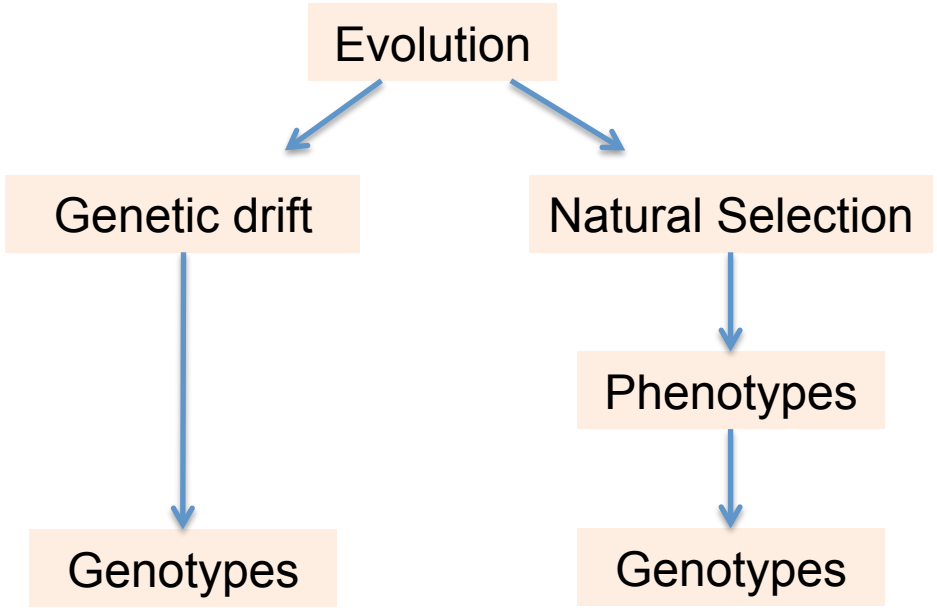
Levels and patterns of nucleotide variability and population differentiation in *Populus*:

insights from transcriptome resequencing and SNP genotyping

Armando Geraldes

PAGXX, San Diego, January 14th 2012

NUCLEOTIDE POLYMORPHISM



POPULUS TRICHOCARPA

Native to western North America

High latitudinal range (California to Alaska)

High growth rate even in marginal lands

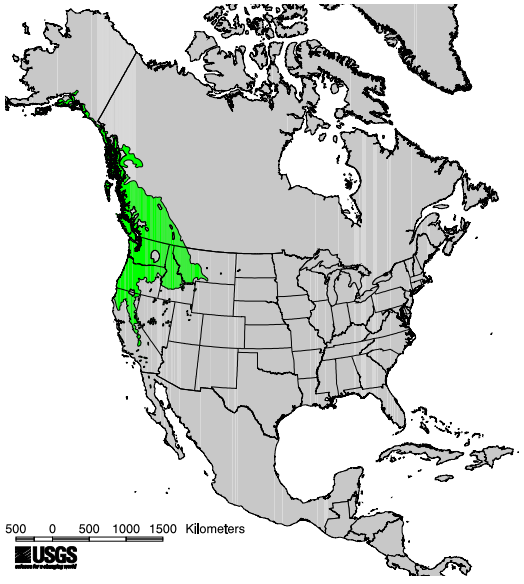
Promising Biofuels source

High phenotypic variability

Genome sequence published in 2006

Large natural accession collections

Several ongoing association studies



QUESTIONS

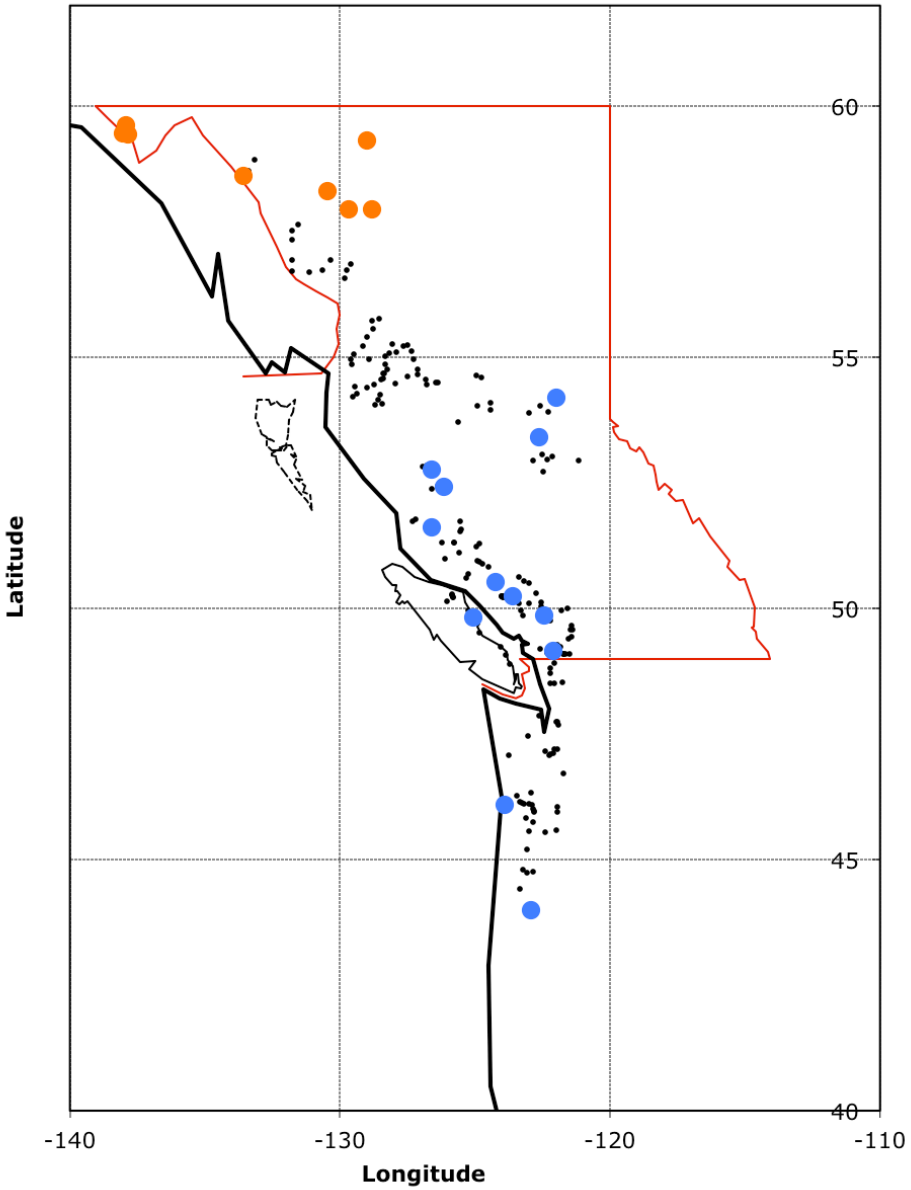
i) What is **the overall level of nucleotide polymorphism** in *P. trichocarpa?*

ii) How is polymorphism **partitioned geographically**?

iii) How is polymorphism **partitioned across different classes of sites**?

iv) How does it compare to **other species in the genus**?

M&M: SAMPLING



- *P. trichocarpa*:
  - Transcriptome n=20
  - SNP genotyping n=745
- *P. balsamifera*:
  - Transcriptome n=3
  - SNP genotyping n=10
- *P. deltoides*:
  - Transcriptome n=1
- *P. tremula*:
  - Transcriptome n=4

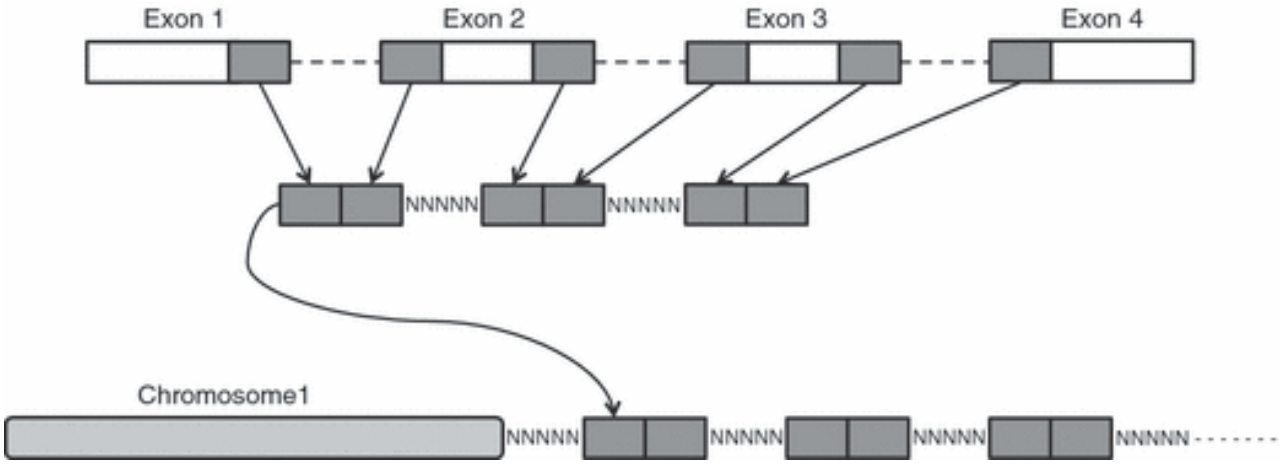
● Northern accessions

● Southern accessions

● SNP genotyping

M&M: Transcriptome resequencing, mapping and SNP calling

- Developing secondary xylem extracted in July 2008/2009
- Total RNA isolated and cDNA synthesized from purified PolyA RNA
- Paired end sequencing libraries prepared from 200-400bp fragments
- Each sample sequenced with four lanes of Illumina's GAII
- Sequences mapped with BWA (Li and Durbin 2009) to the reference genome plus exon-exon junctions
- Only uniquely mapped reads were retained for further analyses



Geraldes et al. 2011

M&M: SNP CALLING AND FASTA FILE GENERATION

- SNPs were called one library at a time using SAMtools/pileup/varFilter (Li et al 2009)
- SNP quality ≥ 20
- RMS mapping quality ≥ 25
- Read depth ≥ 6 (and ≥ 3 per allele)
- Heterozygote positions Minor/Major allele ≥ 0.18

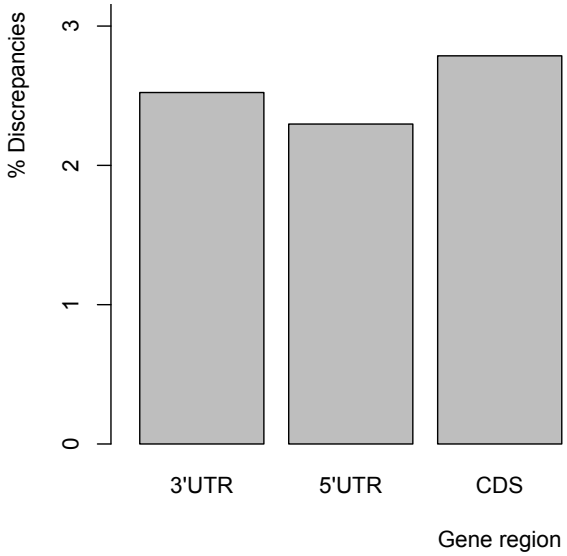
- For each gene in the genome we selected the longest transcript and generated two pseudo-phased fasta sequences per individual
- For heterozygote positions, alleles were assigned to haplotypes randomly
- Bases masked with N if coverage per allele < 3
- Bases mask with N if distance to exon/exon junction ≤ 6 bp
- Insertion deletion polymorphisms ignored and 20 bp around are masked with N

M&M: Population genetic analyses

- Sequence data:
  - Neighbour-Joining trees and bootstrap support estimated in MEGA4 (Tamura et al 2007)
  - Levels of nucleotide polymorphism (π and θ) were estimated for each gene for all sites, for noncoding sites (UTRs), synonymous sites and replacement sites in PolydNdS (scripts by K. Thorton), using the annotation from v2.1 of the genome
  - Tajima's D and F_{ST} across all sites in each gene were calculated in SITES (scripts by J. Hey)

- Genotype data:
  - Used a 32K SNP chip (3700 genes) designed with SNPs ascertained from the 20 transcriptome data and 16 whole genome sequences (BESC; Slavov et al. in prep.)
  - Population structure estimated with STRUCTURE (Pritchard et al 2000)
  - Number of clusters estimated with Structure Harvester (Evanno et al . 2005; scripts by Dent Earl)
  - PCA performed in EigenSoft (Patterson et al 2006, Price et al 2006)

M&M: SNP CALLING ACCURACY

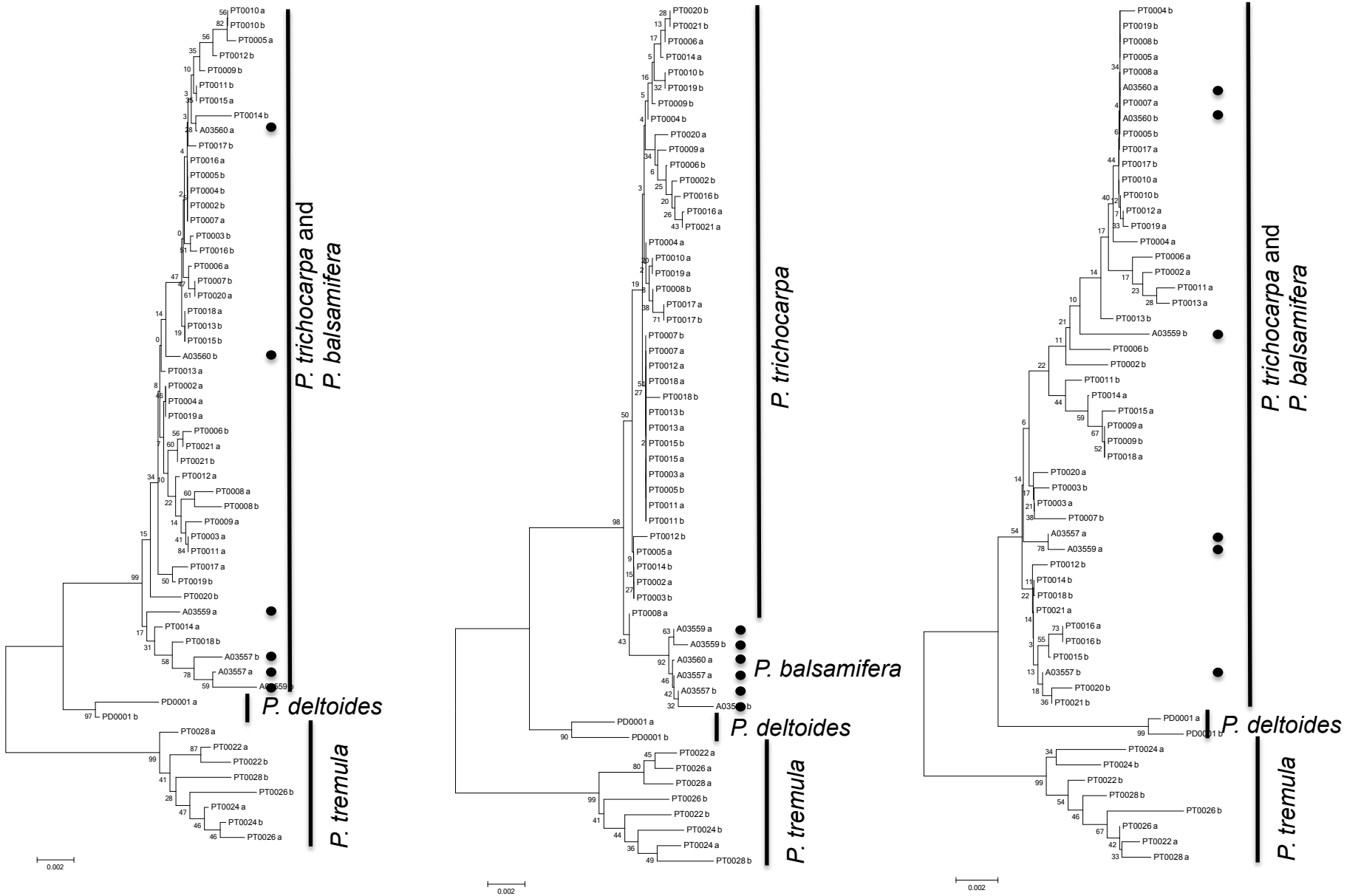


- Comparison of 78,297 genotypes inferred from Illumina transcriptome resequencing and Illumina Infinium array across 11 accessions revealed less than 3% discrepancies.

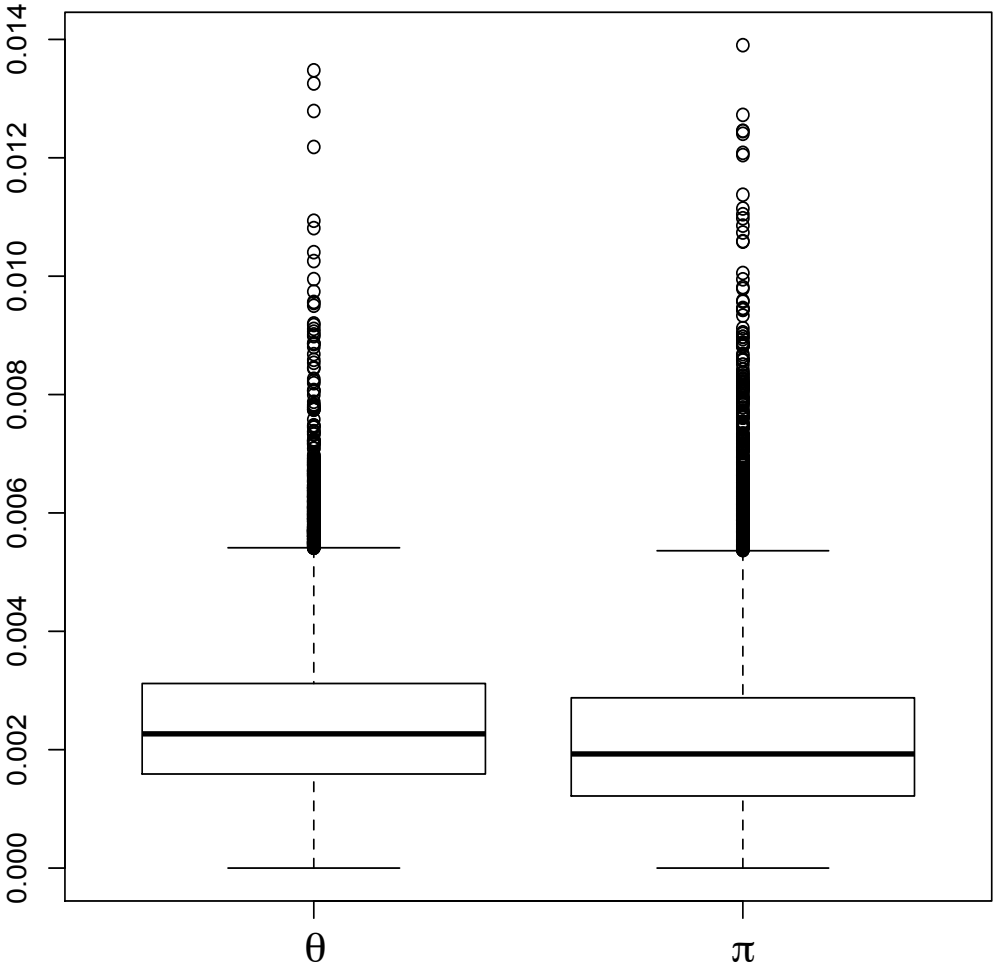
- Comparison 103,663 genotypes inferred from transcriptome and whole genome Illumina sequencing (BESC) of two accessions, revealed 3.1 and 3.2% of discrepancies.

- Levels of silent site polymorphism were well bellow the 9% silent site divergence among paralogs (Tuskan et al 2006)

GENE GENEALOGIES RECOVER SPECIES RELATIONSHIPS



Levels of nucleotide polymorphism

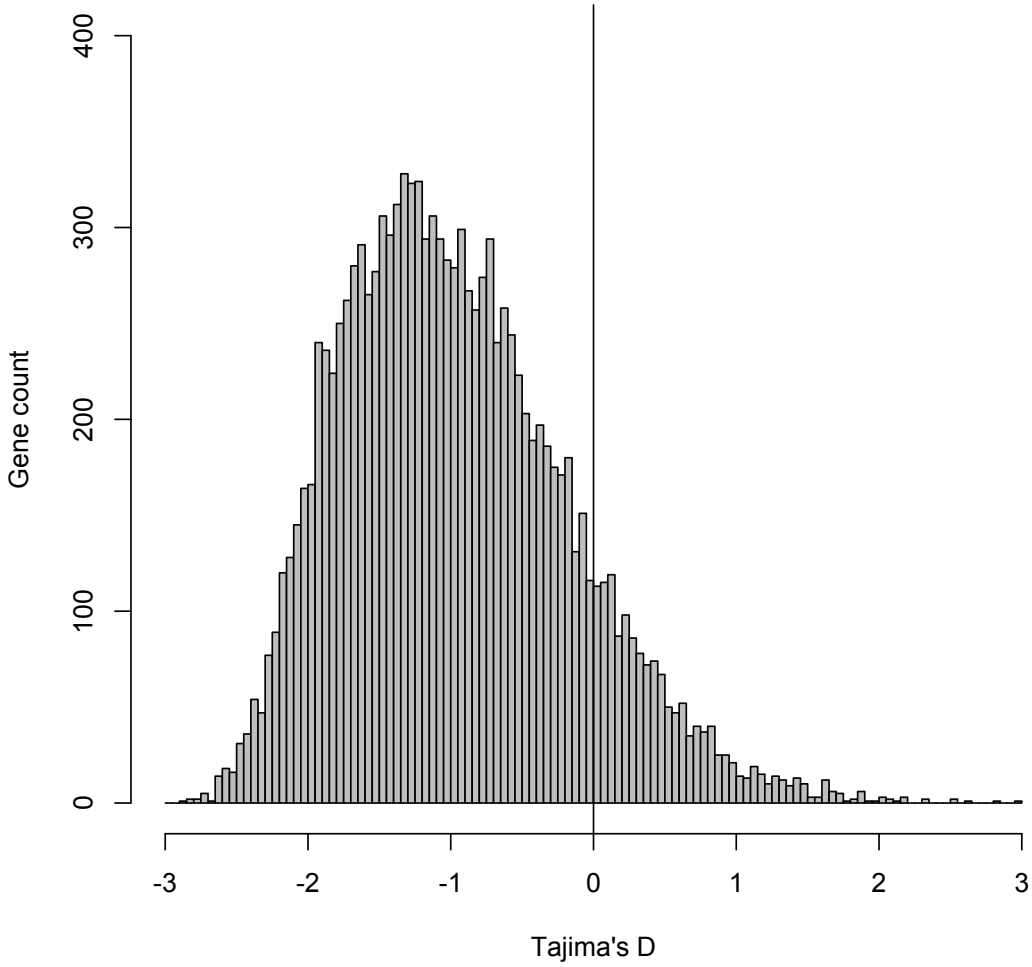


- 12,368 genes for which:
  - MinR ≥ 0.25
  - Bases ≥ 300 bp
  - $Bases_{CDS}$ ≥ 100 bp

- Two estimators of nucleotide polymorphism:
  - θ_w Proportion of segregating sites (Watterson 1975)
  - π Average pairwise differences (Nei and Li 1979)

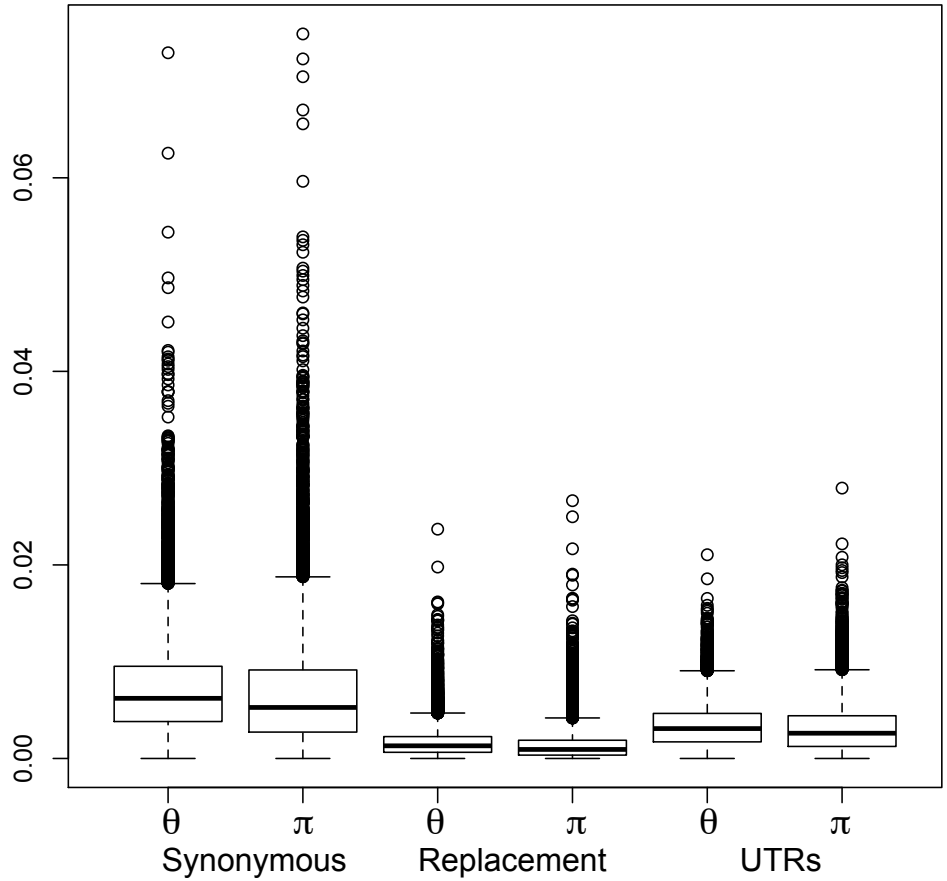
		All Sites	
	n	θ	π
P. trichocarpa	40	0.0024 (0 - 0.0135)	0.0022 (0 - 0.0139)

Frequency spectrum of polymorphisms



	n	All Sites	
		Tajima's D	Fu and Li's D*
P. trichocarpa	40	-0.974 (-2.894 - 2.981)	-0.501 (-4.933 - 1.814)

Purifying selection on coding regions and UTRs

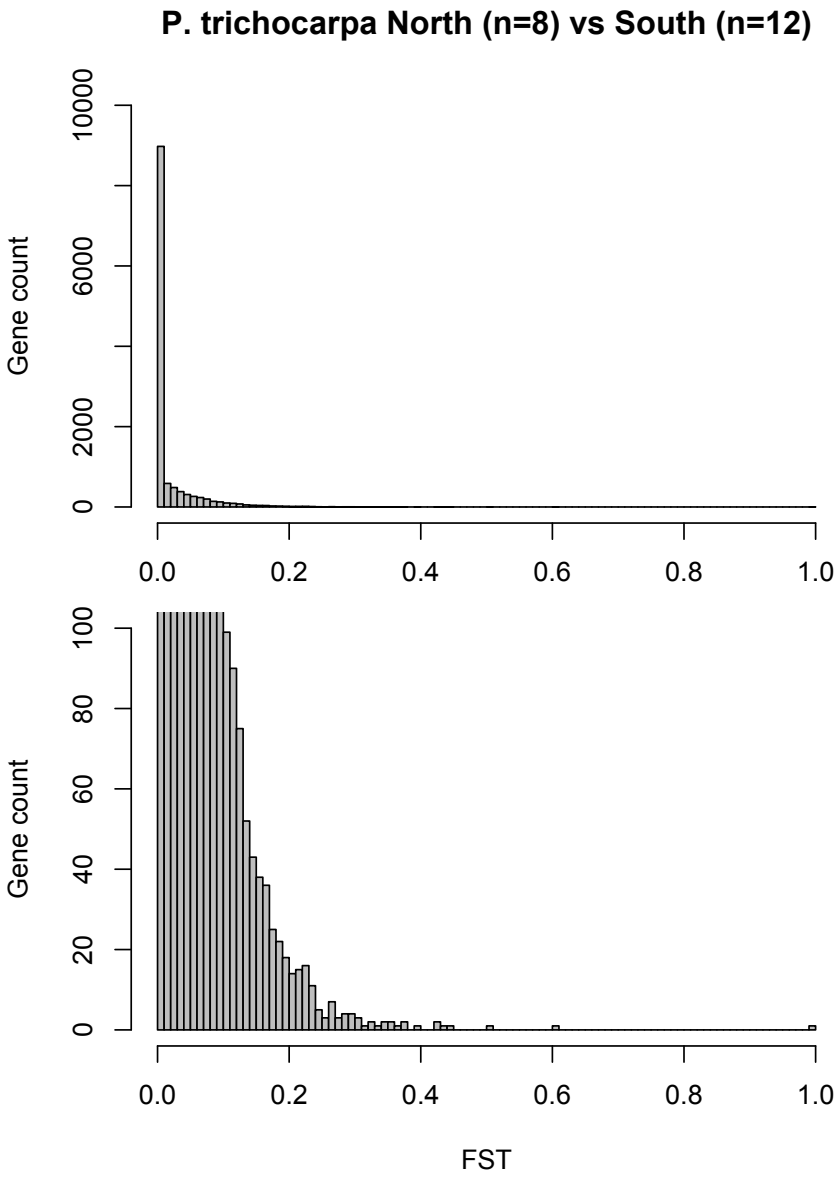


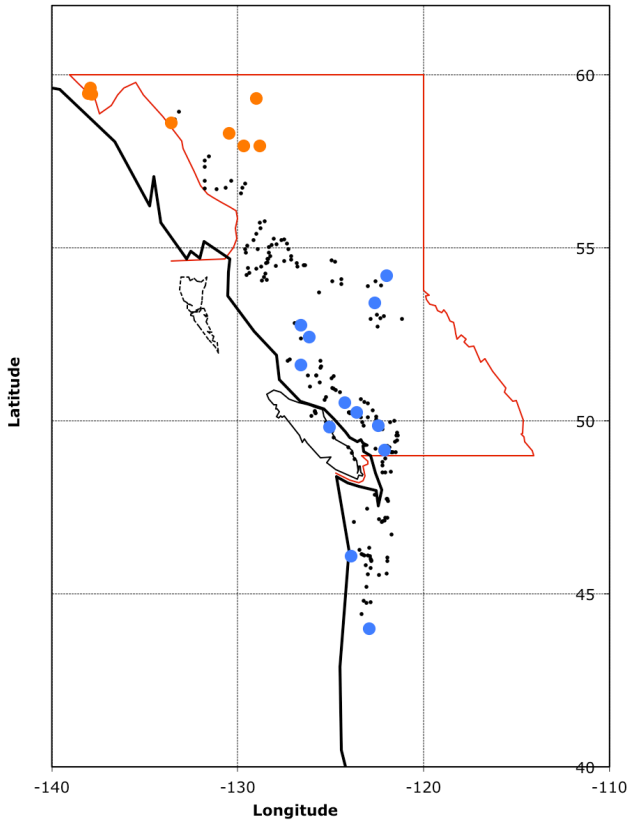
- Median $\pi_A/\pi_S = 0.1814$
- 5.44% of genes $\pi_A/\pi_S > 1$

- Median $\pi_{UTR}/\pi_S = 0.5046$
- 21.70% of genes $\pi_{UTR}/\pi_S > 1$

	n	Synonymous sites		Replacement sites		UTRs	
		θ	π	θ	π	θ	π
P. trichocarpa	40	0.0073 (0 - 0.0729)	0.0069 (0 - 0.0749)	0.0017 (0 - 0.0237)	0.0014 (0 - 0.0266)	0.0034 (0 - 0.0211)	0.0031 (0 - 0.0279)

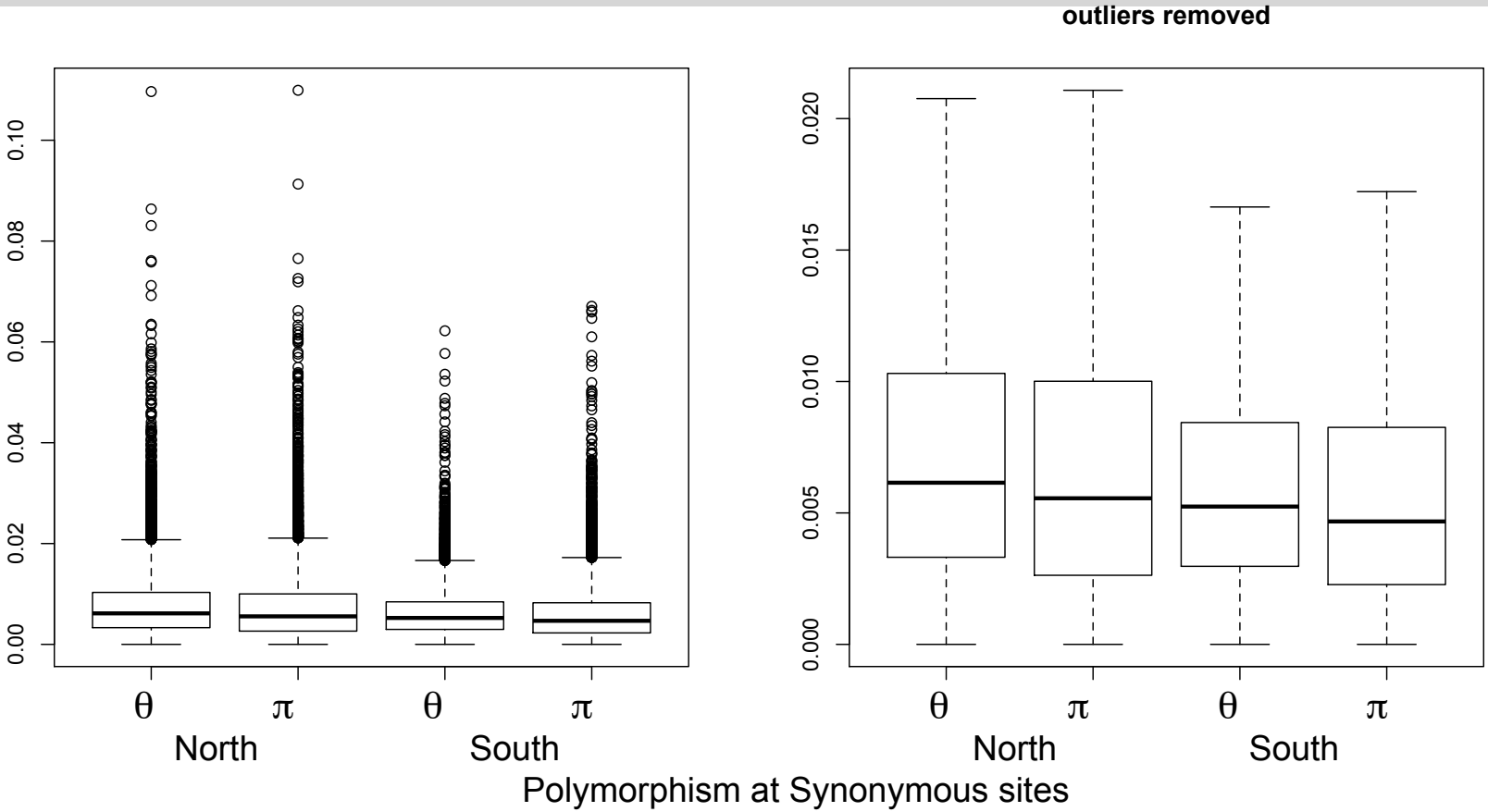
POPULATION DIFFERENTIATION IN *P. TRICHOCARPA*





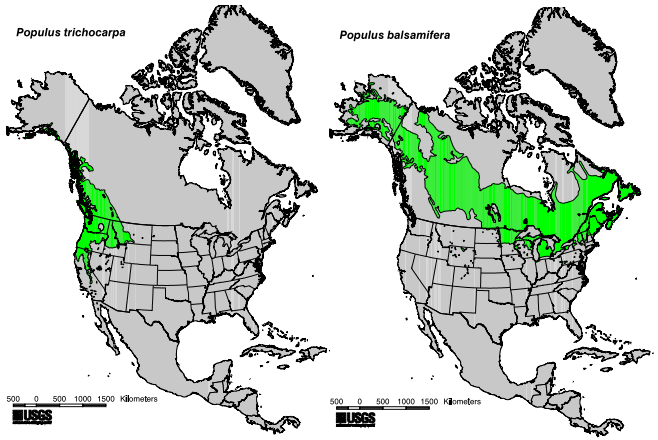
- F_{ST} for sequence data (Hudson et al. 1992, eq. 3) between samples latitude > 57°N (n=8) and samples latitude < 55°N (n=8)
- Average F_{ST} = 0.0137
- Median F_{ST} = 0

LEVELS OF NUCLEOTIDE POLYMORPHISM N VS S

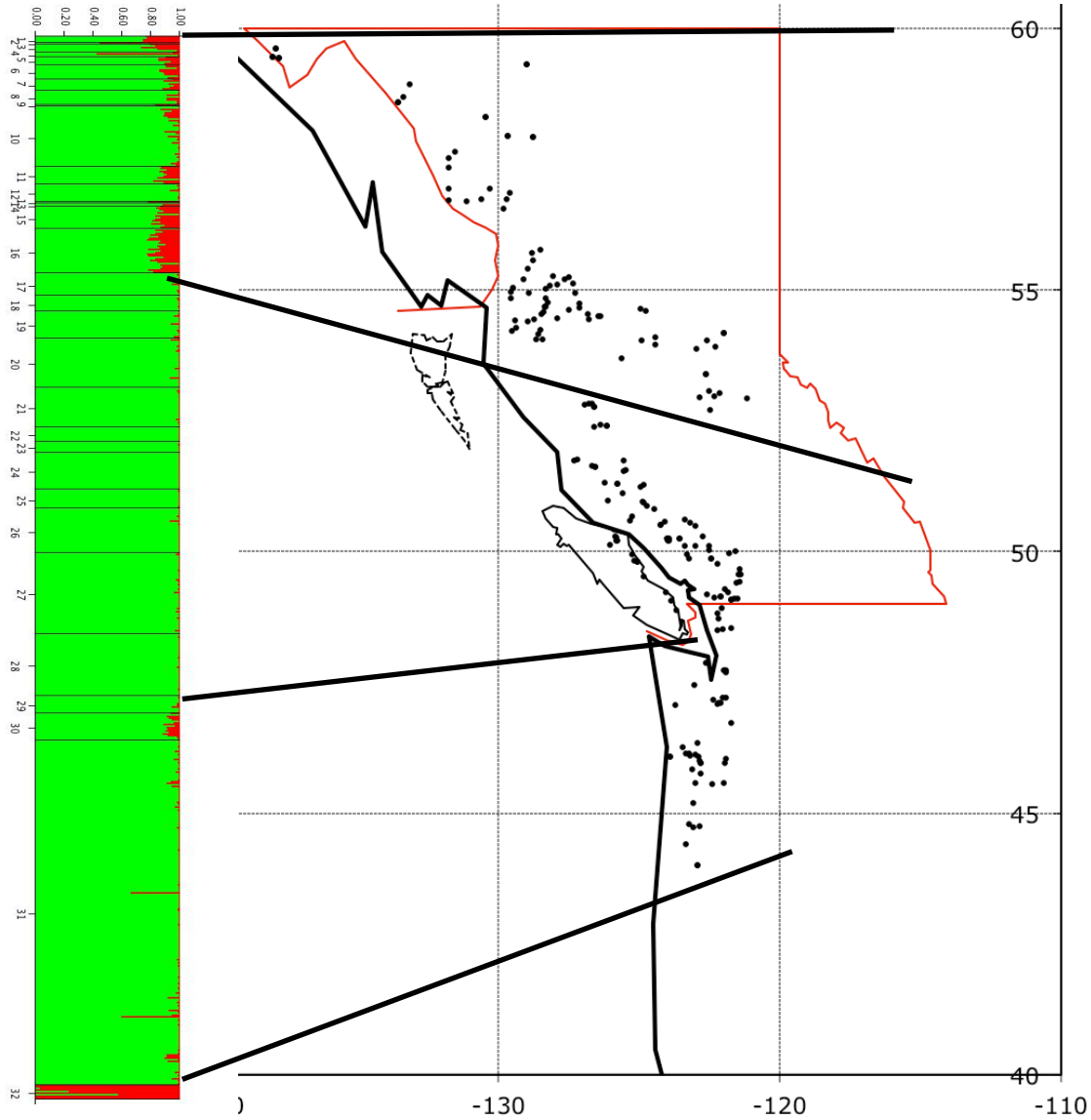


	n	Synonymous sites θ	Synonymous sites π	Replacement sites θ	Replacement sites π	UTRs θ	UTRs π
P. trichocarpa North	16	0.0077 (0 - 0.1097)	0.0075* (0 - 0.1099)	0.0017 (0 - 0.0362)	0.0015* (0 - 0.0367)	0.0030 (0 - 0.0205)	0.0029 (0 - 0.0280)
P. trichocarpa South	24	0.0063 (0 - 0.0622)	0.0061 (0 - 0.0671)	0.0014 (0 - 0.0244)	0.0013 (0 - 0.0269)	0.0031 (0 - 0.0205)	0.0030* (0 - 0.0283)

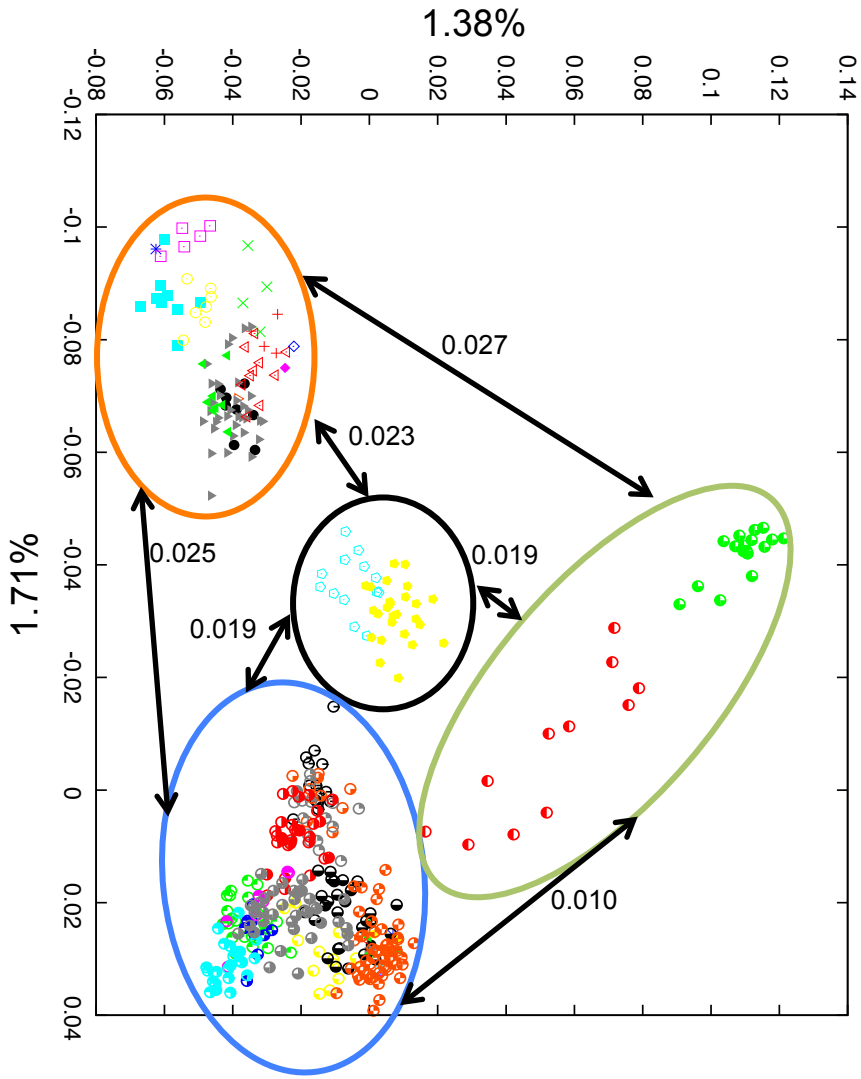
HIGHER POLYMORPHISM IN NORTHERN *P. TRICHOCARPA* - GENE FLOW FROM *P. BALSAMIFERA*?

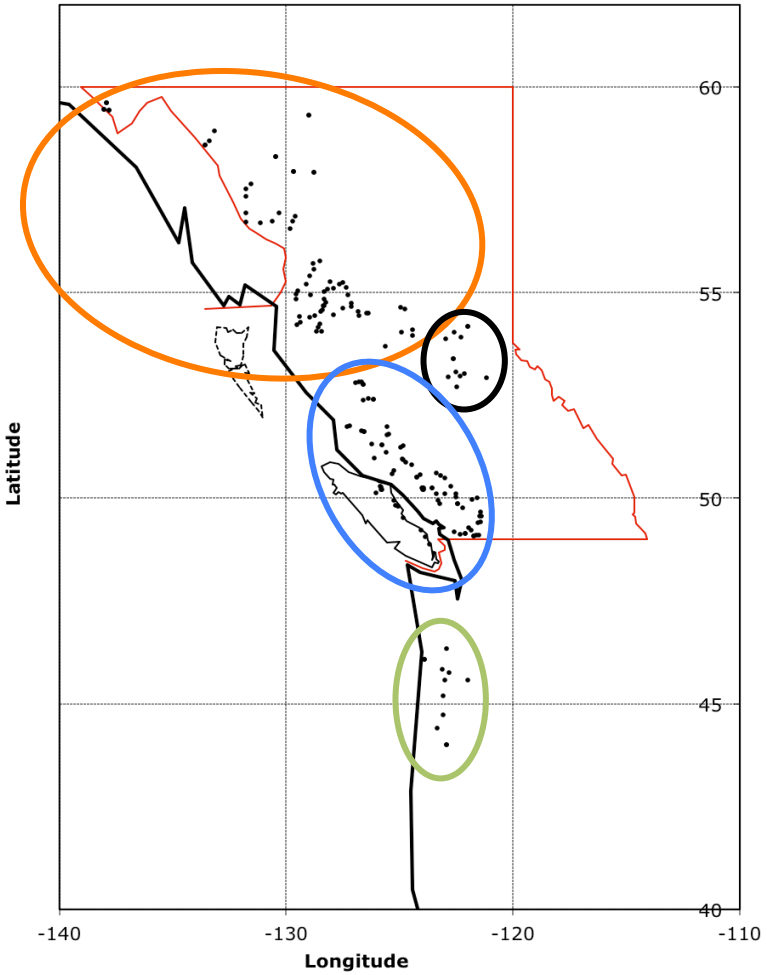


- Structure (K=2):
  - Northern populations more “red”
- North-balsamifera share more alleles and have less fixed differences
- Median FST for sequence data:
  - North-balsamifera FST = 0.294
  - South-balsamifera FST = 0.358



POPULATION STRUCTURE IN *P. TRICHOCARPA*



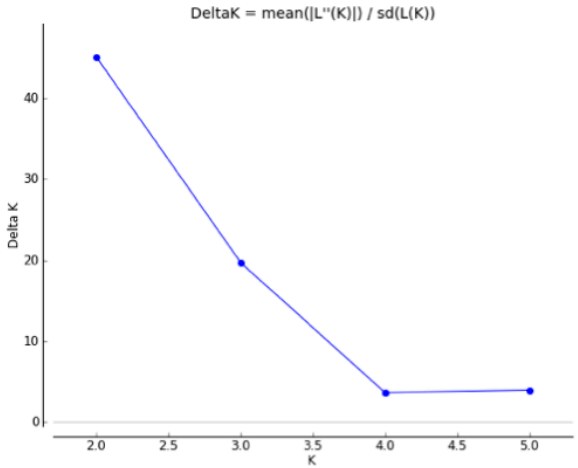


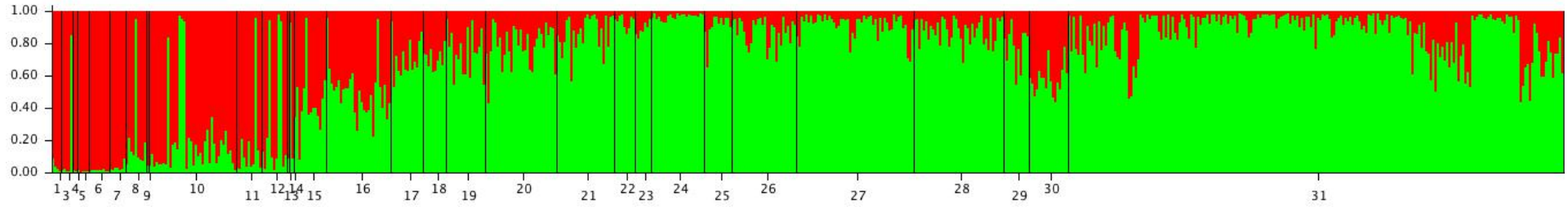
PCA analyses performed in Eigensoft (Price et al 2006; Patterson et al 2006) using ~500 accessions and 29K SNPs

POPULATION STRUCTURE IN *P. TRICHOCARPA*

Structure (Pritchard et al 2000):

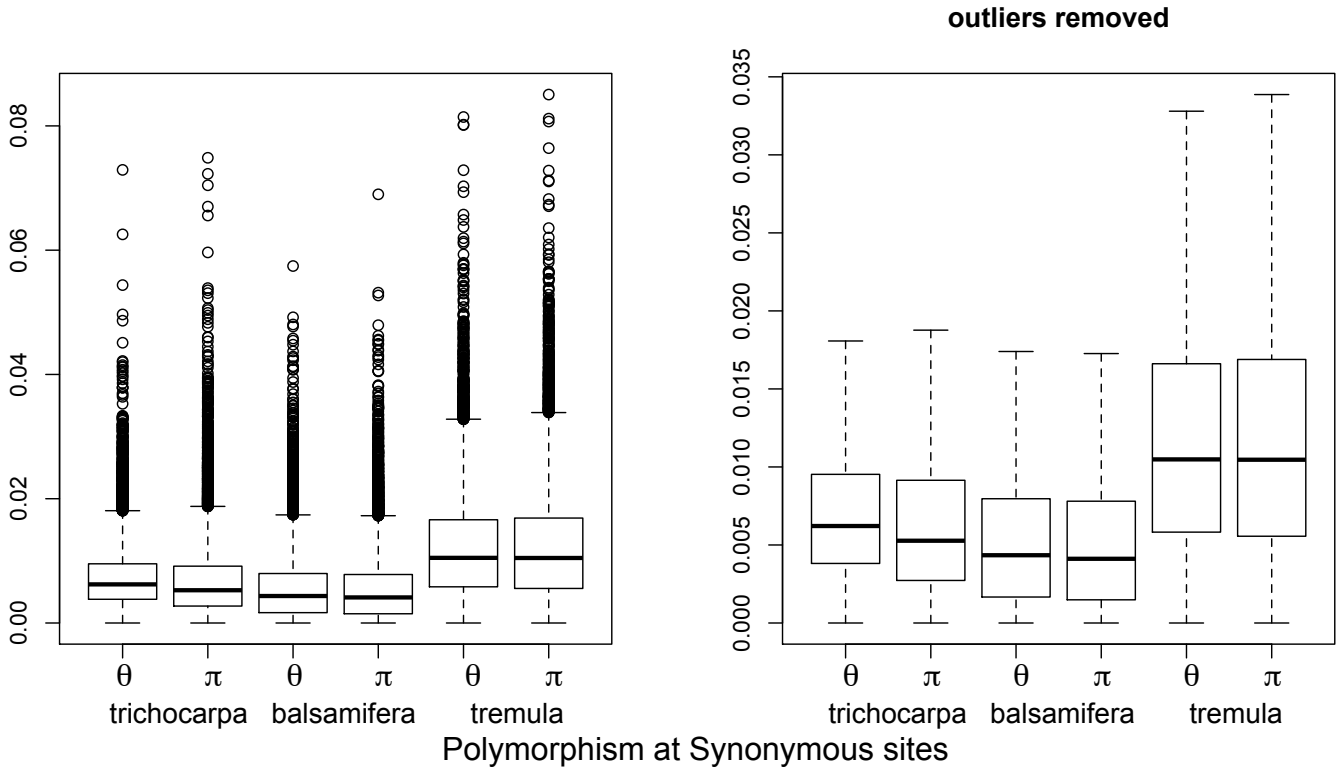
- 656 samples (samples with Q_{balsam} > 0.3 eliminated)
- 824 loci without missing data
- Varied K between 1 and 6
- 5 runs per K with 100k burn-in iterations, plus 100k
- Selected best K (Evano et al. 2005)





North -> South

LEVELS OF NUCLEOTIDE POLYMORPHISM *POPULUS SPP.*



- *P. balsamifera* (Olson 2010)
  π_S=0.0045
- *P. tremula* (Ingvarsson 2008)
  π_S=0.0120

- Median π_A/π_S
  - tricho= 0.181
  - balsam= 0.179
  - tremula= 0.144

		Synonymous sites		Replacement sites		UTRs	
	n	θ	π	θ	π	θ	π
P. trichocarpa	40	0.0073 (0 - 0.0729)	0.0069 (0 - 0.0749)	0.0017 (0 - 0.0237)	0.0014 (0 - 0.0266)	0.0034 (0 - 0.0211)	0.0031 (0 - 0.0279)
P. balsamifera	6	0.0056 (0 - 0.0574)	0.0055 (0 - 0.0690)	0.0012 (0 - 0.0201)	0.0012 (0 - 0.0218)	0.0029 (0 - 0.0365)	0.0029 (0 - 0.0389)
P. tremula	8	0.0121 (0 - 0.0814)	0.0122 (0 - 0.0850)	0.0021 (0 - 0.0431)	0.0020 (0 - 0.0421)	0.0039 (0 - 0.0471)	0.0039 (0 - 0.0464)

CONCLUSIONS

What is **the overall level of nucleotide polymorphism** in *P. trichocarpa?*

Across all sites, P. trichocarpa shows moderate levels of polymorphism for a species that shows a very large amount of phenotypic variability

How is polymorphism **partitioned geographically**?

Overall, levels of population differentiation between southern and northern populations are low (Average F_{ST} =0.01), yet, PCA and structure analyses detect some population structure. This has important consequences for ongoing and future association studies in this species. Northern populations of *P. trichocarpa* have higher polymorphism likely due to gene flow from *P. balsamifera*.

How is polymorphism **partitioned across different classes of sites**?

Polymorphism is lowest at replacement sites and highest at synonymous sites, with UTRs showing intermediate levels. This is consistent with higher constraint at sites that alter the aminoacid composition of proteins than in potentially regulatory untranscribed regions.

How does it compare to **other species in the genus**?

Levels of polymorphism in P. tremula are twice as high as those in P. trichocarpa and P. tremula, suggesting that aspens have higher long-term effective population size.

Co-Authors and Acknowledgements

AGIP team

Shawn Mansfield
JinGui Chen
Jürgen Ehlting
Yousry El-Kassaby
Brian Ellis
Rob Guy
Thomas Maness
Geoff Wasteneys
Shofiul Azam
Hua Bao
Michael Friedmann
Miki Fujita
Jan Hanneman
Howie Harshaw
Peter Kalynyak
Eryang Li
Athena McKown
Ilga Porth
Catalin Ristea
Alex Skyba
Shucai Wang

Genome Science Centre:

Inanc Birol
Reza Falsafi
Yongjun Zhao
Marco Marra
Johnson Pang
Nina Thiessen
Steve Jones

Mike Barker
Nolan Kane
Matt King
Heather Ramsay
Teaghan Mayers
Arnold Limantono

Collaborators

ORNL/BESC – Steve diFazio *et al.*
UPSC/POPLARENERGY – Rishi Bhalerao
USFS Forest Products Lab - Dan Cullen
BC Ministry of Forests and Range - Alvin Yanchuk
Greenwood Resources - Brian Stanton
Kruger Inc. – Dan Carson

GenomeBritishColumbia

Applied Genomics Innovation Program

Co-Authors:

C. Grassa, G. T. Slavov, W. Muchero, R. Priya, P. K. Ingvarsson, Stefan Jansson, G. Tuskan, C. J. Douglas, Q. C. B. Cronk