A review of genetic variation

Karl Schmid

Institute of Plant Breeding, Seed Science and Population Genetics
University of Hohenheim

November 26, 2012

Background

Measures of genetic diversity

Frequency distribution of polymorphisms

Genetic diversity of different site types

Haplotype-based measures of genetic diversity

Variation of genetic diversity in the genome

Challenges arising from genome sequencing (NGS)

Background

Measures of genetic diversity

Frequency distribution of polymorphisms

Genetic diversity of different site types

Haplotype-based measures of genetic diversity

Variation of genetic diversity in the genome

Challenges arising from genome sequencing (NGS)

Key parameters of populations

- Level of polymorphism (Nucleotide diversity, $\pi = 4N_e\mu$)
- Allele frequency (Tajima's D)
- Haplotype structure (Extended haplotype homozygosity, EHH)
- Recombination rate and breakdown of linkage disequilibrium

Background

Measures of genetic diversity

Frequency distribution of polymorphisms

Genetic diversity of different site types

Haplotype-based measures of genetic diversity

Variation of genetic diversity in the genome

Challenges arising from genome sequencing (NGS)

Measures of genetic diversity are based on several criteria

Number of variants:

- Proportion of polymorphic sites (or loci)
- Richness of allelic variants (A)
- Average number of alleles per locus

Frequency of variants:

- Effective number of alleles (A_e)
- Average expected heterozygosity (H_e); Nei's genetic diversity

Groups of variants (haplotypes):

- Haplotype number
- Haplotype frequency and richness
- Extended haplotype homozygosity (EHH)

Gene diversity or Hardy-Weinberg heterozygosity

- Unbiased measure:

$$H_E = 1 - \sum_{i=1}^{n} p_i^2$$

  where p_i is the frequency of allele i.

- For small sample sizes ($n < 50$), the following correction is recommended:

$$H_E = \frac{n}{n-1}(1 - \sum_{i=1}^{n} \hat{p}_i^2)$$

- Note: Gene diversity is identical to the sum of the frequency of heterozygous genotypes in HWE.

Analysis of multiple markers

- Average gene diversity or mean heterozygosity:

$$\hat{H} = \frac{1}{m} \sum_{i=1}^{m} \hat{H}_{E_i}$$

- Haplotype: The combination of alleles on a chromosome in a given individual
- Haplotype count: Number of different haplotypes formed by the markers

Expected heterozygosity under Mutation-Drift equilibrium

The expected heterozygosity of a population is the product of

- gain of new variation by mutation
- loss of existing variation by genetic drift

Under a mutation-drift equilibrium:

$$H_e \approx \frac{4N_e\mu}{1+4N_e\mu} = \frac{\theta}{1+\theta}$$

θ is often thought as the scaled mutation rate, i.e. the mutation rate of the whole population.

Quantification of DNA sequence variation

- Nei's gene diversity is not practical for DNA sequences because H approaches 1 for a locus of reasonable length. (But it is still practical for individual -unlinked- SNPs!)
- Hence alternative measures are used
- Nucleotide diversity
- Nucleotide polymorphism

Quantification of DNA sequence variation

Nucleotide diversity: Two randomly chosen nucleotide sequences differ at a given site

$$\theta \approx \pi = \frac{n}{n-1} \sum_{i=1}^{k} \sum_{j=1}^{k} p_i p_j \pi_{ij}$$

where

- n: number of individuals
- k: number of haplotypes
- p_i: proportion of haplotype i
- π_{ij} as pairwise differences of haplotypes i and j. The difference is measured as the proportion of different SNPs between two haplotypes. If 4 out of 10 sites are different between to haplotypes, $p_{ij} = 0.4$

Quantification of DNA sequence variation

Nucleotide polymorphism:

$$P_n = \frac{n_P}{n_t}$$

- n_P: number of polymorphic nucleotide positions
- n_t: total number of sequenced nucleotide positions

Quantification of DNA sequence variation

Nucleotide polymorphism:

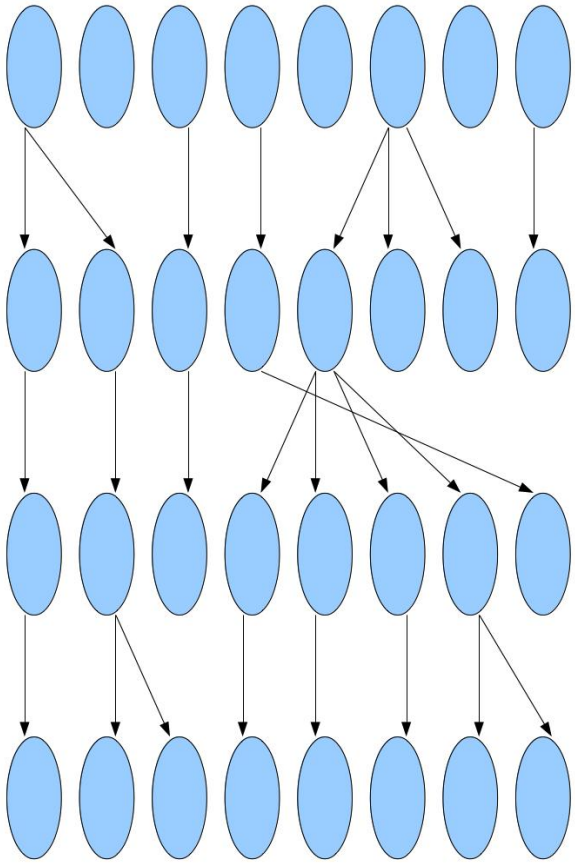
$$P_n = \frac{n_P}{n_t}$$

- n_P: number of polymorphic nucleotide positions
- n_t: total number of sequenced nucleotide positions
- To estimate $\theta = 4N_e\mu$:

$$\theta_W = \frac{S}{\sum_{i=1}^{n-1} \frac{1}{i}}$$

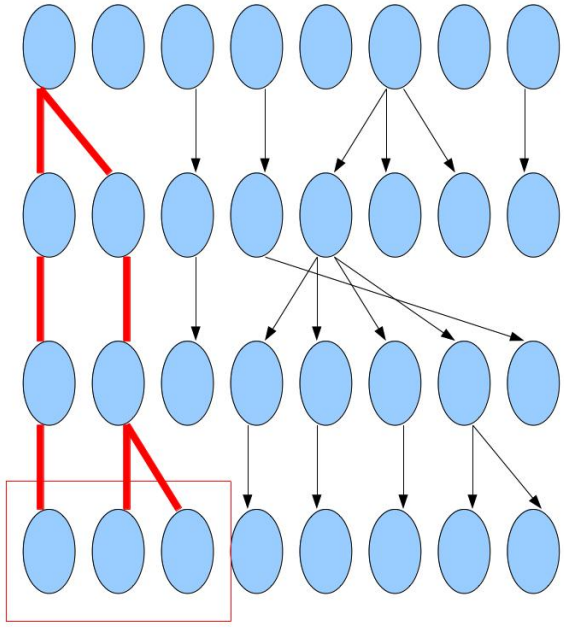
(Watterson 1975)

The Fisher-Wright model reloaded



- Population model for $2N$ alleles
- Random sampling of alleles for the next generation in the present generation
- "Each child picks its parent randomly"
- Probability that two alleles have the same parent allele: $\frac{1}{2N}$
- What is the dynamics of a sample of n alleles from that population backwards in time?
- Samples are small: $n \ll N$

Discrete coalescent



Discrete coalescent, sample size n=3,
population size 2N=8

- We're interested in the genealogy/ancestral tree of the sample of alleles!
- From now on, we measure time backwards

What is coalescence?

- Coalescence event: Two (or more) alleles have the same parent allele in the previous generation

What is coalescence?

- Coalescence event: Two (or more) alleles have the same parent allele in the previous generation
- Probability of coalescent event for two specific alleles in the previous generation: $1/2N$

What is coalescence?

- Coalescence event: Two (or more) alleles have the same parent allele in the previous generation
- Probability of coalescent event for two specific alleles in the previous generation: $1/2N$
- Probability of coalescent event for three or more alleles: much smaller (ignore it)

What is coalescence?

- Coalescence event: Two (or more) alleles have the same parent allele in the previous generation
- Probability of coalescent event for two specific alleles in the previous generation: $1/2N$
- Probability of coalescent event for three or more alleles: much smaller (ignore it)
- Sample of n alleles: How many different allele pairs are possible?

$$n(n-1)/2$$

What is coalescence?

- Coalescence event: Two (or more) alleles have the same parent allele in the previous generation
- Probability of coalescent event for two specific alleles in the previous generation: $1/2N$
- Probability of coalescent event for three or more alleles: much smaller (ignore it)
- Sample of n alleles: How many different allele pairs are possible?

$$n(n-1)/2$$

- Probability of a coalescence event:

$$p \approx \frac{1}{2N}\frac{n(n-1)}{2} = \frac{n(n-1)}{4N}$$

Waiting times between coalescence events

- After one coalescence event in the sample, there are $n-1$ individuals left (with high probability, ignore multiple coalescences)
- Time between first and second coalescence event: T_{n-1}

  $\vdots$

Expected time to coalescence events ($N = 10,000$, $n = 5$)

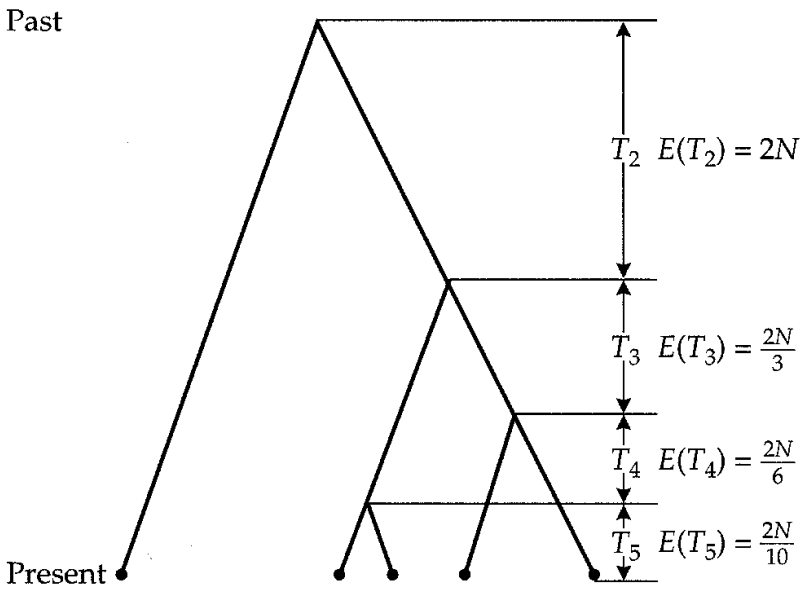
Coalescent event	Generations	per N
1^{st} coalescence event	2,000	2N/10
2^{nd} coalescence event	3,333	2N/6
3^{rd} coalescence event	6,666	2N/3
4^{th} coalescence event	20,000	2N

(Neutral) Mutations in the coalescent

Separate the genealogical from the mutational process (neutral mutation!)

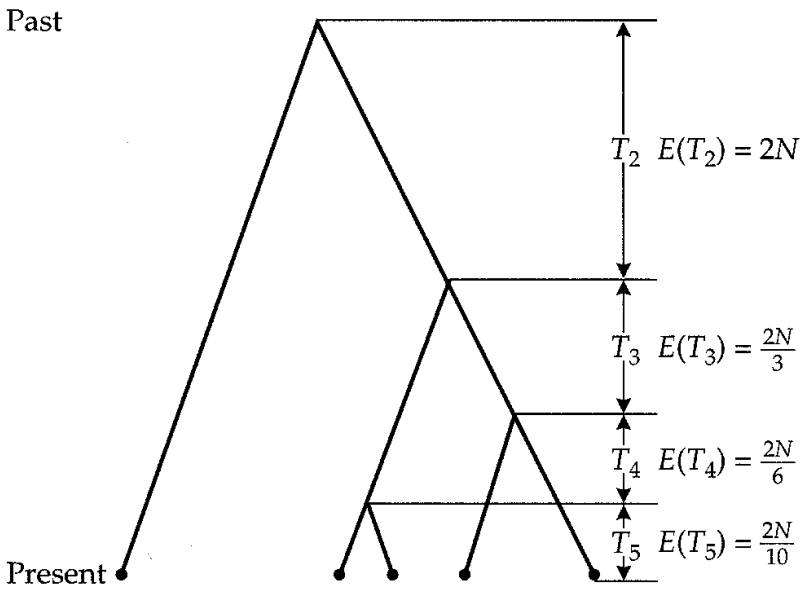
(Neutral) Mutations in the coalescent

Separate the genealogical from the mutational process (neutral mutation!)



(Neutral) Mutations in the coalescent

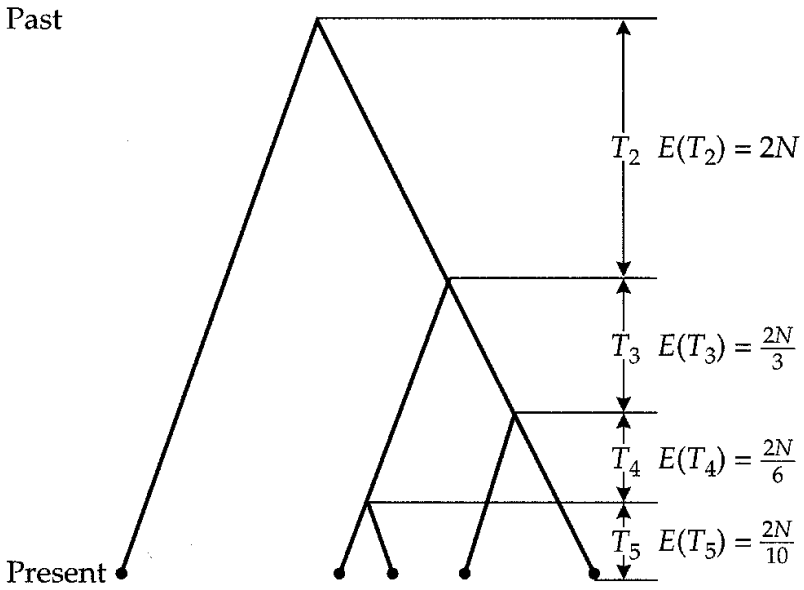
Separate the genealogical from the mutational process (neutral mutation!)



Coalescence time	Generations (approx.)
T_5	$5 \times 2,000 = 10,000$
T_4	$4 \times 3,333 = 13,333$
T_3	$3 \times 6,666 = 20,000$
T_2	$2 \times 20,000 = 40,000$
Total	$83,333$

(Neutral) Mutations in the coalescent

Separate the genealogical from the mutational process (neutral mutation!)



Coalescence time	Generations (approx.)
T_5	$5 \times 2,000 = 10,000$
T_4	$4 \times 3,333 = 13,333$
T_3	$3 \times 6,666 = 20,000$
T_2	$2 \times 20,000 = 40,000$
Total	$83,333$

Total number of expected mutations:

- Mutation rate: $\mu = 10^{-4}$ per generation per locus
- Total number of expected mutations:

$$\mu E(T_c) \approx 10^{-4} \times 83,333 = 8.33$$

Estimating θ from data

- Remember: $\theta = 4N\mu$

Estimating θ from data

- Remember: $\theta = 4N\mu$
- Infinite sites model: each mutation hits another site
  $\Rightarrow$ each mutation causes a polymorphism in the sample

Estimating θ from data

- Remember: $\theta = 4N\mu$
- Infinite sites model: each mutation hits another site
  $\Rightarrow$ each mutation causes a polymorphism in the sample
- S_n: number of polymorphisms in the sample

Estimating θ from data

- Remember: $\theta = 4N\mu$
- Infinite sites model: each mutation hits another site
  $\Rightarrow$ each mutation causes a polymorphism in the sample
- S_n: number of polymorphisms in the sample
- Then:

$$E(S_n) = \mu E(T_c) \approx 4N\mu \sum_{i=2}^{n} \frac{1}{i-1} = \theta \sum_{i=2}^{n} \frac{1}{i-1}$$

Estimating θ from data

- Remember: $\theta = 4N\mu$
- Infinite sites model: each mutation hits another site
  $\Rightarrow$ each mutation causes a polymorphism in the sample
- S_n: number of polymorphisms in the sample
- Then:

$$E(S_n) = \mu E(T_c) \approx 4N\mu \sum_{i=2}^{n} \frac{1}{i-1} = \theta \sum_{i=2}^{n} \frac{1}{i-1}$$

- Solve for θ:

$$\hat{\theta} = \frac{S_n}{1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n-1}}$$

Other measures of genetic diversity

- Interpopulation differentiation for one locus (g_{ST}) and multiple loci (G_{ST})
- Contribution of a population to total genetic diversity
- F statistics after Sewall Wright
- Analysis of molecular variance (AMOVA)

$\Rightarrow$ See lectures on population structure later today!

Background

Measures of genetic diversity

Frequency distribution of polymorphisms

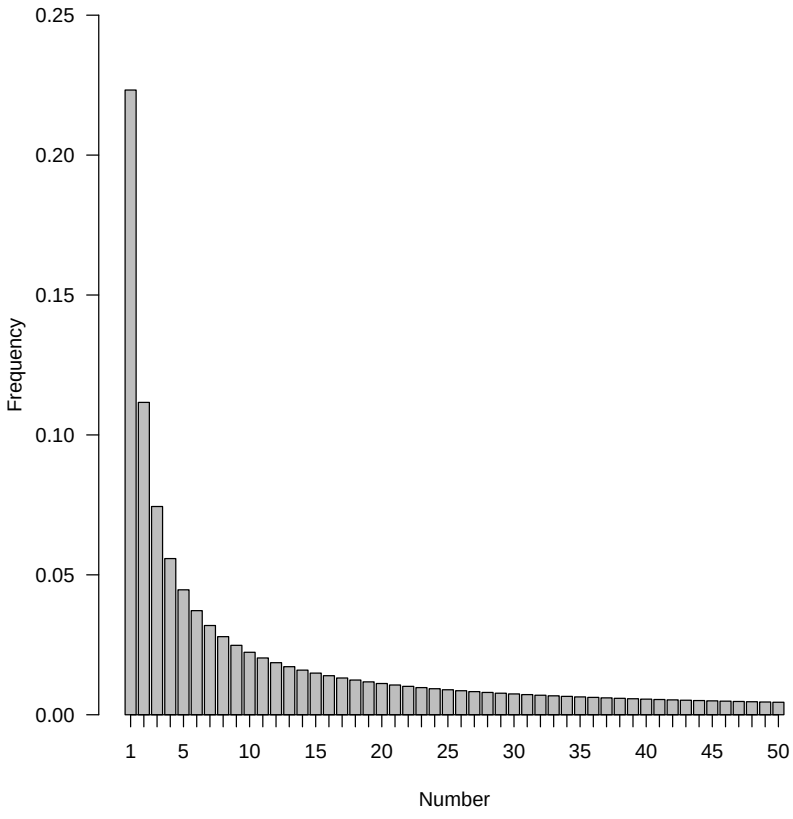
Genetic diversity of different site types

Haplotype-based measures of genetic diversity

Variation of genetic diversity in the genome

Challenges arising from genome sequencing (NGS)

Expected frequency distribution in a population



Sample size= $k = 50$

Expected frequency of i minor alleles in sample of k alleles

$$P = 1/ia_k$$

where

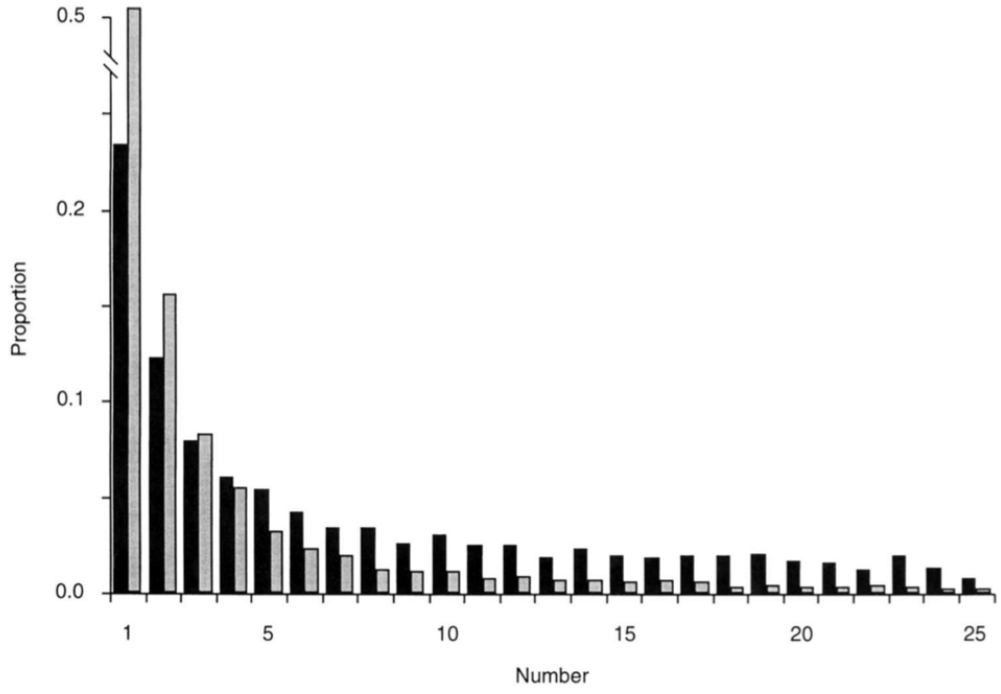
$$a_k = 1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{k-1}$$

(Watterson 1975, Fu and Li 1993, Fu 1995)

- Neutral model
- Distribution is independent of mutation rate
- Singletons are most frequent class
- Test of neutrality by comparing expected to observed distribution

Frequency distribution of under selection

Site frequency spectrum under no (black) and strong hitchiking (grey)



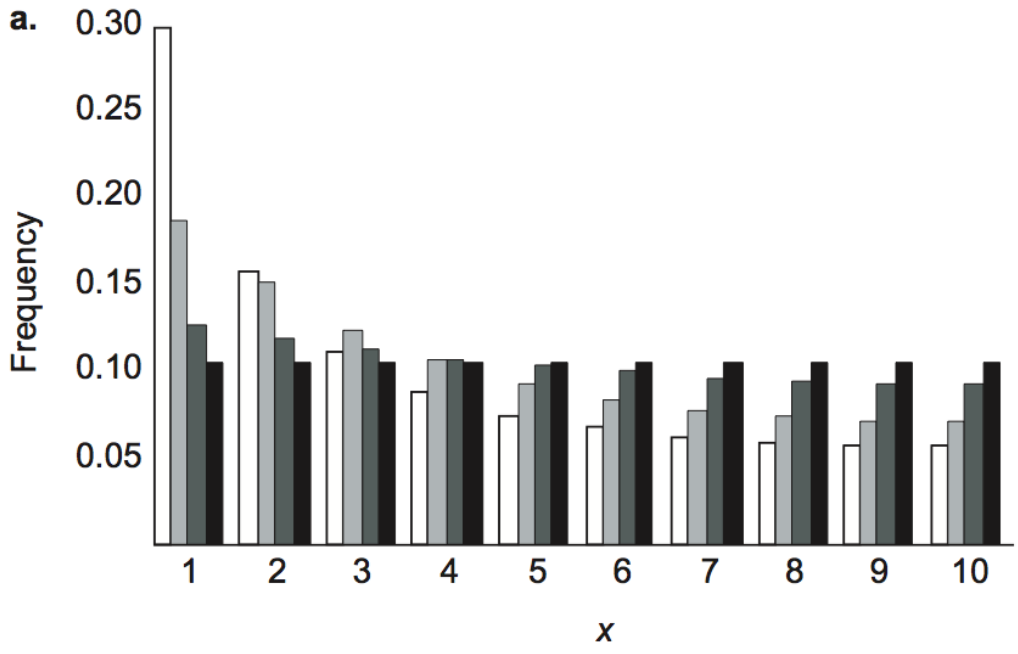
The excess of low frequency variants are new, rare polymorphisms that originated after a sweep.

Braverman et al. Genetics (1995)

Ascertainment bias of SNP markers

- SNPs are frequently used for high-throughput genotyping
- Identification of SNPs from a small panel of individuals
- Application to larger sample of the same or different population
- Ascertainment bias, if not corrected, leads to a wrong inference of parameters (genetic diversity, LD, etc)

Ascertainment bias of SNP markers

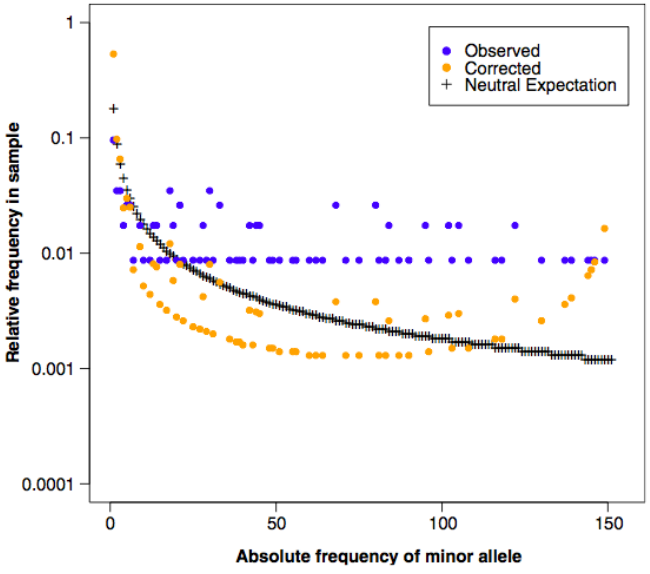


The expected folded frequency spectrum in the standard neutral model assuming a sample of size $n = 20$ chromosomes and an ascertainment sample size of $d = 2$ (black), $d = 5$ (dark grey), $d = 10$ (light grey) and $d = 20$ (white; no ascertainment bias).

Nielsen, Human Genomics (2004)

Examples of SNP ascertainment bias

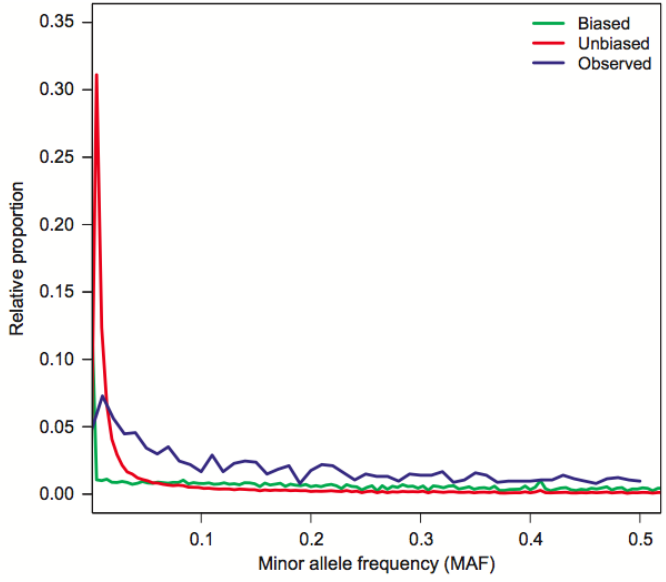
Arabidopsis thaliana



Observed, expected and corrected SNP allele frequency distributions

Schmid et al., TAG (2006)

Wild barley, *Hordeum spontaneum*



Observed and simulated SNP allele distributions under a panmictic and domestication model.

Hubner et al., Mol. Ecol. (2011)

Neutrality test with Tajima's D

Tajima's D is suitable for DNA sequence data

- Compare two estimators of nucleotide diversity, θ_π and θ_W
- Should be equal under neutral evolution
- θ_π is influenced by allele frequency distribution
- Calculated as standardized difference:

$$D = \frac{\theta_\pi - \theta_W}{\sqrt{\hat{V}(\theta_\pi - \theta_W)}}$$

- $D > 0$: Balancing selection (or population admixture)
- $D < 0$: Positive selection (or exponential population growth)

Diversity of different site types in genic regions

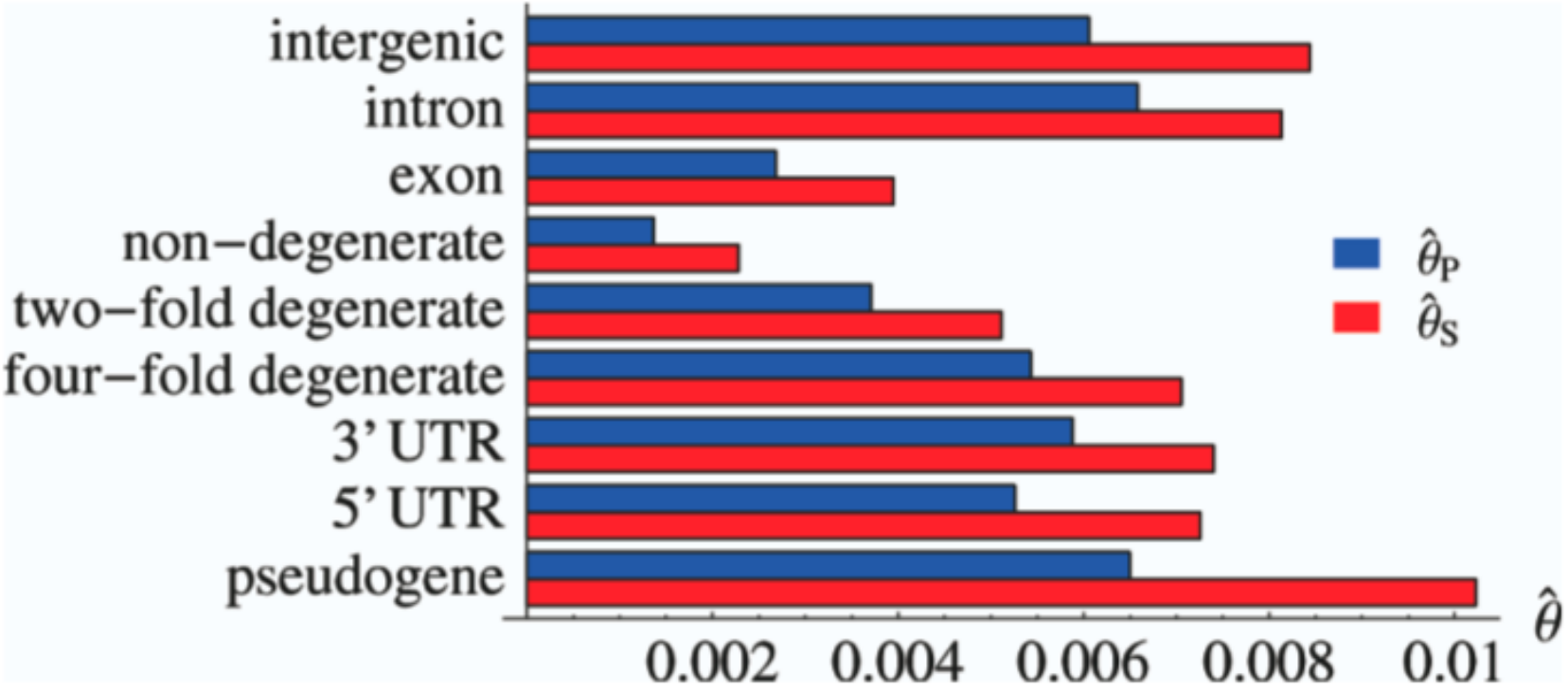
Genic regions can be separated into different site types:

- Coding vs. noncoding
- Coding sites: synonymous vs. nonsynonymous
- Noncoding: 5' UTR, Intronic, 3' UTR
- Upstream or downstream regulatory sequences

Classification of SNP polymorphisms is based on their location in the coding sequence

Site types differ in their level of polymorphisms

Example: *Arabidopsis thaliana*



Two estimates: θ_P uses average of pairwise differences ($= \theta_{pi}$), θ_S: uses the number of polymorphic sites ($= \theta_W$

Nordborg et al. PLoS Biology (2005)

Synonymous and nonsynonymous polymorphisms

The comparison of site types is the basis of some neutrality tests: e.g., **McDonald-Kreitman test**:

Nonsynonymous fixed differences	Nonsynonymous polymorphisms
Synonymous fixed differences	Synonymous polymorphisms

- In a neutrally evolving sequence (i.e., pseudogene) the ratio of Nonsynonymous to synonymous polymorphisms and fixed differences should be the same
- Test with a 2x2 table

Background

Measures of genetic diversity

Frequency distribution of polymorphisms

Genetic diversity of different site types

Haplotype-based measures of genetic diversity

Variation of genetic diversity in the genome

Challenges arising from genome sequencing (NGS)

Haplotype diversity

For defined genetic regions (i.e., exons or coding regions), haplotype diversity can be calculated as

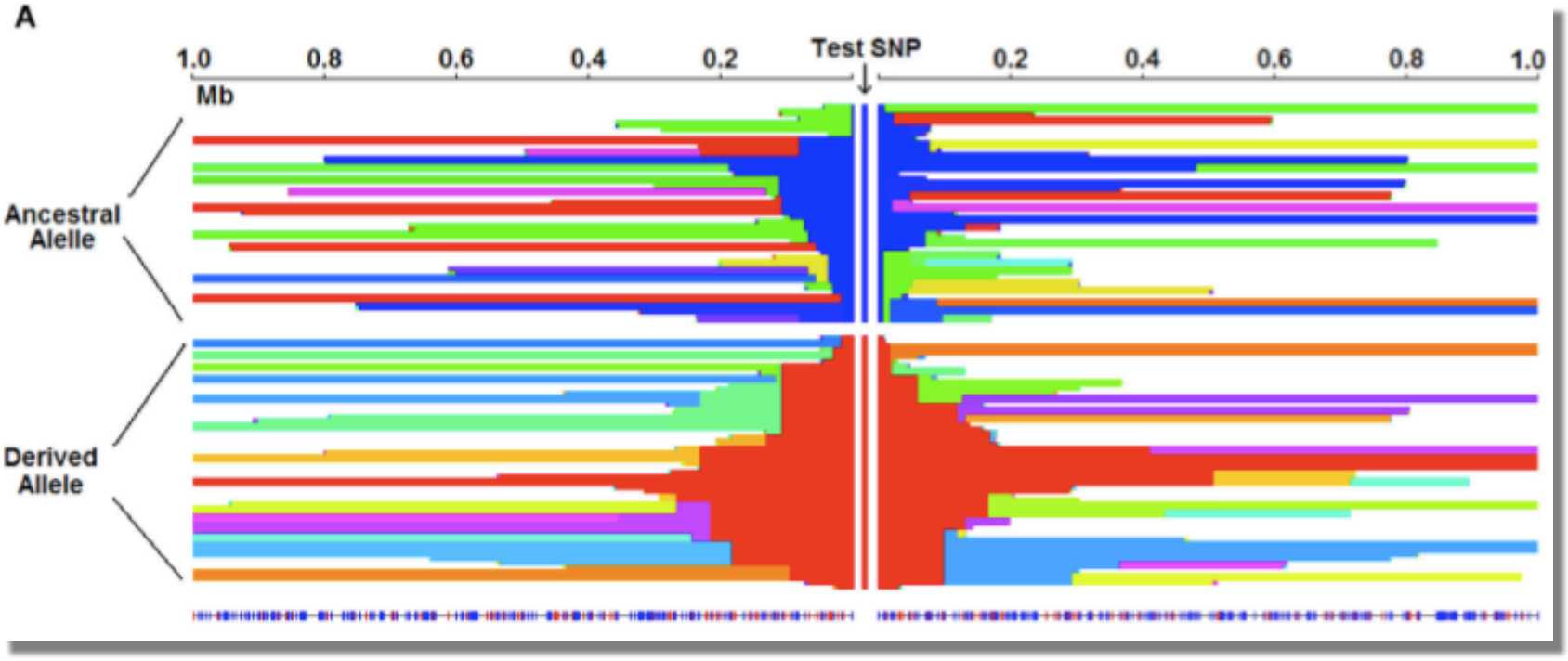
$$H = \frac{n}{n-1}(1 - \sum_i x_i^2)$$

where

- n is the sample size
- x_i is the frequency of haplotype i

Extended Haplotype Homozygosity (EHH)

- EHH is the length of a haplotype around a focal SNP
- The ratio of the areas around the two alleles of a SNP is the integrated haplotype score (iHS)



$\Rightarrow$ More about this in later lectures this week!

Background

Measures of genetic diversity

Frequency distribution of polymorphisms

Genetic diversity of different site types

Haplotype-based measures of genetic diversity

Variation of genetic diversity in the genome

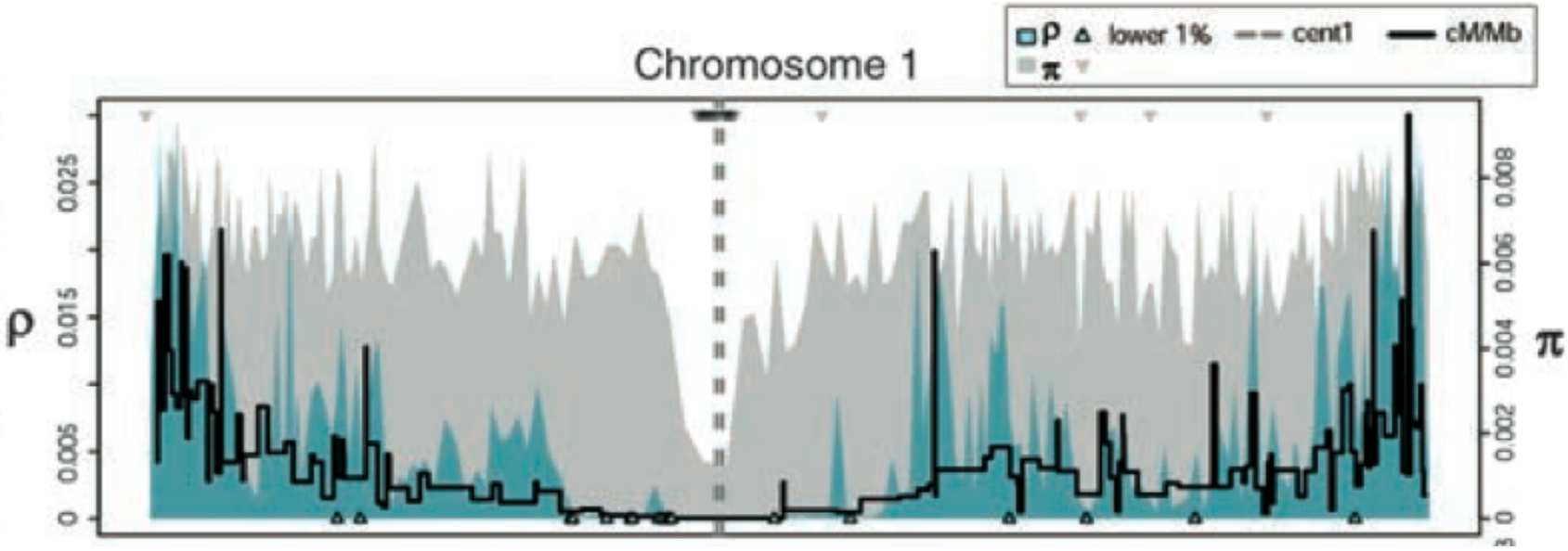
Challenges arising from genome sequencing (NGS)

Variation of genetic diversity in different species

- Genetic diversity varies by 2 orders of magnitude ($0.0001 < \pi < 0.01$)
- Variation in population sizes varies by larger orders of magnitude
- Interaction of (weak) selection and genetic drift likely maintains levels of genetic variation in a small range

Leffler et al., PLoS Biology (2012)

Maize Haplotype Map based on 27 inbred lines



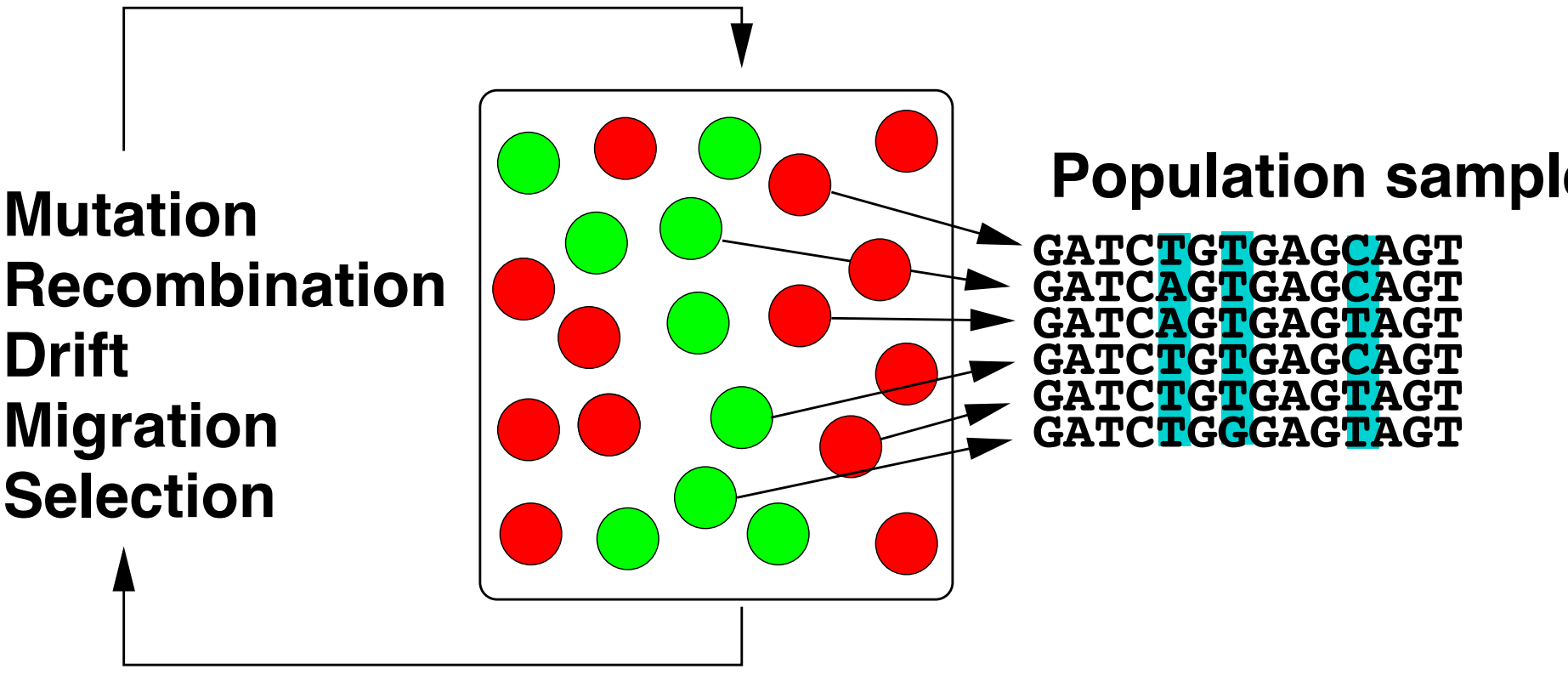
- Variation in both nucleotide diversity and recombination rate

Gore et al., Science 2009

Which process cause variation in genetic diversity?

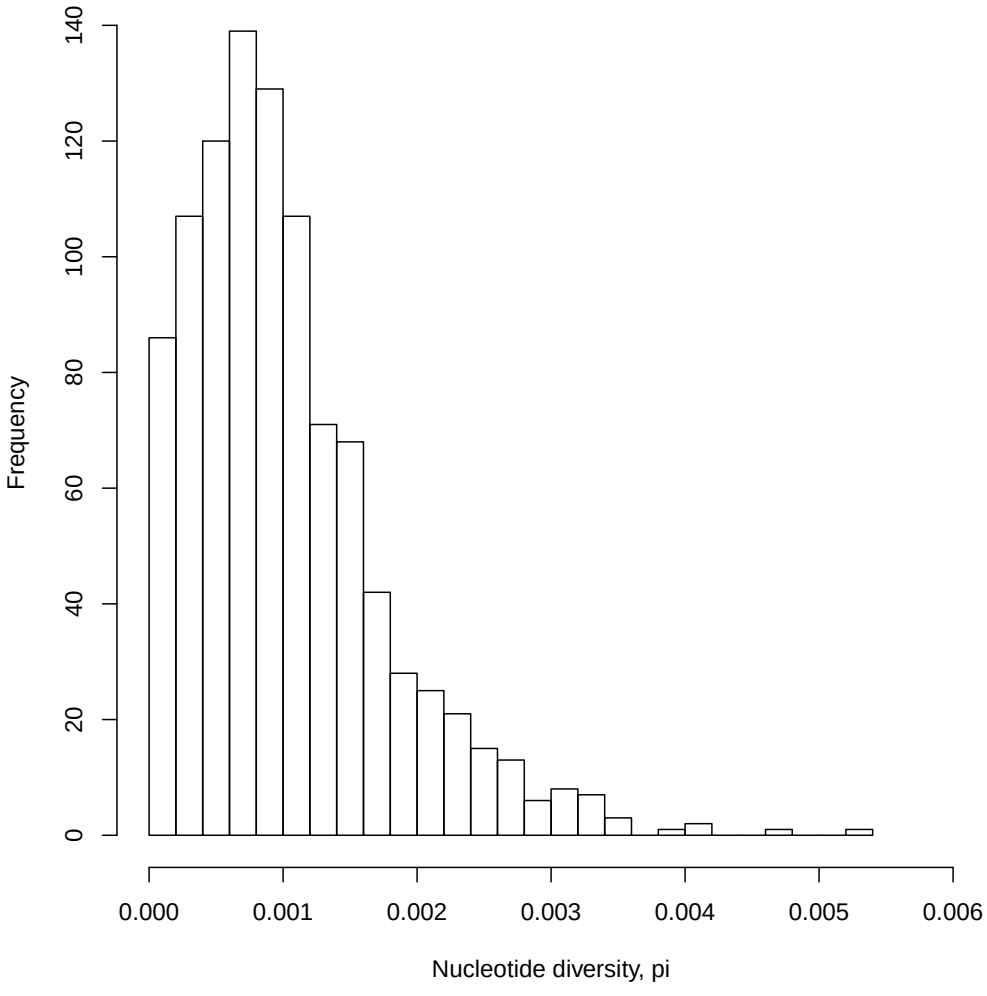
- **Random variation:** Genetic drift and sampling variance
- **Selection:** Positive and purifying (or background) selection
- **Recombination:** Variation in recombination rates throughout the genome
- **Structural variation:** Heterchromatic vs. euchromatic regions; inversions and duplications, ...

Random variation as a source of differences



Random variation at a locus

Coalescent simulation of a standard neutral model



Simulation parameters:

- Sample size = 100
- 1,000 Simulations
- θ (per nt) = 0.001
- $\rho = 0.001$
- Gene length = 1,000 nt

Reduction of genetic variation by selection

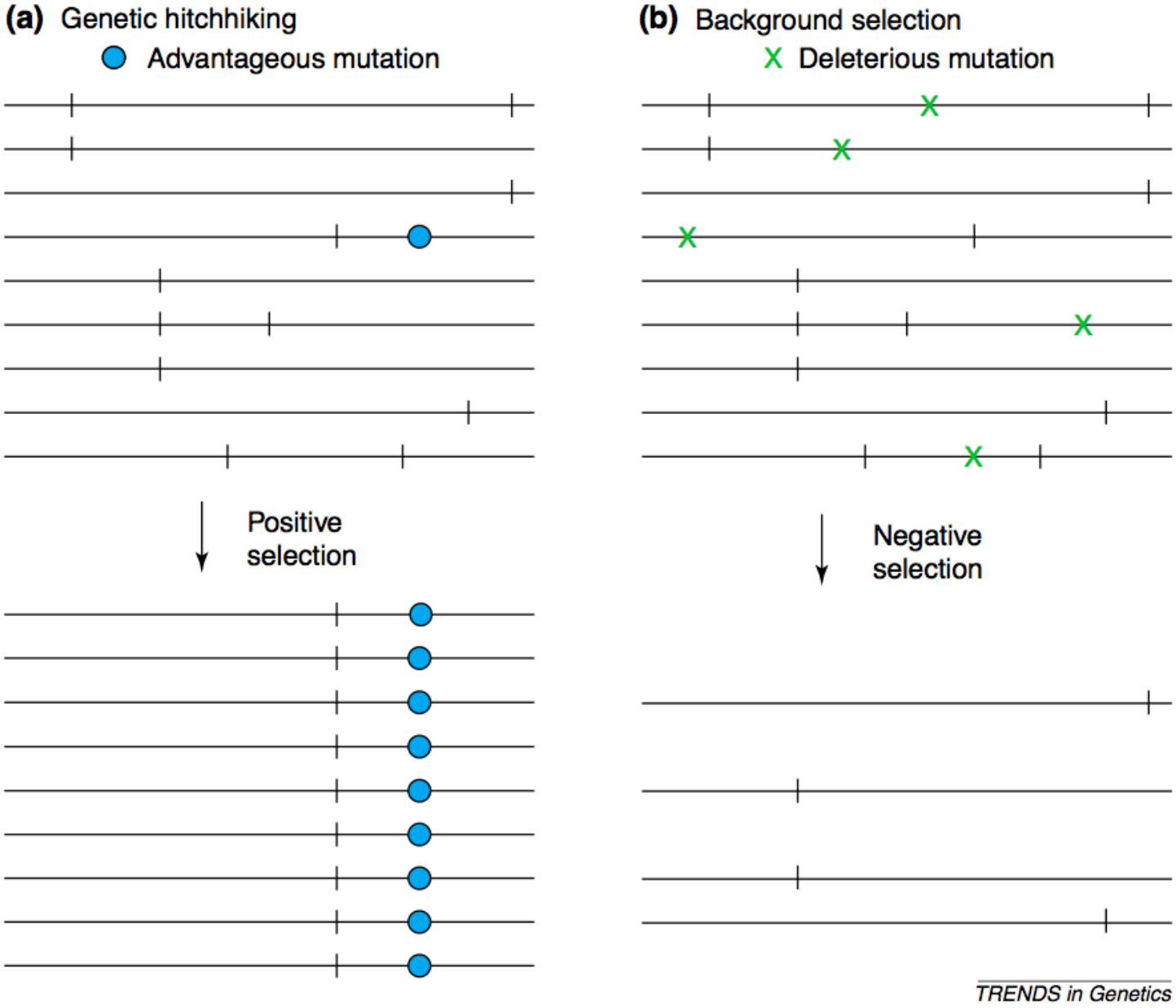
Positive selection:

- **Selective sweep:** selective fixation of a new, advantageous polymorphism
- **Hitchhiking** of linked neutral variation $\rightarrow$ Loss of variation
- Length of affected region depends on recombination rate and selection strength

Purifying or background selection:

- Removal of deleterious variation by selection
- Linked neutral variation is removed as well
- Removal of chromosomes with deleterious mutations $\rightarrow$ Reduced N_e $\rightarrow$ Lower genetic diversity
- Size of affected regions depends on recombination rate

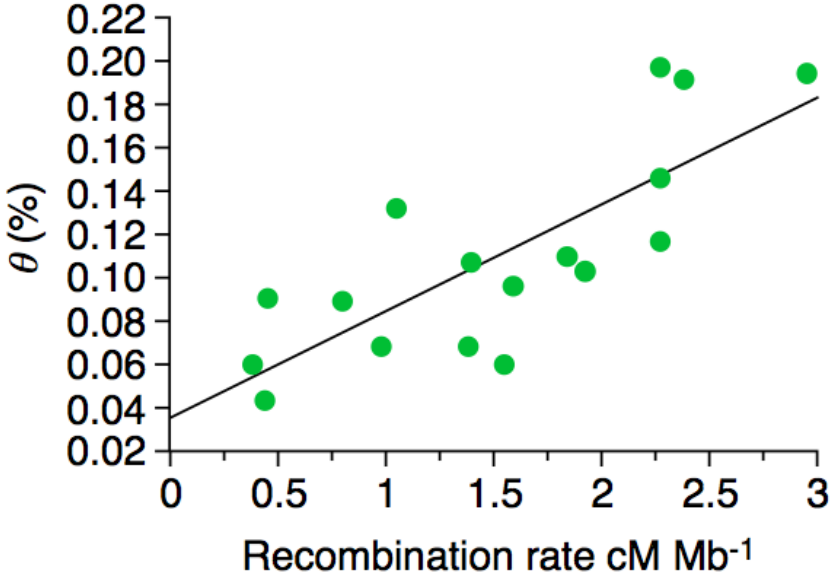
Positive vs. background selection



Nachman, Trends in Genetics (2001)

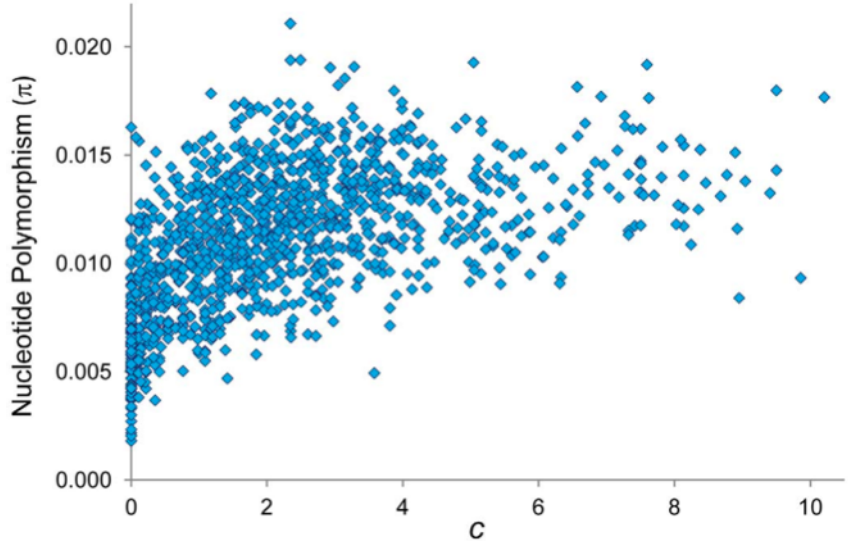
Genetic diversity vs. local recombination rate

Homo sapiens



Nachman, Trends in Genetics (2001)

Drosophila melanogaster

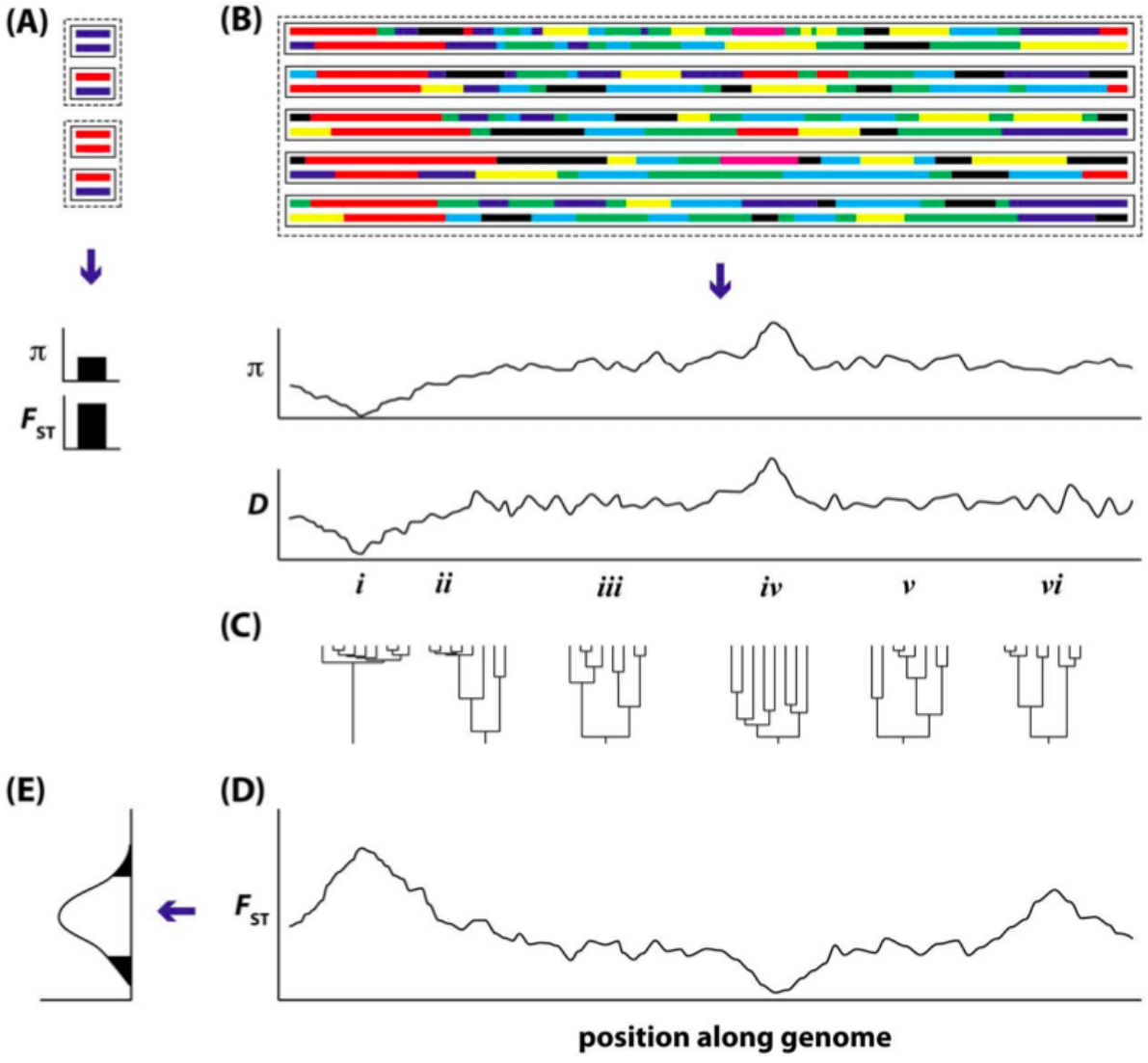


Comeron et al., PLoS Genetics (2012)

Explanations:

- Stronger background selection in low recombination regions
- Mutagenic nature of high recombination
- Complex interplay of biased gene conversion, mutation and recombination

NGS and population genomics



Hohenlohe et al. Int. J. Plant. Sci. (2010)

Genetic diversity in the NGS age

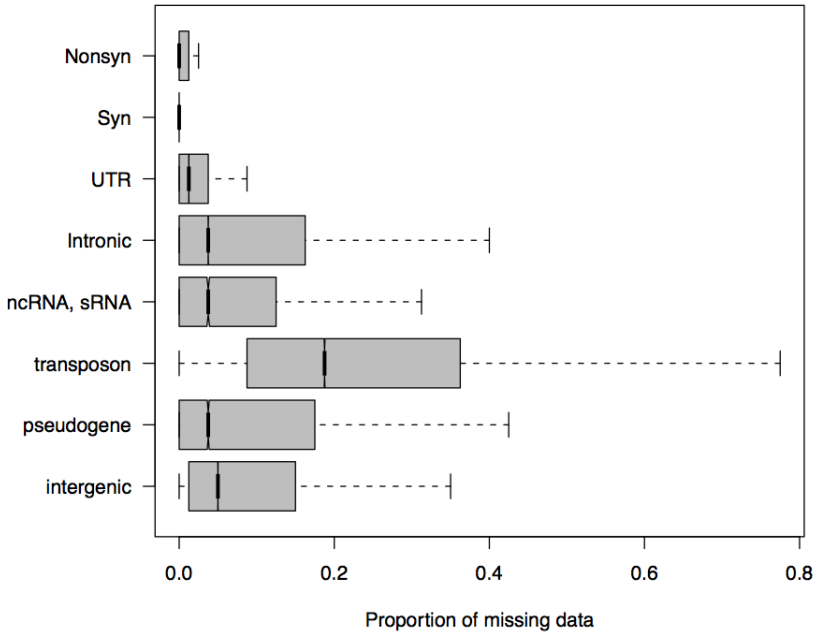
Key assumption of classical measures of nucleotide diversity:
$\Rightarrow$ The alleles included in the calculation of a single locus are homologous!

Adaptation of classical diversity statistics to the NGS age

- Joint estimation of sequence quality and diversity
- Short reads: Allelic or paralogous variation?
- Repetitive regions are difficult to sequence
- Divergent alleles are difficult to align to reference Reduced sequence quality for repetitive regions (i.e. no mapping to reference possible)
- A large proportion of variation is located in non-genic regions

$\Rightarrow$ Session with Christian Schlötterer

Data quality is an issue in NGS



Cao et al., Nature Genetics (2011)

- Based on mapping to high-quality reference!
- Highly divergent alleles are masked
- Genes under balancing selection are missed
- $\Rightarrow$ Need to wait for 3rd generation sequencers?

Frequency of structural variants

An integrated map of genetic variation from 1,092 human genomes

The 1000 Genomes Project Consortium*

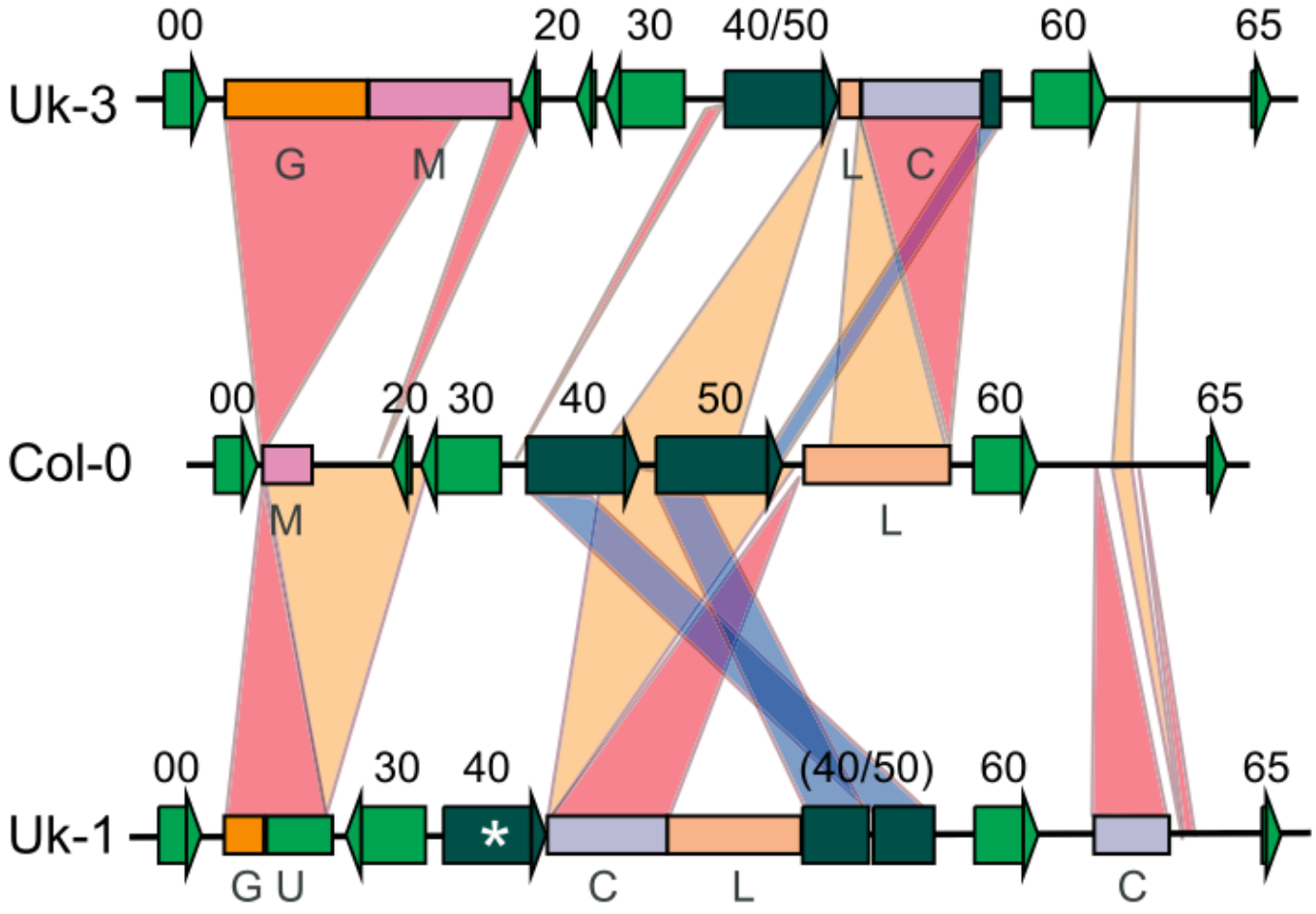
By characterizing the geographic and functional spectrum of human genetic variation, the 1000 Genomes Project aims to build a resource to help to understand the genetic contribution to disease. Here we describe the genomes of 1,092 individuals from 14 populations, constructed using a combination of low-coverage whole-genome and exome sequencing. By developing methods to integrate information across several algorithms and diverse data sources, we provide a validated haplotype map of 38 million single nucleotide polymorphisms, 1.4 million short insertions and deletions, and more than 14,000 larger deletions. We show that individuals from different populations carry different profiles of rare and common variants, and that low-frequency variants show substantial geographic differentiation,

56 | NATURE | VOL 491 | 1 NOVEMBER 2012

Structural variation in NGS is measured as read depth variants (RDV)

Complex patterns of allelic variation

Changes in disease resistance genes lead to autoimmunity diseases in Arabidopsis



Bomblies et al., PLoS Biology (2007)

Complex patterns of allelic variation

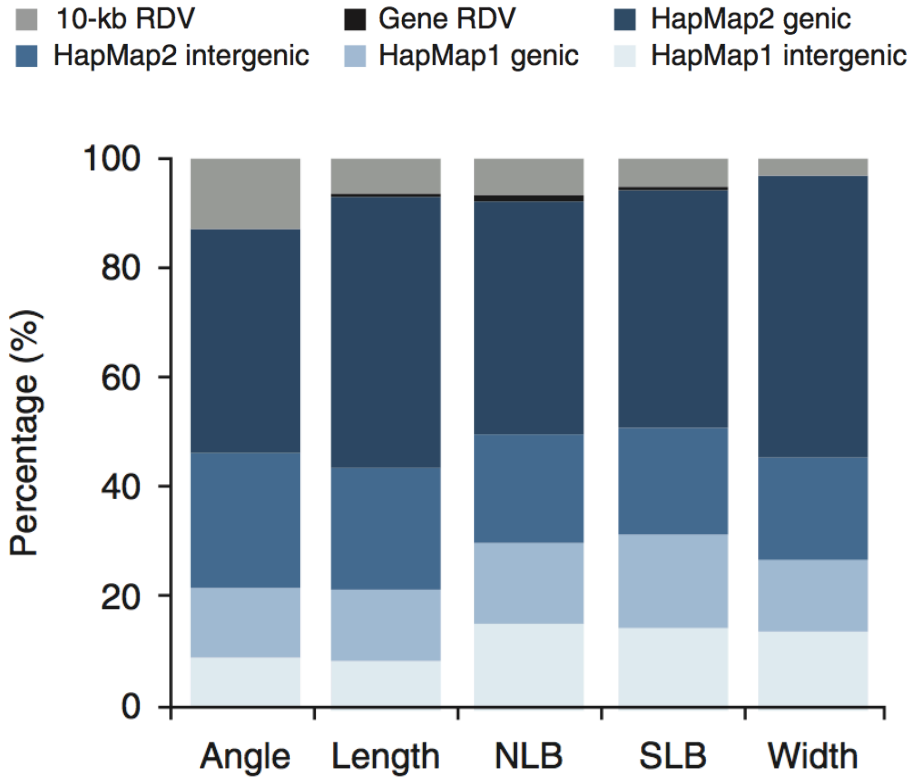
Differences in the *bronze* region of two maize varieties

- Genes are conserved (orange/yellow)
- Transposable elements are highly variable (red/blue/green)

Modified after Dooner et al.

Structural variation in maize

- Maize HapMap2: Resequencing of 103 inbred lines
- Proportion of significant SNPs and RDVs in a GWAS of five traits



Traits:

- Leaf angle
- Leaf length
- Leaf width
- Resistance to Northern Corn Blight
- Resistance to Southern Corn Blight

Chia et al., Nature Genetics 2012

Summary

- Genetic variation can be quantified with different metrics
- Close relationship with theoretical models
- Genetic variation is simultaneously affected by different processes
- NGS has a big impact on diversity estimation
- “Dark matter” of genomes is challenging also for population genomics

Thank you!