# Conifer Translational Genomics Network
# Coordinated Agricultural Project

## Genomics in Tree Breeding and Forest Ecosystem Management

---

## Module 6 – Genetic Markers

***Nicholas Wheeler and David Harry – Oregon State University***
***Jill Wegrzyn – University of California, Davis***

# Background

- Markers denote the detection of some type of sequence variation (a polymorphism) in a stretch of DNA within or among individuals in a taxa.
- Nuclear markers reflect variation in diploid chromosomes; allelic forms segregate in a Mendelian manner
- Organelle genomes are typically inherited uniparentally (a single allele from only one parent)
  - *In angiosperms, both chloroplasts and mitochondria are inherited maternally*
  - *In conifers (with some exceptions), chloroplasts are paternally inherited, mitochondria are maternally inherited*

# Historical significance of markers

- Genetic markers were key to the development of genetics as a science
- Mendelian genetics is the study of how markers are transmitted from parents to offspring
- Markers provide the empirical foundation of population genetics, with many basic and applied uses
- Until the 1970's most markers were morphological
  - *A few were based on immunology (e.g., blood types) or simple chemistry*
  - *Few markers existed for trees*

# History of marker technology

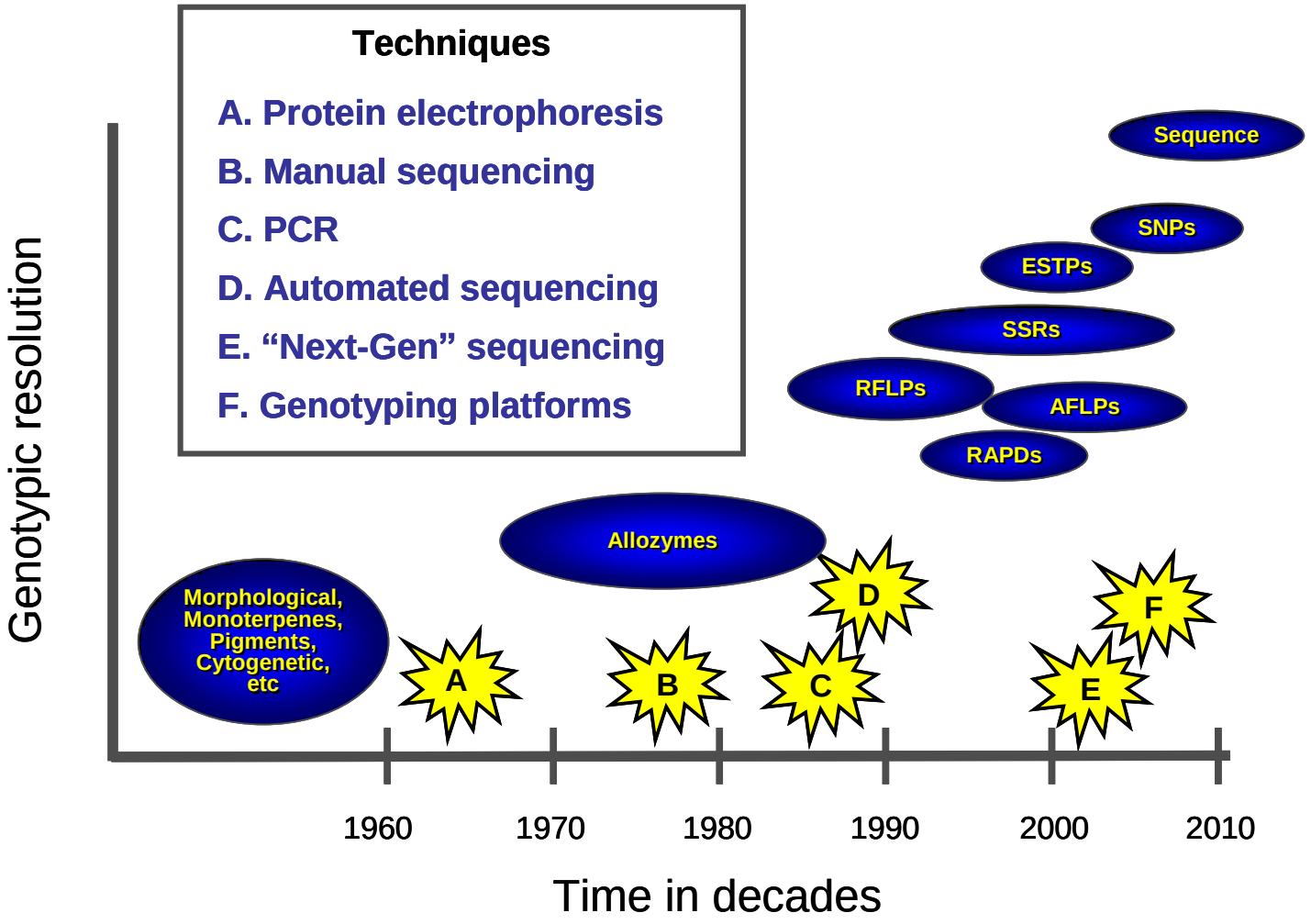





Figure Credit: David Harry, Oregon State University

# How are markers useful to tree geneticists

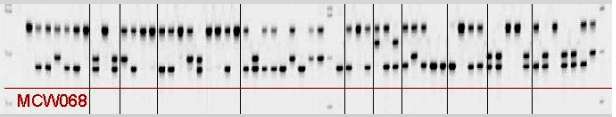



Image Credit: David Harry, Oregon State University

- Fingerprinting / quality control / IP
- Characterizing and managing seed sources
- Assessing relationships among populations and species
- Characterizing within-stand dynamics
- Parentage determination and reconstruction
- Guiding gene resource management and conservation
- Genetic mapping and phenotypic associations
- Marker informed breeding

# What to consider when selecting a marker?

- Reliability and interpretability
- Level of polymorphism (variability)
- Number of markers required for the application
- Cost
- Technical requirements

# Allozymes: Starch gel electrophoresis

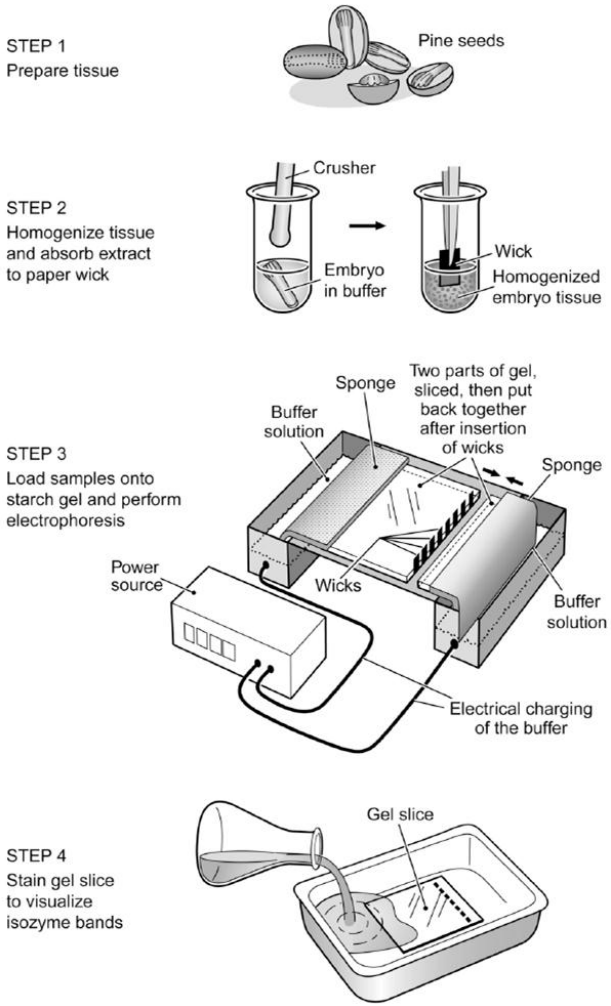



Figure Credit: White, T. L, W. T. Adams, and D. B. Neale. 2007. Forest genetics. CAB International, Wallingford, United Kingdom. Used with permission.

Shikimate Dehydrogenase from *Calocedrus decurrens*

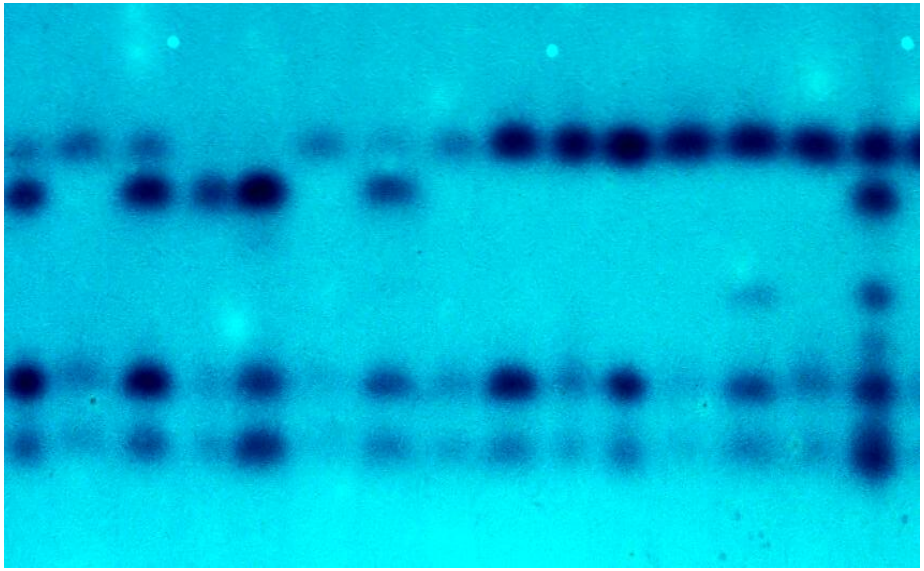


Image Credit: David Harry, Oregon State University

# Allozyme characteristics

## Advantages

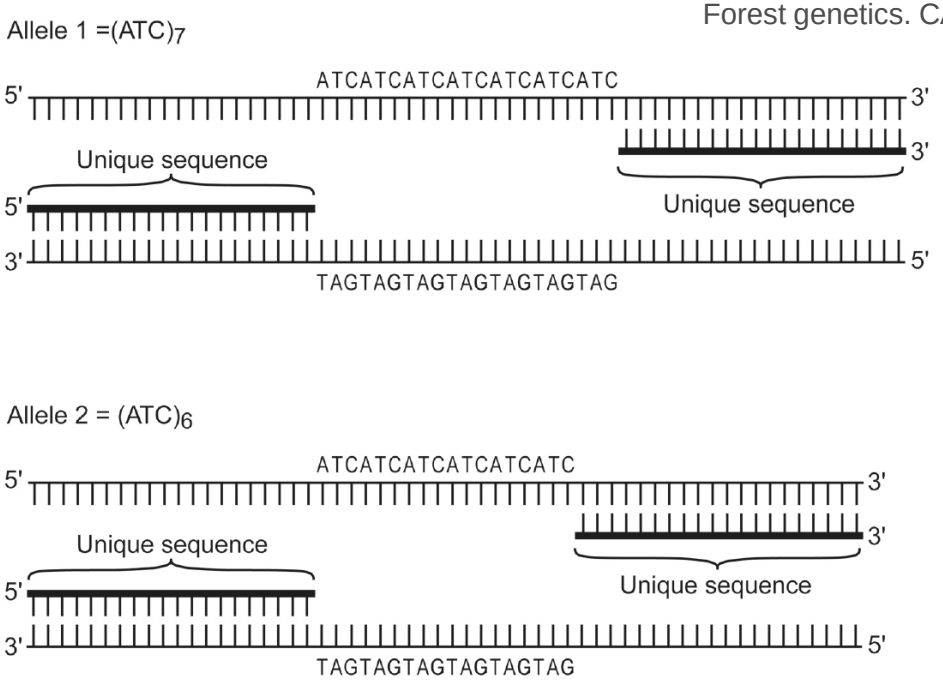
- Widely adaptable across taxa
- Codominant, multi-allelic
- Low cost
- Easy to do (undergrads)
- Captures haploid – diploid biology of conifer seed

## Limitations

- Modest number of markers
- Some environmental influences
- Short shelf life for samples
- Manual scoring

# Microsatellites
## SSR – (Simple sequence repeat)

Figure Credit: White, T. L, W. T. Adams, and D. B. Neale. 2007. Forest genetics. CAB International, Wallingford, United Kingdom. Used with permission.




Sequence Variation

Marker segregation

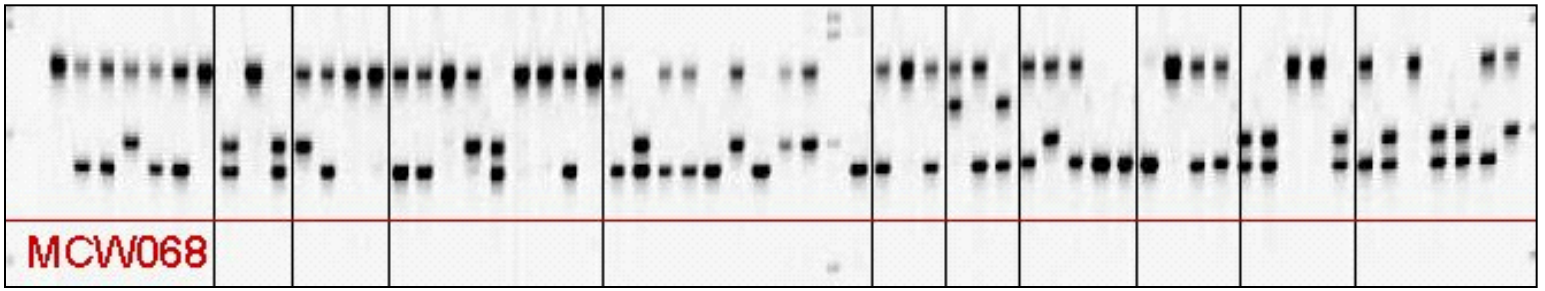



Image Credit: David Harry, Oregon State University

# SSR characteristics

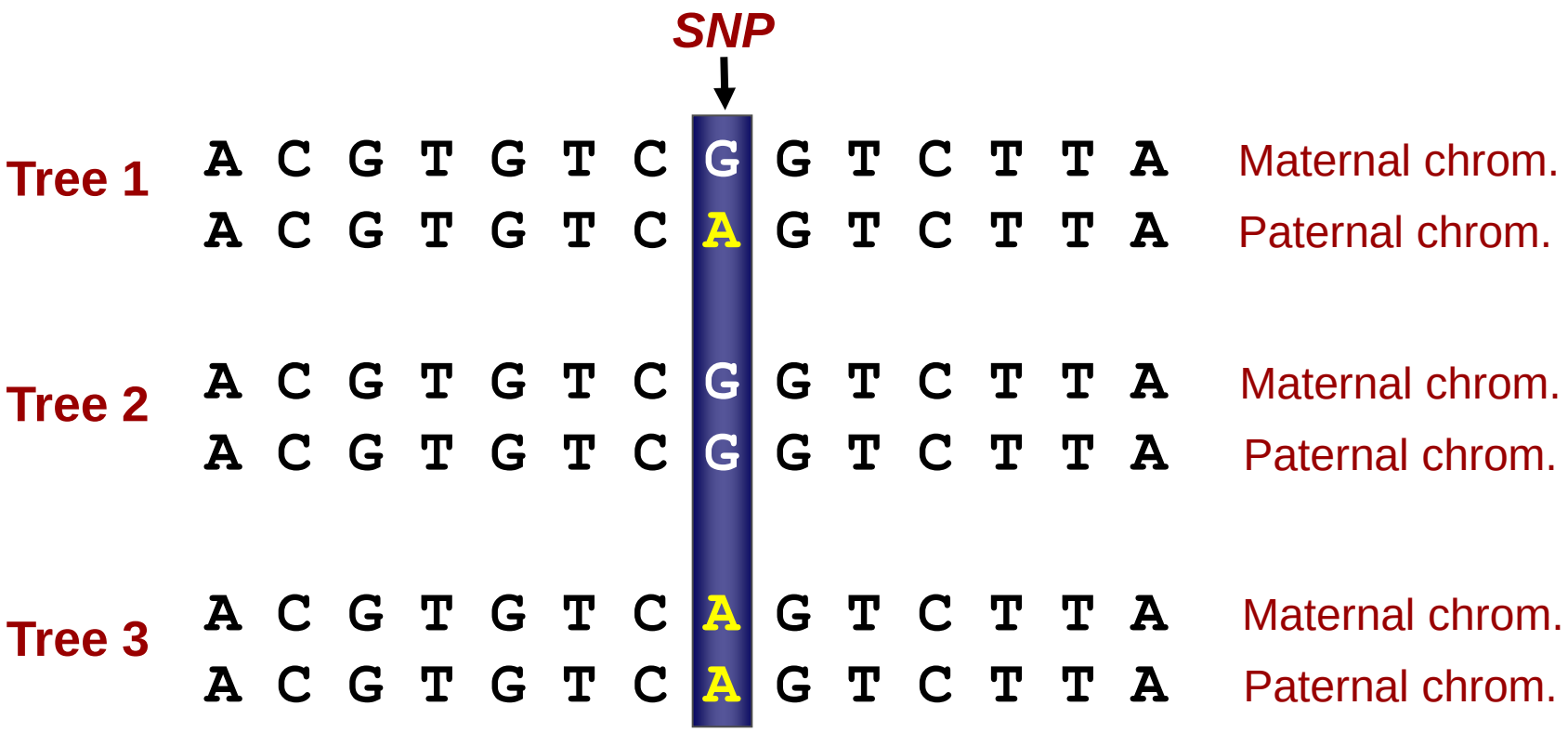
## Advantages

- Reveals many alleles per locus
- Co-dominant, although nulls can be troublesome
- Broadly applicable: mapping, population structure, parentage, fingerprinting
- Numerous

## Limitations

- High development costs
- Development requires technical skills: cloning, sequence, primer design (although vendors can now do this)
- Markers may not transfer easily between related taxa

# Single Nucleotide Polymorphism (SNP)




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



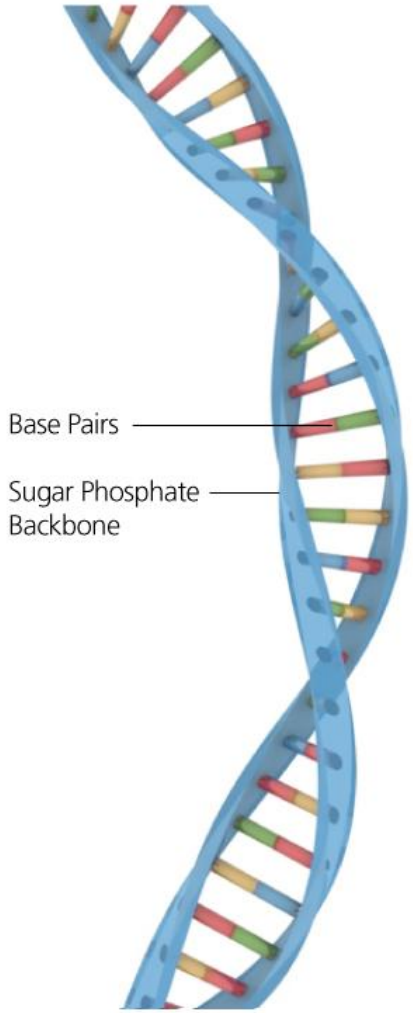
Figure Credit: Glenn Howe, Oregon State University

# Single Nucleotide Polymorphisms (SNPs)

## Single nucleotide differences between allelic sequences

- SNP characteristics
  - *Most common genetic polymorphism*
  - *Result from nucleotide substitutions (e.g. C to T))*
  - *Found in coding and non-coding DNA*
  - *SNPs can alter amino acids and resulting phenotype (but most are "neutral")*
  - *Usually bi-allelic (only two alleles per SNP locus)*
  - *Once assembled, haplotypes are typically multi-allelic*
- Advantages of SNPs
  - *Fundamental unit of inheritance*
  - *Codominant*
  - *Simplicity allows high-throughput detection*

# SNPs result from single-base changes



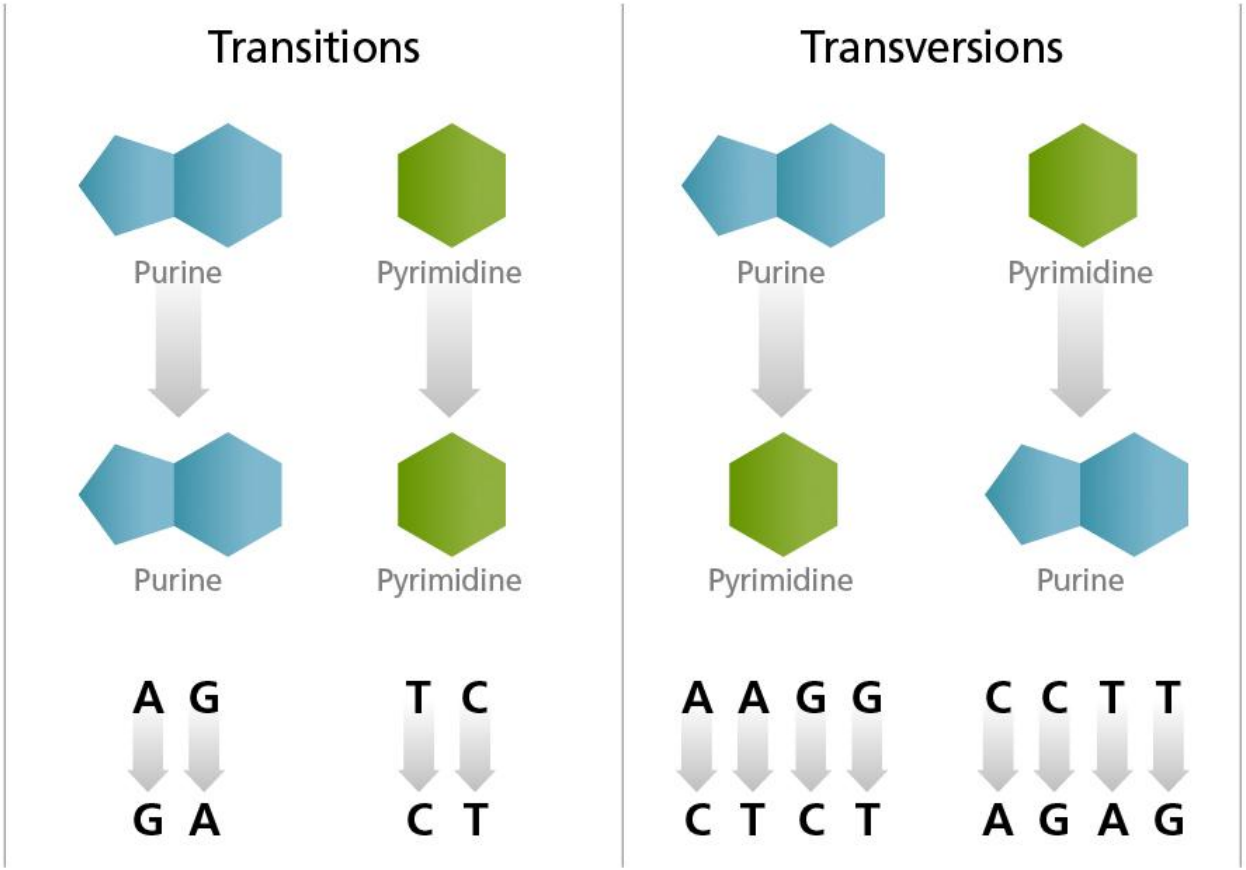




Figure Credit: Modified from Pierce, 2010

# SNPs occur throughout the genome

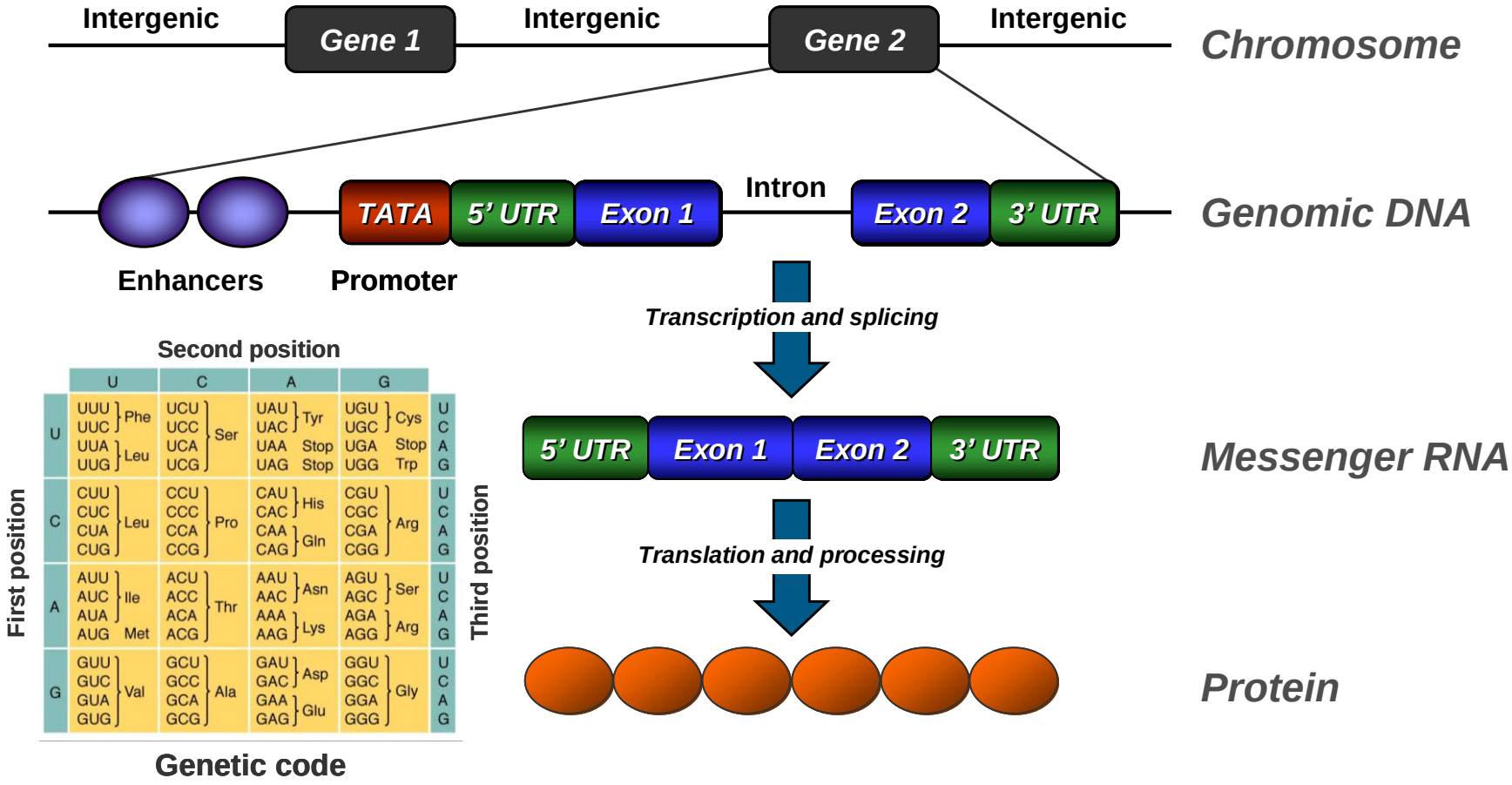
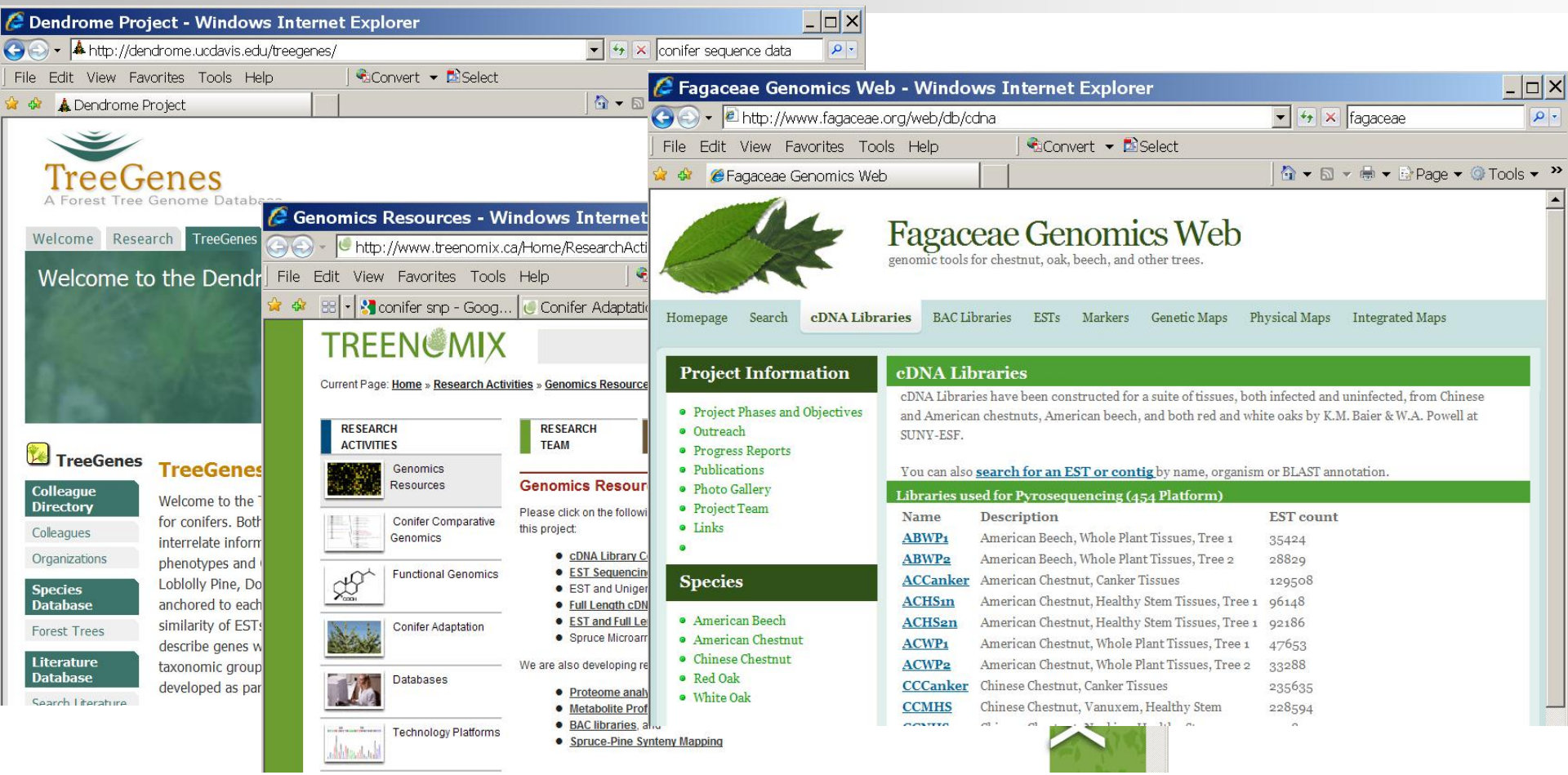



Figure Credit: Glenn Howe, Oregon State University

# Finding, verifying, and using SNPs

- SNP discovery requires genetic sequence of the same chromosomal region from two or more chromosomes
- Sequence can be obtained
  - *From existing sources (online)*
  - *By re-sequencing*
  - *By obtaining new DNA sequence*
- Sequence processing requires several steps. Doing this for many (100's to 1000's) SNPs requires specialized bioinformatic tools
- SNP genotyping: a wide array of genotyping platforms are now available

# Finding existing SNPs




Also, the National Center for Biotechnology Information, the ultimate clearing house for genomic information: www.ncbi.nlm.nih.gov

# Targeted SNP discovery via re-sequencing

- Obtain DNA sequence from the region or gene of interest
- Design overlapping PCR amplicons across the entire gene
  - *PCR amplicon = gene fragment synthesized using PCR*
- PCR-amplify the gene fragments from diverse individuals
- Use various computer programs to find and evaluate SNPs

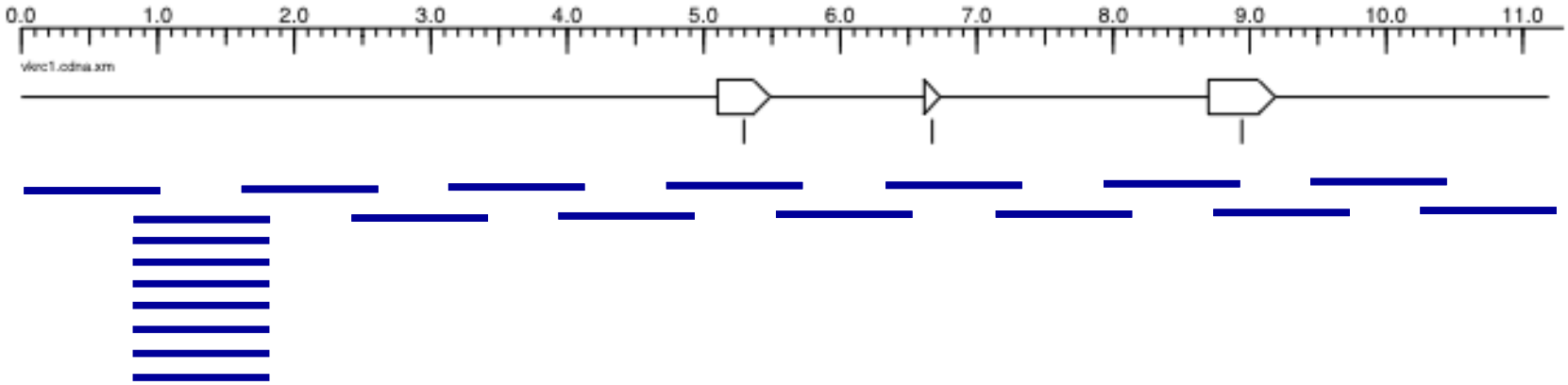
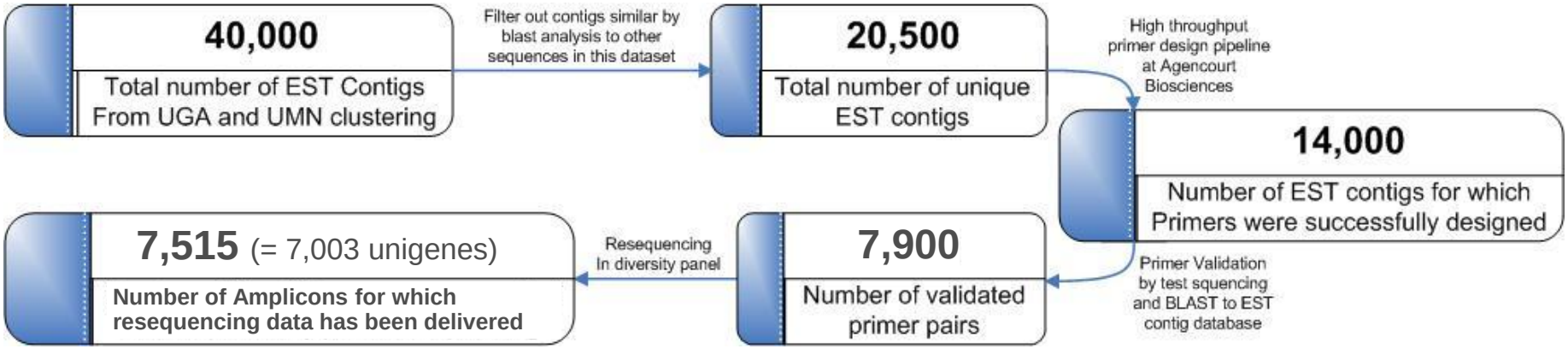



Figure Credit: Jill Wegrzyn, University of California, Davis

# SNP discovery: Loblolly pine example




- >3 million bases of aligned sequence
- ~40,000 SNPs detected in ***7,003 unigenes***
- ~6 SNPs per gene fragment (amplicon)
- Amplicon size averages 450 bp
- 1 SNP per 75 bp on average

Figure Credit: Jennifer Lee, University of California, Davis

# Sample size is important for finding SNPs

Minimum = 2 chromosomes

```
GTTACGCCAATACAGGATCCAGGAGATTACC
GTTACGCCAATACAGCATCCAGGAGATTACC
```

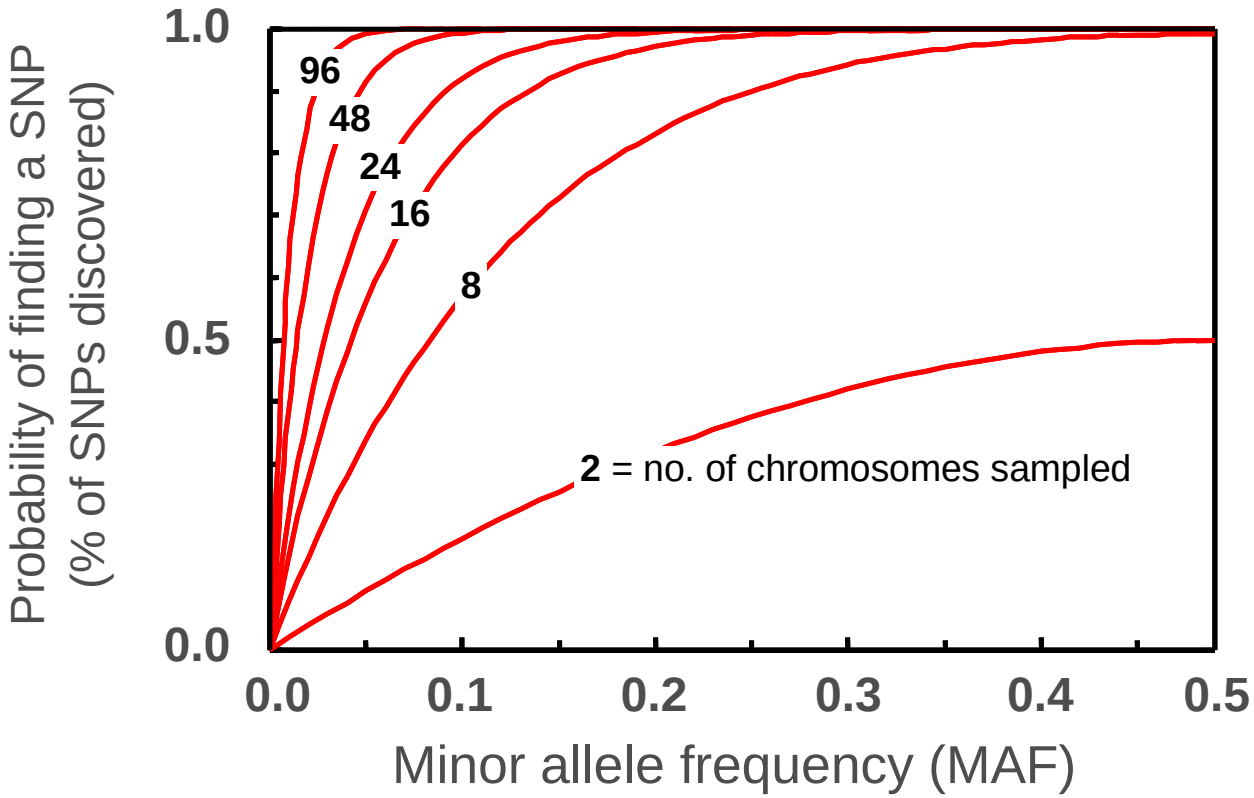
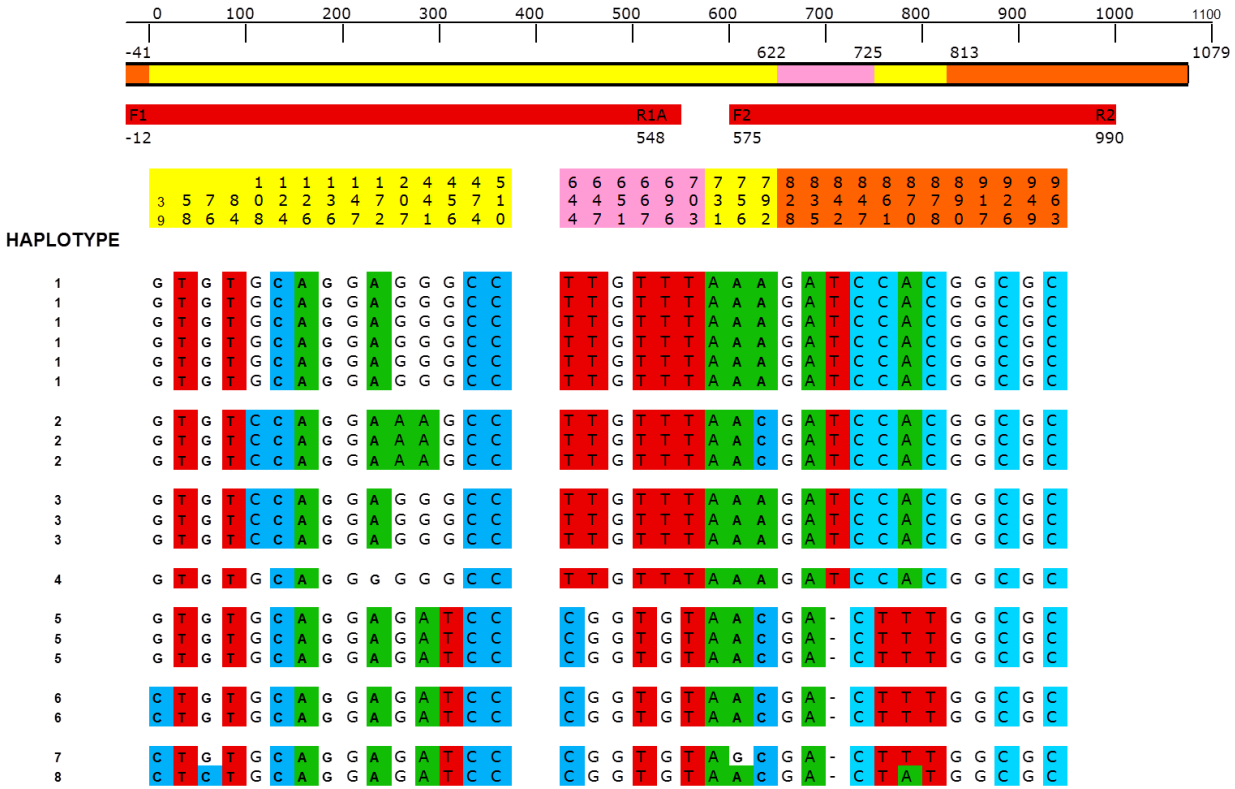



Figure Credit: Modified from Glenn Howe, Oregon State University

# SNP detection in the pine AGP6 gene

## Direct detection of haplotypes using haploid pine tissues




***AGP6 gene***


*Pine seed with haploid megagametophyte*

Figure Credits: Modified from White et al 2007.

# SNP detection using de novo sequencing

When no prior knowledge of sequence is available to guide SNP discovery, new sequence may be obtained.

Some things to think about

- *What DNA source to use?*
- *What population will be represented by the sequence?*
- *What method should be used for acquiring sequence?*

# SNP discovery

**BAC** = bacterial artificial chromosome
**Fosmid** = Bacterial F - plasmid
**RRG** = reduced representation genomic
**Shotgun** = many small DNA sequences
**cDNA** = complementary DNA made from mRNA

## Which DNA sequence?

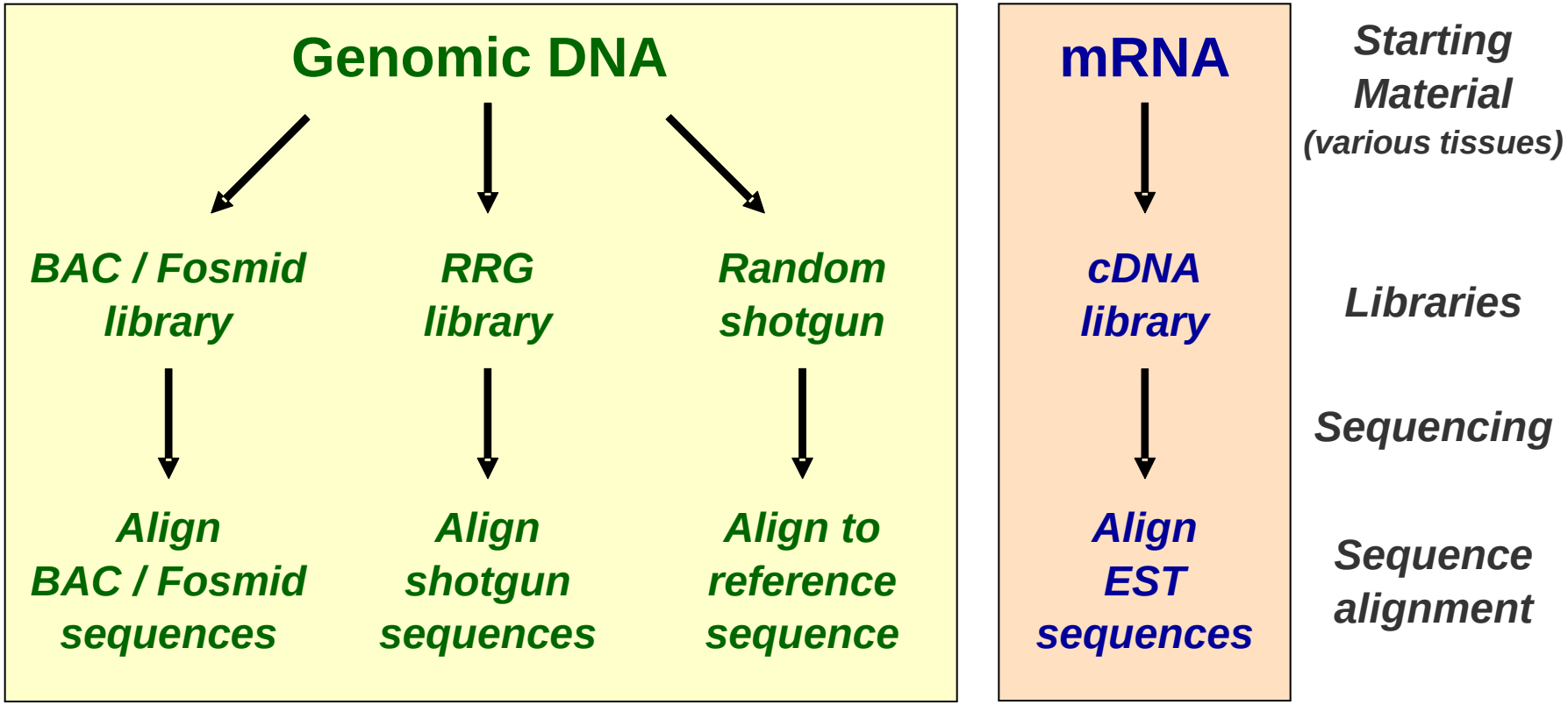
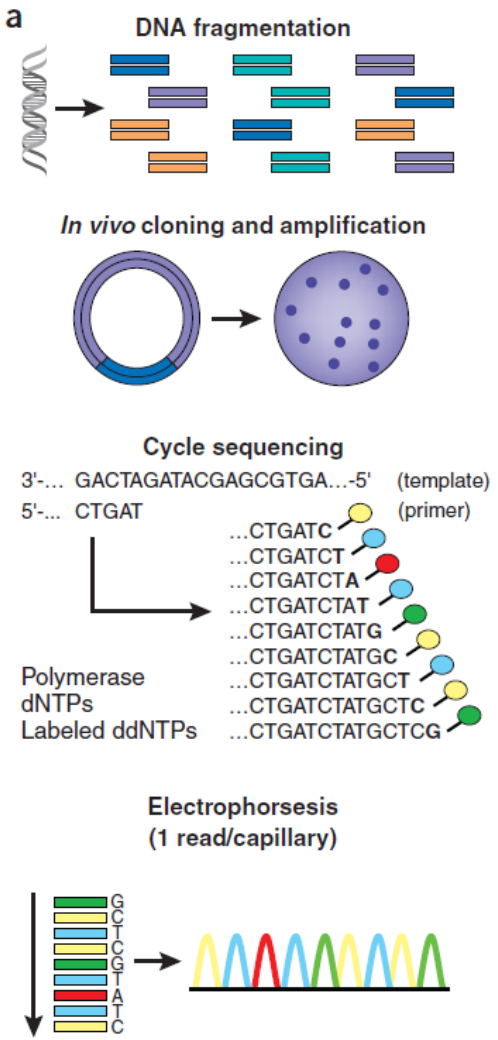



Figure Credit: David Harry, Oregon State University

# Sanger sequencing




- DNA is fragmented
- Cloned to a plasmid vector
- Cyclic sequencing reaction
- Separation by electrophoresis
- Readout with fluorescent tags

Figure Credit: Jill Wegrzyn, University of California, Davis

# Next Generation Sequencing Technologies: Current!

| | Roche (454) | Illumina (Solexa) | SOLiD |
|---|---|---|---|
| Chemistry | Pyrosequencing | Polymerase-based (CRT) | Ligation-based |
| Amplification | Emulsion PCR | Bridge Amp | Emulsion PCR |
| Paired ends/sep | Yes/3kb | Yes/200 bp | Yes/3 kb |
| Mb/run | 600 Mb | 1300 Mb | 2000 Mb |
| Time/run | 7 h | 4 days | 5 days |
| Read length | 300-800 bp | 32-150 bp | 50 bp |
| Cost per run (total) | $8439 | $8950 | $17447 |
| Cost per Mb | $84.39 | $5.97 | $5.81 |

**Released recently: Illumina HiSeq 1000: will deliver 100+ Gb of data/run using paired 100 bp reads, enabling the sequencing of a complete human genome in a single run**

# Cost of sequencing a base dropping faster than storing a byte of data

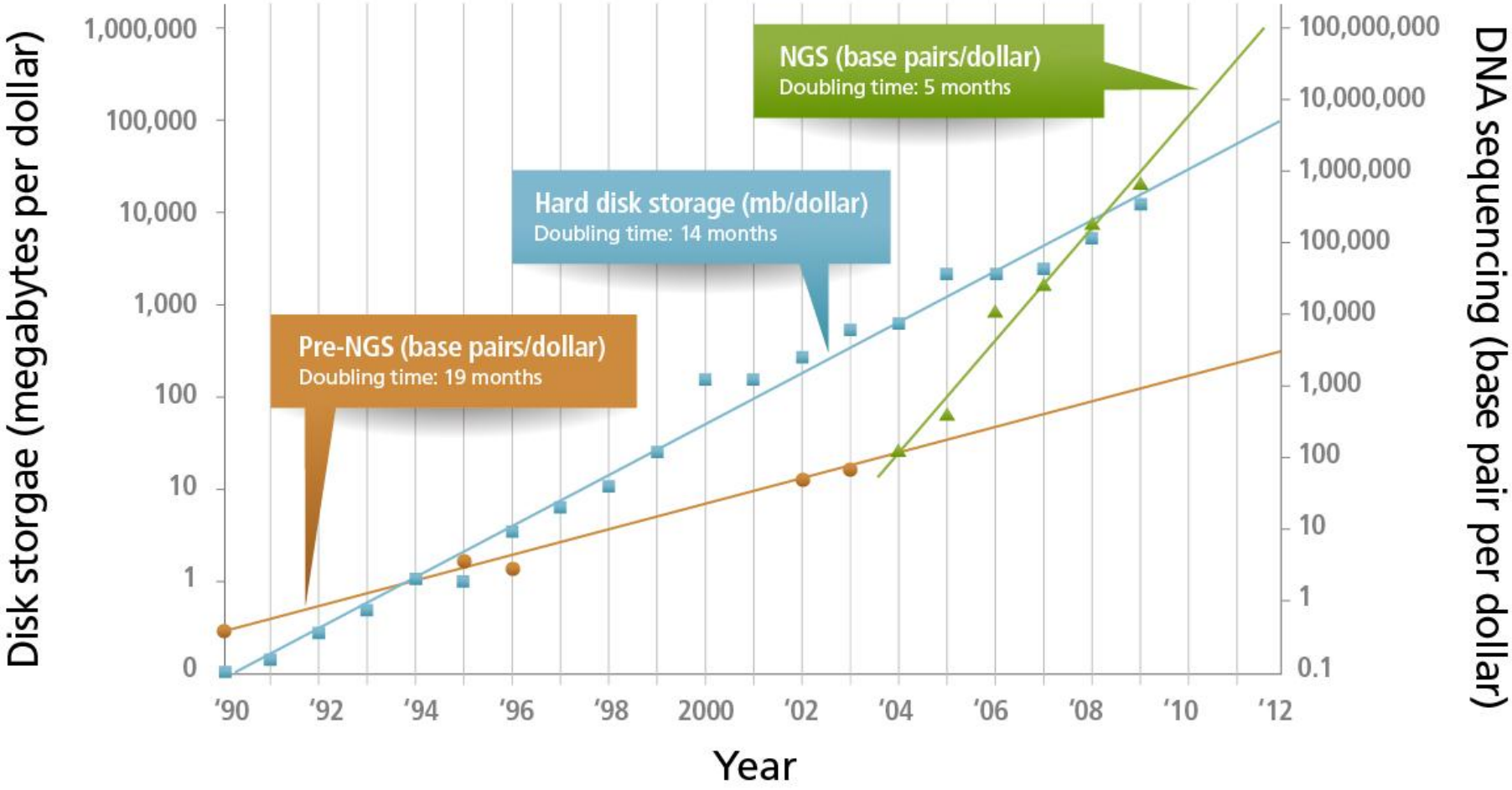



Figure Credit: Modified from Baker, 2010

# A SNP discovery example – Chinese chestnut

- In 2008 a series of cDNA libraries based on RNA collected from different tissues of three plant sources were submitted to sequencing on the 454 FLX platform.
- Cost: ~$20K  Time: ~2-3 weeks for acquisition, 4 weeks for data processing

Yield: 838,472 Reads

171.9 Million bp of sequence

40,039 contigs

2 to 20+ copies of a given sequence

**Detected: > 15,000 potential SNPS, from expressed genes**

# Sequence processing

## Like taking a sip of water from a fire hydrant

- Sequence is now abundant and inexpensive (relatively)
- Making sense of sequence requires time, computer resources, skilled personnel, and a collection of bioinformatic tools
- Chief Concerns
  - *How reliable is the quality of your sequence?*
  - *Can you assemble and align sequences representing alleles?*
  - *Can you reliably identify polymorphisms as candidate SNPs?*

# Sequence assembly

- The **sequence assembly problem** is essentially one of constructing a DNA sequence superstring that explains the observed set of sequence reads.
- This superstring might be…any type of DNA that was subject to sequencing
- If the data were completely error-free, then we would expect every sequence read to be contained within the superstring

# Sequence analysis software

## Traditional DNA sequence analysis programs for 1st generation sequencing technology

- **Phred** - Calls bases and assigns quality scores
- **Phrap** - Forms contigs (unigenes) and assigns new quality scores
- **Consed** - Allows users to view and edit alignments (contigs), and view trace files (Visual X-Windows graphic interface)
- **Polyphred** - Finds SNPs in phrap contigs
- **Polybayes** - Calculates probability that putative SNPs are true SNPs and not sequencing errors

# Sequence analysis for NGS (next generation sequencing)

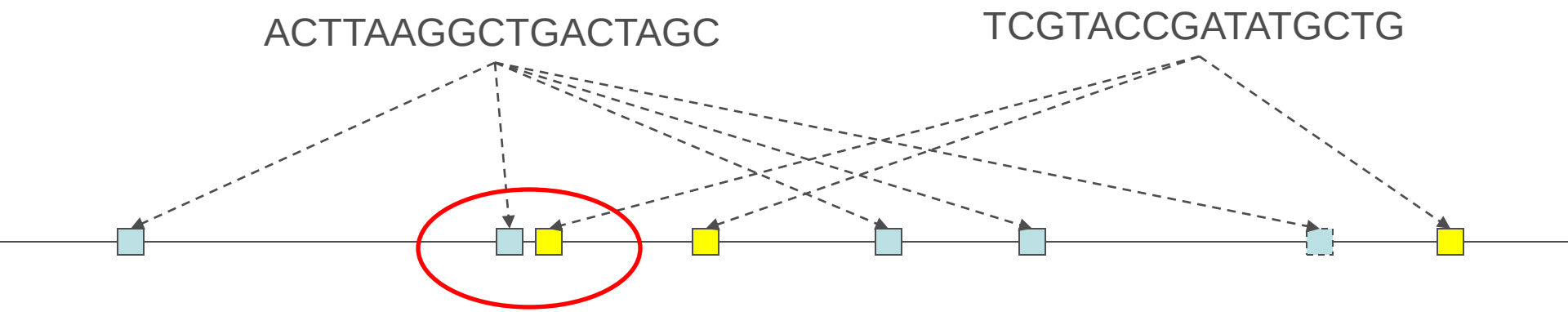
- NGS use a variety of strategies that rely on a combination of:
  - *template preparation*
  - *sequencing/imaging*
  - *genome alignment (if available)*
  - *assembly methods*

- The short sequence read lengths (32 – 150 bp) for many of the NGS present challenges for alignment, assembly, and SNP detection. All functions are improved by having a reference sequence.

- (Assembly): One review site lists 27 programs, most developed over the last 3 years (See http://en.wikipedia.org/wiki/Sequence_assembly)

- (SNP Detection): Equally abundant (currently favored for Illumina sequence –GATK and GigaBayes)

# Sequence analysis for NGS (next generation sequencing)




- Short reads are problematic, because short sequences do not map uniquely to the genome/transcriptome
- Solution #1: Get longer reads
  *i.e. Pair with longer read technologies such as 454*
- Solution #2: Get paired reads (sequenced from both ends)

Shendure and Ji, 2008

# SNP validation and genotyping

- Validation: Do SNPs segregate in a Mendelian fashion
  - *Re-sequence array of progeny (haploid seed tissue*
- Selecting a genotyping platform
  - *Factors to consider*
    - *Cost per genotype*
    - *Number of individuals to genotype*
    - *Number of SNPs*
- Selecting population
  - *Preliminary survey*
  - *Full-fledged study*

# In summary

- Markers have played a pivotal role in the development of genetics
- Marker development has proliferated in recent decades, driven by technology improvements
- Relative abundance of markers has brought about increased utility
- High-throughput sequencing has delivered SNPs, a class of abundant and fundamental genetic markers
- Bioinformatic tools and genotyping vendors play key roles
- SNP markers make rigorous association genetic  and genomic selection applications feasible

# References cited


- Baker, M. 2010. Next-generation sequencing: adjusting to data overload. Nature Methods 7: 495–499. (Available online at: http://dx.doi.org/10.1038/nmeth0710-495) (verified 31 May 2011).

- Pierce, B. A. 2010. Genetics Essentials: Concepts and Connections 1st Edition. W. Freeman and Co., New York.

- Shendure, J., and H. Ji. 2008. Next-generation DNA sequencing. Nature Biotechnology 26: 1135-1145. (Available online at: http://dx.doi.org/10.1038/nbt1486) (verified 23 March 2011).

- White, T. L, W. T. Adams, and D. B. Neale. 2007. Forest genetics. CAB International, Wallingford, United Kingdom. (Available online at: http://bookshop.cabi.org/?page=2633&pid=2043&site=191) (verified 27 Apr 2011).

# External Links

- National Center for Biotechnology Information [Online]. U.S. National Library for Medicine, National Institutes of Health. Available at: http://www.ncbi.nlm.nih.gov (verified 23 March 2011).

# Additional Resources

| | |
|---|---|
| Dendrome | Neale, D., J. Wegrzyn, B. Figueroa, and J. Yu. Tree genes: A forest tree genome database [Online]. University of California at Davis. Available at: http://dendrome.ucdavis.edu/treegenes (verified 31 May, 2011). |
| Treenomix | Treenomix [Online]. University of British Columbia. Available at: http://www.treenomix.ca (verified 31 May, 2011). |
| Chestnut | Fagaceae genomics web: Genomic tools for chestnut, oak, beech, and other trees [Online]. Available at: http://fagaceae.org (verified 31 May, 2011). |

# Thank You.

Conifer Translational Genomics Network
Coordinated Agricultural Project

CTGN CAP

UCDAVIS

USDA

United States
Department of
Agriculture

National Institute
of Food and
Agriculture