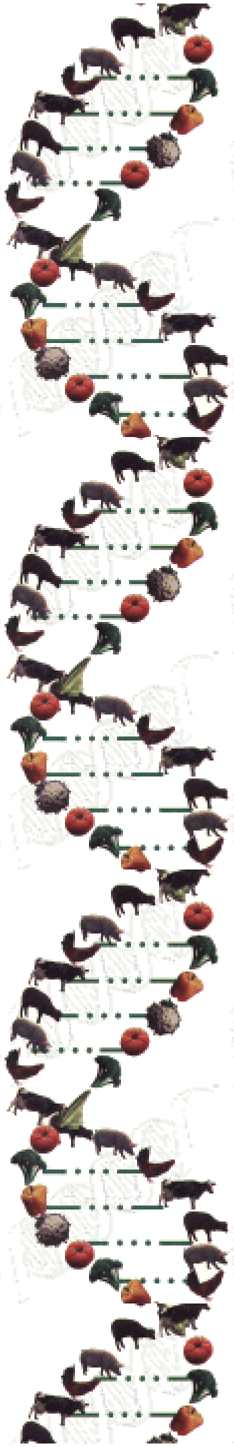
ГЕНОМИКА (6 лекций и семинаров, 30 акад. часов)

- ***Лектор:*** Константин Валерьевич Крутовский
  - Профессор Гёттингенского университета, Германия (http://www.uni-goettingen.de/en/414626.html) и Техасского АМ университета (http://essm.tamu.edu/people/faculty/adjunct-faculty/krutovsky-konstantin
  - Ведущий научный сотрудник Института общей генетики им. Н. И. Вавилова РАН, Москва
  - Профессор базовой кафедры защиты и современных технологий мониторинга, зав. лабораторией лесной геномики и руководитель Научно-образовательного центра геномных исследований Сибирского федерального университета, Красноярск (http://genome.sfu-kras.ru/en/krutovsky)
- ***Тел.:*** +7 965 912 1959 (моб. для вотцапа) ***Скайп***: k-krutovsky
- ***E-mail:*** kkrutovsky@gmail.com или kkrutov@gwdg.de
- ***Office:*** ЦЗЛ, Академгородок 50а, корп.2, ауд. 133
- ***Office Hours:*** You can contact by e-mail or phone to make an appointment.
- ***Textbook (not required)***.
- ***Required e-mail***: You will need to send an e-mail to kkrutov@gwdg.de from your preferred address. This would allow me to distribute class announcements, lecture notes, readings, problem sets, etc. Please, provide me with:
  - your full name
  - name by which you prefer to be called
  - phone number(s) where you can be reached
  - e-mail address that you check daily
  - academic & career interests
  - what you hope to get from this course
- ***Course web page***: http://genome.sfu-kras.ru/ru/lectures
- ***Lecture notes***: PowerPoint lecture notes for most of the class sessions will be available on the Web site prior to each class session. **I expect you to print out and bring the notes with you to class (bring also your laptop with you to class)**

ТЕМЫ ЛЕКЦИЙ ПО КУРСУ «ГЕНОМИКА»

№ п/п	Наименование лекций	Объем в акад. часах	Дата и время проведения
1	**Введение в геномику.** Содержание и организация геномной информации. Геномика, транскриптоника, протеомика, метаболомика. Программа "Геном человека"	4	Пон. 16.03 16:00-16:45 16:50-17:35 18:35-19:20 19:25-20:10
2	**Технология секвенирования ДНК.** Полногеномное *de novo* секвенирование, ресеквенирование, целевое и метагеномное секвенирование.	6	Сре. 18.03 15:00-21:10
3	**Методы работы с нуклеотидными сиквенсами и геномными базами данных.** Программа поиска гомологий – BLAST. Формат Genbank, выравнивание (BioEdit) и аннотирование нуклеотидных последовательностей (BLAST2GO, Augustus).	4	Пон. 23.03 16:00-20:10
4	**Популяционная геномика.** Генотипирование ДНК-полиморфизмов (SSRs, SNPs). Тесты на селективную нейтральность (DNASP). Гены-аутсайдеры (LOSITAN).	6	Сре. 25.03 15:00-21:10
5	**Полногеномное ассоциативное картирование.** Подходы и методы полногеномного ассоциативного картирования (TASSEL).	4	Пон. 30.03 16:00-20:10
6	**Практические приложения геномики**: филогеномика, экогеномика, природоохранная геномика, палеогеномика, персонифицированная медицина, геронтогеномика, метагеномика, эпигеномика и геномная селекция.	6	Сре. 1.04 15:00-21:10



ГЕНОМИКА

16 марта 2020, Понедельник

1. **Введение в геномику.** Содержание и организация геномной информации. Геномика, транскриптоника, протеомика, метаболомика.

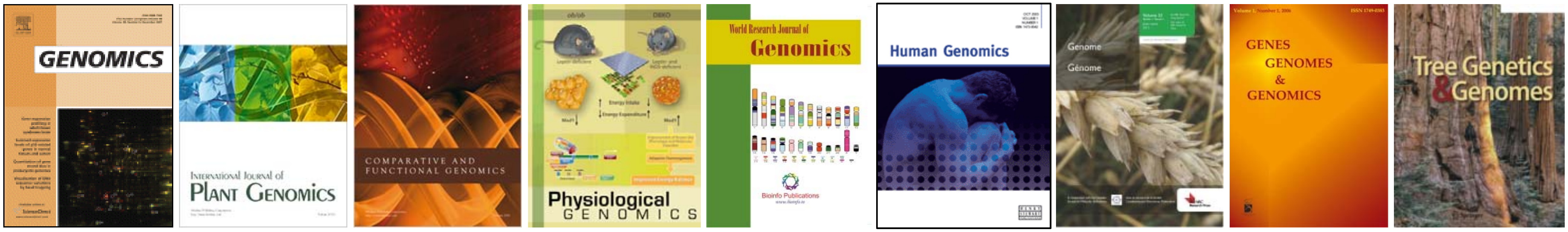
18 марта 2020, Среда

2. **Технология секвенирования ДНК.** Полногеномное *de novo* секвенирование, ресеквенирование, целевое и метагеномное секвенирование.

What is Genomics?

- The term "genome“ was introduced in 1920 by German botanist Hans Winkler (1877-1945) who combined “gene” and “chromosome” to refer to all genes on all chromosomes in the nucleus of a cell.

- The term "genomics" was coined in 1986 by Thomas Roderick (Jackson Laboratory, USA) to describe the scientific discipline of mapping, sequencing, and analyzing genomes and to provide a name for the new journal *Genomics*.


GENOMICS
International Journal of Plant Genomics
Comparative and Functional Genomics
Physiological Genomics
World Research Journal of Genomics
Human Genomics
Genome
Génome
Genes Genomes & Genomics
Tree Genetics & Genomes

- Genomics is more comprehensive now and includes comparison of individual genomes in a population or from different populations (population genomics) and species (comparative genomics), study of their evolution (evolutionary genomics) and function (functional genomics).

Genomics studies genomes, genes and their functions using global genome-wide experimental approaches

Milestones in Genetics that led to Genomics

1944: DNA was identified as the genetic material of all living organisms (Avery et al., J. Exp. Med. **79**, 137: 1944)

1953: The genetic code was deciphered (Watson & Crick, *Nature* **171**, 737: 1953).

1977: The first complete DNA sequence of an entire genome - the bacteriophage phiX174 (Sanger et al. 1977 *Nature* **265**, 687-695) of only 5386 nucleotides, which is 60000 times smaller than the human nuclear genome.

mid-1980s: Major advances in DNA sequencing, laboratory automation & computing.

1990: The project on complete sequencing of the human genome was launched.

1997: complete sequencing of the yeast genome (12 Mbp)

1998: nematode genome (97 Mbp)

2000: fruit fly (180 Mbp) & the first plant *Arabidopsis* (125 Mbp) genomes

2001: human genome (3,200 Mbp)

2002: mouse (3,500 Mbp) & rice (420 Mbp) genomes

2006: the first forest tree - poplar genome (550 Mbp)

mid-2000s: Next generation sequencing (NGS) platforms - high-throughput massively parallel sequencing

The Genomes OnLine Database (GOLD, https://gold.jgi-psf.org):

- Complete Projects: 19,165
- Organisms: 373,486 (Archaea: 3,691; Bacteria: 328,310; Eukarya: 32,234; Viruses 9,251)
- Incomplete Projects: 102,049
- Targeted Projects: 1,082

https://www.earthbiogenome.org

2013: neandertal genome (3,200 Mbp)

2013: Norway spruce (~19,570 Mbp) 2014: Loblolly pine (~21,610 Mbp)

2016: Sugar pine (~28,900 Mbp) 2019: Siberian larch (~12,340 Mbp)

Major areas of Genomics

Structural Genomics

- ***DNA libraries and complete genome sequence***
- ***Gene annotation and homology search***
- ***Linkage analysis, genetic and physical mapping***
- ***Development of genome-wide genetic markers***

Functional Genomics

- ***Gene expression analysis*** (*transcriptome, proteome & metabolome profiling*)
- ***Gene function, gene-trait*** *and gene-environment* ***relationships***

Comparative & Evolutionary Genomics

- ***Comparative mapping and search for orthology and synteny***
- ***Gene and sequence comparison across different species***
- ***Signatures of selection, evolutionary footprints***

Gene discovery

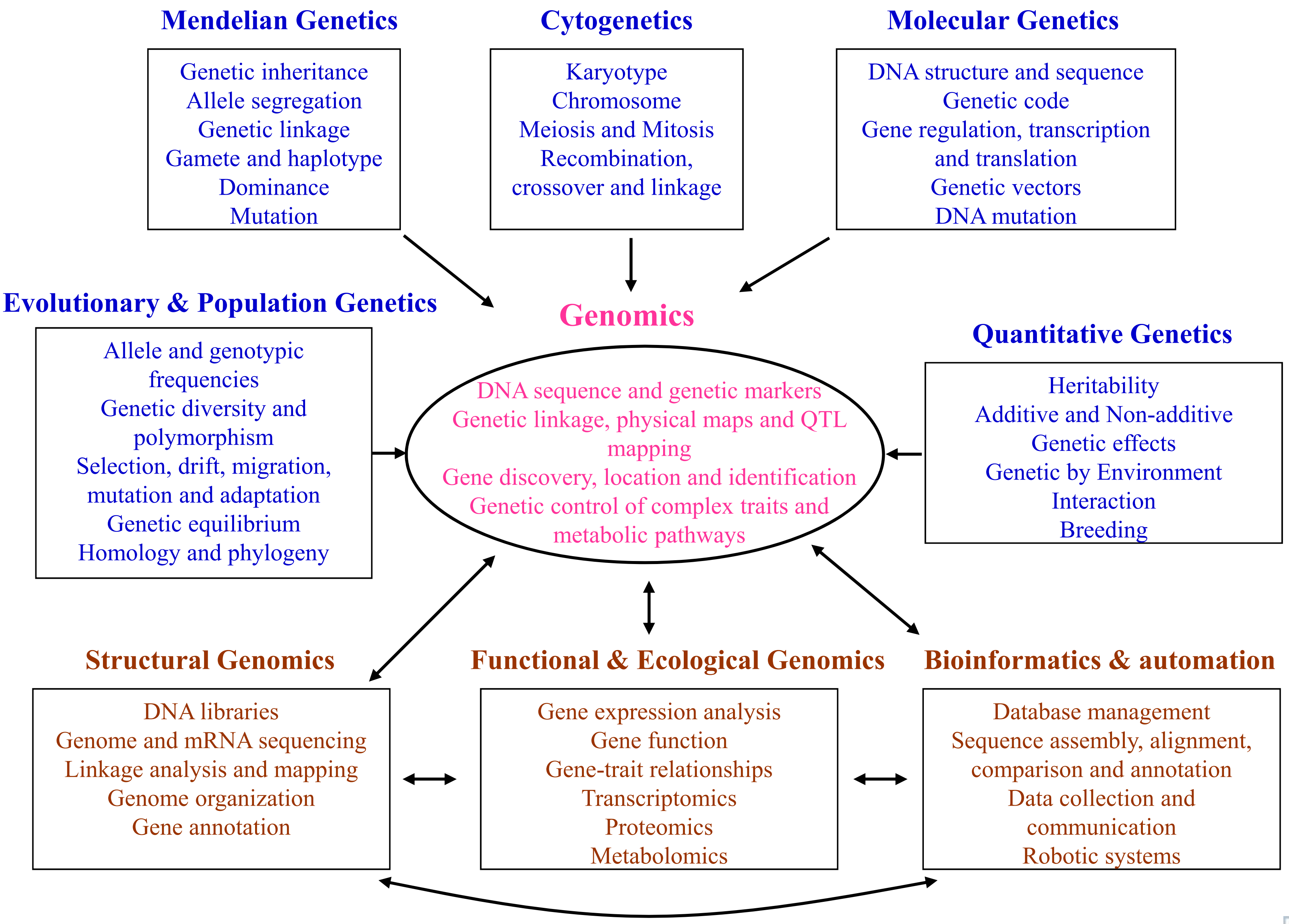
Statistical Genomics

- ***Mapping algorithms and associative analysis***
- ***Database management, data collection and communication***
- ***Sequence assembly, alignment, comparison and annotation***

Population & Ecological Genomics

- ***Genome wide scan for nucleotide diversity***
- ***Genome wide and candidate gene based association mapping***
- ***Assessment of association between alleles and phenotypes and environments via association mapping***

Genomics as an integrative science



Mendelian Genetics

Genetic inheritance
Allele segregation
Genetic linkage
Gamete and haplotype
Dominance
Mutation

Cytogenetics

Karyotype
Chromosome
Meiosis and Mitosis
Recombination,
crossover and linkage

Molecular Genetics

DNA structure and sequence
Genetic code
Gene regulation, transcription
and translation
Genetic vectors
DNA mutation

Evolutionary & Population Genetics

Allele and genotypic
frequencies
Genetic diversity and
polymorphism
Selection, drift, migration,
mutation and adaptation
Genetic equilibrium
Homology and phylogeny

Genomics

DNA sequence and genetic markers
Genetic linkage, physical maps and QTL
mapping
Gene discovery, location and identification
Genetic control of complex traits and
metabolic pathways

Quantitative Genetics

Heritability
Additive and Non-additive
Genetic effects
Genetic by Environment
Interaction
Breeding

Structural Genomics

DNA libraries
Genome and mRNA sequencing
Linkage analysis and mapping
Genome organization
Gene annotation

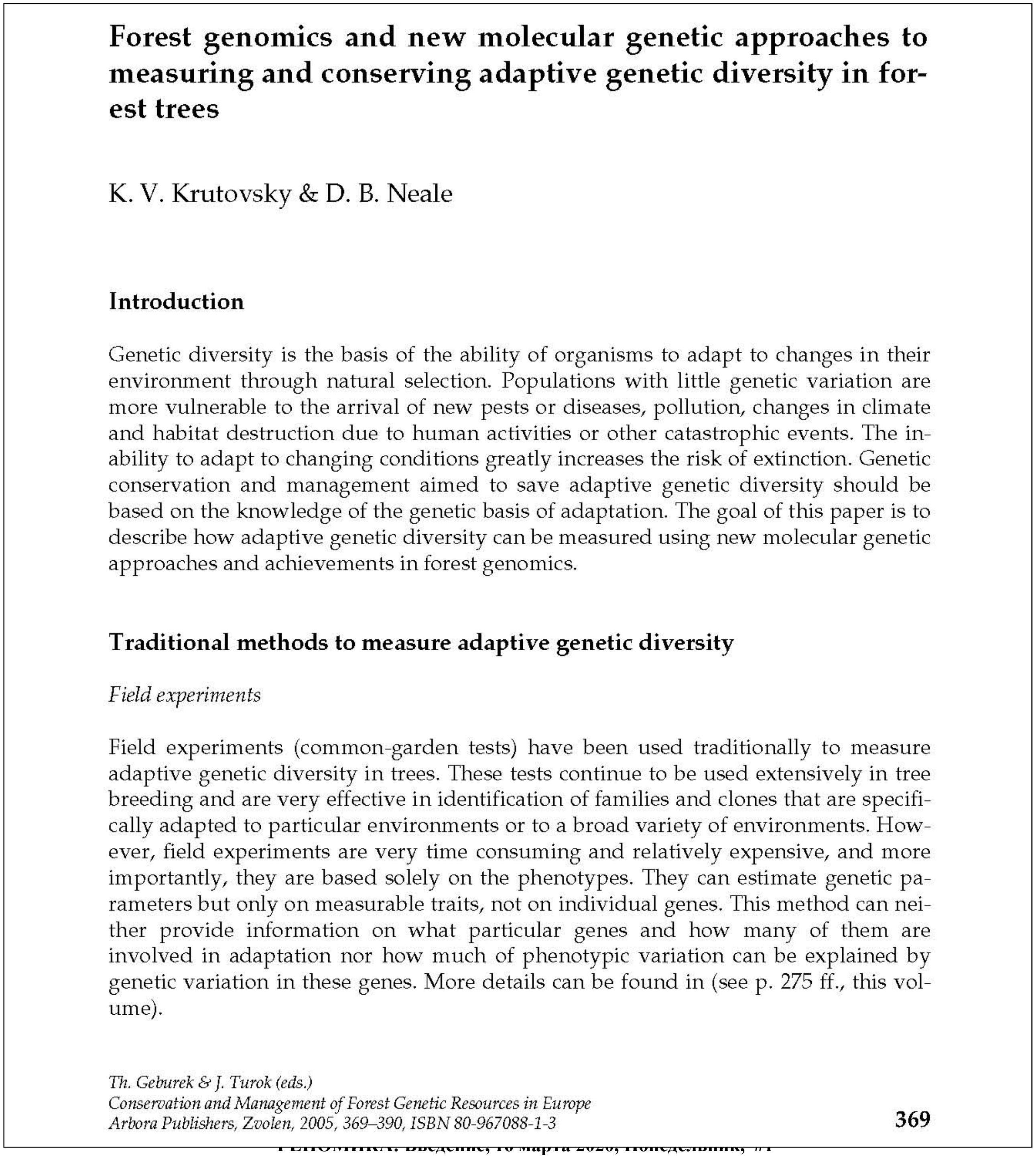
Functional & Ecological Genomics

Gene expression analysis
Gene function
Gene-trait relationships
Transcriptomics
Proteomics
Metabolomics

Bioinformatics & automation

Database management
Sequence assembly, alignment,
comparison and annotation
Data collection and
communication
Robotic systems

Krutovsky K.V. & D.B. Neale. 2005 Forest genomics and new molecular genetic approaches to measuring and conserving adaptive genetic diversity in forest trees, pp. 369-390 in *Conservation and Management of Forest Genetic Resources in Europe*, edited by Th. Geburek and J. Turok. Arbora Publishers, Zvolen.



Forest genomics and new molecular genetic approaches to measuring and conserving adaptive genetic diversity in forest trees

K. V. Krutovsky & D. B. Neale

Introduction

Genetic diversity is the basis of the ability of organisms to adapt to changes in their environment through natural selection. Populations with little genetic variation are more vulnerable to the arrival of new pests or diseases, pollution, changes in climate and habitat destruction due to human activities or other catastrophic events. The inability to adapt to changing conditions greatly increases the risk of extinction. Genetic conservation and management aimed to save adaptive genetic diversity should be based on the knowledge of the genetic basis of adaptation. The goal of this paper is to describe how adaptive genetic diversity can be measured using new molecular genetic approaches and achievements in forest genomics.

Traditional methods to measure adaptive genetic diversity

Field experiments

Field experiments (common-garden tests) have been used traditionally to measure adaptive genetic diversity in trees. These tests continue to be used extensively in tree breeding and are very effective in identification of families and clones that are specifically adapted to particular environments or to a broad variety of environments. However, field experiments are very time consuming and relatively expensive, and more importantly, they are based solely on the phenotypes. They can estimate genetic parameters but only on measurable traits, not on individual genes. This method can neither provide information on what particular genes and how many of them are involved in adaptation nor how much of phenotypic variation can be explained by genetic variation in these genes. More details can be found in (see p. 275 ff., this volume).

Th. Geburek & J. Turok (eds.)
Conservation and Management of Forest Genetic Resources in Europe
Arbora Publishers, Zvolen, 2005, 369–390, ISBN 80-967088-1-3

Climate change is a global systemic threat to ecosystems and needs a systemic (holistic) approach

- To be able to predict and mitigate effects of climate change and to breed resilient crops we have to understand **evolutionary responses** and **molecular mechanisms of genetic adaptation.**

- **Evolutionary response** is a **genetic adaptation** via genetic change that promotes adaptation of plants and animals to their natural environment, including their ecosystem interactions with members of their own and other species (the biotic environment) as well as the physical environment (the abiotic environment).

- Multiple genes are involved in **genetic adaptation**, so its study requires genomic methods and genome-wide approaches.

How genomics can help us identify evolutionary responses and study molecular mechanisms of genetic adaptation?

- **Structural genomics** provides practically unlimited number of genetic markers for population genetic studies via new methods of high-throughput massively parallel sequencing and genotyping.

- **Population genomics** 1) provides detailed nucleotide and allelic variation in numerous adaptive trait related candidate genes, 2) identifies genes under selection (via **whole genome scans**, **neutrality tests, outliers**, etc.), and 3) links genotypes to phenotypes (via **genome-wide association study - GWAS**).

- **Ecological genomics** helps to identify genes responsive to environmental factors via genome wide differentiation expression analysis (using mRNA/cDNA sequencing or transcriptome profiling) and associating genotypes with environmental variables.

(**Spatial** or **landscape genomics** is one of applications of **population genomics**

What is population genomics?

- foundation for **ecological genomics** that links genotypes to environment via linking genomic variation with environmental variables

- provides detailed data and description of genome-wide nucleotide and allelic variation for all types of molecular genetic markers (SNPs, SSR, etc.) in numerous selectively neutral loci (supposedly non-coding regions) as well as adaptive trait related candidate genes in multiple individuals and populations

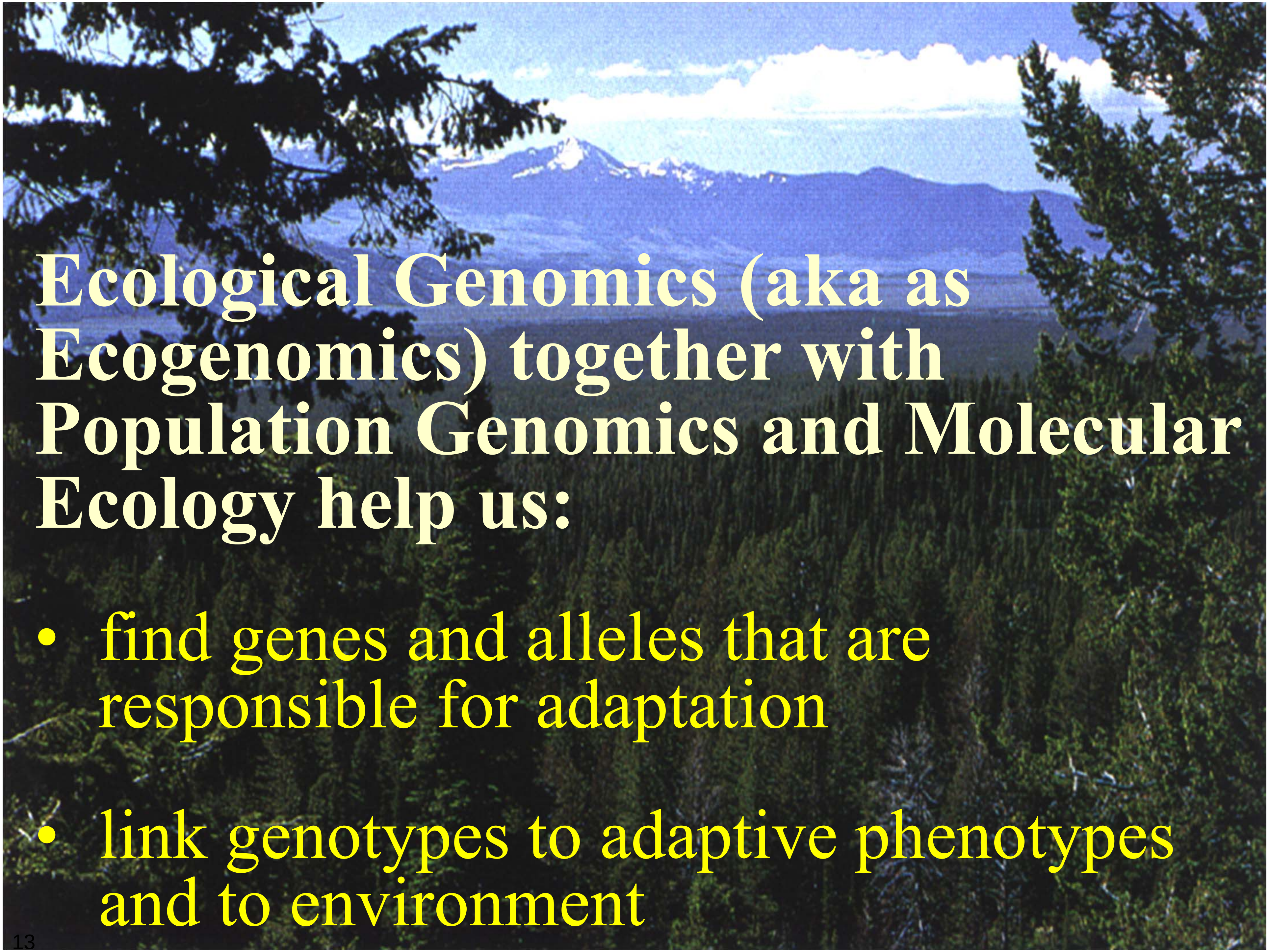
- identifies genes under selection via **neutrality tests**, **outliers**, etc., using **genome wide scans**

- links genotypes to phenotypes via **genome wide association studies (GWAS)**

(**Spatial** or **landscape genomics** is one of applications of **population genomics**).

What is ecological genomics?

- a novel, fast-developing biological discipline, combining traditional ecological and population genetic approaches with the genome-wide level of analysis
- thousands of genes with known function and sometimes known genome-wide localization can be simultaneously studied in many individuals
- applies genomic tools and approaches to traditional ecological and population genetics questions (local adaptation, mating system, gene exchange, reproductive population size, population disequilibrium, interaction among populations and species, population–environment relationships, community composition, conservation and assessment of genetic diversity, etc.)
- these traditional problems of ecological and population genetics can be now studied using data on variation in many genes
- an interdisciplinary approach to a full understanding of interaction between genotypes, phenotypes, and environment
- a truly integrative discipline that embraces many related disciplines, such as ecology, phylogenetics, population and conservation genetics, molecular evolution, etc.



Ecological Genomics (aka as Ecogenomics) together with Population Genomics and Molecular Ecology help us:

- find genes and alleles that are responsible for adaptation

- link genotypes to adaptive phenotypes and to environment

Nature (*Genome*) vs. Nurture (*Environment*)

$$P = G + E + G \times E$$

Phenotype = *Genotype* + *Environment* + *Interaction*

Organisms are different because of the:

- ***genetic* differences among individuals**
- **different *environments* where individuals are growing**
- **and *interactions* between the *genotypes* and the *environments* in which they grow**

Nature (*Genome*) vs. Nurture (*Environment*)

ZEYERS
MY GENOME MADE ME DO IT!
STAYSKAL 2000 TAMPA TRIBUNE

How much can we blame our genotype for our phenotype?

Nature vs. Nurture

Nature

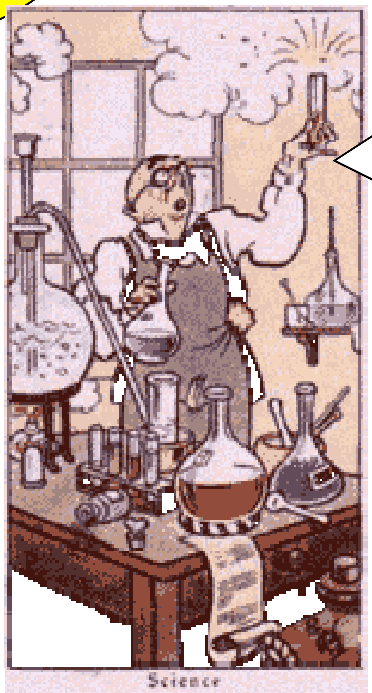
Phenotypic Expression

Nurture

Proteins & Regulatory Factors

Environment

Genome (blueprint)



I found the gene that makes us believe all traits are based on genes

Simple single gene (Mendelian) vs. Complex multiple genes (Quantitative) variation

- Mendelian = Qualitative
  - single gene responsible for most of the observed phenotypic variance

- Complex = Quantitative
  - with gene × gene, gene × environment interactions contributing to phenotypic variance

Single vs. Multiple Genes in Population

$$P^n = G^n + E^n + G^n \times E^n$$

n – multiple phenotypes, genes and environments

Great Fermat's Theorem: $Z^n = X^n + Y^n$

does not have integer solutions X, Y, Z **for** $n > 2$

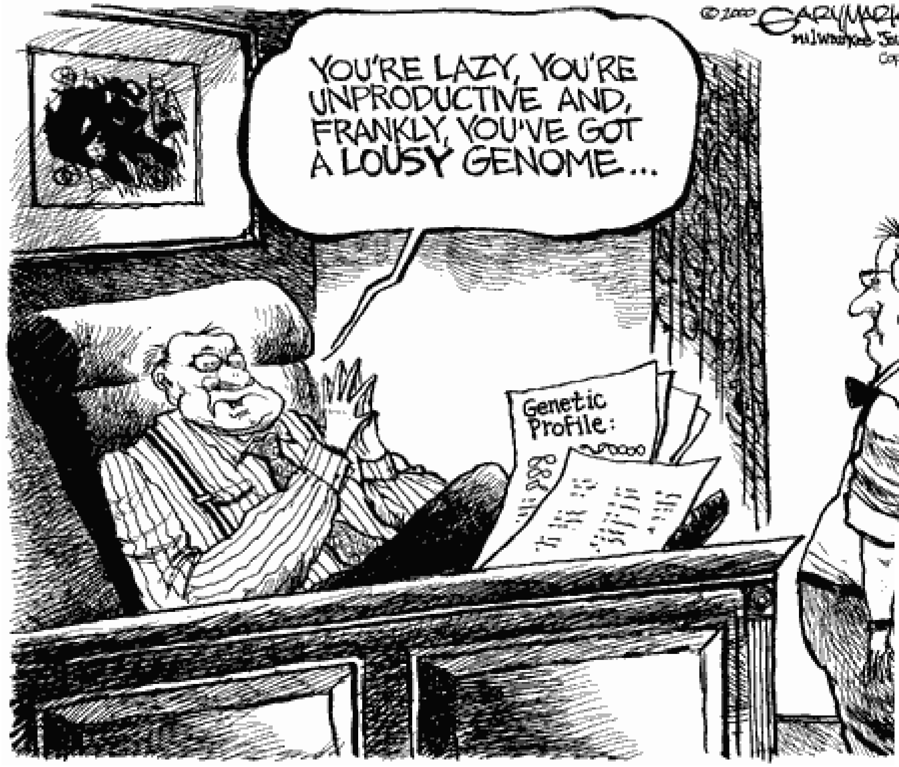
Andrew Wiles, 1994

Life Theorem: $P^z = G^x + E^y$

Great Life Theorem: $P^z = G^x + E^y + G^x \times E^y$

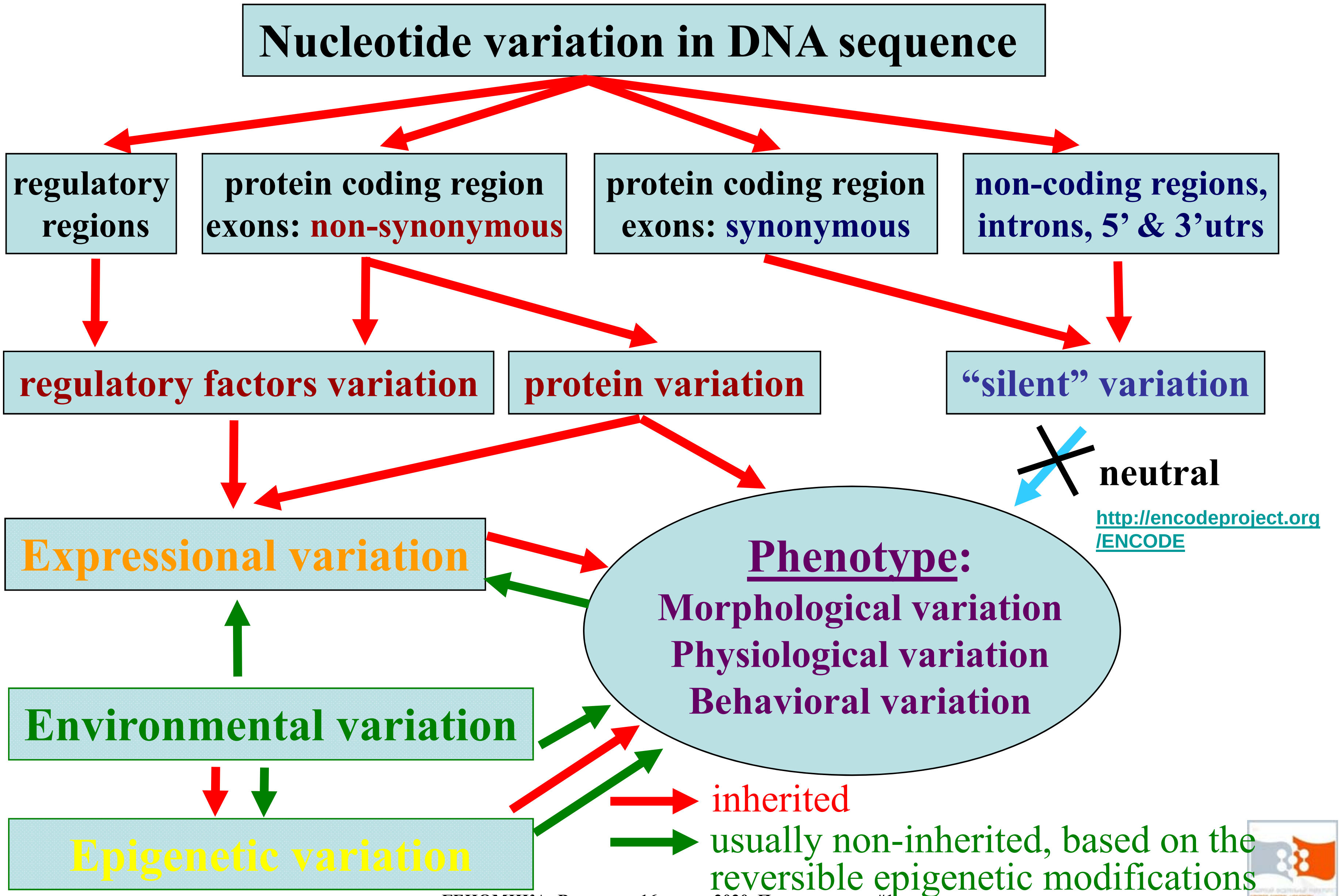
Genomics is the solution!

Can we predict phenotypes based on the genotypes?


©2000 GARY MARKSTEIN
Milwaukee Journal Sentinel
Copley News Service
YOU'RE LAZY, YOU'RE UNPRODUCTIVE AND, FRANKLY, YOU'VE GOT A LOUSY GENOME...
Genetic Profile:

JOB INTERVIEW 2010
..YOUR DNA CHIP WILL TELL US IF YOU HAVE A PREDISPOSITION TO CERTAIN ILLNESSES, OUR LAB TECHS ARE TRYING TO ISOLATE "THE WILLING TO WORK LONG HOURS FOR LITTLE PAY" GENE...
EMPLOYER
WILLIS ©2000 SAN JOSE MERCURY NEWS 6/20/00

Expression of genetic variation



Nucleotide variation in DNA sequence

regulatory regions

protein coding region exons: non-synonymous

protein coding region exons: synonymous

non-coding regions, introns, 5’ & 3’utrs

regulatory factors variation

protein variation

“silent” variation

neutral

http://encodeproject.org/ENCODE

Expressional variation

Phenotype:

Morphological variation

Physiological variation

Behavioral variation

Environmental variation

Epigenetic variation

inherited

usually non-inherited, based on the reversible epigenetic modifications

How to measure evolutionary response in populations and ecosystems?

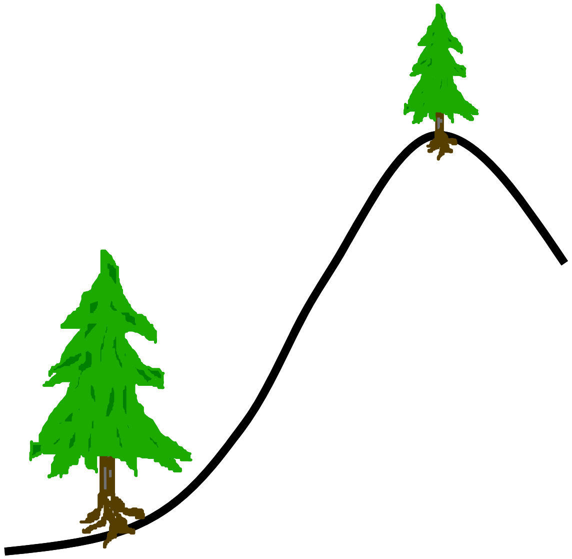
- **Traditional methods**
  - field or common garden experiments (provenance, progeny and clonal tests)
  - molecular genetic markers: advantages and disadvantages; summary statistics
  - Quantitative Trait Locus (QTL) Mapping
- **Modern population genomics approaches**
  - new type of functional genomic markers
  - use of adaptive trait related candidate genes in population studies
  - association mapping with phenotypic and environmental variation
  - detecting selective signatures and loci under adaptive genetic divergence in natural populations using neutrality tests and outlier-detection approaches
  - differential expression, transcriptome profiling, etc.

Integrated approach: 1) & 2)

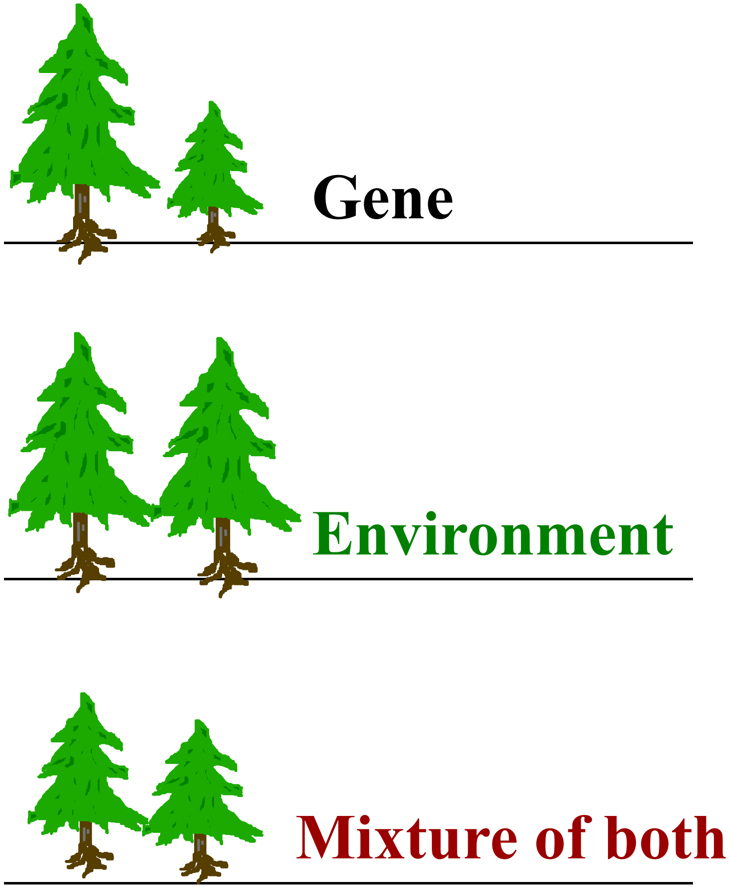
Nature vs. Nurture

How to separate the two?

Example: *Altitudinal gradient*



Common Garden Exp.



Field or common garden experiments (provenance, progeny and clonal tests)

USDA Forest Service Pacific Northwest Research Station runs an ecogeographic study of Douglas-fir

Weyerhaeuser Company runs clonal evaluation of phenotypes in loblolly pine and Douglas-fir

USDA Forest Service Dorena Tree Improvement Center runs a white pine blister rust screening program in sugar pine

Common garden experiments

- helped to learn a lot about patterns of adaptive variation in complex traits, both at the macro- and micro-environmental level
- often shows geographical patterns in populations, such as steep latitudinal or altitudinal clines
- time consuming and relatively expensive, and are based solely on the phenotypes
- can estimate genetic parameters but only on measurable traits, not on individual genes
- can provide neither information on what particular genes and how many of them are involved in adaptation nor how much of phenotypic variation can be explained by genetic variation in these genes

Quantitative Trait Locus (QTL) Mapping

- **One of the first genome-wide approaches to link genotypes and phenotypes**

- **It directs to the chromosomal regions (and sometime genes) responsible for the observed phenotypic variation**

What is a Quantitative Trait Locus (QTL)?

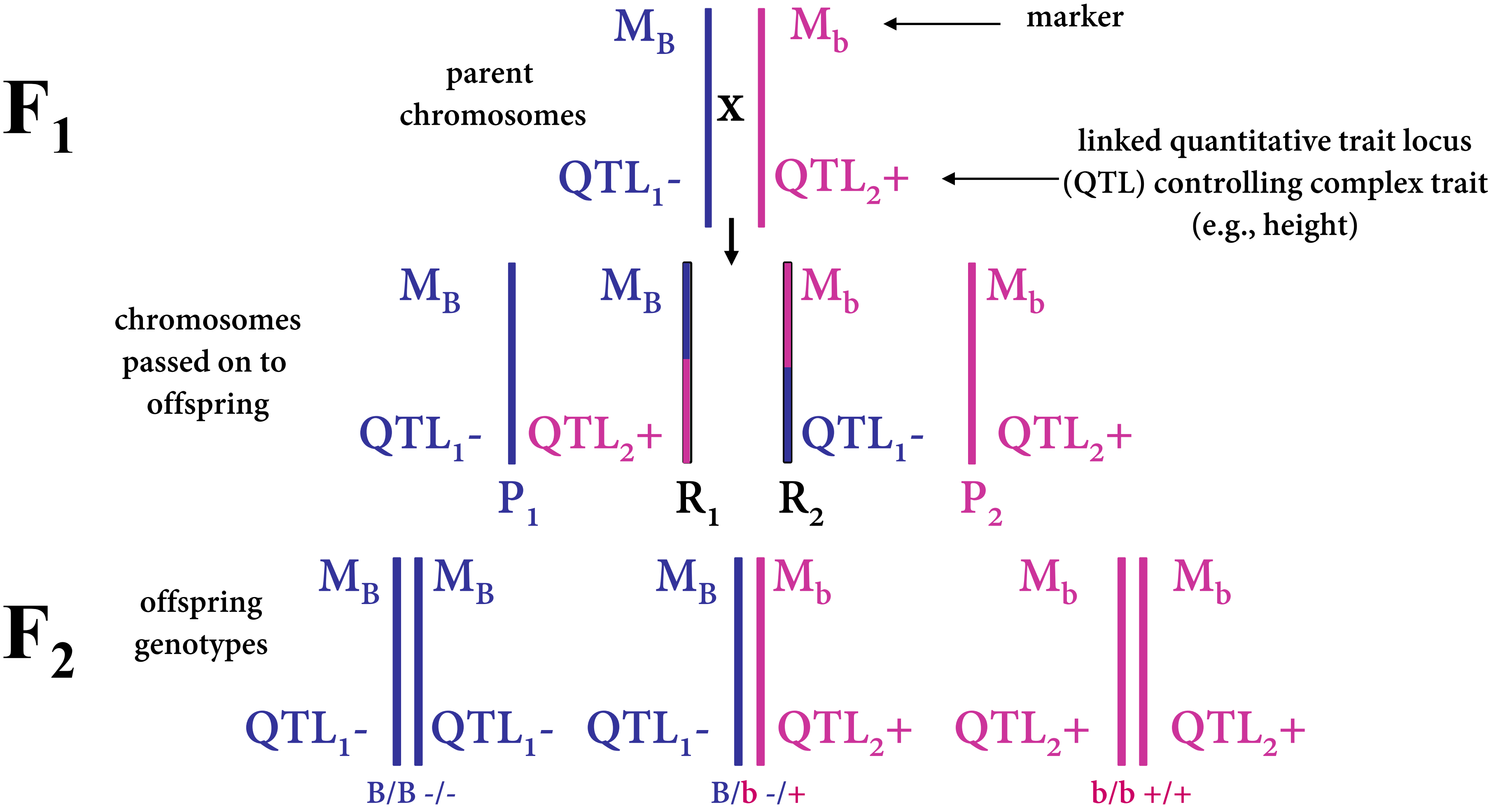
- A *QTL* is a chromosomal region supposedly containing a gene (or cluster of genes) that contributes to the variation observed at a *quantitative trait*

- It must be polymorphic (have allelic variation) to have an effect in a population

- It must be linked to a polymorphic marker allele to be detected

- QTLs are detected through QTL mapping experiments

Quantitative Trait Locus (QTL) mapping



F_1

parent chromosomes

M_B x M_b ← marker

QTL_1- QTL_2+ ← linked quantitative trait locus (QTL) controlling complex trait (e.g., height)

chromosomes passed on to offspring

M_B QTL_1- P_1

M_B QTL_2+ R_1

M_b QTL_1- R_2

M_b QTL_2+ P_2

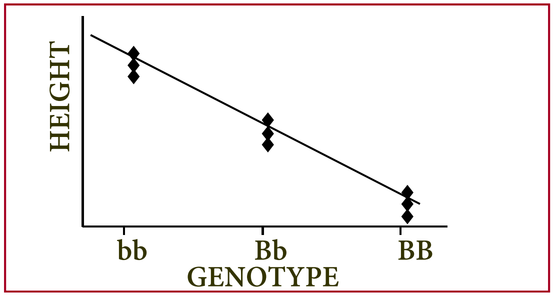
F_2

offspring genotypes

M_B M_B QTL_1- QTL_1- B/B -/-

M_B M_b QTL_1- QTL_2+ B/b -/+

M_b M_b QTL_2+ QTL_2+ b/b +/+



Basic concept

- The closer together are two markers or genes linked on a parental chromosome, the less likely the parental alleles at the two loci will be split up in gametes as a result of meiotic recombination (crossing over).

- Genes and genetic markers that are closely linked together on a chromosome will tend to co-segregate in the $\boldsymbol{F_2}$ - the same allele combinations (haplotypes) that occurred in one of the parents will tend to occur together in the offspring.

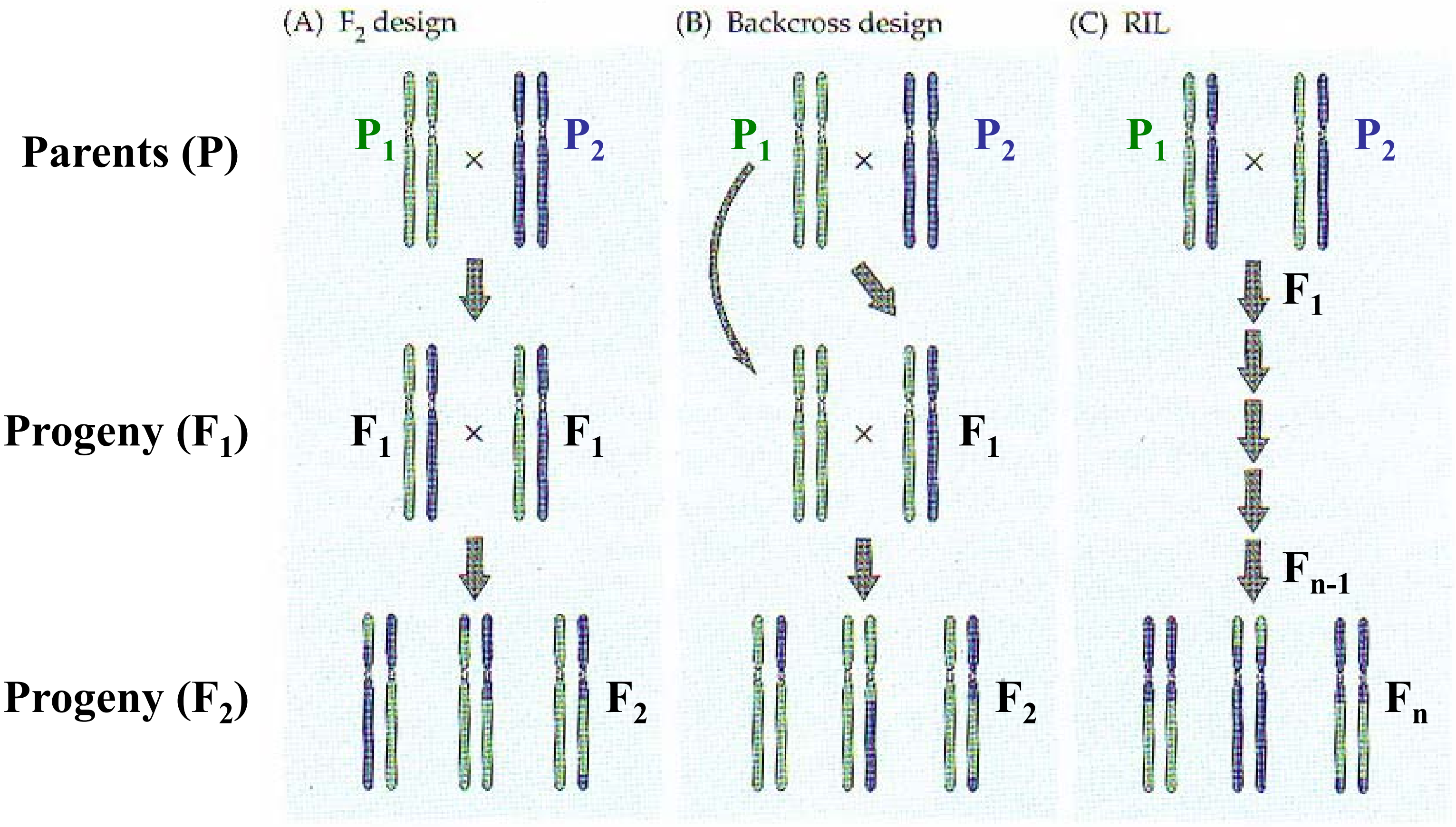
- This will lead to a statistical association between a gene segregating for alleles that have a measurable difference in their affect on a quantitative trait (QTL) and segregating alleles at closely linked marker loci.

QTL mapping populations

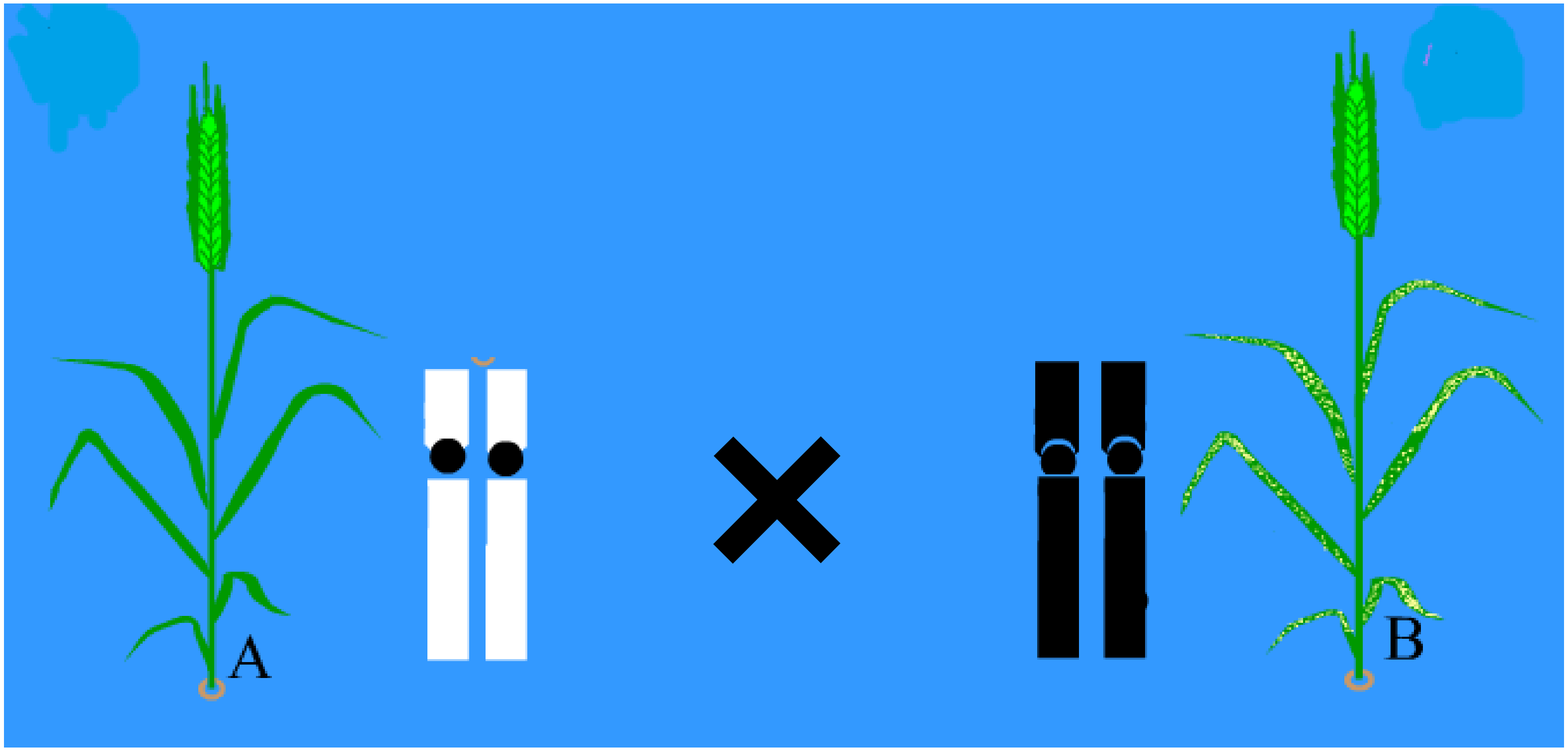
A. F_2 populations

B. Backcrosses

C. Recombinant Inbred Lines (RILs)

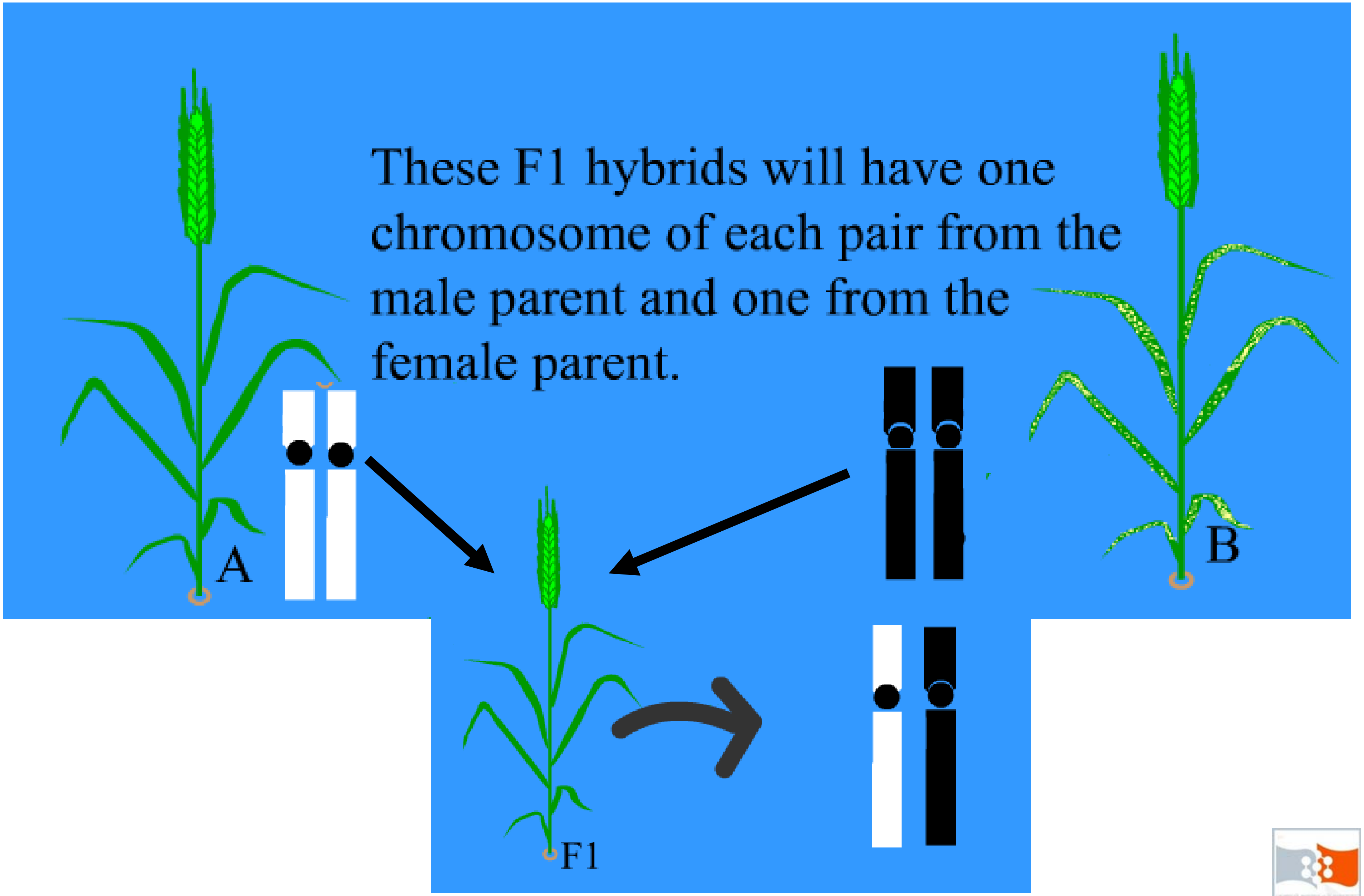


F_2 mapping populations



- First step: cross line A (e.g., disease resistant) with line B
- This can be done with placing pollen from one parent to the stigmas of the other in order to produce hybrid seeds

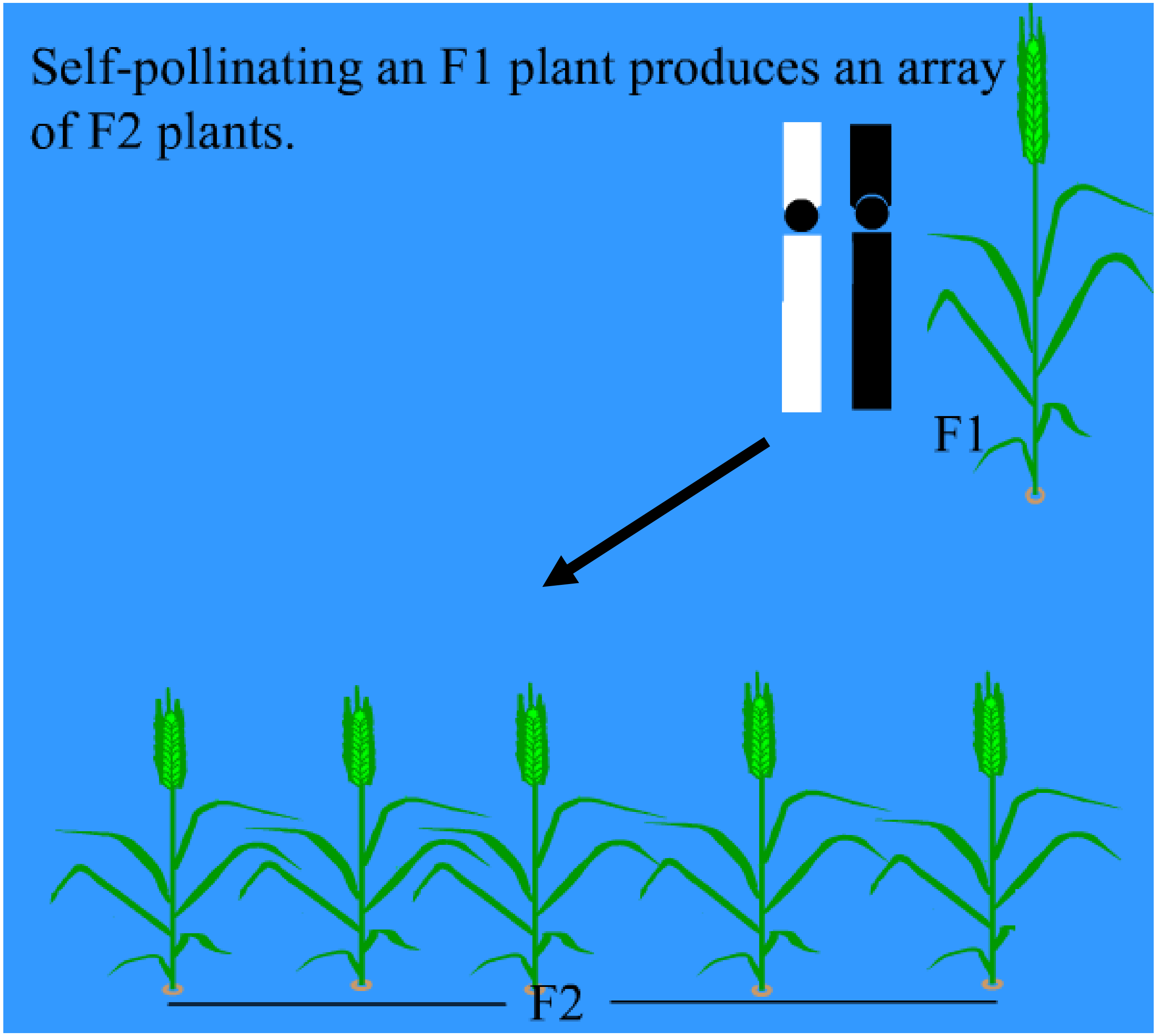
F_2 mapping populations



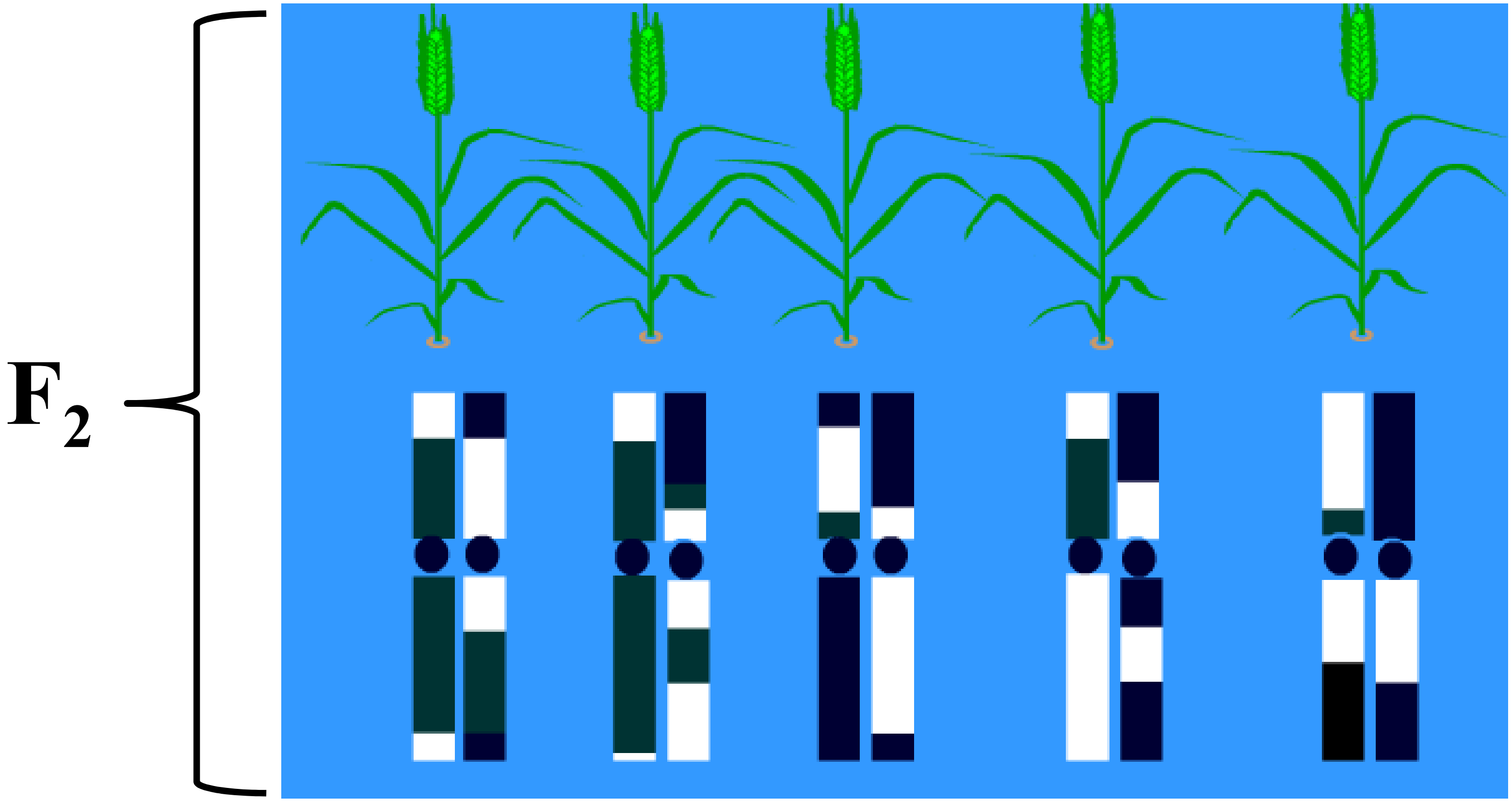
These F1 hybrids will have one chromosome of each pair from the male parent and one from the female parent.

F_2 mapping populations

Self-pollinating an F1 plant produces an array of F2 plants.

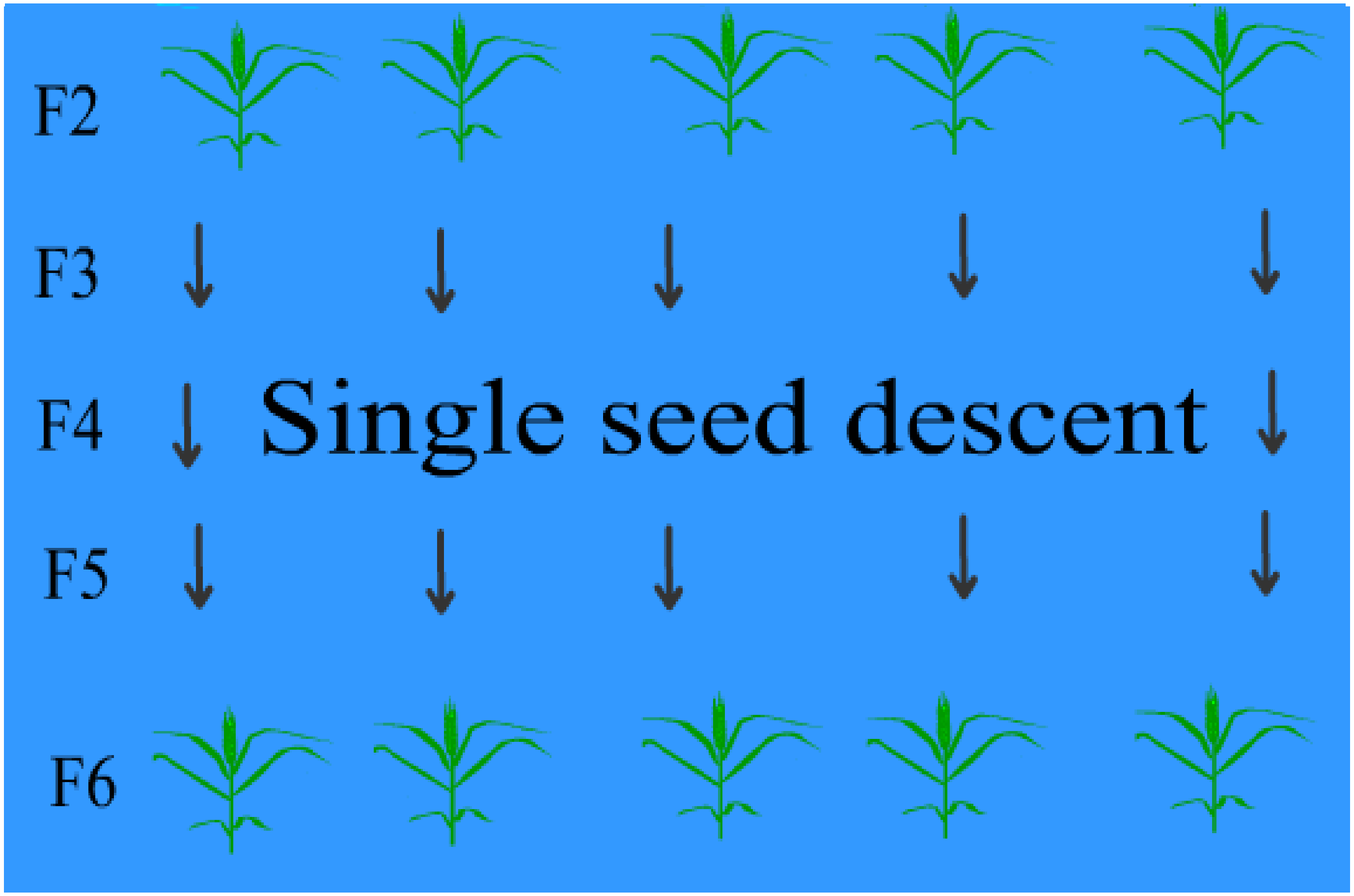


F_2 mapping populations



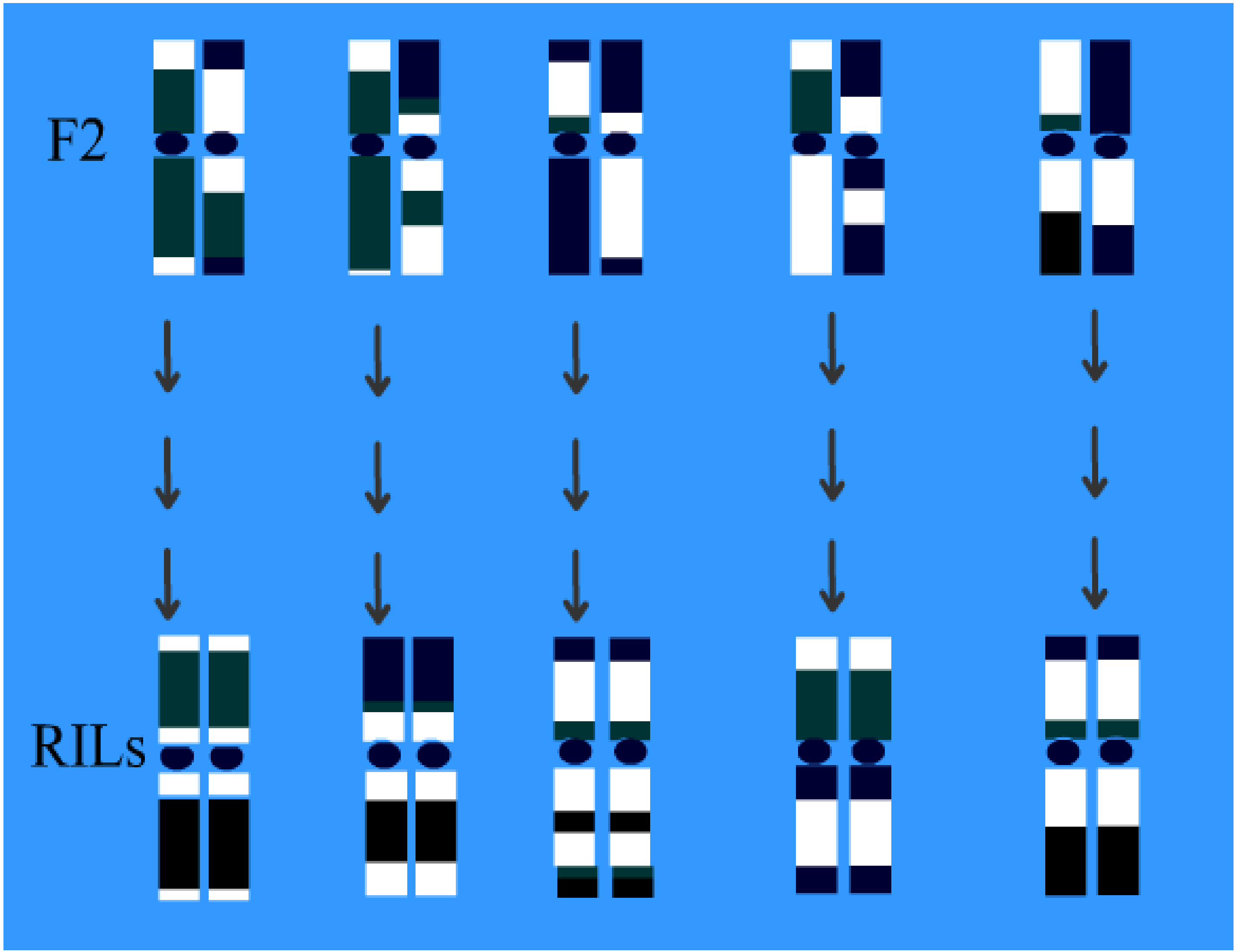
- Unique mosaic of chromosome segments from each parent
- Typical QTL mapping population (F_2 design) 200-300 plants

Recombinant Inbred Lines (RILs)



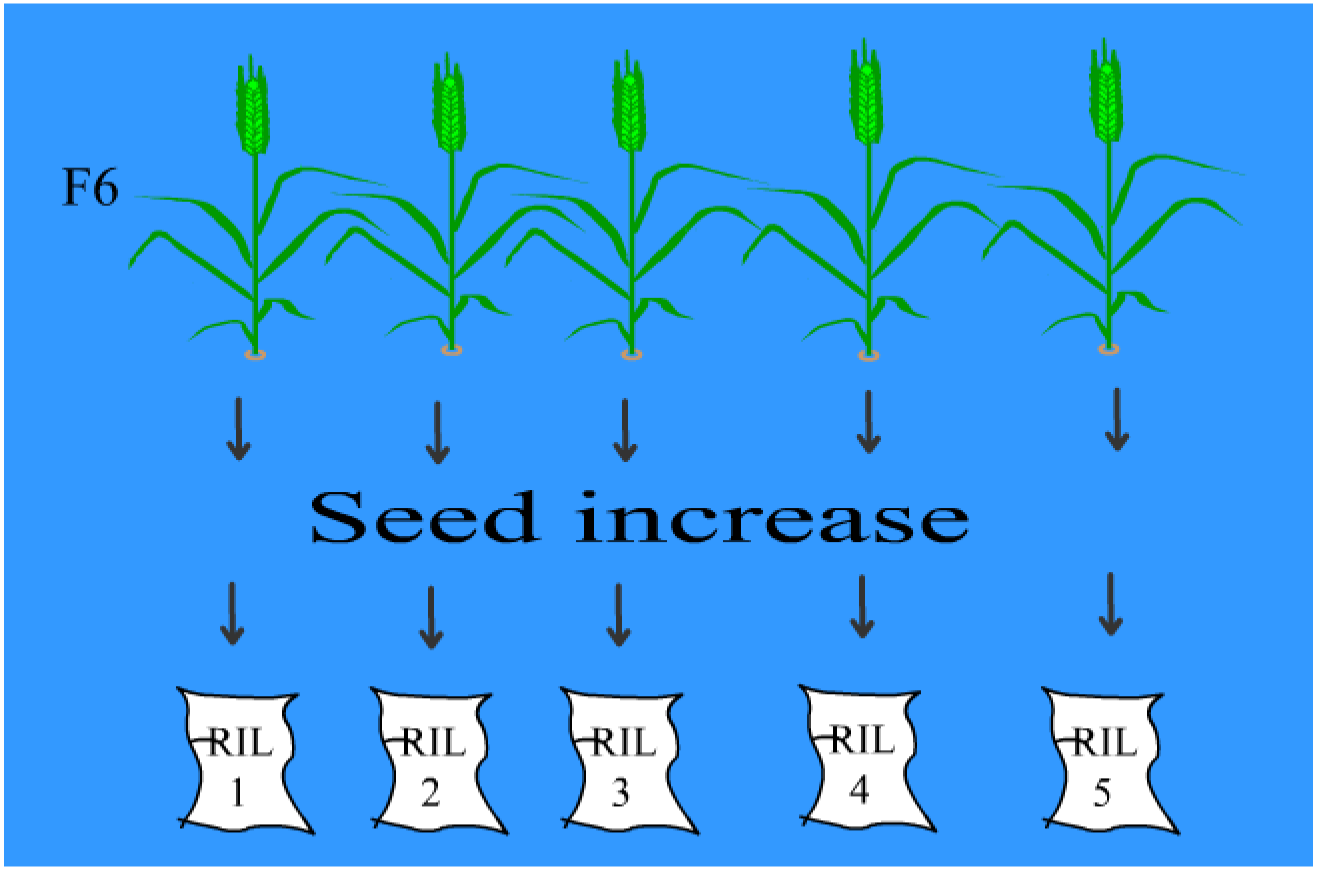
- RILs are developed by several generations of self-pollination from F_2 plants, through a process known as “single seed descent”

Recombinant Inbred Lines (RILs)



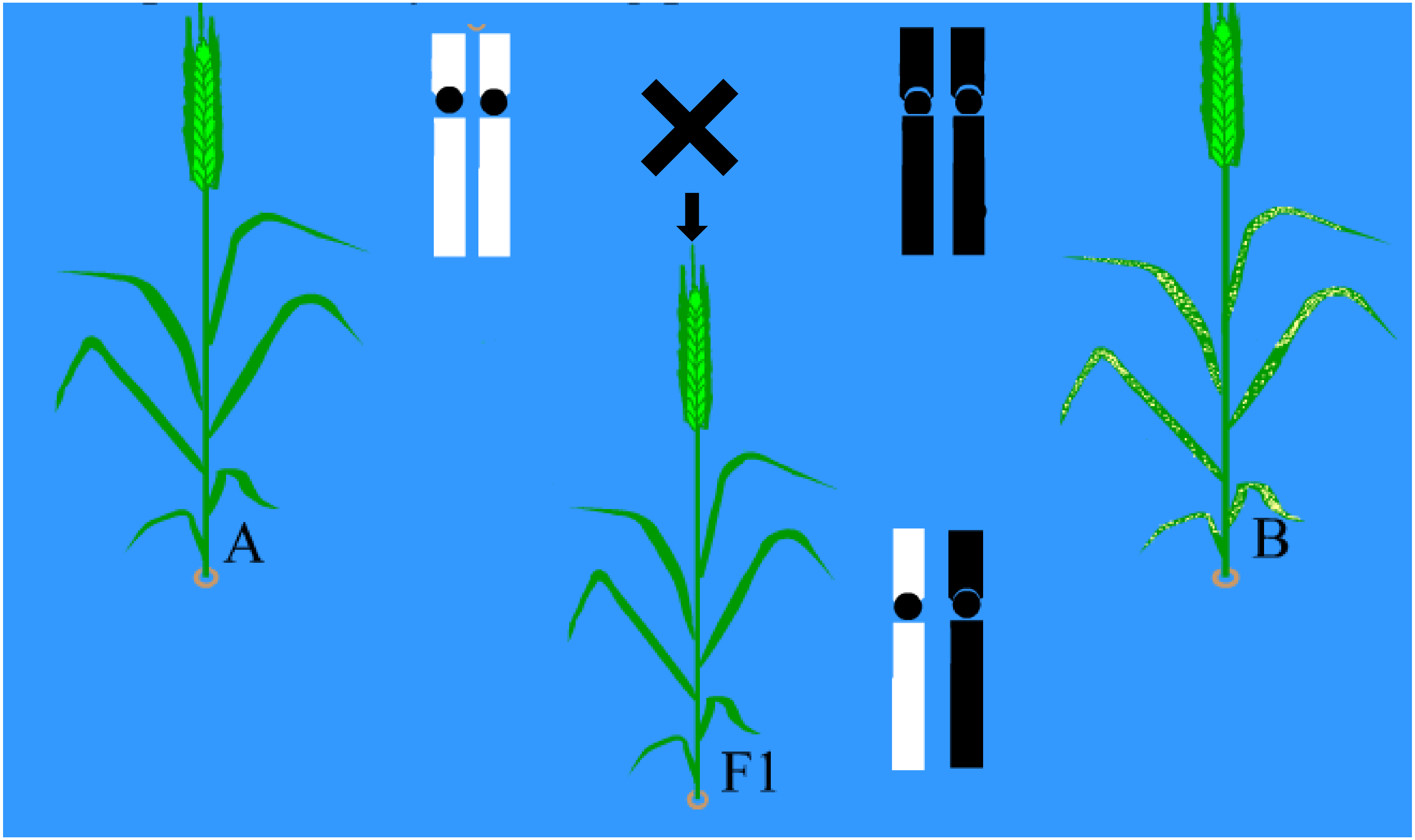
- More generations → more meiosis → more recombination

Recombinant Inbred Lines (RILs)

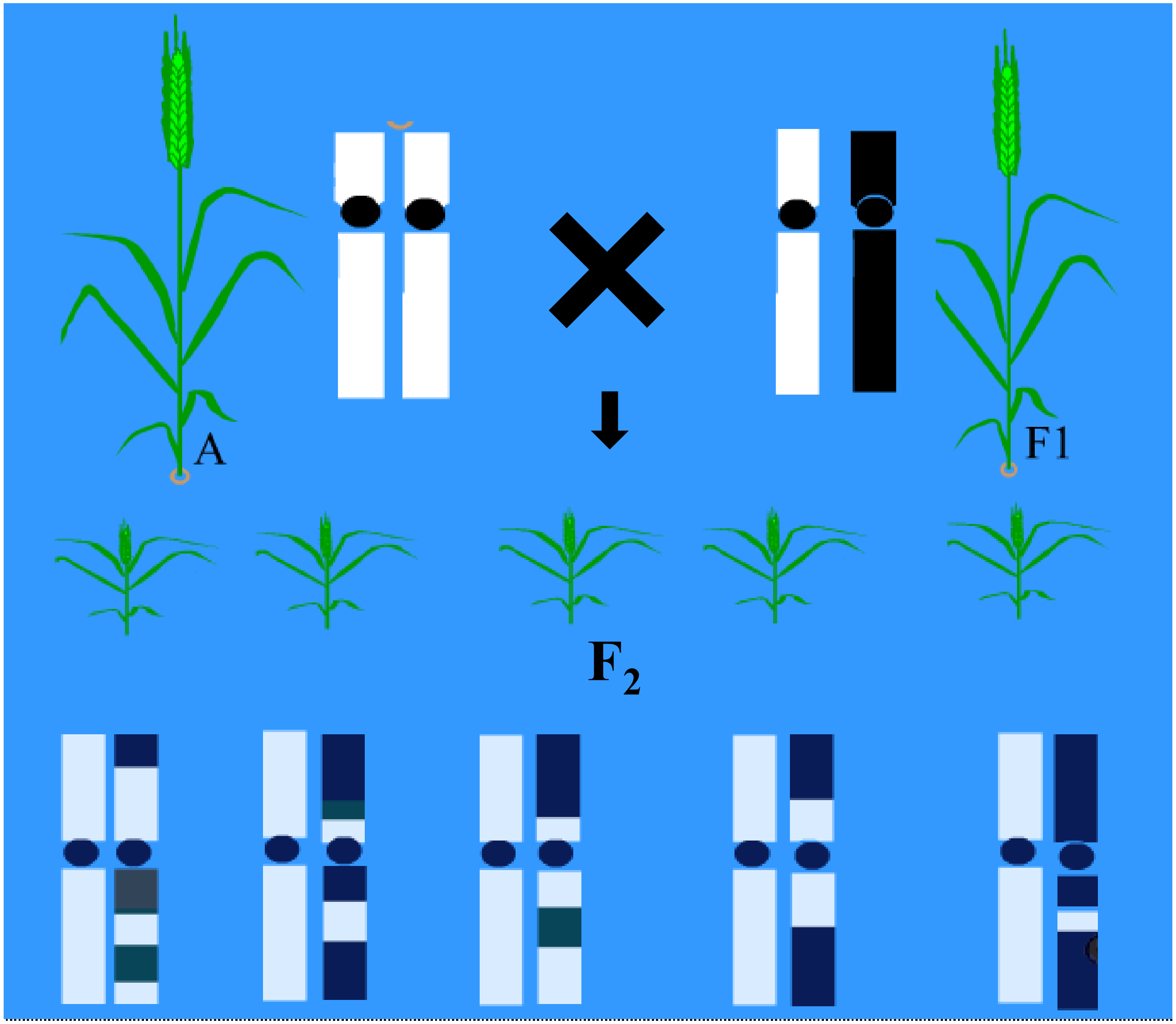


- <u>Important advantages:</u>
  - multiple genetically identical individuals (decreasing the environmental contribution to one trait)
  - multiple traits analysis (from the same line)

Backcross populations



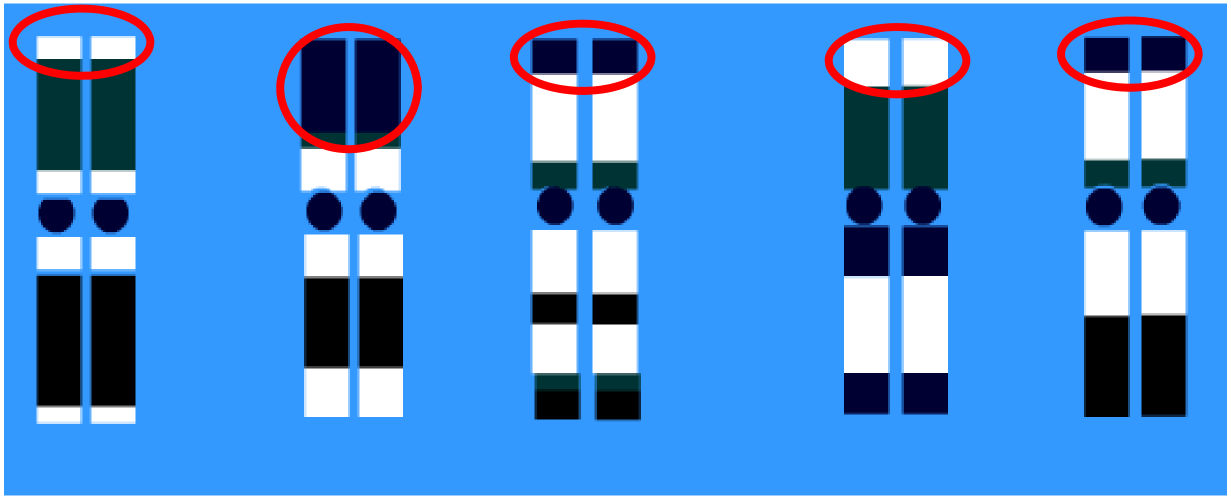
Backcross populations



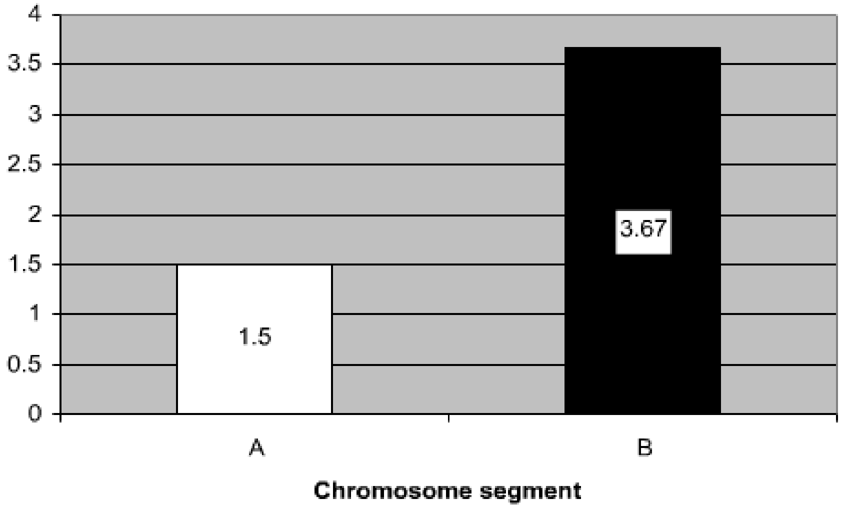
- Result: population of individuals that have one chromosome from parent A and a recombined chromosome with segments of both parents A and B

Example: Calculate the average trait value of individuals having the same segment (either parent A or parent B) in a specific chromosome region

RILs:

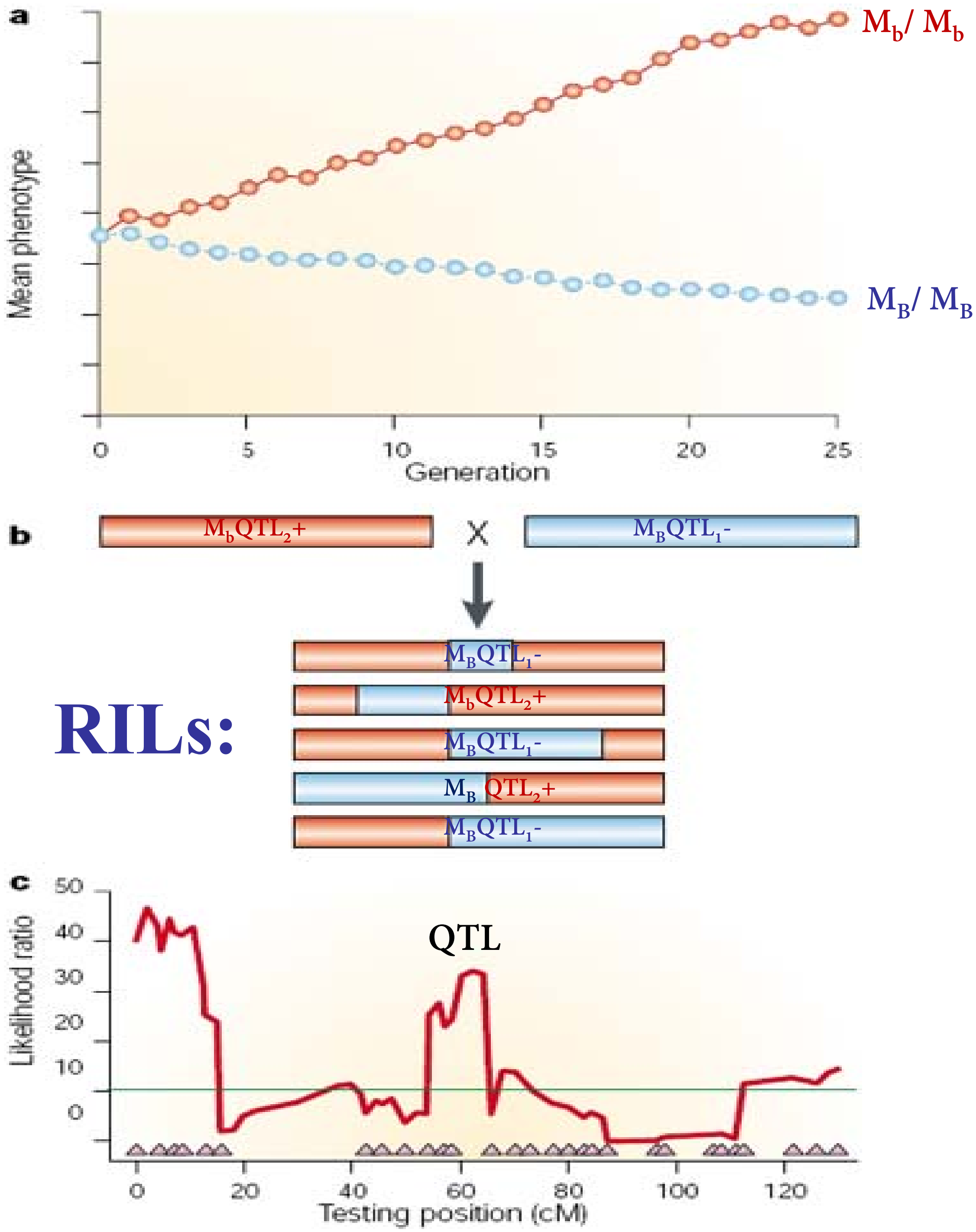


Parental segment:	**A**	**B**	**B**	**A**	**B**
Scale of Resistance (1-4)	**1**	**3**	**4**	**2**	**4**



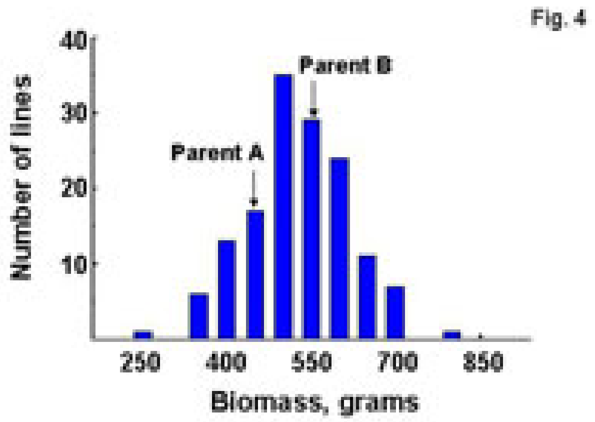
- can be statistically tested if the difference is significant
- can be one gene in this chromosome segment that influences the trait (in this case, resistance)

Quantitative Trait Locus (QTL) Mapping



Phenotypic data evaluation

- Frequency distribution?
  - normal
  - transgressive segregation



- ANOVA
  - are there significant differences for the trait among the plants?

- Heritability analysis
  - the higher the heritability the higher proportion of variation in the trait is due to genetic causes

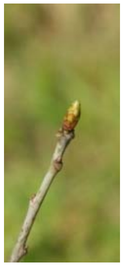
- If two traits are highly correlated, it may indicate that the same QTLs influence both traits (pleiotropy)

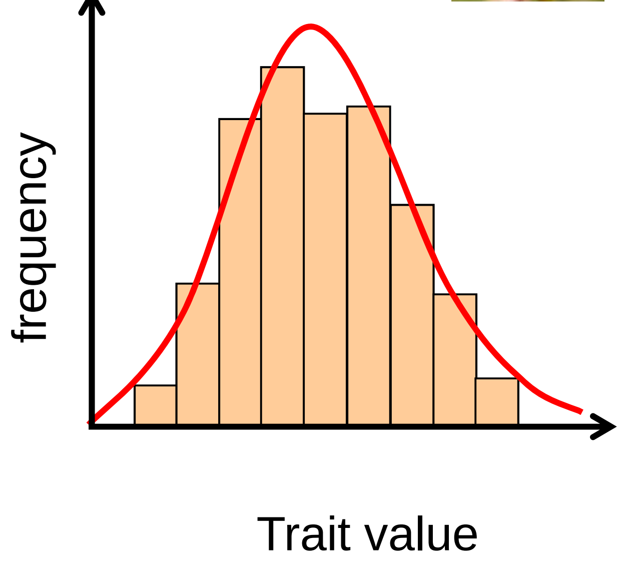
Genetic data analysis

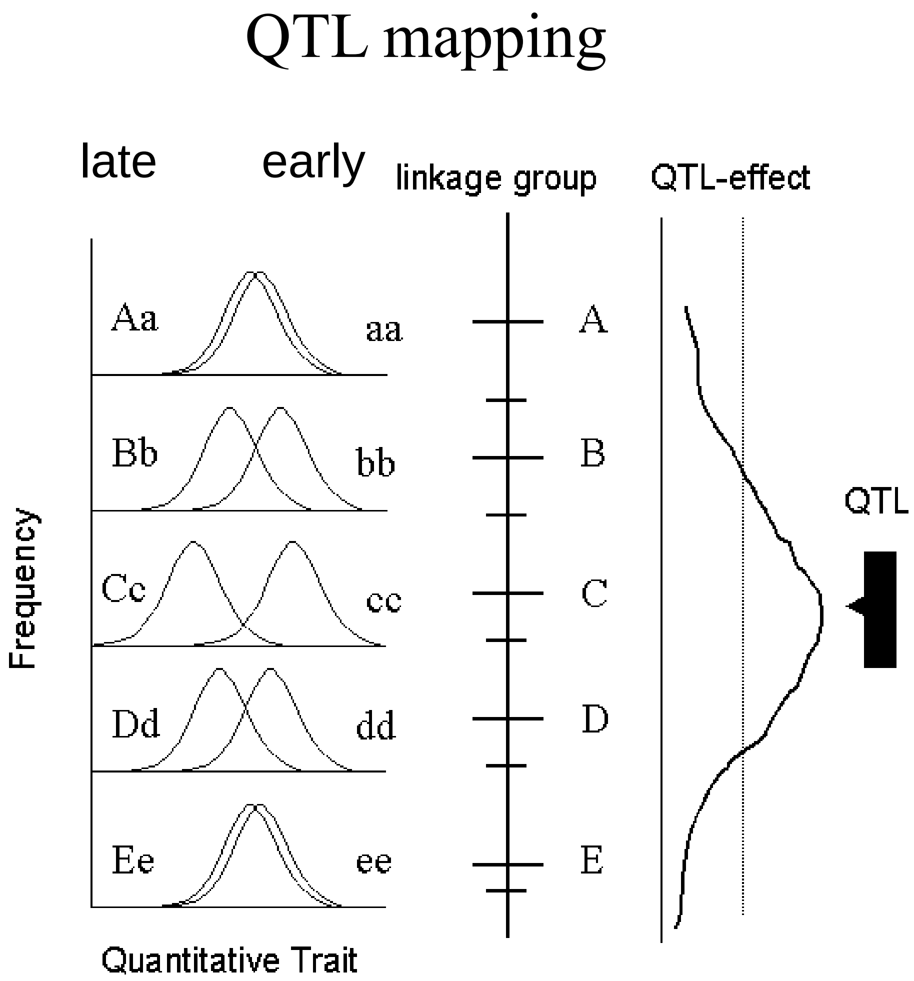
- Parental screening (only markers polymorphic among parents can be useful)
- Genotyping markers in the segregating population
- Check for segregation distortion
- Linkage mapping

QTL mapping in oaks

late early







Adaptive phenotypic quantitative traits

- **Stomatal traits (stomatal density)**

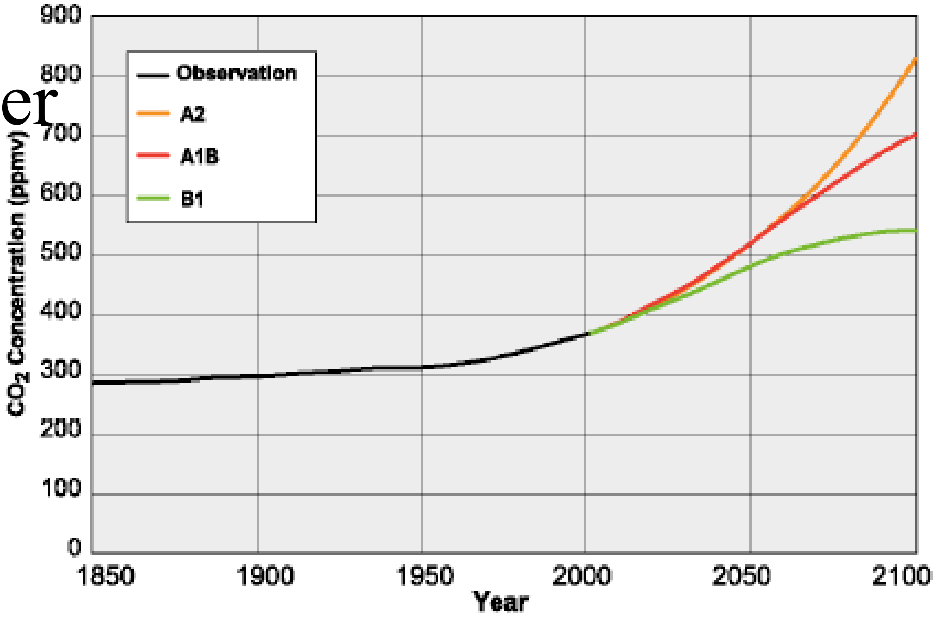
Important to cope with:

– changes in levels of atmospheric CO_2

– changes in availability of water

– higher growth rate

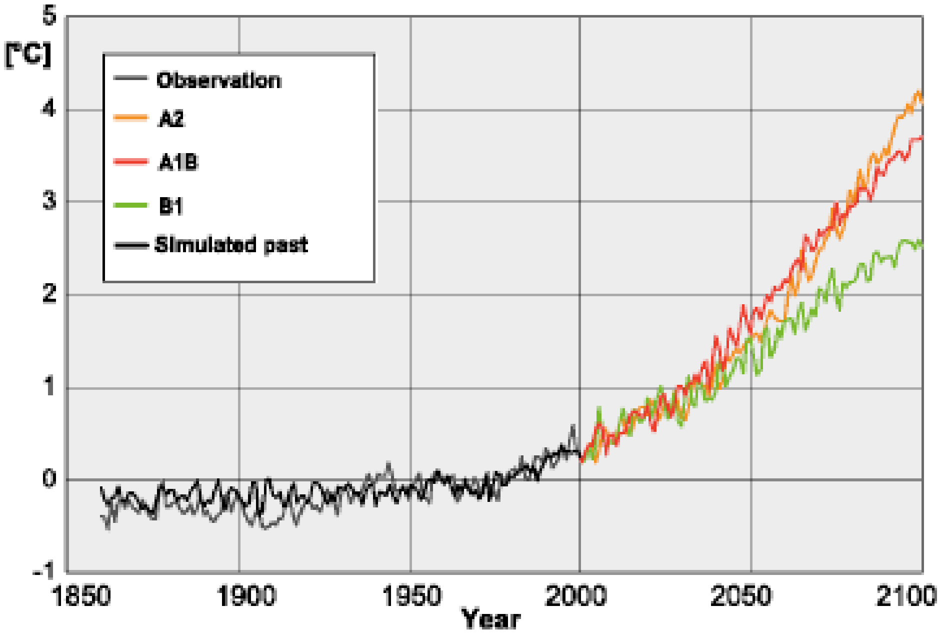
– higher biomass



Roeckner et al. (2006) Climate Projections for the 21st century. Max Planck Institute for Metereology.

Adaptive phenotypic quantitative traits

- **Bud phenology traits:**
  - longer growth season
  - changes in the time of bud burst
- **Frost and drought injuries**
- **Insect damages**

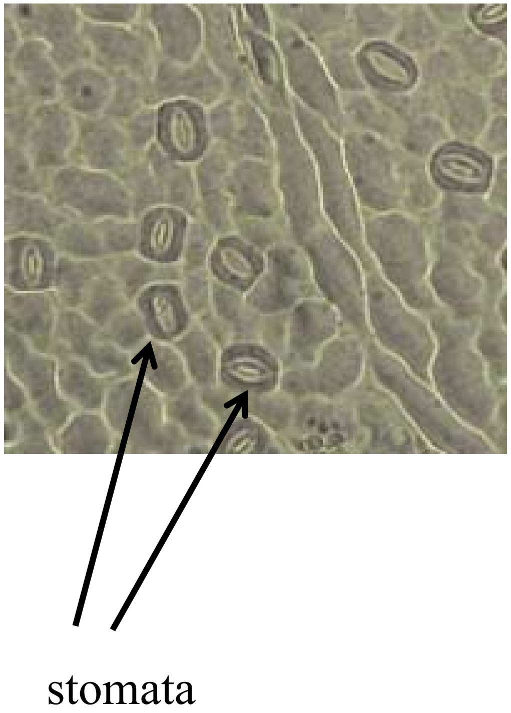


Roeckner et al. (2006) Climate Projections for the 21st century. Max Planck Institute for Metereology.

I. Stomatal density

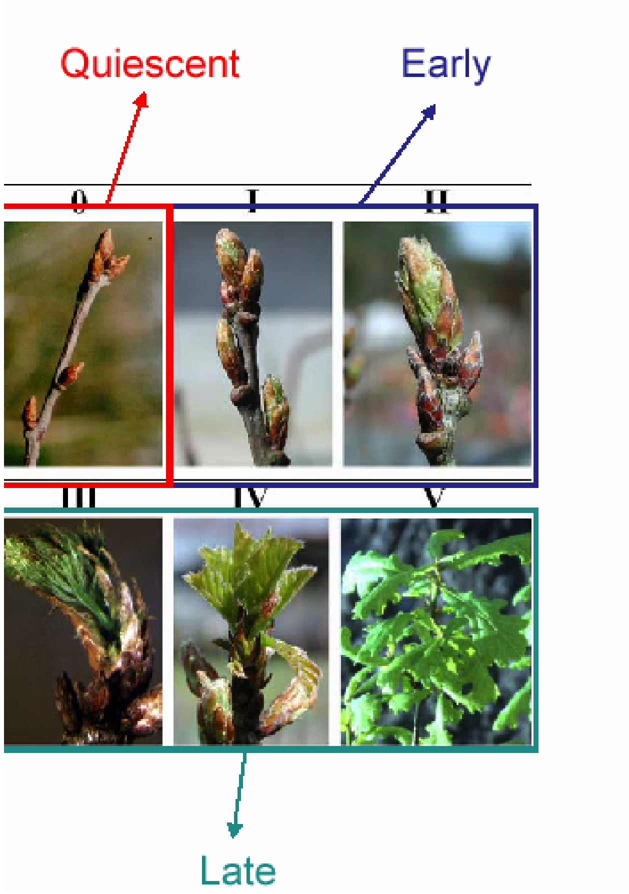
- Aim of the study:

Characterization of the genetic basis of stomatal density and plant growth under non-waterstress conditions



Q. robur x *Q. robur* subsp. *slavonica*, 387 full-sibs

II. Bud burst



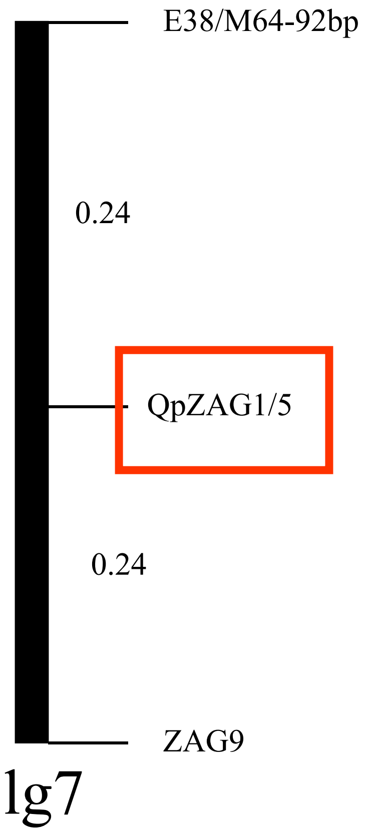
Bud stages 1 to 5
(according to Derory *et al.* 2006)

Q. robur x *Q. robur* subsp. *slavonica*, 387 full-sibs

Comparative QTL mapping

- significant association with bud burst in different *Quercus* spp. progeny (Gailing *et al*. 2005, Scotti-Saintagne *et al*. 2004).

- in six full-sib families the 166 bp allele is associated with an early bud burst

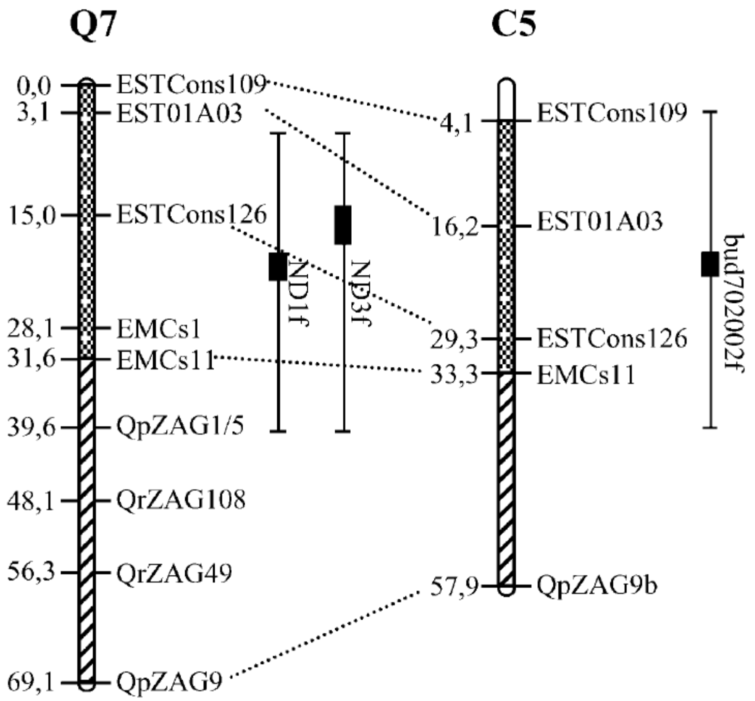


Gailing, O., Kremer, A., Steiner, W., Hattemer, H.H. & R. Finkeldey. 2005. Results on quantitative trait loci for flushing date in oaks can be transferred to different segregating progenies. *Plant Biology* 7: 516-525.

Comparative QTL mapping in Fagaceae

Co-location of QTL in different species

Quercus robur *Castanea sativa*



QTL limitations

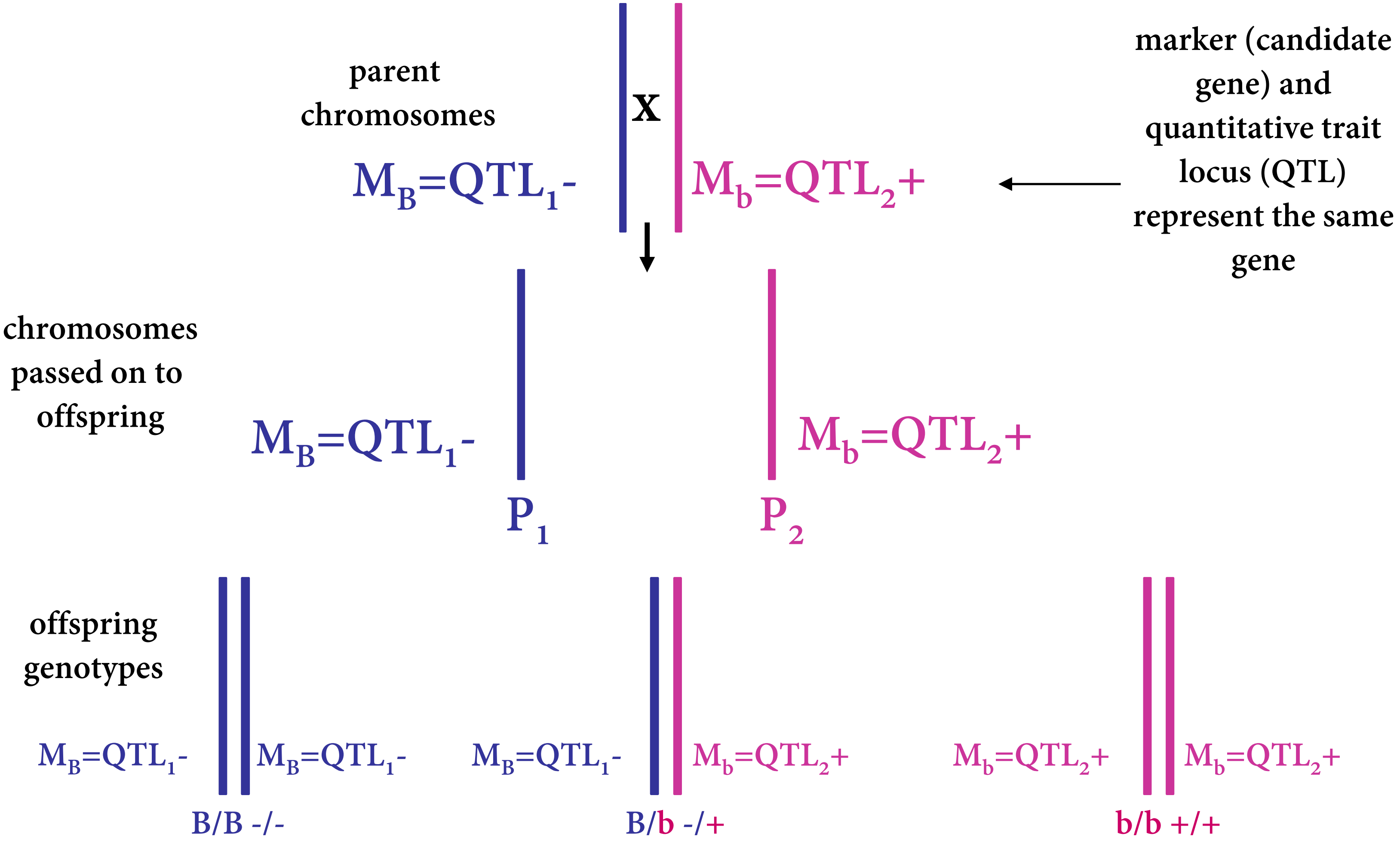
- Information on QTL locations and effects is specific to a particular population and cannot be readily transferred to another population
- QTL analysis detects chromosome regions, not genes, that influence traits. QTL locations have large confidence intervals, often greater than 30 cM
- It is difficult to distinguish two closely linked QTLs, those that are less than 20 cM apart.
- When two QTLs are linked “in repulsion”, it may not be possible to detect the QTL, because the effects of the associated alleles cancel each other out.

Quantitative Trait Locus (QTL) Mapping

- ***Much information has been learned from QTL mapping, but we still mostly do not know what genes and alleles are involved !***

- **Situation is changing now due to**
  - **QTL mapping using candidate genes, and**
  - **genome-wide association studies based on high-density genotyping via sequencing**

Candidate gene based QTL mapping



parent chromosomes

$M_B=QTL_1-$ x $M_b=QTL_2+$

marker (candidate gene) and quantitative trait locus (QTL) represent the same gene

chromosomes passed on to offspring

$M_B=QTL_1-$ P_1

P_2 $M_b=QTL_2+$

offspring genotypes

$M_B=QTL_1-$ $M_B=QTL_1-$

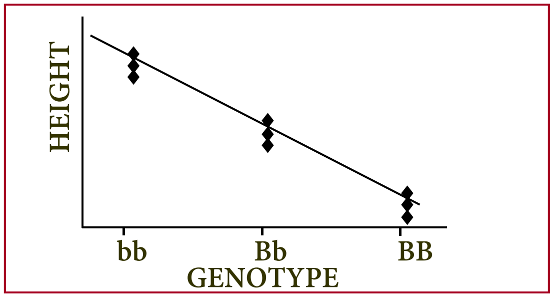
B/B -/-

$M_B=QTL_1-$ $M_b=QTL_2+$

B/b -/+

$M_b=QTL_2+$ $M_b=QTL_2+$

b/b +/+



Criteria to select candidate genes

1) Physiology and biochemistry, high similarity/homology to genes with well known effects, mutations

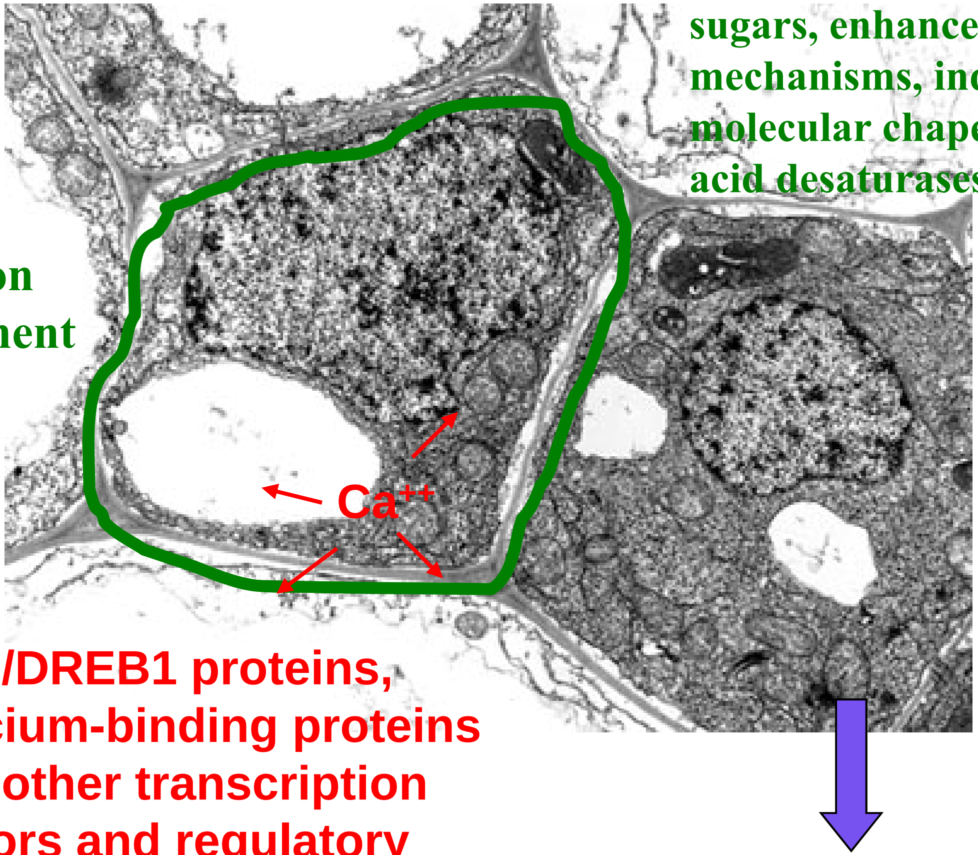
2) Differential expression

3) Positional association with quantitative trait loci (QTLs)

1) Physiological mechanisms involved in low temperature and drought tolerance

Stabilization of membranes and membrane fluidity via changes in lipid composition, accumulation of sucrose and other simple sugars, enhancement of antioxidative mechanisms, induction of genes encoding molecular chaperones, dehydrins, fatty acid desaturases etc.

Cytoskeleton rearrangement



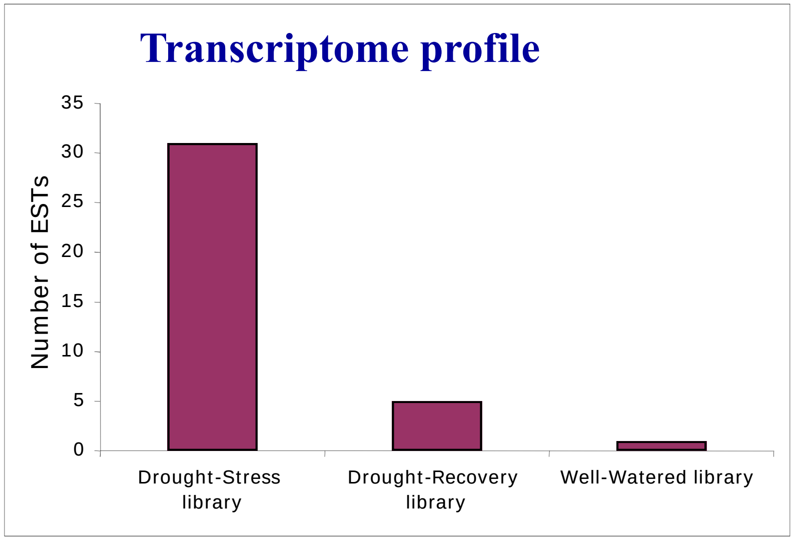
CBF/DREB1 proteins, Calcium-binding proteins and other transcription factors and regulatory proteins in a low temperature signal transduction pathway

Increase of antifreeze proteins and level of cryoprotectants, such as proline, for instance, that prevent ice formation

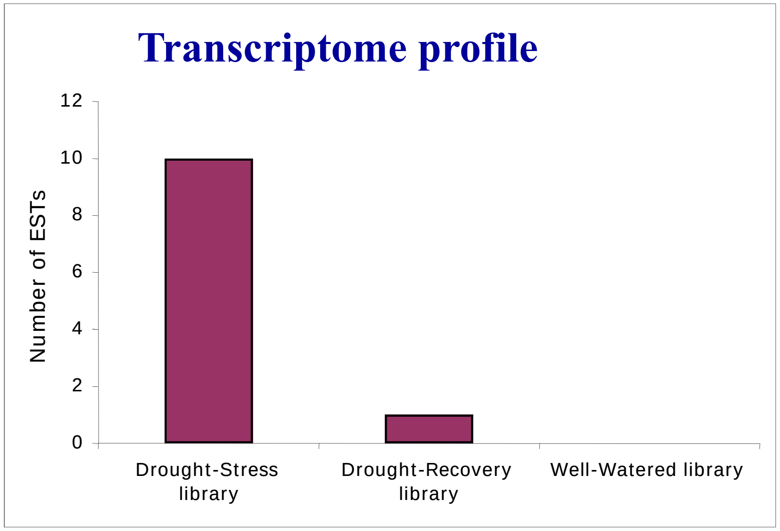
2) Differential gene expression *in silico*: transcriptome profiling using RNA-seq and NGS

Candidate genes for drought resistance in loblolly pine:

ABA and WDS induced gene (*pLP3-3*)



Dehydrin (*Dhn-1*)



Lorenz W.W., Sun F., Liang C., Kolychev D., Wang H., Zhao X., Cordonnier-Pratt M.-M., Pratt L.H., Dean J.F.D. 2006. Water stress-responsive genes in loblolly pine (*Pinus taeda*) roots identified by analyses of expressed sequence tag libraries. *Tree Physiology* **26**: 1–16.

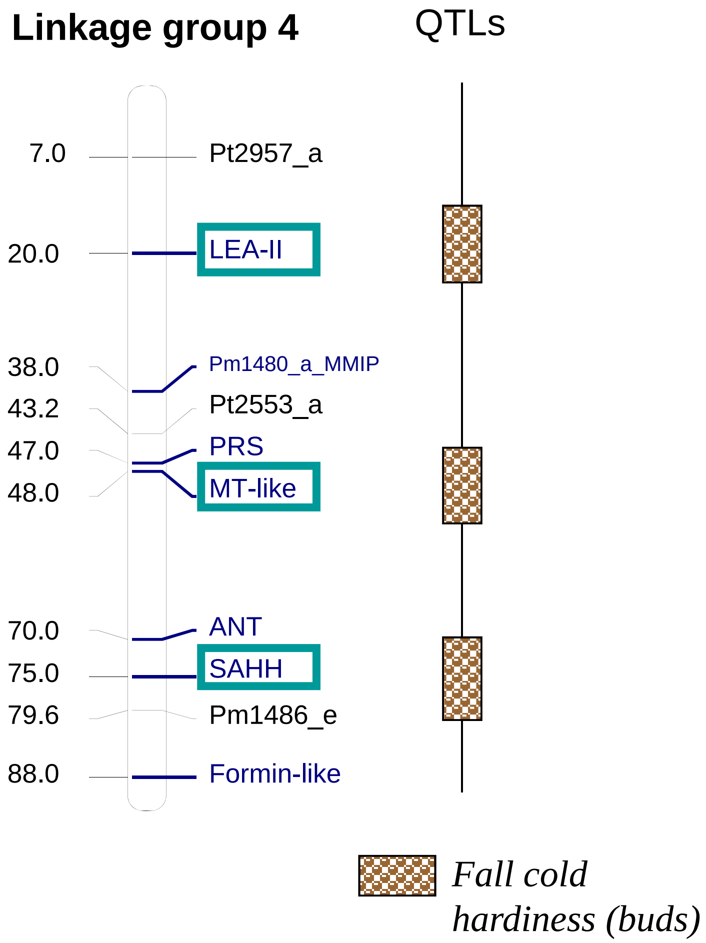
3) Collocation of adaptive trait related genes with QTLs controlling adaptive traits

Positional candidate genes:

LEA-II
(late embryogenesis abundant type II)
dehydrin-like protein
Induced by cold

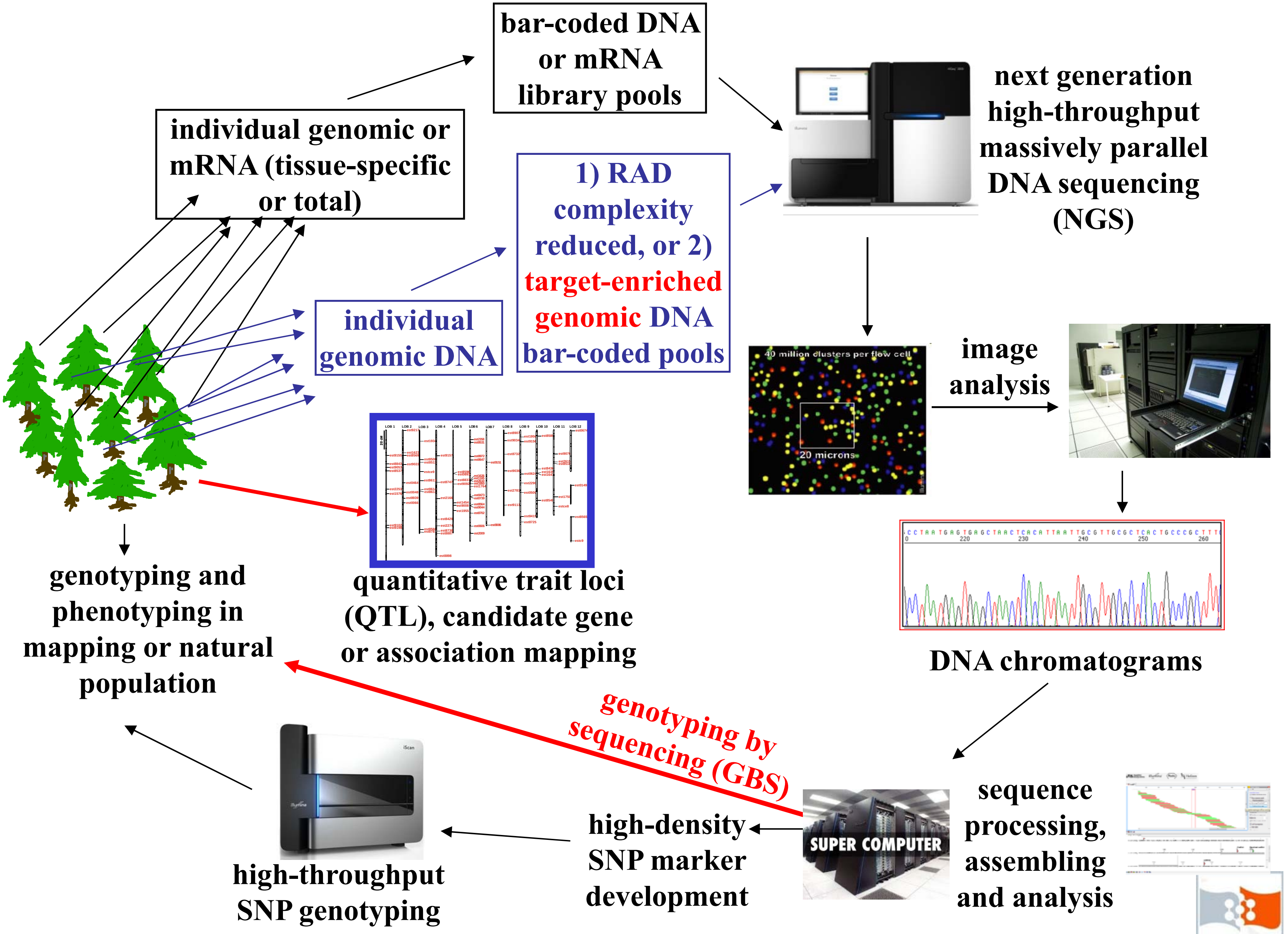
MT-like
(metallothionein-like protein)
stress-induced;
downregulated under water deficit

SAHH
(S-adenosyl-L-homocysteinas hydrolase)
upregulated under water deficit



Wheeler N.C., Jermstad K.D., Krutovsky K.V. *et al.* 2005. Mapping of quantitative trait loci controlling adaptive traits in coastal Douglas-Fir. IV. Cold-hardiness QTL verification and candidate gene mapping. *Molecular Breeding* 15: 145-156.

Genomic markers development and genotyping using next generation sequencing



bar-coded DNA or mRNA library pools

individual genomic or mRNA (tissue-specific or total)

1) RAD complexity reduced, or 2) target-enriched genomic DNA bar-coded pools

individual genomic DNA

next generation high-throughput massively parallel DNA sequencing (NGS)

image analysis

DNA chromatograms

sequence processing, assembling and analysis

high-density SNP marker development

high-throughput SNP genotyping

genotyping by sequencing (GBS)

genotyping and phenotyping in mapping or natural population

quantitative trait loci (QTL), candidate gene or association mapping

Association mapping

- AKA: Linkage disequilibrium mapping

- Aim: (like in QTL) to find a statistical association between a genetic marker and a quantitative trait

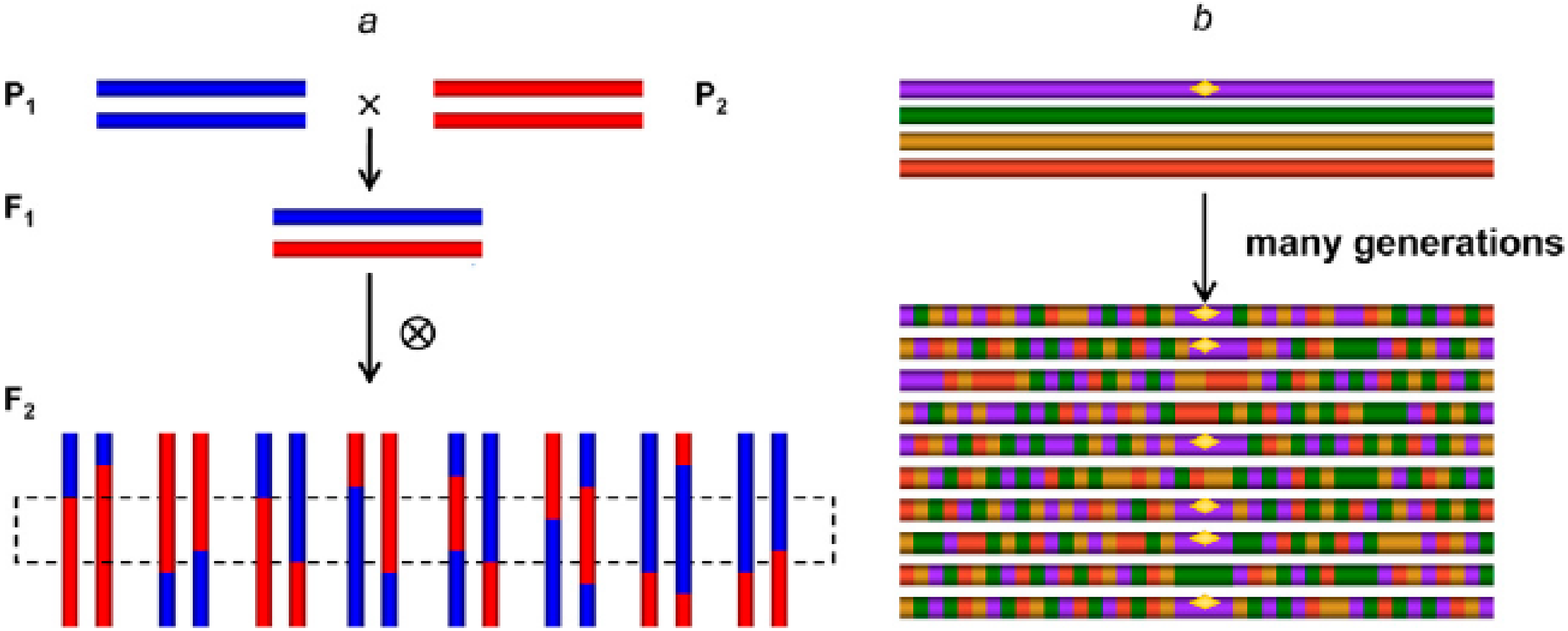
BUT

- Association mapping is performed at the population level (unrelated or distantly-related individuals sampled from a population)

- In association mapping, the genetic markers usually must lie close to or within genes responsible for a measured trait

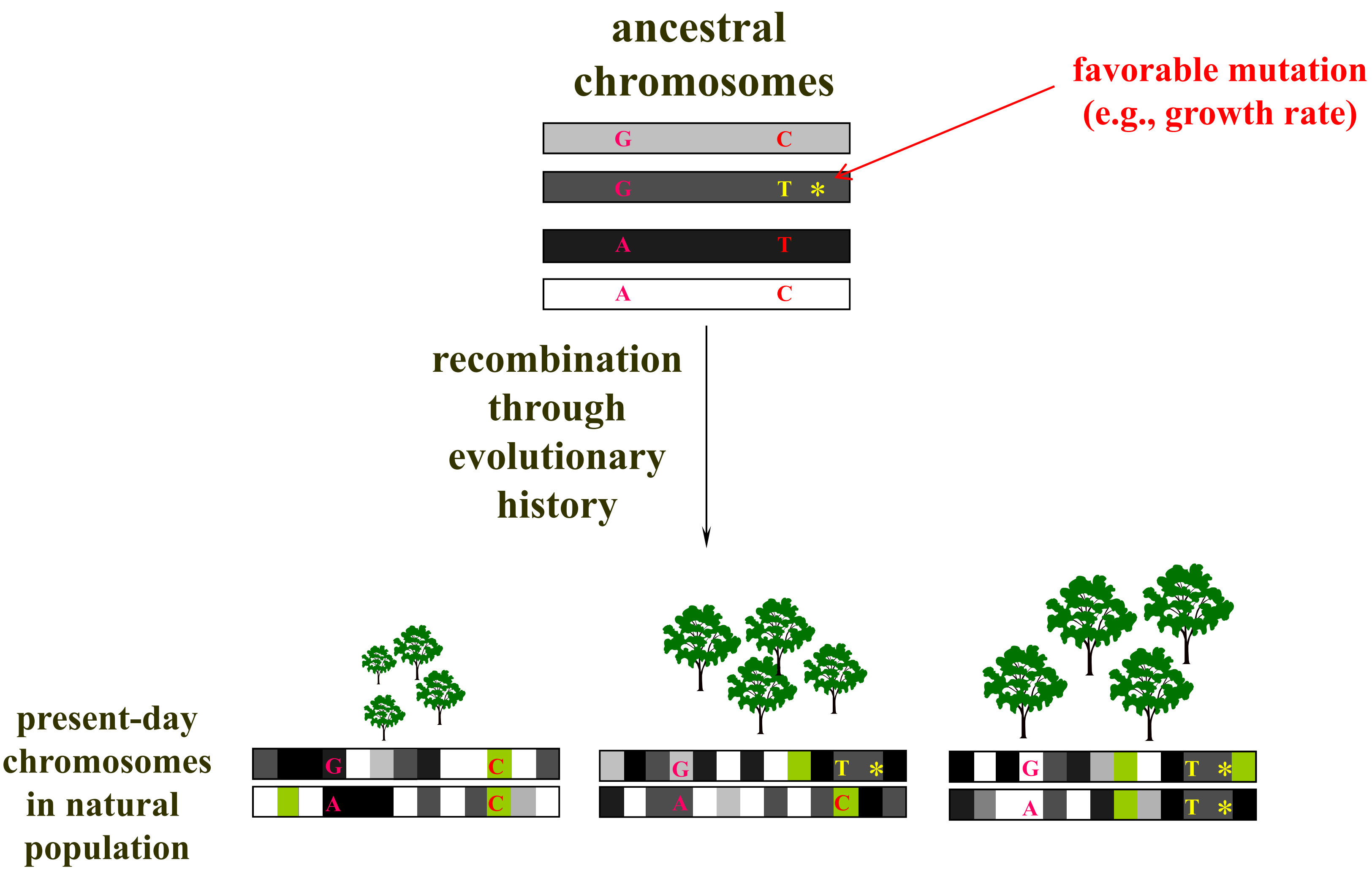
- The goal is to identify the actual genes affecting that trait, rather than just (relatively large) chromosomal segments

QTL mapping (*a*) vs. Associations Mapping (*b*)



Zhu et al. 2008

Association Mapping using SNPs



ancestral chromosomes

favorable mutation (e.g., growth rate)

recombination through evolutionary history

present-day chromosomes in natural population

Linkage disequilibrium (LD)

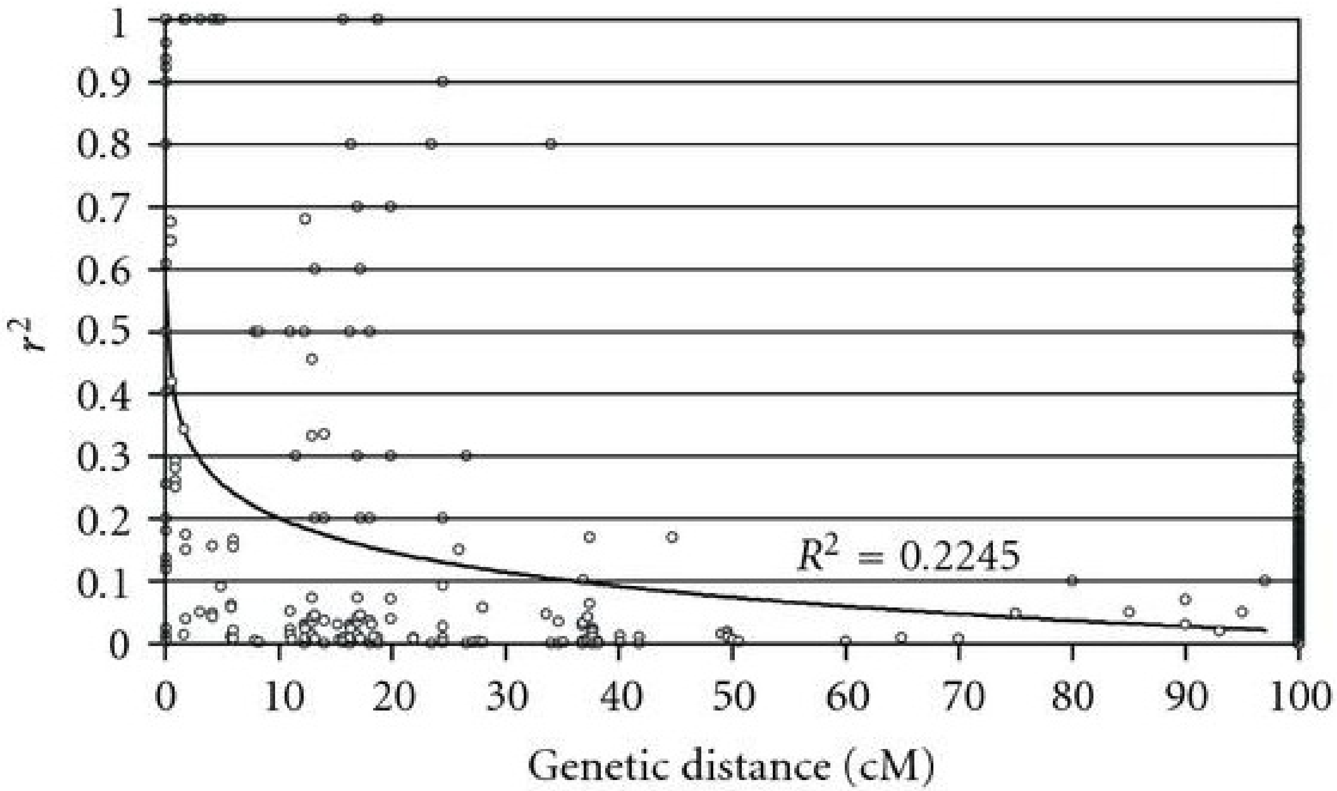
- LD: The non-independence of alleles at different loci (certain combinations of alleles across loci occur more often than expected by chance alone)

- LD can be caused by various combinations of many different factors, including selection, new mutations, population genetic structure, interbreeding of genetically divergent populations (admixture) and population bottlenecks

- LD is reduced by recombination, which "shuffles" the combinations of alleles at different loci

Linkage disequilibrium (LD)

- LD decay very fast in most sexually reproducing species with large effective population size:



Linkage disequilibrium (**LD**) decay plot depicted from the pairwise **LD** (***r***^{2}) and genetic distance (**cM**) values measured between all pairs of linked markers. A pairwise **LD** values (measured as ***r***^{2} parameters) are plotted against a pairwise genetic distances (**cM**). Inner fitted trend line is a nonlinear logarithmic regression curve of on genetic distance (***R***^{2}).

Genome-wide random marker based association mapping
vs.
Candidate gene based association mapping

- **Association mapping based on random genetic markers** relies on linkage disequilibrium (LD) between gene markers and the actual causative polymorphism in genes that causes the differences in the phenotypic trait.

- **Association mapping based on candidate genes** relies on direct association of the actual causative polymorphism in the candidate gene with the differences in the phenotypic trait.

Use of candidate genes (functional markers) increases chances that association is due to causative gene

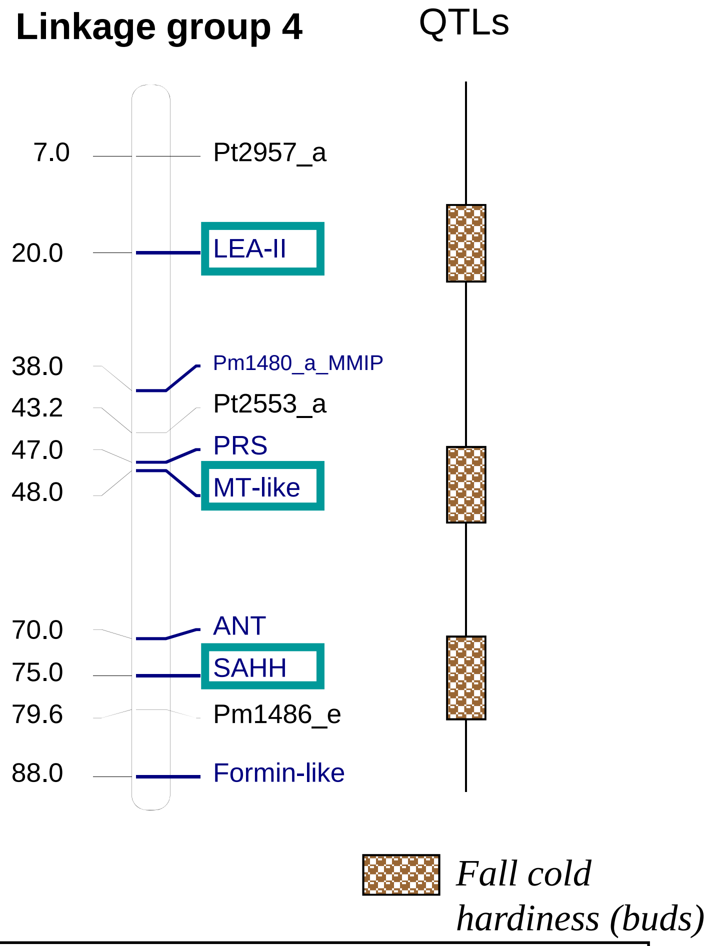
Collocation of adaptive trait related candidate genes with QTLs controlling adaptive traits:

Functional candidate genes:

LEA-II
(late embryogenesis abundant type II)
dehydrin-like protein
Induced by cold

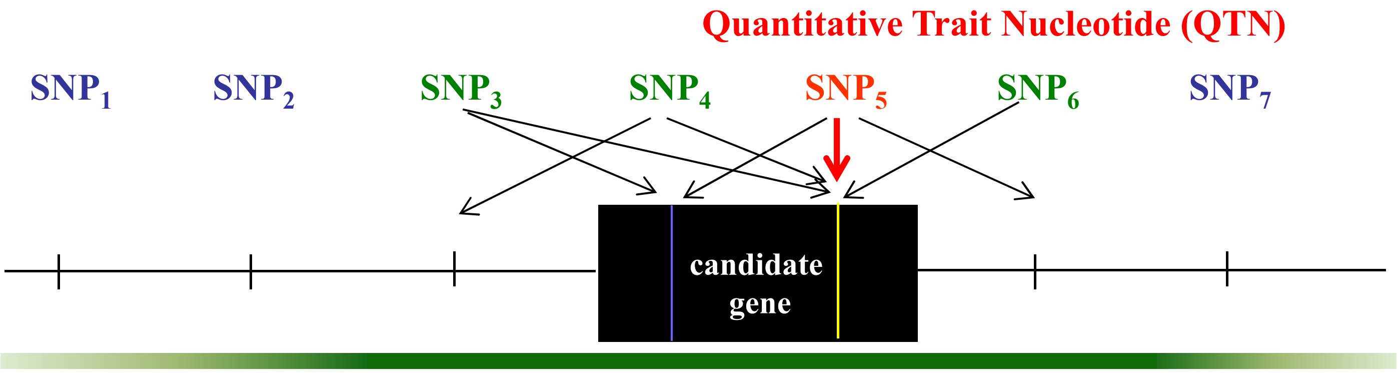
MT-like
(metallothionein-like protein)
stress-induced;
downregulated under water deficit

SAHH
(S-adenosyl-L-homocysteinas hydrolase)
upregulated under water deficit



Wheeler N.C., Jermstad K.D., Krutovsky K.V. *et al.* 2005. Mapping of quantitative trait loci controlling adaptive traits in coastal Douglas-Fir. IV. Cold-hardiness QTL verification and candidate gene mapping. *Molecular Breeding* 15: 145-156.

Genome-wide random marker based association mapping vs. Candidate gene based association mapping



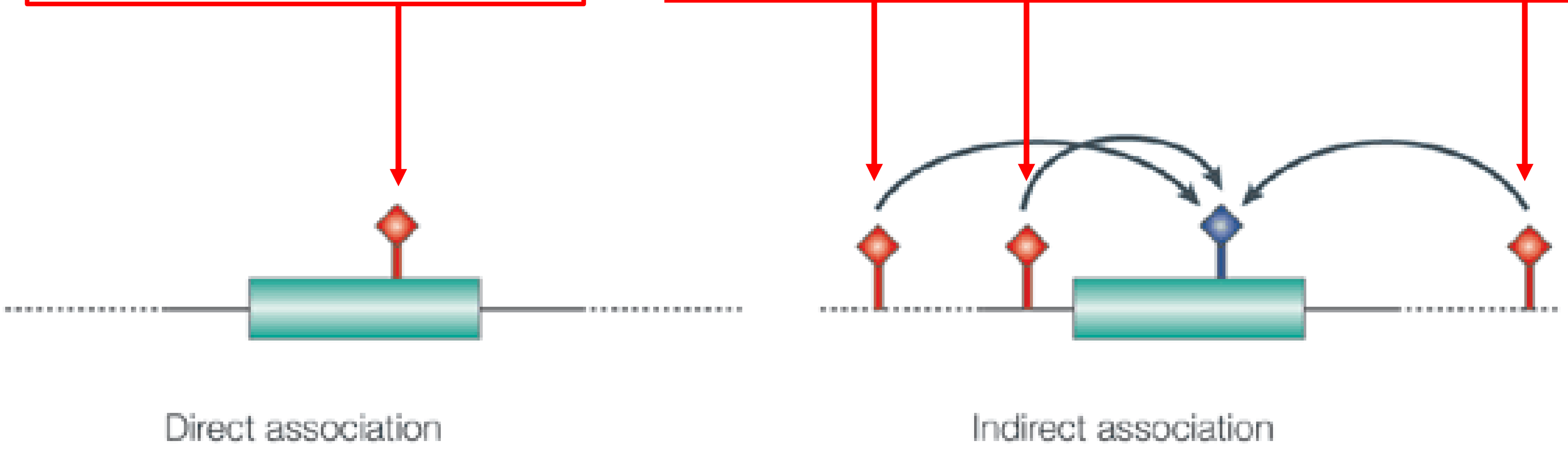
Extent of linkage disequilibrium

Linkage Disequilibrium (LD) is a nonrandom association of alleles at linked loci

SNP Association Testing: Direct or Indirect

A candidate, functional SNP is directly tested for association with disease

SNPs to be genotyped are chosen randomly or on the basis of linkage disequilibrium patterns to provide as much information about other SNPs as possible



Nature Reviews | Genetics

2005; 6:95-108

Genomic approaches to study complex adaptive traits

- High-density genome-wide genotyping via high-throughput NGS sequencing or high-density single nucleotide polymorphism (SNP) assays
- Genome-wide Association Studies using SNPs
- Functional genomic markers (candidate genes)
- Detection genes-outliers:
  - with unusually high or low differentiation
  - with unusually high or low expression

Krutovsky K.V. & D.B. Neale. 2005. Forest genomics and new molecular genetic approaches to measuring and conserving adaptive genetic diversity in forest trees, pp. 369-390 in *Conservation and Management of Forest Genetic Resources in Europe*, edited by Th. Geburek and J. Turok. Arbora Publishers, Zvolen.

González-Martínez S.C., Krutovsky K.V., Neale D.B. 2006. Forest tree population genomics and adaptive evolution. *New Phytologist* **170**(2): 227-238.

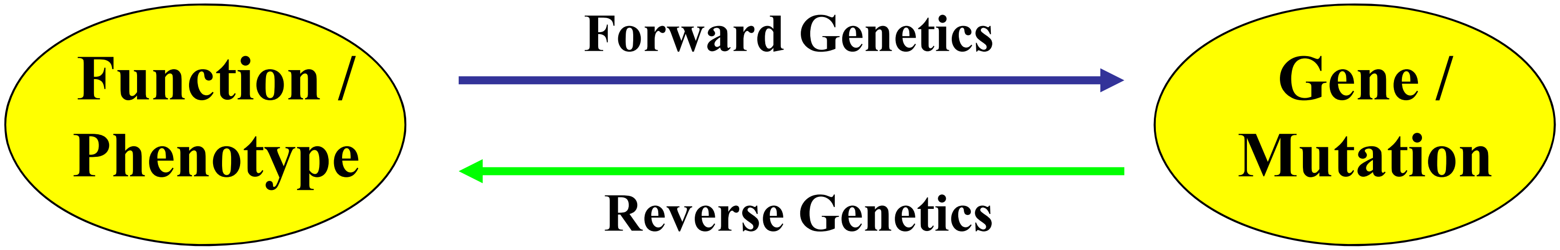
Genomic approaches to study complex adaptive traits

- Forward Genetics (from trait to genotype) vs. Reverse Genetics (from genotype to trait)

- Top down (pattern to process) and bottom up (process to pattern) approaches

- Neutral vs. functional markers

Forward vs. Reverse Genetics



Function / Phenotype

Forward Genetics

Reverse Genetics

Gene / Mutation

Forward Genetics (traditional): Starts with a phenotype and moves towards the gene

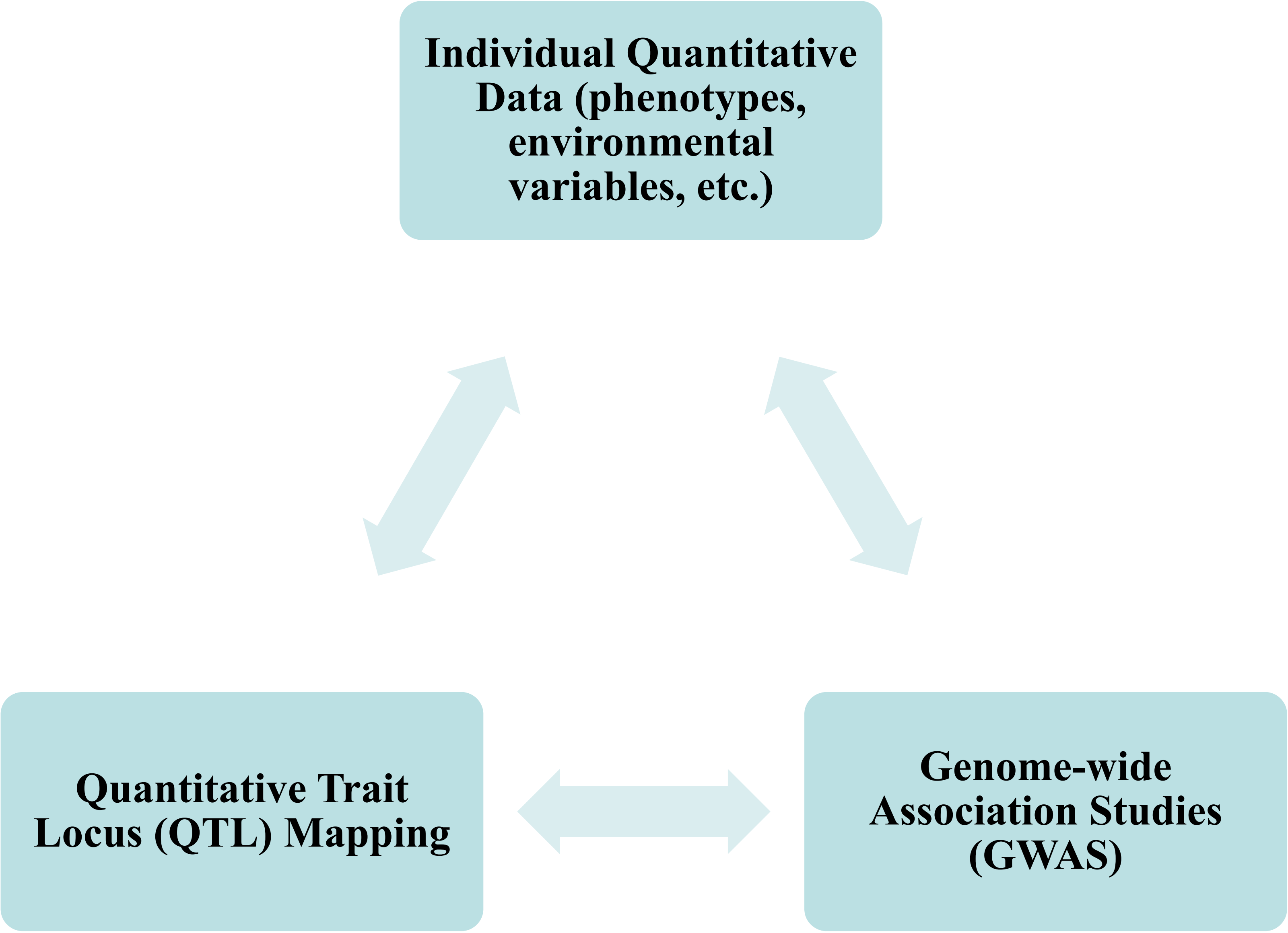
Reverse Genetics: Starts with a particular gene and assays the effect of its disruption or allelic effects

Genomics (via Association Genetics): $P \rightleftarrows G$

What is the genetic basis of a particular phenotype?

(How to determine the function of a gene, or the identity of genes responsible for a trait?)

Association Genetics



Individual Quantitative Data (phenotypes, environmental variables, etc.)

Quantitative Trait Locus (QTL) Mapping

Genome-wide Association Studies (GWAS)

Integrated approach to study evolutionary response

1) association of population substructure with particular environments and ecozones;
2) identification of outliers - genes with unusually high differentiation (that could be a signature of positive or diversifying selection), or genes with unusually low differentiation (that could be a signature of balancing selection) that significantly above or below levels that are expected for selectively neutral markers;
3) correlation of allele frequencies with environmental gradients; clinal variation, etc.;
4) QTL and/or association mapping of candidate genes with adaptive trait related phenotypes;
5) identification of genome-wide signatures of selection (LD, selective sweeps, etc.);
6) intra- and inter-specific selective neutrality tests

- **XX century:**

Evolutionary theory + Genetics
= Synthetic theory of evolution
(Genetic theory of evolution or
Evolutionary Genetics)

↓ **paradigm shift**

Theodosius Dobzhansky
(1900-1975)

from an individual as a main unit of evolution to population genetics level of thinking

---

- **XXI century:**

Molecular genetics + Bioinformatics
= Genomics

↓ **paradigm shift**

from genetics to genomics level of thinking

Krutovsky, K.V. 2006 From Population Genetics to Population Genomics of Forest Trees: Integrated Population Genomics Approach. *Russ. J. of Genetics* **42**(10): 1088–1100

Key terms and definitions

- **Quantitative Trait Locus (QTL) mapping**: A QTL is a chromosomal region suspected to contain a gene (or cluster of genes) that contributes to the variation observed at a quantitative trait. QTLs are detected through linkage mapping experiments using progeny usually obtained in experimental crosses or pedigrees that segregate for both quantitative traits and genetic markers. QTL and genetic markers that are close together on a chromosome will tend to co-segregate.
- **Association mapping**: As in QTL mapping, the goal of association mapping is to find a statistical association between genetic markers and a quantitative trait. However, unlike QTL mapping, which is performed in the context of a pedigree, association mapping is performed at the population level: the genotypes of the candidate gene markers and the phenotypes of the corresponding trait are determined in a set of unrelated or distantly-related individuals sampled from a population. Association mapping relies on linkage disequilibrium (LD) between the markers and the actual causative genes (i.e., the actual polymorphism that causes the differences in the phenotypic trait). Hence association mapping is also referred to as 'LD mapping'. For association to be detected the genetic markers usually must be closely linked to genes (lie within or directly upstream or downstream of them) that contribute to the variation in that trait, and the goal is to identify the actual genes affecting that trait, rather than just (relatively large) chromosomal segments. Since population genetic structure (genetic differences that accumulate between populations) can cause LD even at unlinked loci, association analyses must account for population genetic structure whenever it is present in the population from which your sample has been drawn (Pritchard et al. 2000; Thornsberry et al. 2001).

Key terms and definitions

- **Amino-acid or nucleotide multiple sequence alignment**: is a way of arranging the sequences of DNA, RNA, or protein to identify regions of similarity that may be a consequence of functional, structural, or evolutionary relationships between the sequences. Aligned sequences of nucleotide or amino acid residues are typically represented as rows within a matrix.

- **Contig**: (from *contiguous*) is a set of overlapping DNA sequences in a **multiple sequence alignment** that together represent a **consensus** region of DNA. In sequencing projects, a contig refers to overlapping sequence reads or to the overlapping clones that form a physical map of the genome that is used to guide sequencing and assembly. Contigs can thus refer both to overlapping DNA sequence and to overlapping physical segments (fragments) contained in clones depending on the context.

- **Consensus sequence**: is the calculated order of most frequent residues, either nucleotide or amino acid, found at each position in a **multiple sequence alignment**. It represents the results of a **multiple sequence alignment,** in which related sequences are compared to each other, and most frequent residues are calculated.

- **Basic Local Alignment Search Tool (BLAST)**: is an algorithm for comparing amino-acid or nucleotide sequences using their alignment. A BLAST search enables a researcher to compare a query sequence with a library or database of sequences, and identify library sequences that resemble the query sequence above a certain threshold.

- **Expressed Sequence Tag (EST):** is a short sub-sequence of a mRNA/cDNA sequence. They are used to identify gene transcripts, and are instrumental in gene discovery, gene sequence determination and differential gene expression analysis (transriptome profiling).

- **Unigene:** is a supposedly unique transcript that represents the same transcription locus (expressed gene or pseudogene), often inferred as a **consensus sequence** from **EST** based **multiple sequence alignment.**

- **SNP**: stands for *S*ingle *N*ucleotide *P*olymorphism. This refers to a particular nucleotide (or "base") in a DNA sequence that is variable within a species (or between related species). For example, at a certain position in a DNA sequence there may be a **C** (cytosine) present in some individuals but a **T** (thymine) present in others (**C/T** polymorphism). SNPs represent the most basic form of genetic polymorphism. There are tens of millions of SNPs present in the genome of a typical organism and can be used as genetic markers (SNP markers).