DNA sequencing and conservation genomics

Konstantin (Kostya) Krutovsky

Department of Forest Genetics and Forest Tree Breeding
Georg-August-University of Göttingen, Germany

Department of Ecosystem Science & Management
Texas A&M University, College Station, USA

Genome Research and Education Center
Siberian Federal University, Krasnoyarsk, Russia

N.I. Vavilov Institute of General Genetics
Russian Academy of Sciences, Moscow, Russia

ProCoGen Workshop "Genomics and the conservation of conifer genetic resources" Sept. 3rd 2014, Kámoni Arboretum Szombathely, Hungary

Specifics of forest conservation

Major objectives:

- sustainable reproduction
- maintaining of genetic variation and adaptive potential

Different levels or targets:

- population, species, ecosystem (forest trees are often keystone species)

Depends on:

a) species mating system (insect- vs. wind-pollinated, etc.)

b) population system (isolated, fragmented, connected, disconnected, continuous, marginal, etc.)

c) forest type (tropical, temperate, boreal)

Specifics of forest conservation

A revised concept of forest management and conservation different from other species and ecosystems is needed because of:

- conflict between practical vs. conservation objectives (growing demand for timber and wood products)
- conflict between tree improvement and breeding vs. native protection (domesticated animals & plants are not well-adapted to the wild environment due to the narrow selection that target and improve some traits, but can simultaneously lose or worsen others)
- conflict between evolving, dynamic nature of forest ecosystems and desire of environmentalists to keep it "as it is"

A revised concept should be based on dividing forests into different management units based on the objectives and reproduction mode:

a) naturally reproducing forest with natural succession (no breeding; no reforestation)

b) semi-managed forests (reforestation and ecological restoration using mostly local or the most adapted seed sources, and sometimes genetically improved seeds)

c) seed orchards and forest plantations based on genetically improved trees resulted from intense breeding and domestication

What is “conservation genomics”?

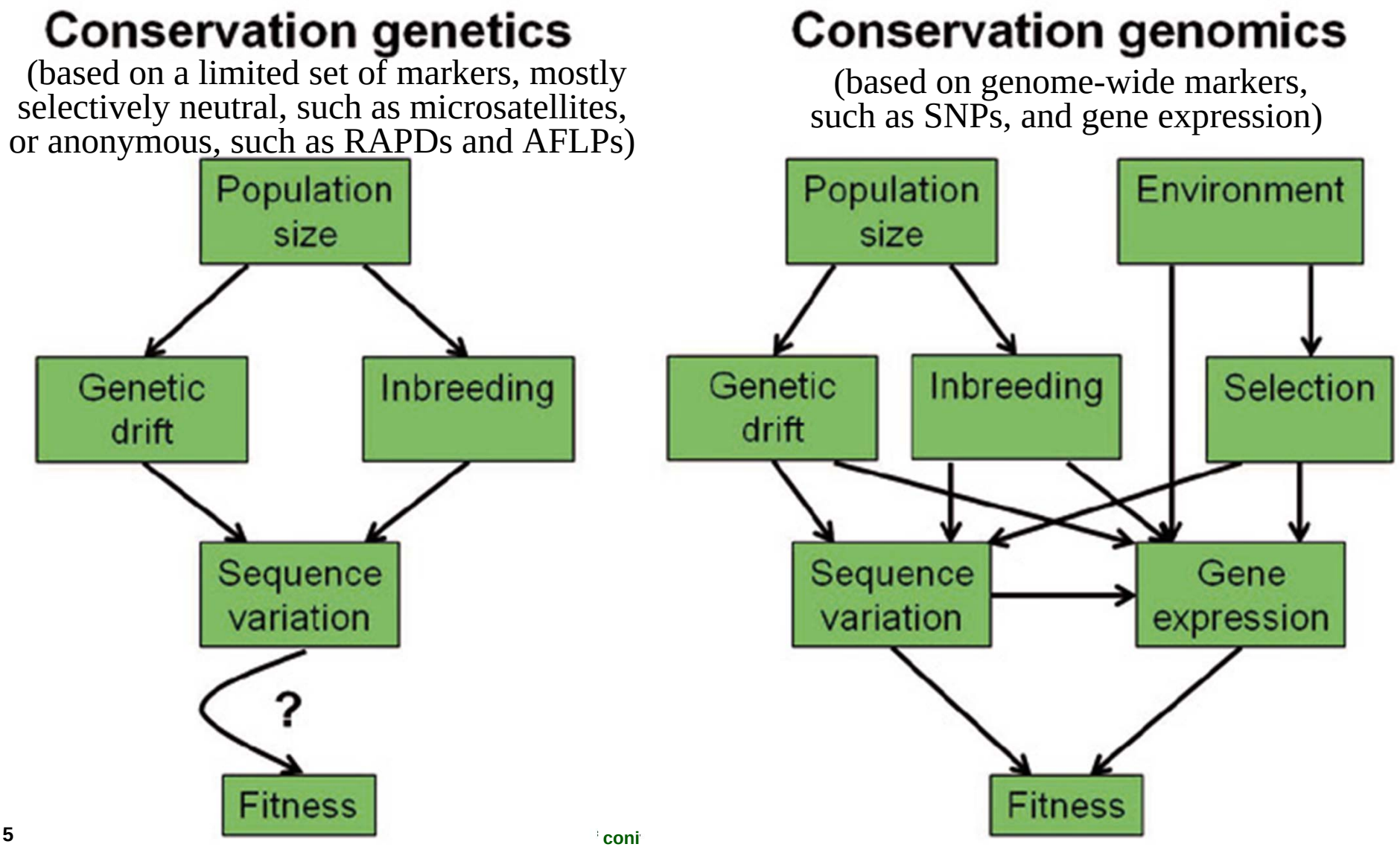
a new multidisciplinary research field that **integrates conservation genetics with ecological and evolutionary genomics** with the same objectives as in conservation biology to **conserve biodiversity, including genetic variation within populations and species**, but **it uses the latest genomic and DNA sequencing technologies** to:

- identify functionally important genomic variation;
- apply genome-wide markers to reliably estimate demographic, mating system and population genetic parameters in a conservation context;
- apply epigenomic and gene-expression tools to study the mechanisms behind important conservation genetic processes, like adaptation and phenotypic plasticity;
- apply metagenomic approaches to step from population level approaches up to species and community level assessments.

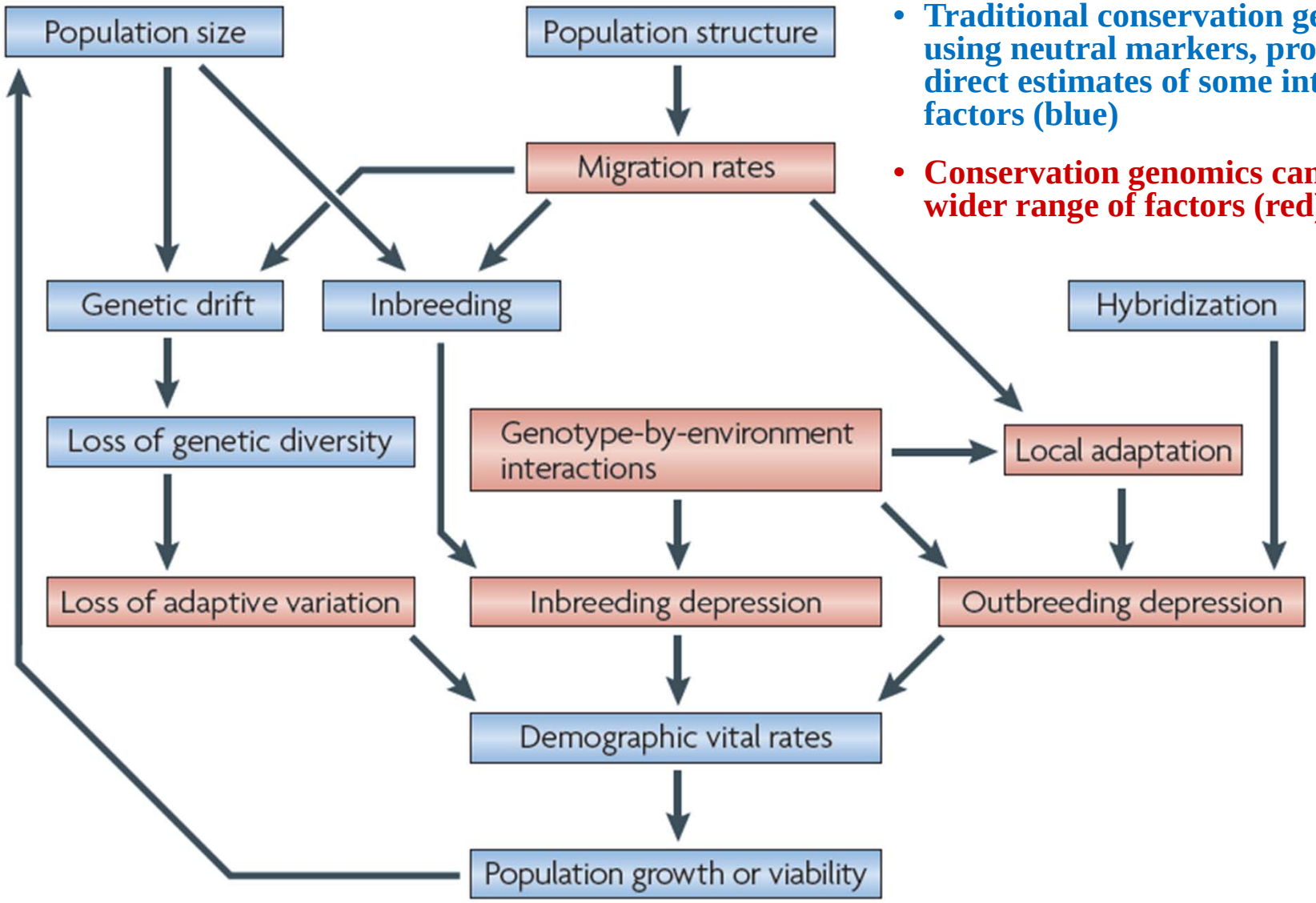
The field has been designated by the former ESF-CONGEN networking program as the **future of conservation genetics**.

Conservation genomics vs. Conservation genetics

The conceptual difference between a conservation genetics and conservation genomics research approach (modified Fig. 2 in Ouborg et al. 2010, Trends in Genetics 26: 177–187):



Main interacting factors in conservation of natural populations



- **Traditional conservation genetics, using neutral markers, provides direct estimates of some interacting factors (blue)**
- **Conservation genomics can address a wider range of factors (red)**

Fig. 1 in Allendorf et al. (2010) *Nature Reviews Genetics* 11: 697-709

Conservation genomics paradigm

integrates two views (paradigms) on threats to biodiversity:

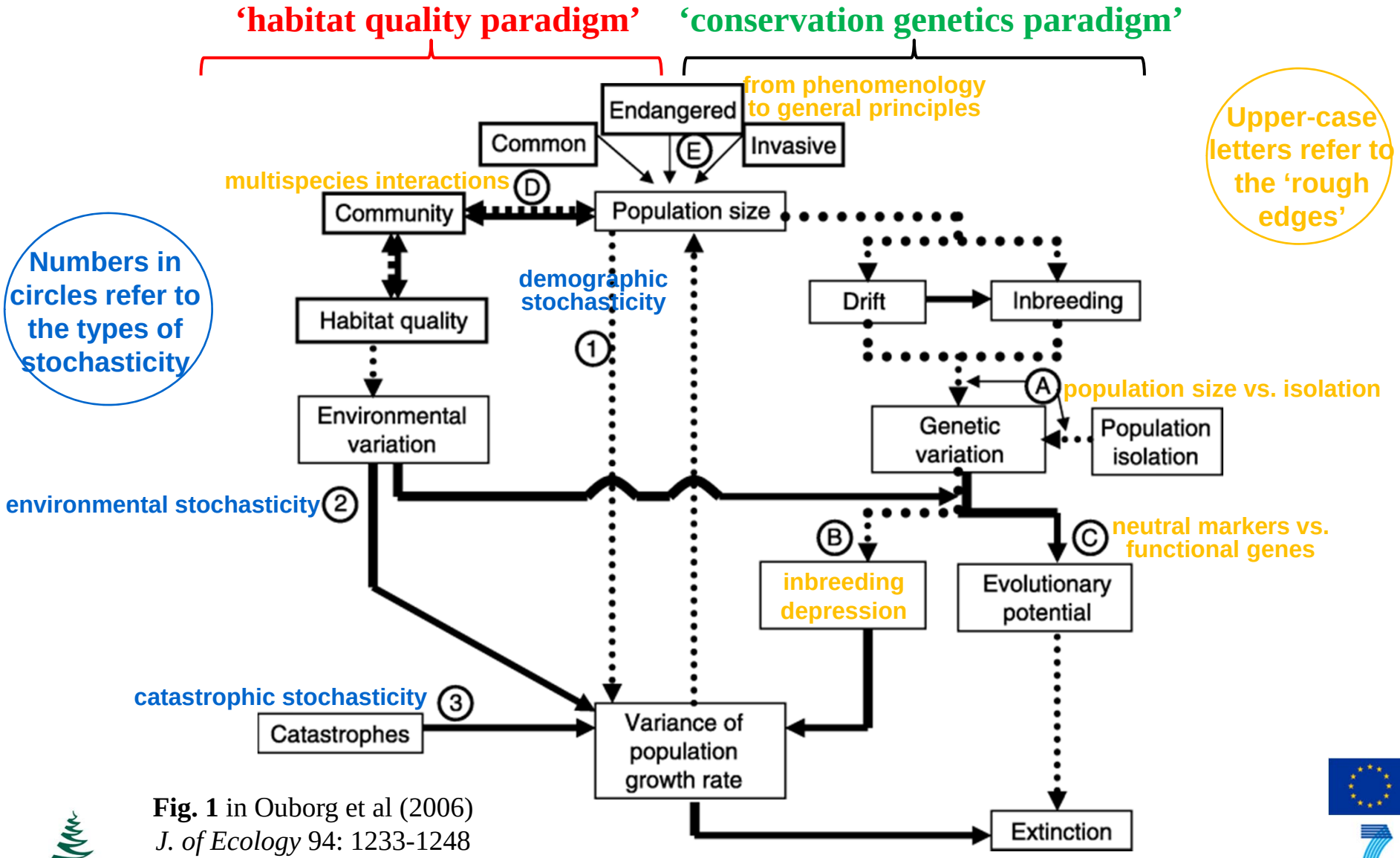


Fig. 1 in Ouborg et al (2006)
J. of Ecology 94: 1233-1248

Primary genetic problems in conservation and how genomics can contribute to their solution

Primary problem	Possible genomic solution
Estimation of N_e, m and s	Increasing the number of markers, reconstructing pedigrees and using haplotype information will provide greater power to estimate and monitor N_e and m, as well as to identify migrants, estimate the direction of migration and estimate s for individual loci within a population
Reducing the amount of admixture in hybrid populations	Genome scanning of many markers will help to identify individuals with greater amounts of admixture so that they can be removed from the breeding pool
Identification of units of conservation: species, evolutionarily significant units and management units	The incorporation of adaptive genes and gene expression will augment our understanding of conservation units based on neutral genes. The use of individual-based landscape genetics will help to identify boundaries between conservation units more precisely
Minimizing adaptation to captivity	Numerous markers throughout the genome could be monitored to detect whether populations are becoming adapted to captivity
Predicting harmful effects of inbreeding depression	Understanding the genetic basis of inbreeding depression will facilitate the prediction of the effectiveness of purging. Genotyping of individuals at loci associated with inbreeding depression will allow the selection of individuals as founders or mates in captive populations. Pedigree reconstruction will allow more powerful tests of inbreeding depression
Predicting the intensity of outbreeding depression	Understanding the divergence of populations at adaptive genes will help to predict effects on fitness when these genes are combined. Detecting chromosomal rearrangements will help to predict outbreeding depression
Predicting the viability of local populations	Incorporating genotypes that affect vital rates and the genetic architecture of inbreeding depression will improve population viability models
Predicting the ability of populations to adapt to climate change and other anthropogenic challenges	Understanding adaptive genetic variation will help to predict the response to a rapidly changing environment or to harvesting by humans and allow the selection of individuals for assisted migration

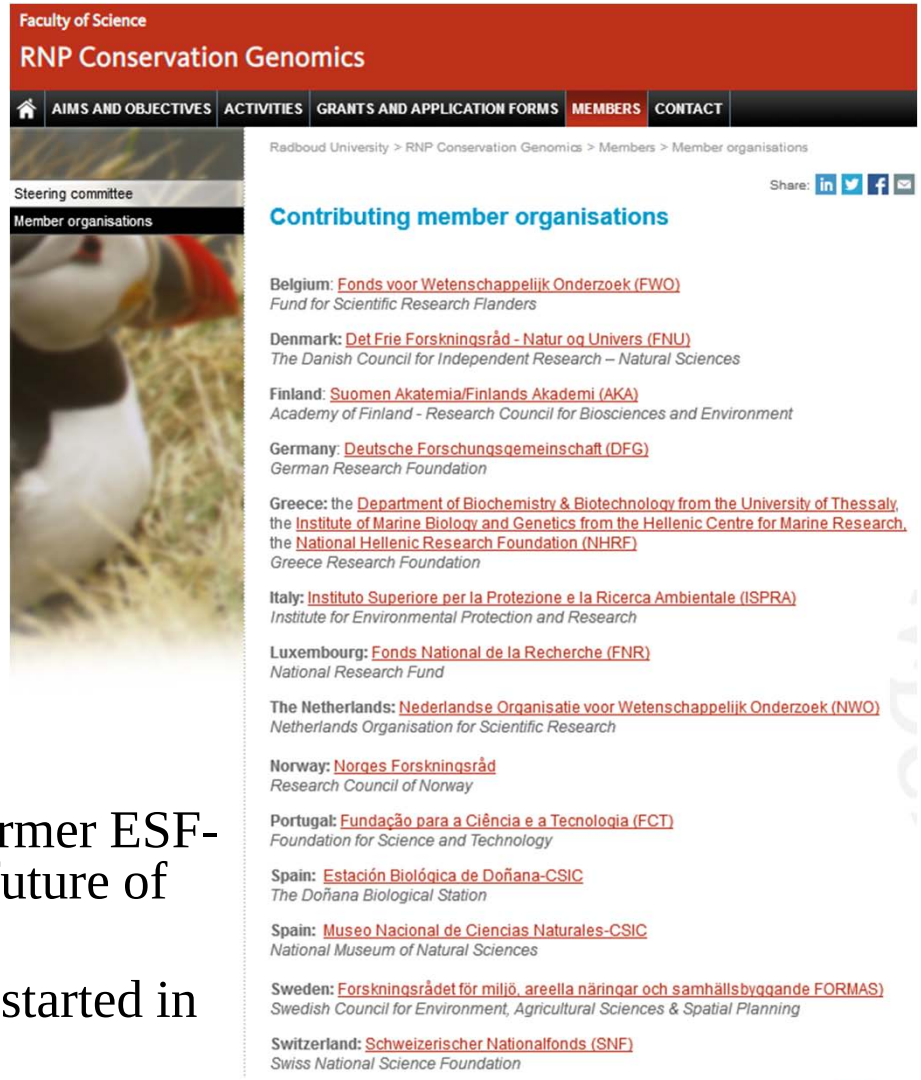
*These problems are listed from top to bottom in sequence of those that can be immediately addressed to those that will become more feasible to address in the future. m, migration rate; N_e, effective population size; s, selection coefficient.

Table 1 in Allendorf et al. (2010) *Nature Reviews Genetics* 11: 697-709

Conservation genomics' challenges

There are several challenges that have to be overcome, including:

- the transfer of genomic tools to non-model, threatened species;
- the transfer of knowledge from genomic oriented labs to conservation oriented labs;
- the equal sharing of genomic resources and knowledge between European labs;
- the design of multidisciplinary approaches to data management and analyses of the vast amounts of genomic data on threatened species likely to become available in the near future.



The field has been designated by the former ESF-CONGEN networking program as the future of conservation genetics.

The 5 year CONGENOMICS program started in 2011 http://www.ru.nl/congenomics/

Forest conservation genomics objectives

- protect forest species and their habitats from the negative environmental effects due to human activity (traditional) and global climate change (new);
- increased focus on improving *predictions* about what changes are expected to occur in the future, what biodiversity needs to be preserved to maximize the chances of species to adapt to these changes, the rate at which these adaptations can occur, and understanding how species are expected to respond to these changes
- identification of functionally important genomic variation;
- estimate demographic, mating system and population genetic parameters in a conservation context;
- understand the mechanisms behind important conservation genetic processes, like adaptation, phenotypic plasticity, inbreeding depression, etc.;
- asses metacommunity composition and its contribution into ecosystem functions and services;
- build tools of communication between researchers and end users in the field.

Advantages of next generation sequencing (NGS) techniques

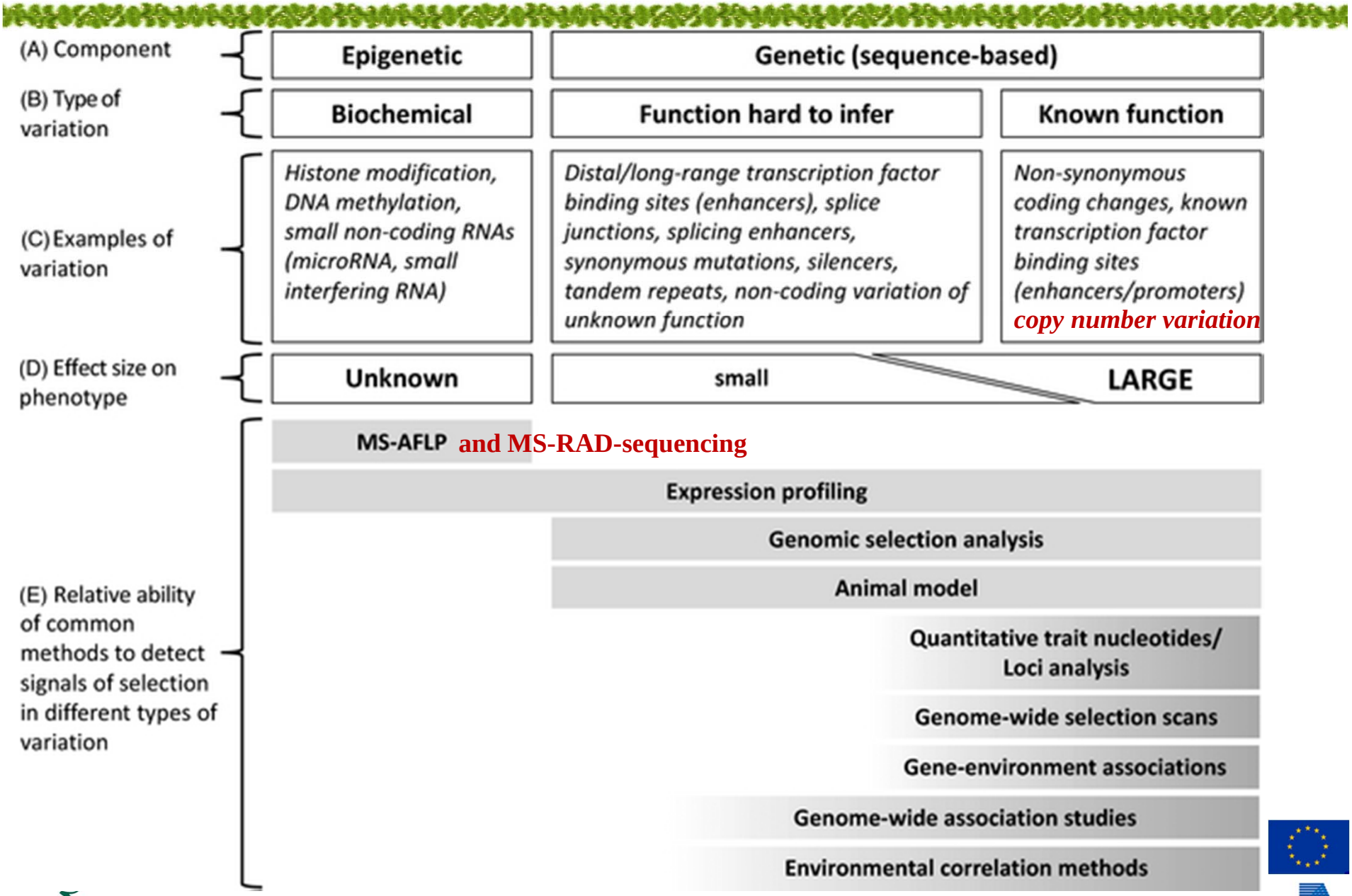
- deeper inferences on demographic history;
- higher resolution inferences of population structure;
- subtracting population structure into adaptive component caused by natural selection and selectively neutral component due to genetic drift and isolation;
- opportunities to identify genomic regions under selection;
- greater coverage and more representative estimates of genetic variation;
- the experimental study of the genomic mechanisms behind biological processes important for conservation, such as, for instance, overdominance, inbreeding and outbreeding depression, genotype-by-environment interactions, the genomic signature and mechanisms of local adaptation, epigenetic mechanisms of adaptation and evolution, dynamics of functional gene variation in small populations likely affected by genetic drift;
- applications, such as genomic selection, for tree improvement programs and breeding trees for the managed forests and forest plantations

Forest conservation metagenomics

Soil metacommunities (fungi and microorganisms) greatly affect forests. So, to integrate species and community level genomic assessments in forest conservation genomics approaches it is important to carry out studies that:

- incorporate metagenomic monitoring in conservation genetics assessments,
- aim to disentangle the relationship between genetic and genomic variation at population level, biodiversity and ecosystem functioning,
- target the influence of species interactions on genetic variation and functional genomic activity,
- aim to disentangle the genomic interactions between pathogens, parasites and herbivores on one side and hosts on the other.

Genome basis of evolutionary potential



Modified **Fig.** 1 from Harrisson *et al.* (2014) Evolutionary Applications. doi: 10.1111/eva.12149

Transcriptome (RNA-seq) in conservation genomics

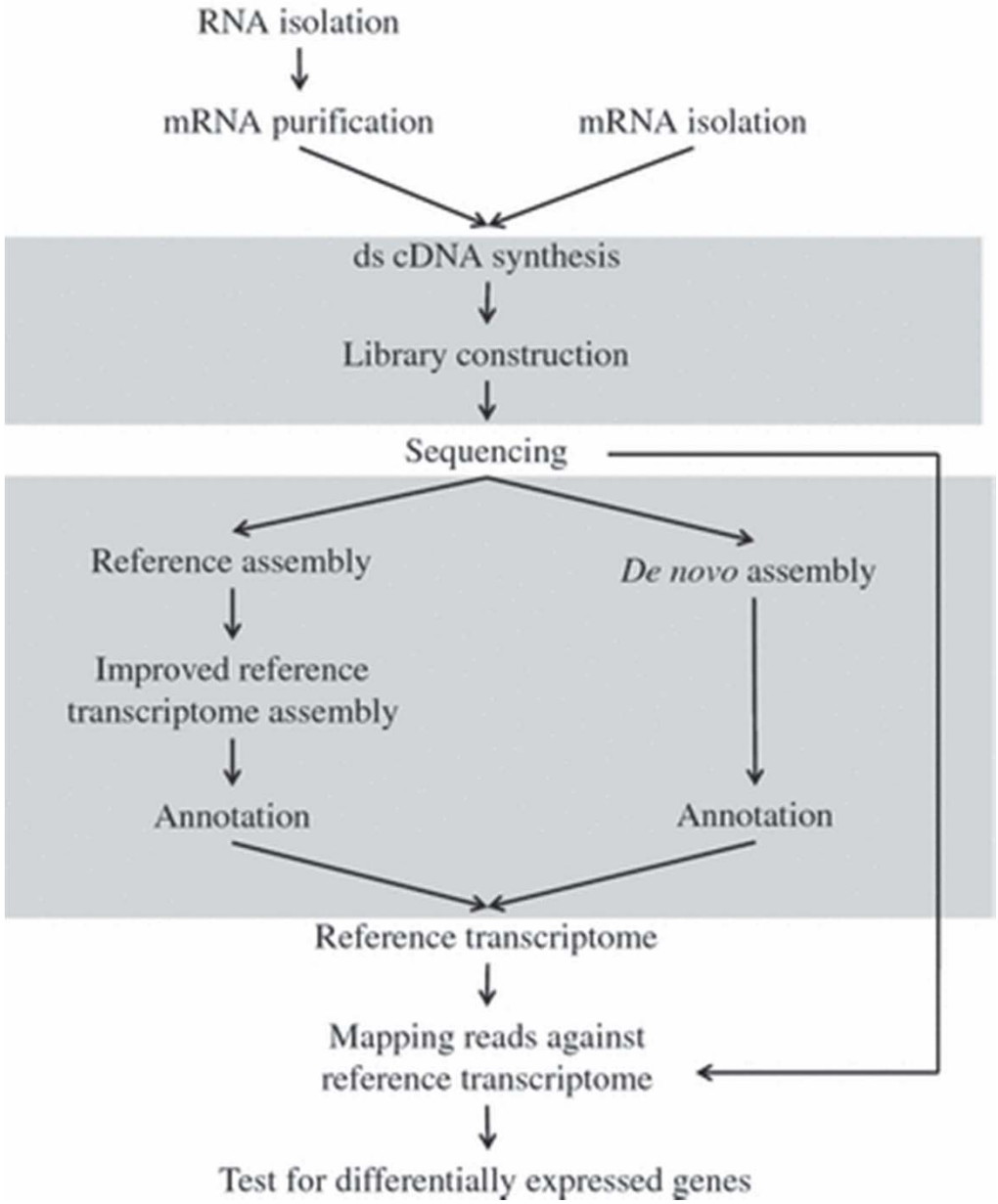
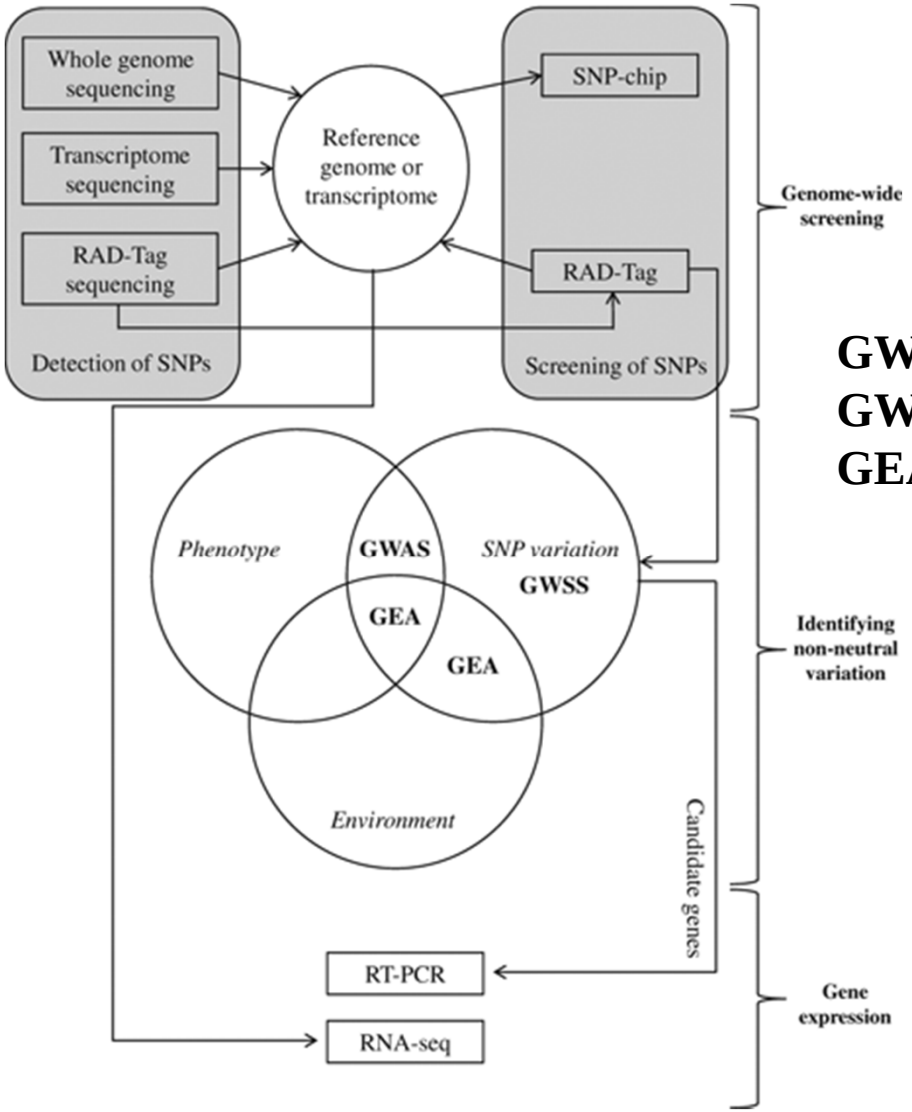


Fig. 2 in Angeloni et al. (2012) Genomic toolboxes for conservation biologists. *Evolutionary Applications* 5: 130–143.

NGS approaches in conservation genomics

A scheme of how various next-generation sequencing approaches relate to the three main categories of questions in conservation genomics and how to feed their results into each other



GWAS: genome-wide association studies
GWSS: genome-wide selection scans
GEA: gene–environment association studies

Fig. 3 in Angeloni et al. (2012) Genomic toolboxes for conservation biologists. *Evolutionary Applications* 5: 130–143.

Workflow of a typical *de novo* whole-genome sequencing project

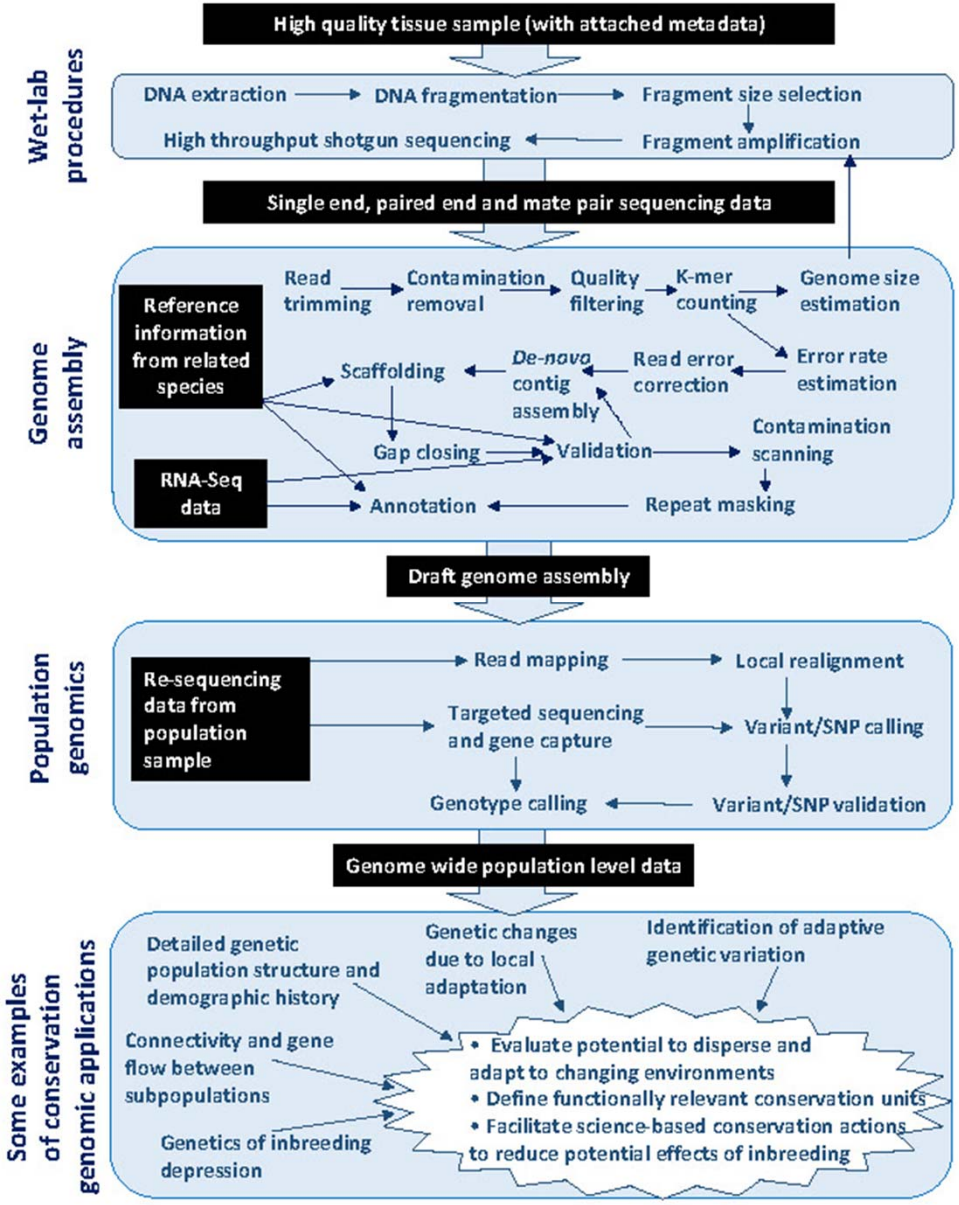


Fig. 1 in Ekblom R & Wolf JBW (2014) A field guide to whole-genome sequencing, assembly and annotation. *Evolutionary Applications* 24 June 2014 DOI: 10.1111/eva.12178

Simplified illustration of the assembly process and terminology

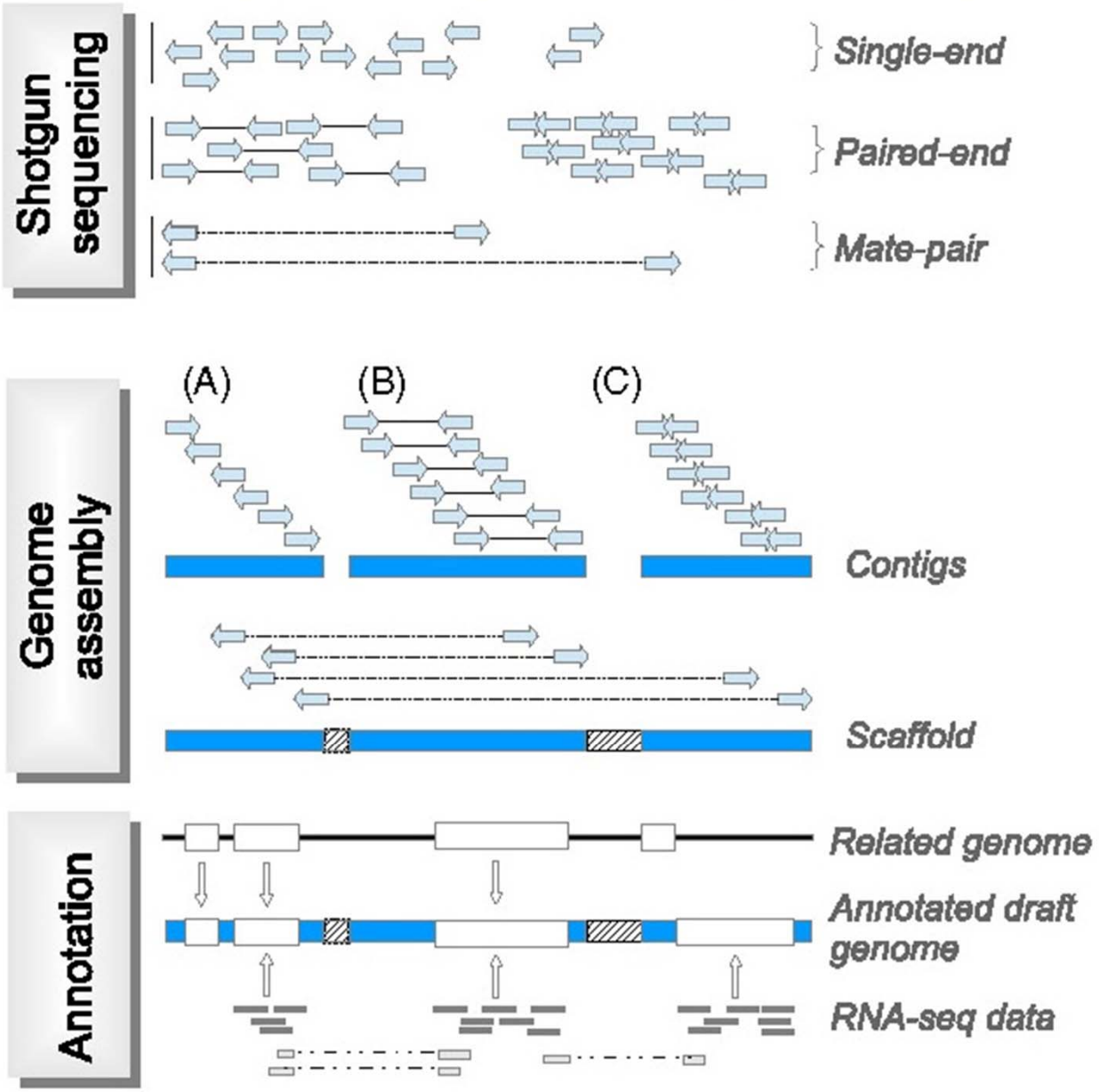


Fig. 2 in Ekblom R & Wolf JBW (2014) A field guide to whole-genome sequencing, assembly and annotation. *Evolutionary Applications* 24 June 2014 DOI: 10.1111/eva.12178

Genomic markers development and high-throughput genotyping using NGS

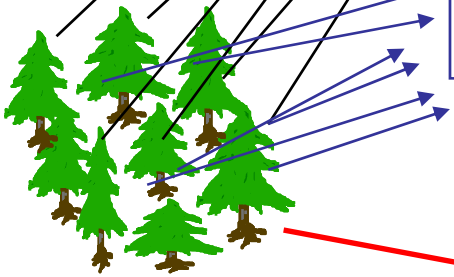
bar-coded DNA or mRNA library pools

individual genomic or mRNA (tissue-specific or total)

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

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

individual genomic DNA



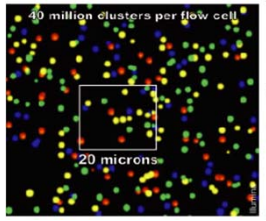
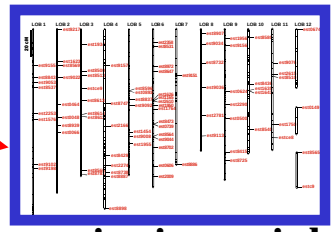
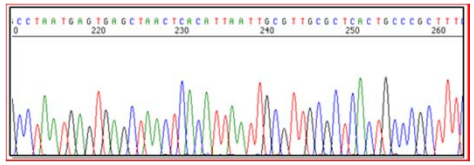


image analysis



genotyping and phenotyping in mapping or natural population

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



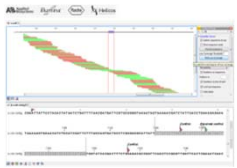
DNA chromatograms

genotyping by sequencing (GBS)

high-throughput SNP genotyping

high-density SNP marker development

sequence processing and analysis



Examples and descriptions of genetic and genomic approaches commonly used in conservation genetics

Approach	Description	Reference
Mapping genes associated with traits		
Quantitative trait nucleotides/loci programs	Use experimental crosses to look for physical location of regions of genome underlying complex phenotypic traits.	(Barton and Keightley 2002)
Identifying loci putatively under selection		
Genome-wide selection scans (GWSS)	Look for regions of the genome where genetic variation between populations differs relative to the genome-wide average (e.g. F_{ST}-outliers)	(Oleksyk et al. 2010)
Associating genetic variation with selective pressures		
Genome-wide association studies (GWAS)	Look for associations between genetic variants and particular phenotypic traits	(Stranger et al. 2011)
Genetic–environment associations (GEA)	Look for associations between candidate loci (e.g. outliers identified using GWSS) and environmental variables	(Bierne et al. 2011)
Environmental correlation methods	Look for correlations between allele frequencies and environmental variables. Some methods control for population structure.	(Joost et al. 2007; Coop et al. 2010; Eckert et al. 2010; Hancock et al. 2010b)
Directly identifying the genes involved in adaptation		
Expression profiling	Looks for differential expression of genes under different conditions	(Harrison et al. 2012; Smith et al. 2013)
Estimating additive genetic variance and genetic correlations and predicting phenotypes without knowledge of underlying genotypes		
Animal model	Employed in animal/plant breeding. Uses sparse or dense genome-wide markers to estimate additive genetic variance and genetic correlations and to predict breeding value for phenotypes without knowing particular loci underlying traits.	(Wilson et al. 2010)
Genome-selection	Employed in animal/plant breeding. Uses dense genome-wide markers to estimate additive genetic variance and genetic correlations and to predict breeding value for phenotypes without knowing particular loci underlying traits. Requires a reference population.	(Meuwissen et al. 2013)
Characterizing genome-wide methylation patterns		
Methylation-sensitive amplified fragment length polymorphism (MS-AFLP)	Detects variation in methylation at restriction sites (loci) using methylation-sensitive enzymes.	(Schrey et al. 2013)

Table 1 in Harrisson et al. (2014) Evolutionary Applications. doi: 10.1111/eva.12149

Unresolved questions and possible conservation genomic approaches for tackling them

To assess the impact of habitat fragmentation on selectively important variation

Possible conservation genomic approaches:

1. Use of genome-wide SNPs to obtain a representative estimate of genetic variation
2. Perform a genome scan to distinguish neutral from non-neutral markers
3. Comparison of patterns of neutral (microsatellite, AFLP) and non-neutral (as identified above) variation
4. Undertake an association-mapping approach to find correlations between markers and phenotypic traits important for adaptation
5. Candidate-gene studies can be used to search for frequency changes of alleles in relation to environmental change

To identify genetic mechanisms underlying inbreeding depression

Possible conservation genomic approaches

1. Population transcriptomics can help to identify genes associated with inbreeding depression, in different life-history stages and many genotypes
2. QTL mapping will help to identify genomic regions associated with inbreeding-depression phenotypes
3. Selection experiments on gene-expression phenotypes

To characterize the role of gene–environment (G x E) interactions

Possible conservation genomic approaches:

1. Population transcriptomics can be performed in combination with full factorial experiments to identify genetic, environmental and G x E effects in transcript profiles
2. Perform epigenetic screening, using methylation-sensitive AFLP or high-throughput bisulfite sequencing, of small and large populations in high and low quality habitats.

To identify the role of phenotypic plasticity in the response to environmental challenges

Possible conservation genomic approaches:

1. Epigenetic manipulation experiments (5-azacytidine) to manipulate methylation levels, and study phenotypic effects in relation to population size, inbreeding level and environmental variation
2. Screening of methylation levels as a function of the level of phenotypic plasticity in relation to level of inbreeding

To characterize the effects of habitat fragmentation on gene expression and genomic pathways

Possible conservation genomic approaches:

1. Use microarrays or RNA-seq to screen for changes in genome-wide gene expression profiles in response to inbreeding and population size
2. Screen gene-expression variation in high- and low-diversity populations and genotypes to disentangle direct gene effects from regulatory changes

Ouborg et al. (2010) Trends in Genetics 26: 177 - 187

Why is it so important for Forest conservation to study adaptive genetic variation?

- To understand and unlock the adaptive potential
- To be able to predict effects of climate change
- To mitigate these effects via breeding more resilient trees and promoting assisted migration
- To understand **evolutionary responses** and molecular mechanisms of genetic adaptation

Why genomics?

- **evolutionary response** is a **genetic adaptation** via genetic change that increase fitness of plants and animals and promotes their adaptation to their natural environment, including their biotic and abiotic interactions

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

Conclusions

Genomics will make a difference primarily in

- partitioning population structure into selectively neutral structure caused by genetic drift and adaptive structure caused by natural selection
- determining which genes, alleles and parts of the genomes are responsible for local adaptation and therefore important to preserve
- more accurately estimating effective population size
- more accurately estimating past demographics such as population size fluctuations and disentangling demographic events such as population size bottlenecks from selection allowing us to elucidate whether endangered species have been endangered and bottlenecked also during their past evolutionary history or whether their present threat status is a consequence of what is currently happening therefore providing information as to how the situation can be rectified

High-Throughput Genome-Wide Genotyping, Targeted Sequencing and Association Mapping of Adaptive and Breeding Traits in Loblolly Pine (*Pinus taeda* L.) Populations

Department of Ecosystem Science & Management
Texas A&M University, College Station, USA

ProCoGen Workshop "Genomics and the conservation of conifer genetic resources" Sept. 3rd 2014, Kámoni Arboretum Szombathely, Hungary

How to study genetic adaptation in forest tree populations?

1) Traditional methods

- field or common garden experiments (provenance, progeny and clonal tests)
- Quantitative Trait Locus (QTL) mapping

2) Population and ecological genomics approaches

- use of **functional genomic markers** and **adaptive trait related candidates genes** for QTL mapping, population and association studies
- **association mapping** with phenotypic and environmental variation using **high-density genome-wide genotyping** via high-throughput sequencing (NGS) in field and common garden experiments (provenance, progeny and clonal tests)
- detecting **genome-wide signatures of selection** (LD, selective sweeps, etc.) and loci under adaptive genetic divergence in natural populations using **neutrality tests and outlier-detection approaches**

3) Genomic selection using intense phenotyping, high-density genome-wide genotyping and regression models to predict phenotypes and breeding values in the progeny based on their genome-wide genotypes alone

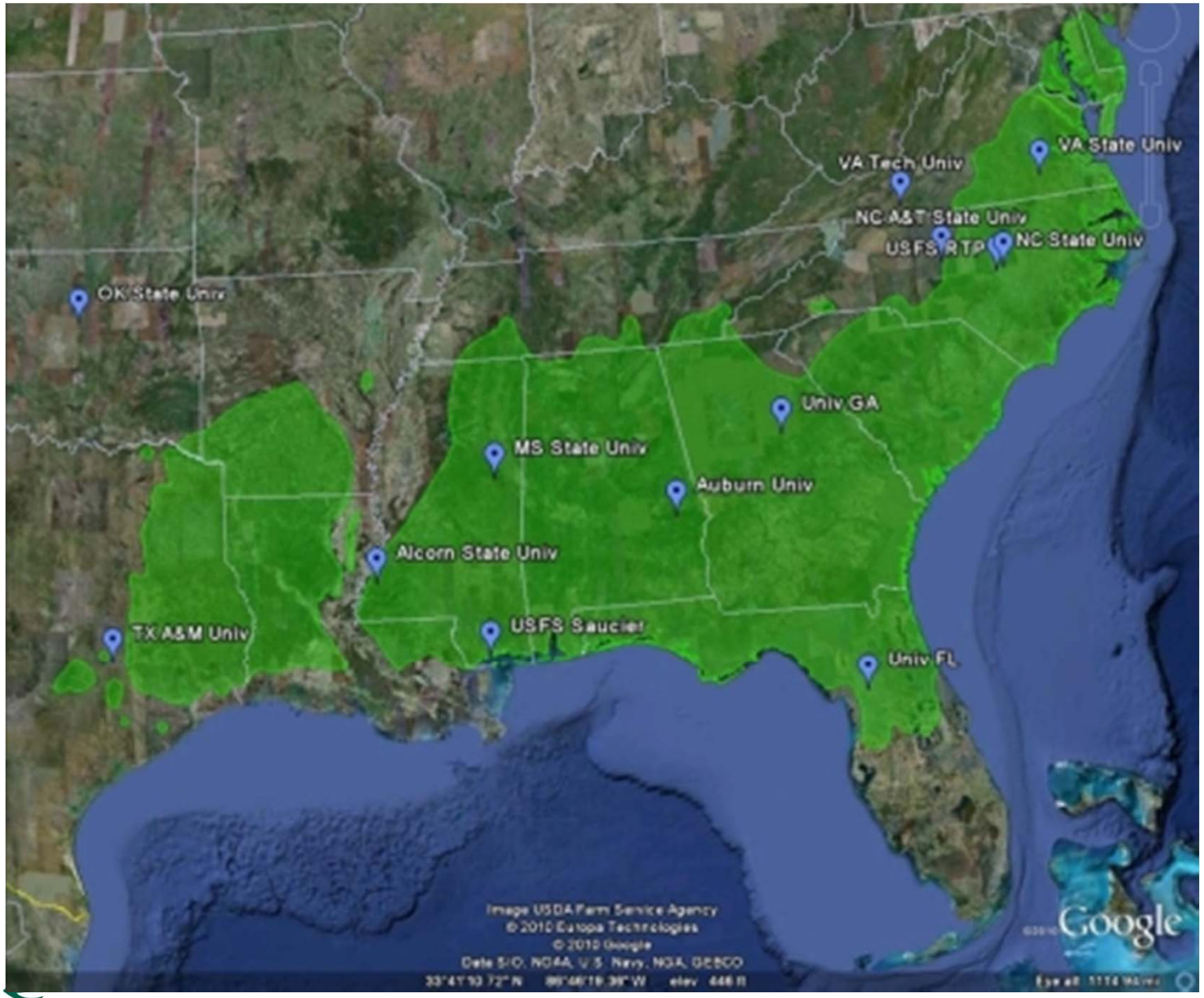
Integrated approach: 1) - 3)

Loblolly pine case studies

- USDA NRICGP Plant Genome Program / National Institute of Food and Agriculture (NIFA) Agriculture and Food Research Initiative (AFRI) Competitive Grants Program, Applied Plant Genetics Coordinated Agricultural Project (CAP), #CA-D-PLS-2038-CG, PI: D. B. Neale (dendrome.ucdavis.edu/ctgn/people), **2004-2011**, $6,000,000; "**Conifer Translational Genomics Network**" (**CTGN**).

- USDA NIFA AFRI Competitive Grants Program, CAP, Climate Change Program 1: Regional Approaches to Climate Change, Program Area Code – A3101, #2011-68002-30185, PI: Timothy Martin (www.pinemap.org/about/team-members), 5 years, **3/1/2011-2/28/2016**, $19,976,825; "**Integrating research, education and extension for enhancing southern pine climate change mitigation and adaptation**" (**PINEMAP**).

Loblolly pine area



Conifer Translational Genomics Network (CTGN)

- **CTGN** is a multi-state, multi-institution **Coordinated Agricultural Project (CAP)** funded by USDA and US Forest Service

USDA United States Department of Agriculture National Institute of Food and Agriculture

- **Project goal: Apply genomic resources and tools to practical breeding** by linking experimental research with tree breeding

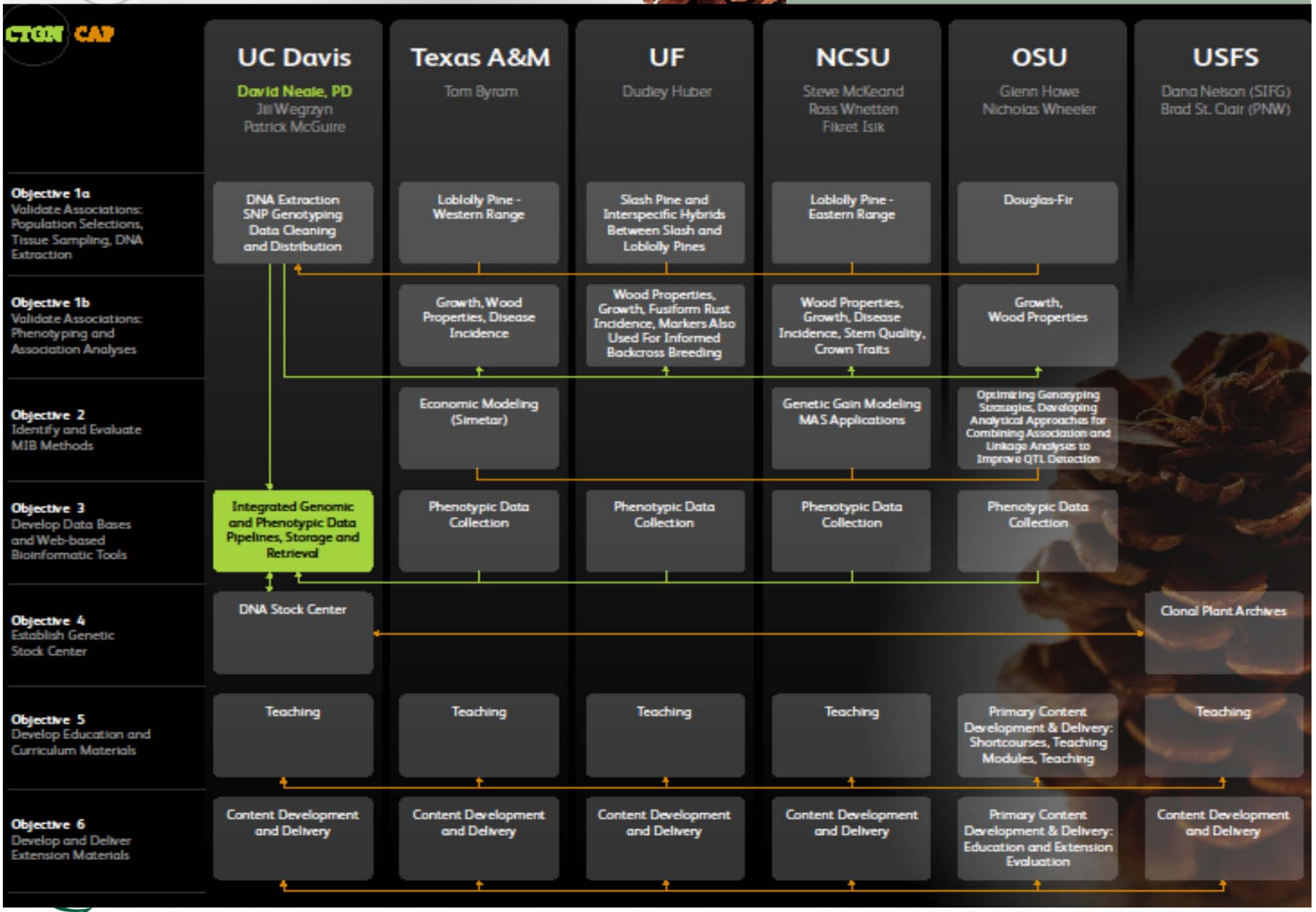
www.pinegenome.org/ctgn

- **Project Approach**:
  - large scale genotyping in elite loblolly pine and Douglas-fir populations belonging to tree improvement cooperatives
  - finding & validating genetic marker - phenotypic trait associations
  - modeling, outlining and implementing optimal approaches for incorporating markers in breeding programs (genomic selection)

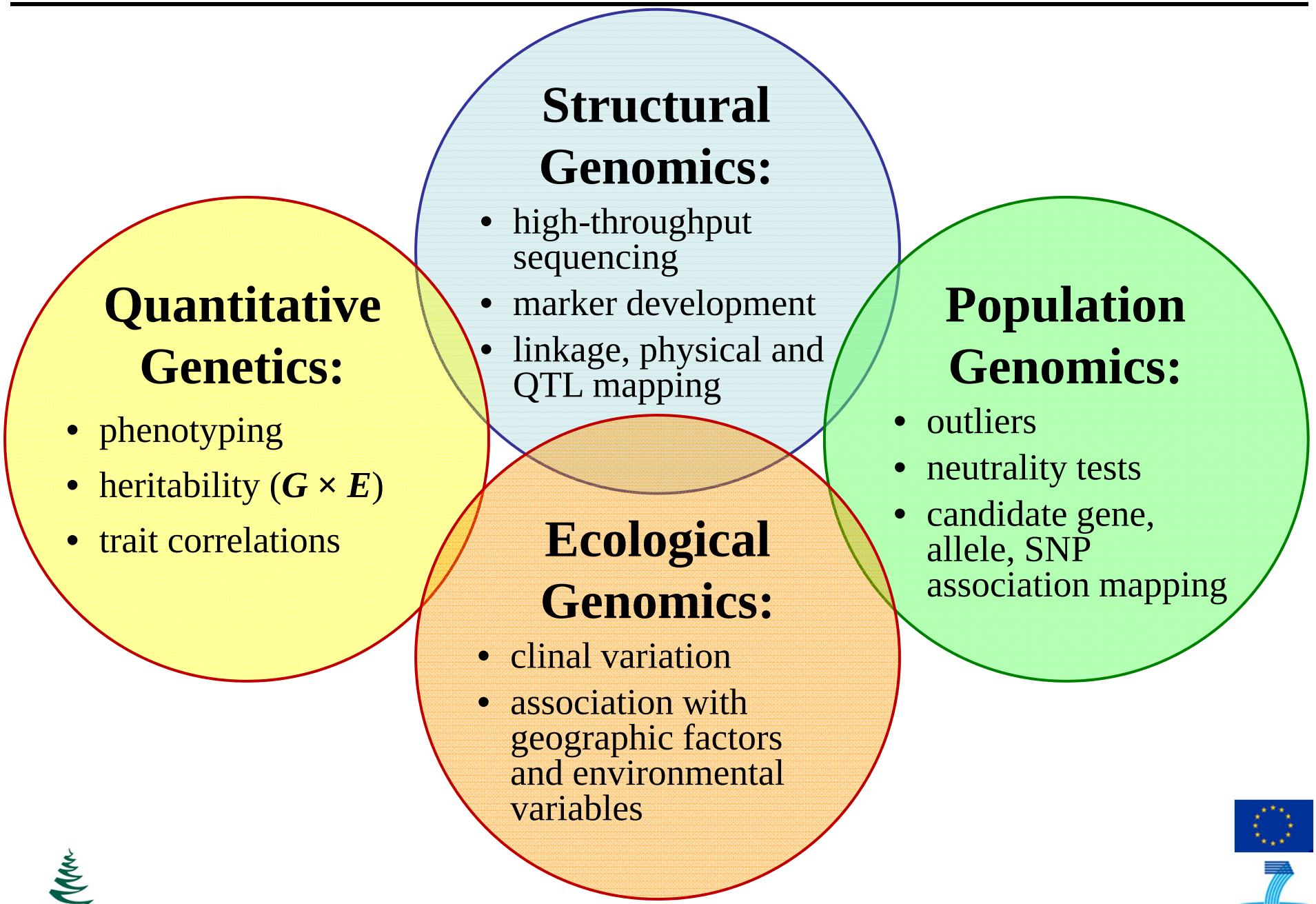
Conifer Translational Genomics Network
Coordinated Agricultural Project

GCTGCTGN CAPTCATCCATGATTAGCTTAGCTGGACCTA

dendrome.ucdavis.edu/ctgn

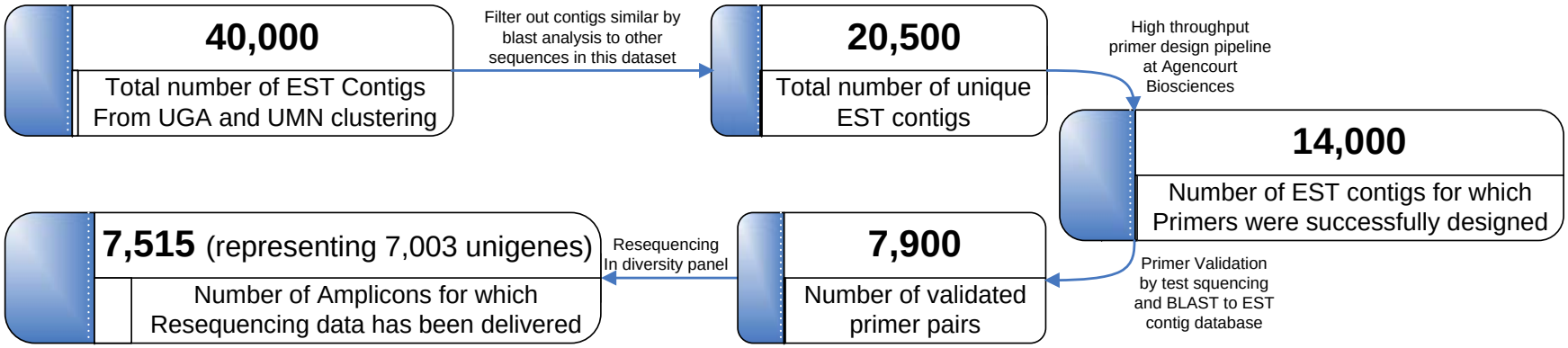


CTGN: Linking Genotype to Phenotype & Environment



Sequencing & SNP Genotyping

NSF Allele Discovery of Economic Pine Traits 2 (ADEPT2) resequencing & SNP discovery project



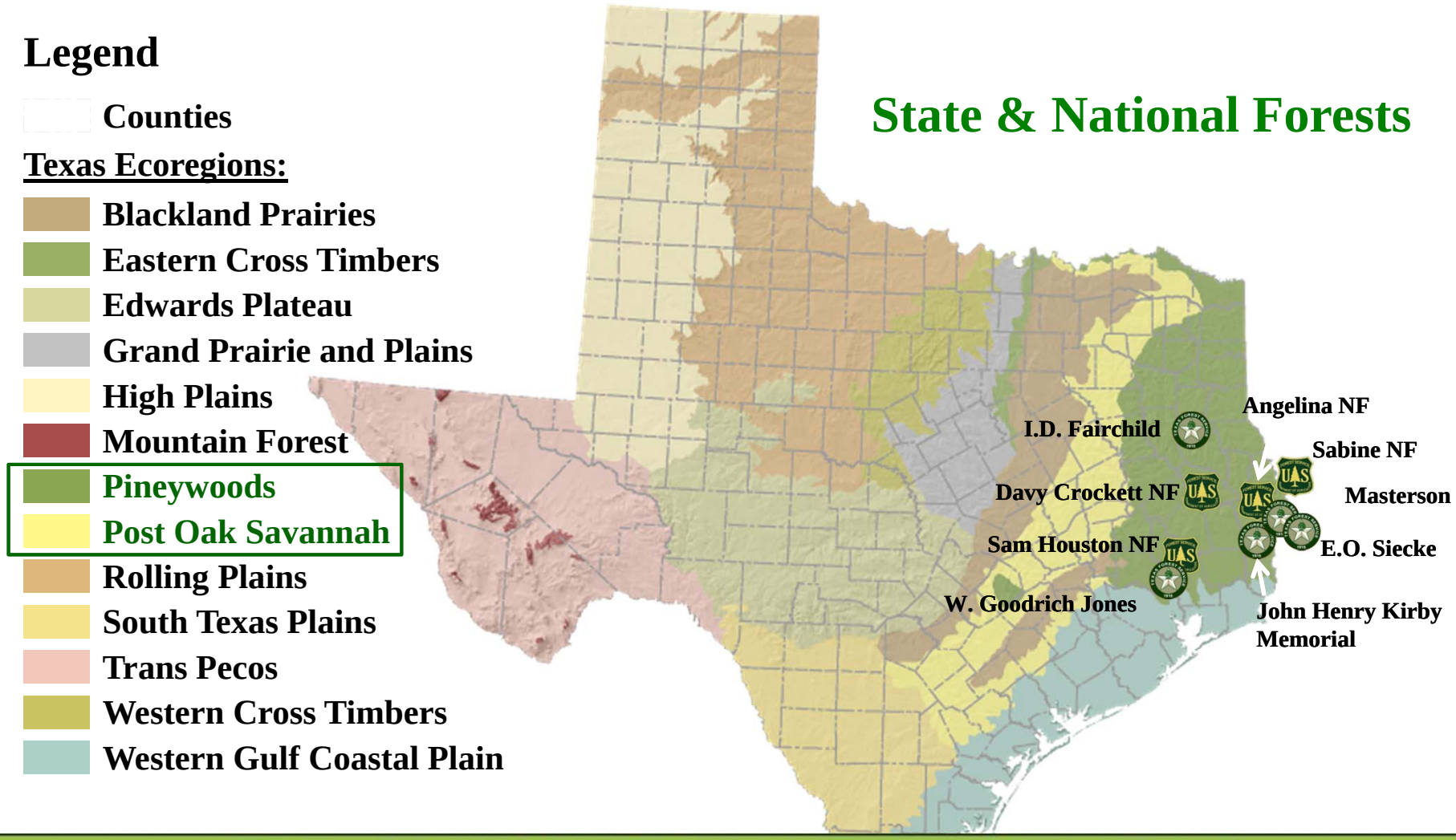
~23,000 SNPs discovered in ~7,000 partly amplified unique genes sequenced in 18 loblolly pine haploid megagametophytes

USDA Conifer Translational Genomics Network (CTGN) project

5,379 SNPs were genotyped in >4500 trees from multiple association and breeding populations using Illumina Infinium platform;

4,264 SNPs were polymorphic in East Texas populations

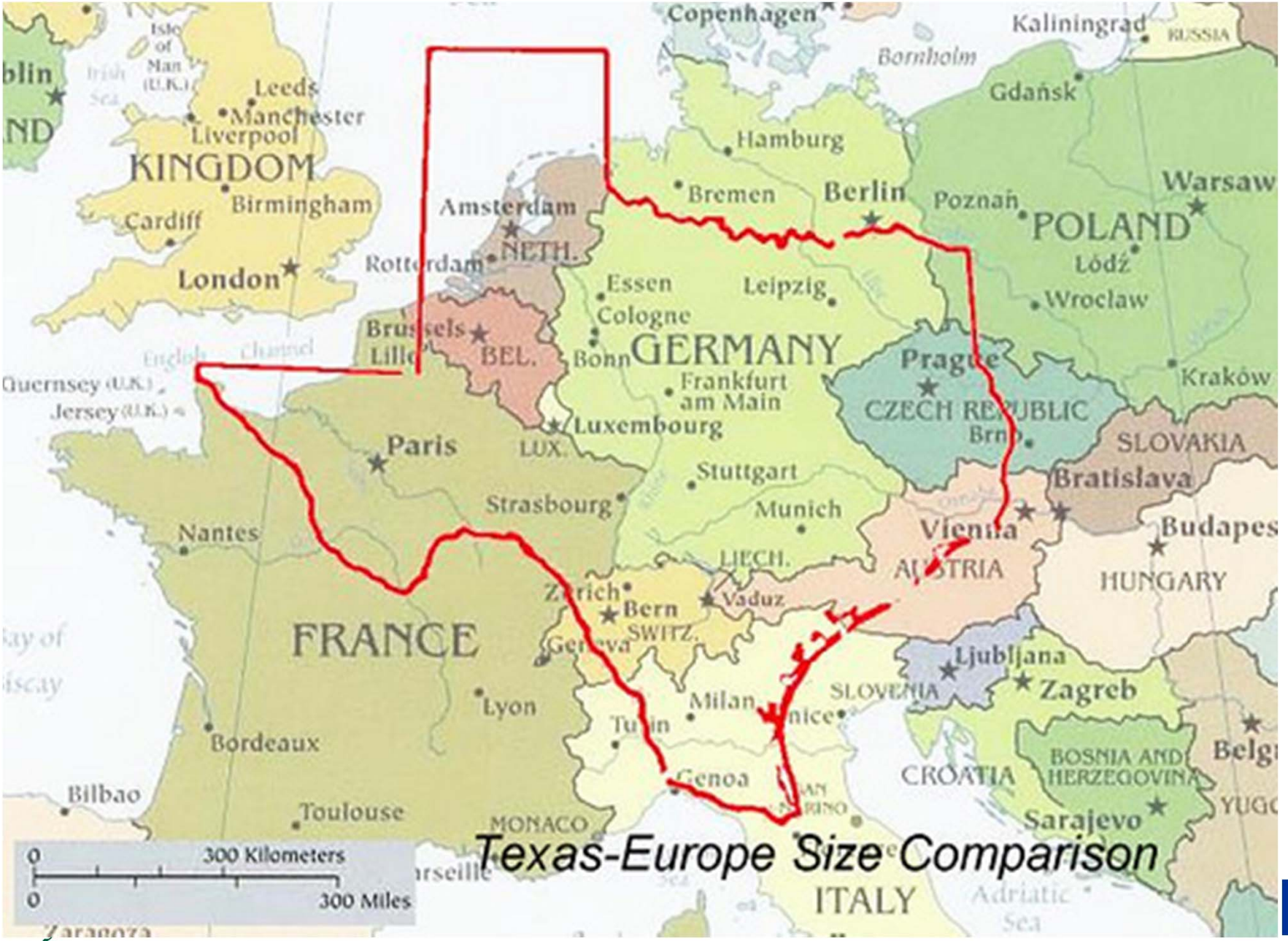
Texas Ecoregions



Texas has **60 million acres** of forestland
- more than any other state in the lower 48 United States

EVERYTHING IS BIGGER AND BETTER IN TEXAS — EVEN THE FORESTS

http://txforestservice.tamu.edu/main/popup.aspx?id=7002


Texas-Europe Size Comparison
Copenhagen
Kaliningrad
RUSSIA
Bornholm
Gdańsk
Hamburg
Bremen
Berlin
Poznań
Warsaw
POLAND
Łódź
Wrocław
Amsterdam
NETH.
Rotterdam
Leeds
Manchester
Liverpool
KINGDOM
Birmingham
Cardiff
London
Essen
Cologne
Leipzig
Brussels
BEL.
Bonn
GERMANY
Frankfurt am Main
Prague
CZECH REPUBLIC
Kraków
SLOVAKIA
Bratislava
Budapes
HUNGARY
Luxembourg
LUX.
Paris
Stuttgart
Strasbourg
Munich
Vienna
AUSTRIA
LIECH.
Vaduz
Zurich
Bern
SWITZ.
Nantes
FRANCE
Lyon
Milan
Turin
Genoa
Ljubljana
SLOVENIA
Zagreb
CROATIA
BOSNIA AND HERZEGOVINA
Sarajevo
Bordeaux
Bilbao
Toulouse
MONACO
ITALY
Adriatic Sea
Guernsey (U.K.)
Jersey (U.K.)
English Channel
Irish Sea
Isle of Man (U.K.)
0
300 Kilometers
0
300 Miles

Experimental Populations & Study Design

14 populations of the 1st generation selection

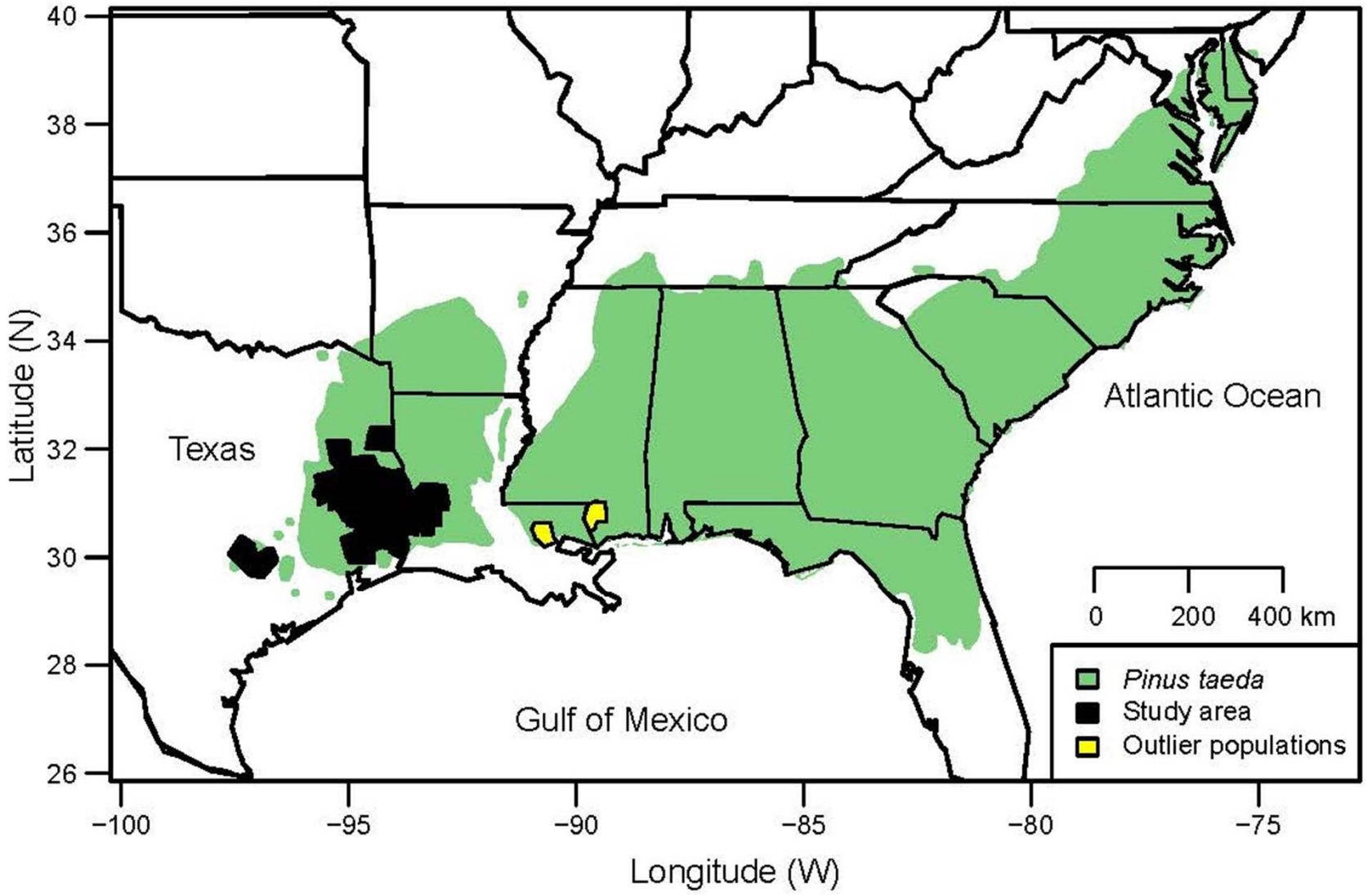
8 populations of the 2nd generation selection

30 breeding groups

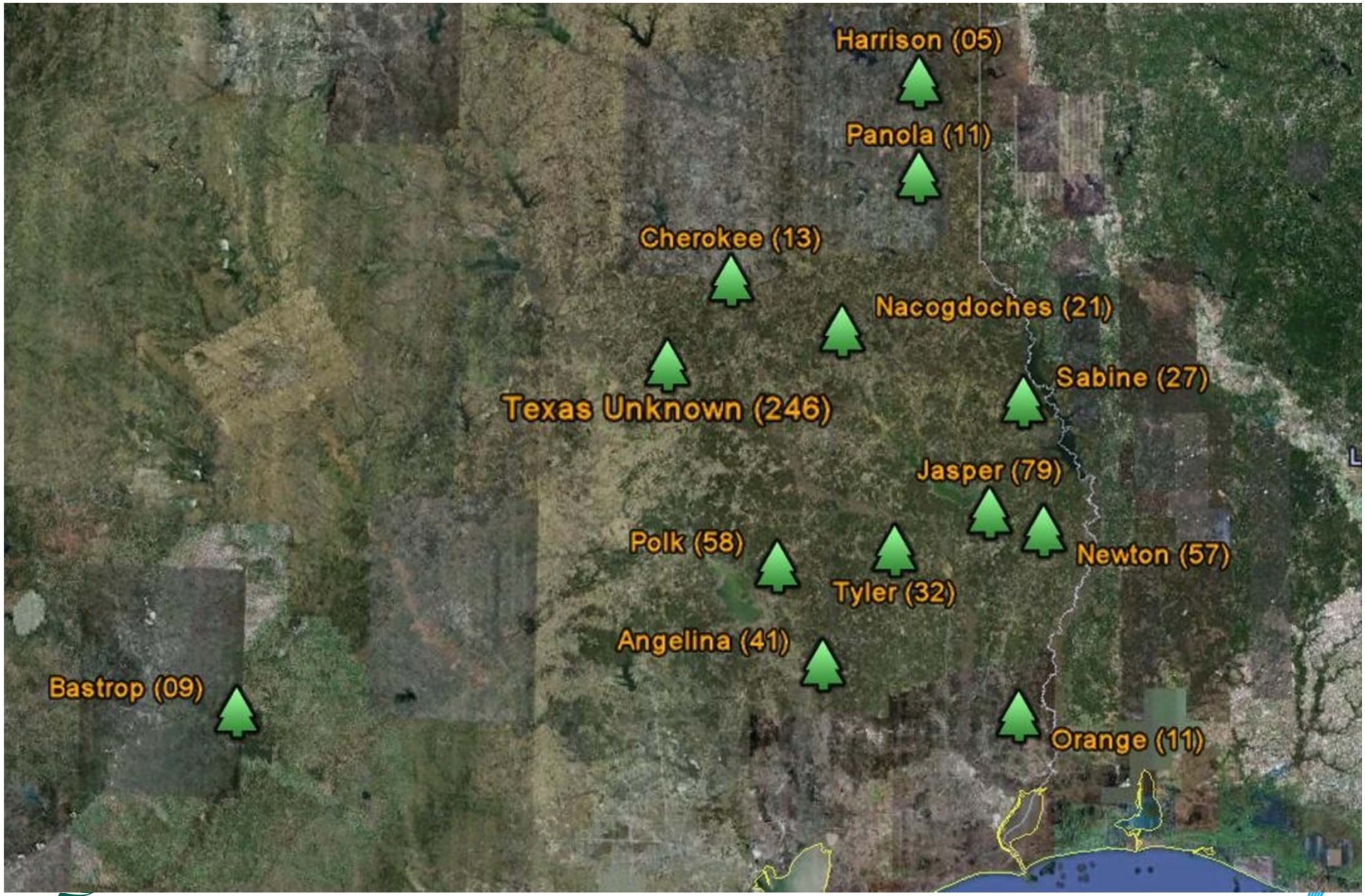
4,264 SNPs / >1,700 trees + **Phenotypes: *DBH, height, volume, etc.***
Environmental variables: *t, prec., arid.*

- **Genetic diversity: H_e, F_{IS}, etc.**
- **Genetic differentiation index: F_{ST}**
- **Selection detection via search for F_{ST} outliers**
- **Population structure inference using STRUCTURE & ΔK**
- **Association mapping & functional annotation of significant SNPs**
- **Genome wide LD**

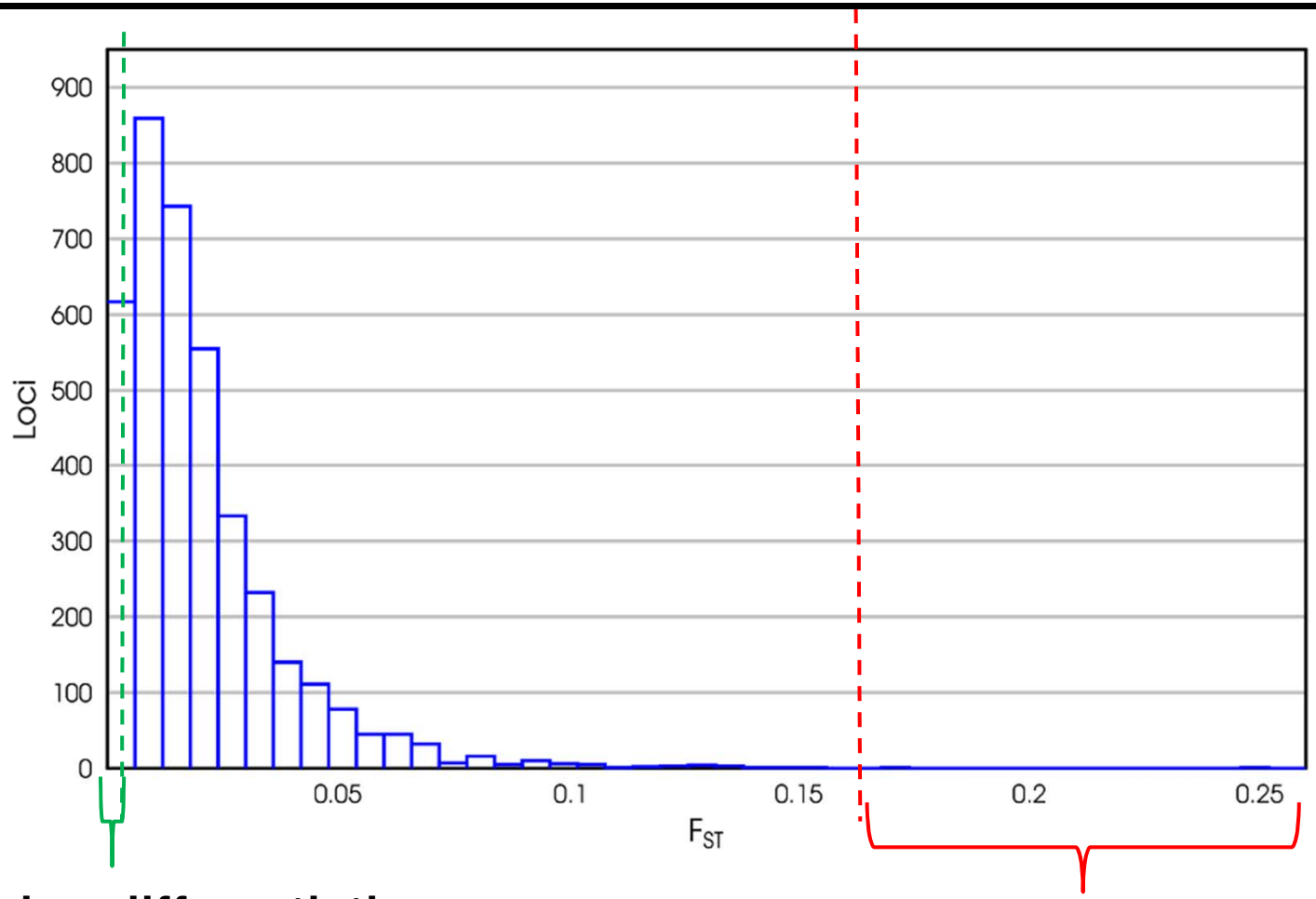
Loblolly pine study sites in East Texas



Geographic origin of the First Generation Selections



Genetic differentiation (F_{ST}) for 4,264 SNPs in the first generation of selections

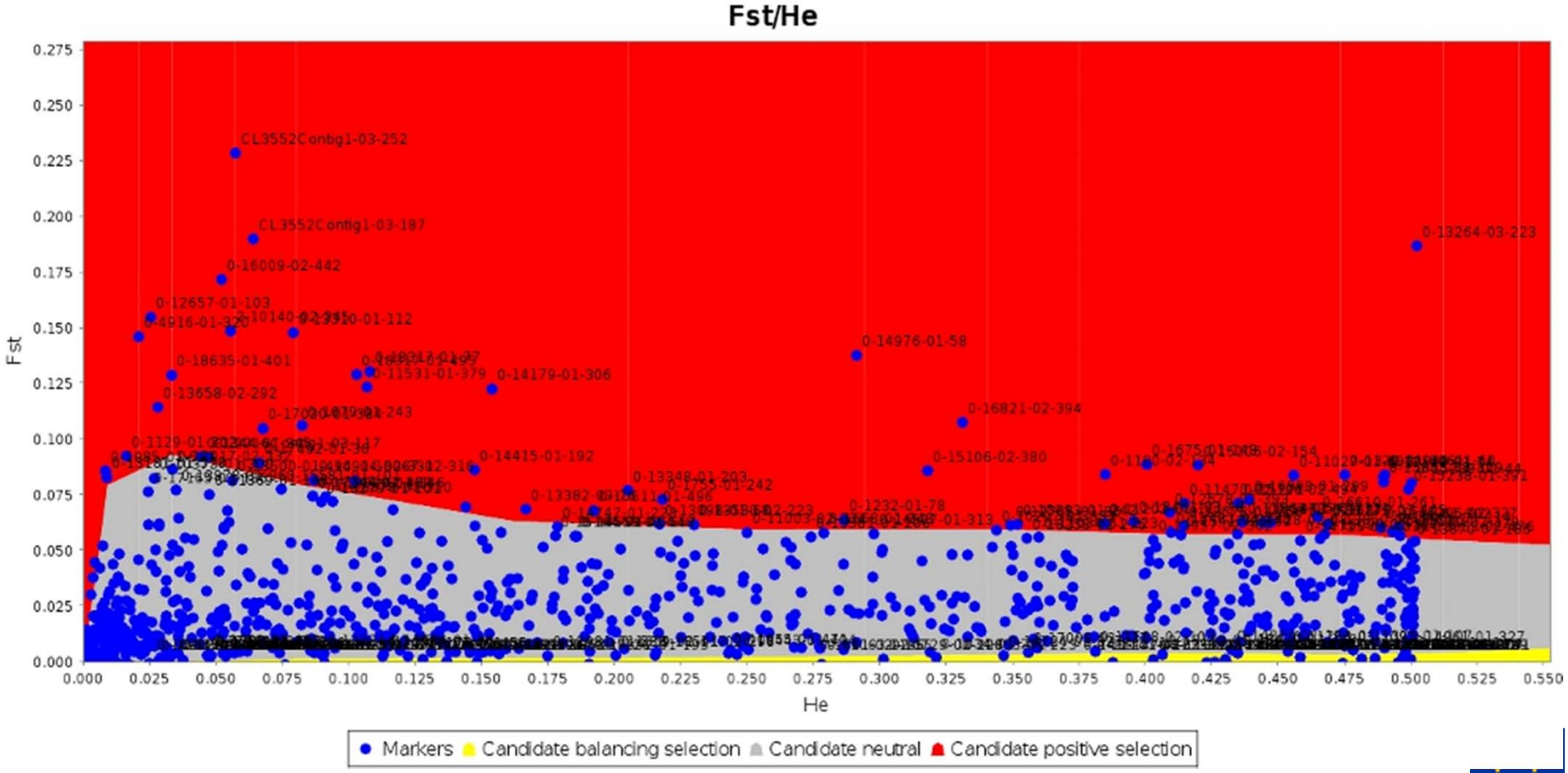


Very low differentiation; under balancing or stabilizing selection?

Very differentiated; under positive or divergent selection?

Search for outlier SNPs under selection and supposedly neutral SNPs using coalescence simulation to find thresholds for selectively neutral markers

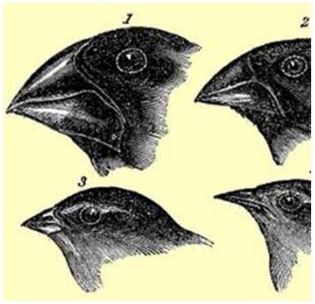
4,264 SNPs, distribution of F_{ST} & expected heterozygosity (H_e) in the loblolly pine populations (1st generation selections), 95K simulations to find thresholds



Beaumont *et al.* (1996) Royal Soc. B; Antao *et al.* (2008) Bioinformatics

F_{ST} «outliers»

Examples of candidate genes after correction for false positive:



Balancing selection (244):

Arabinofuranodisase, Glycoprotease protein, Dehydrin, Lypoxygenase, Cytokinin oxidase, Transmembrane transporters, Endoglucanase, MYB transcription factor, Pinus taeda Heme oxygenase I, etc.

Dirigent protein, Homeobox-leucine zipper, Cytochrome p450, Gras transcription factor, Gigantia protein, Ethylene-forming enzymes Histone H4, Reductases, etc.

Directional selection (74):

Geranyldiphosphate, Disease resistance proteins, Arabinogalactan-like proteins, Pinus Expansin, Pinus alpha-Xyloxidase, etc.

Potassium/proton antiporter, Laccase 90DProtein kinases, Histone H3.2, etc.

Chhatre *et al.* 2013 *Tree Genetics and Genomes* 9: 1161-1178

F_{ST} outliers and neutrality tests

Gene	Map: LG,cM	SNP	Het	Fst	*P*(Simul Fst<sample Fst = 0.010)	Neutrality test
4cl (*4-coumarate:CoA ligase*)	-	0-7767-01-191	0.0073	0.0000	0.7356	
	-	UMN-CL379Contig1-12-117	0.0013	0.0000	0.0000	Tajima's *D* +
ccoaomt(*caffeoyl CoA O-methyltransferase 1*)	V, 32.7	CL544Contig1-03-112	0.4992	0.0097	0.5488	Tajima's *D* +
ccr1(*cinnamoyl CoA reductase*)	-	CL594Contig1-06-236	0.4643	0.0051	0.4217	
comt2(*caffeate O-methyltransferase 2*)	-	0-10914-02-331	0.0397	0.0174	0.7236	
	-	0-10914-02-55	0.0097	0.0022	0.8055	
cpk-3 (*calcium-dependent protein kinase*)	-	CL2332Contig1-01-175	0.1571	0.0000	0.0086	
	-	CL2332Contig1-01-314	0.4127	0.0229	0.8561	
lp3-3(*water-stress inducible protein 3*)	-	CL1740Contig1-03-78	0.3703	0.0351	0.9621	
pal1(*phenylalanine ammonia-lyase 1*)	-	CL863Contig1-03-164	0.0093	0.0005	0.7761	
ppap12 (*putative wall-associated protein kinase*)	-	CL3898Contig1-04-256	0.0329	0.0165	0.7109	Tajima's *D* +
ptlim1 (*LIM domain protein 1 (LIM transcription factor)*)	-	CL1905Contig1-06-353	0.0142	0.0000	0.3565	
	-	CL1905Contig1-03-377	0.0142	-0.0007	0.3565	
ptlim2 (*LIM domain protein 2 (LIM transcription factor)*)	II, 3.5	CL711Contig1-04-212	0.1766	0.0010	0.2830	

Koralewski TE, Brooks JE and Krutovsky KV (2014) Molecular evolution of drought tolerance and wood strength related candidate genes in loblolly pine (*Pinus taeda* L.). *Silvae Genetica* **63**(1-2): 59-66

Association mapping of SNPs and phenotypic variation in adaptive and breeding traits (such as growth rate, wood density, disease resistance, drought tolerance, etc.)

Significant associations were found, for example, for:

- Arabinofuranosidase
- Xylosidase
- Protein kinases
- Chloroplast proteins
- Metallothionein-like protein
- Chlorophyll binding protein
- Glucuronase 4-epimerate
- Clavata-like receptor
- RNA polymerases
- Decarboxylases
- Sodium simporter family protein
- Acyl CoA synthetase
- Tubulin beta-chains
- NBS disease resist. protein - P. taeda Transmembrane protein
- Universal stress protein
- Cyclin-D like protein
- cdc2 protein kinases
- Synaptotagmins etc.

Chhatre *et al.* 2013 *Tree Genetics and Genomes* 9: 1161-1178

Clinal variation and association with environmental variables : Logistic regression (LR) data

Number of SNPs genotyped in 463 trees from 27 populations in East Texas:

- total = **5379**
- polymorphic = **4264**
- used for LR = **3667**

Clinal variation - significant correlation with:

- Latitude = **210**
- Longitude = **293**
- both latitude and longitude = **34**

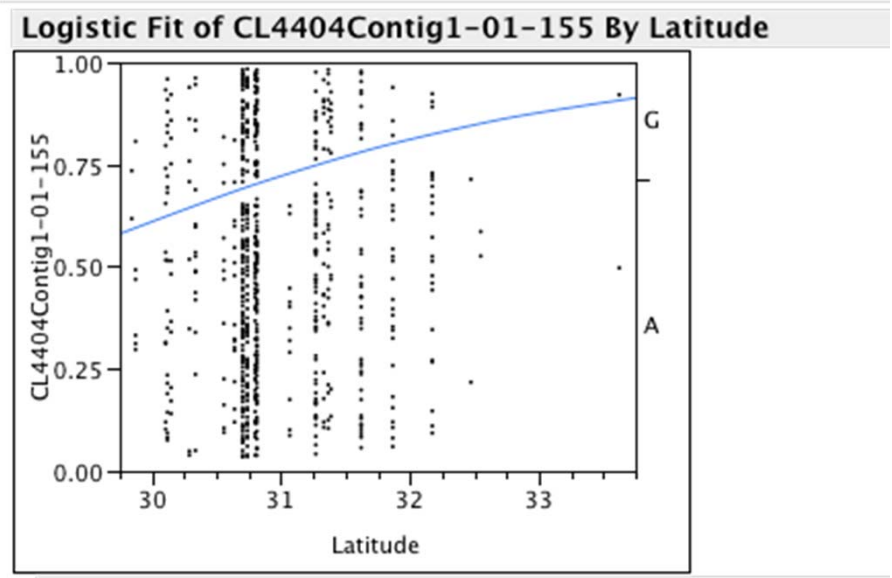
Environmental variables- significant correlation with:

- Monthly mean total annual temperature above 5˚C or Growing Degree Days (MEAN_annGDD5)* = **245**
- Mean Annual Precipitation (MEAN_annP)** = **268**
- Mean Annual Temperature (MEAN_MAT)*** = **259**
- Aridity index = *in progress*

Chhatre *et al.* 2014 *Molecular Ecology* (in prep.)

Clinal variation: Logistic regression (LR) data

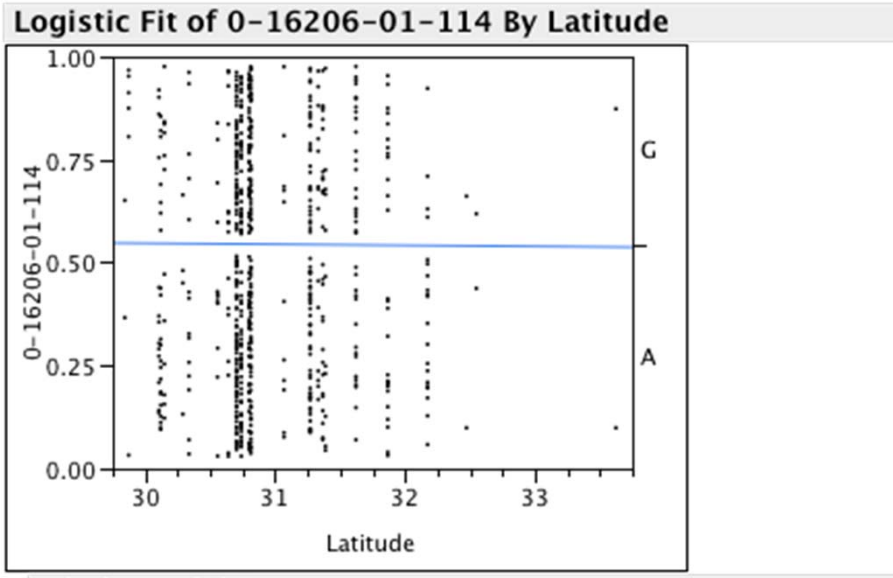
Significant association of the A/G alleles of the CL4404Contig1-01-155 SNP with Latitude:



Whole Model Test

Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	4.42624	1	8.85247	0.0029*

Insignificant association of the A/G alleles of the 0-16206-01-114 SNP with Latitude:

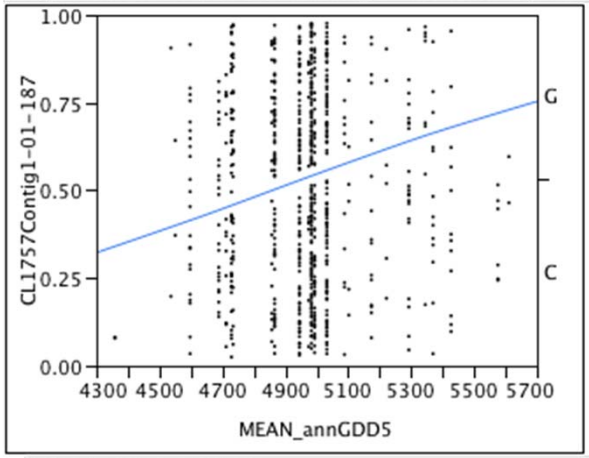


Whole Model Test

Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	0.00239	1	0.004776	0.9449

Association with environmental variables: LR data

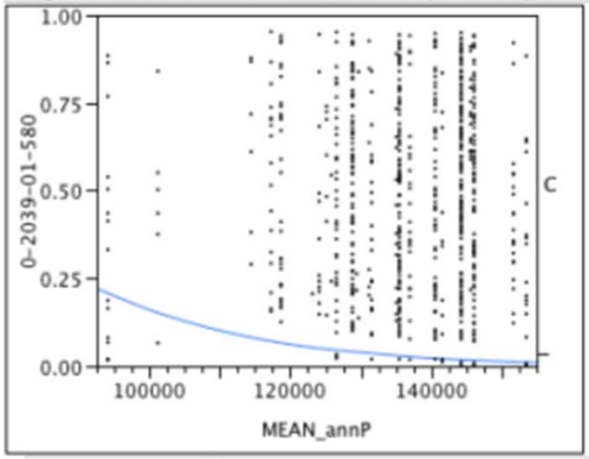
Logistic Fit of CL1757Contig1-01-187 By MEAN_annGDD5



Whole Model Test

Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	5.44054	1	10.88108	0.0010*

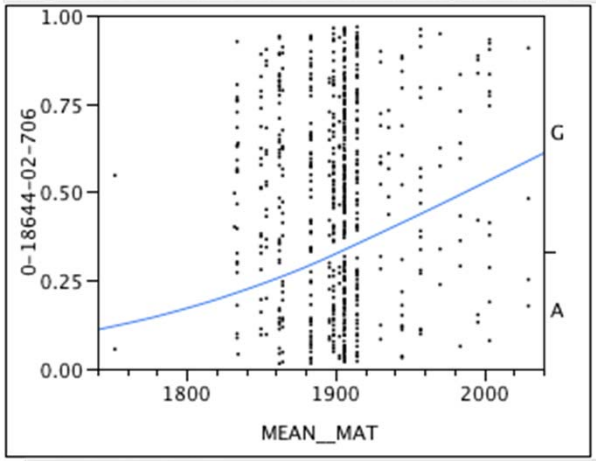
Logistic Fit of 0-2039-01-580 By MEAN_annP



Whole Model Test

Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	6.76393	1	13.52786	0.0002*

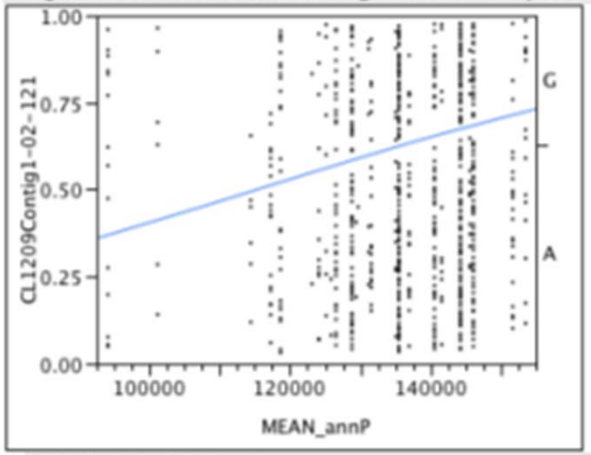
Logistic Fit of 0-18644-02-706 By MEAN_MAT



Whole Model Test

Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	6.40695	1	12.81391	0.0003*

Logistic Fit of CL1209Contig1-02-121 By MEAN_annP



Whole Model Test

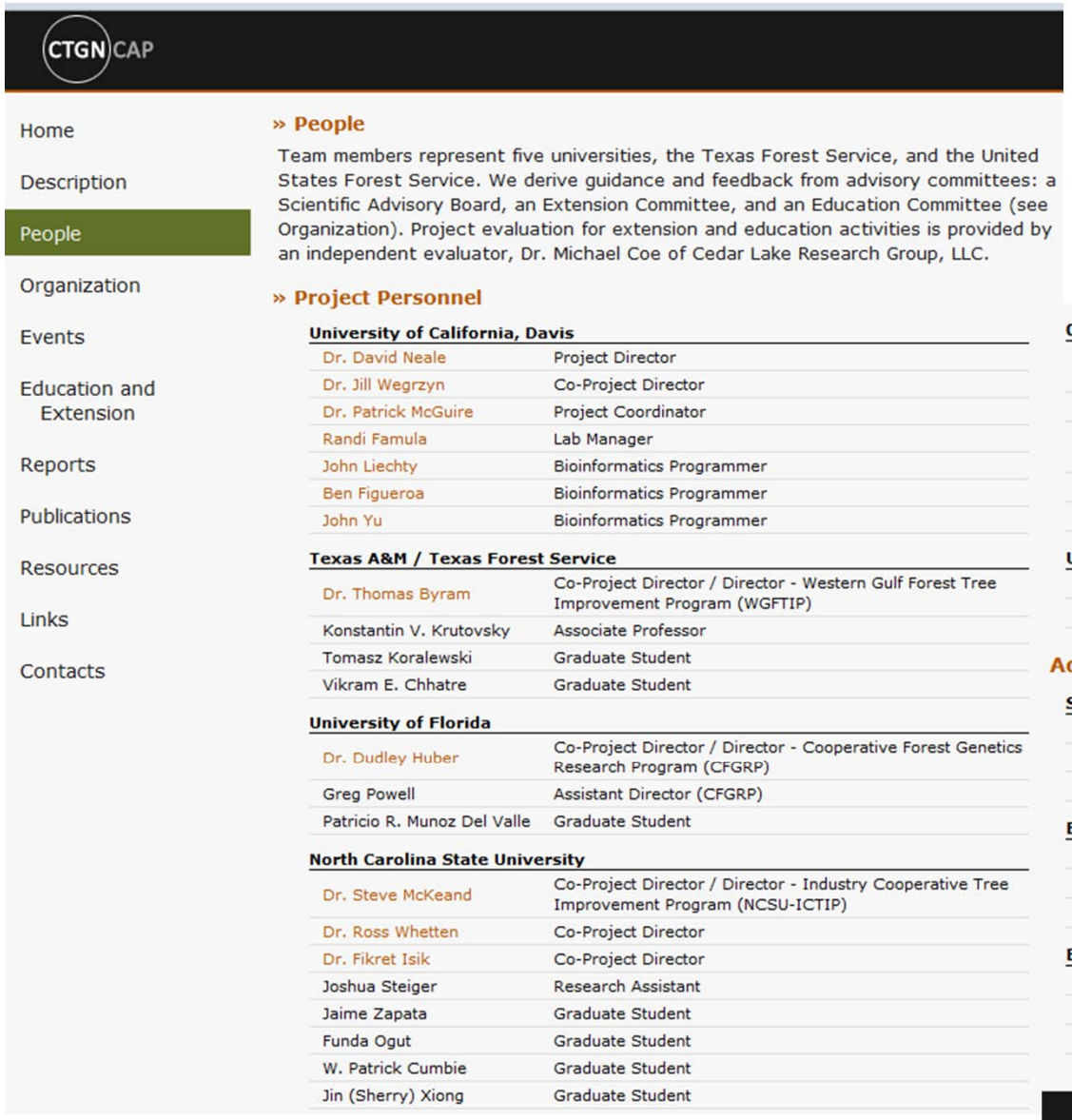
Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	7.23560	1	14.47121	0.0001*

CTGN: Results of ecological & population genomic studies of loblolly pine (*Pinus taeda* L.)

- SNPs from >5000 genes were genotyped in >4500 trees sampled from numerous natural and breeding populations covering the full-range of the species
- Significant associations were found between adaptive trait phenotypes, geographic and environmental variables (temperature, growing degree-days, precipitation and aridity) and a diverse sets of genes including abiotic stress response genes ranging from trans-membrane proteins to proteins involved in sugar metabolism and transcription factors
- Numerous genes under selection were found (outliers)
- Multiple allele candidates for local adaptation were discovered

Eckert *et al.* 2010 *Genetics* 185: 969–982
Eckert *et al.* 2010 *Molecular Ecology* 19: 3789–3805
Chhatre *et al.* 2013 *Tree Genetics and Genomes* 9: 1161-1178
Chhatre *et al.* 2014 *Molecular Ecology* (in prep.)

Acknowledgements



» People

Team members represent five universities, the Texas Forest Service, and the United States Forest Service. We derive guidance and feedback from advisory committees: a Scientific Advisory Board, an Extension Committee, and an Education Committee (see Organization). Project evaluation for extension and education activities is provided by an independent evaluator, Dr. Michael Coe of Cedar Lake Research Group, LLC.

» Project Personnel

University of California, Davis

Dr. David Neale	Project Director
Dr. Jill Wegrzyn	Co-Project Director
Dr. Patrick McGuire	Project Coordinator
Randi Famula	Lab Manager
John Liechty	Bioinformatics Programmer
Ben Figueroa	Bioinformatics Programmer
John Yu	Bioinformatics Programmer

Texas A&M / Texas Forest Service

Dr. Thomas Byram	Co-Project Director / Director - Western Gulf Forest Tree Improvement Program (WGFTIP)
Konstantin V. Krutovsky	Associate Professor
Tomasz Koralewski	Graduate Student
Vikram E. Chhatre	Graduate Student

University of Florida

Dr. Dudley Huber	Co-Project Director / Director - Cooperative Forest Genetics Research Program (CFGRP)
Greg Powell	Assistant Director (CFGRP)
Patricio R. Munoz Del Valle	Graduate Student

North Carolina State University

Dr. Steve McKeand	Co-Project Director / Director - Industry Cooperative Tree Improvement Program (NCSU-ICTIP)
Dr. Ross Whetten	Co-Project Director
Dr. Fikret Isik	Co-Project Director
Joshua Steiger	Research Assistant
Jaime Zapata	Graduate Student
Funda Ogut	Graduate Student
W. Patrick Cumbie	Graduate Student
Jin (Sherry) Xiong	Graduate Student

USDA United States Department of Agriculture | National Institute of Food and Agriculture

Oregon State University

Dr. Glenn Howe	Co-Project Director / Director-The Pacific Northwest Tree Improvement Research Cooperative (PNWTIRC)
Dr. Nicholas Wheeler	Co-Director
Dr. Keith Jayawikrama	Director - Northwest Tree Improvement Cooperative (NWTIC)
Dr. Terrance Ye	NWTIC Statistician
Dr. Jianbin Yu	Post-Doc

USFS

Dr. Dana Nelson	Co-Project Director (Southern Institute of Forest Genetics)
Dr. Brad St. Clair	Co-Project Director (Pacific Northwest Experiment Station)

Advisory Personnel

Scientific Advisory Board

Dr. Luca Comai	University of California, Davis
Dr. Jack Dekkers	Iowa State University
Dr. Julie Ho	Pioneer

Education Committee

Dr. Bert Abbott	Clemson University
Dr. Bill Beavis	Iowa State University
Dr. Toby Bradshaw	University of Washington

Extension Committee

Dr. Peggy Lemaux	University of California, Berkeley
Dr. James Johnson	Oregon State University
Dr. JB Jett	North Carolina State University, Emeritus

2012

http://dendrome.ucdavis.edu/ctgn/people/

PINEMAP project: Goals & Aims

Overall goal is to create, synthesize, and disseminate knowledge that enables southern forest landowners to

- manage forests to increase carbon sequestration by 15% by 2030;
- increase the efficiency of nitrogen and other fertilizer inputs by 10% by 2030; and
- adapt forest management approaches and **plant improved tree varieties to increase forest resilience and sustainability under variable climates.**

www.pinemap.org

PINEMAP project: Investigators, Organizations & Teams

- 57 scientists, educators, and extension professionals
- 11 southeastern land grant universities and the USDA Forest Service
- 6 disciplinary groups
- 4 research teams:
    - (1) Ecosystem Ecology / Silviculture
    - (2) Modeling
    - **(3) Genetics and Breeding**
    - (4) Economics and Policy

www.pinemap.org

PINEMAP project: Genetics & Breeding

Main objectives:

- Discovering adaptive genetic variation in association mapping studies via **genotyping-by-sequencing (GBS)** and using it in Loblolly pine forest management and genomic selection

GBS needs genome complexity reduction

www.pinemap.org

PINEMAP project:
Genome complexity reduction followed by Genotyping-by-Sequencing (GBS)

1. Restriction-enzyme-based double digest **R**estriction site-**A**ssociated **D**NA **Seq**uencing (**RAD-Seq**) using Illumina Hiseq2000 (R. Whetten, F. Isik: the North Carolina State University Cooperative Tree Improvement Program will genotype the Plantation Selection Seed Source Study (PSSSS) association mapping population planted across the region from four different coastal provenances).

2. Single Digest **R**estriction site-**A**ssociated **D**NA **Seq**uencing (**RAD-Seq**) using Illumina Hiseq2000 (G. Peter: the Cooperative Forest Genetics Research Program at the University of Florida will genotype the Comparing Clonal Lines ON Experimental Sites (CCLONES) association mapping population).

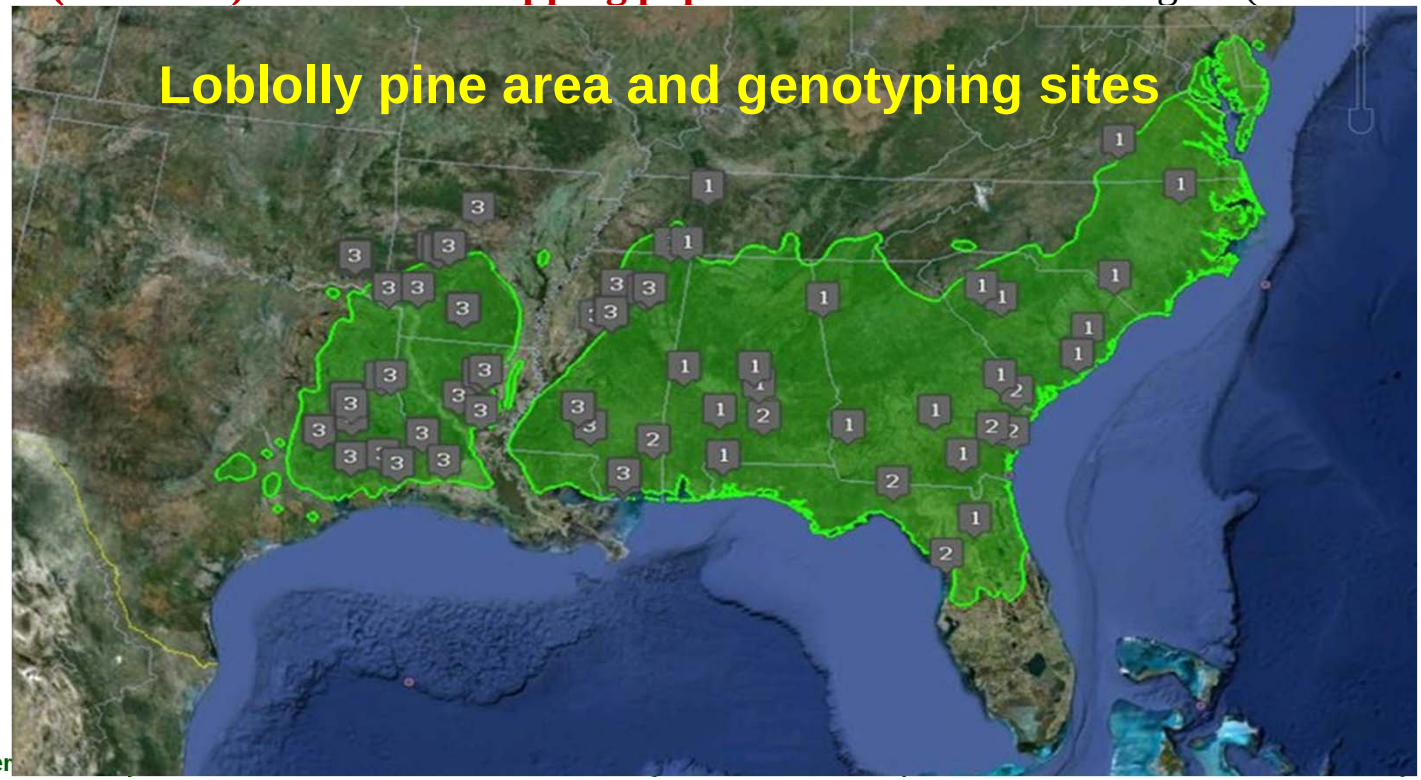
3. In solution hybridization-capture based **Agilent SureSelect Target Enrichment** followed by Illumina Hiseq2000 sequencing (K. V. Krutovsky, C. Loopstra, T. Byram: the Western Gulf Forest Tree Improvement Program at Texas A&M University and Texas Forest Service will genotype the Allele Discovery of Economic Pine Traits2 (ADEPT2) association mapping population for the western region; Jason Holiday: Virginia Tech).

PINEMAP project: Genetics & Breeding Teams

1. **Ross Whetten, Fikret Isik, Steve McKeand**: the North Carolina State University Cooperative Tree Improvement Program will genotype the **Plantation Selection Seed Source Study (PSSSS) association mapping population** planted across the region from four different coastal provenances ("**1**" in map).

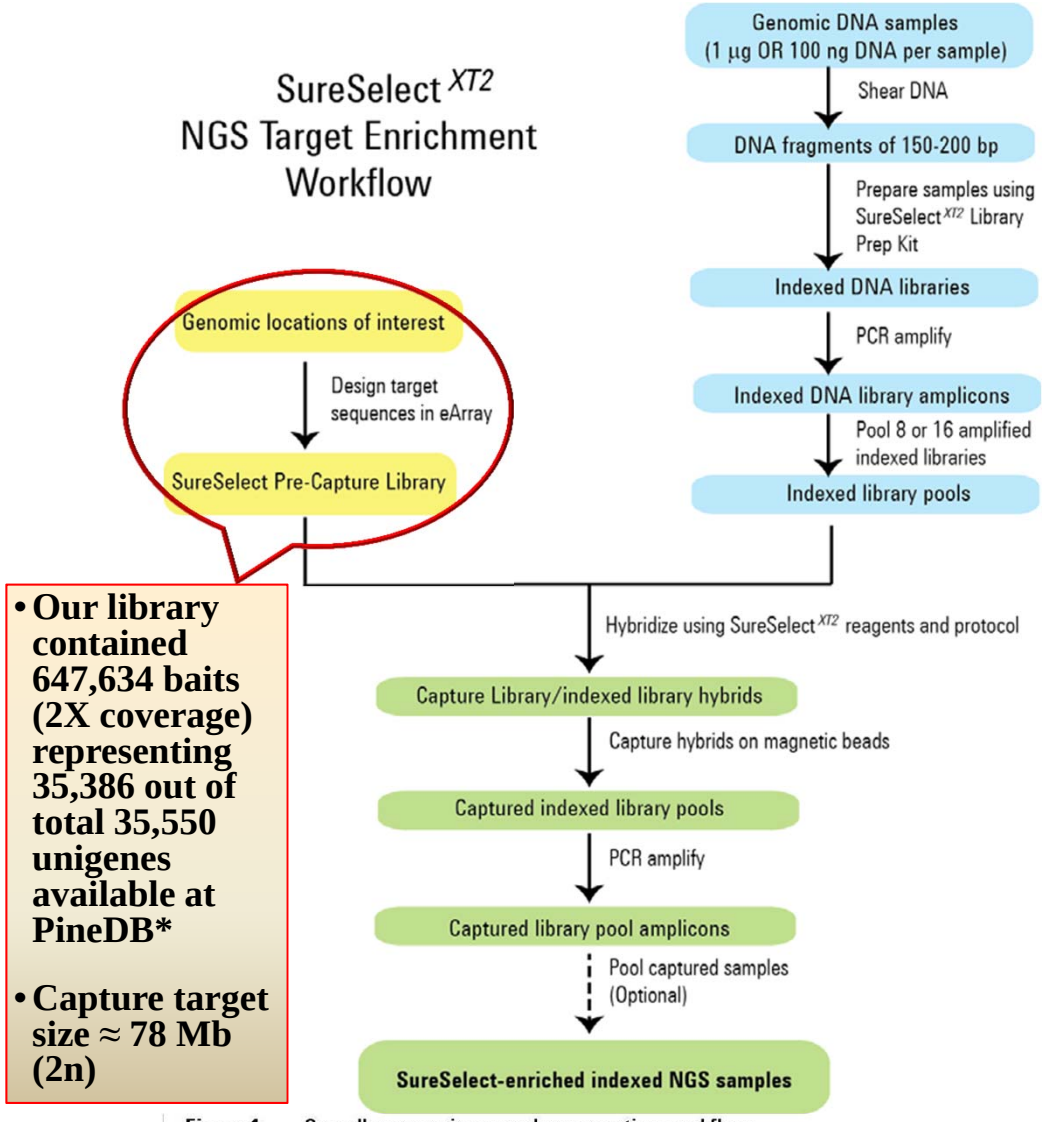
2. **Gary Peter, John Davis**: the Cooperative Forest Genetics Research Program at the University of Florida will genotype the **Comparing Clonal Lines ON Experimental Sites (CCLONES) association mapping population** ("**2**" in map).

3. **Kostya Krutovsky, Carol Loopstra, Tom Byram**: the Western Gulf Forest Tree Improvement Program at Texas A&M University and Texas Forest Service will genotype the **Allele Discovery of Economic Pine Traits2 (ADEPT2) association mapping population** for the western region ("**3**" in the map)



PINEMAP

Exome target enrichment using the Agilent's SureSelect Target Enrichment System for genotyping by sequencing (GBS) using NGS



- **Our library contained 647,634 baits (2X coverage) representing 35,386 out of total 35,550 unigenes available at PineDB***
- **Capture target size ≈ 78 Mb (2n)**

Figure 1 Overall sequencing sample preparation workflow.

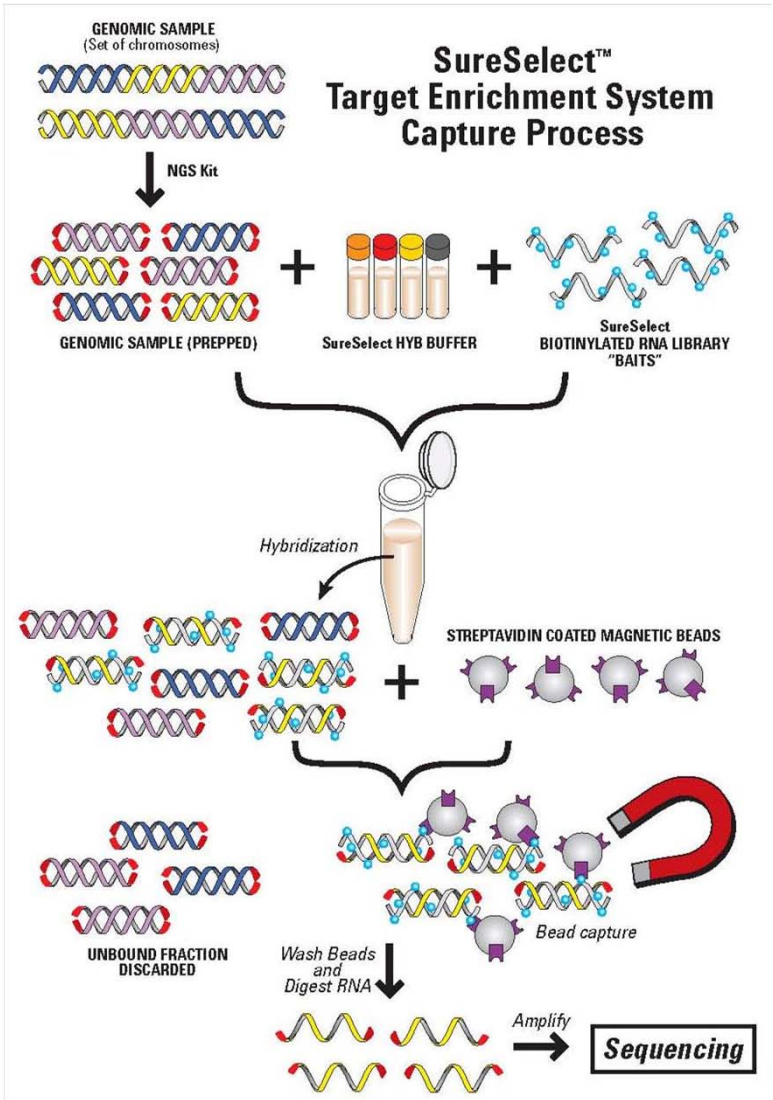
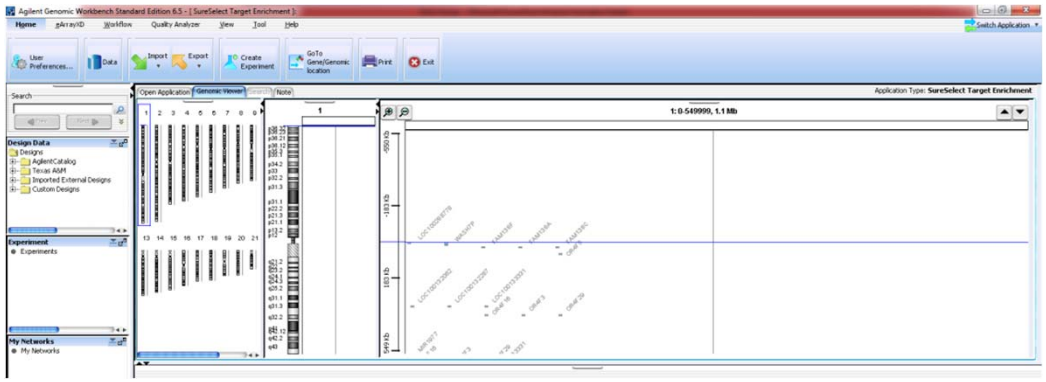


Figure 2 SureSelect Target Enrichment System Capture Process

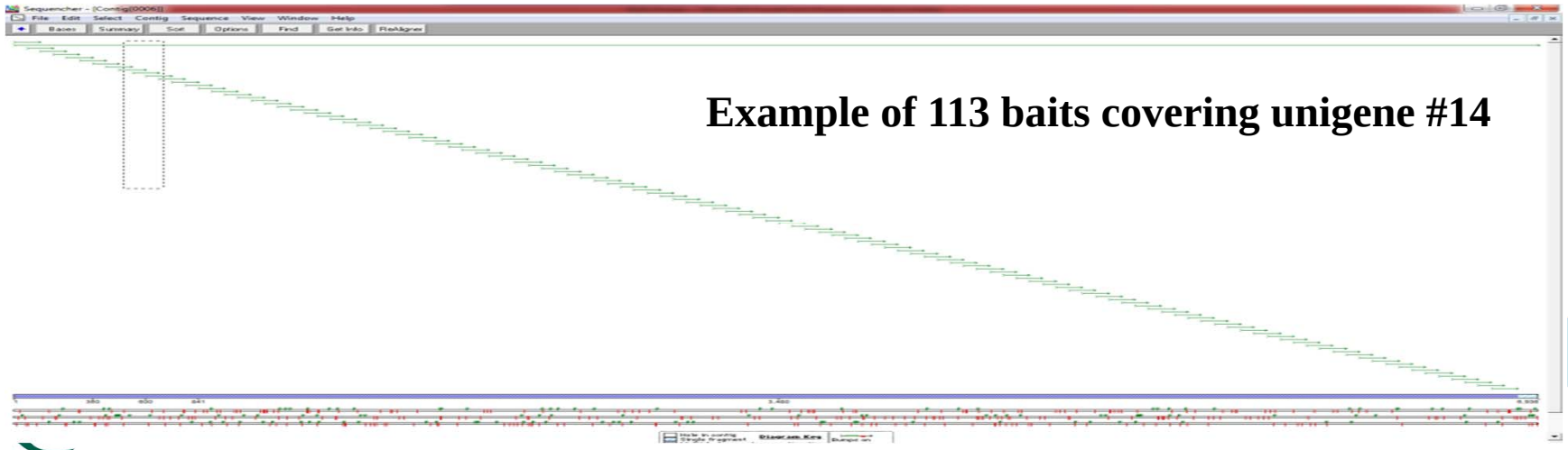
http://www.chem.agilent.com/library/usermanuals/Public/G9630-90000.pdf

* http://bioinfolab.muohio.edu/txid3352v1

Exome enrichment for 35,386 genes in loblolly pine for NGS using bar-coding and the Agilent's SureSelect Target Enrichment System



647,634 oligonucleotide hybridization 120 bp long probes (baits) based on 35,386 unigenes build by Dr. Chun Liang (Miami University; PineDB v.1 bioinfolab.muohio.edu/txid3352v1) were designed to target 78 Mb of gene space using Agilent Genomic Workbench software to gene enrich DNA libraries for sequencing



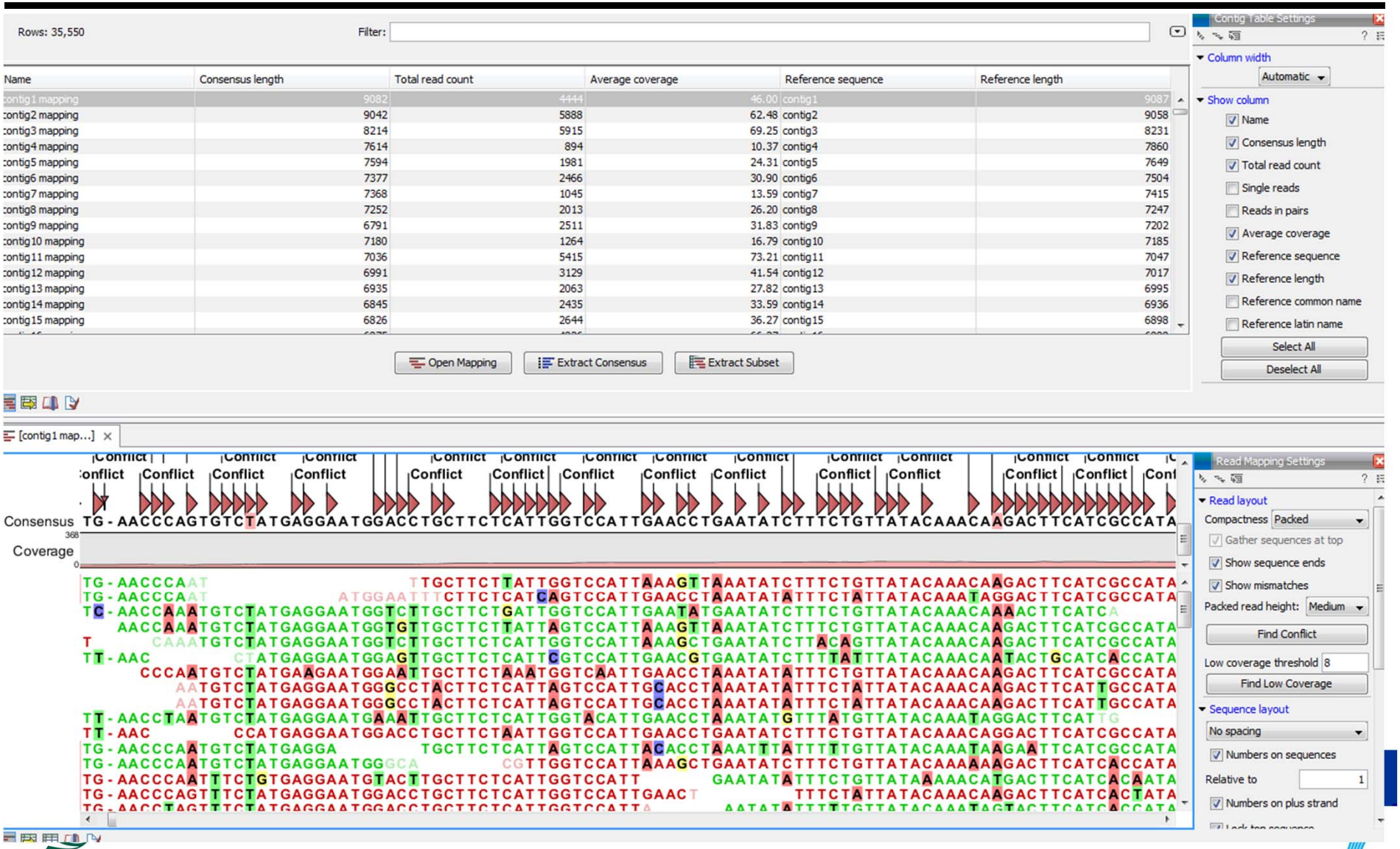
Example of 113 baits covering unigene #14

Illumina HiSeq2000 paired-end (2×100 bp) sequencing data

# Pooled samples per HiSeq2000 flowcell lane	Total reads, mln (coverage, X)	Mapped reads, mln	Uniquely mapped reads, mln
2	123 (12.3Gb ≈ 158X)*	119 (97%)	69 (56%)
4	404 (40.4Gb ≈ 259X)	376 (93%)	226 (56%)
8	342 (34.2Gb ≈ 110X)	320 (93%)	184 (54%)

*(123 Gb / 72 Mb ≈ **158X** coverage per target on average)

Mapping short reads back to the unigenes and to the draft loblolly pine reference genome assembly (v1.01, provided by the PineRefSeq project; http://pinegenome.org/pinerefseq) using the *CLCbio* software



SNPs discovery using SAMtools and only uniquely mapped Illumina HiSeq2000 short sequence reads

# Pooled samples	#SNP detected	SNP diversity, π
2	1,905,814	0.0026
4	1,870,997	0.0061
8	1,816,724	0.0050

Mapping criteria for SAMtools and Freebayes: minimum read depth = 10, at least 30% of total uniquely mapped reads contain an alternate allele

Conclusions

- Agilent SureSelect Target Enrichment method is very effective to capture target sequences in large genomes such as pine (≈21 Gb)

- Under a relatively stringent conditions, the rate of SNPs was 1 SNP per every 100 bp on average (close to expected from earlier studies of nucleotide variation in pine coding regions)

- Most of the unigenes were effective for baits design, but well-annotated reference complete genome sequence is needed for final verification and better bait design

Conclusions

- Integrative genomic approach helps to study local adaptation and to find genes and alleles under selection
- Genome-wide association mapping helps to identify genes responsible for adaptive trait variation, to study local adaptation, and to find genes and alleles under selection
- Genomic selection should help with **breeding for resilient trees**
- Technology of target gene enrichment is effective for genome-wide gene **genotyping by sequencing (GBS)**

Acknowledgements

- US Department of Agriculture
- National Science Foundation
- Western Gulf Forest Tree Improvement Program
- Texas Forest Service and Industry Partners
- CTGN Members & Collaborators
- Genetics Graduate Program, Texas A&M University

TAMU Department of Ecosystem Science and Management

Laboratory of Forest Genomics

Dr. Kostya Krutovsky
Adjunct Professor
k-krutovsky@tamu.edu
(979) 458-1417

Dr. Judy Brooks
Research Assistant
jebrooks@tamu.edu

Dr. Tomasz Koralewski
Former Graduate Student
now Research Associate
tkoral@tamu.edu

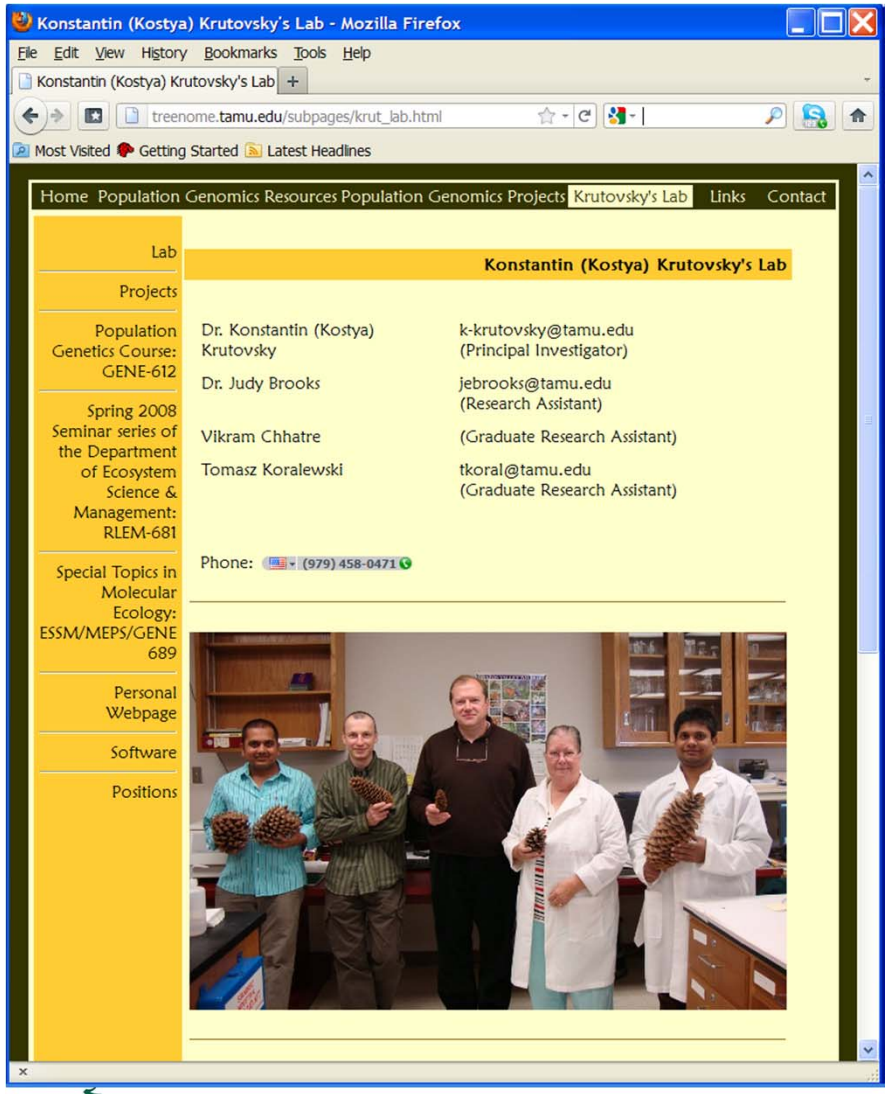
Mengmeng Lu
Graduate Student
mira0501@tamu.edu

Dr. Vikram Chhatre
Former Graduate Student
crypticlineage@gmail.com
(now Research Associate at the Center for Environmental Science, University of Maryland)

Habibul Islam
Graduate Student
milon237@tamu.edu

Acknowledgements

TAMU Department of Ecosystem Science and Management



Dr. Carol Loopstra

Texas A&M University ESSM

Dr. Tom Byram

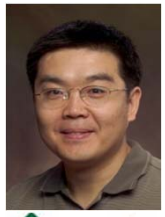
TEXAS FOREST SERVICE
The Texas A&M University System

Dr. David Neale

UCDAVIS
DEPARTMENT OF PLANT SCIENCES
College of Agricultural and Environmental Sciences

Dr. Jill Wegrzyn

UCONN
Biotechnology•Bioservices Center



Dr. Chang Liang
Department of Botany

MIAMI UNIVERSITY
OXFORD OHIO

SIFG
Southern Institute of Forest Genetics

Dr. Dr. Dana Nelson
Dr. Craig Echt
Sedley Josserand

FOREST SERVICE US DEPARTMENT OF AGRICULTURE

Take home message

Population Genomics together with Molecular Ecology (Ecogenomics) help us:

- discover genes and alleles that are responsible for adaptation
- link genotypes to adaptive phenotypes and to environment

Why complete genome sequence is important?

How would forest genetics and forest protection benefit from complete genome sequence for major conifers?

- identify and annotate genes, other functional elements (sRNA, transcription factors, regulatory elements, etc.) and genetic networks that control adaptation and disease resistance
- develop highly informative genetic markers that can be used in population genetic studies to create database of barcodes for individual populations to fight illegal timber harvest and trade
- develop genome-wide genetic markers for association studies for linking genetic variation (SNPs, alleles, haplotypes, and genotypes) with environmental factors, adaptive traits and phenotypes for better understanding genetic control of agronomically and economically important traits
- develop genome-wide genetic markers for genomic-assisted selection to breed for better adapted and desirable quality trees
- integrate proteomics, transcriptomics and metabolomics
- reference genome for resequencing

Current complete *de novo* conifer genome sequencing projects

Species	Leading organization (PI, budget, start year)
Norway spruce (***Picea abies***)	Umeå Plant Science Centre, Sweden (Dr. Pär Ingvarsson, $12M, 2010)
Loblolly pine (***Pinus taeda***), Douglas-fir (***Pseudotsuga menziesii***), Sugar pine (***Pinus lambertiana***)	University of California, Davis, USA (Dr. David Neale, $15M, 2011)
White, Sitka and Black spruce (***Picea glauca*** , ***P. sitchensis*** & ***P. mariana***)	Université Laval, Canada (Dr. John MacKay, $10M, 2010)
Maritime pine (***Pinus pinaster***), Scots pine (***Pinus sylvestris***)	European Union (Drs Carmen Diaz-Sala, University of Alcalá, Spain & María-Teresa Cervera, INIA CIFOR, Spain, $10M, 2013)
Siberian larch (***Larix sibirica***), Siberian pine (***Pinus sibirica***)	Siberian Federal University, Russia (Dr. Konstantin Krutovsky, $3M, 2014)

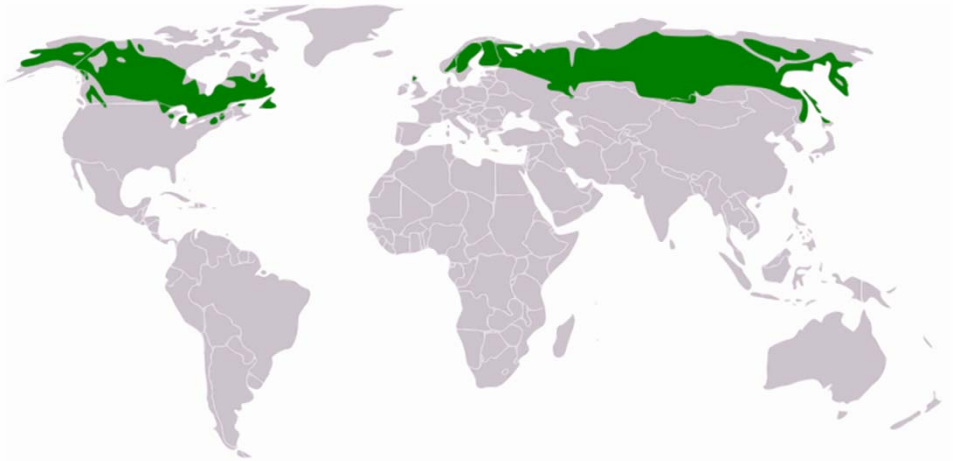
Scale of the problem: gigantic size of the genome!

Comparative genome sizes in conifers that are objects of genome sequencing in current projects

Species [1]	DNA (1C)		Ratio to human genome
	pg	*Gbp*	
Human (***Homo sapiens***) [2]	**3.47**	**3.20**	**1**
Siberian larch (***Larix sibirica***) [3]	**12.30**	**12.03**	**4**
Douglas-fir (***Pseudotsuga menziesii***) [4]	**19.05**	**18.63**	**6**
Norway spruce (***Picea abies***) [5]	**20.02**	**19.57**	**6**
White spruce (***Picea glauca***) [3]	**20.20**	**19.76**	**6**
Loblolly pine (***Pinus taeda***) [4]	**22.10**	**21.61**	**7**
Scots pine (***Pinus sylvestris***) [5]	**22.98**	**22.47**	**7**
Siberian pine (***Pinus sibirica***) [6]	**24.15**	**23.62**	**7**
Maritime pine (***Pinus pinaster***) [3]	**24.35**	**23.81**	**7**
Sugar pine (***Pinus lambertiana***) [7]	**29.55**	**28.90**	**9**

[1]All **Pinaceae species** have 12 chromosome pairs except Douglas-fir that has 13 (http://data.kew.org/cvalues); [2]IHGSC 2004; [3]Ohri & Khoshoo 1986; [4]O'Brien *et al*. 1996; [5]Fuchs *et al*. 2008; [6]Siberian pine genome size hasn't been studied yet, and the data are given for the closest species *P. cembra* (Greilhuber 1986); [7]Wakamiya et al. 1993.

The major boreal forest trees in Northern Eurasia



Species	Territory (1000 km²)
Pinus sylvestris	1143.26
Picea abies and *P. obovata*	758.66
Larix spp. (**mostly *Larix sibirica* & *L. gmelinii***)	2633.48
Pinus sibirica (**can be considered as an Alpine species in mountain regions; close relative to *P. cembra***)	397.98
Abies sibirica	143.71
Betula ermanii	83.40
Betula spp.	877.33
Populus tremula	189.08

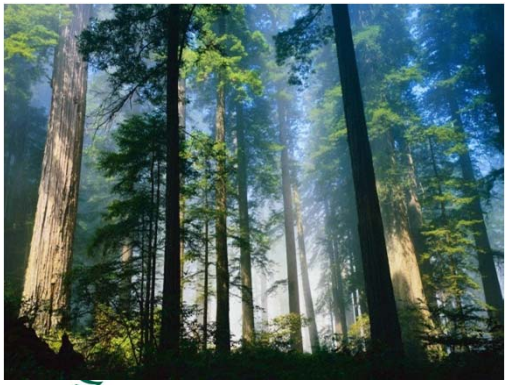
(Isaev *et al.* 1995)

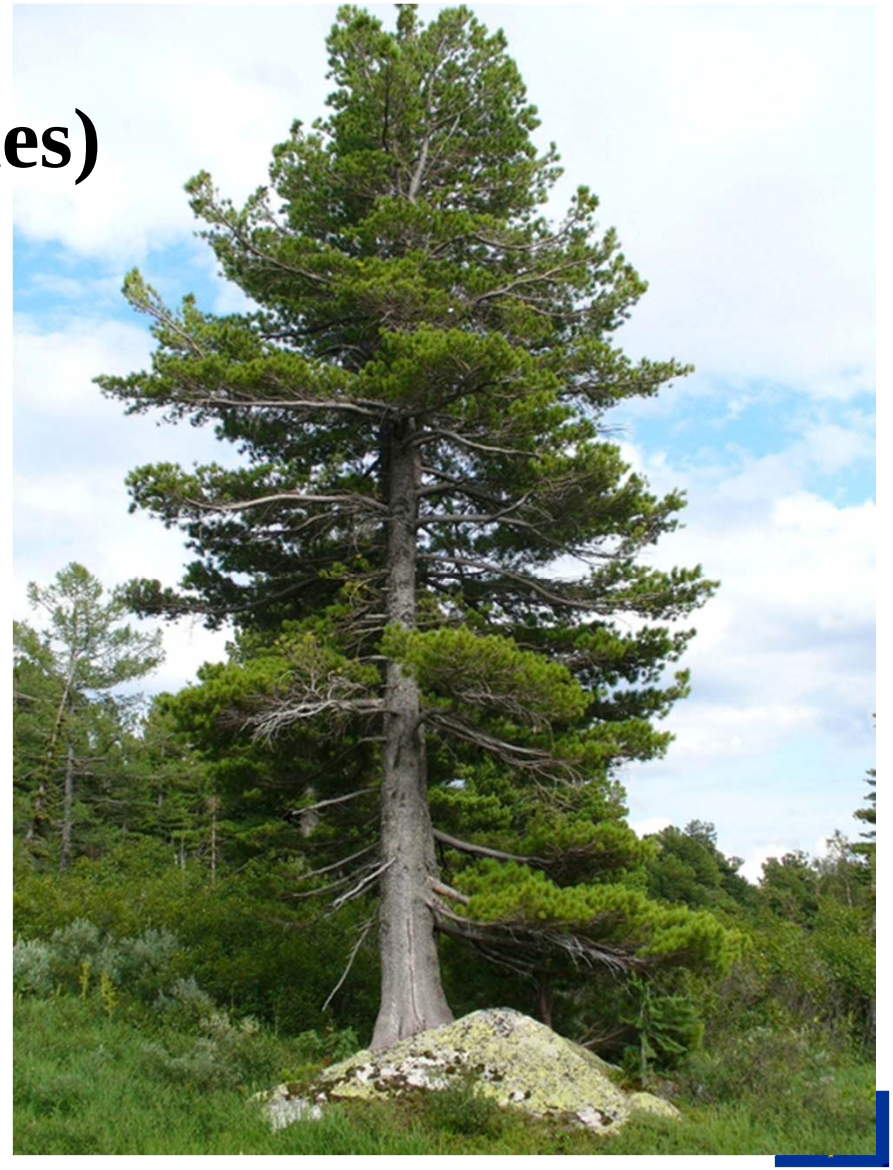
Siberian stone pine (*Pinus sibirica* Du Tour)

<u>Subgenus</u>: *Strobus*
(5-needle, soft or white pines)

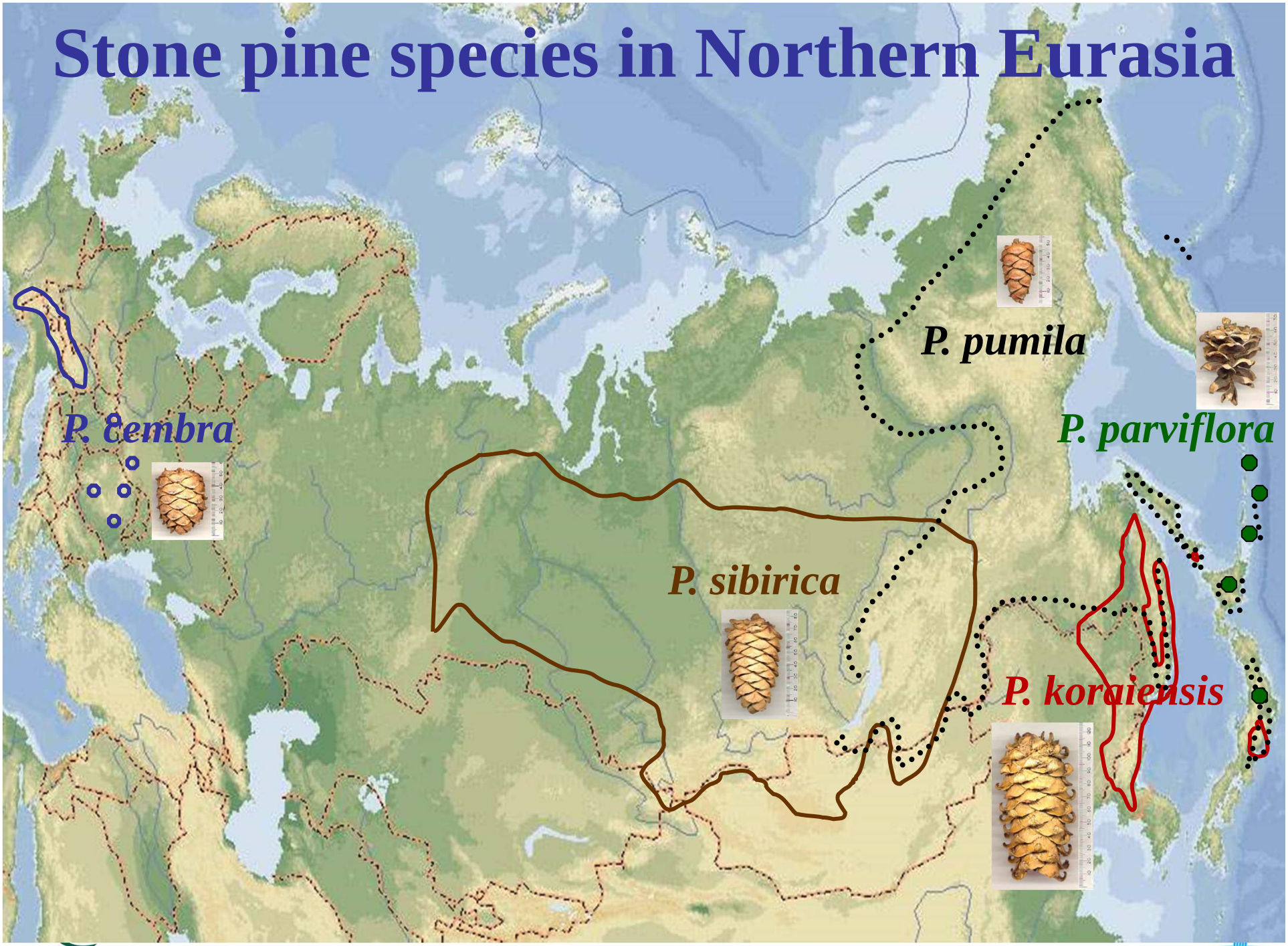
<u>Section</u>: *Strobus*

<u>Subsection</u>: *Cembrae*
(stone pines, 5 species)





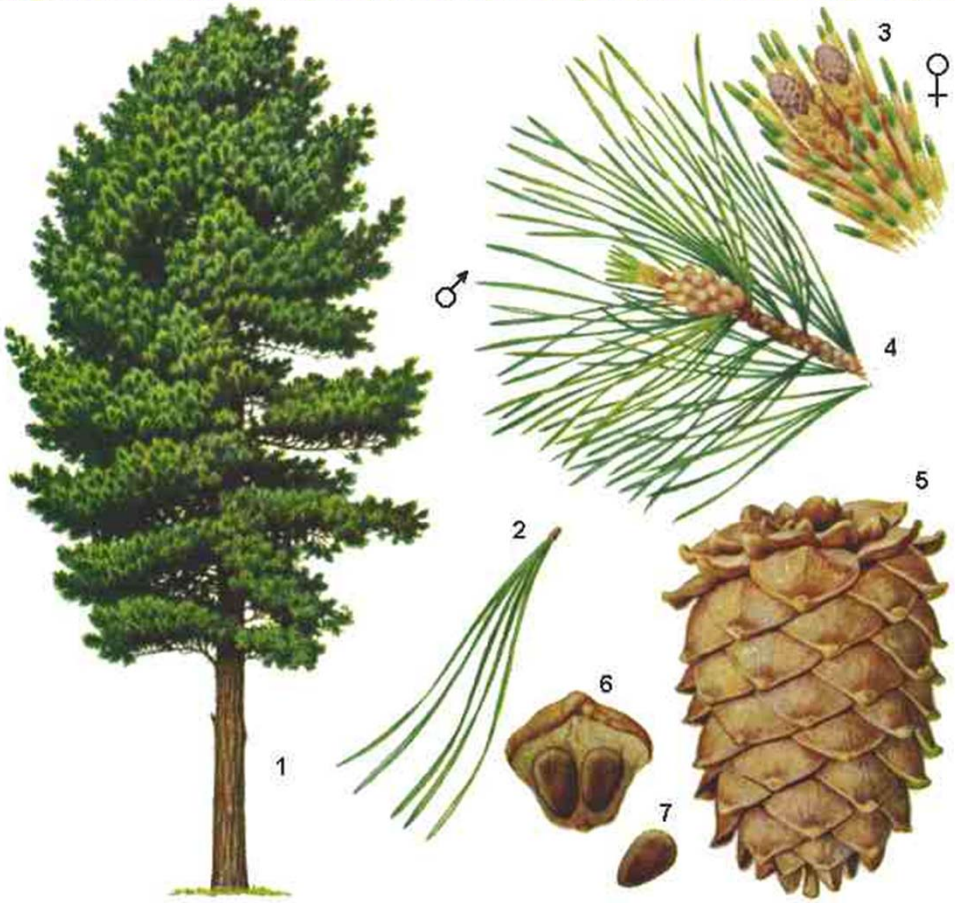
Stone pine species in Northern Eurasia



Pinus sibirica Du Tour (Siberian stone pine)

Pinus sibirica Du Tour (Siberian stone pine)

Pinus sibirica Du Tour (Siberian stone pine)



coevolved with spotted nutcracker (*Nucifraga caryocatactes*)

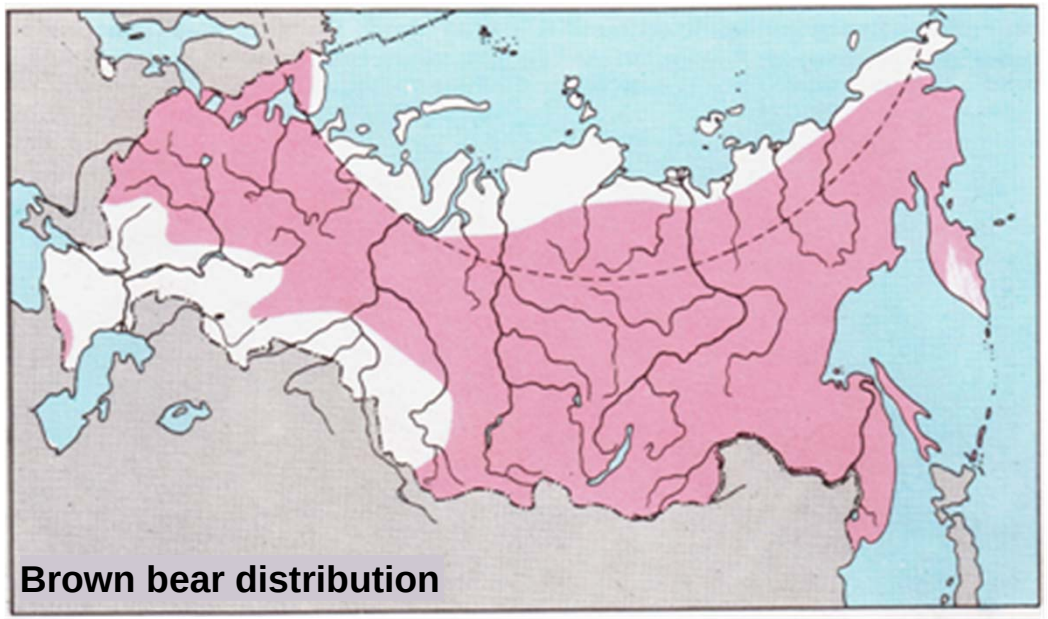
Spotted nutcracker (*Nucifraga caryocatactes* L.)

Pinus sibirica Du Tour (Siberian stone pine)

Siberian chipmunk

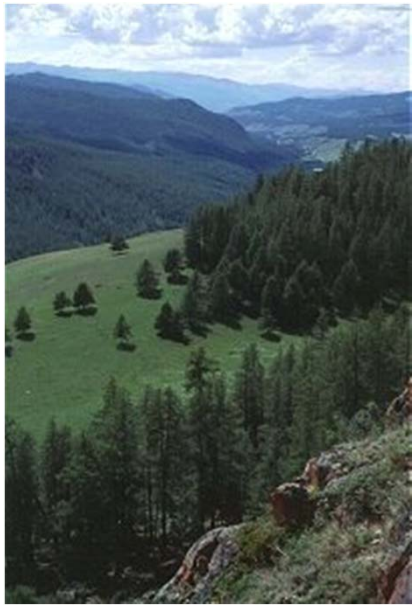
Siberian squirrels

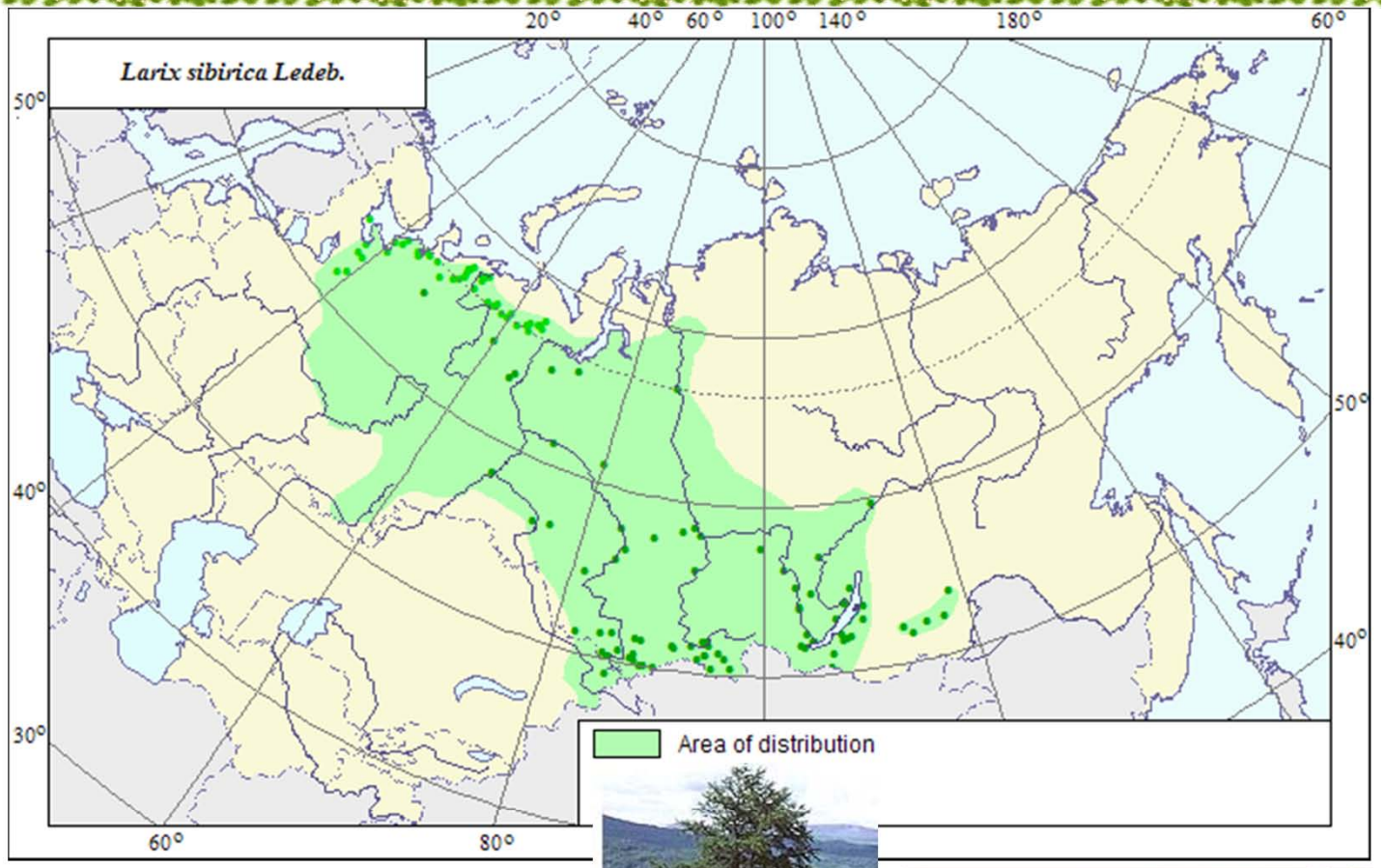
Brown bear (*Ursus arctos*)



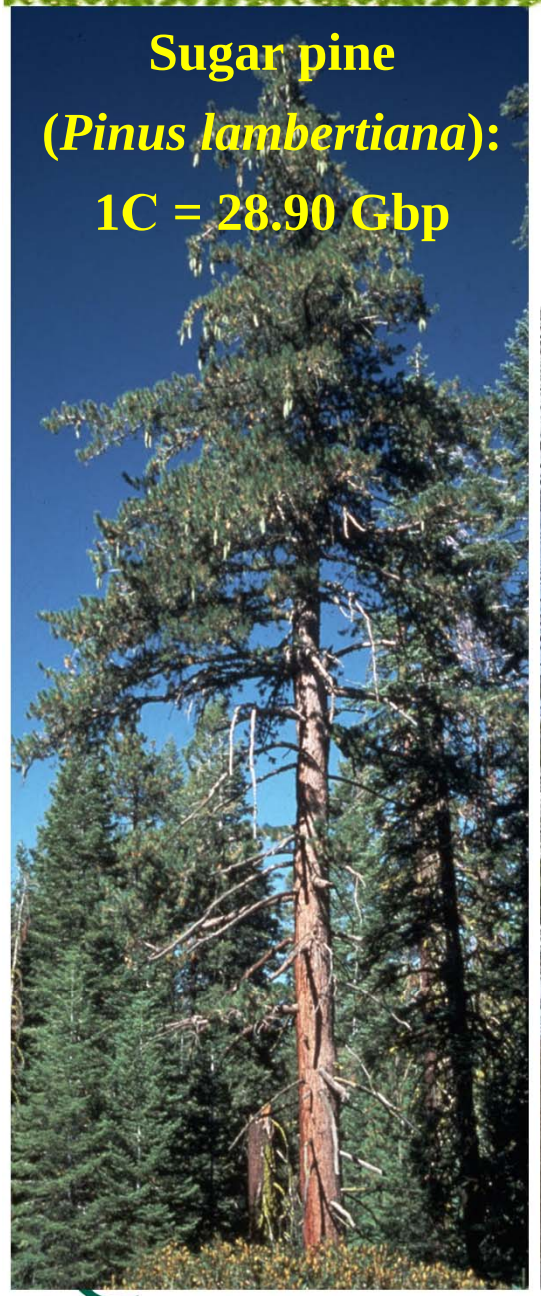
Brown bear distribution

Siberian Larch (*Larix sibirica* Ledeb.)





Scale of the problem: gigantic size of the genome!



Human genome, *Homo sapiens*: 1C = 3.20 Gbp

Other problems

- **highly repetitive** (75-80% of entire genome)
- **high allelic variation**
- **large gene families**

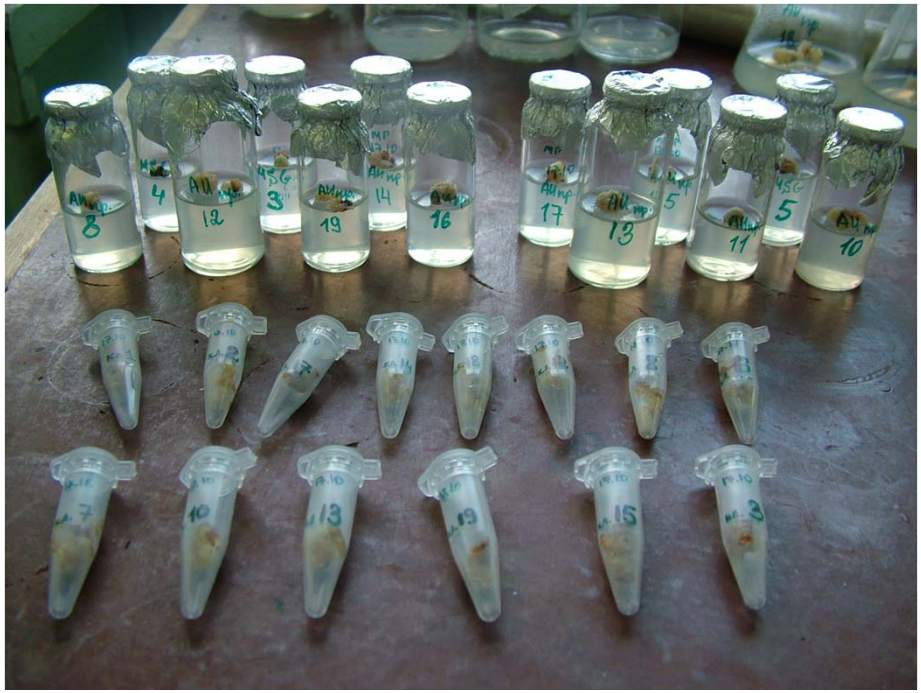
Optimization and innovative approaches are needed!

Modern and prospective innovative approaches

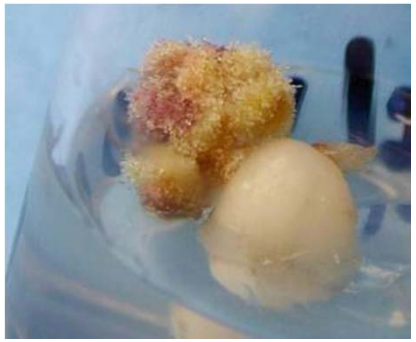
- Combination of different size libraries: 200-800 bp paired-end tags (PET), 2-10 Kb mate paired-end tags (MPET) or paired-end jumping libraries, 454 long reads, barcoded pooled fosmid libraries (30-40 Kb)
- Haploid tissue from a single seed megagametophyte
- New assemblers (i.e., based on de Bruijn graphs, string graphs, etc.)
- Optical mapping (OpGen, Inc.; www.opgen.com)
- Haploid tissue culture
- Genome partitioning:
  - chromosome microdisection using laser capture microscopy (LCM) followed by
  - whole genome amplification (WGA)

Genome complexity reduction via haploid callus

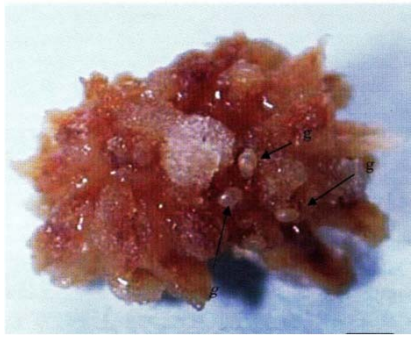
2 week old, from Siberian larch (*Larix sibirica*) megagametophyte of immature seeds (Krutovsky *et al*. 2014)



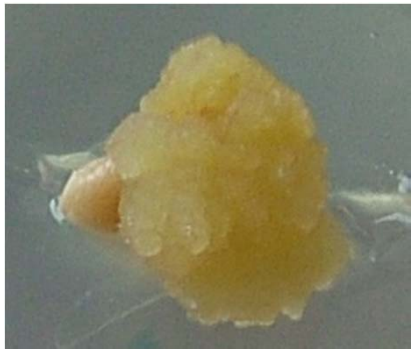
18 lines from Siberian larch (*Larix sibirica*) megagametophyte of immature seeds (Krutovsky *et al*. 2014)



from apical part of the Siberian pine (*Pinus sibirica*) megagametophyte (Tretyakova & Izhboldina 2008)



from Siberian larch (*Larix sibirica*) microstrobili (Tretyakova *et al*. 2006)

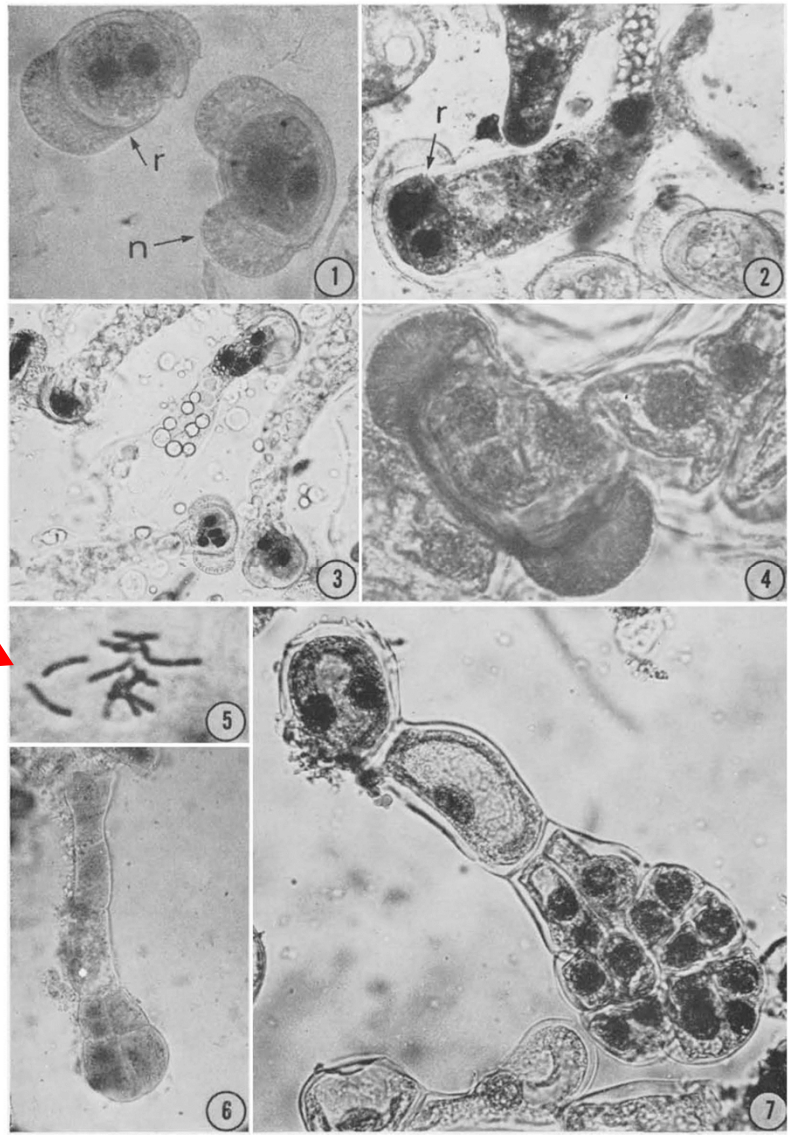


from Siberian larch (*Larix sibirica*) megagametophyte of immature seeds (Krutovsky *et al*. 2014)

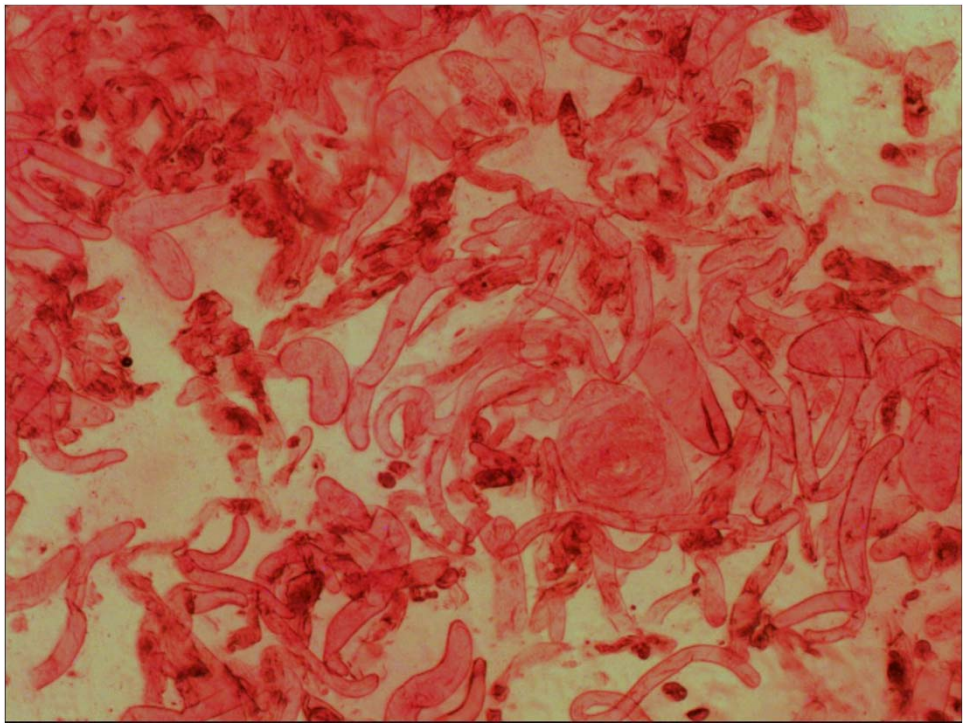
Krutovsky *et al.* (2014) Somaclonal variation of haploid *in vitro* tissue culture obtained from Siberian larch (*Larix sibirica* Ledeb.) megagametophytes for whole genome *de novo* sequencing. *In Vitro Cellular and Developmental Biology – Plant* (http://link.springer.com/article/10.1007%2Fs11627-014-9619-z)

Haploid callus: Cytogenetic data

Haploid metaphase (900X)



1*N*=12



loosely bundled aggregates of long cells that form a suspensor-like callus structure (Krutovsky *et al*. 2014)

Haploid callus: SSR genotyping

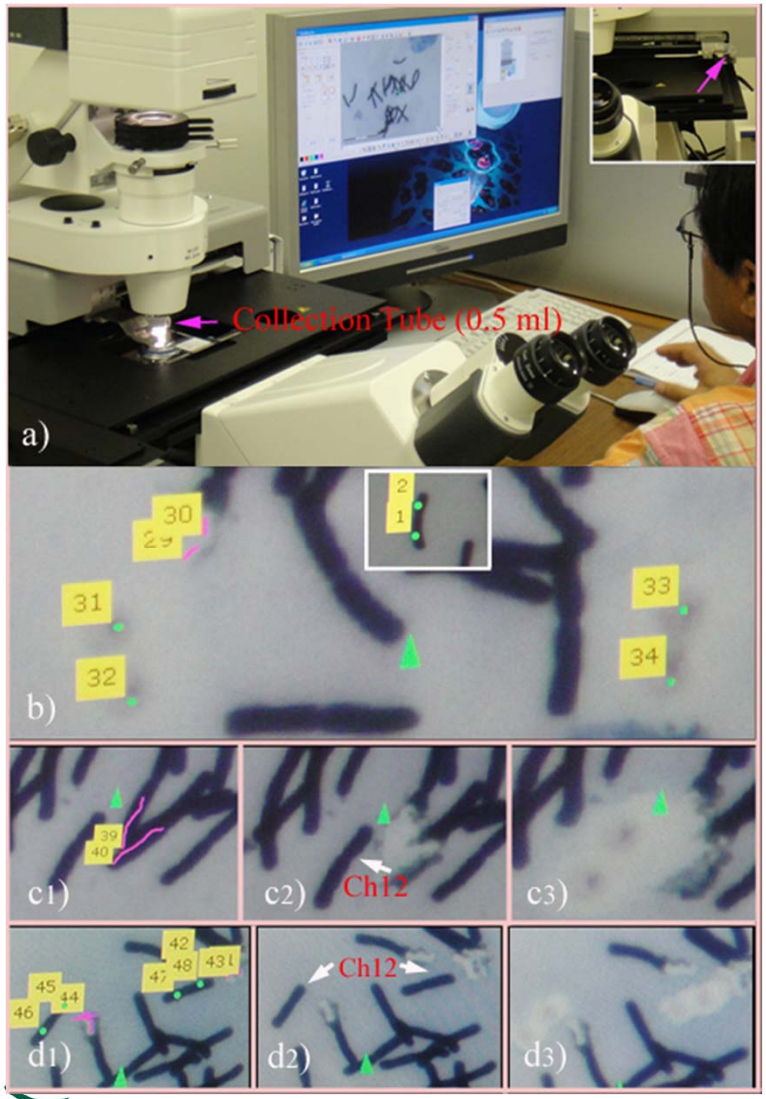
Loblolly pine	*PtRIP_0619*	*PtSIFP_0737*	*PtRIP_0079*	*PtRIP_0968*	*NZPR0143*
Maternal tree 7-56	221/223	449/455	156/175	217/231	124/130
Line 1-3	223	449	156	217	130

Siberian larch	*bcLK232*	*bcLK056*
Maternal tree #3	135/145	146/174
Line 10	135	174
Line 16	135	146
Line 18	145	146
Maternal tree #7	135/145	146/146
Line 2	145	146

Krutovsky *et al*. (2014) Somaclonal variation of haploid *in vitro* tissue culture obtained from Siberian larch (*Larix sibirica* Ledeb.) megagametophytes for whole genome *de novo* sequencing. *In Vitro Cellular and Developmental Biology – Plant* (http://link.springer.com/article/10.1007%2Fs11627-014-9619-z)

Genome complexity reduction via partitioning

Isolation and microdissection of individual chromosomes and their fragments using mitotic chromosome slides and a laser capture microscopy (LCM)



<u>Microdissection of individual chromosome №12 of *Pinus taeda* using Zeiss Laser Capture Microscopy (LCM) by Dr. Nurul Islam-Faridi at Texas A&M University</u>:

a) LCM microscope, the arrow is pointing at the µ-tube for sample collection (0.5 ml);

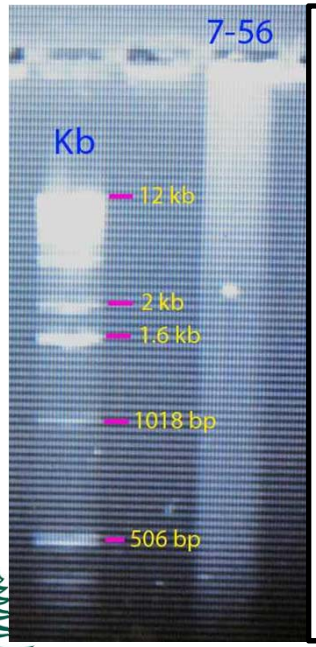
b) chromosome № 12 is marked in the white square by two green dots and yellow numbers 1 and 2, its two copies, marked correspondingly by the numbers 31, 32, and 33, 34, have been already extracted (catapulted), and the nearest other chromosome arms, marked by numbers 29 and 30, have been eliminated to avoid accidental contamination by other chromosomes;

c) and d) the adjacent chromosomes have been eliminated before the extraction of the12th chromosome.

Amplification of individual chromosomes or their fragments

Amplification of the whole DNA of individual chromosomes or their fragments can be done using the whole genome **amplification (WGA) methods, such as** developed recently by the companies:

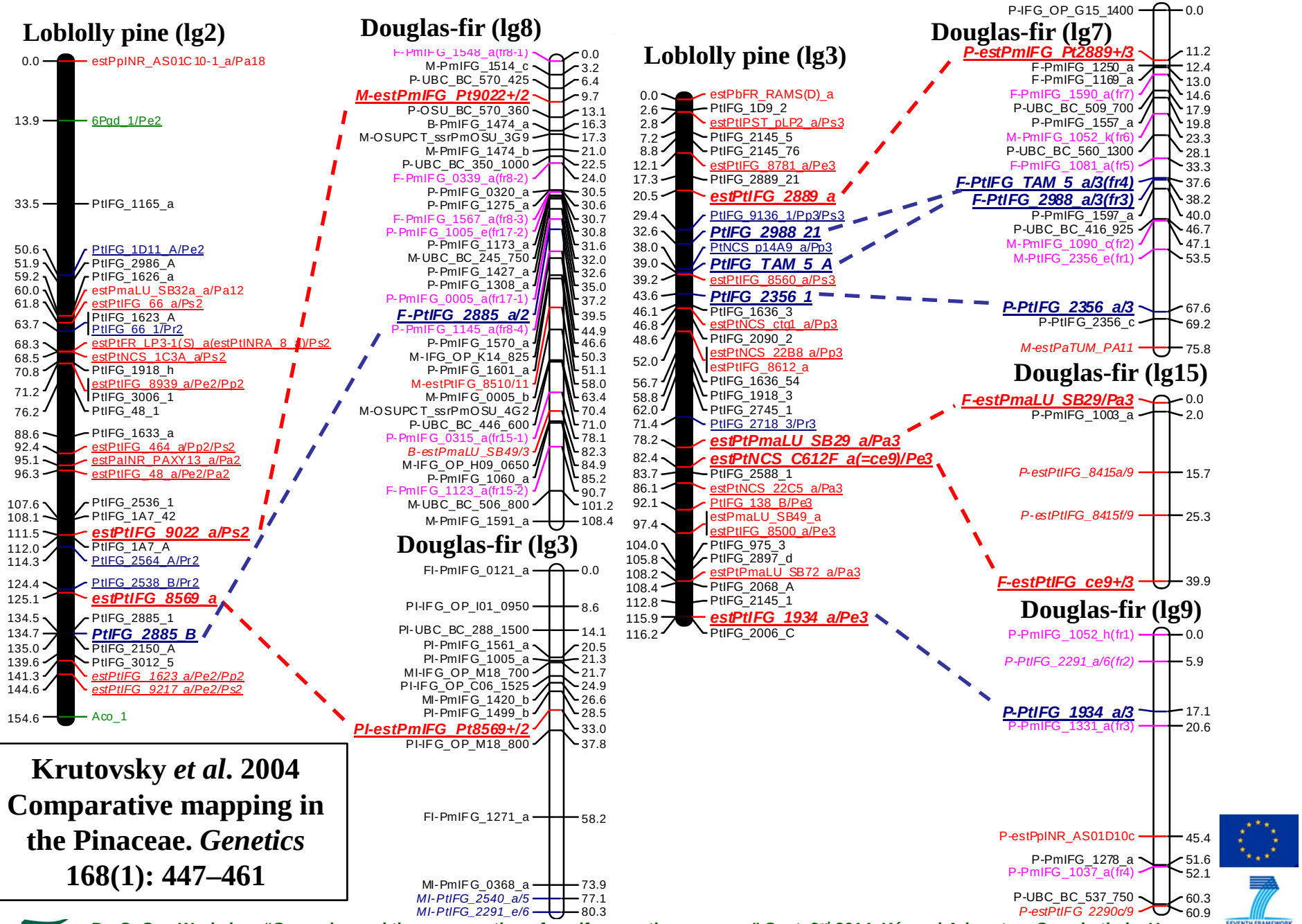
- **Rubicon Genomics**, Inc., Ann Arbor, MI, USA (the Rubicon Genomics PicoPlex NGS WGA Kit; http://www.rubicongenomics.com/products/picoplexngs), and
- **Sigma-Aldrich Co.** (St. Louis, MO, USA), modified for individual chromosome amplification (the GenomePlex Single Cell Kit; http://www.sigmaaldrich.com/life-science/molecular-biology/whole-genome-amplification.html).
- Promising results on pine individual chromosome amplification using this method have already been obtained recently.



Electrophoresis in 1.5% agarose gel of DNA amplified from sample №7-56 DNA extracted from LCM isolated individual chromosome №12 of *Pinus taeda* using the GenomePlex® Single Cell Whole Genome Amplification Kit.

The left lane contains DNA size standards (75-1500 bp); the right lane contains 4 µl WGA reaction of amplified chromosome #12 DNA.

Syntenic linkage groups in Douglas-fir and Loblolly pine



Krutovsky *et al.* 2004 Comparative mapping in the Pinaceae. *Genetics* 168(1): 447–461

Whole genome sequence data

- ***Larix sibirica*** (12.03 Gbp):

  needles and a haploid tissue culture

  PE libraries: 180, 250, 400 and 500 bp

  MPE library: 5 Kbp

  total good quality **576 Gbp (48X)**

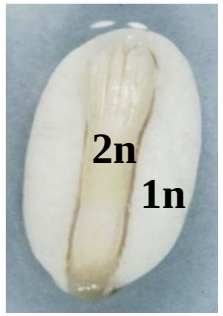
- ***Pinus sibirica*** (23.62 Gbp):

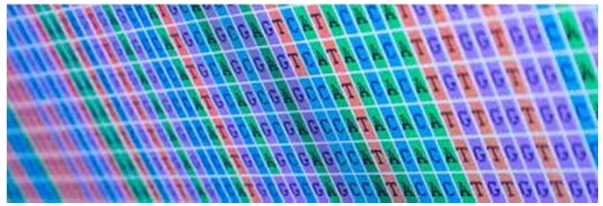
  a single megagametophyte

  PE libraries: 250 and 500 bp

  MPE libraries: 3 and 5 Kbp

  total good quality **679 Gbp (29X)**



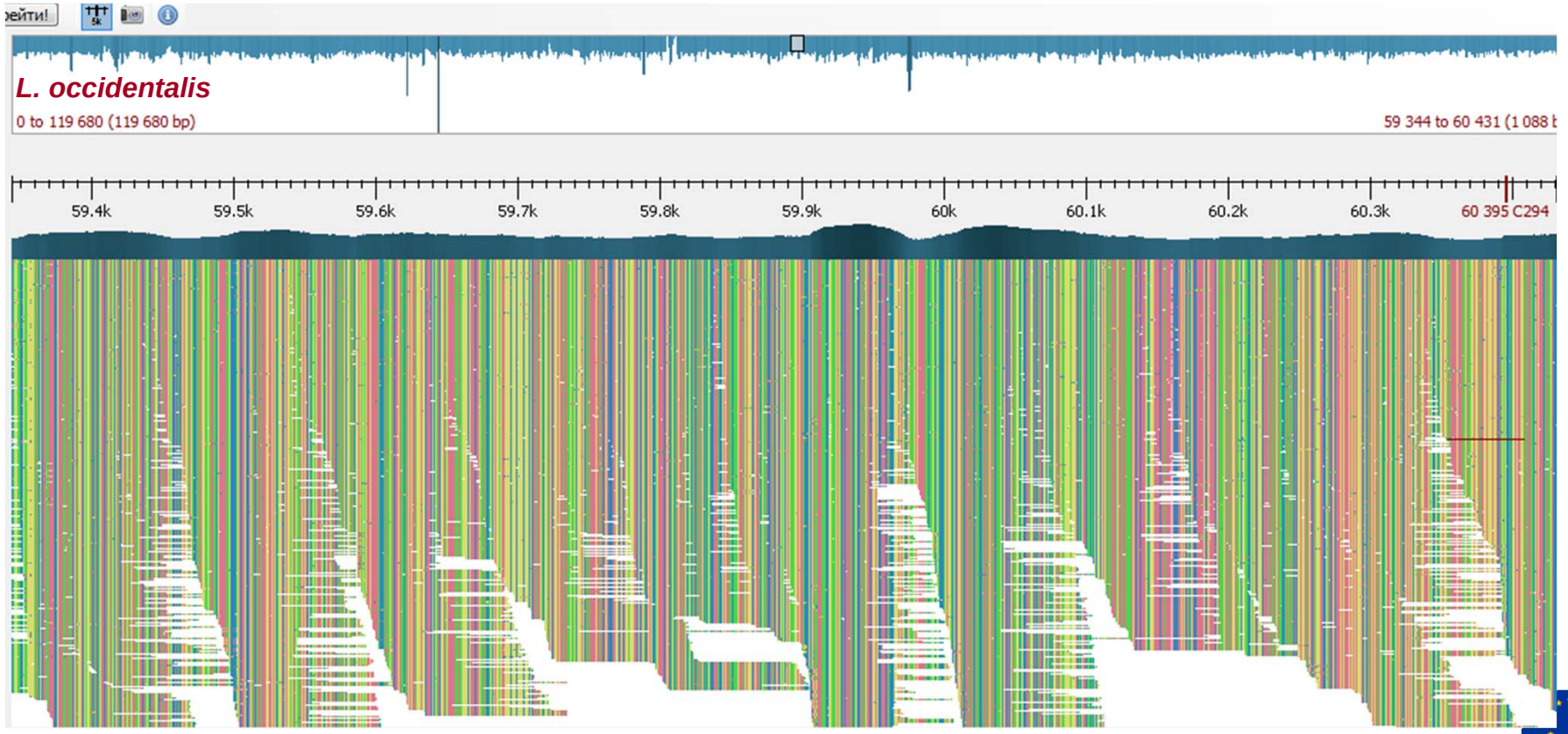


Microsatellite loci in Siberian larch (*Larix sibirica*)

Motif, bp	Loci selected	Mean locus length, bp	Per 1 Gbp
2	563	20.6	28,570
3	140	31.8	7,100
4	7	53.9	360
5	3	60.0	150
6	10	59.9	510

Mapping the Siberian larch and pine reads to the larch and pine chloroplast genomes available in Genbank

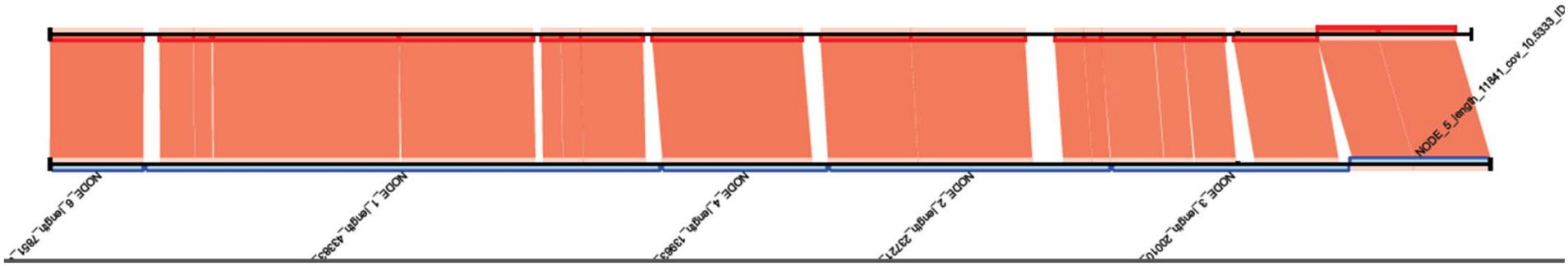
Larix	*Pinus*
L. decidua (NC_016058, 122,474 bp, complete) *L. occidentalis* (FJ899578, 119,680 bp, incomplete)	*P. sibirica* (116,593 bp, incomplete) *P. cembra*, *P. pumila*, *P. koraiensis*, etc.



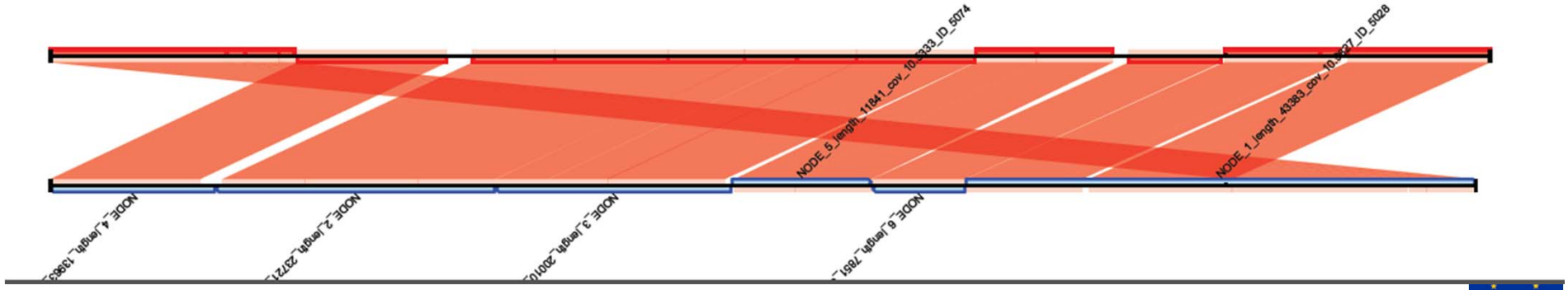
(using Bowtie, Bowtie2, and BWA)

Mapping the Siberian larch reads to the European and Asian larch chloroplast genomes available in Genbank

L. decidua (122,474 bp, complete)



L. occidentalis (119,680 bp, incomplete)

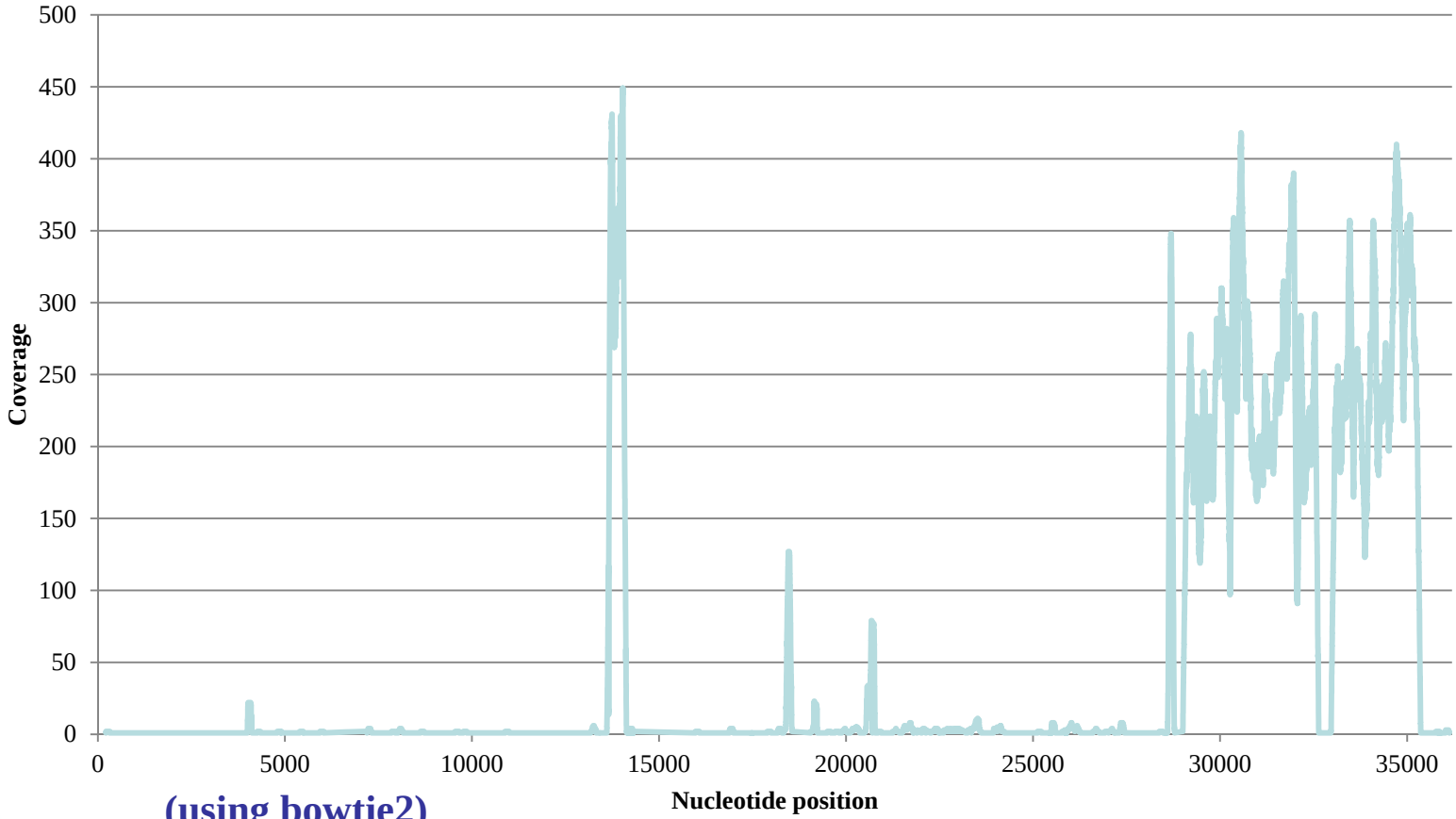


Mapping the Siberian pine reads to the Loblolly pine mitochondrial genome assembly

P. taeda **mitochondrial genome assembly (loblolly.ucdavis.edu/bipod/ftp/Genome_Data/genome/pinerefseq/Pita):**

Scaffolds	Contigs	Total length, bp	Max length, bp	Min length, bp	N50, bp	N90, bp
4	31	1,263,957	256,879	124	193,087	58,893

P. sibirica **reads mapped to the** ***P. taeda*** **scaffold 2 (36,185 bp, 46X)**



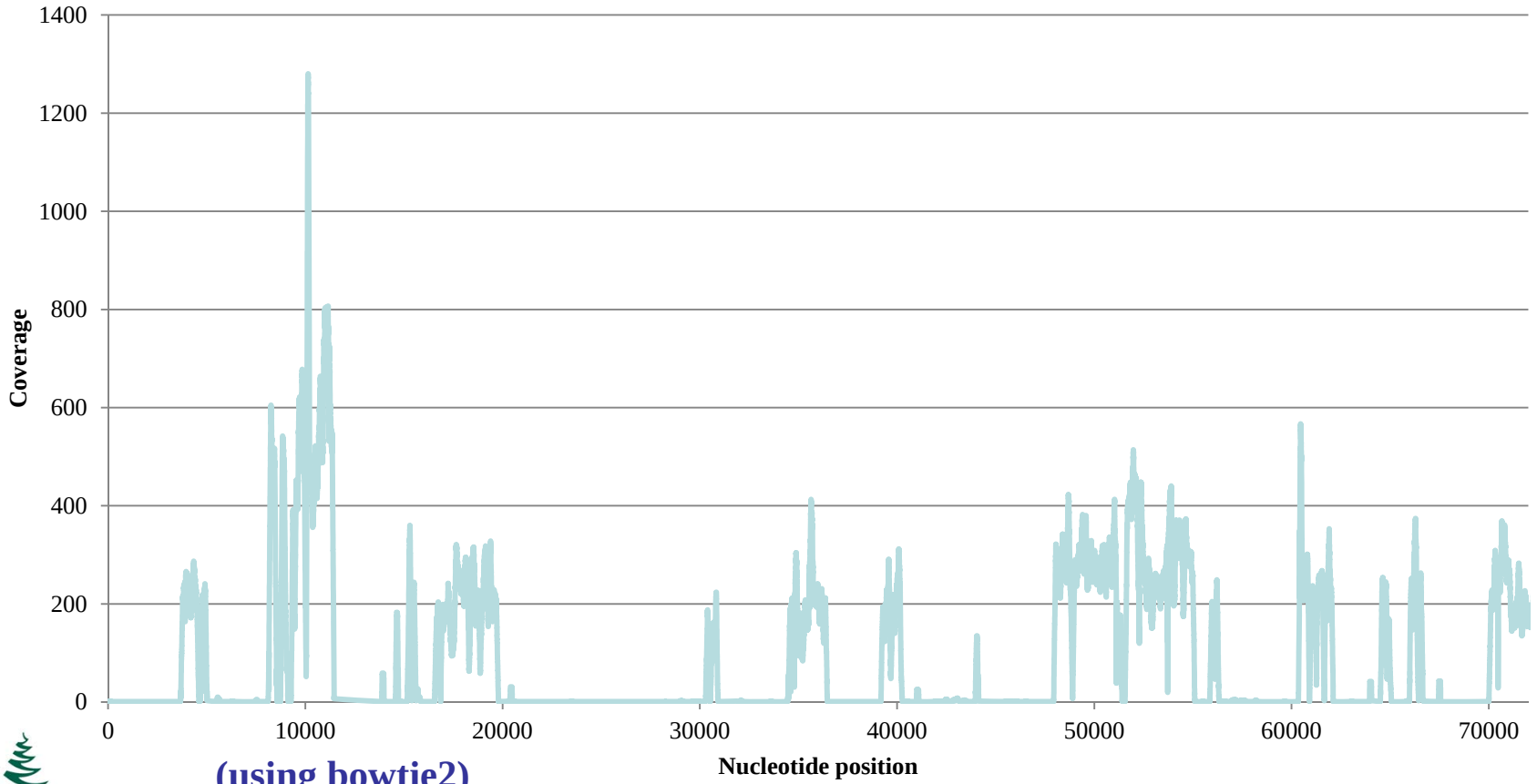
(using bowtie2)

Mapping the Siberian pine reads to the Loblolly pine mitochondrial genome assembly

P. taeda **mitochondrial genome assembly (loblolly.ucdavis.edu/bipod/ftp/Genome_Data/genome/pinerefseq/Pita):**

Scaffolds	Contigs	Total length, bp	Max length, bp	Min length, bp	N50, bp	N90, bp
4	31	1,263,957	256,879	124	193,087	58,893

P. sibirica **reads mapped to the** ***P. taeda*** **scaffold 4 (76,864 bp, 90X)**



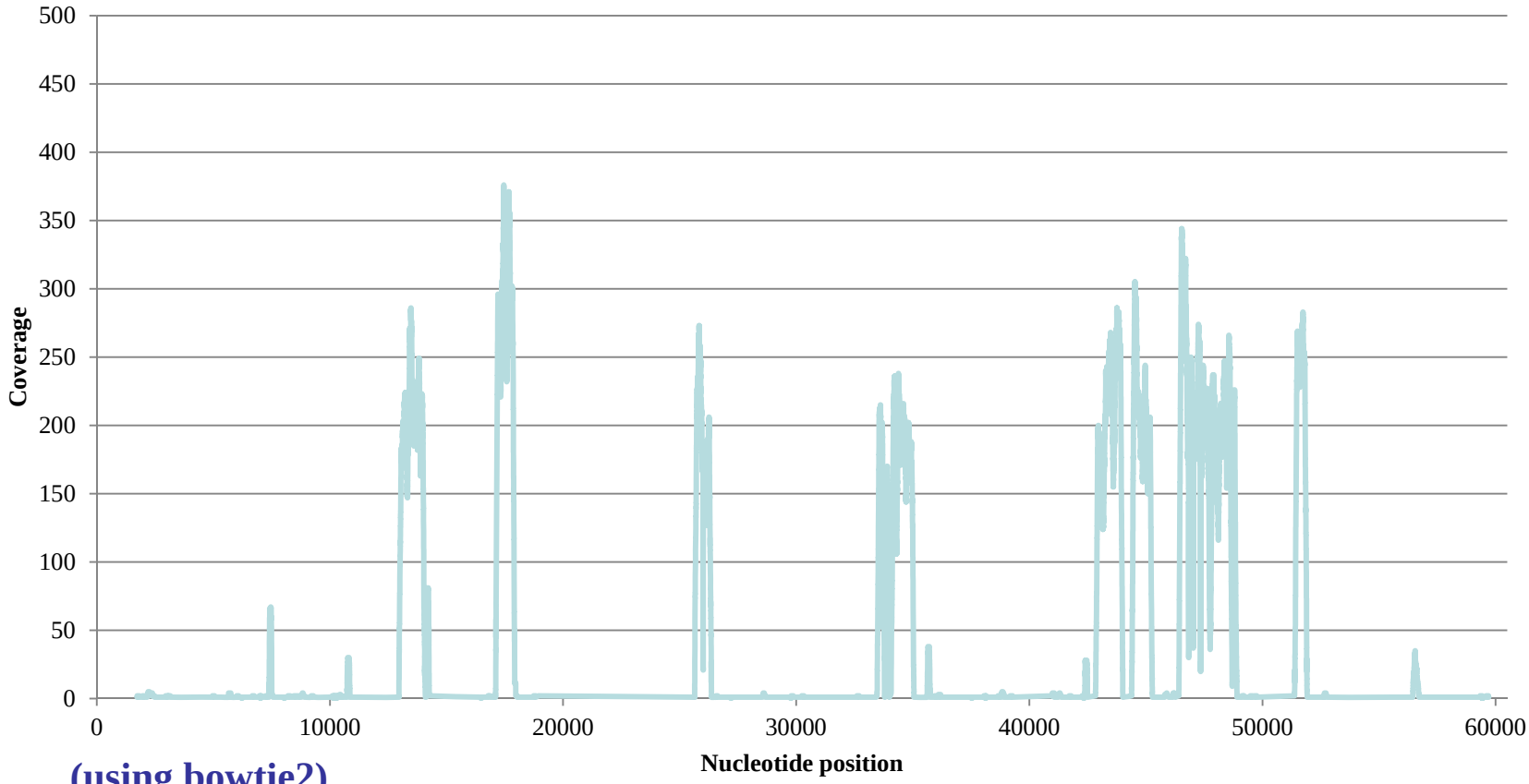
(using bowtie2)

Mapping the Siberian pine reads to the Loblolly pine mitochondrial genome assembly

P. taeda **mitochondrial genome assembly (loblolly.ucdavis.edu/bipod/ftp/Genome_Data/genome/pinerefseq/Pita):**

Scaffolds	Contigs	Total length, bp	Max length, bp	Min length, bp	N50, bp	N90, bp
4	31	1,263,957	256,879	124	193,087	58,893

P. sibirica **reads mapped to the *P. taeda* contig 67 (60,321 bp, 30X)**



(using bowtie2)

Total coverage for all reads - 81X

Complex plant mitochondrial genomes

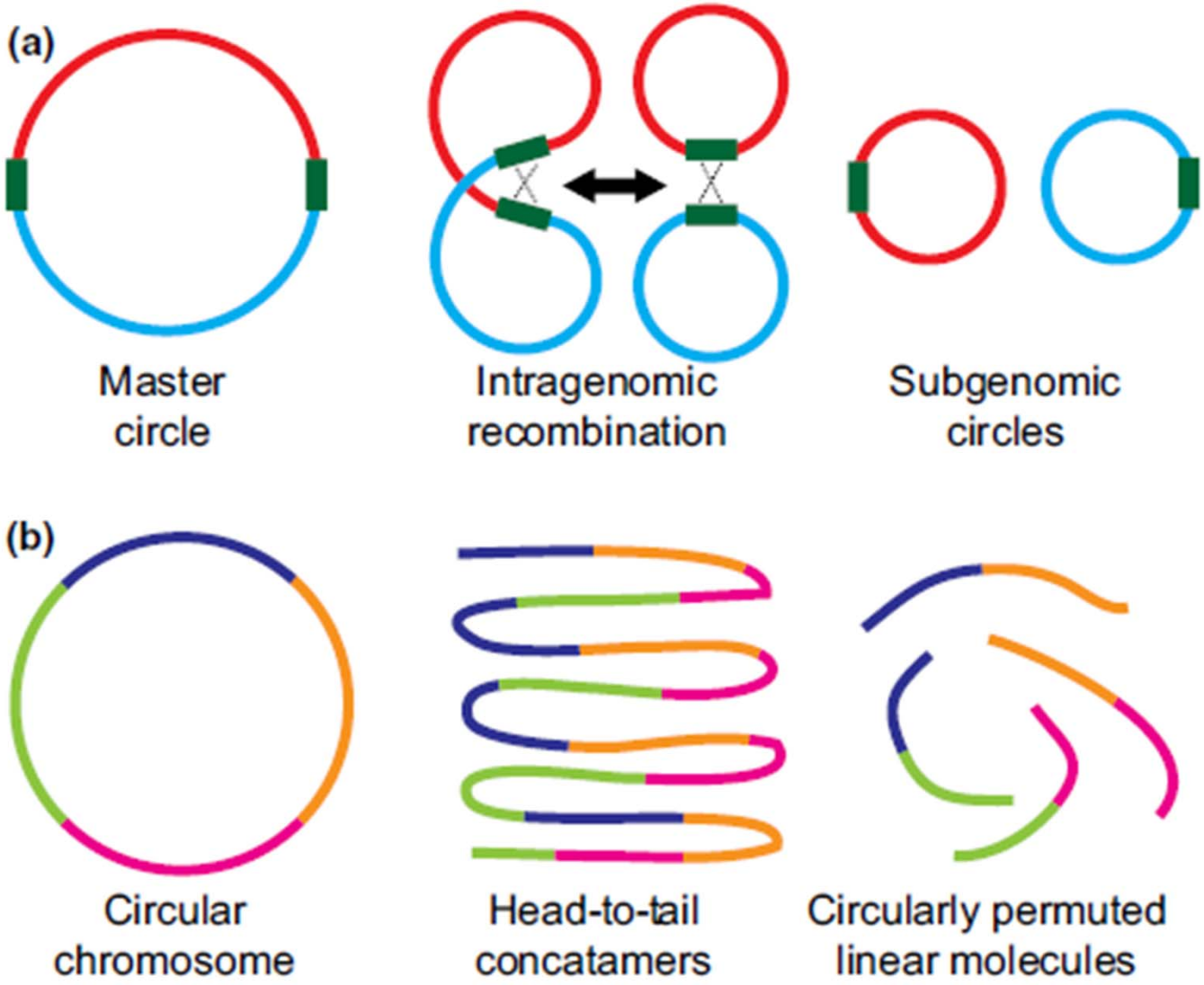
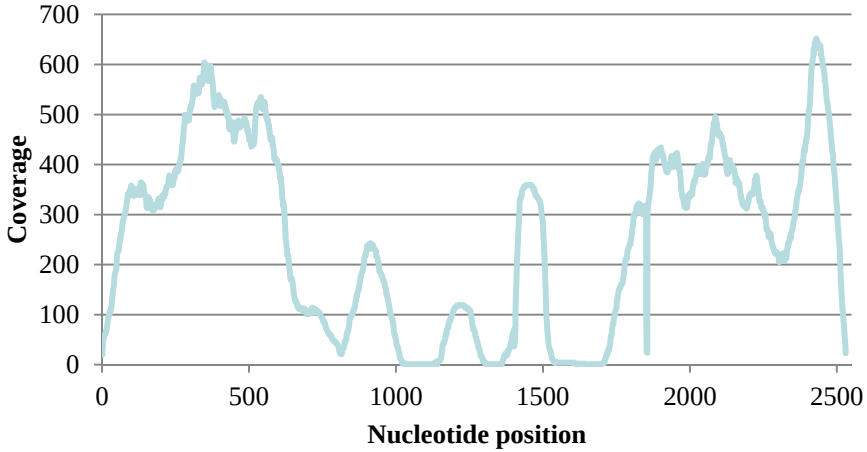


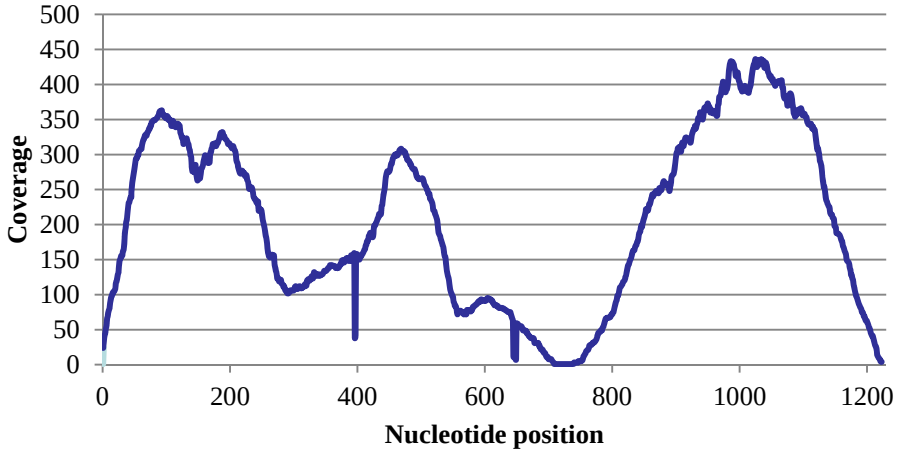
Fig. 1 in Sloan (2013) *New Phytologist* 200: 978–985

Mapping the Siberian pine reads to the Siberian pine genes in Genbank

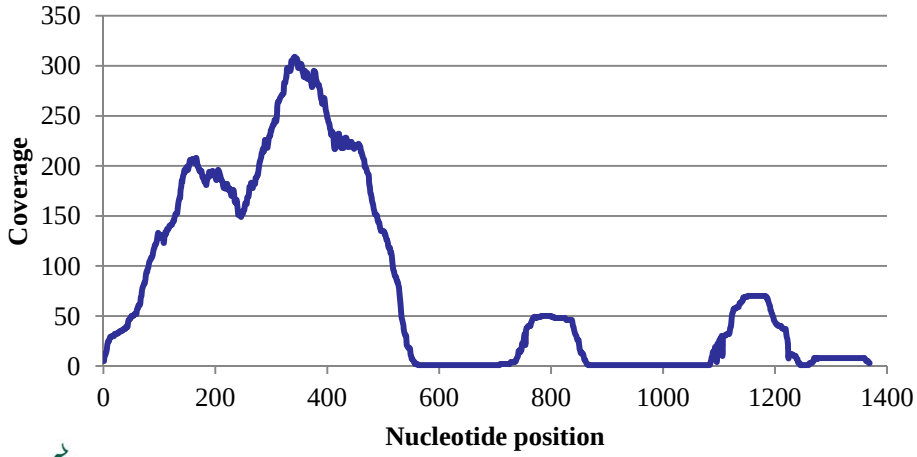
nad1



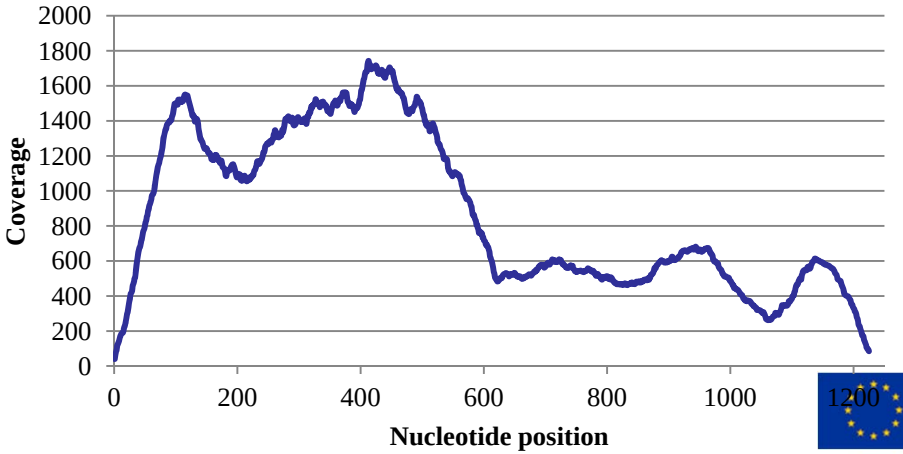
nad5



nad7



cox1



(using bowtie2)

Mapping the Siberian pine reads to the Loblolly pine mitochondrial genome assembly

P. taeda **mitochondrial genome assembly (loblolly.ucdavis.edu/bipod/ftp/Genome_Data/genome/pinerefseq/Pita):**

Scaffolds	Contigs	Total length, bp	Max length, bp	Min length, bp	N50, bp	N90, bp
4	31	1,263,957	256,879	124	193,087	58,893

Picea abies **mitochondrial genome assembly (dl.dropboxusercontent.com/u/8576950/putative_mitochondrial_scaffold.fasta.zip):**

Scaffolds	Contigs	Total length, bp	Max length, bp	Min length, bp	N50, bp	N90, bp
?	105	2,400,824	78,675	?	30,157	12,361

Species	Reads mapped to *Pinus taeda* mitochondrial genome, %	Reads mapped to *Picea abies* mitochondrial genome, %
P. sibirica	0.93	0.63
P. sylvestris	3.57	0.84
L. sibirica		0.12
A. sibirica	0.22	0.33

Mitochondrial genome assemblies for 4 conifer species in our study:

Species	Contigs	Total length, bp	Max length, bp	N50, bp	N90, bp
P. sibirica	431	482,064	9,882	2,274	398
P. sylvestris	2,586	1,190,450	19,705	2,003	177
A. sibirica	248	147,965	3,432	1,031	240
L. sibirica	209	69,615	1,039	333	227

Conclusions

- Haploid tissue culture can be generated, but its use for whole genome *de novo* sequencing is still questionable due to high mutation rates

- LCM approach is feasible, but needs additional testing

- High chromosomal conservatism and synteny in conifers should facilitate genome assembly across different species

Long term objectives

- Complete genome sequence, assembly and annotation
- Discover all expressed genes via RNA-seq of multiple tissues at different developmental stages and after different treatments
- Landscape and ecological genomics: find associations of SNPs, alleles, haplotypes, and genotypes with environmental factors, adaptive traits and phenotypes
- Genomic Selection (selection based on genome wide genotyping): use genome wide markers to infer kinship relationships and to develop a regression model between markers and phenotypes in a training population and apply it to other breeding populations to predict and select the best performing prospective trees

Acknowledgements

- Siberian Federal University

- Ministry of Science and Education, Russian Federation
- Russian Foundation for Basic Research
- Russian Government

Acknowledgements

Dr. Iraida N. Tretyakova
Dr. Nataliay V. Oreshkova

Dr. Eugene A. Vaganov
Dr. Yuliya A. Putintseva

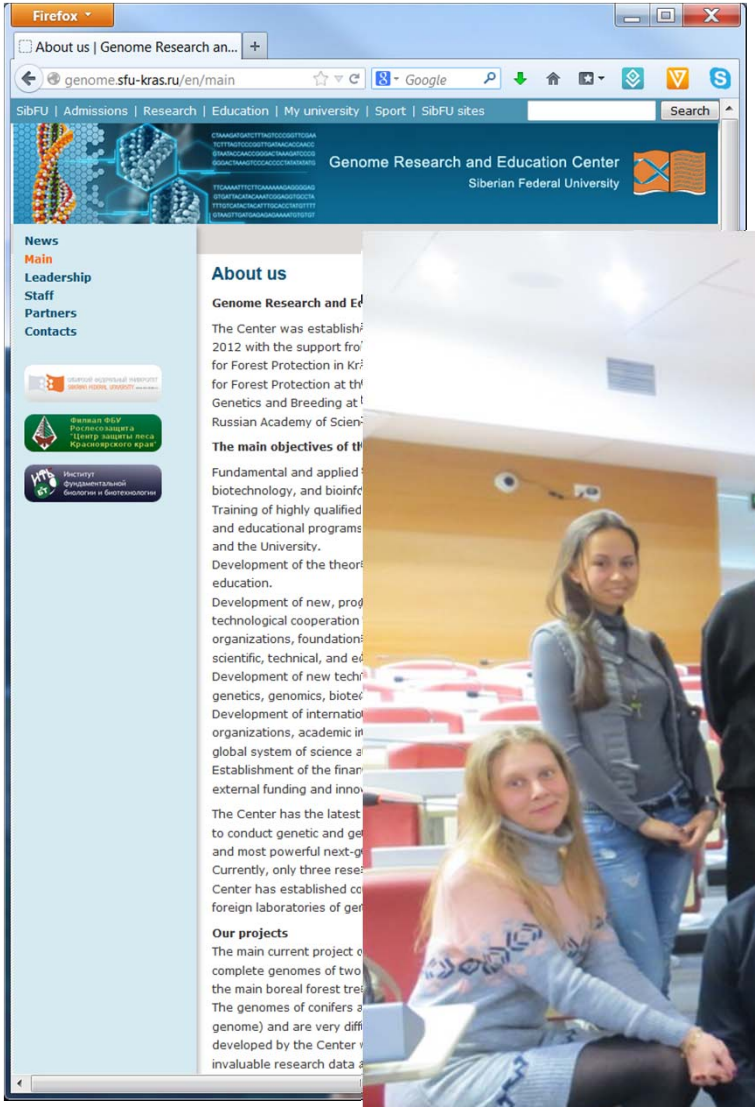
Dr. Vladimir V. Soldatov
Dr. Irina V. Chubugina
Dr. Elena A. Shilkina
Dr. Alexey A. Ibe
Kseniya O. Deich

Dr. Nikolai K. Yankovsky

Acknowledgements

Genome Research and Education Center at the Siberian Federal University in Krasnoyarsk

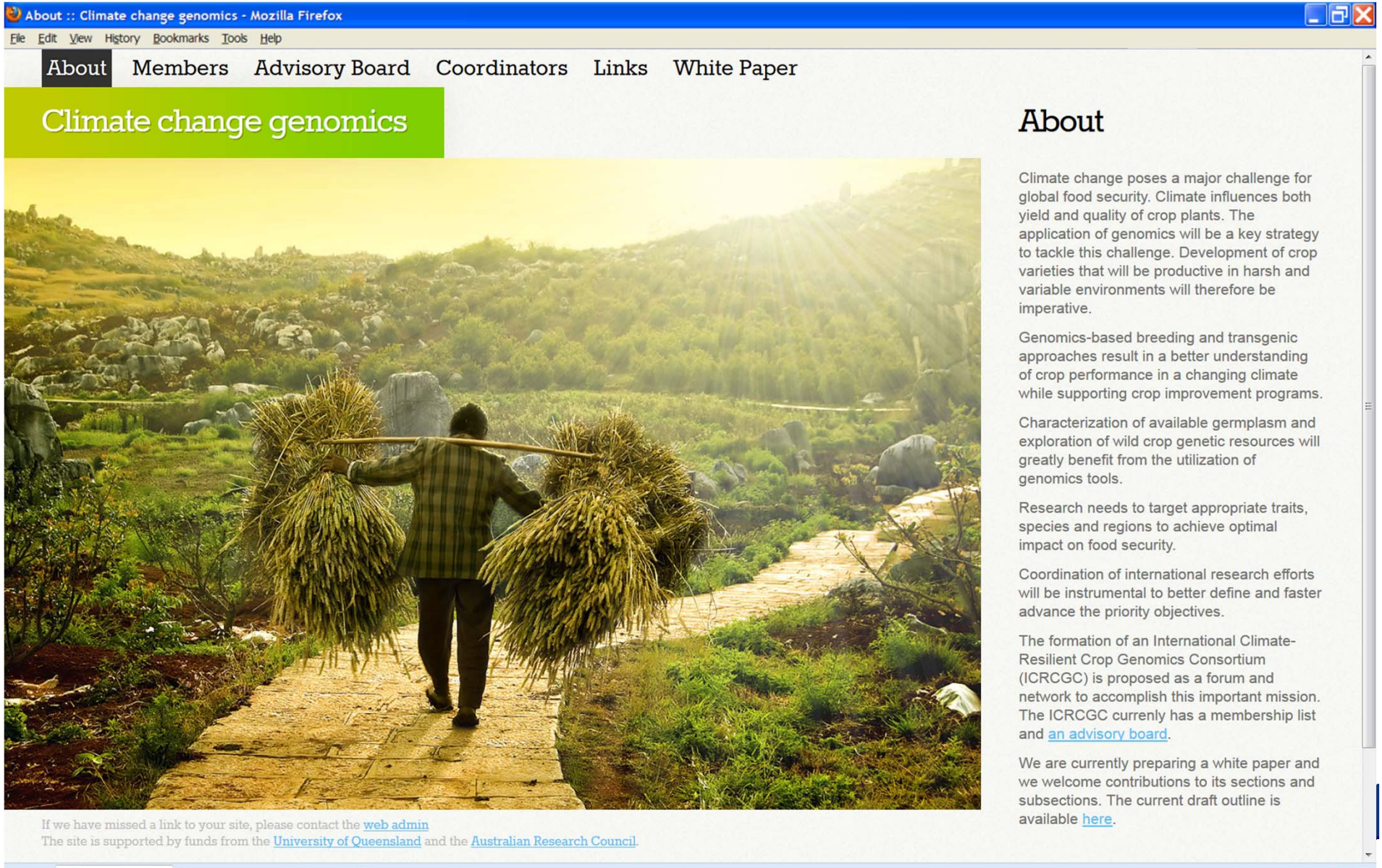
http://genome.sfu-kras.ru/main



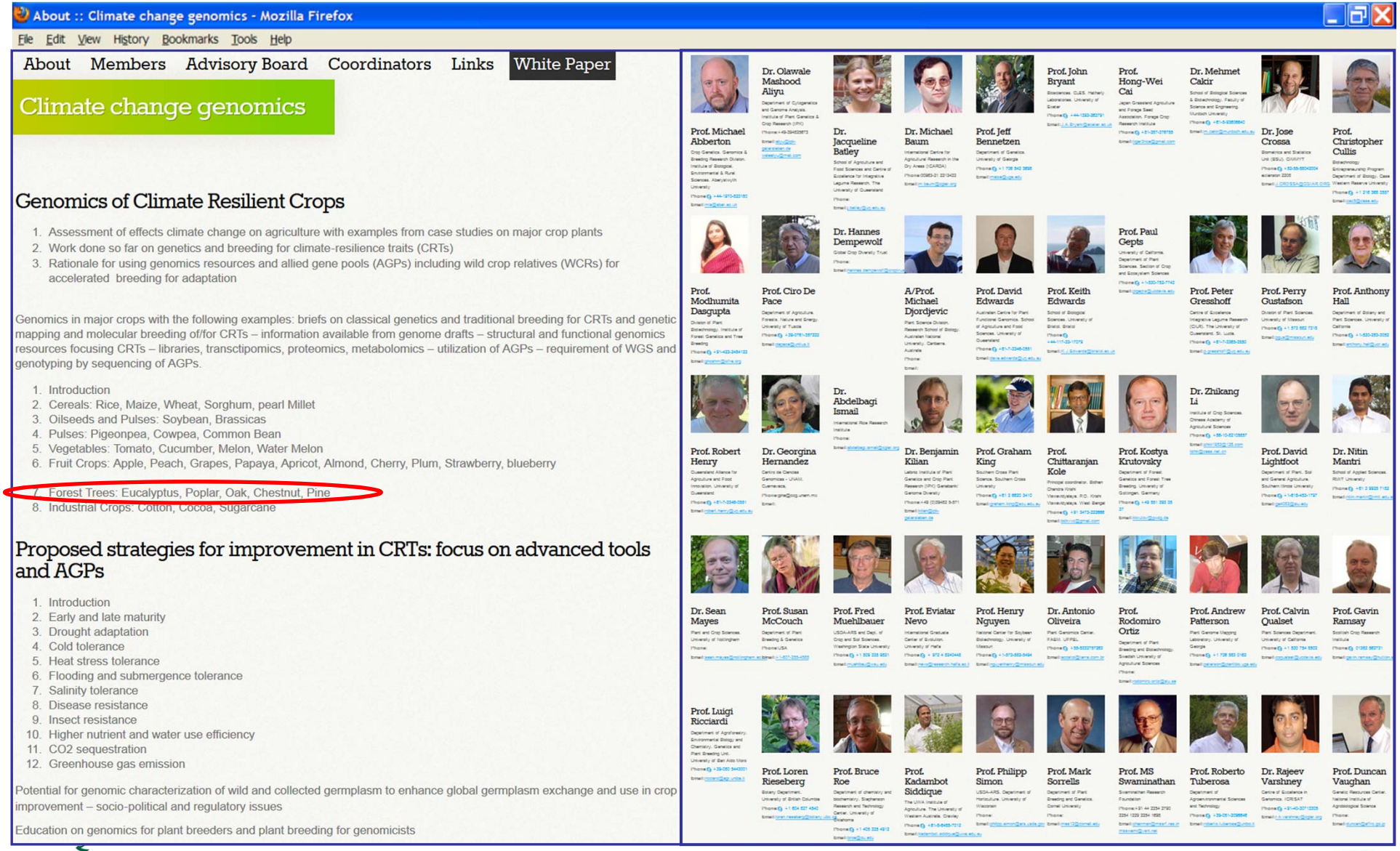
Thank you for your attention!

The Bastrop State Park “Lost Pines” in Central Texas

The International Climate-Resilient Crop Genomics Consortium (ICRCGC) http://www.climatechangegenomics.org



The International Climate-Resilient Crop Genomics Consortium (ICRCGC) http://www.climatechangegenomics.org/members.php



Assisted migration

- **To mitigate climate change and** to help populations cope with climate change genotype-environment controlled assisted migration is considered and should be more studied

  ✓Do we need assisted migration – yes or no?

  ✓Are we ready for this – yes or no?

- The International Climate-Resilient Crop Genomics Consortium might help to raise awareness and funds

Krutovsky et al. 2012 Gene flow, spatial structure, local adaptation and assisted migration in trees, pp. 71-116, Ch. 4 in *Genomics of Tree Crops*, edited by R.J. Schnell and P.M. Priyadarshan. Springer Science, Inc.

- **XX century:**

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

↓

population genetics level of thinking

Theodosius Dobzhansky (1900-1975)

- **XXI century:**

Molecular genetics + Bioinformatics
= Genomics

↓

population 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