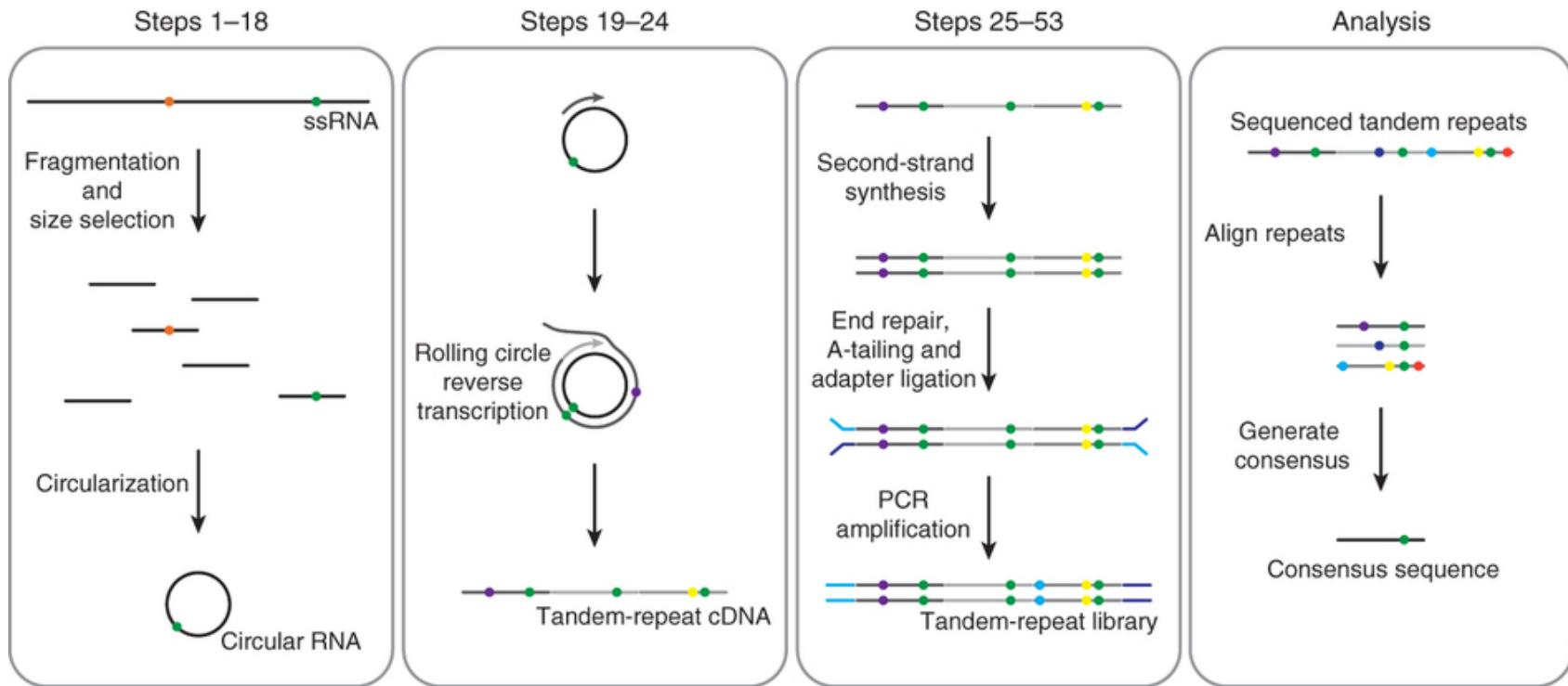
Several clever methods have been developed to get beyond the limit of detection due to sequencing errors

Some protocols take advantage of PCR duplicates to sequence the same original molecule multiple times

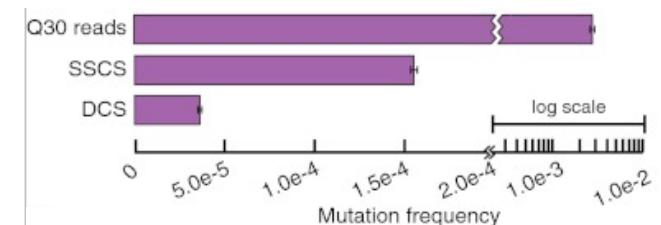
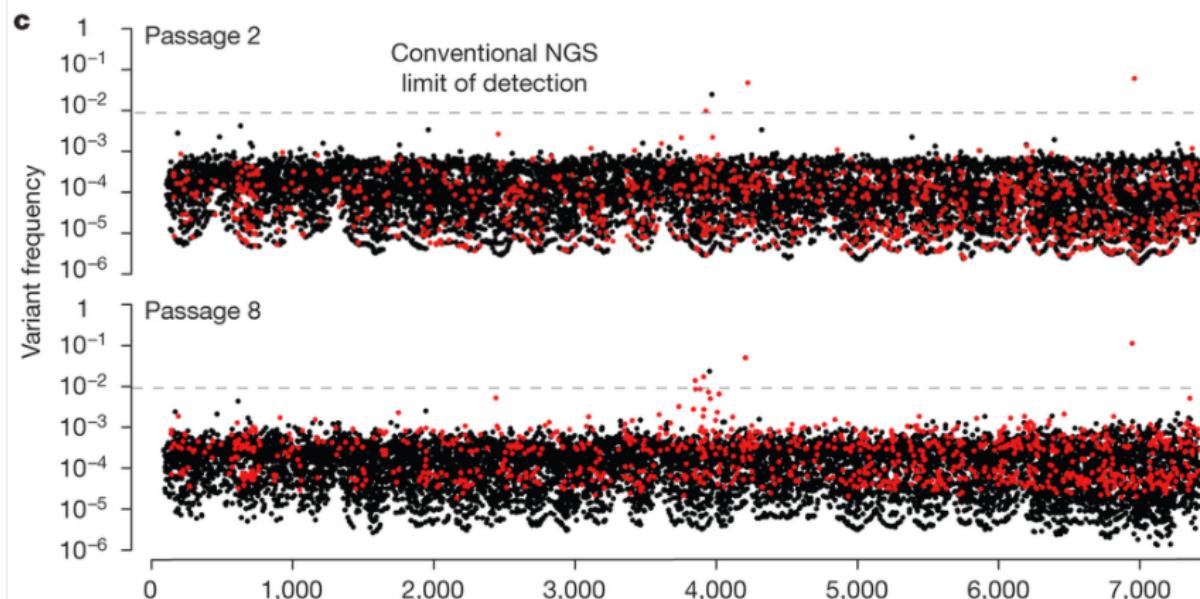


Kennedy et al (2014) Nature Protocols

CirSeq aims to measure lower frequency variants in RNA viruses



These methods aim to decrease variant frequency limit of detection



Acevedo *et al* (2014) *Nature*

Schmitt *et al* (2012) *PNAS*

These approaches have practical limitations

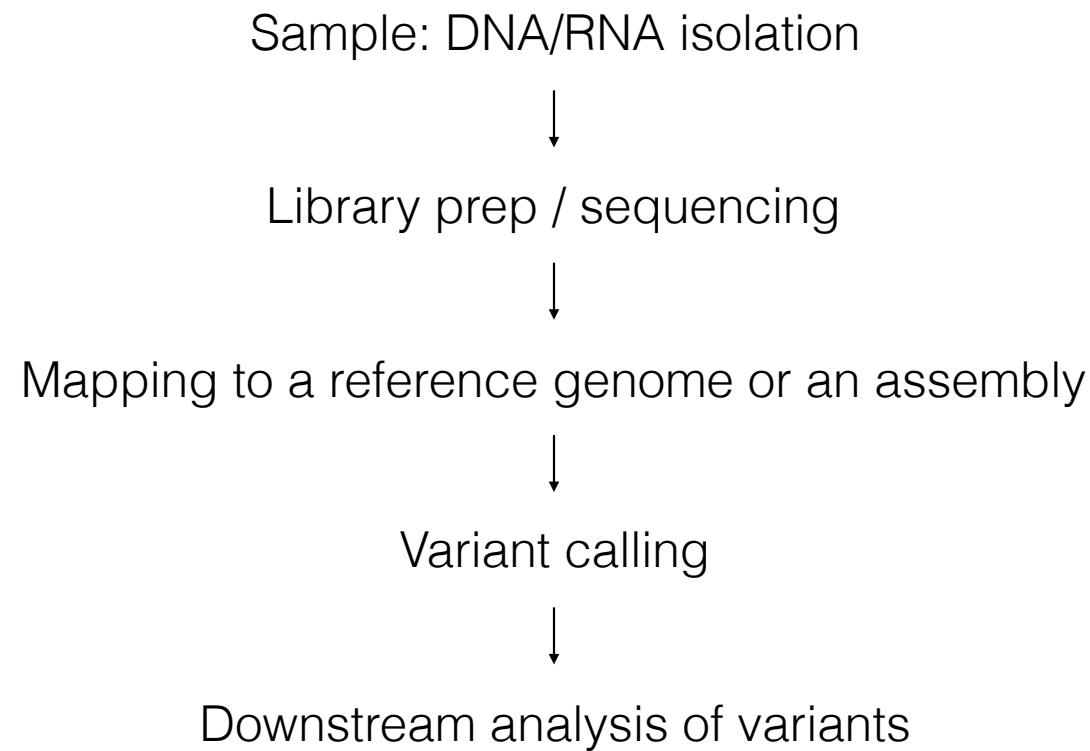
That's a lot of (poly-A) RNA

according to the manufacturer guidelines. Then 2–5 µg of poly(A)-containing RNA was fragmented with fragmentation reagent (Ambion) for 7.5 min at 70 °C. A practical minimum for this library preparation is 1 µg to ensure that enough fragmented RNA is obtained to produce a library with sufficient complexity and handle reproducibly. Approximately 80–90-base RNA fragments

The good news!

You don't necessarily or even often need linked variants or ultra low frequency variants to infer population genetic parameters (or otherwise answer your question of interest)

## A typical workflow for variant identification



The standard format for variant data is the vcf file (variant call format)

```
##fileformat=VCFv4.3
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT NA00001 NA00002 NA00003
20 14370 rs6054257 G A 29 PASS NS=3;DP=14;AF=0.5;DB;H2 GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51 1/1:43:5:...
20 17330 . T A 3 q10 NS=3;DP=11;AF=0.017 GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3 0/0:41:3
20 1110696 rs6040355 A G,T 67 PASS NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2 2/2:35:4
20 1230237 . T . 47 PASS NS=3;DP=13;AA=T GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51 0/0:61:2
20 1234567 microsat1 GTC G,GTCT 50 PASS NS=3;DP=9;AA=G GT:GQ:DP 0/1:35:4 0/2:17:2 1/1:40:3
```

<https://github.com/samtools/hts-specs>

## A couple reviews to get you started

Schlötterer et al (2014) *Nat Rev Gen*  
doi:10.1038/nrg3803

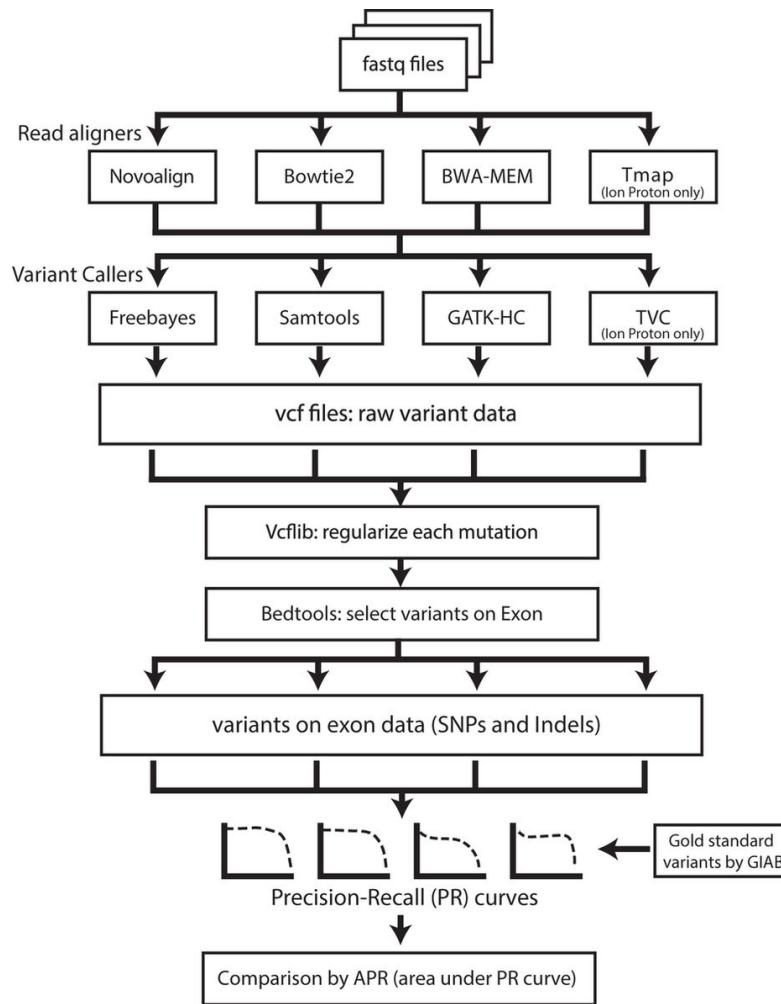
Posada-Cespedes et al (2016) *Virus Res*  
doi:10.1016/j.virusres.2016.09.016

Table 2 | To pool or not to pool?

Scenario	Pool-seq recommended?
Small sample size (<40 individuals)	Yes, but only appropriate when carried out on genomic windows containing multiple SNPs instead of on individual SNPs
Phenotypes of individuals are or will be available	RAD-seq of individuals is probably better suited for many cases
Linkage disequilibrium is key to data analysis	RAD-seq of individuals is probably better suited for many cases
High confidence about low-frequency SNPs is needed	Not with current protocols; sequencing of individuals is preferred
Simple population genetic analyses, such as population differentiation or average heterozygosity	Yes, but when coverage is low it results in a lower confidence of the allele frequency estimate of individual SNPs
Identification of selective sweeps	Yes, but only limited information about linkage disequilibrium can be obtained
Time series with large sample sizes and many replicates	Yes
Mapping of induced mutations	Yes, identification of the causative site is possible
GWAS	Yes, provided that replicates and large pool sizes are available, but other approaches should also be considered
QTL mapping	Yes, but no effect sizes are estimated
Intraspecific polymorphism of bacterial and viral populations	Yes
Information about dominance and effect size is important	No
Cancer	Pool-seq is a natural approach to analyse the cell population

GWAS, genome-wide association study; QTL, quantitative trait locus; RAD-seq, restriction-site-associated DNA sequencing; SNP, single-nucleotide polymorphism.

Several papers fairly recently compared variant calling software



Fairly good overlap from different pipelines using Illumina data

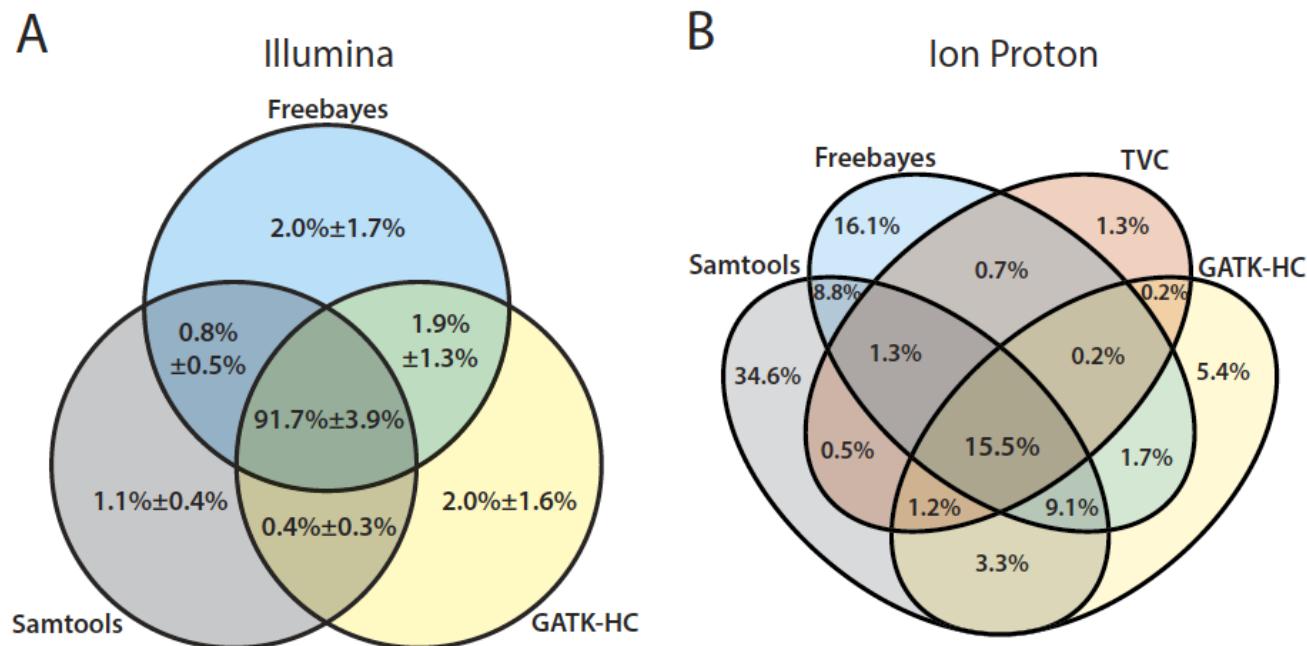


Figure 3. Venn diagrams summarizing called variants by different callers. The mean percentage with

Different s/w had different biases  
Avoid using Ion Proton data for variant calling

We'll call some variants later this morning

GGAATAATGGA **A**CATATGCA **CTCT** AAGGGGACGGCGCTCATGTAT

GAATAATGGA **A**CATATGCA **CTCC** AAGGGGACGGCGCTCATGTAT

GGAATAATGGAGCATATGCATTCCAAGGGGACGGCGCTCATGTAT

GGAATAATGGAGCATATGCATTCCAAGGGGACGGCGCTCATGTAT

GGAATAATGGAGCATATGCATTCCAAG**T**GGACGGCGCTCATGTAT

GGAATAATGGA **A**CATATGC                                    GCGCTCATGTAT

GGAATAA    **CTCT** AAGGGGACGGCGCTCATGTAT  
   

GGAATAATGGA **A**CATATGCA **CTCT** AAGGGGACGGCGCTCATGTAT

GGAATAATGGA **A**CATATGCA **CTCT** AAGGGGACGGCGCTCATGTAT

GGAATAATGGA **A**CATATGCA **CTCT** AAGGGGACGGCGCTCATGTAT

GAATAATGGA **A**CATATGCA **CTCC** AAGGGGACGGCGCTCATGTAT