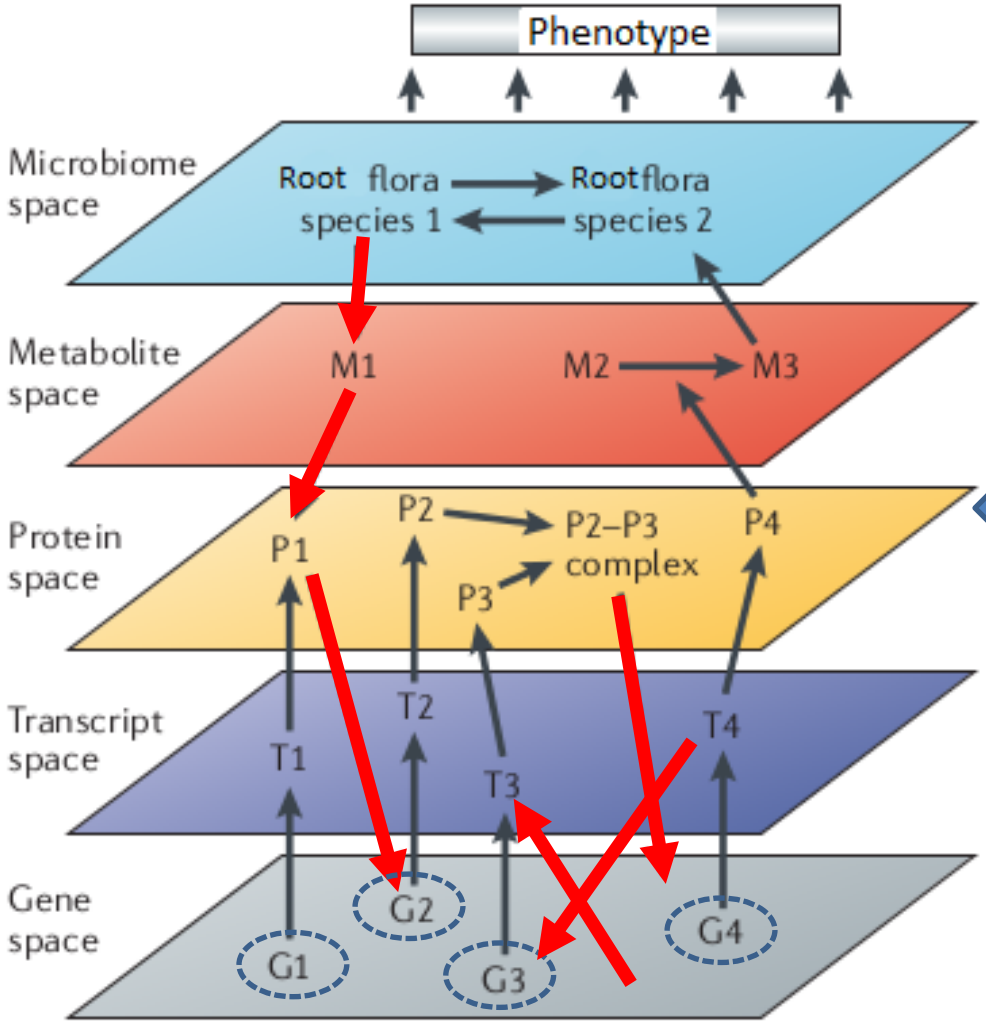
Genomic Technologies in Tree Breeding

Ross Whetten
Professor
Forestry & Environmental Resources
NC State University

Overview

- "Genomic technologies" – what is that?
- Status of conifer genome sequences
- Potential applications to breeding
- Cost-effectiveness

A General Model



Environment

Modified from Civelek & Lusis, Nat Rev Genet 15:34-48, 2014

Genomic technologies

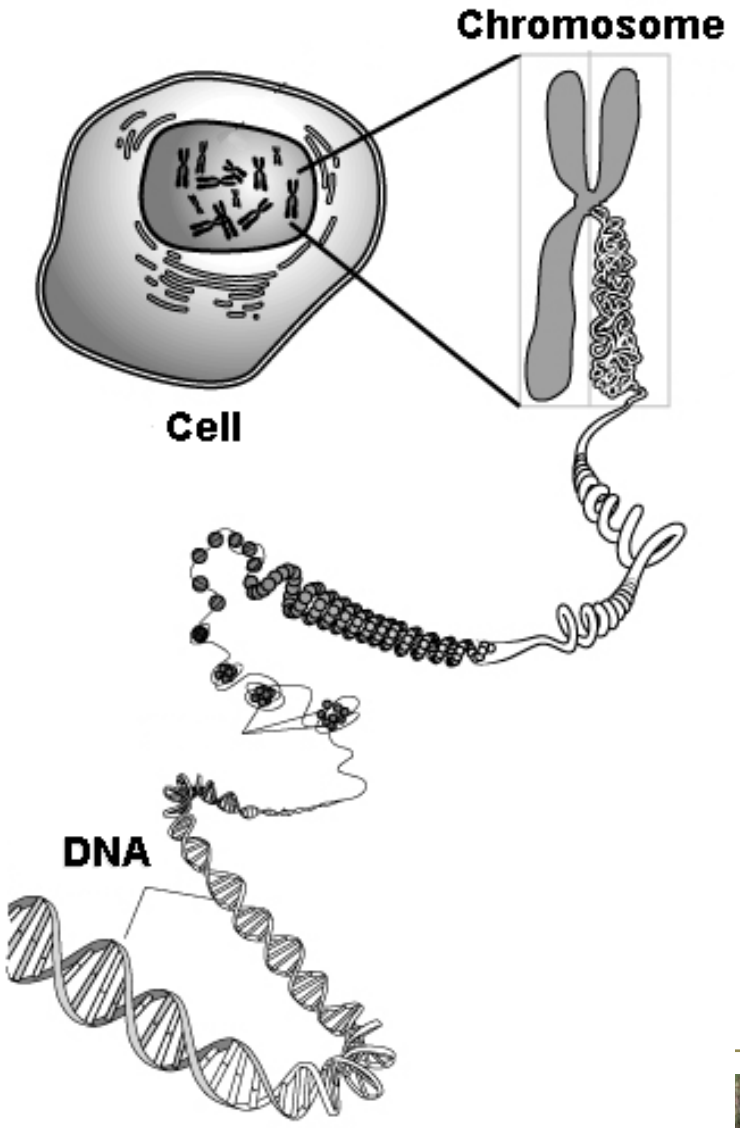
- DNA sequencing using massively-parallel methods
  - Total genomic DNA
  - Functional complexes of DNA and proteins
  - DNA copies of RNA molecules
- Array hybridization – gene expression and variant detection
- Proteomics - separation and analysis of proteins
- Metabolomics - separation and analysis of metabolites
- Metagenomics - Analysis of mixed populations of microbes by DNA sequencing

What and where is the genome?

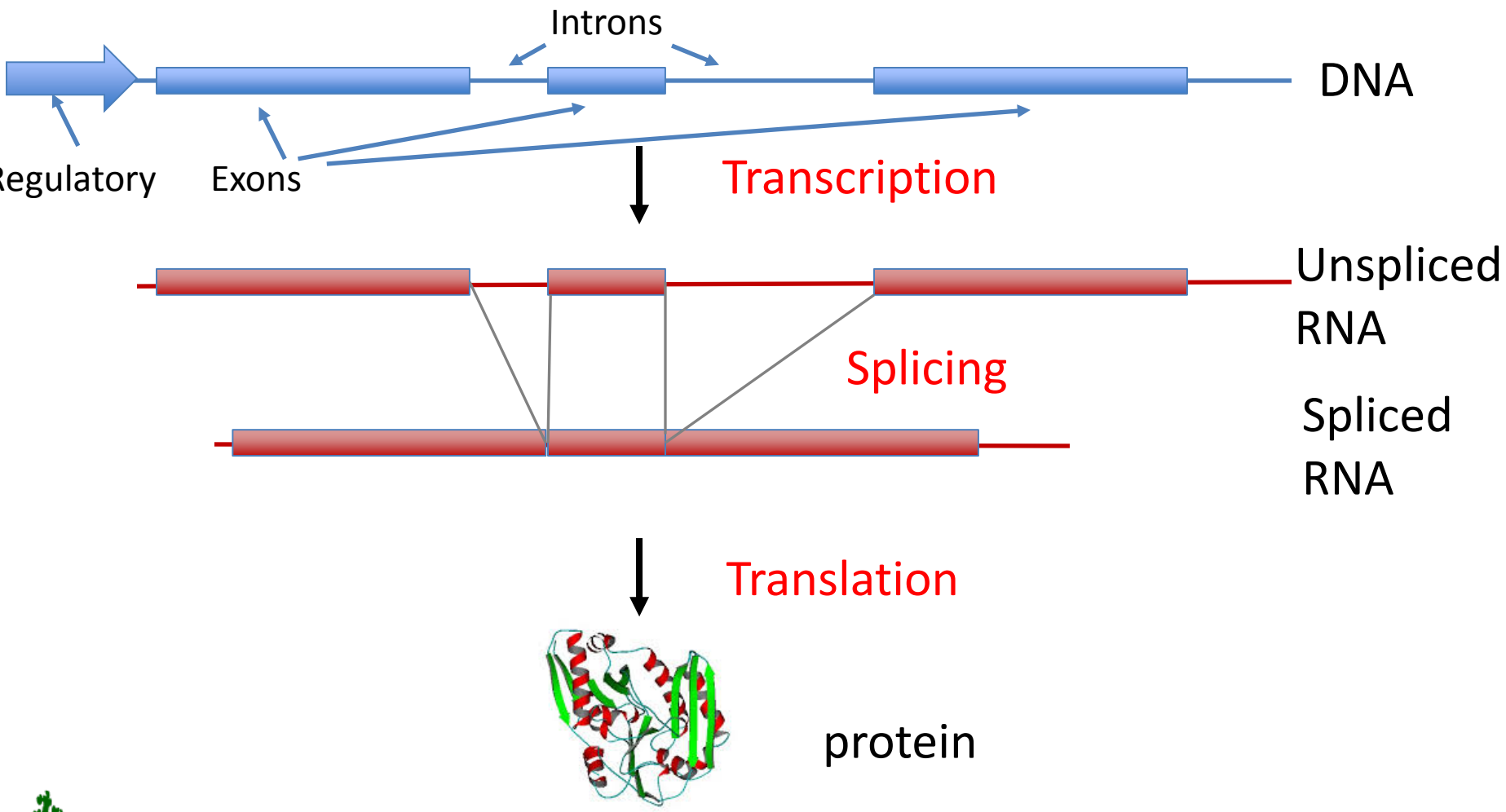
Diploid pine cells each contain about 42 linear feet of DNA, packed into 24 chromosomes in the cell nucleus.

How DNA is packed into chromatin, and unpacked to allow gene expression, is an important part of how genes are regulated, but we know little about how this occurs in trees

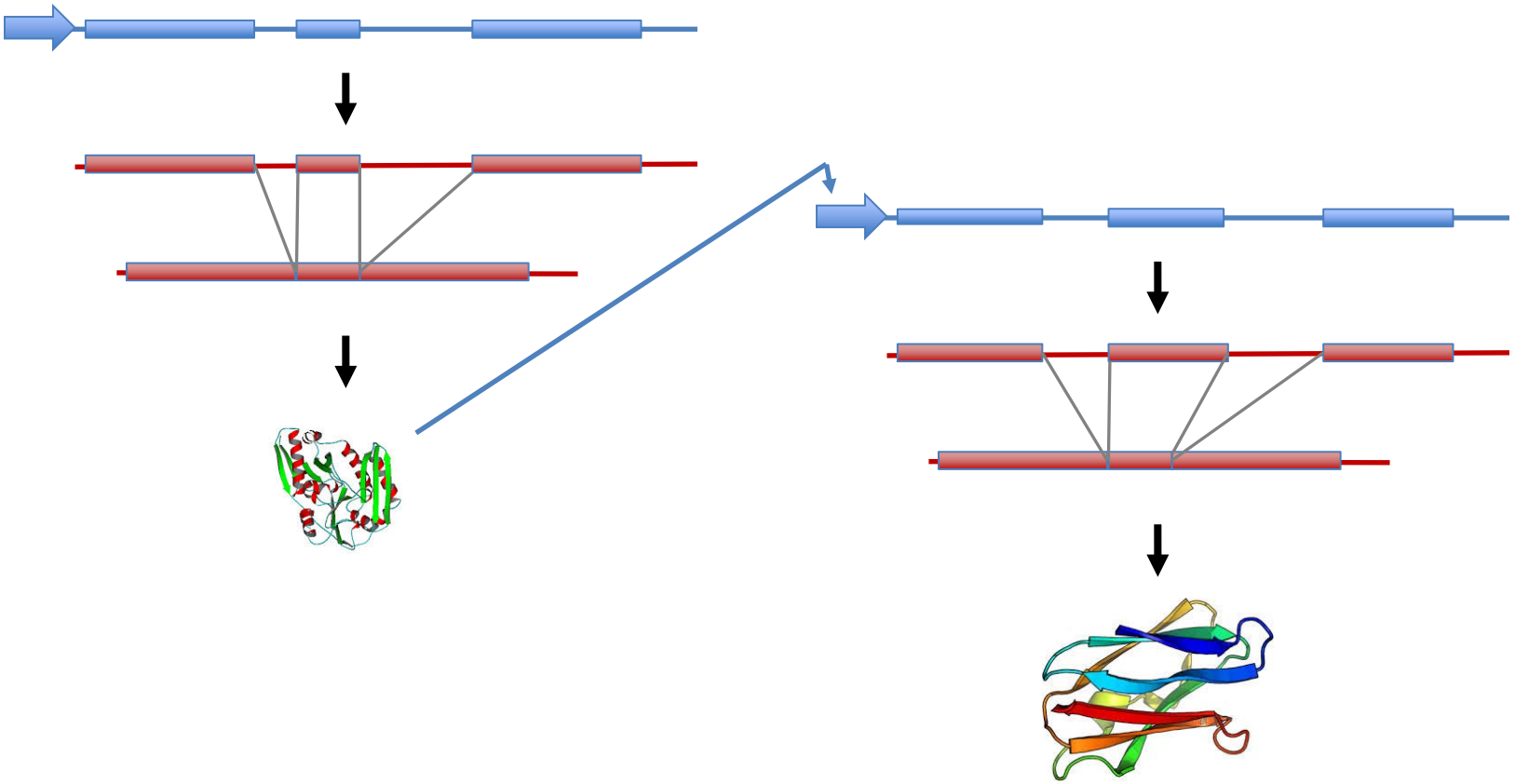
Methods to analyze chromatin structure rely on availability of an assembled genome sequence



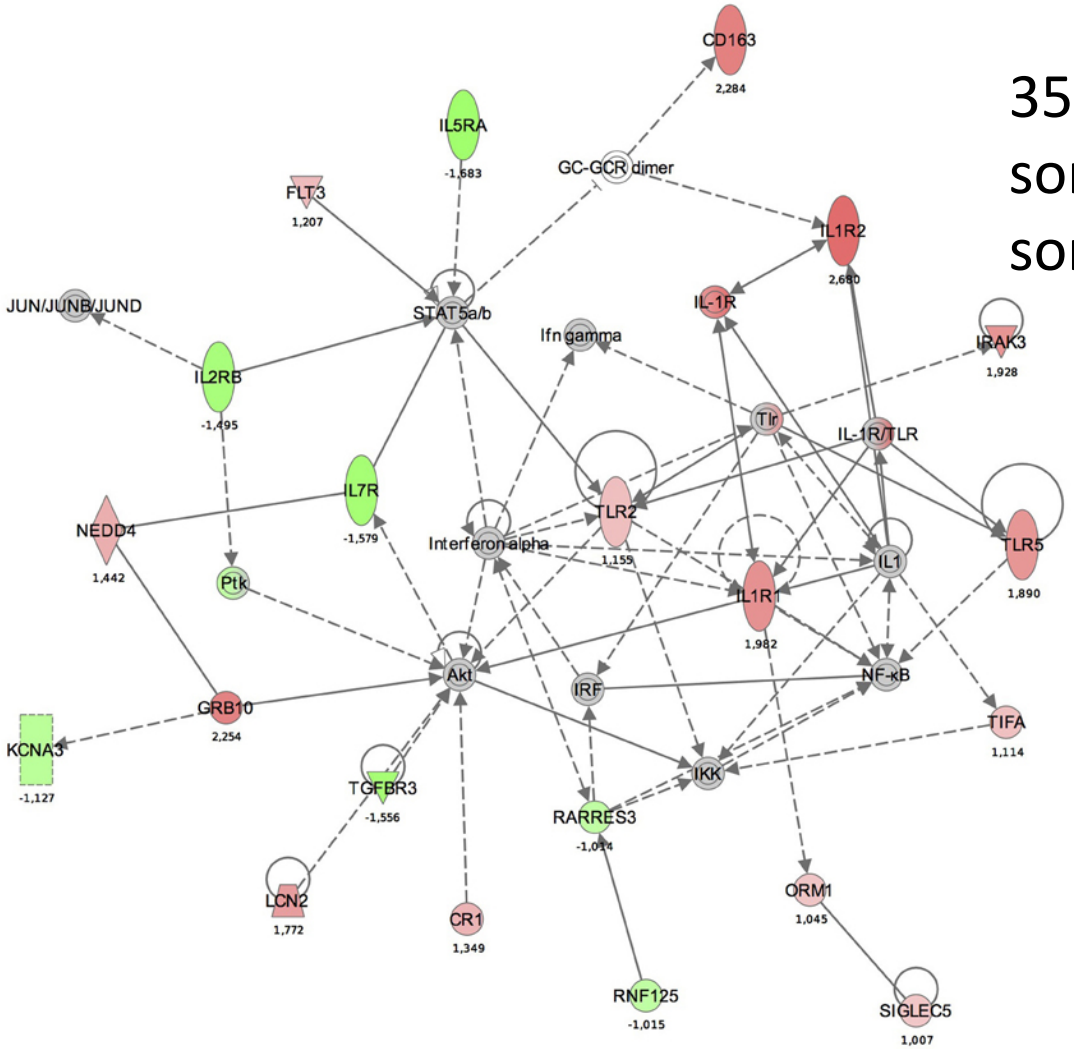
What do genes do?



Some genes regulate other genes…

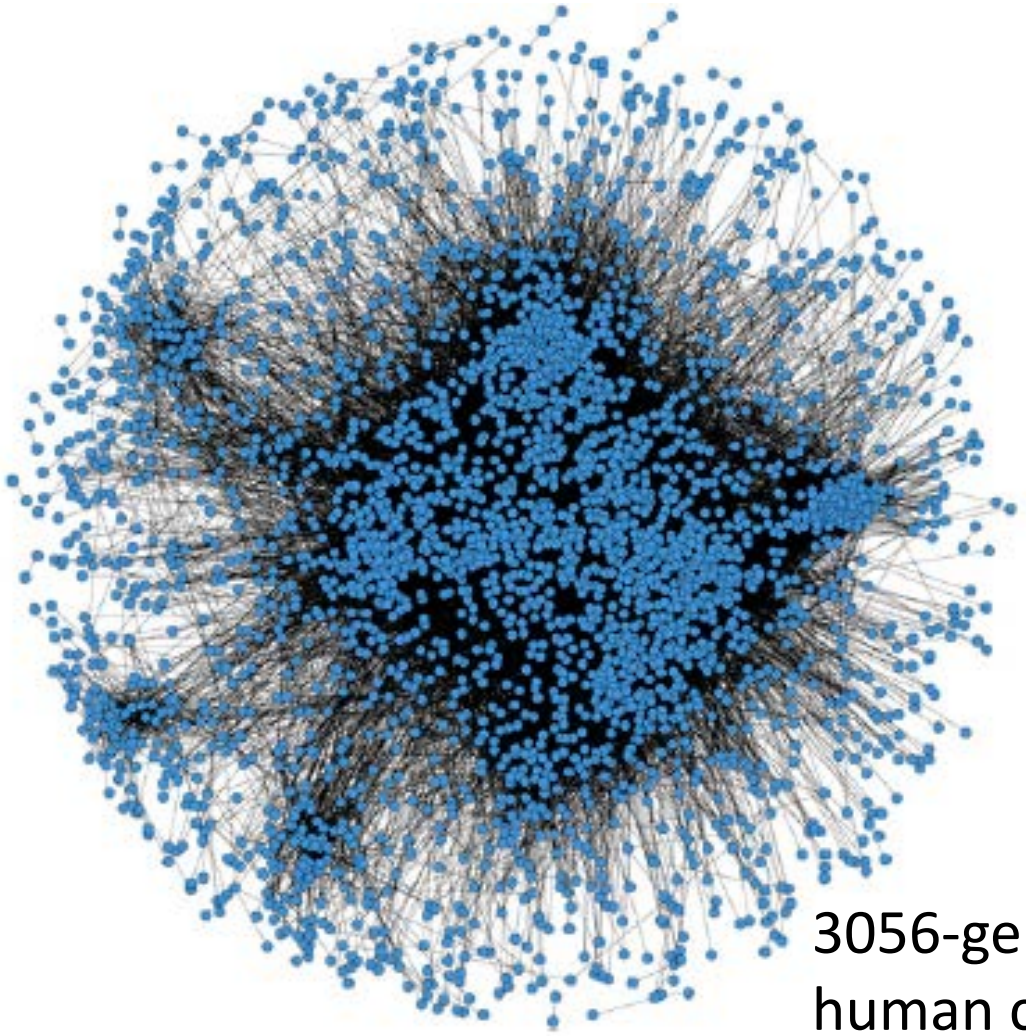


…and those may regulate others …



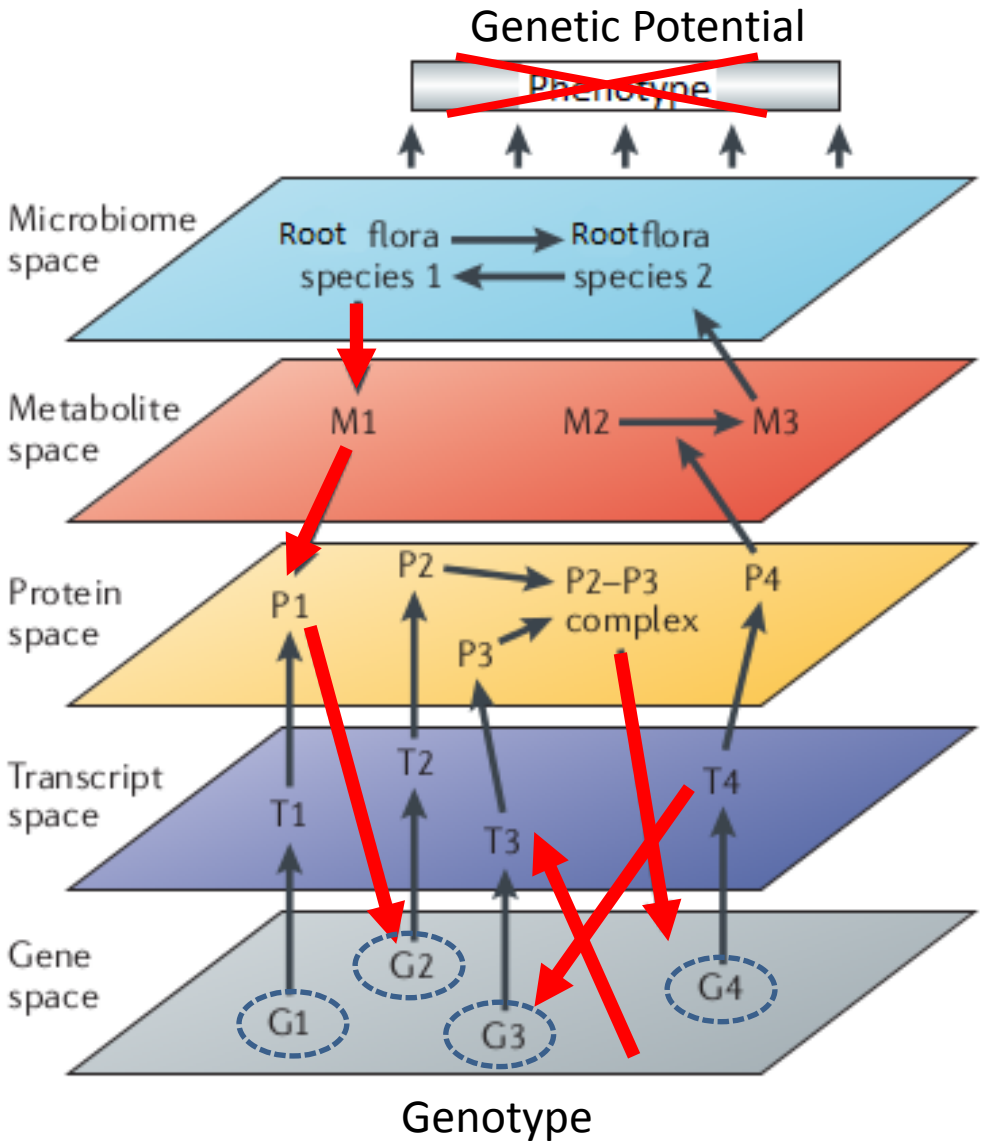
35-gene interaction network: some positive interactions, some negative interactions, some auto-regulation.

…and it gets complicated



3056-gene interaction network, human cell lines

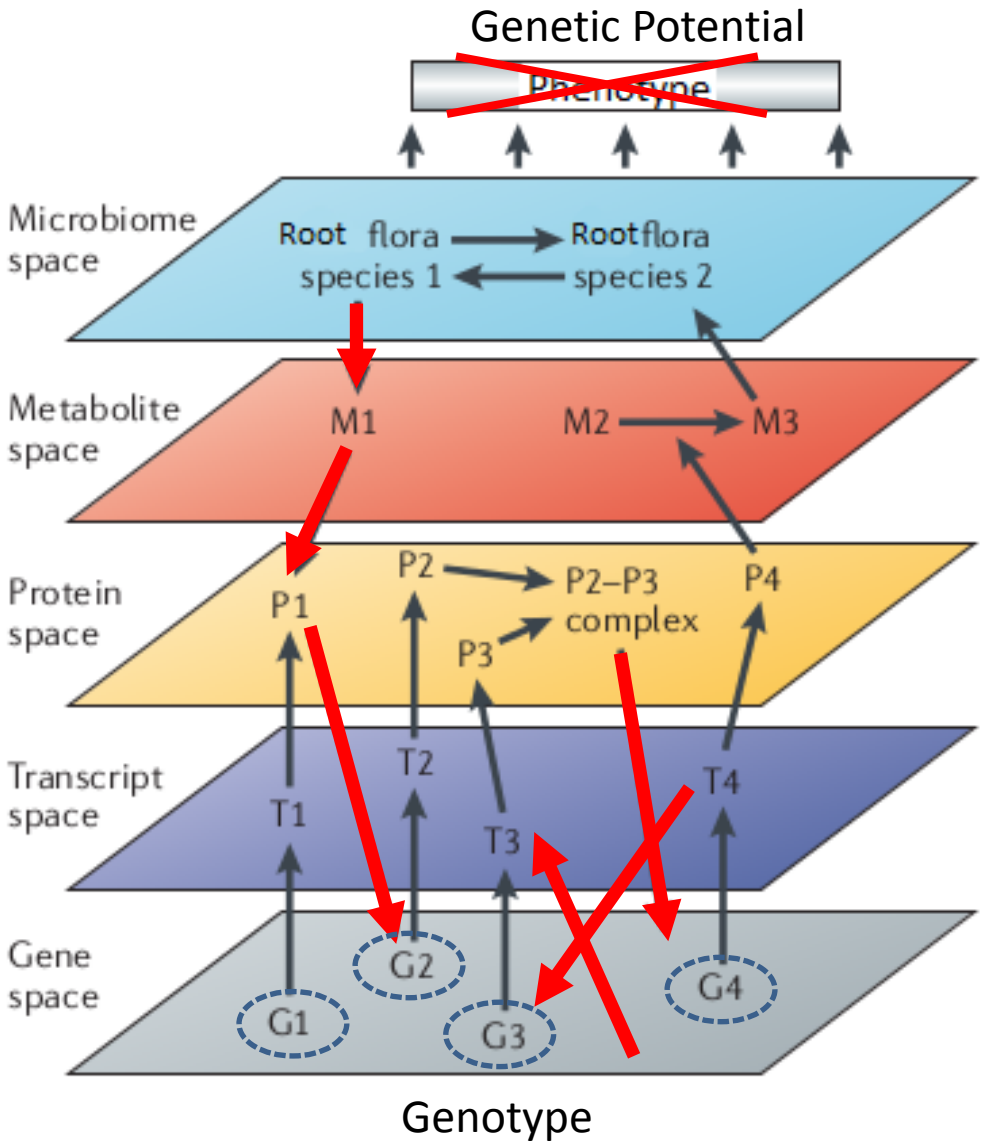
A Less General Model



Components

Neutral sequences
Coding sequences ~ 1%
Regulatory sequences ~ 1%

A Less General Model

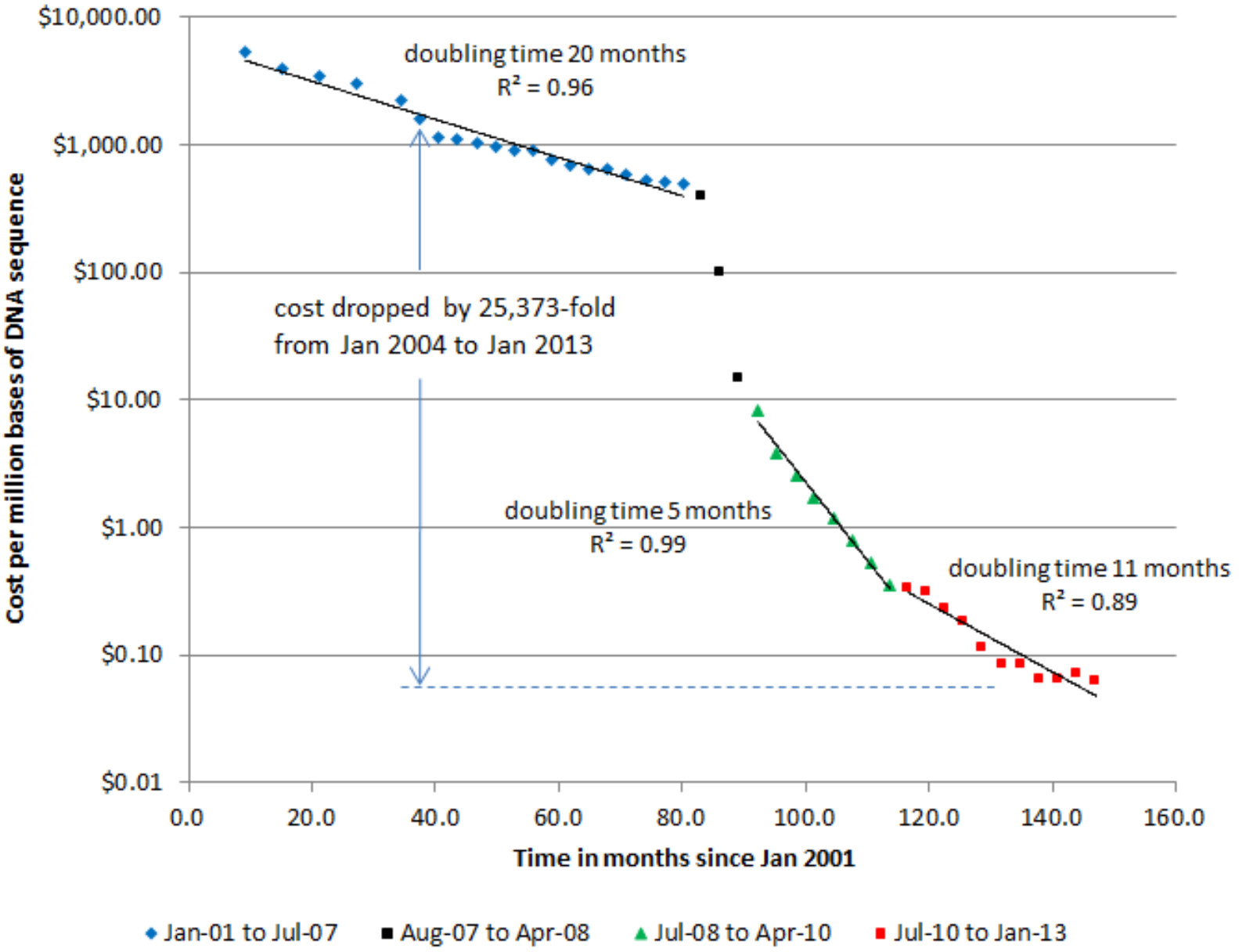


Components

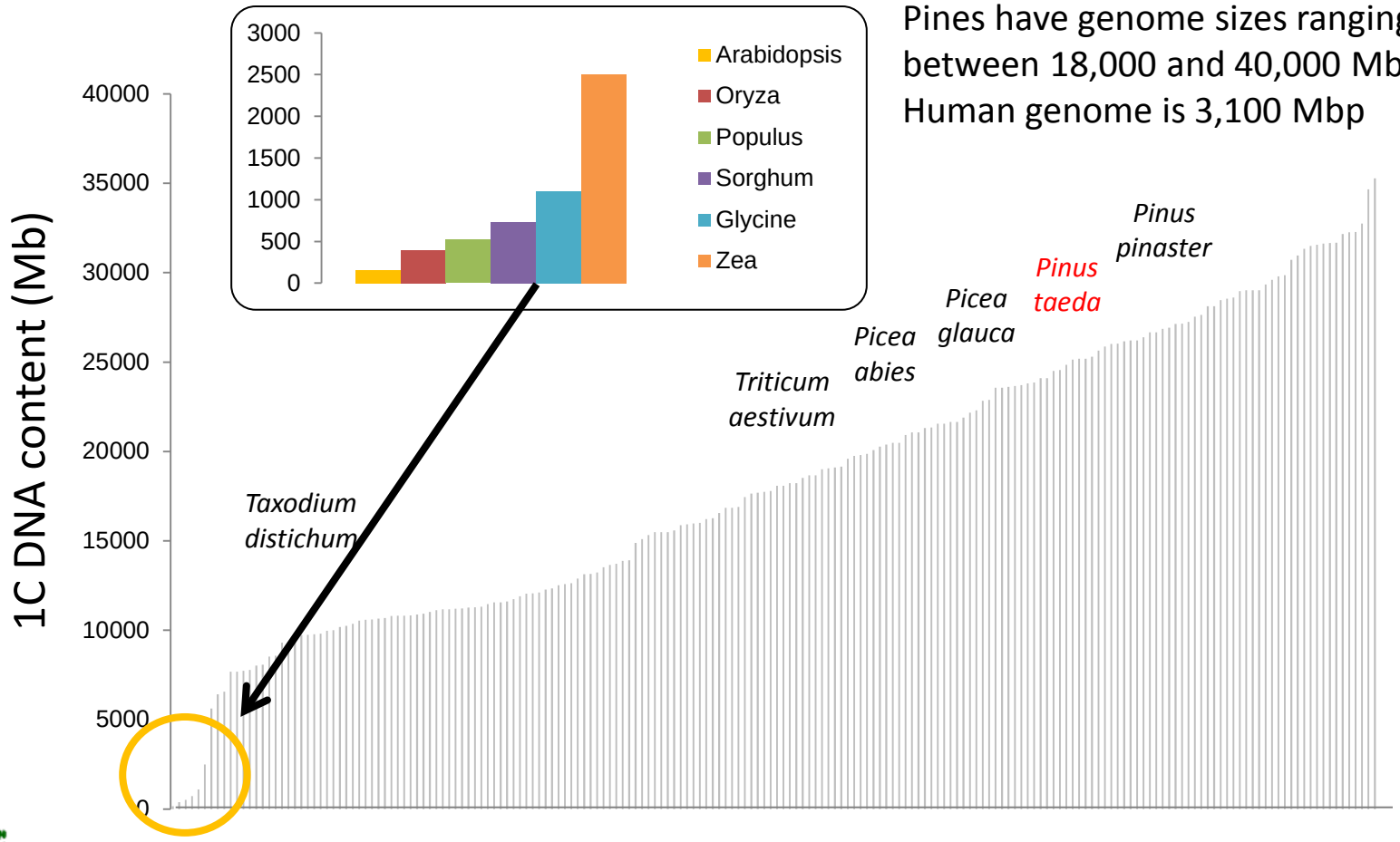
Messenger RNA – protein-coding
Short non-coding – splicing factors
microRNAs – regulatory
Long non-coding – regulatory?

Extensively modified from Civelek & Lusis, Nat Rev Genet 15:34-48, 2014

DNA sequencing cost at NIH genome centers



Genome size



Pines have genome sizes ranging between 18,000 and 40,000 Mbp. Human genome is 3,100 Mbp

PineRefSeq

Assembling the Reference Sequence

Based on Whole Genome Shotgun Sequencing

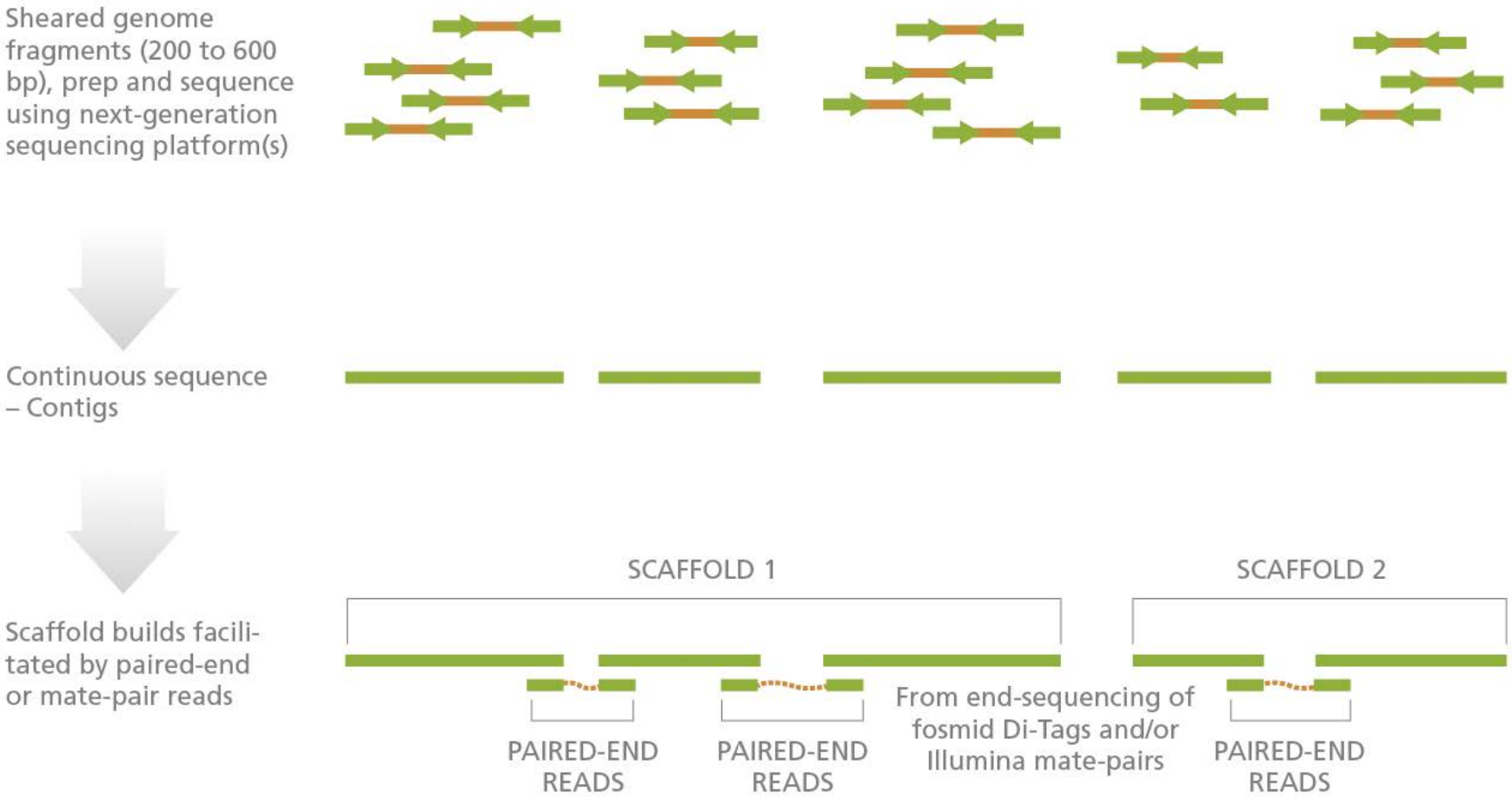


Figure Credit: Nicholas Wheeler, University of California, Davis

Genome sequencing progress

Sugar pine – 34 billion base-pairs

- 1.4×10^{12} bases of DNA sequence
- 58.4 million scaffolds in v 1.0 assembly

Loblolly pine – 23 billion base-pairs

- 1.9×10^{12} bases of DNA sequence
- 14.4 million scaffolds in v1.01 assembly

Douglas fir – 18.6 billion base-pairs

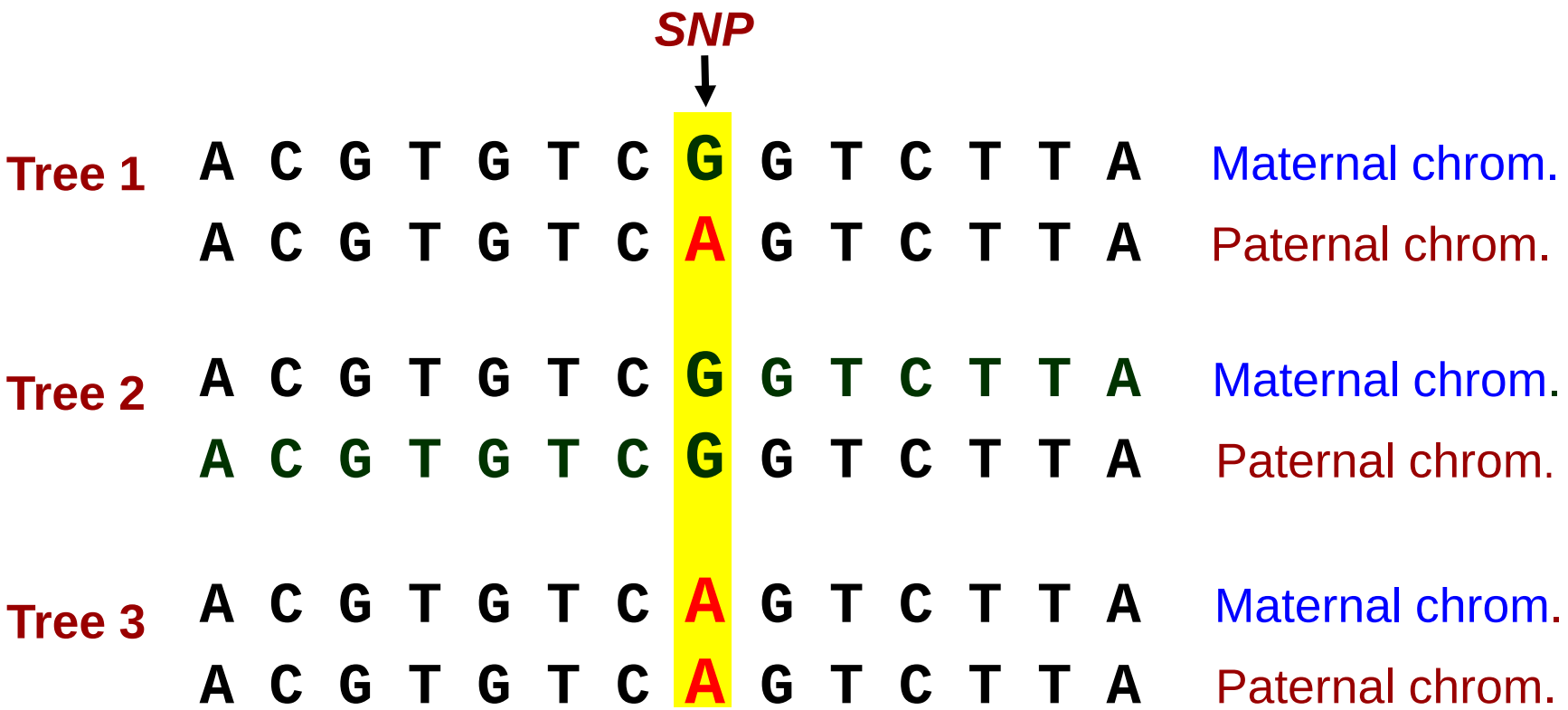
- 1.1×10^{12} bases of DNA sequence
- 39.6 million scaffolds in v 0.5 assembly

Data from http://pinegenome.org/pinerefseq/ and http://www.ncbi.nlm.nih.gov/SRA/

From first principles

- Three major classes of genetic variation
  - –Variation in coding sequences may cause changes in gene product function
  - –Variation in regulatory sequences may cause changes in timing or location of gene product expression
  - –Neutral variation has no effect at all
- Mapping genotype to phenotype
  - –Genetic covariance usually estimated by allele-sharing
  - –Similarity in gene expression patterns is another level of genetic covariance that integrates environmental information and genetic interactions
  - –Neutral variation has value only if in LD with causative variants

Single-nucleotide polymorphisms
SNPs



Tree 1 is ***heterozygous*** Trees 2 and 3 are ***homozygous***

What to measure?

- DNA sequence variation – stable in development
  - SNP = single-nucleotide polymorphisms
  - Structural variation – in maize, any single inbred line has only about 80% of the total “pan-genome”
- Epigenetic variation – may be unstable
- Gene expression – will this be stable enough?
  - Individual genes
  - Networks or modules of coexpressed genes
- Metabolites or proteins
  - May also be unstable over development

Applied Breeding Options

Pedigree control and population management

- Measuring pollen contamination in orchards
- Confirming validity of controlled-cross offspring

Pedigree reconstruction or estimation of relationships

- Pedigree reconstruction requires fewer markers
- Estimation of realized relationships can add more value

Prediction of breeding value based on markers

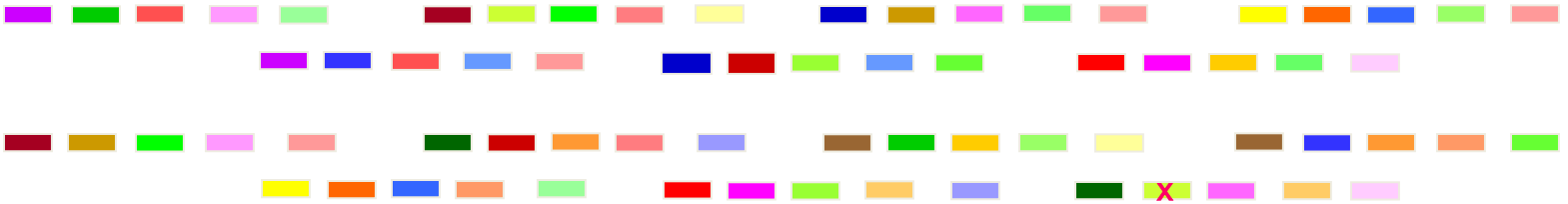
- QTL mapping or association genetics
- Genomic selection based on high-density marker genotyping

Mutations and linkage disequilibrium

Many copies of a particular chromosome exist in a population – in this example, a chromosome has five different genes, and each gene has seven haplotypes

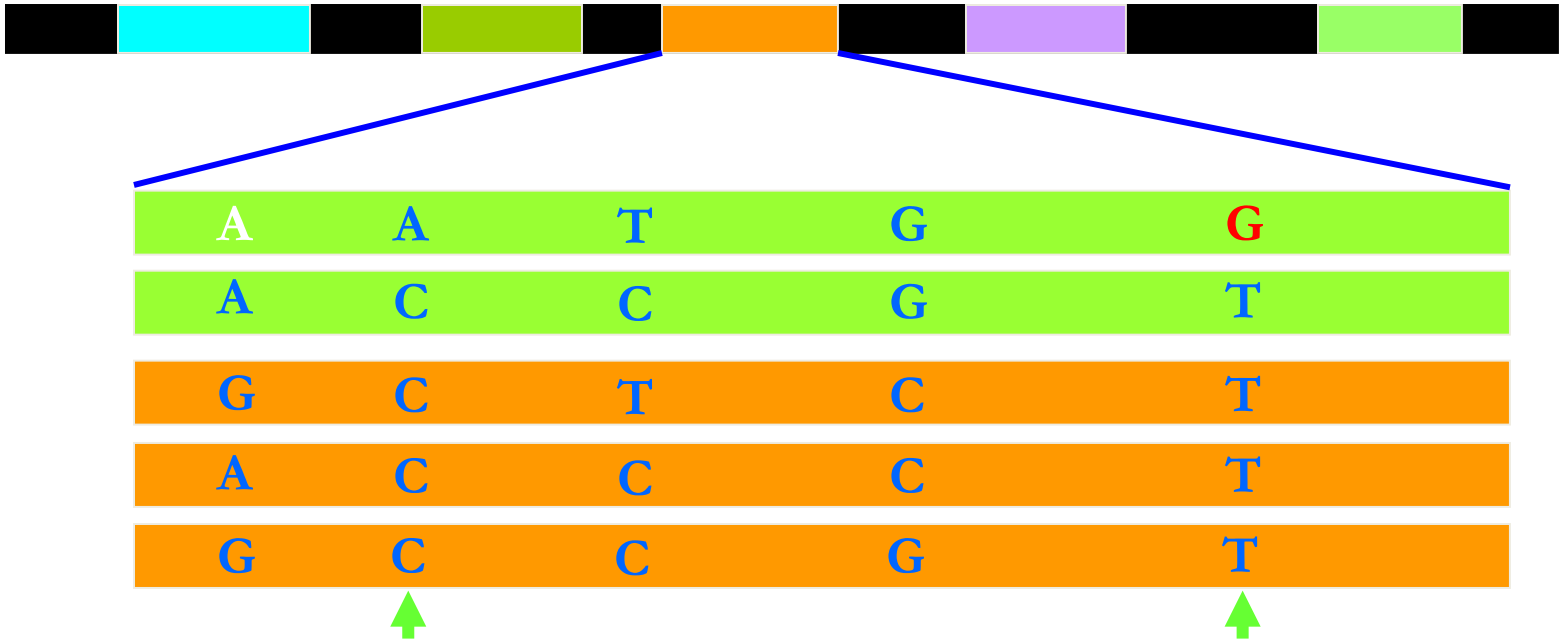
A mutation occurs in a particular haplotype that causes a genetic difference in one of the five genes on the chromosome, and results in a new phenotype.

That mutation will remain associated with that same haplotype of that gene for many generations, but recombination will exchange all the other haplotypes so that there is no association between any other gene and the new phenotype caused by that particular mutation in that particular gene.



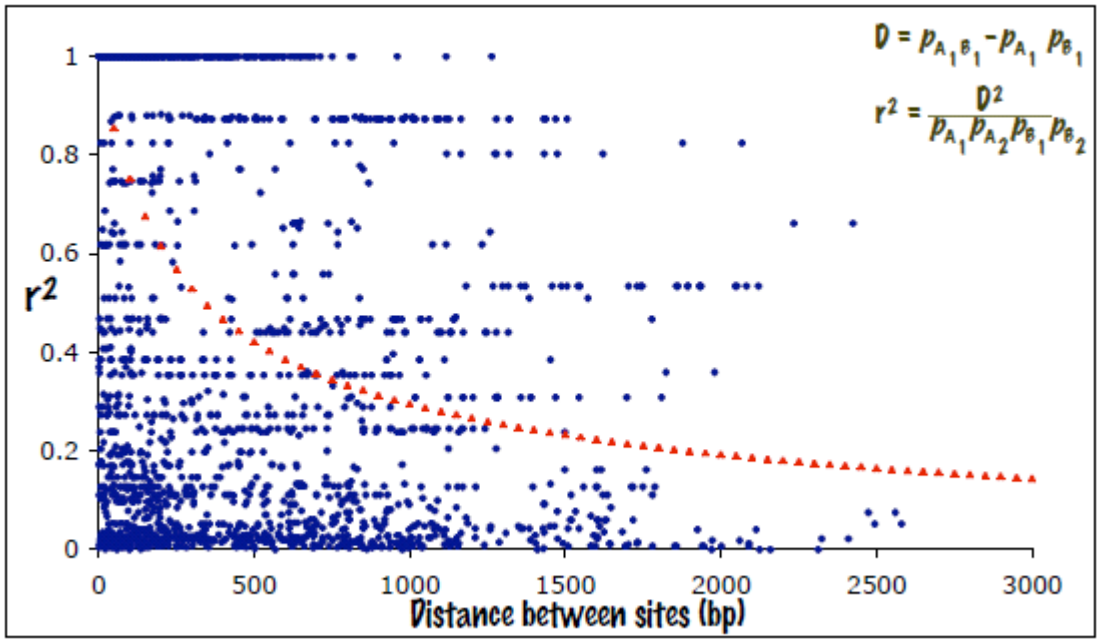
Linkage disequilibrium

How is genetic variation distributed in populations?



SNPs in the same haplotype may be in linkage disequilibrium – the presence of a particular allele at one locus is associated with the presence of a particular allele at the other locus across the population

Linkage disequilibrium decreases with distance in *Pinus taeda*



SNPs in the same gene can be in disequilibrium; SNPs in different genes are likely to be in complete equilibrium

How to measure it?

- SNP variation
  - –Arrays – closed system, only assay what is on array
  - –DNA sequencing – open system, detects what is present, also provides haplotype information
- Epigenetic variation – hard to do high-throughput
- Gene expression
  - –Arrays
  - –Sequencing – provides data on both sequence variants and relative expression levels
- Metabolites – chromatography and mass spec

Summary

- Conifer genomes are complex
  - Low signal-to-noise ratio
  - Much remains to be learned
  - Work is underway to improve the assembly
- Technology platforms vary in cost and value
  - Convenience of array platforms may be offset by closed nature of system and cost structure
  - Sequencing costs are decreasing and sequence data have value for discovery of rare variants and haplotypes
  - Scale of costs is an important consideration for cost-effectiveness

Summary

- "Intermediate phenotypes" interact
  - Within-level and among levels
  - With environment
  - The network of interactions can be quite robust
- Predicting genetic potential from genomic data
  - Genetic covariance usually estimated by allele-sharing
  - Similarity in gene expression patterns is another level of genetic covariance that integrates environmental information and genetic interactions
  - Metabolite variation may contain additional information, or it may simply add more predictors

Summary

- Cost-effectiveness
  - Do the right experiment first
  - Look for opportunities to scale efficiently
  - Successive approximations to the ideal
- Breeding applications
  - Perfect predictive power is not required
  - Breeding strategies may need to change for some applications
  - Willingness to change breeding strategies may depend on cost and predictive power of the genomic tools to be used

Why bother?

- Ecologically important
  - –Clean air & water, other “ecosystem services”
  - –Wildlife habitat
  - –Recreation
- Economically important
  - –Timber production is an important part of the US economy, particularly the southeastern US
- Climate change puts forests at risk
  - –Better understanding of adaptability is needed
  - –Phenotypic plasticity versus genetic adaptation

Systems genetics approach

"Systems genetics is an approach to understand the flow of biological information that underlies complex traits. It uses a range of experimental and statistical methods to quantitate and integrate intermediate phenotypes, such as transcript, protein, or metabolite levels, in populations that vary for traits of interest."

- Civelek & Lusis

Nat Rev Genet 15:34-48, 2014

The hard part…

- Can we predict genetic potential based on DNA variation in the pine genome?
  - Expect about 20 million SNPs to choose from; most are probably neutral
  - Coding sequences are ~1% of the genome, will contain variants affecting gene structure
  - Regulatory regions of pine genome have not been mapped
- Not clear which genetic variants will be informative or should be assayed

Our approach

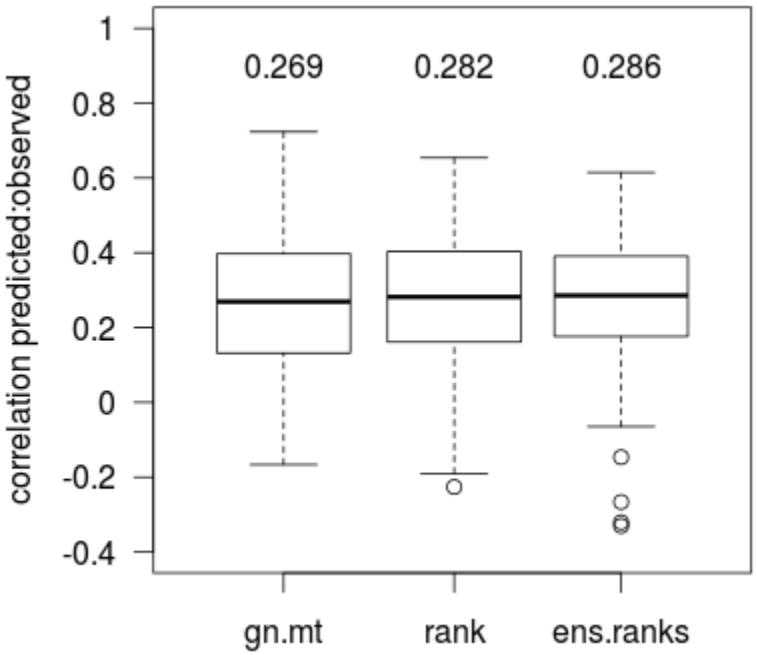
- Measure at a level of detail appropriate to the scope of the system
  - No inbred lines, so half-sib or full-sib families are the appropriate "genetic entries"
  - Distinguishing signal from noise is a challenge
- Predictive power is a higher priority than mechanistic understanding
  - "Missing heritability" a problem with GWAS approach
  - Modeling at whole-genome level can have more predictive power

Overview

- Status of conifer genome sequences
- "Genomic technologies" – what is that?
- Methods for discovery and analysis of
  - DNA sequence (genetic) variation
  - Chromatin structure (epigenetic) variation
  - Gene expression variation
  - Protein variation
  - Metabolite variation
  - Associated epiphytic or endophytic microbes
- Potential applications to breeding
- Cost-effectiveness

Preliminary results with pine

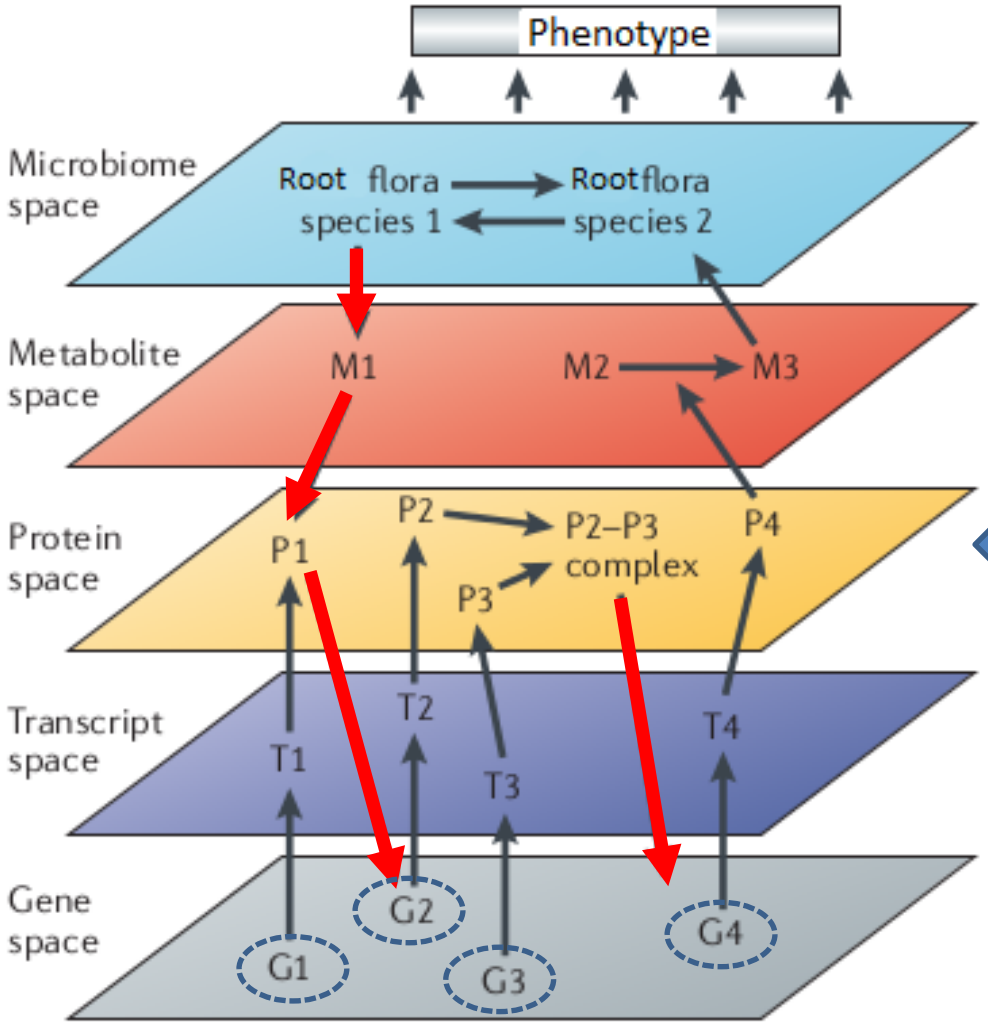
- 243 unrelated clonally-replicated pine genotypes grown in two locations
- Height growth over two years measured for one set of two replicates/clone
- Gene expression (110 genes) and metabolite levels (383 metabolites) measured at age 1 year on two more replicates at a different location
- H^2 of height measurement ~ 0.43



A cross-validation analysis using only small numbers of genes and metabolites accounts for >7% of phenotypic variation, or ~ 17% of heritable variation.

This suggests that gene expression and metabolite covariances are robustly associated with phenotypic variation, even when the association is not causal.

A General Model

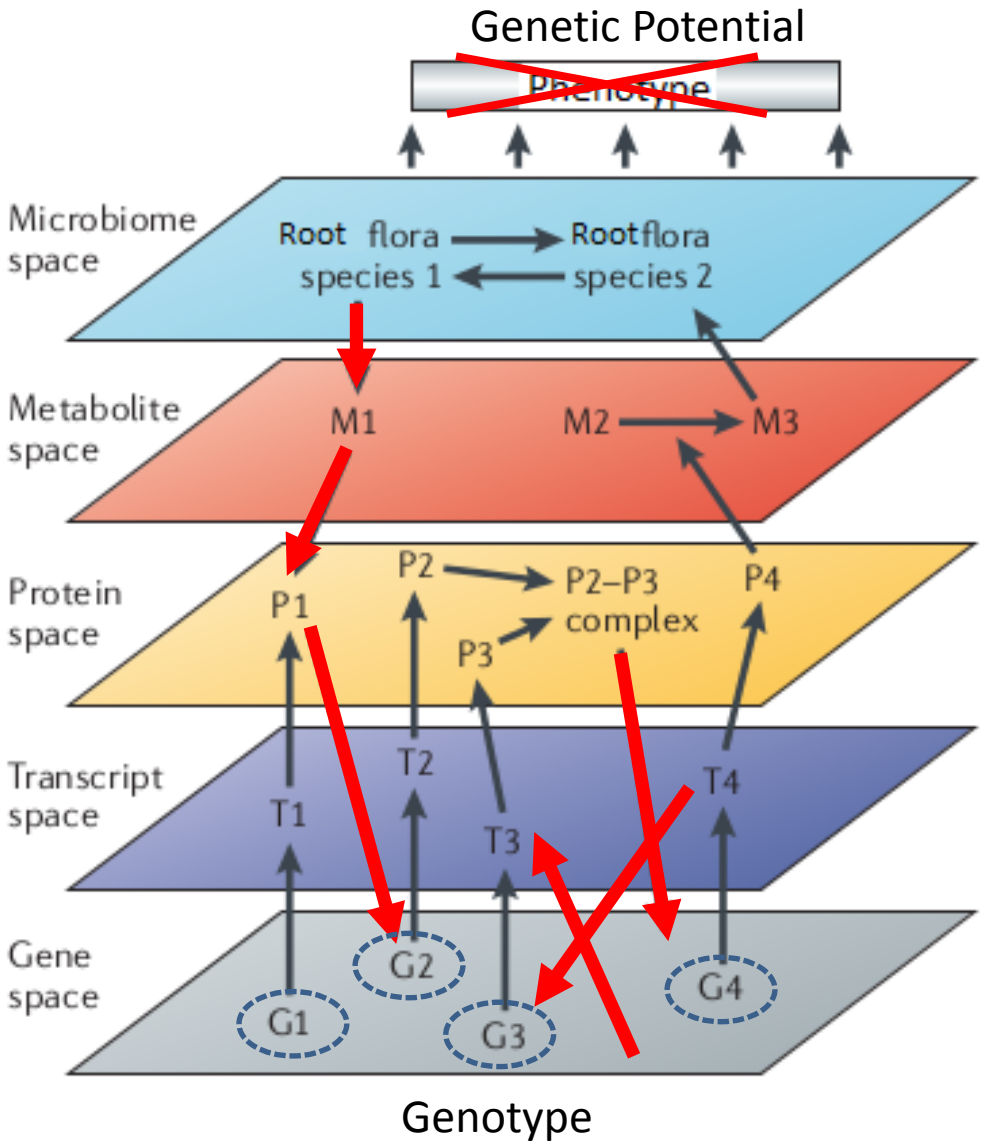


Genotype

Environment

Modified from Civelek & Lusis, Nat Rev Genet 15:34-48, 2014

A Less General Model

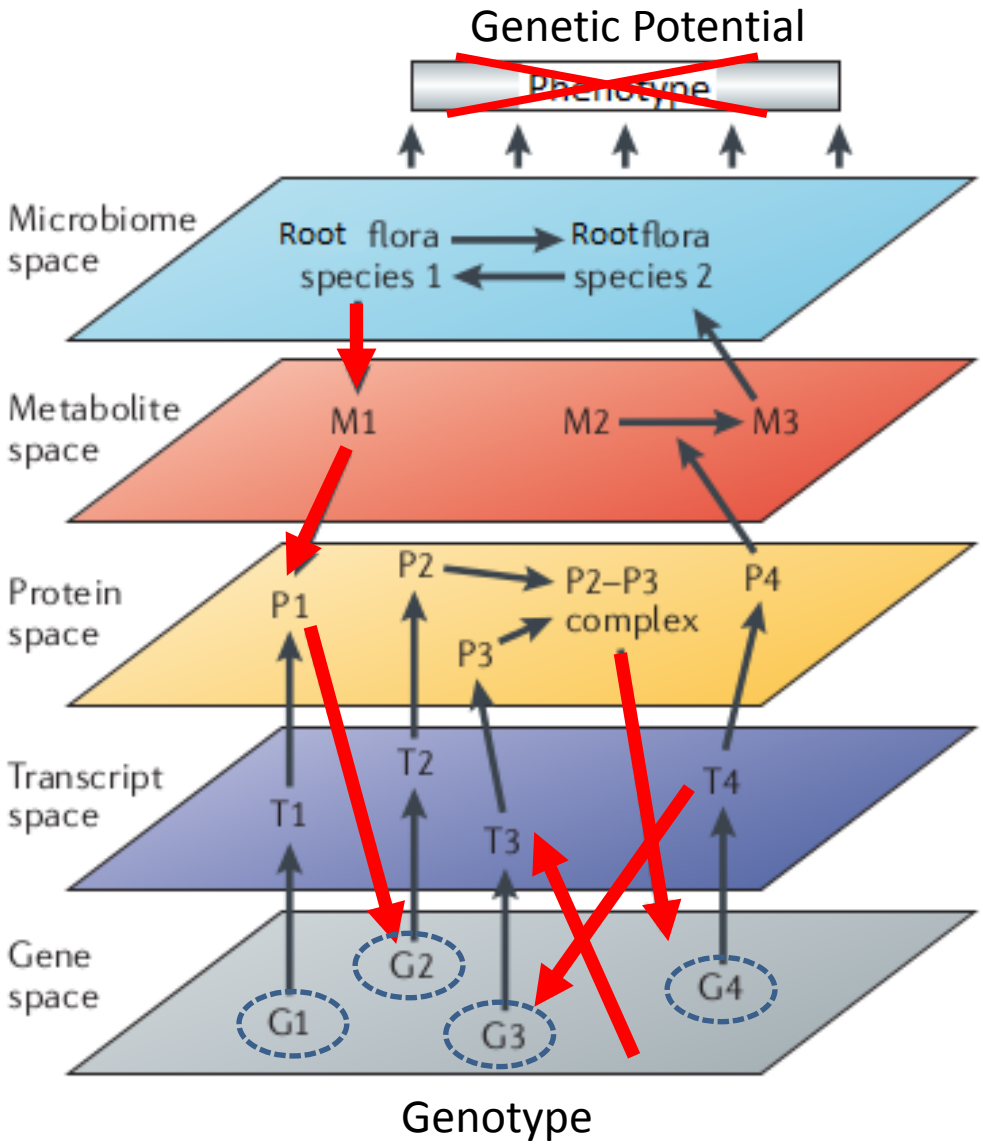


Components

Enzymes – convert metabolites
Structural proteins – cell structure
Regulatory proteins – affect activity of genes or other proteins

Extensively modified from Civelek & Lusis, Nat Rev Genet 15:34-48, 2014

A Less General Model



Components

Small molecules – sugars, amino acids, nucleotides, intermediates
Large molecules – polymers of small molecules

Extensively modified from Civelek & Lusis, Nat Rev Genet 15:34-48, 2014

Key concepts for genetic analysis

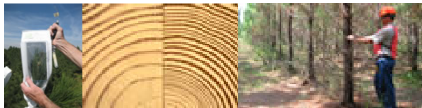
Example: Imagine a chromosome that contains five genes

After many generations of DNA replication and cell division, the population will contain many different versions of this chromosome and the genes that it contains. Some mutations affect the function of a gene (x), while others do not (o). Mutations can be insertions, deletions, or substitutions of one base for another in the DNA sequence. Most are Single Nucleotide Polymorphisms, or SNPs

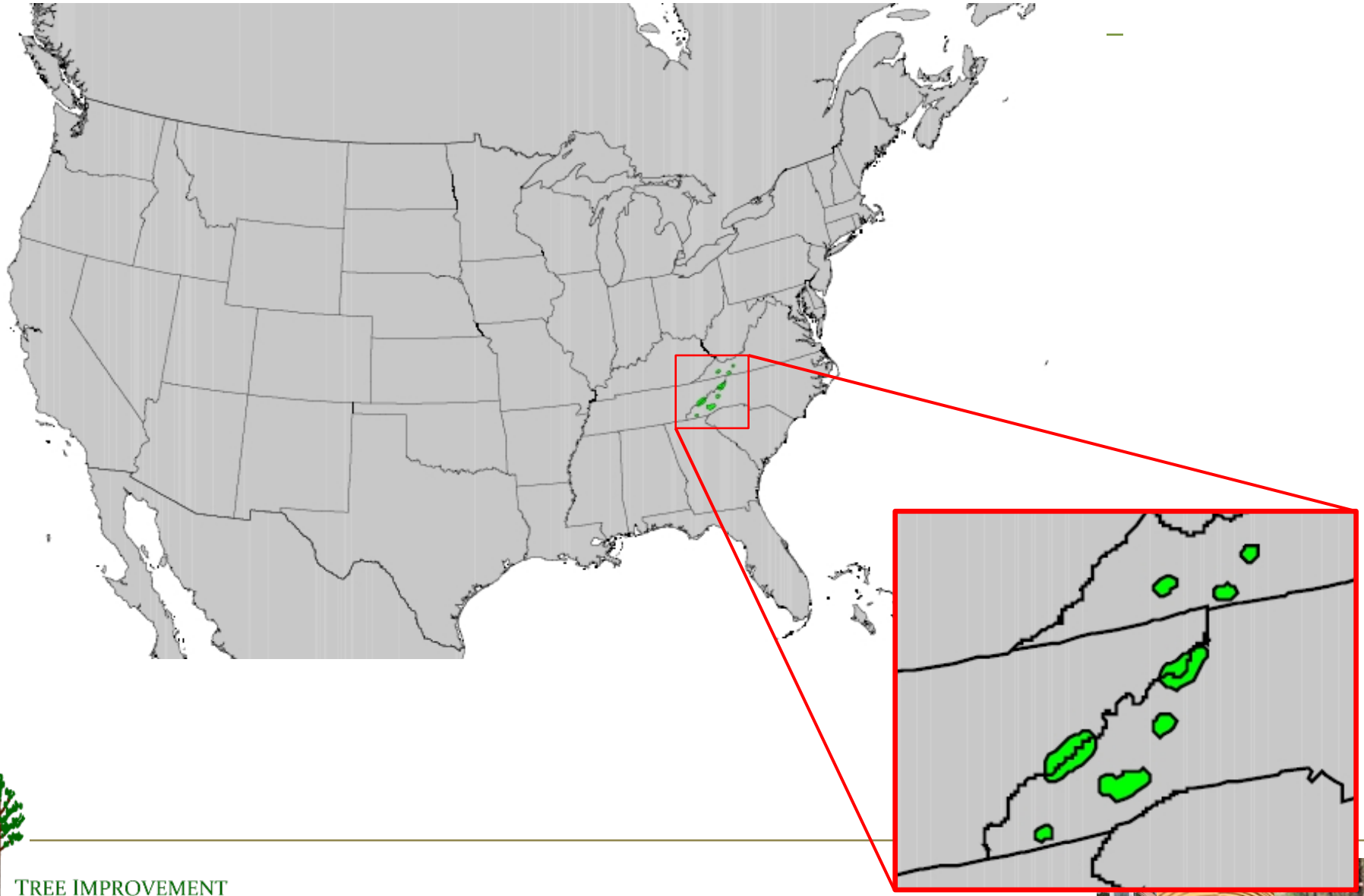
Single-nucleotide polymorphisms (SNPs)

28 SNPs in 1530 bp of DNA, 8 haplotypes

```
A,G,T,T,C,T,G,C,G,C,G,C,T,C,C,G,A,G,C,A,A,G,G,T,A,G,A,A
A,G,T,T,C,T,G,C,G,C,G,C,T,C,C,G,A,G,C,A,A,G,G,T,A,G,A,A
A,G,T,T,C,C,G,C,T,C,G,C,C,T,C,G,A,G,C,A,A,T,G,T,A,G,A,G
A,G,T,T,C,C,G,C,T,C,G,C,C,T,C,G,A,G,C,A,A,T,G,T,A,G,A,G
A,G,T,T,C,C,G,C,T,C,G,C,C,T,C,G,A,G,C,A,A,T,G,T,A,G,A,G
A,G,T,T,C,C,G,C,T,C,G,C,C,T,C,G,A,G,C,A,A,T,G,T,A,G,A,G
A,G,T,T,C,C,G,C,T,C,G,C,C,T,C,G,A,G,C,A,A,T,G,T,A,G,A,G
A,G,T,T,C,C,G,C,T,C,G,C,C,T,C,G,A,G,C,A,A,T,G,T,A,G,A,G
A,G,T,T,C,C,G,C,G,C,G,C,C,T,C,G,A,G,C,A,A,T,G,T,A,G,A,G
A,G,T,T,C,C,G,C,G,C,G,C,C,T,C,G,A,G,C,A,A,T,G,T,A,G,A,G
A,G,T,T,C,C,G,C,G,C,G,C,C,T,C,G,A,G,C,A,A,T,G,T,A,G,A,G
A,G,T,T,C,T,G,C,G,C,A,C,C,T,C,G,G,G,C,A,A,T,G,T,A,G,A,G
A,G,T,T,T,C,G,C,G,C,G,C,T,T,C,T,A,G,C,A,A,T,G,T,A,G,A,G
A,A,T,T,C,T,G,C,G,C,G,C,T,T,C,G,A,G,C,A,A,T,G,T,A,G,A,G
T,G,A,A,C,T,A,T,G,G,G,A,T,T,A,G,A,A,T,G,G,T,C,G,G,C,G,G
T,G,A,A,C,T,G,T,G,G,G,A,T,T,A,G,A,A,T,G,G,T,C,G,G,C,G,G
T,G,A,A,C,T,G,T,G,G,G,A,T,T,A,G,A,A,T,G,G,T,C,G,G,C,G,G
T,G,A,A,C,T,G,T,G,G,G,A,T,T,A,G,A,A,T,G,G,T,C,G,G,C,G,G
T,G,A,A,C,T,G,T,G,G,G,A,T,T,A,G,A,A,T,G,G,T,C,G,G,C,G,G
T,G,A,A,C,T,G,T,G,G,G,A,T,T,A,G,A,A,T,G,G,T,C,G,G,C,G,G
```

Natural Range of *Abies fraseri*



Threatened in Native Stands

SNP genotyping by amplicon resequencing in *Abies fraseri*

Target SNP

SNP genotyping by amplicon resequencing in *Abies fraseri*

Non-Target SNPs

Target SNP