Ecological Genetics – HS18

v0.5

Gleb Ebert

November 13, 2018

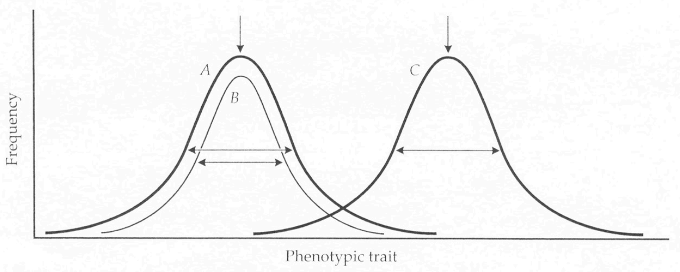
This document aims to summarize the lecture Ecological Genetics as it was taught in the autumn semester of 2018. Unfortunately I can't guarantee that it is complete and free of errors. You can contact me under `glebert@student.ethz.ch` if you have any suggestions for improvement. The newest version of this summary can always be found here: `http://www.glebsite.ch`

Contents

1 Introduction

> *"Nothing in biology makes sense except in the light of evolution"*
> — Theodosius Dobzhansky

In 1971 Ford wrote that ecological genetics deal with the adjustments and adaptations of wild populations to their environment. According to Conner and Hartl (2004) ecological genetics is the study of the process of phenotypic evolution occurring in present-day natural populations and is concerned with the genetics of ecologically important traits, that is, those traits related to fitness.
Phenotypic evolution is the change in the mean or variance of a trait across generations due to changes in allele frequencies.



Ecologically important traits are closely tied to fitness and are important in determining an organisms adaptation to its natural environment. **Adaptation** is a heritable phenotypic trait that has evolved in a population in response to a specific environmental factor and improves the survival or reproduction of its carriers. It can also be seen as a process whereby the members of a population become better suited to some feature of their environment through change in a characteristic that affects their survival or reproduction. Of the four key evolutionary processes, only natural selection consistently leads to adaptation (mutations, genetic drift and gene flow don't).

Uses of ecological genetics include

- agriculture (crop improvement)
- medicine (e.g. antibiotics)
- conservation measures (assisted migration)
- geographical differences between populations
- changes in species composition
- habitat adaptation & speciation

Fields related to ecological genetics include

- population genetics
- ecology
- evolutionary biology
- phylogenetics
- quantitative genetics
- statistics
- molecular biology
- epigenetics
- genomics

2 Species

Species are the fundamental unit in ecology, evolution and conservation legislation. Depending on the research question, adequate species identification, assignment of samples to populations or discrimination of individuals may be of relevance.

2.1 Species concepts

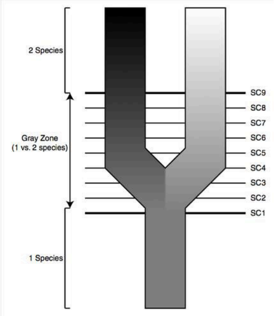
- **Morphological / typological species concept**: focus on similar morphology
- **Biological species concept**: group of potentially or actually interbreeding populations which are **reproductively isolated** from other such groups
- **Phylogenetic species concept**: focuses on monophyletic lineages

2.2 Operational Taxonomic Unit

An **OTU** is a group of organisms that is treated as a distinct evolutionary unit for the purposes of research underway. They are often applied when one or several species concepts fail. Once identified and research has been completed, OTUs should receive full taxonomic treatment and be given a scientific name if possible. OTUs are sometimes called **molecular operational taxonomic unit** (MOTU) when molecular methods are used.

2.3 Unified Species Concept

The only necessary property of species is that they form a separately evolving matepopulation (involves dynamics of gene flow and separation) lineage. The concept separates species conceptualization and separation. All criteria can be used for species delimination and any one of the properties is accepted as evidence for the existence of a species. More properties provide a higher degree of corroboration.



2.4 Identification of Species

Difficulties may include

- species-specific traits are not (always) visible
- differences are cryptic
- direct observation may be difficult and traces may be confused
- undescribed species may occur

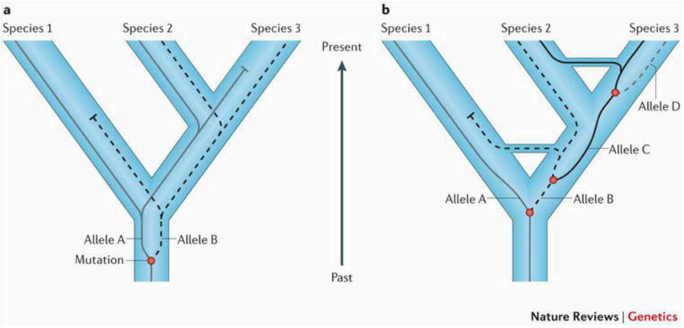
Parataxonomy may be used when identification is difficult. It sorts the material to species on the basis of external morphology without considering taxonomy.

2.5 Species delimination

Species delimitation is the act of identifying species-level biological diversity (independent evolutionary lineages). Most methods fit models to collected data to make (often different) simplifying assumptions. Incongruence across methods may occur due to differences in the power to differentiate lineages or due to violations of one or several assumptions made by a given method. Fundamentally there are two approaches. Some models can assign samples to groups without being given information first (STRUCTURE, Structurama, Geneland). Others need the user to assign samples to putative lineages (BPP, iBPP, spedeSTEM, DISSECT, tr2).

2.5.1 Problems of Species Trees

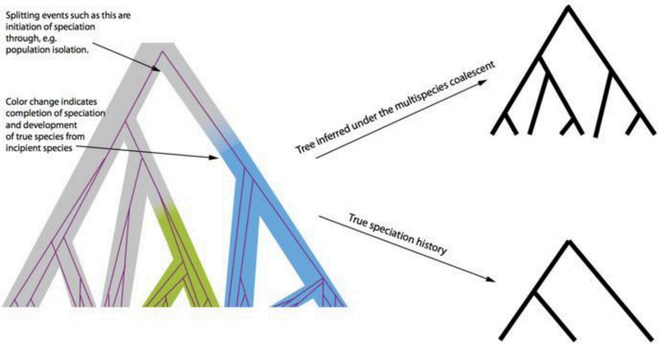
Problems like imcomplete lineage sorting (left) or gene flow (right) may occur.



2.5.2 Bayesian Species Identification under the Multispecies Coalescent

This method is currently the most used approach for species delimination. The multispecies coalescent describes the genealogical relationships between DNA samples from several species. Simplifying assumptions include:

- species phylogeny unknown
- complete isolation after divergence
- no recombination


Splitting events such as this are initiation of speciation through, e.g. population isolation.
Color change indicates completion of speciation and development of true species from incipient species
Tree inferred under the multispecies coalescent
True speciation history

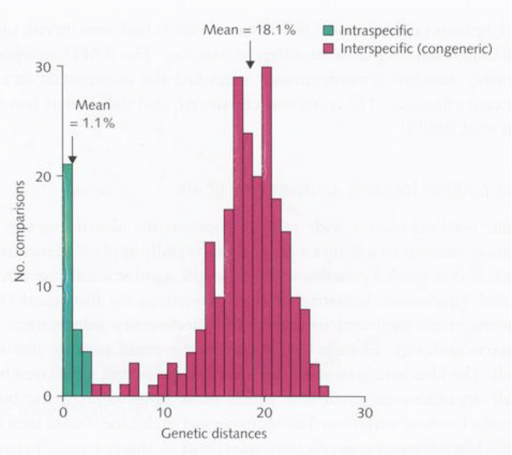
The above graph shows that the MSC approach identified populations as separate species from a simulated data set. MSC delimits structure, not species.

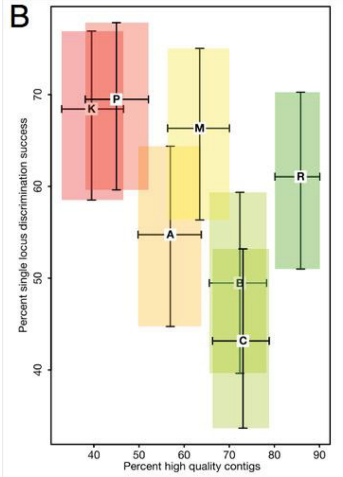
2.5.3 Recommendations for Species Delimination

- use at least 10 samples from all lineages
- simulate before analysis
- use several complementary methods
- combine genetic with nongenetic data
- define the used species concept
- be cautious

2.5.4 DNA Barcoding

DNA barcoding is a method for rapid identification of species through the analysis of short, standardized gene regions. These have to be universal for all animals or plants, have to be amenable to the production of bidirectional sequences with little ambiguity and allow for discrimination of most species. The barcoding gap defines the distance between species. Genetic distances are based on the extent of nucleotide sequence devergence. Some groups like orchirds lack the barcoding gap.





Examples of application are the identification of amphibian species diversity and abundance after epidemic diseases in Panama or restriction of shaving brushes to only use hair from the Hog badger instead of the Eurasian badger.

3 Molecular Markers

Molecular markers are polymorphic proteins or DNA sequences and reveal different alleles within individuals, populations or species. Ideally they can be used as indicators of genome-wide variation.

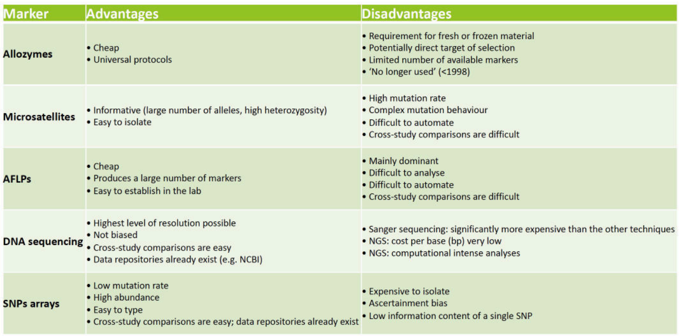
- **Chromosome based markers**
  - Numbers and staining patterns
- **Enzyme based**
  - Allozymes
- **DNA based**
  - Restriction fragment length polymorphisms (RFLPs)
- **DNA & PCR based**
  - Random amplified polymorphic DNA (RAPD)
  - Amplified fragment length polymorphism (AFLP)
  - Microsatellites (SSRs)
  - DNA sequencing and SNPs

3.1 Genome

The size of genomes can differ between species by large factors. Size and complexity are not coupled, as non-coding regions are the main reason for big genomes. These regions contain repeated sequences like tandem repeats or interspersed repeats (transposable elements). The number of

chromosomes is limited, because at some point there would be problems with the spindle apparatus. Introns are spliced after transcription while intergenic regions aren't.
Mitochondria and chloroplasts have their own genome (mtDNA and cpDNA respectively). Depending on the species, different organelle genomes should be used for comparisons. There is much more organelle DNA in a cell than there is nuclear nDNA and it is easier to amplify because of its high conservation.

3.2 Widely used genetic markers



Marker	Advantages	Disadvantages
Allozymes	• Cheap • Universal protocols	• Requirement for fresh or frozen material • Potentially direct target of selection • Limited number of available markers • 'No longer used' (<1998)
Microsatellites	• Informative (large number of alleles, high heterozygosity) • Easy to isolate	• High mutation rate • Complex mutation behaviour • Difficult to automate • Cross-study comparisons are difficult
AFLPs	• Cheap • Produces a large number of markers • Easy to establish in the lab	• Mainly dominant • Difficult to analyse • Difficult to automate • Cross-study comparisons are difficult
DNA sequencing	• Highest level of resolution possible • Not biased • Cross-study comparisons are easy • Data repositories already exist (e.g. NCBI)	• Sanger sequencing: significantly more expensive than the other techniques • NGS: cost per base (bp) very low • NGS: computational intense analyses
SNPs arrays	• Low mutation rate • High abundance • Easy to type • Cross-study comparisons are easy; data repositories already exist	• Expensive to isolate • Ascertainment bias • Low information content of a single SNP

3.2.1 Microsatellites

- SSR (simple sequence repeat) and STR (short tandem repeat)
- highly polymorphic: mutation rates between 10^{-6} and 10^{-2} per locus per generation
- widely used to assess genetic variation in animals, plants and fungi as they are highly variable between individuals
- codominant
- mostly evolutionary neutral, as they are in intragenic regions
- PCR-based (primers, agarose-gel electrophoresis)
- capillary sequencers use fluorescence labelled primers
- 10-20 SSRs per study
- the mutation mechanism is called slipped-strand mispairing: polymerase slips of and when rejoining doesn't know which repeat it already copied; insertions and excisions happen, but the latter seem to be corrected in nature

3.2.2 Structural variation (SV)

- Microsatellite repeats
- 1bp indels
- More complex insertions and deletions
- Copy number variants (CNVs) (> 1kb)
- ...

3.2.3 Amplified fragment-length polymorphisms (AFLPs)

- 100-4'000 random markers
- dominant (homo- and heterozygotes indistinguishable)

1) DNA extraction
2) Digestion by restriction enzyme MseI and EcoRI
3) Ligation of the adaptors MseI and EcoRI
4) Selection amplification (+3/+3 bp)
5) Sequencing
6) Binary data matrix (peak present/absent)

3.2.4 SNP microarrays

DNA is hybridised to predefined probes. Only common and known SNPs are called. This is the ascertainment bias.

4 Sampling

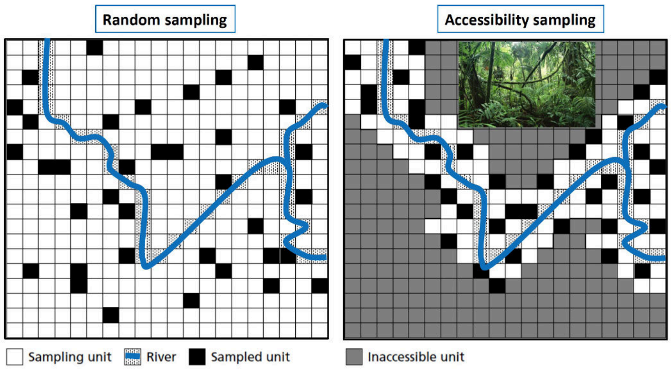
Sampling determines...

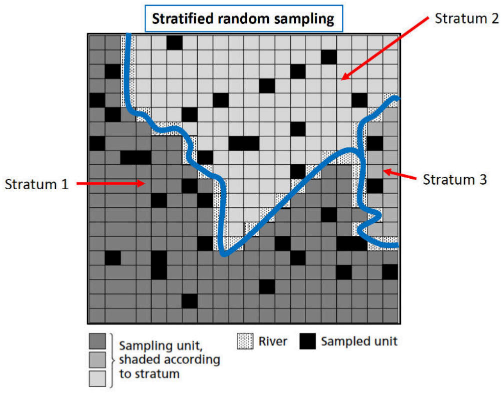
- The chances of answering your research questions
- The time you may have to spend to answer your questions
- Research costs
- The likelihood of obtaining research permits
- ...

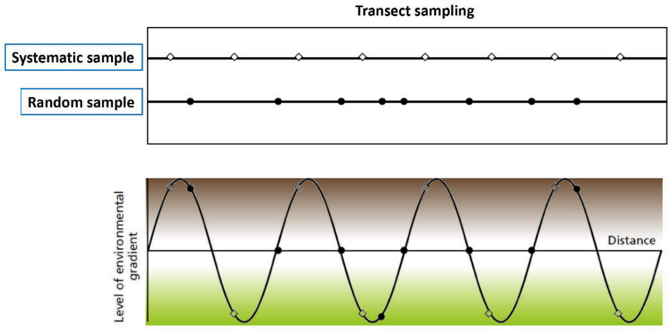
4.1 Populations

Population type	Definition
Genetic	All individuals which are connected by gene-flow
Ecological	Group of organisms occuring in a particular area at a particular time
Statistical	The universe of times that are under study

4.2 Individuals




Stratified random sampling
Stratum 2
Stratum 1
Stratum 3
Sampling unit, shaded according to stratum
River
Sampled unit



Do you sample proportionally or equally across strata? Randomly or systematically? There is no single right solution. One should always consider the circumstances.
Roughly 20-30 individuals represent a population (> 80% of all alleles).

4.3 Important considerations

- Documentation: archived, reproducable, verifiable, new technologies
- Storage: adequate storage and transport, test in advance!
- Permits: sampling, handling animals, export and import

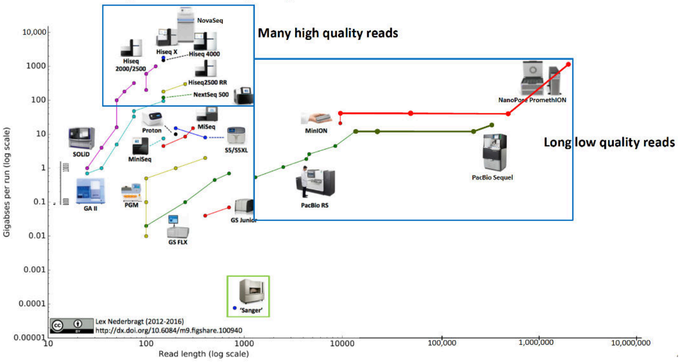
4.4 Nagoya protocol

The convention on biological diversity from 1993 had the objectives of conservation and sustainable use of biological diversity as well as the sharing to benefits arising from the utilisation of genetic resources. The Nagoya protocol from 2010 is a supplementary agreement that expands on the fair and equitable sharing of benefits arising out of the utilisation of genetic resources.

- **Access obligations**: Provide fair and non-arbitrary application procedures and issue permits when access is granted
- **Benefit-sharing obligations**: Share the value of genetic resources and traditional knowledge with developing countries
- **Compliance obligations**: Ensure that genetic resources and traditional knowledge have been accessed in accordance with prior informed consent

5 Genomic Methods

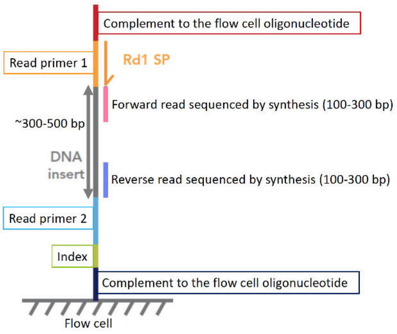
Sequencing methods evolved rapidly since the development of pyrosequencing in 1993. The cost per raw megabase is sinking rapidly as well. Data is being gathered faster and cheaper than it can be analysed nowadays.



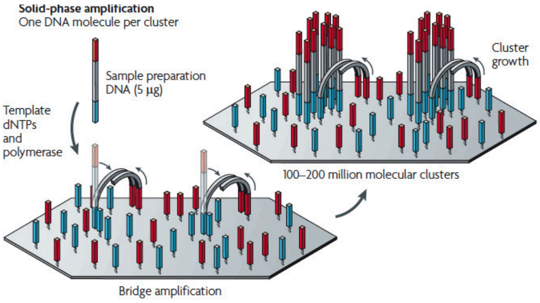
5.1 Illumina high-throughput sequencing

5.1.1 Sample / library preparation

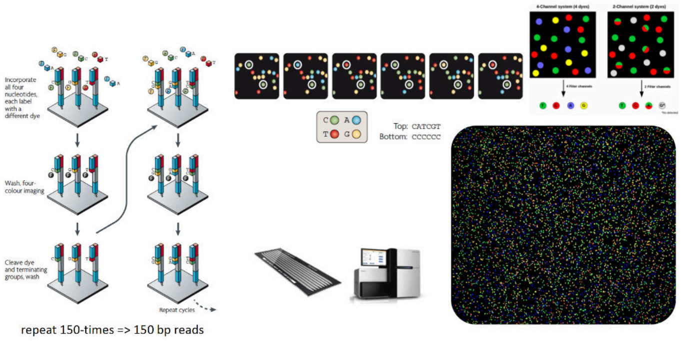
1) Fragmentation of genomic DNA
    1) Mechanical shearing: e.g. sonification
    2) Tagmentation: enzymes
2) Size selection: 300-500bp
3) Adapter ligation
    1) PCR amplification
    2) Individual barcoding


Complement to the flow cell oligonucleotide
Read primer 1
Rd1 SP
Forward read sequenced by synthesis (100-300 bp)
~300-500 bp
DNA insert
Reverse read sequenced by synthesis (100-300 bp)
Read primer 2
Index
Complement to the flow cell oligonucleotide
Flow cell

5.1.2 Bridge amplification and cluster generation

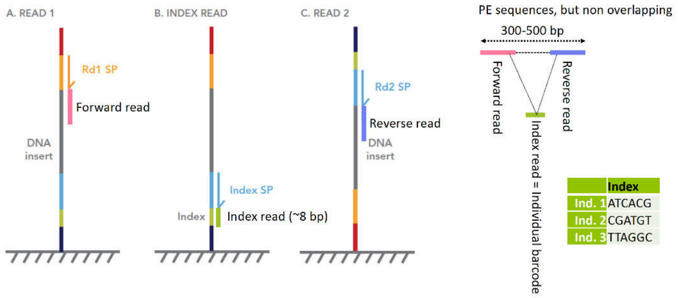

Solid-phase amplification
One DNA molecule per cluster
Sample preparation DNA (5 μg)
Template dNTPs and polymerase
Cluster growth
100–200 million molecular clusters
Bridge amplification

5.1.3 Sequencing by synthesis


Incorporate all four nucleotides, each label with a different dye
Wash, four-colour imaging
Cleave dye and terminating groups, wash
Repeat cycles
4-Channel system (4 dyes)
2-Channel system (2 dyes)
C A
T G
Top: CATCGT
Bottom: CCCCCC

repeat 150-times => 150 bp reads

5.1.4 Paired-end sequencing



5.2 3rd generation sequencing

3rd gen methods sequence single molecules and generates long reads of 10'000 - 2 million bp. They don't use PCR and thus avoid PCR artefacts. One player in the market is **Pacific Biosciences**. They use single molecule real-time analysis (SMART). Another one is **Oxford NanoPore**, whose **MinION** costs only around $300 and generates even longer reads than PacBio is able to. Sequencing happens through voltage changes in the membrane when DNA passes the nano pore. NanoPore sequencing is useful for *de novo* genome assembly, CNV detection, real-time identification of samples and mRNA splicing variant identification.

5.3 Uses of NGS data

NGS inferences require fully annotated high quality reference genomes, draft reference genomes, reference transcriptomes or *de novo* assembled STACKS (RADseq). Methods used for high-throughput marker discovery include:

- Sequencing of individuals and populations
- Whole-genome re-sequencing
- Transcriptome sequencing (RNA-seq)
- Reduced representation sequencing (RADseq; no reference genome required)
- Target capture sequencing methods
  - Sequencing of ultra-conserved elements
  - Exome capture

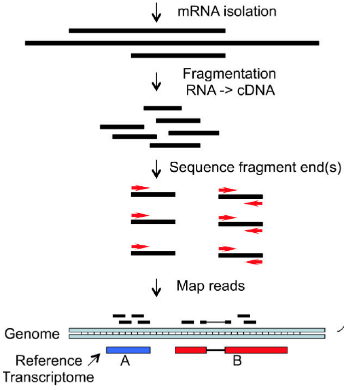
5.3.1 Whole genome re-sequencing

No bias from insufficient marker density or distribution occurs when re-sequencing whole genomes.

- **Individual sequencing**: 1 Flow cell NovaSeq S4 ($\sim 35'000$CHF+ ~ 100CHF per individual library
- **Sequencing pools of individuals** (Pool-seq)
  - Cost effective: population libraries (~ 100CHF per pool library)
  - Lower coverage per individual
  - Population allele frequenciy estimates (no individuals genotypes)

5.3.2 Transcriptome sequencing (RNA-seq)

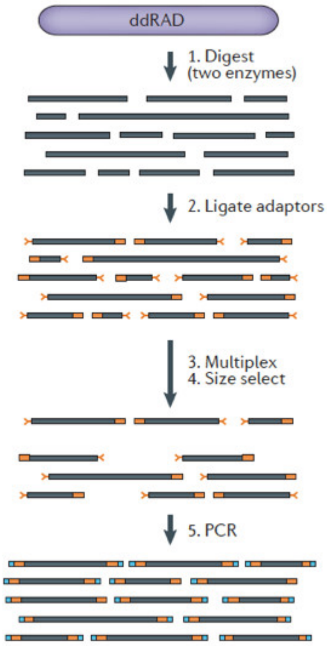
This method reduces the complexity of the genome through only sequencing the expressed exons (mRNA).

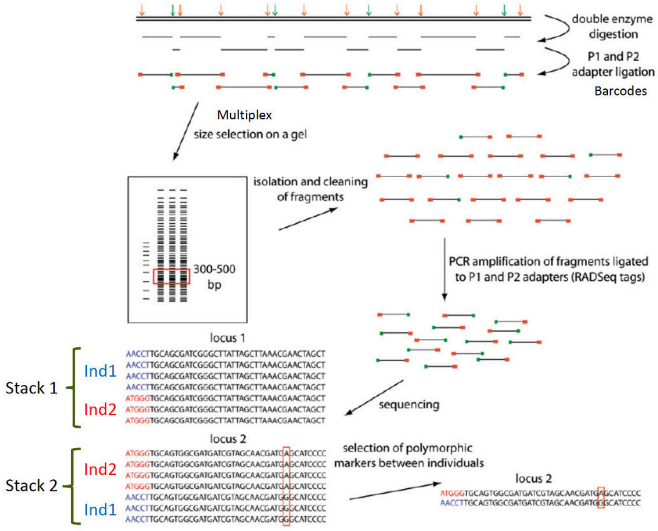


5.3.3 Reduced representation sequencing (RADseq)

One method of reduced representation sequencing is **Restriction site-Associated DNA Sequencing (RADseq)**. Restriction enzymes are used to obtain DNA sequences adjacent to a large number of restriction cut sites. As these site are conserved and no reference genome is required, RADseq is used extensively for high-throughput SNP discovery and genotyping in ecological and evolutionary studies. The sequencing depth is considered to be high and the cost per sample lower. 1-200 loci per 1 Mb of genome with lengths of 100-150bp result in a $\frac{100\text{ loci } *100\text{ bp}}{1'000'000} = 100$x reduction of genome complexity.

ddRAD: double digest RADseq





5.3.4 Marker identification

If possible, reads are mapped to a genome (e.g. through Burrows-Wheeler Alignment). Other algorithms then do the **SNP calling**. They take into account the **coverage** (# of reads per base), base and mapping qualities as well as many other factors.

5.3.5 Gene expression differences

Differentially expressed genes (DEGs) can be inferred from RNAseq data. Expression is measured in **FPKM**, which stands for **F**ragments (reads) **P**er **K**ilobase per **M**illion mapped fragments. It is then corrected for sequencing depth and gene / exon length.

5.4 The question of usefullness

The computational and storage needs when dealing with NGS data are enormous. Whether a few microsatellites are enough is a valid question. SSRs (see chapter 3.2.1) only reflect a limited portion of the genome and have very different mutation rates compared to SNPs and there's also a lot less of the former. Microsatellites in general suffer from **ascertainment bias**. It is of statistical nature and is introduced during collection of the data, when only markers that were found to be polymorphic are used. When markers with little variance are ignored, genetic diversity is generally overestimated. Also, rare alleles are often missed which may lead to incorrect inferences of demographic parameters. Additional benefits of whole genome re-sequencing information include

- More than anonymous markers
- Candidate genes can be studied
- Signatures of adaptation / selection can be detected
- Genetic and genomic diversity (e.g. estimates through exome-wide diversity)
- Demographic history

5.5 NGS pro and contra

	Advantages	Disadvantages
Data quantity	huge amounts of data	huge amounts of data
Data quality	high quality	multiple sequencing required; storage costs
Costs	cost per bp relatively cheap	high costs for individual reads; expensive IT infrastructure
Data analyses	almost everything possible	lots of IT infrastructure

6 Genetic variation

The ultimate source of genetic variation are **mutations**. They occur at random positions in the genome but rates can vary across genomes. In sexual life cycles, existing genetic variation is being re-shuffled continuously through random gamete fertilization, random chromosome segregation and recombination.

6.1 Forms of genetic variation

6.1.1 Singe-nucleotide polymorphism

SNP refers to variation in a single nucleotide at a specific position in the genome. Adjacent nucleotides or indels (insertions and deletions) can have substantial effects on SNP mutation rates. Roughly 15% of all polymorphisms are small indels.

6.1.2 Structural variation

- Balanced nucleotide variation
  - Inversion
  - Intrachromosomal translocation
  - Interchromosomal translocation (up to whole arms of chromosomes)
- Unbalanced nucleotide variations
  - Duplication
  - Deletion

6.1.3 Sizes of variations

- SNP: 1bp
- Microsatellites and minisatellites: 14-200bp
- Indels: <1kb
- Copy number variations (CNVs): >1kb

6.1.4 Other forms of genetic variation

- Gene expression variation
- Methylation variation
- Post-transcriptional modification

6.2 Levels of genetic variation

The main axes of variation in diversity are among species and within genomes. The neutral theory of molecular evolutions postulates that the vast majority of evolutionary change at the molecular level is maintained by the interaction between mutation, which creates variation, and genetic drift, which eliminates it. It also predicts, that in a population of constant size, diversity should be proportional to N_e. The neutral theory is usefull as a null hypothesis for test whether natural selection is occuring.

6.2.1 Population size

In an idealized, panmictic population, also known as **Wright-Fisher population**, with an equal expected contribution of individuals to reproduction and equal survival, the strength of genetic drift is inversely proportional to the size of the population. Real populations depart from the concept. Therefore the following two concepts are used. The **census population size N_c** is the number of individuals in a population. It varies by several orders of magnitude across taxa. The **effective population size N_e** is the size of an idealized population that would show the same amount of genetic diversity as the population of interest. It varies over time, with long-term N_e explaining current levels of genetic diversity in populations but contemporary N_e explaining how strong drift currently is.

The observed differences in population size are expected to determine differences in genetic diversity across species. However, across-species variation in genetic diversity is much narrower than the variation in abundance. This conflict has been termed the **paradox of variation** by Lewontin and is also known as **Lewontin's paradox**. Possible solutions include

- Demographic fluctuations
- Natural selection and genetic hitchhiking
- Molecular constraints on heterozygosity (in yeast, recombination is impeded when heterozygosity is too high)
- Variation in mutation rate

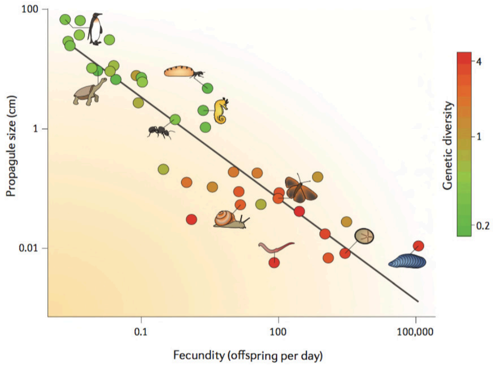


Figure 1 | **Genetic diversity and the r/K gradient in animals.** The average per-day fecundity is on the x axis and the average size of eggs or juveniles is on the y axis; each dot is for a family (one to four species each). The colour scale indicates the average nucleotide diversity at synonymous positions, expressed in per cent. The negative correlation reflects a trade-off between quantity and size of offspring. r-strategists (bottom right; for example, blue mussels, heart urchins and lumbricid earthworms) are more polymorphic than K-strategists (top-left; for example, penguins, Galapagos tortoises and subterranean termites). Figure from REF. 37, Nature Publishing Group.

6.3 Determinants of genetic variation

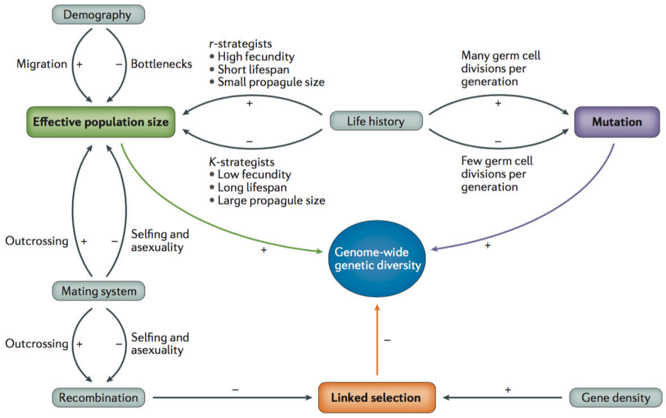


Figure 2 | **Overview of determinants of genetic diversity.** Effective population size, mutation rate and linked selection are the main factors affecting diversity. These factors are in turn governed by several other parameters. The direction of correlation is indicated by the + and – symbols. Selfing, self-fertilization.

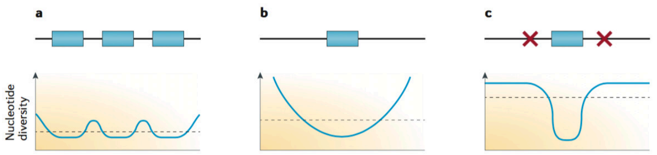
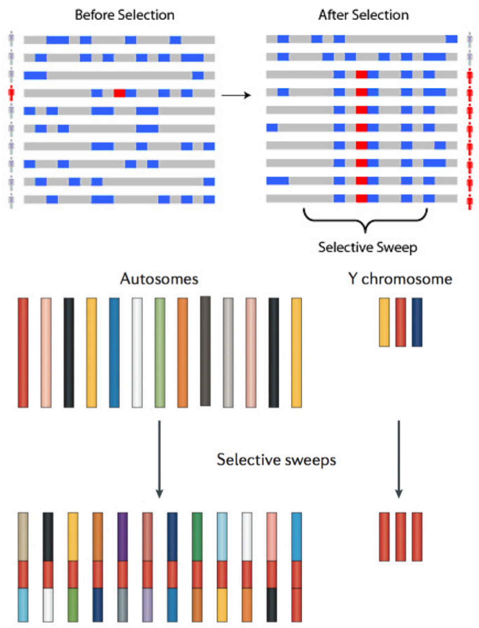


Figure 3 | **Genetic diversity affected by the density of targets for selection and by recombination rate.** A schematic illustration of the effects of linked selection on genetic (nucleotide) diversity around genes or other functional elements (boxes; upper panels). In the lower panels, solid lines indicate the local variation in diversity level and dashed lines indicate the average diversity in the whole region in question. In regions with a high density of targets of selection (part **a**), linked selection is pervasive and significantly reduces diversity compared with regions with a lower density of selection targets (part **b**). When the recombination rate is high (part **c**), the effect of linked selection becomes less prevalent, allowing maintenance of high diversity levels.

6.3.1 Selective sweeps

Selective sweeps happen, when a beneficial allele carries other neutral alleles close to it along through hitchhiking. It results in less diversity in the beneficial alleles vicinity.

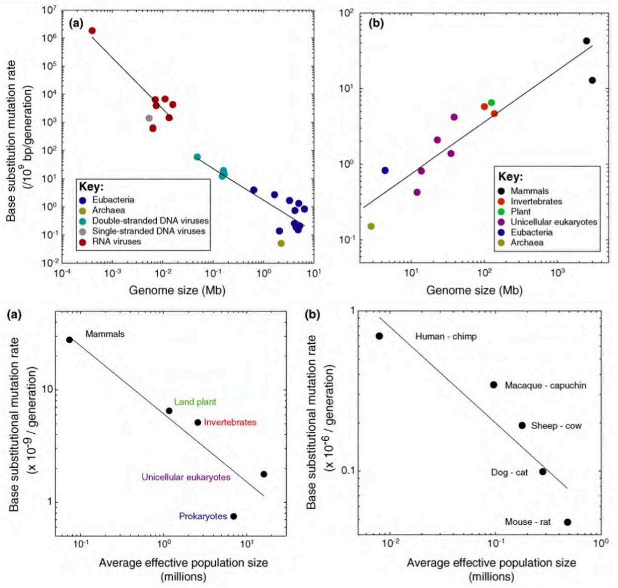


6.4 Loss of variation

Relative population size	Rate at which variation is lost each generation
Haploid	$1/N_e$
Diploid	$1/(2N_e)$
Tetraploid	$1/(4N_e)$
Plastid DNA	$1/N_{ef}$*
mtDNA	$1/N_{ef}$*

*True for taxa in which plastid DNA (including cpDNA) and mtDNA are maternally inherited, since N_{ef} is the effective number of females in the population.

6.5 Mutation rate variation



Mutation accumulation experiments generate multiple lines of one acestor, that reproduce through selfing or inbreeding to accumulate mutations. They useful to estimate mutation rates and mutation variation.

6.5.1 Direct sequencing of families

Analysis of Genetic Inheritance in a Family Quartet by Whole-Genome Sequencing *(Roach et al, Science (2010), 328, 636-639.)* We analyzed the whole-genome sequences of a family of four, consisting of two siblings and their parents. Family-based sequencing allowed us to delineate recombination sites precisely, identify 70% of the sequencing errors (resulting in > 99.999% accuracy), and identify very rare single-nucleotide polymorphisms. We also directly estimated a human intergeneration mutation rate of approximately 1.1×10^{-8} per position per haploid genome. Both offspring in this family have two recessive disorders: Miller syndrome, for which the gene was concurrently identified, and primary ciliary dyskinesia, for which causative genes have been previously identified. Family-based genome analysis enabled us to narrow the candidate genes for both of these Mendelian disorders to only four. Our results demonstrate the value of complete genome sequencing in families.

7 Genetic architecture of adaptive traits

7.1 Linking phenotype with genotype

A fundamental problem in evolutionary biology and ecological genetics is to understand the genetic basis of adaptation and adaptive traits in natural populations.

7.1.1 Forward genetics

A forward genetics approach investigates the genetic basis of a phenotypic trait. Classical forward genetic screens start by mutagenizing individuals. Those with the phenotype of interest are sought and the mutated gene is identified. Detailed studies of the mutant, together with molecular analyses of the gene allow identification of gene function.

7.1.2 Reverse genetics

The goal of a reverse genetics approach is to identify the phenotype(s) that are associated with particular nucleotide sequences. Using various techniques, a gene's function is altered and the effect on the development or behaviour of the organism is analysed. Methods include gene inactivation through RNA interference (RNAi), gene silencing mediated by virus infection (virus-induced gene silencing, VIGS) or gene inactivation via CRISPR/Cas9.

7.1.3 Reverse ecology

Reverse ecology is analogous to reverse genetics and is the application of genomic approaches to living systems to uncover the genetic bases of functional variation in nature. By identifying genetic polymorphisms that are associated with a particular habitat or phenotypic trait, one can find targets of natural selection – without a priori knowledge about how selection acted and without knowing the trait that was the target of selection.

7.2 QTL Analysis

The term **quantitative trait locus (QTL)** refers to a specific DNA region that influences the expression of a quantitative phenotype (trait). They are typically identified in the offspring of crosses between parental individuals that differ clearly in the traits of interest. QTL analysis can be considered a forward genetics approach. Results of QTL analyses provide information about the genetic architecture of traits, including:

- The number of loci that contribute to trait variation
- The positions of these loci in the genome
- Their effect sizes
- Interactions (additive and epistatic) among loci

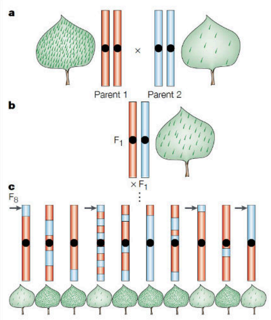
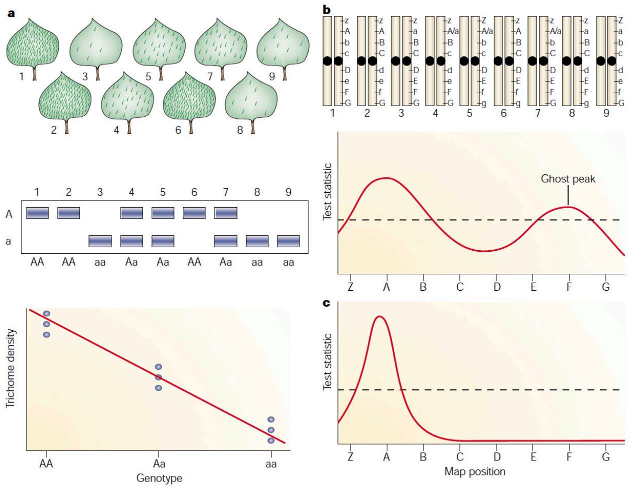


Figure 1 | **Principles of mapping quantitative trait loci.** The basic strategy behind mapping quantitative trait loci (QTL) is illustrated here for **a** | the density of hairs (trichomes) that occur on a plant leaf. Inbred parents that differ in the density of trichomes are crossed to form an F_1 population with intermediate trichome density. **b** | An F_1 individual is selfed to form a population of F_2 individuals. **c** | Each F_2 is selfed for six additional generations, ultimately forming several recombinant inbred lines (RILs). Each RIL is homozygous for a section of a parental chromosome. The RILs are scored for several genetic markers, as well as for the trichome density phenotype. In **c**, the arrow marks a section of chromosome that derives from the parent with low trichome density. The leaves of all individuals that have inherited that section of chromosome from the parent with low trichome density also have low trichome density, indicating that this chromosomal region probably contains a QTL for this trait.



Limitations of QTL mapping include

- Controlled crosses are either impossible or time-consuming in many species
- Genetic variation in the mapping population is restricted with only two (or four, e.g. in outcrossing species) parents used to initiate the QTL mapping population
- Resolution is limited because typically early-generation crosses are used and the number of recombination events per chromosome is small.
- QTL intervals often span tens of centiMorgans and thus several Megabases and correspondingly many genes (tens to thousands).
- Many QTL regions contain multiple, closely-linked QTLs that may have smaller or even opposite effects.
- Phenotypes and their QTL are frequently affected by interactions, e.g. epistatic interactions, genotype-by-sex, genotype-by-environment,but most QTL studies do not allow testing for such effects.
- Effect sizes are difficult to estimate (Beavis effect) and alleles withsmall effect sizes are typically not identified.

7.2.1 LOD Score

Statistical support for a QTL at a specific position is estimated using the LOD score

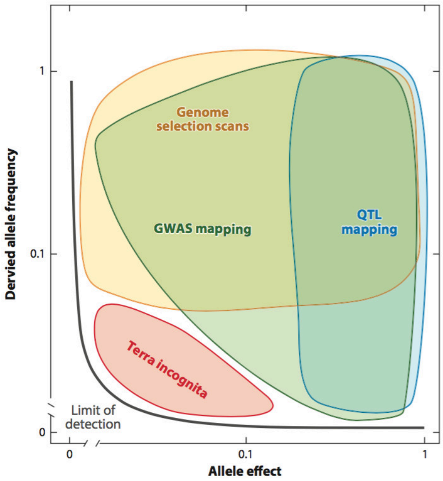
$$\text{LOD} = log_{10}\frac{L_1}{L_0}$$

7.3 Linkage mapping

Linkage maps are constructed using large numbers of segregating markers and information about the recombination rate (linkage disequilibrium) between different markers in a population with known pedigree. Distances between markers do not indicate physical distances (e.g. in Mb), but instead indicate recombination distances, are estimated from observed recombination frequencies, and are typically given in cM (centiMorgan). Difficulties in constructing such linkage maps include

- Sufficient number of suitable markers
- Estimates of recombination distance may be biased by multiple (e.g. double) crossovers
- Interference: the interaction between neighboring crossover events
- Incomplete information from some offspring genotypes when parents are not fully homozygous. This requires statistical estimate of recombination distance between markers.

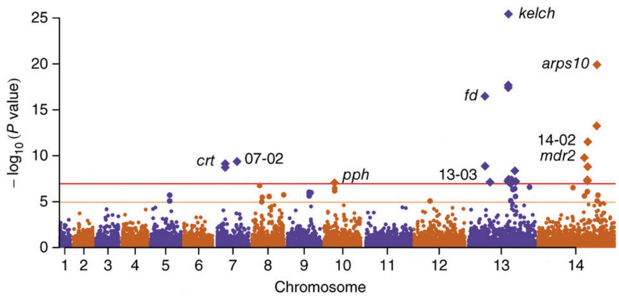
7.4 Methods selection



7.5 Genome-wide assiciation studies

In GWAS, also known as genome-wide association mapping, the **association between each genotyped marker and the phenotype of interest** scored across a large number of individuals is analysed. They can provide insights into trait architecture as well as candidate genes for certain traits or for functional analysis and can compliment QTL analysis. They also provide much higher resolution because associations in natural populations reflect historical recombination events.

1) Select panel of accessions
2) Assess genomic variants
3) Phenotype traits of interest
4) Run statistical methods
5) Rank candidates
6) Validate genes



The power of GWAS to identify a true association is dependent on the phenotypic variance within the population explained by the SNP. The phenotypic variance is determined by how strongly the two allelic variants differ in their phenotypic effect (the effect size), and their frequency in the sample. Problems for GWAS are caused by population structure, rare variants or small effect sizes. The effects of rare variants may be easier to analyse in a QTL experiment because crossing elevates rare variants to intermediate frequency.

Key challenges for GWAS in natural populations are, that quantitative phenotypes are often strongly affected by the environment, outcrossing organisms often have low LD and the lack of *a priori* knowledge of trait architecture complicates planning.

GWAS can only identify a small fraction of causal genes. This is called the problem of **missing heritability**. Possible explanations are

- Complex epistatis interactions among genes are relevant for many traits
- Studies do not have suffiecient power to detect small-effect loci
- Importance of epigenetic variation ignored
- Genetic effect due to rare mutations
- Traits are diagnosed incorrectly or inconsistently

7.5.1 GWAS and population structure

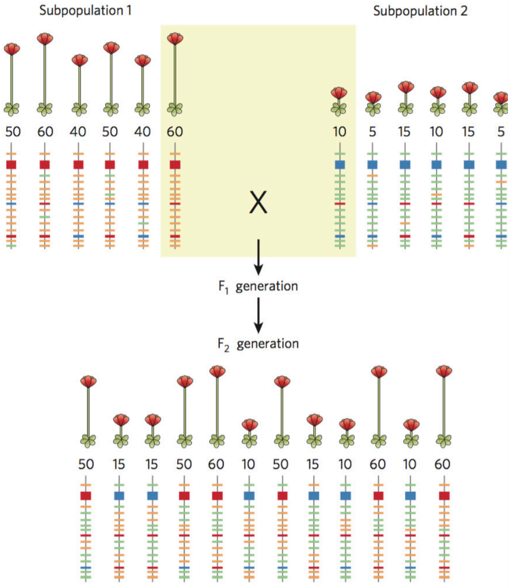


Figure 1 | GWA mapping is ineffective if there is strong genetic differentiation between subpopulations (that is, if there is structure in the population). In this example, two subpopulations of plants are depicted, one tall and one short (as illustrated and indicated by the numerical measurement), together with a schema of the genotype of each plant. The presence of red alleles increases the height of a plant, whereas blue alleles decrease the height; one locus has a major effect, and two have a minor effect. The many background markers (orange and green) are mostly exclusive to a specific subpopulation but are also strongly associated with height, even though they are not causal. By crossing the plants (shaded area) and generating an experimental population of F_2 generation or recombinant inbred lines, any linkage disequilibrium between background markers and causal markers is broken up, and the causal loci can then easily be mapped, albeit with relatively poor resolution.

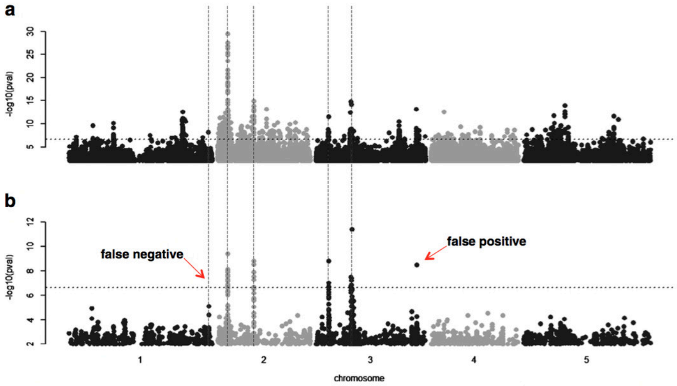
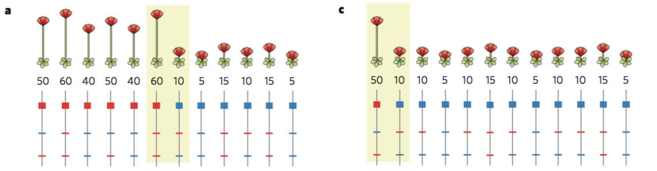


Figure 3 Taking genetic background into account improves the performance of GWAS. Manhattan plots for a simulated trait, in which each data point represents a genotyped SNP, ordered across the five chromosomes of *Arabidopsis*. Five SNPs (indicated by vertical dashed lines) were randomly chosen to be 'causative' and account for up to 10% of the phenotypic variance each. GWAS using **a)** a linear model, and **b)** a mixed model that accounts for population structure and other background genomic factors. The simple linear model leads to heavily inflated p-values and the five causative markers are not the strongest associations. The mixed model is superior, but still leads to one false negative and one false positive. A dashed horizontal line denotes the 5% Bonferroni threshold.

7.5.2 GWAS and trait structure



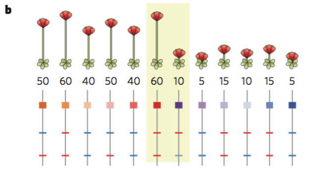


Figure 2 | The performance of GWA mapping depends strongly on the genetic architecture of the trait. Plants are depicted as in Fig. 1. Red alleles increase plant height, whereas blue alleles decrease height; one locus has a major effect, and two have a minor effect. **a**, Provided that there is no confounding effect of population structure, the major causal locus of a trait can easily be mapped by a GWA scan (otherwise an approach such as that in Fig. 1 is needed). **b**, Even if there is a major locus for the trait, the presence of many different alleles of the causal gene (multiple colours) makes naive GWA scans (such as in **a**) ineffective. **c**, GWA scans miss rarely occurring alleles that have a large effect (the top locus) but locate common alleles with smaller effects (the centre locus). In all three cases (**a**–**c**), the causal loci can be identified in experimental crosses (shaded areas) by using linkage mapping instead of GWA mapping, but alleles identified in this way might be rare in the population and therefore not informative for most individuals and not representative of the population.