Measuring genetic variation

There are many ways to summarize molecular variation, and therefore many different *statistics* that describe the data.

A *statistic* is a summary of the sample of observations (the DNA sequences).

In molecular population genetics we often focus on statistics that are estimators of the theoretical parameter θ.

$\theta \equiv 4Ne\mu$ = represents the amount of variation found at autosomal loci in hypothetical populations in which all variation is neutral and at mutation-drift equilibrium.

Key statistics for measuring genetic diversity:

Heterozygosity

Nucleotide diversity (sometimes called $\theta\pi$)

Watterson’s theta θw

Tajima’s D, a comparison of $\theta\pi$ and θw

Tajima’s D **compares different populations** by evaluating whether observed genetic diversity follows neutral expectations or has been influenced by **demographic events** or **selection**. The sign and magnitude of **D** help infer population history

$D = 0$ neutral evolution

$D < 0$ excess of rare variants

$D > 0$ excess of intermediate frequency variants

How to measure genetic diversity (from NGS data):

ANGSD: Pro — incorporates genotype likelihoods, so great for low-coverage data. Con — difficult to use

https://www.popgen.dk/angsd/index.php/ANGSD

input is BAM files (mapped reads)

How to measure genetic diversity (from NGS data):

ANGSD: Pro — incorporates genotype likelihoods, so great for low-coverage data. Con — difficult to use

https://www.popgen.dk/angsd/index.php/ANGSD

input is BAM files (mapped reads)

pixy: Much more user-friendly! “Software for painlessly estimating average nucleotide diversity within and between populations”

https://github.com/ksamuk/pixy

π, θ_W, Tajima’s D

d_{xy}, F_{ST} →

in windows across the genome — for genome-wide averages, vcftools; hierfstat or SNPRelate in R; other scripts. Note: Different algorithms behave very differently with small sample sizes.

needs a VCF containing invariant sites

Detecting adaptation

Genotype-environment analysis (GEA)

What variants are associated with an environmental gradient?

Genome scan approaches

Which variants show evidence for being under selection?

Genome-wide association studies (GWAS)

Which variants are associated with a quantitative trait?

Detecting adaptation

Genotype-environment analysis (GEA)

What variants are associated with an environmental gradient?

—> Environment is the predictor; genotype is the response

Genome scan approaches

Which variants show evidence for being under selection?

Genome-wide association studies (GWAS)

Which variants are associated with a quantitative trait?

—> Genotype is the predictor; phenotype is the response

Genotype-environment analysis (GEA)

What variants are associated with an environmental gradient?

LFMM (Latent Factor Mixed Models) — regresses genotype on an environmental variable, taking into account population structure

Genotype-environment analysis (GEA)

What variants are associated with an environmental gradient?

LFMM (Latent Factor Mixed Models) — regresses genotype on an environmental variable, taking into account population structure

RDA (Redundancy Analysis) — multivariate ordination relating environmental variables and allele frequencies. Good for polygenic selection.

Often these are used in combination!

Genotype-environment analysis (GEA)

What variants are associated with an environmental gradient?

LFMM (Latent Factor Mixed Models) — regresses genotype on an environmental variable, taking into account population structure.

RDA (Redundancy Analysis) — multivariate ordination relating environmental variables and allele frequencies. Good for polygenic selection.

Often these are used in combination!

Key issues:

These methods will (almost) always give you loci! You should be highly skeptical of them! They are candidate loci = hypotheses.

A single GEA is not proof that loci are involved in adaptation — what are ways to verify this?

Correlated environmental variables and population structure can lead to false positives

How to do GEAs:

There is a great tutorial on lfmm and RDA by Brenna Forester associated with the Landscape Genetics course (see link on our Github site). This can all be done in R.

$$\text{Genotype}_i \sim f(\text{Environment}_i) + \text{structure}$$

Genome scan approaches

Which variants show evidence for being under selection?

Back in the day, lots of work simply looked for Fst outliers and told stories about these loci being under selection (e.g. approaches like BayeScan)

Now there is a lot of skepticism about that approach

Genome scan approaches

Which variants show evidence for being under selection?

Back in the day, lots of work simply looked for Fst outliers and told stories about these loci being under selection (e.g. approaches like BayeScan)

Now there is a lot of skepticism about that approach

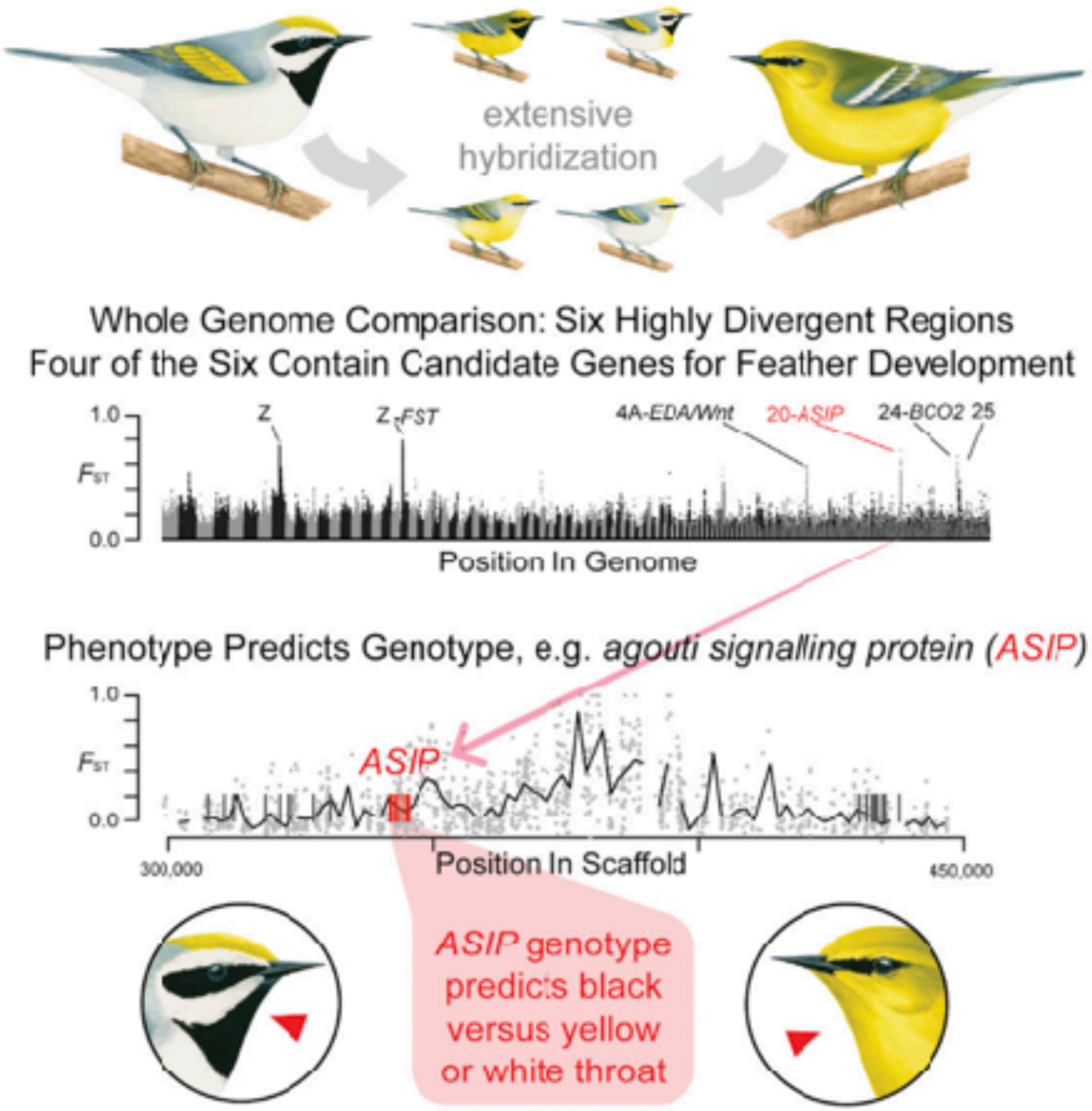
Fst is a *relative measure* (it depends on genetic diversity)

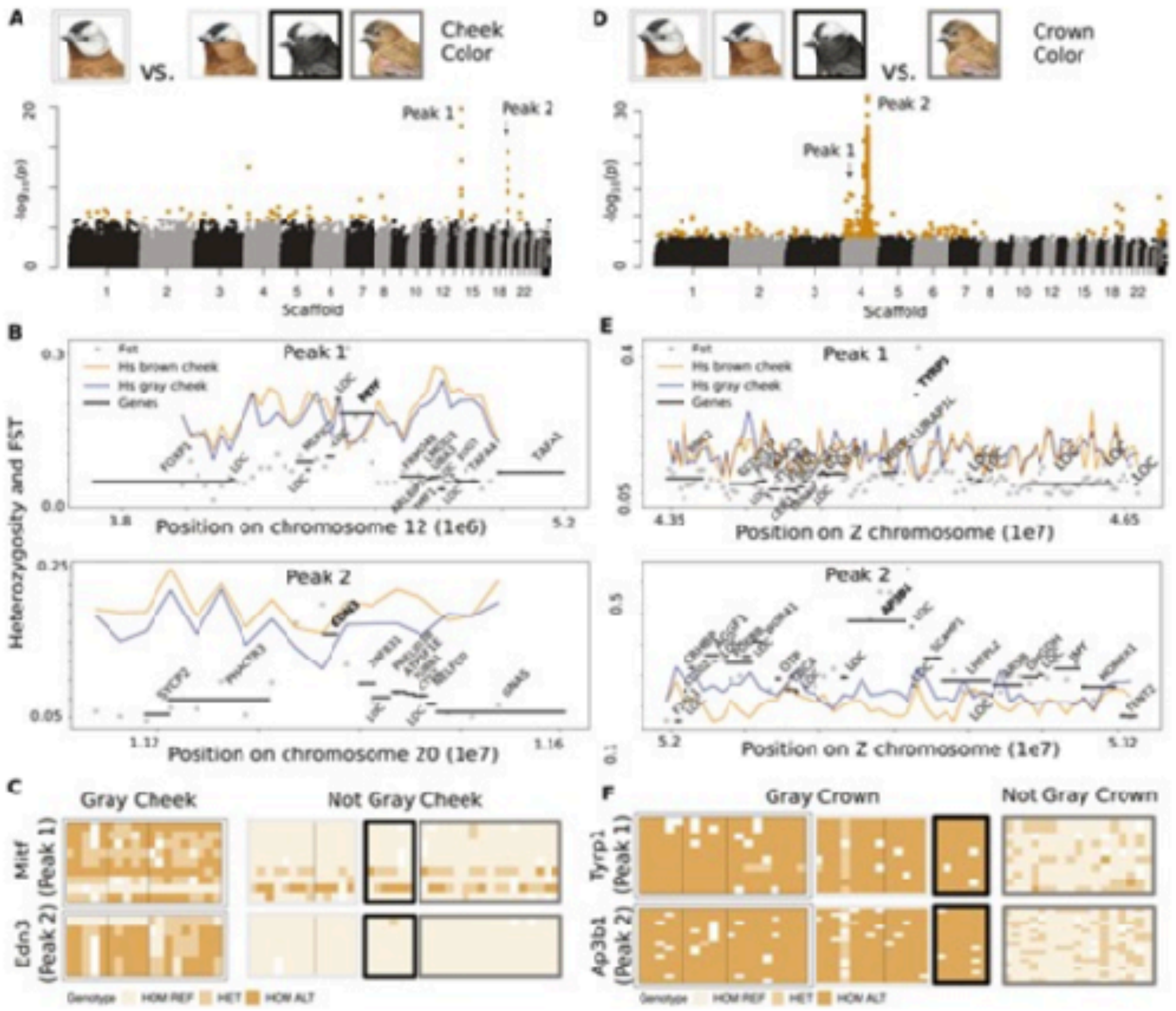
At minimum, we also want to calculate a measure of absolute divergence, e.g. Dxy

Sometimes these approaches when used in sliding windows across the genome *will* identify clear regions that are plausibly under selection. (Many more times, the situation is not that simple).

Plumage Genes and Little Else Distinguish the Genomes of Hybridizing Warblers (Toews et al 2016)

The genetic basis of plumage coloration and elevation adaptation in a clade of recently diverged alpine and arctic songbirds (Funk et al 2022)





How to do genome scans:

It is straightforward to calculate Fst / Dxy in sliding windows in pixy

How to do genome scans:

It is straightforward to calculate Fst / Dxy in sliding windows in pixy

Population structure / demography can be problematic and needs to be accounted for

Can simulate a neutral situation in e.g. fastsimcoal, compare to this expectation

How to do genome scans:

It is straightforward to calculate Fst / Dxy in sliding windows in pixy

Population structure / demography can be problematic and needs to be accounted for

Can simulate a neutral situation in e.g. fastsimcoal, compare to this expectation

As with GEA, loci are *candidates* = hypotheses needing further validation

Genome-wide association studies (GWAS)

Which variants are associated with a quantitative trait?

$$\text{Phenotype}_i \sim f(\text{Genotype}_{ij}) + \text{covariates}$$

How to do GWAS:

A MAF filter is good here — rare variants decrease power

It is essential to account for population structure

Linear mixed models are commonly used, among these GEMMA is frequently used

Must account for multiple tests (e.g. FDR correction)

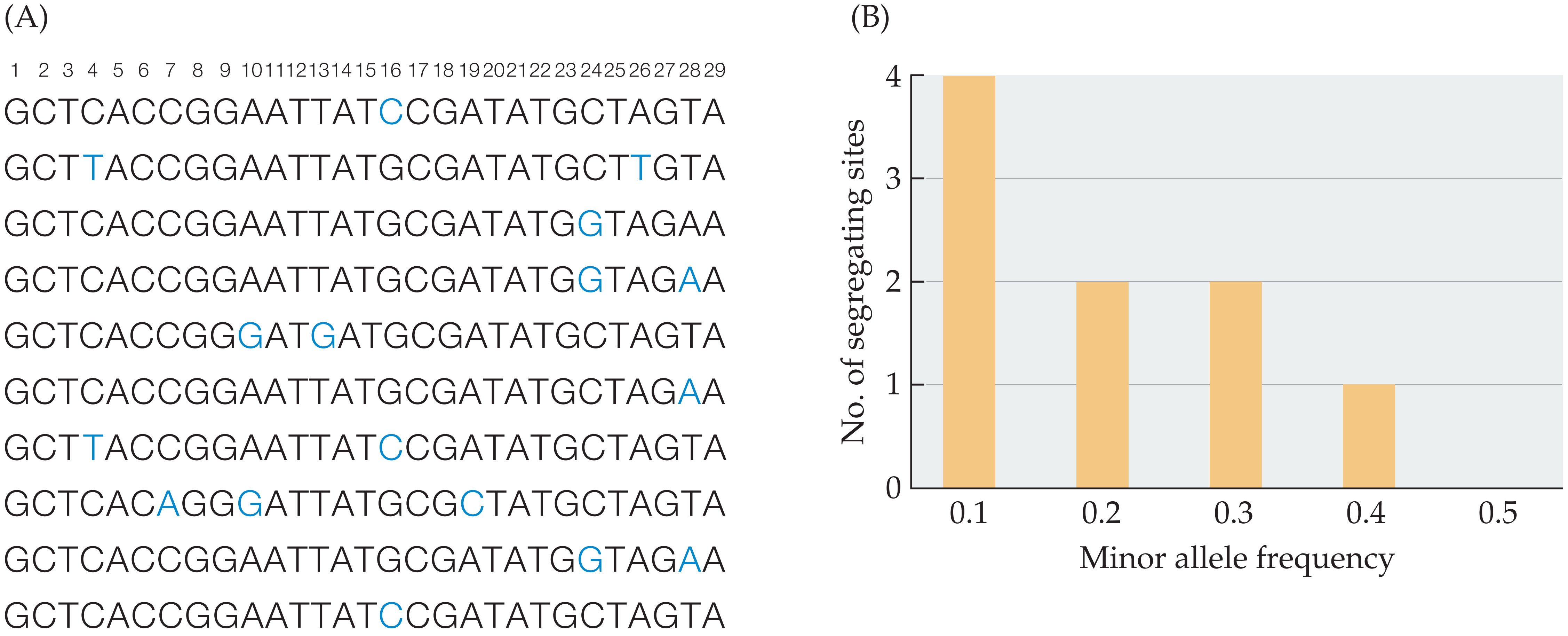
Power is limited without large sample sizes

To be a broken record: loci are *candidates* = hypotheses needing further validation

The site frequency spectrum

A graphical way to describe nucleotide variation

The minor allele frequency (MAF) is the frequency of the less common allele - it range from 1/n to 0.5.



```
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29
GCTCACCGGAATTATCCGATATGCTAGTA
GCTTACCGGAATTATGCGATATGCTTGTA
GCTCACCGGAATTATGCGATATGGTAGAA
GCTCACCGGAATTATGCGATATGGTAGAA
GCTCACCGGGATGATGCGATATGCTAGTA
GCTCACCGGAATTATGCGATATGCTAGAA
GCTTACCGGAATTATCCGATATGCTAGTA
GCTCACAGGGATTATGCGCTATGCTAGTA
GCTCACCGGAATTATGCGATATGGTAGAA
GCTCACCGGAATTATCCGATATGCTAGTA
```

FIGURE 3.2 (A) For each segregating site the minor allele is blue and the major allele is black. For this sample, $n = 10$, $L = 29$, and $S = 9$. (B) The "folded" allele frequency spectrum for the alignment shown in (A).

Heterozygosity

$$h = \frac{n}{n-1}\left(1 - \sum p_i^2\right) \tag{3.1}$$

n = the number of sequences in a sample

Nucelotide diversity

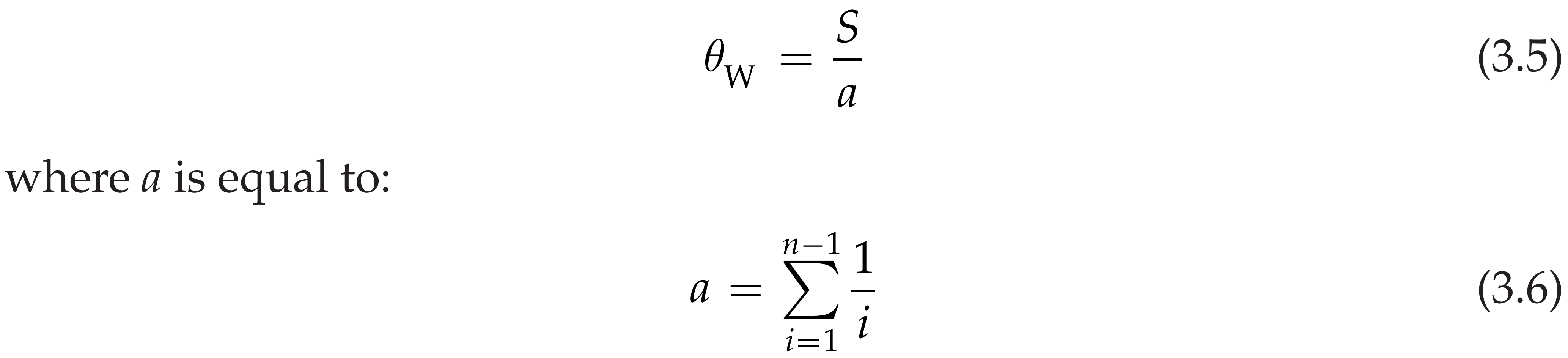
$$\pi = \sum_{j=1}^{S} h_j \tag{3.2}$$

S = the number of segregating sites

h = heterozygosity

Under the infinite sites model for a diploid Wright-Fisher population at equilibrium, $E(\pi) = \theta$, which is why this statistic is sometimes called $\theta\pi$

Watterson’s theta



$$\theta_W = \frac{S}{a} \qquad (3.5)$$

where a is equal to:

$$a = \sum_{i=1}^{n-1} \frac{1}{i} \qquad (3.6)$$

S = total number of segregating sites

- a arises from coalescent theory and corrects for the fact that larger sample sizes increase the number of expected segregating sites in a non-linear way.
- It represents the **expected total branch length** of a neutral genealogy in a sample of size n under the **standard neutral model** (Wright-Fisher with neutrality and constant population size).

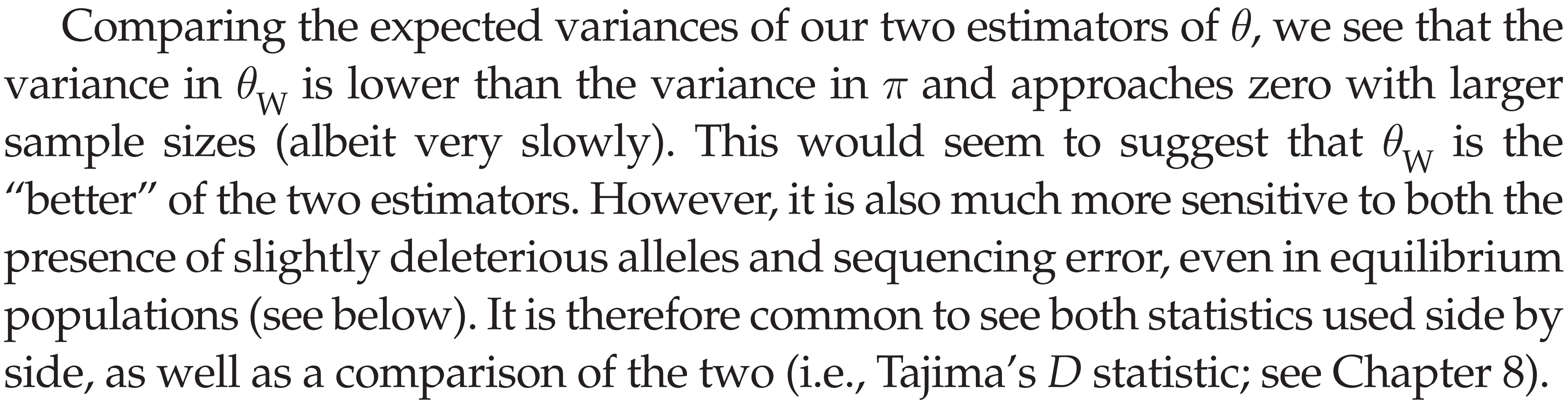
For example:

For a sample of n = 5 sequences:

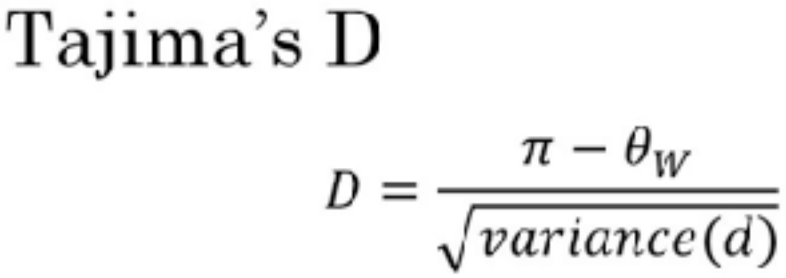
a = 1/1+1/2+1/3+1/4 = 1 + 0.5 + 0.333 + 0.25 = 2.083

If S = 10 segregating sites were observed:

ThetaW = 10/2.083 = 4.8



Comparing the expected variances of our two estimators of θ, we see that the variance in θ_W is lower than the variance in π and approaches zero with larger sample sizes (albeit very slowly). This would seem to suggest that θ_W is the "better" of the two estimators. However, it is also much more sensitive to both the presence of slightly deleterious alleles and sequencing error, even in equilibrium populations (see below). It is therefore common to see both statistics used side by side, as well as a comparison of the two (i.e., Tajima's *D* statistic; see Chapter 8).



Tajima's D

$$D = \frac{\pi - \theta_W}{\sqrt{variance(d)}}$$

Tajima's D **compares different populations** by evaluating whether observed genetic diversity follows neutral expectations or has been influenced by **demographic events** or **selection**. The sign and magnitude of **D** help infer population history

D = 0 neutral evolution

D < 0 excess of rare variants

D > 0 excess of intermediate frequency variants