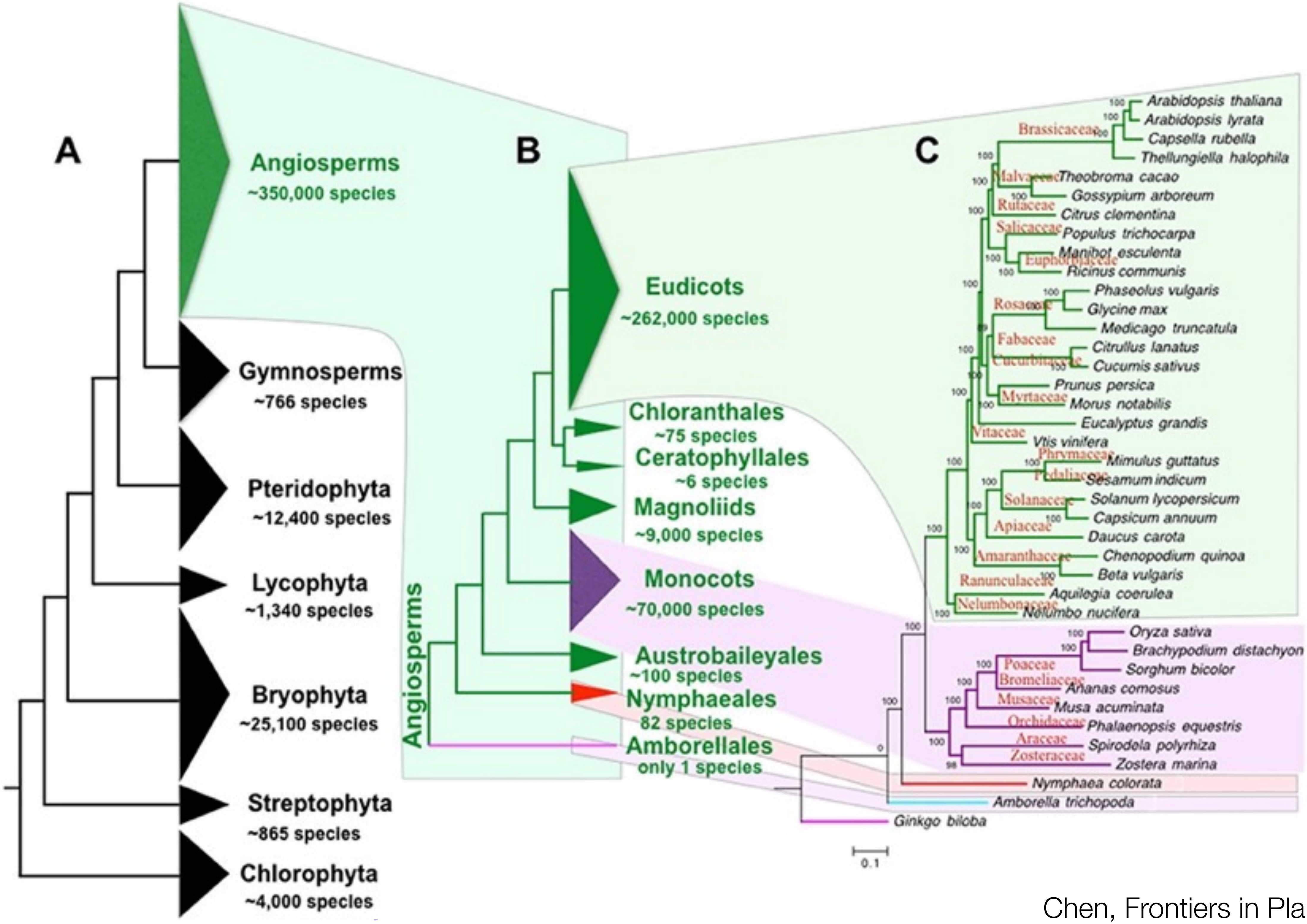
Genomic insights to evolution

Jarkko Salojärvi
Nanyang Technological University

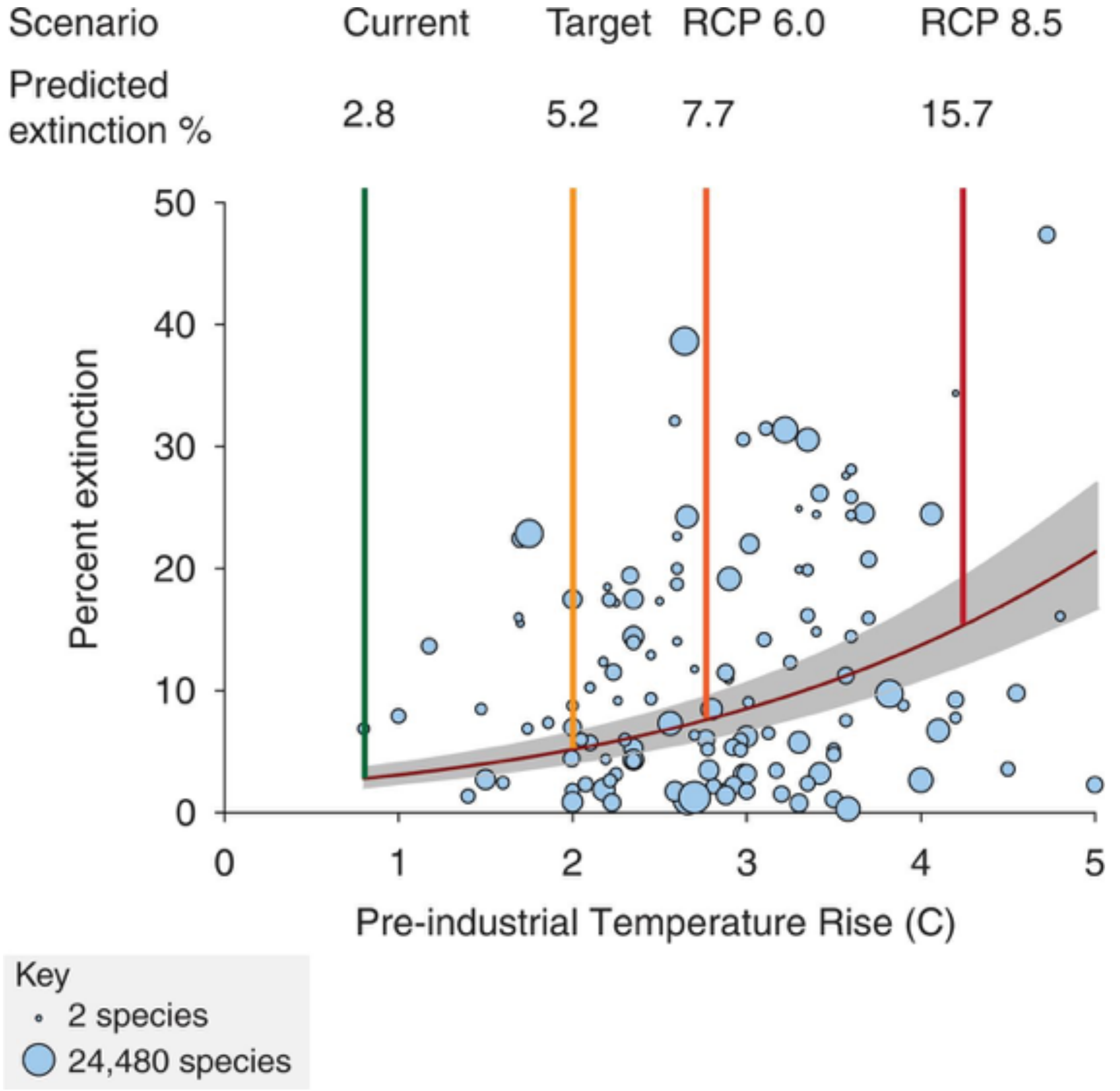
Tokyo University of Science
Bioinformatics workshop
4.12.2018

An immense plant biodiversity



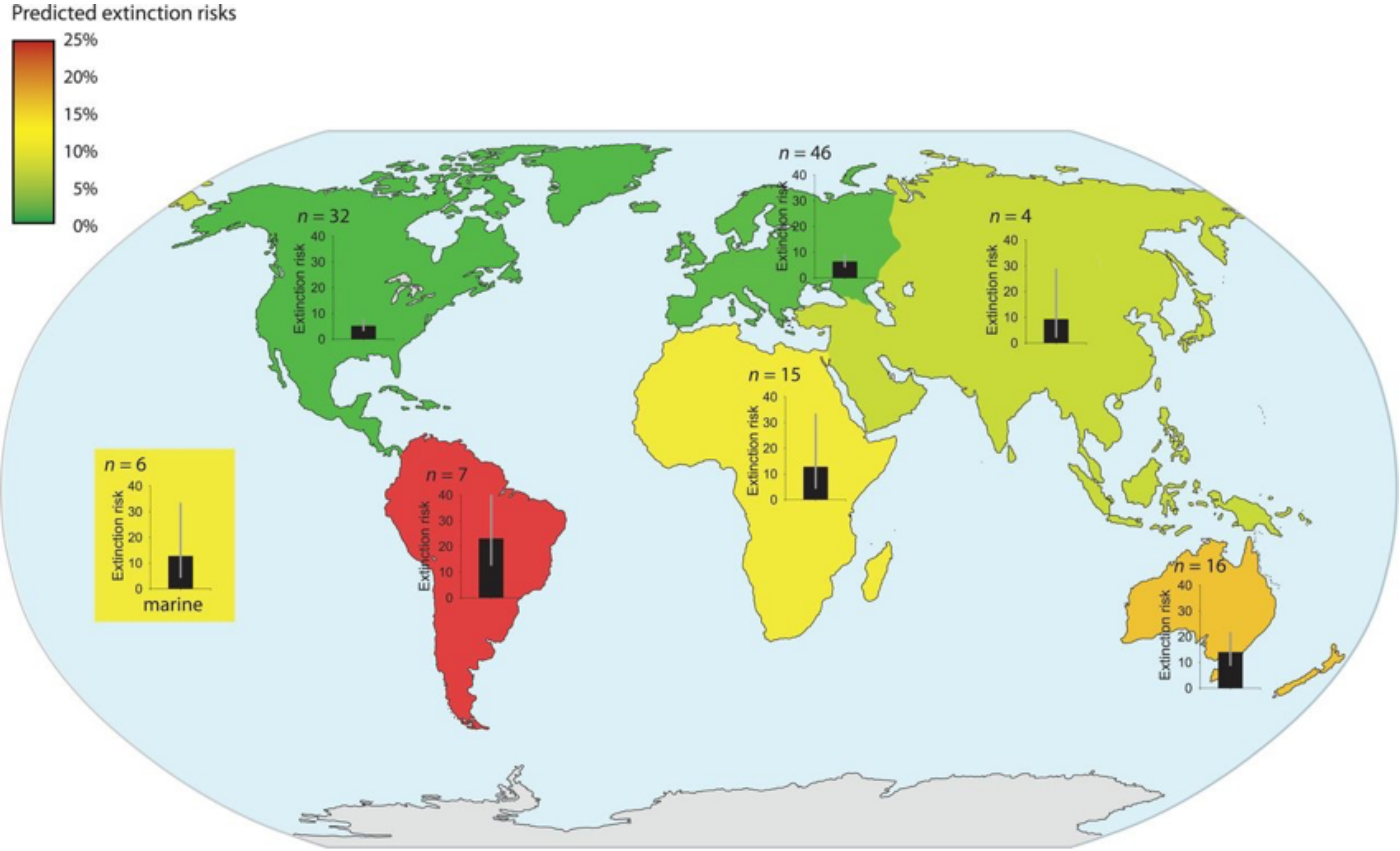
Chen, Frontiers in Plant Science 2018.

..but how much longer..?

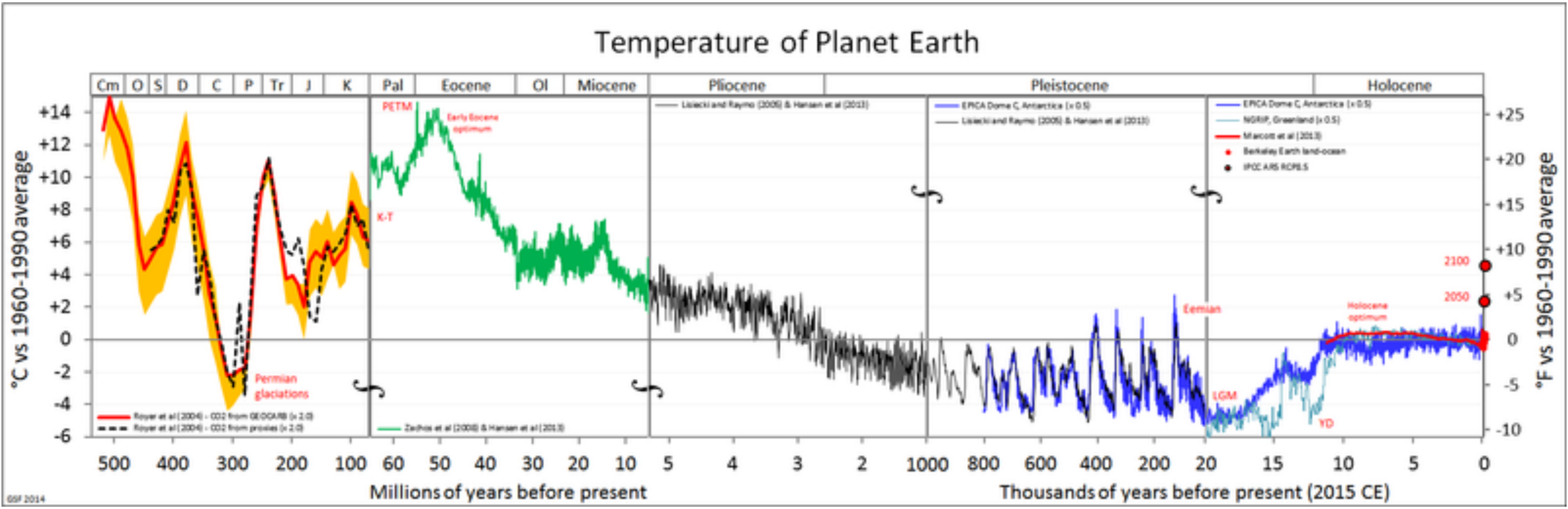


Urban, Science 2015.

..especially in the tropics

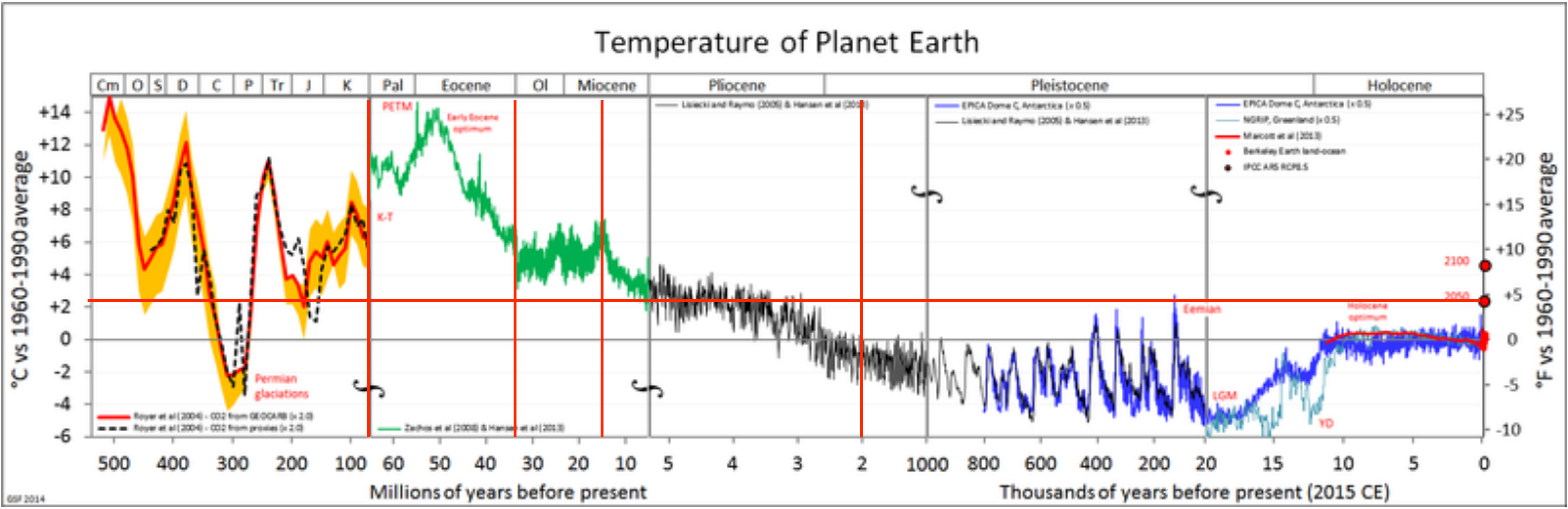


Insights from history?



Pliocene	Pliocene–Pleistocene boundary marine extinction	2 Ma
Neogene	Middle Miocene disruption	14.5 Ma
Palaeogene	Eocene–Oligocene extinction event	33.9 Ma
	Cretaceous–Paleogene extinction event	66 Ma

Insights from history?



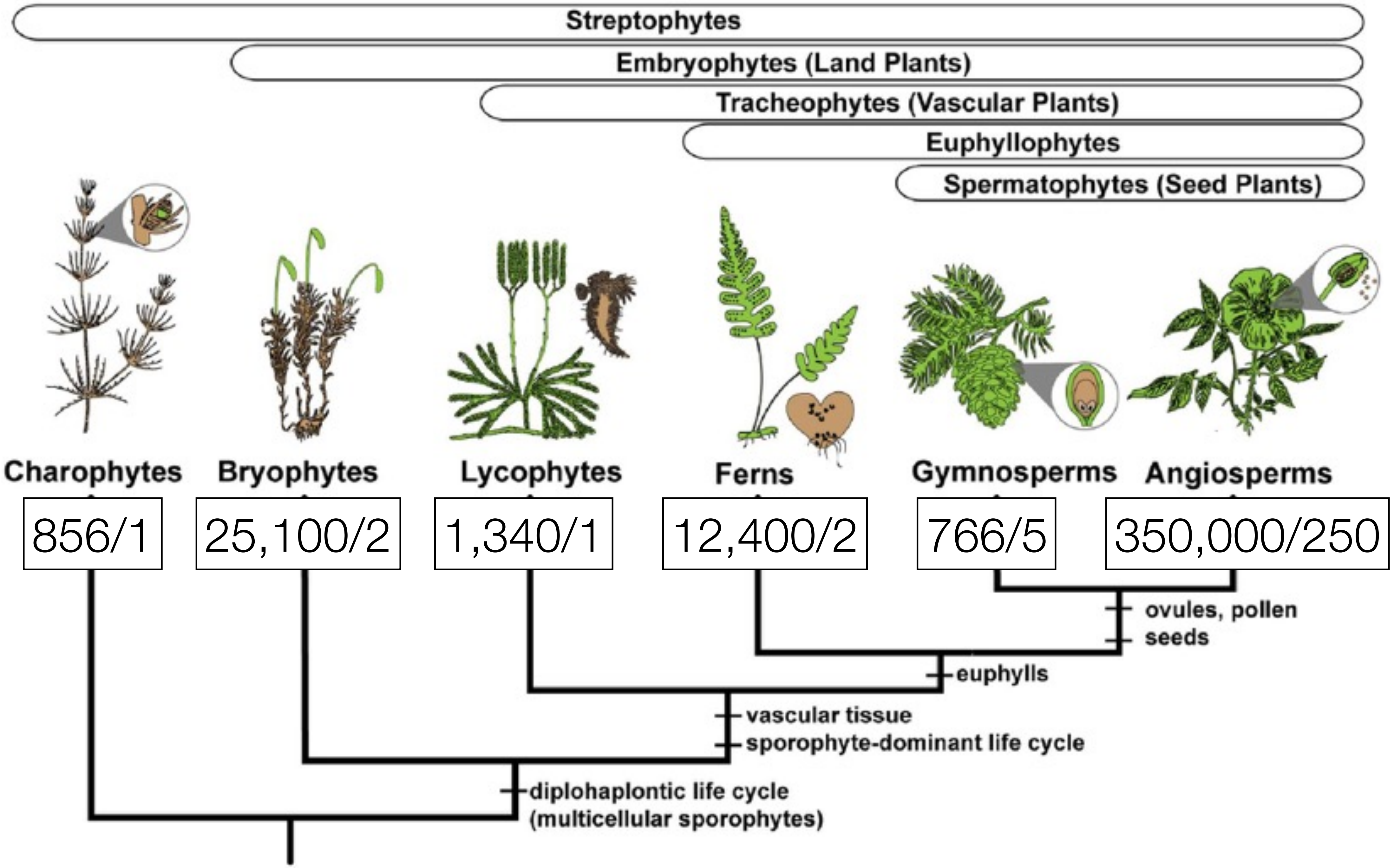
Pliocene	Pliocene–Pleistocene boundary marine extinction	2 Ma
Neogene	Middle Miocene disruption	14.5 Ma
Palaeogene	Eocene–Oligocene extinction event	33.9 Ma
	Cretaceous–Paleogene extinction event	66 Ma

Evolutionary questions

- How much and how fast can evolution compensate for the climate change?
- The evolutionary “winners”?
- Human influence; landscape fractionation.

- Plant genomes are the result of millions of years of selection and adaptation.
  - What can we see from the past?
  - How does genome evolution work?

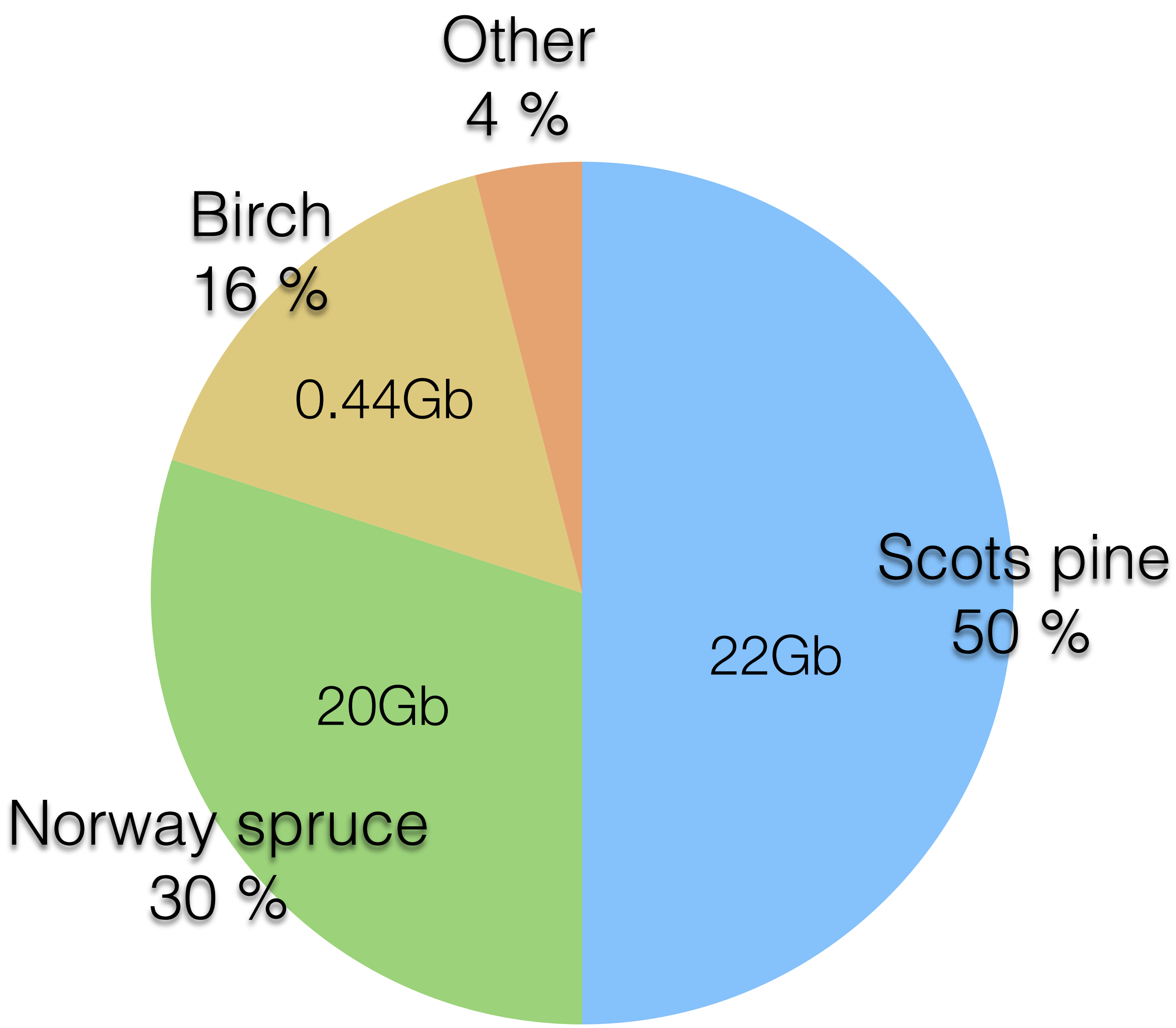
Sequencing the world biodiversity… not there yet



Jarkko Salojärvi

Chen, Frontiers in Plant Science 2018.

Boreal forest consists of only few foundation species



Using silver birch as a model to study genome evolution in trees

Ecology:

- Fast-growing pioneer tree.
- Important species in the boreal ecosystem.

Genome:

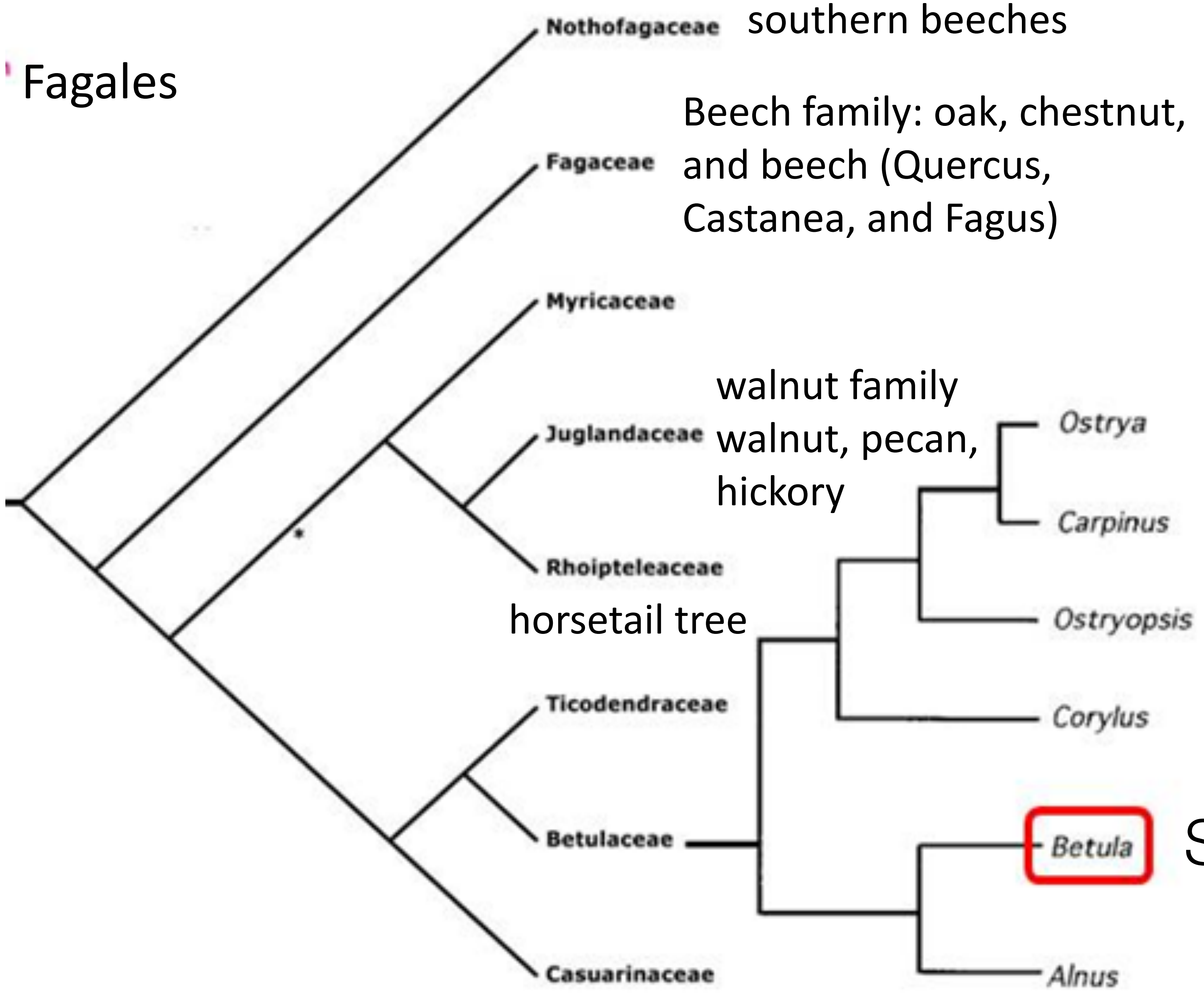
- Only tree with a reference genome which can be made to flower within a year.
- Small reference genome.

Genetics:

- Selfing possible - can construct highly inbred lines.
- Low interspecific incompatibility - crosses between different Betula species possible.

Betulaceae and higher order Fagales contain several important tree species

Fagales



southern beeches

Beech family: oak, chestnut, and beech (Quercus, Castanea, and Fagus)

walnut family walnut, pecan, hickory

horsetail tree

Hop-hornbeam

Hornbeam

Hazel-hornbeam

Hazels

Silver birch (*Betula pendula*)

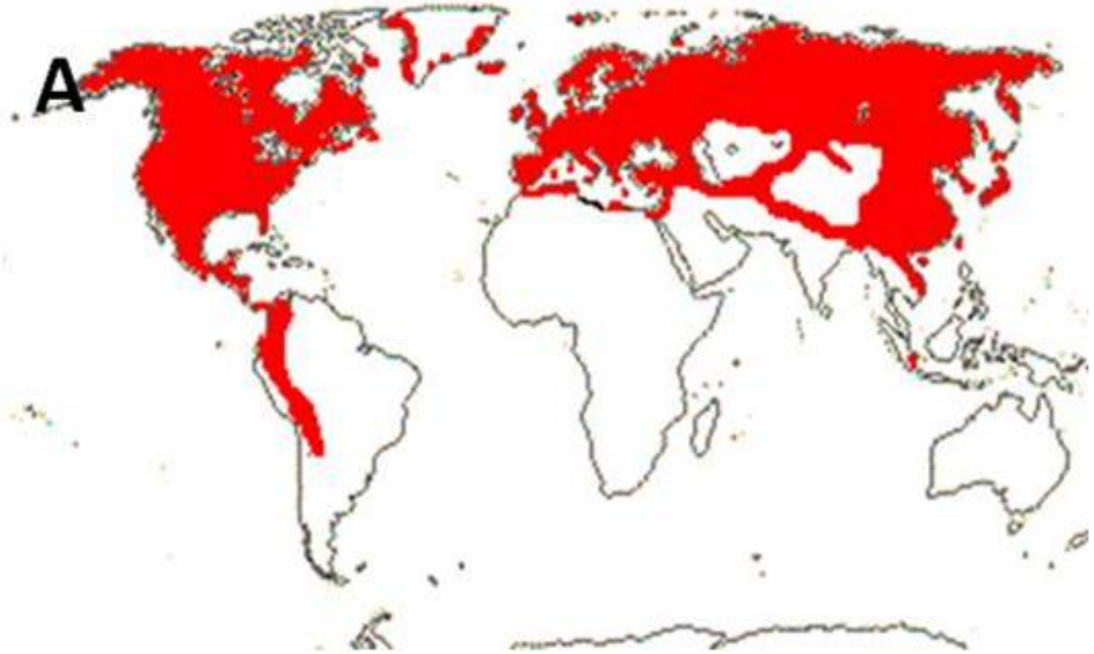
Alders

www.mobot.org

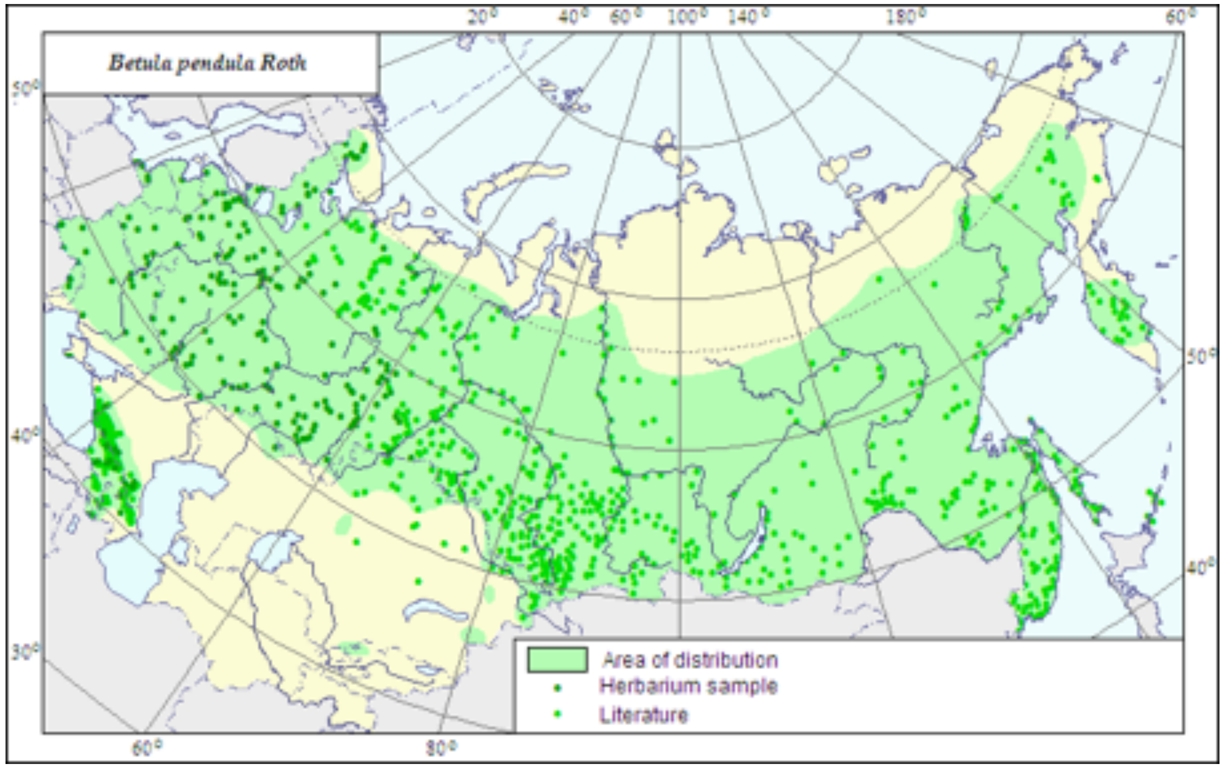
Betulaceae is a widely spread family

- Birch highly prevalent - ecologically important.
- Most important deciduous tree for forestry in Finland - 16% of total tree volume.

Range of the family *Betulaceae*

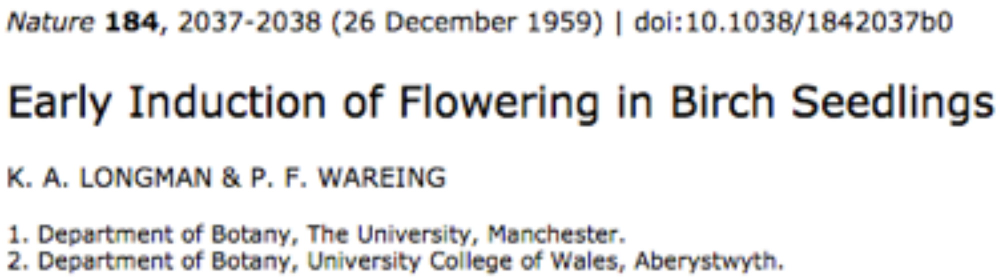


Range of *Betula pendula*



Model for Tree Genetics

- **Only tree which can be made to flower within a year.**
- Selfing possible - can construct highly inbred lines.
- Speciation not resolved - crosses between different Betula species possible.

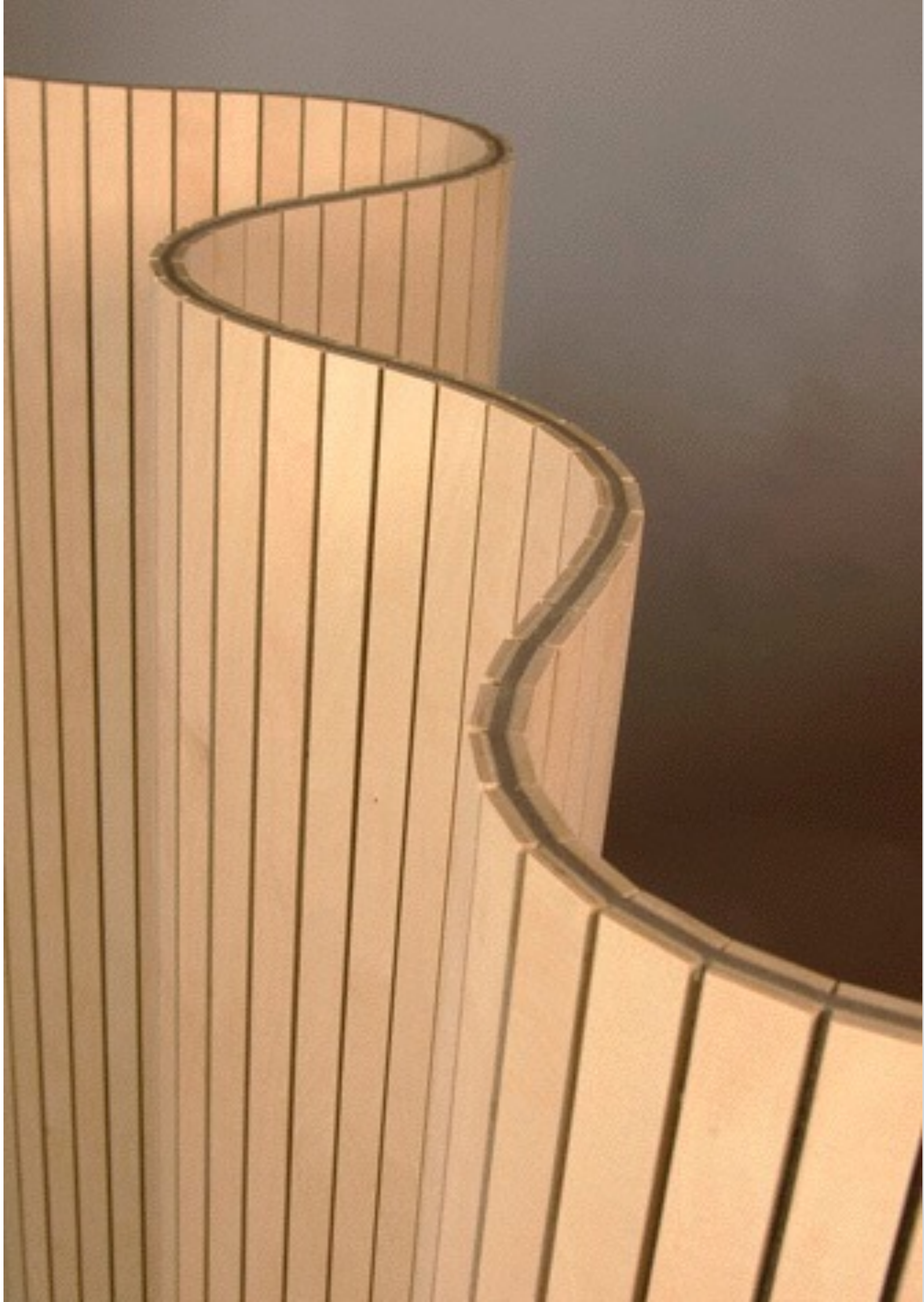


Nature **184**, 2037-2038 (26 December 1959) | doi:10.1038/1842037b0

Early Induction of Flowering in Birch Seedlings

K. A. LONGMAN & P. F. WAREING

1. Department of Botany, The University, Manchester.
2. Department of Botany, University College of Wales, Aberystwyth.



Birch Genome

- Flow cytometry: 440 Mb.
- Assembly v1.2 435.0 Mb.
  - 5,524 scaffolds >1,500 b.
  - N50 length 239,8 kb.
  - Longest scaffold: 3,17 Mb
- 89% of the assembly mapped to 14 chromosomes
  - Ultra-high density linkage map, 3.6 M markers.

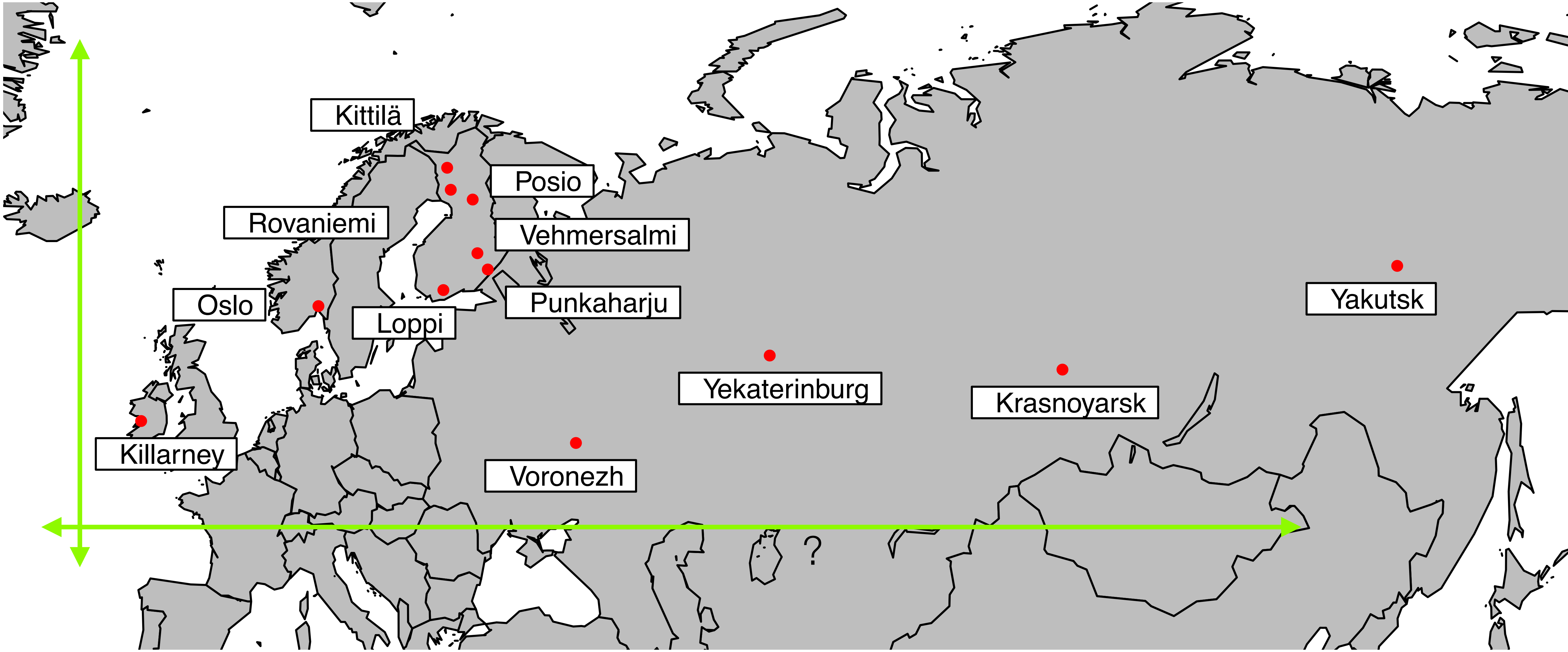
Jarkko Salojärvi

Genomic insights to population history

NANYANG TECHNOLOGICAL UNIVERSITY

Jarkko Salojärvi

Population Genomics of Birch: 12 Sites Representing the Geographic Range



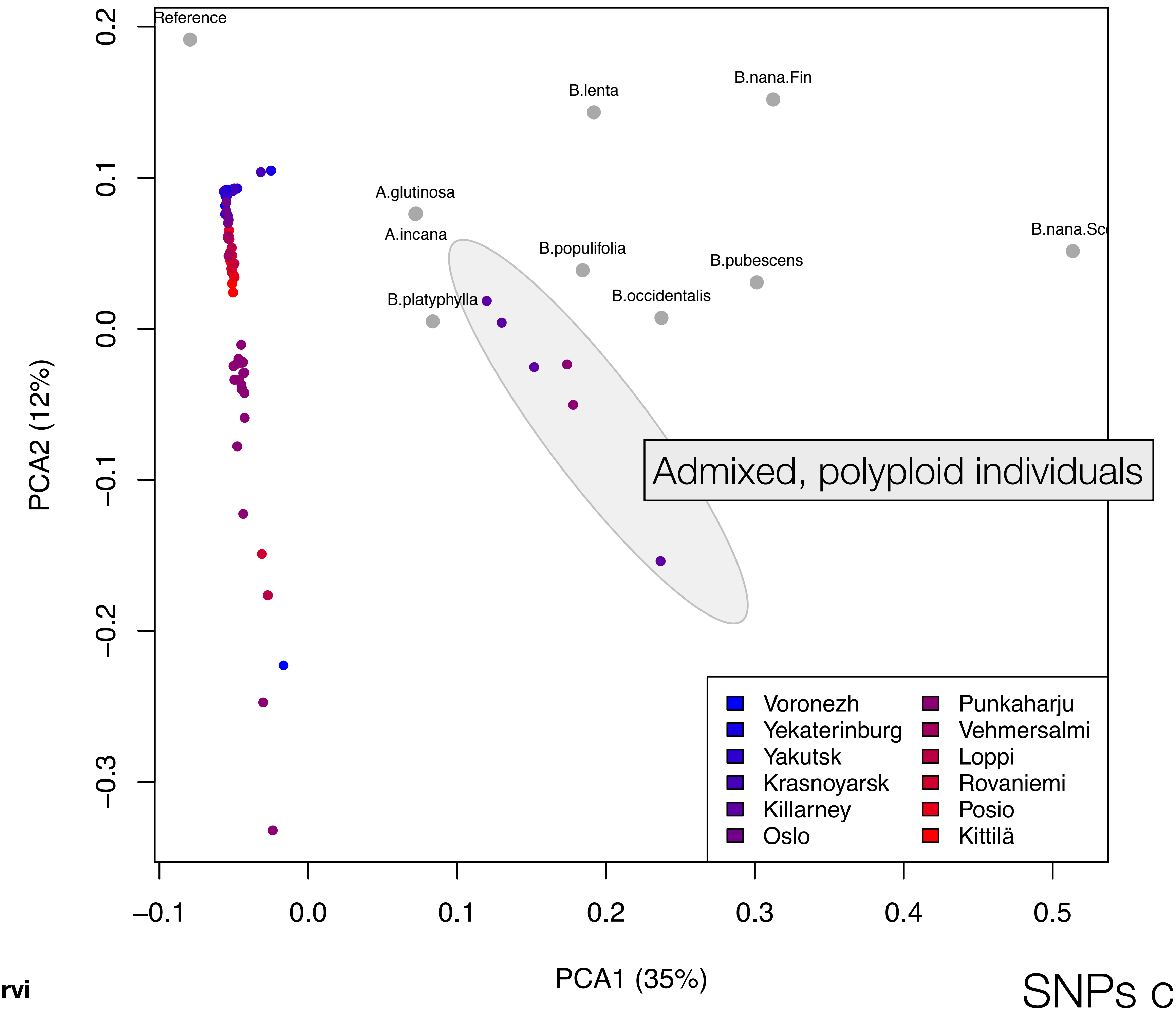
4-5 individuals/site

Total 80 diploid individuals, ~8x coverage.

8 other Betula species with ~20x coverage.

Jarkko Salojärvi

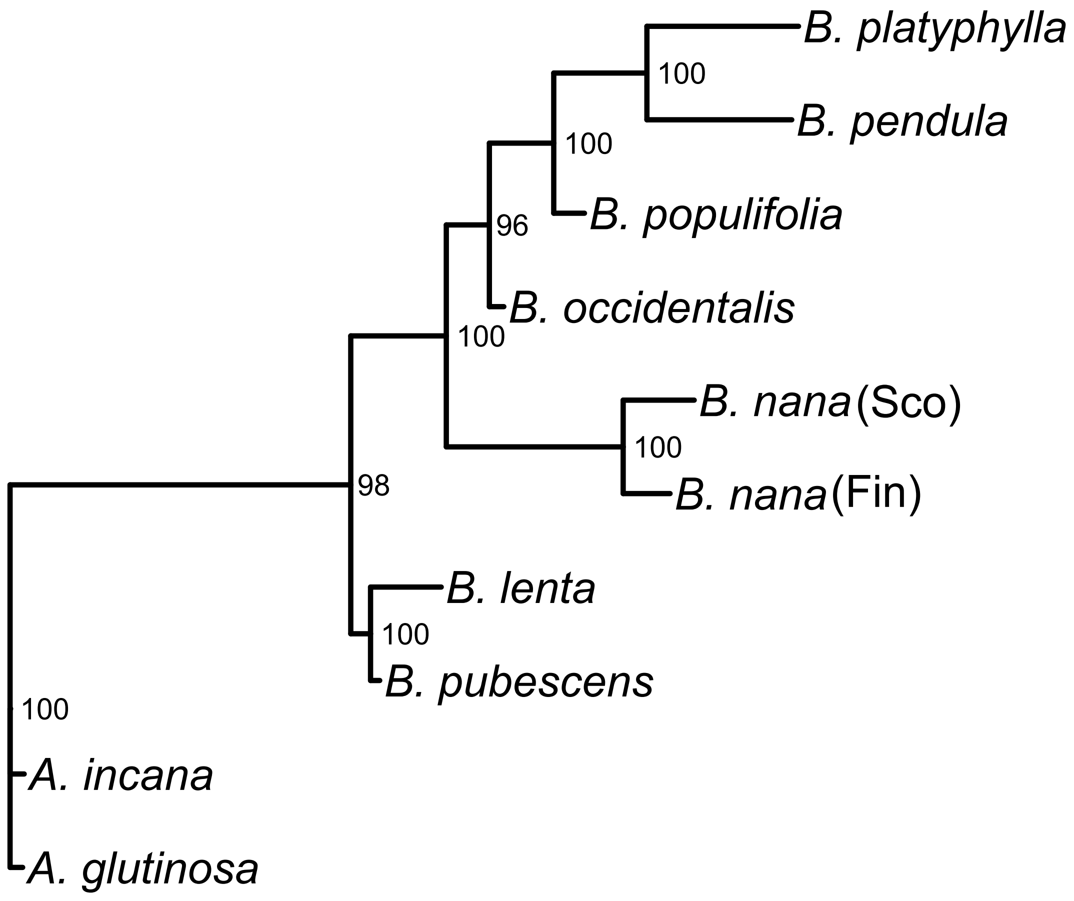
PCA of SNPs Shows Admixed Individuals



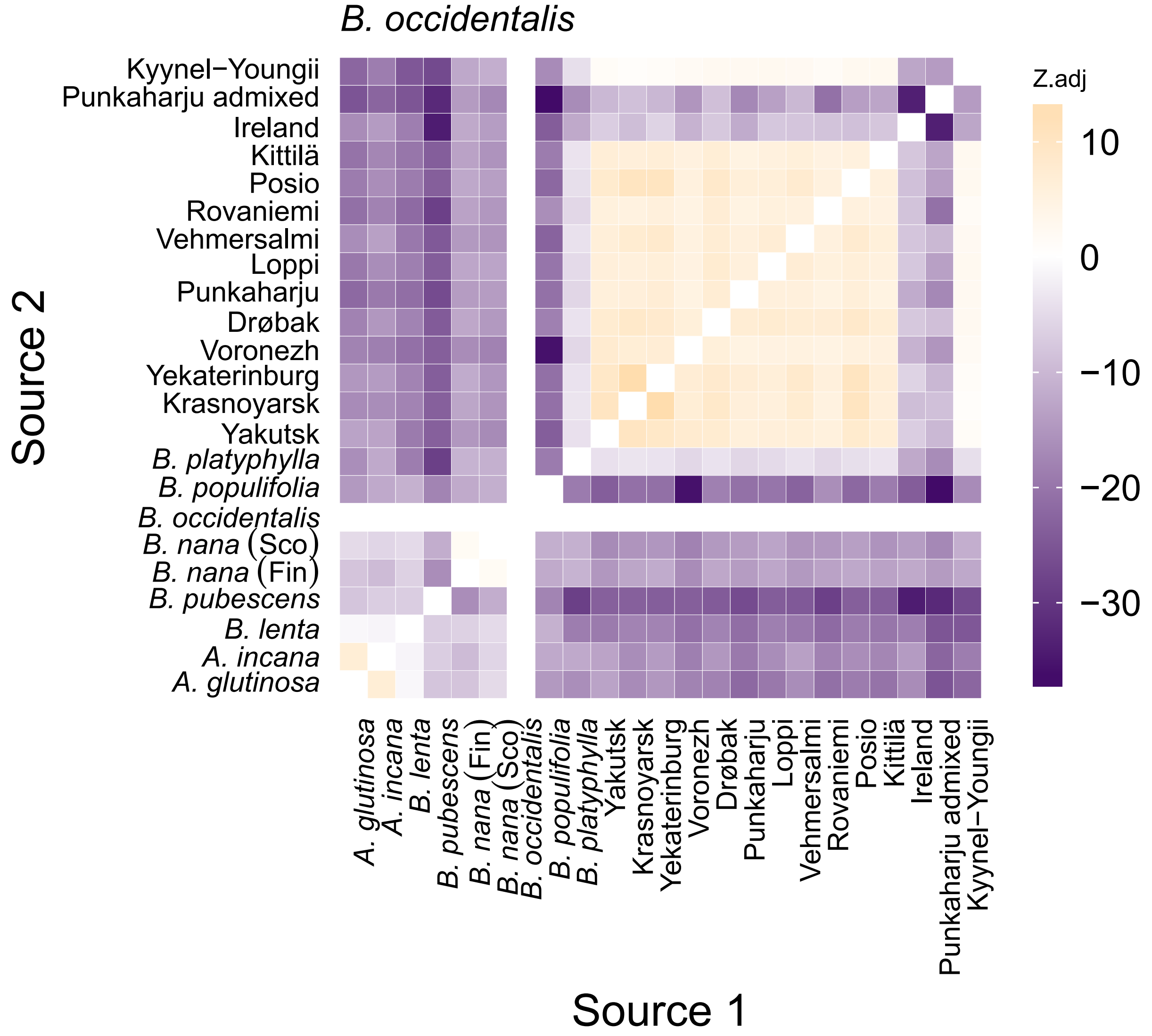
Jarkko Salojärvi

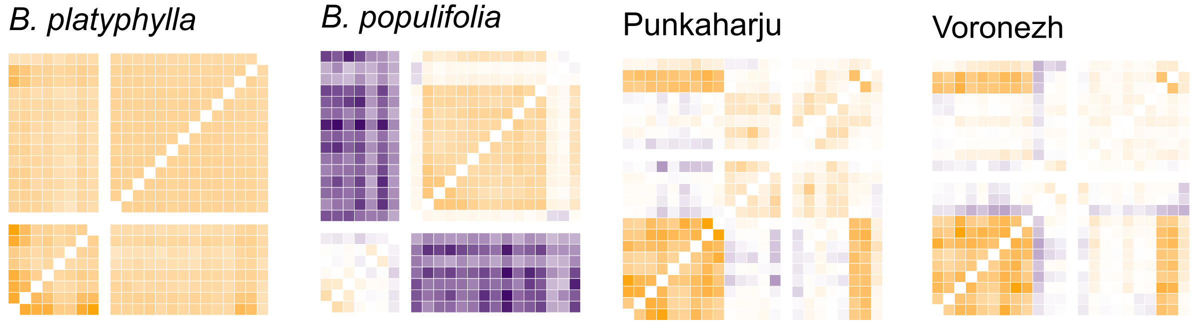
SNPs called with GATK

Birches Have a Complex Phylogeny with Admixed Species



F3 statistic



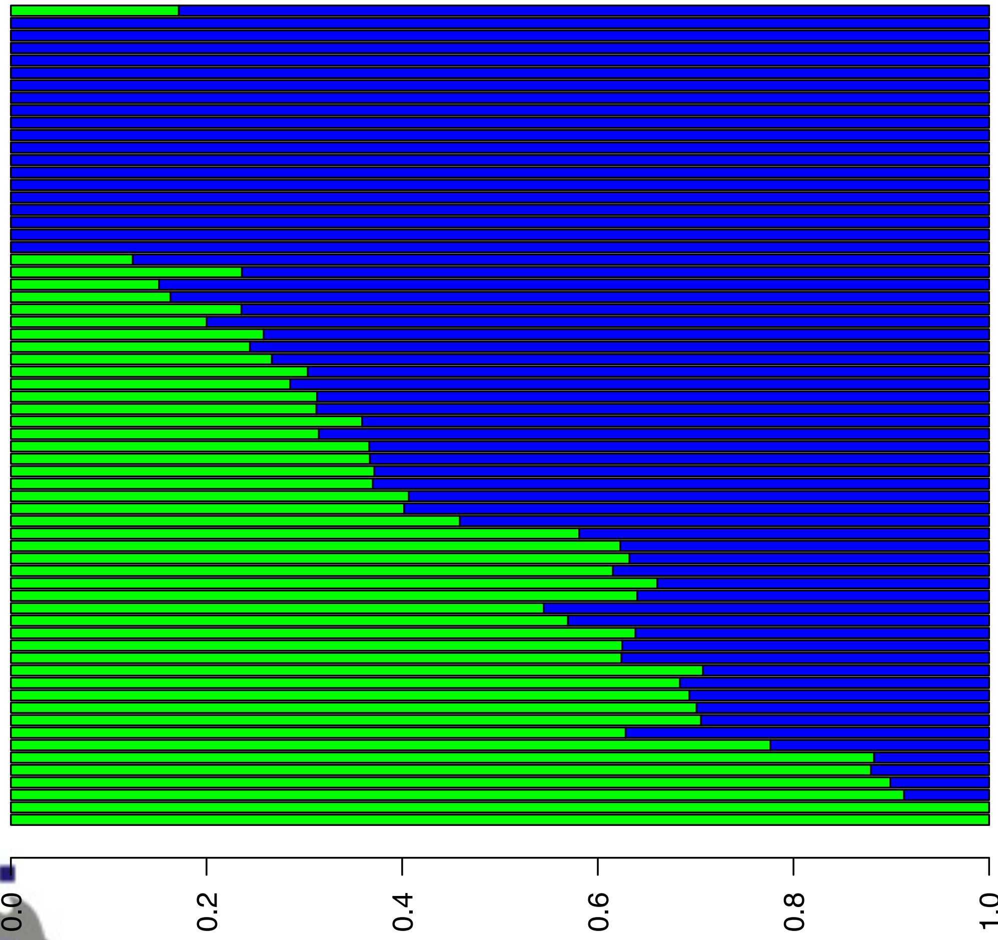


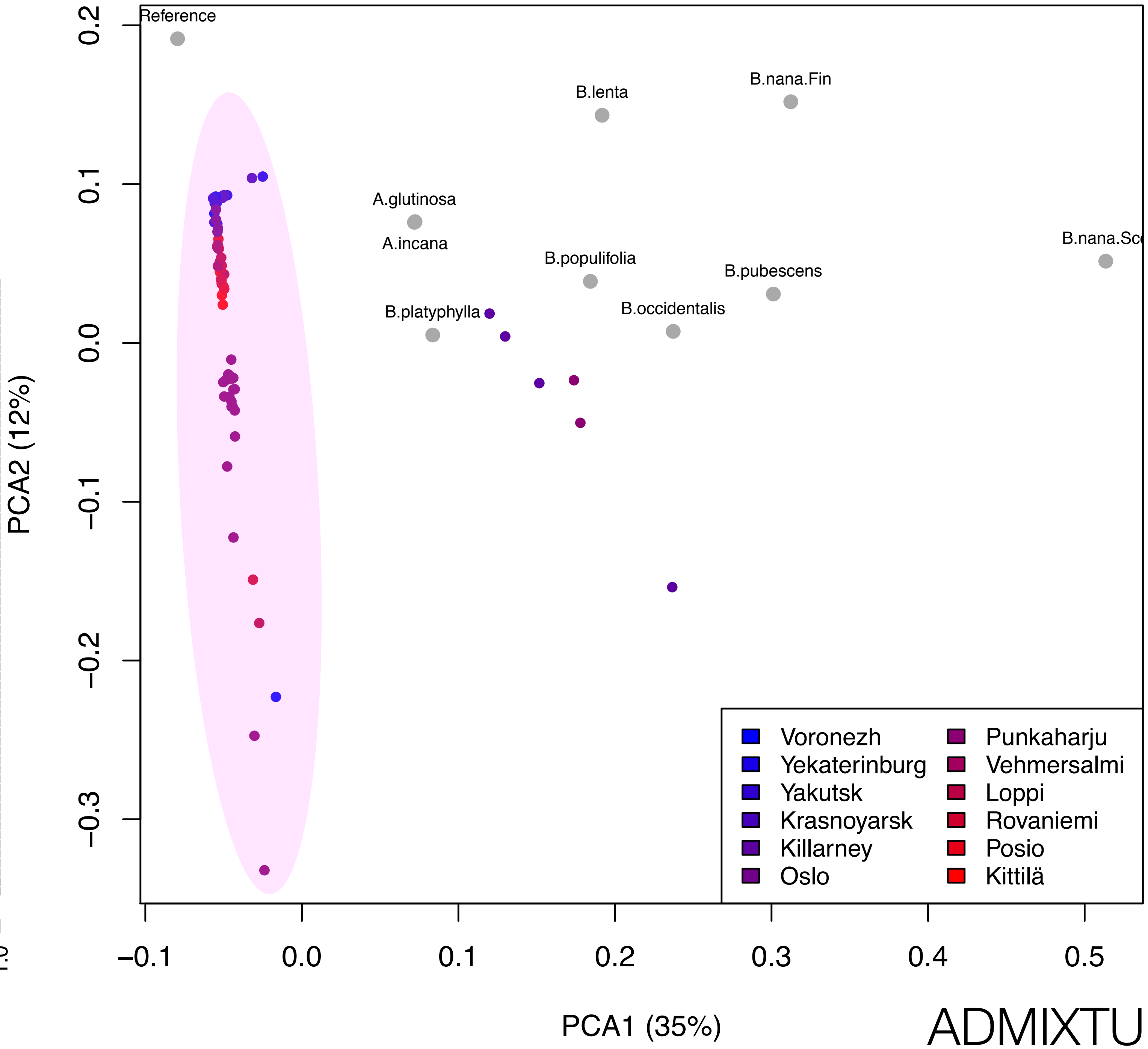
SNPs called with GATK

Jarkko Salojärvi

Birch Has Weak Population Structure

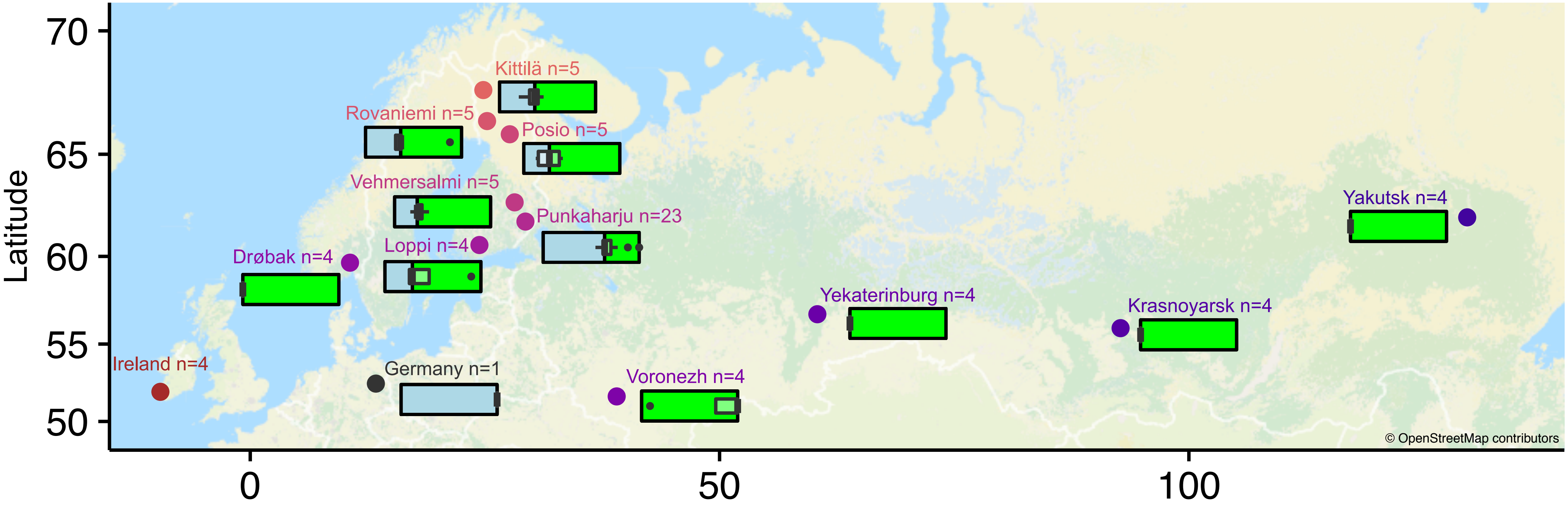
- Outcrossing, wind-pollinated species with over 1,000 km pollen dispersal.

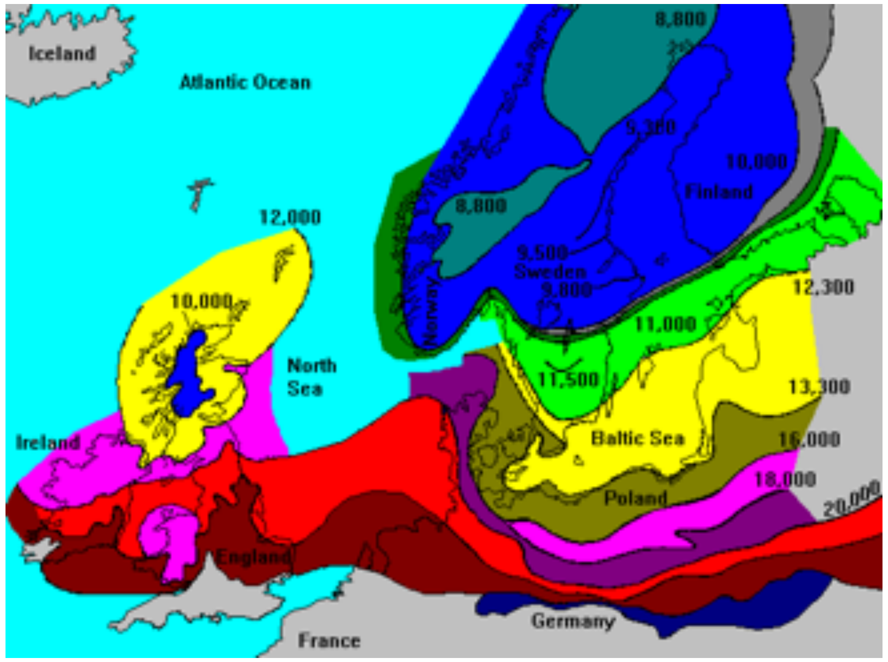




ADMIXTURE

Split into European and Siberian Populations



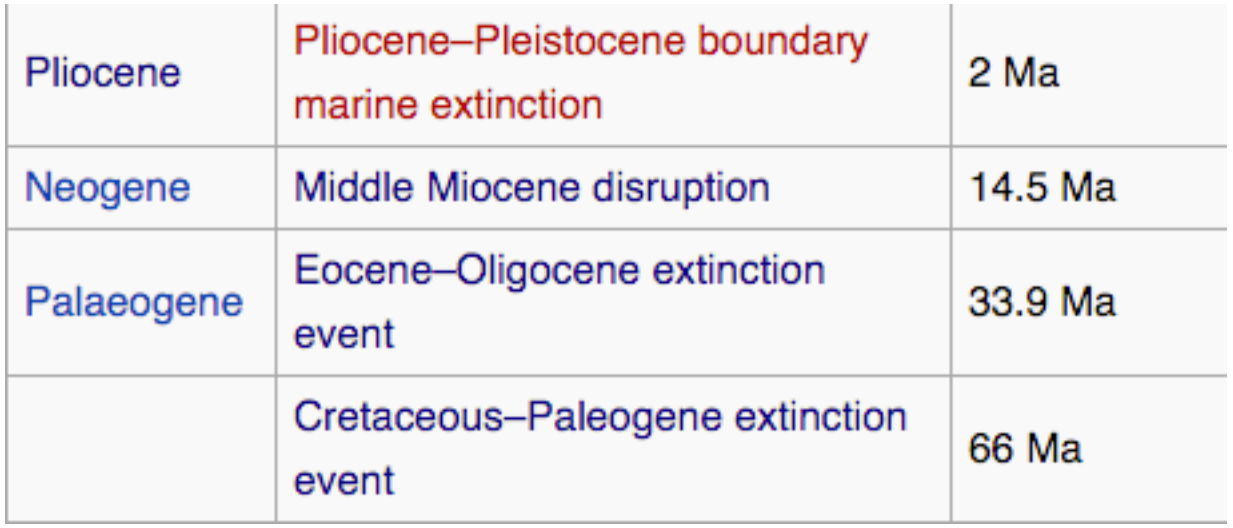


Population History

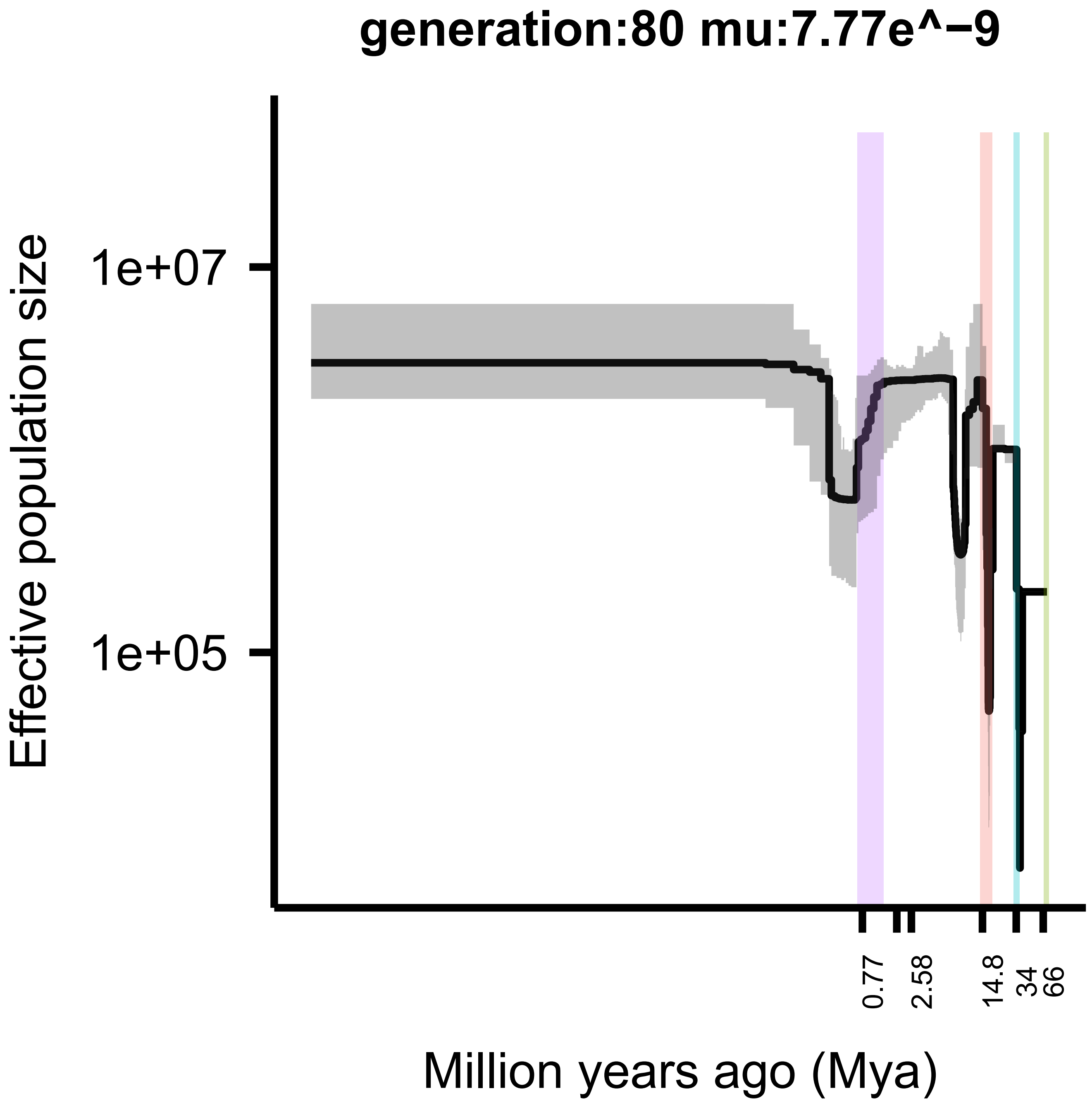
- Demographic history estimated with Stairway Plot.
  - Unfolded site-frequency spectrum estimate (ANGSD) with ancestral state reconstruction (Phangorn, R).
- Substitution rate and generation time of birch is not known.
  - Large range of possible substitution rates, from 2.7×10^{-9} to 3.5×10^{-8}.
    - Best guess from peach: 7.77×10^{-9} mutations / (Bp x generation).
      - Same rate as Arabidopsis.
  - Generation time also unknown, could be between 10 and 80 years.

Population History - Four Major Bottlenecks

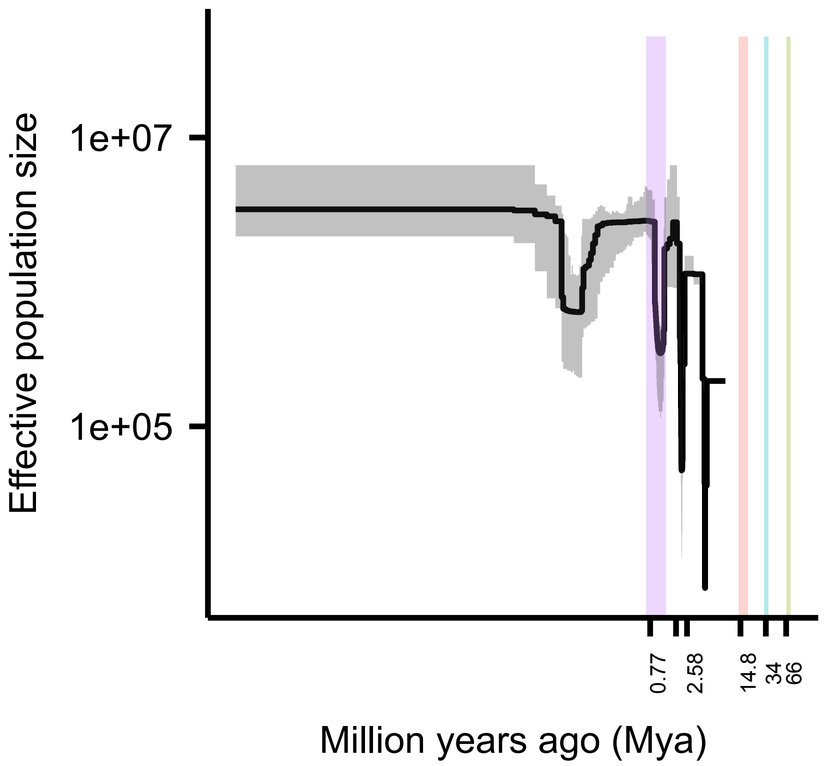
- Population bottlenecks appear to co-occur with events of climatic upheaval.

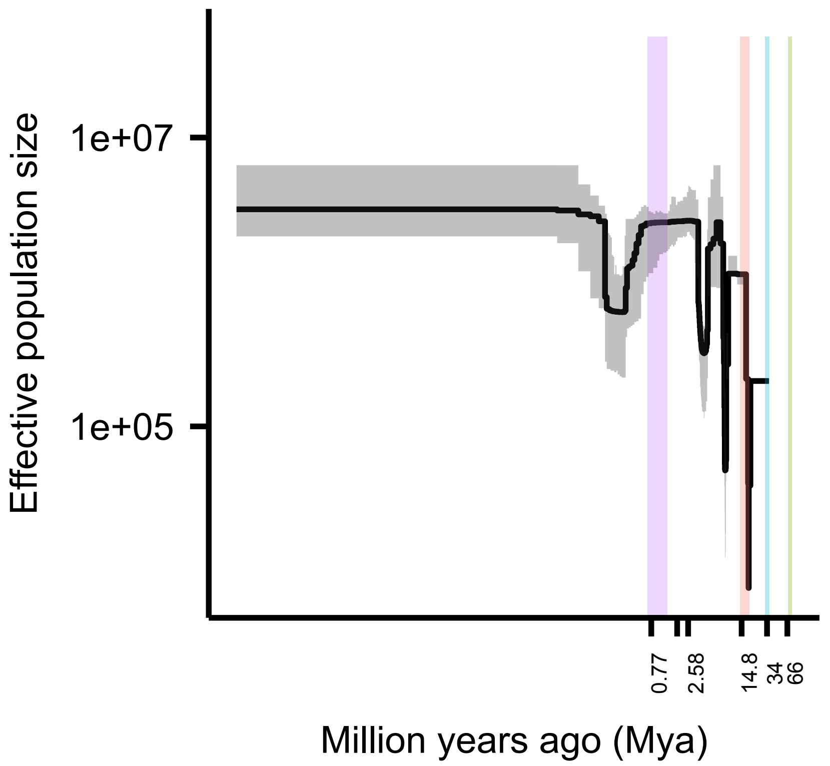


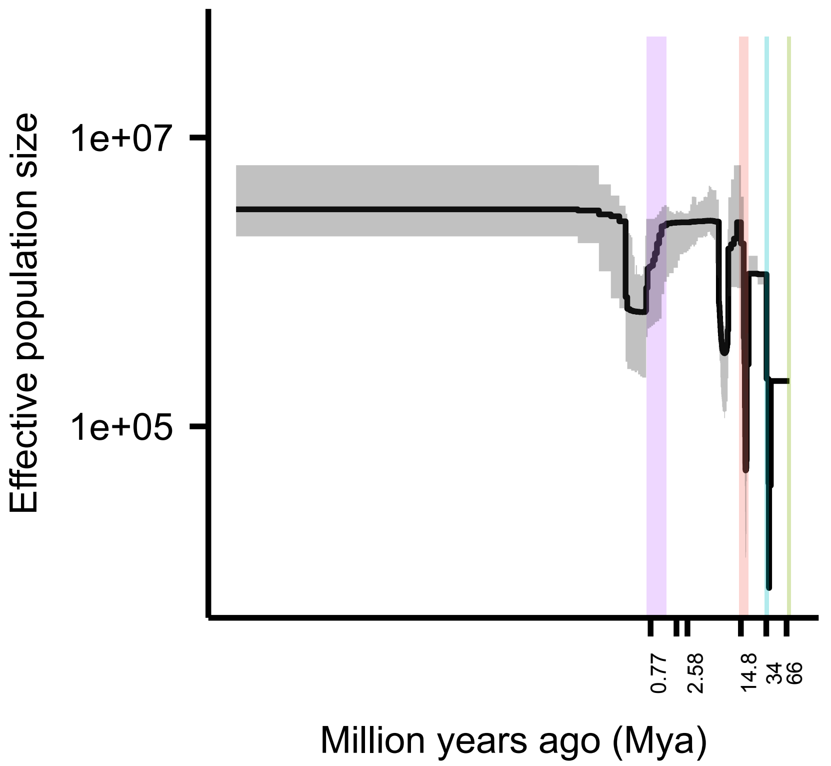
Pliocene	Pliocene–Pleistocene boundary marine extinction	2 Ma
Neogene	Middle Miocene disruption	14.5 Ma
Palaeogene	Eocene–Oligocene extinction event	33.9 Ma
	Cretaceous–Paleogene extinction event	66 Ma



Timing of bottlenecks?

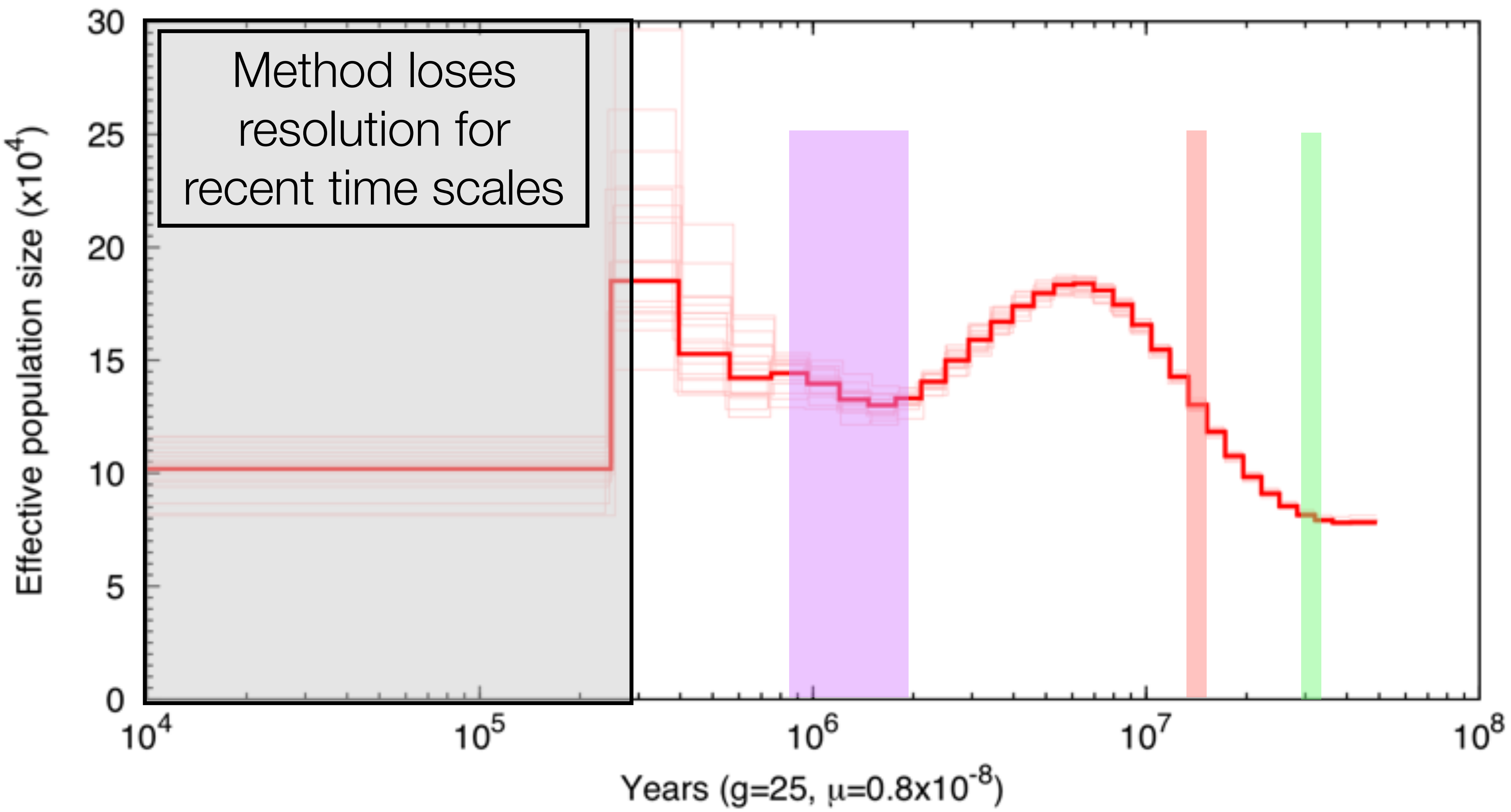

generation:10 mu:7.77e^−9


generation:40 mu:7.77e^−9


generation:80 mu:7.77e^−9

Jarkko Salojärvi

PSMC



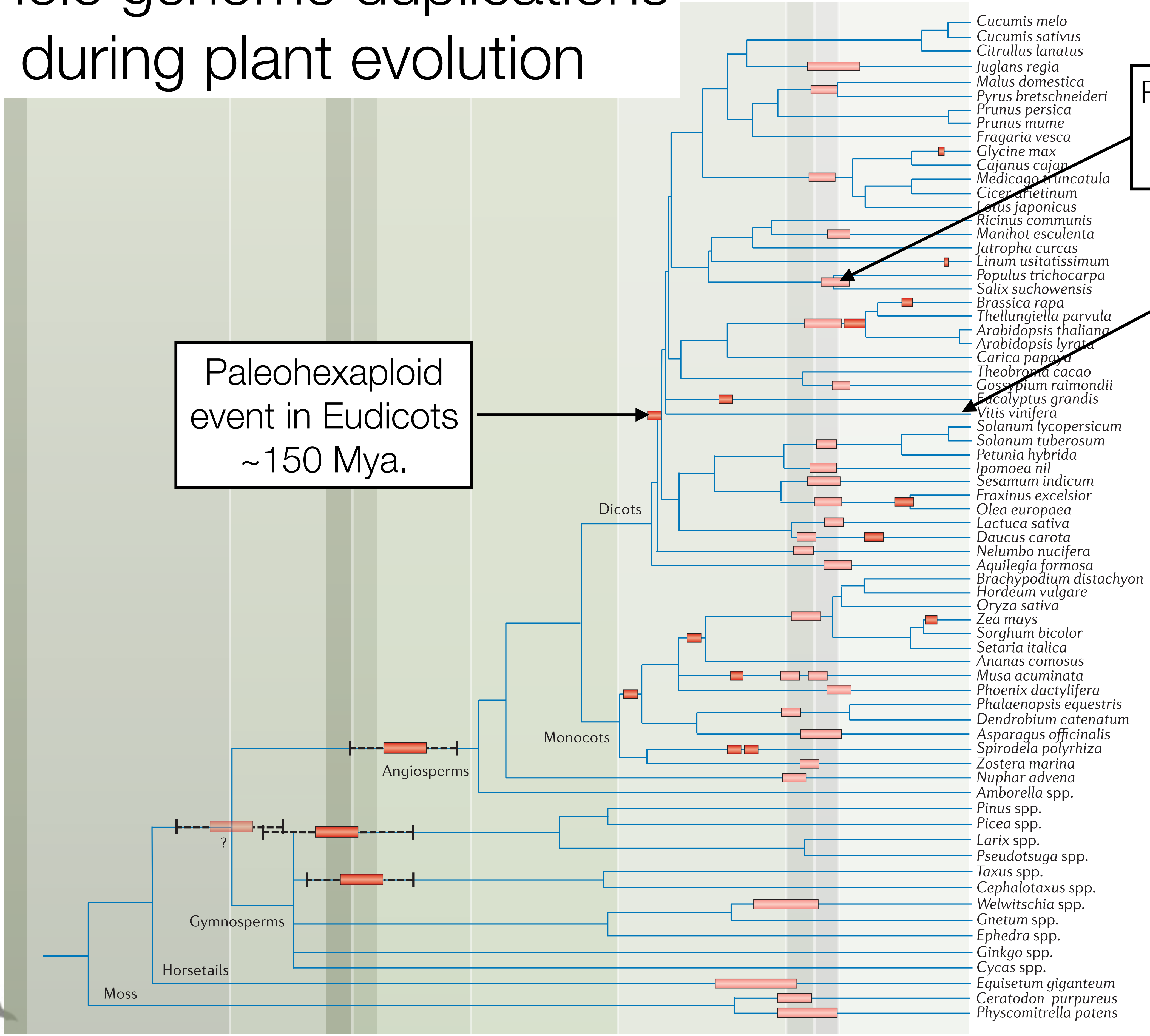
Method loses resolution for recent time scales

Jarkko Salojärvi

Luis Leal & Martin Lascoux (Unpublished)

Plant adaptation and evolution

Whole genome duplications during plant evolution



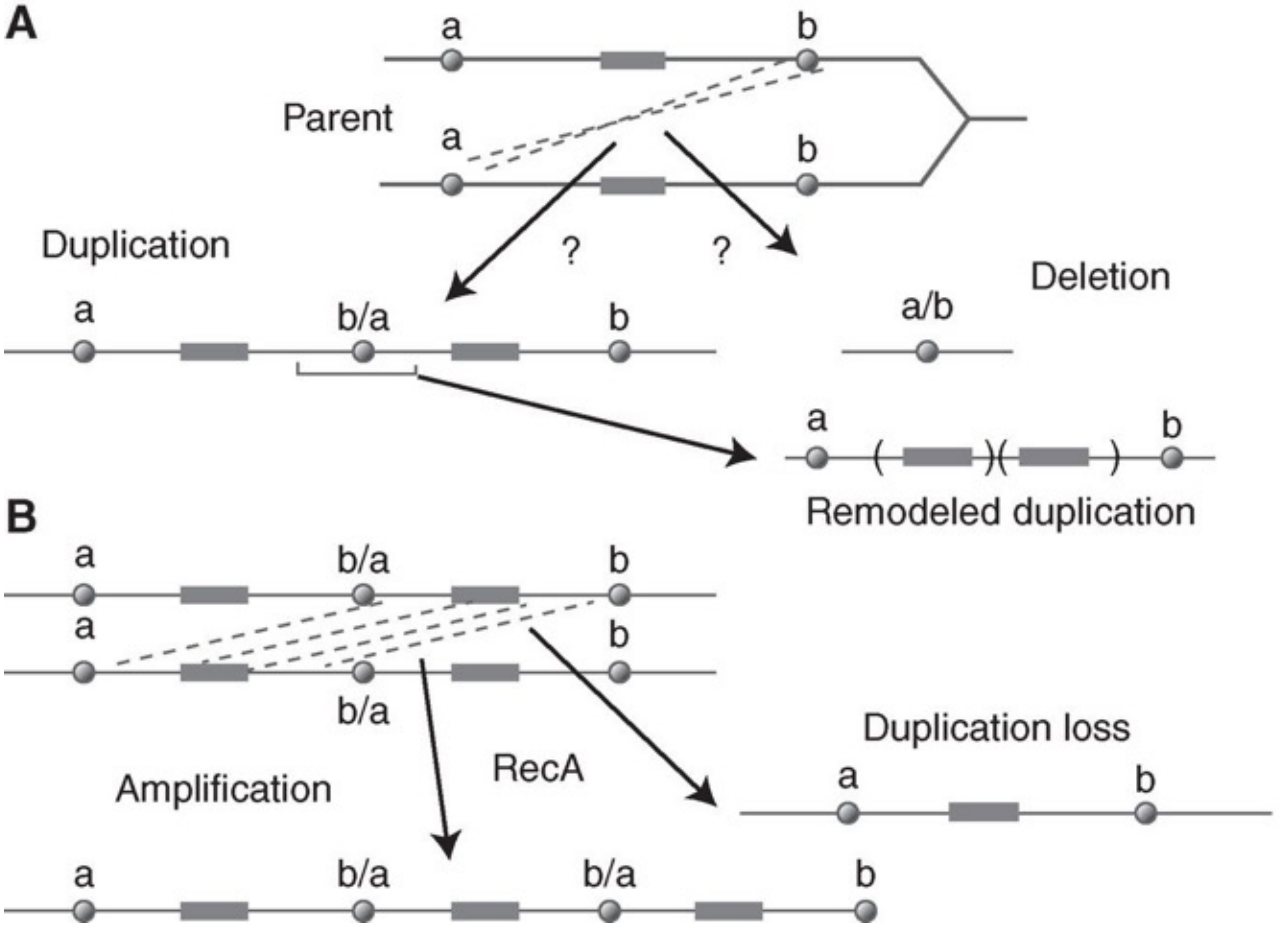
Poplar has a recent whole genome duplication event.

No recent WGDs in grape

Paleohexaploid event in Eudicots ~150 Mya.

Tandem duplication: local amplification

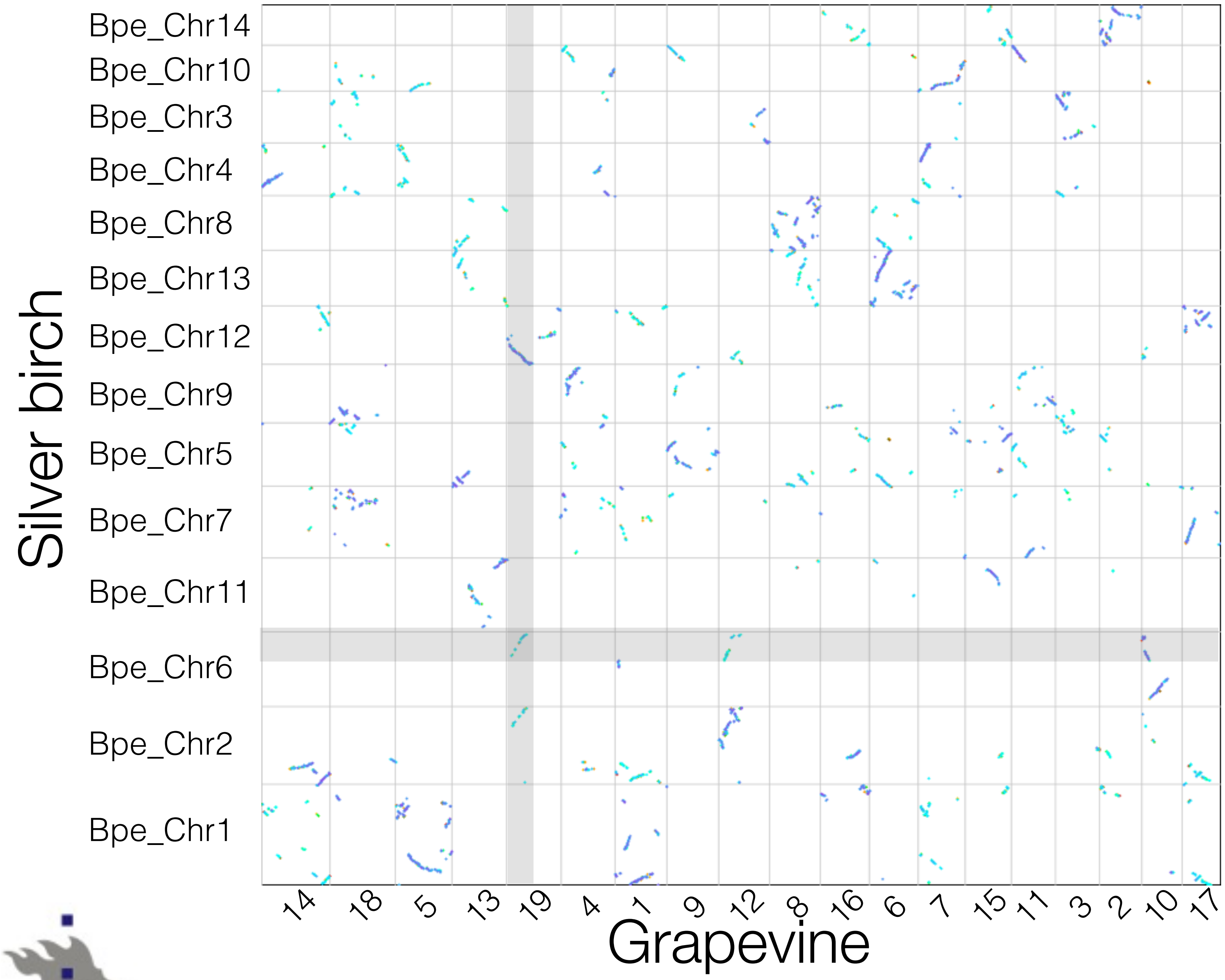
- Mostly mediated by homologous recombination

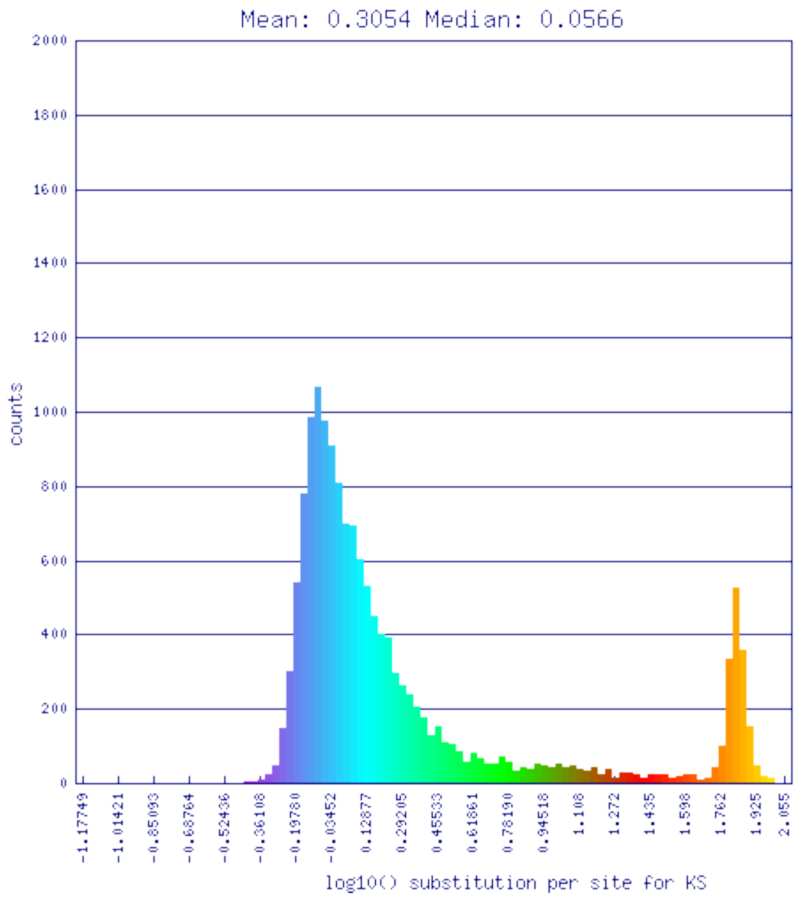


Jarkko Salojärvi

Beams&Roth, 2015

No Recent Whole Genome Duplication Events in birch

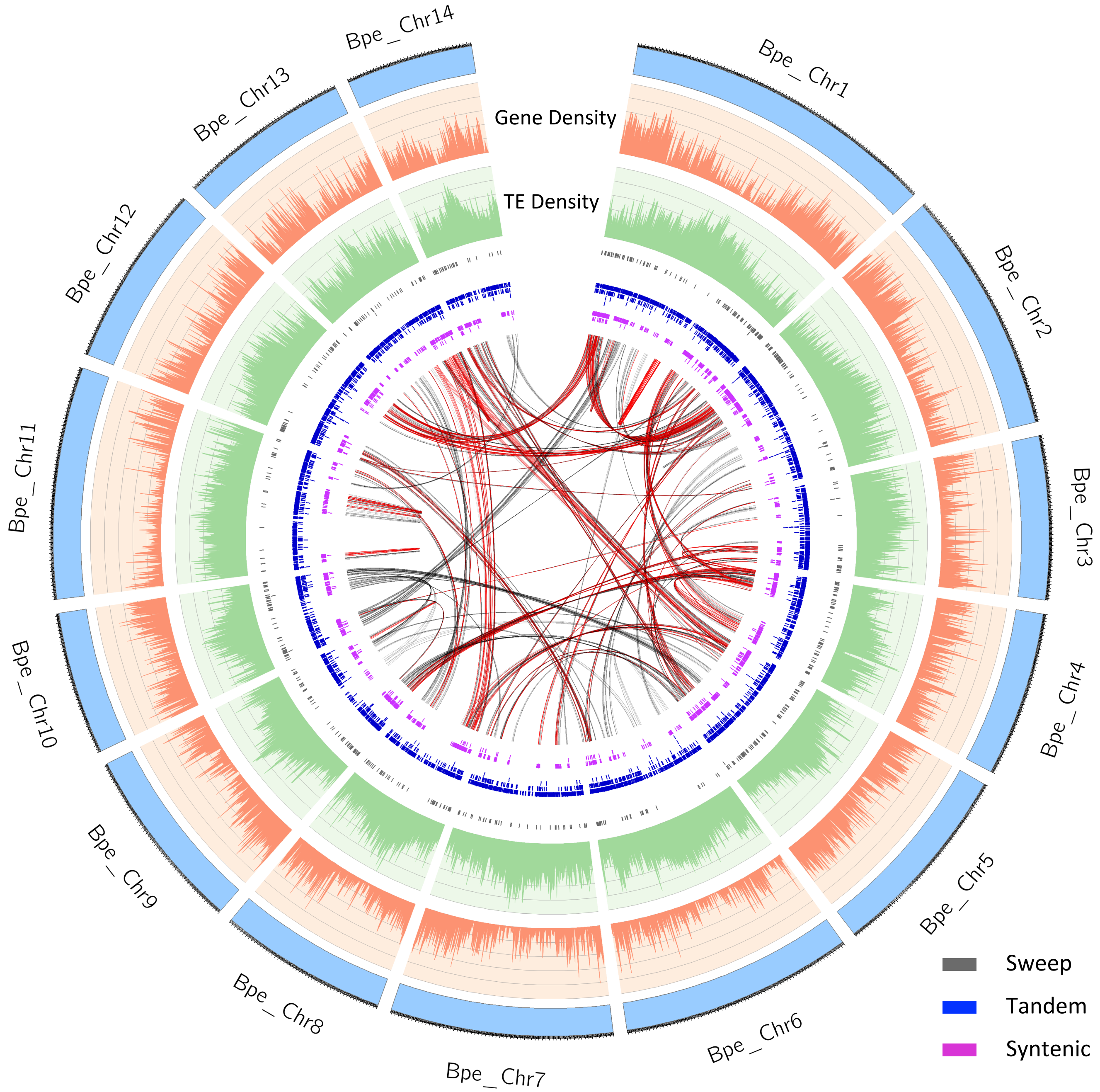




- 9,973 genes follow 3:3 syntenic relationship with Grapevine
- Longest syntenic block 358 genes

Jarkko Salojärvi

Functional Bias According to the Duplication Origins



Syntenic: Enriched for transcriptional regulation

Tandem expansions: Enriched for environmental responses and secondary metabolism.

Holds also for most other species.

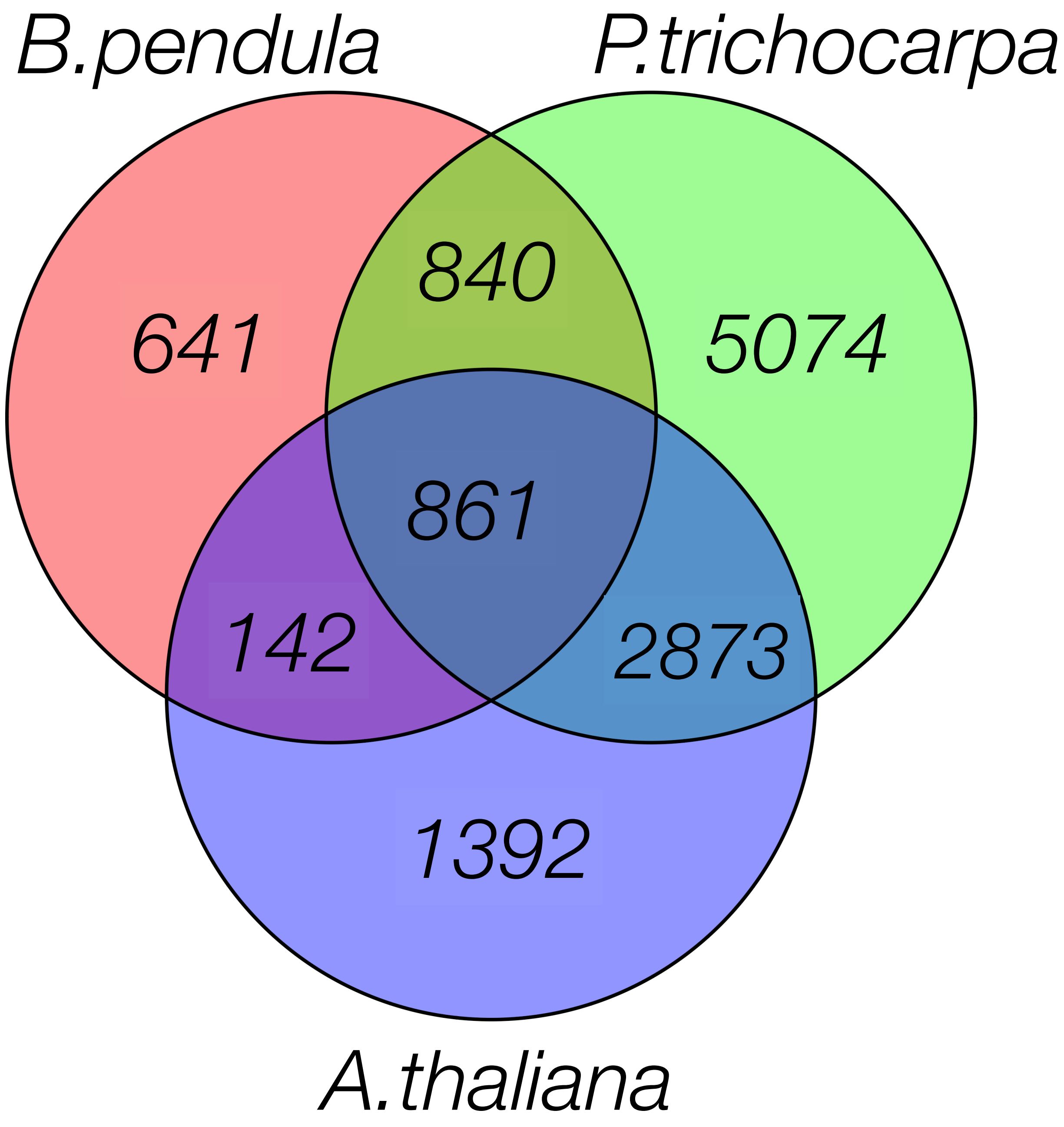
Jarkko Salojärvi

Salojärvi et al. Nat.Genetics 2017

Convergent Tandem Expansions

- Significant overlap of orthogroups (Fisher test, p<2.2e-16)
- Enriched GOs associated with **environmental responses**.
- **Adaptations to tree-like lifestyle** visible in tandems shared by *B.pendula* and *P.trichocarpa* (fungal defences, cell wall biogenesis, cellulose synthase).

N orthogroups



Salojärvi et al. Nat.Genetics 2017

How does the genome evolve..?

- Genes retained after WGDs appear to follow a dosage balance hypothesis:
  - In complex regulatory networks and protein complexes the stoichiometric balance between the different components needs to be preserved.
  - Therefore selection acts against losses after WGDs and against tandem duplications.
- Other genes evolve through local expansions, such as tandem duplications.

Selective Sweeps: Recent Adaptation

- Sweepfinder2 + post processing pipeline adapted from the analysis in great tit genome (Laine et al 2016).
    - Careful curation of sweeps - removal of organellar insertions and transposable elements.
- This resulted in 913 genes, out of which 841 had orthologs in Arabidopsis.

In Birch, Selection Acts by Tuning the Timing and Crosstalk Between Different Processes

- **Regulators and receptors** holding key positions in triggering developmental or physiological chains of events.
- **3 GOs** significantly enriched (44 genes), related to related to wood and fiber development, light sensing, embryogenesis, and reproductive isolation.

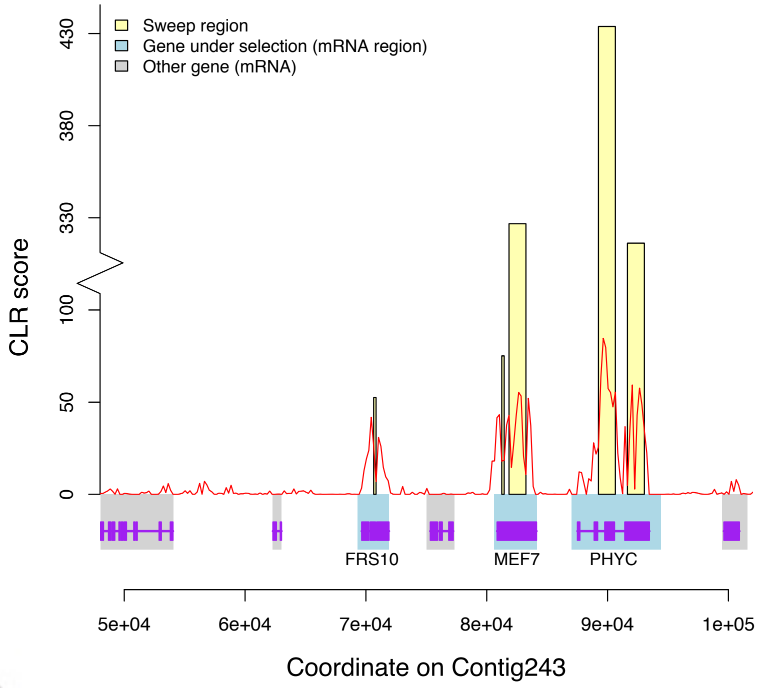
In Birch, Selection Acts by Tuning the Timing and Crosstalk Between Different Processes

- Further analysis was carried out to test how many of the sweeps can be explained by population structure.
    - Significant contribution in 423 genes, 490 genes unexplained.
- We further correlated the SNPs around sweep regions with temperature, precipitation, and geographic coordinates of the sites.
    - SIX genes with high statistical significance: PHYTOCHROME C, SWEETIE, KAKTUS, MED5A, RESPONSE REGULATOR 1, FAR RED SEQUENCE 10

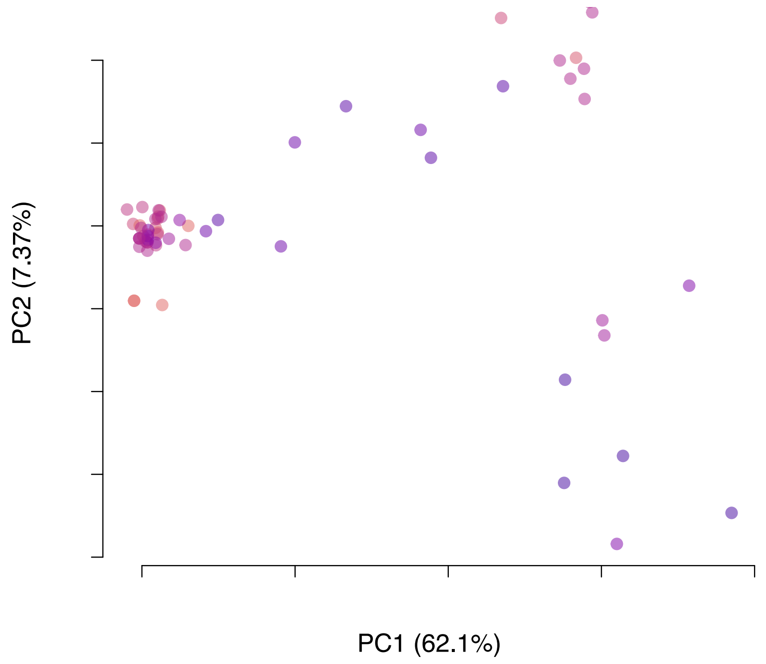
Example Sweep Region: PHYC

- In Arabidopsis, Phytochrome C is connected to temperature-specific regulation of the circadian clock, and strongly linked to altitudinal, and latitudinal-longitudinal clines in flowering time.
- Population structure does not explain the SNP pattern around the sweep (5.3%, NS).
- Maximum temperature explains a significant proportion (26% RDA, p.adj=0.01)

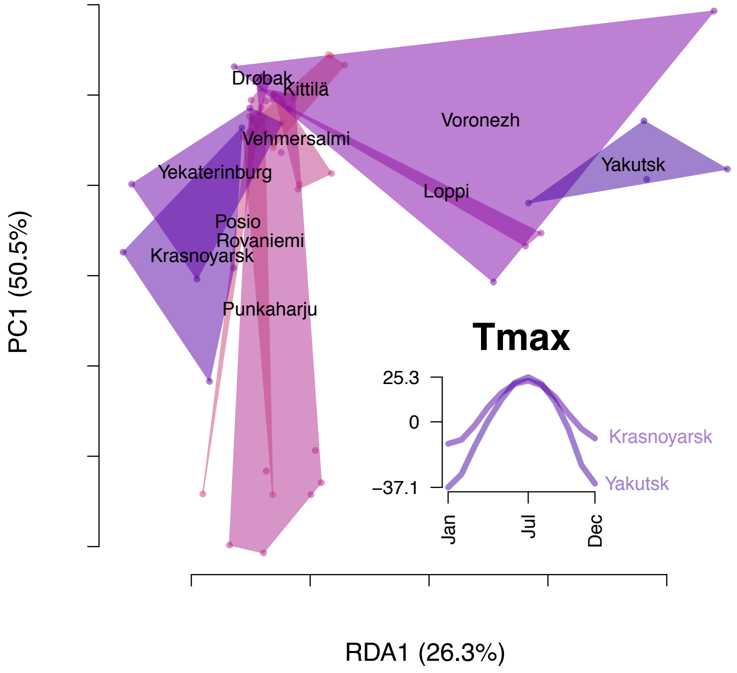
Sweep region



PCA of SNPs around sweep region



RDA



Jarkko Salojärvi

Connection between sweeps and gene origins?

- Enrichment of **sweeps in syntenic genes** from paleohexaploid event, as well as in non-tandemly duplicated genes.
    - Young birch-specific genes significantly depleted.
    - Tandems not enriched.

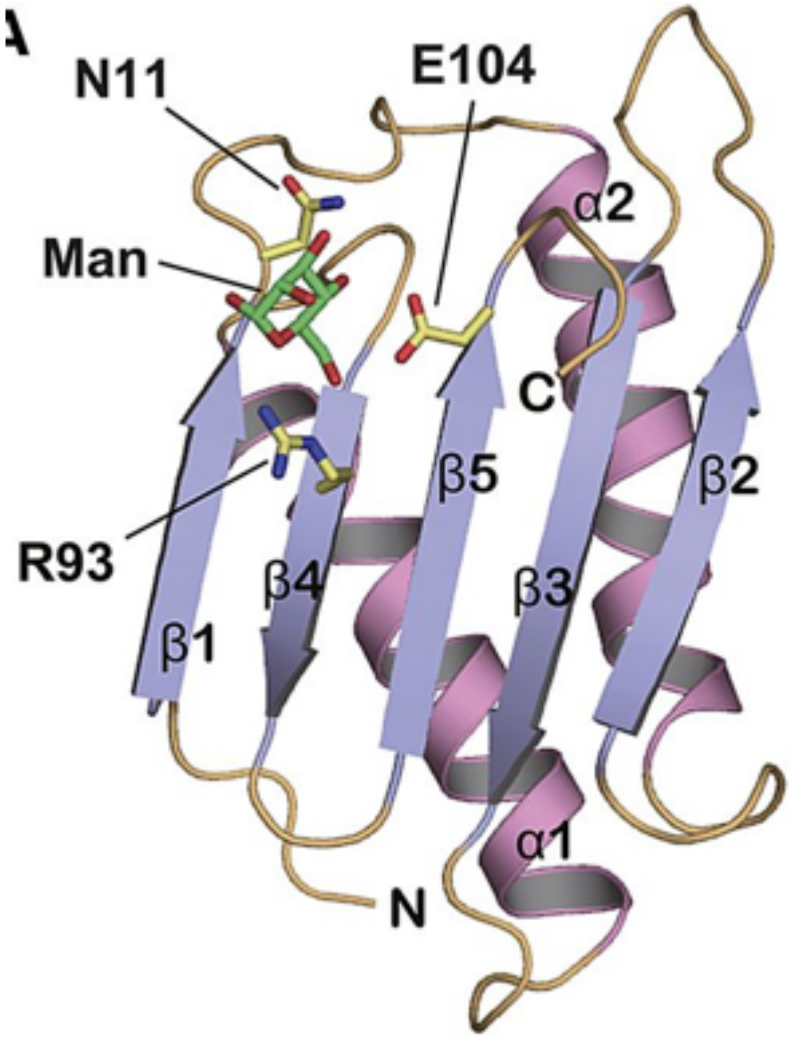
How does the genome evolve..?

- Genes retained after WGDs appear to follow a dosage balance hypothesis:
  - In complex regulatory networks and protein complexes the stoichiometric balance between the different components needs to be preserved.
  - Therefore selection acts against losses after WGDs and against tandem duplications.
- Tandems: Exploration - exploitation hypothesis:
  - Exploration: tandems are evolving with low evolutionary selection pressure (copy number variants)
    - Different mechanism for selection (homologous recombination, decreased LD)
  - Exploitation: at specific times in history, rapid environmental changes require exploitation of the exploratory gene space -> fixation, sweeps.

Evolution of a gene family

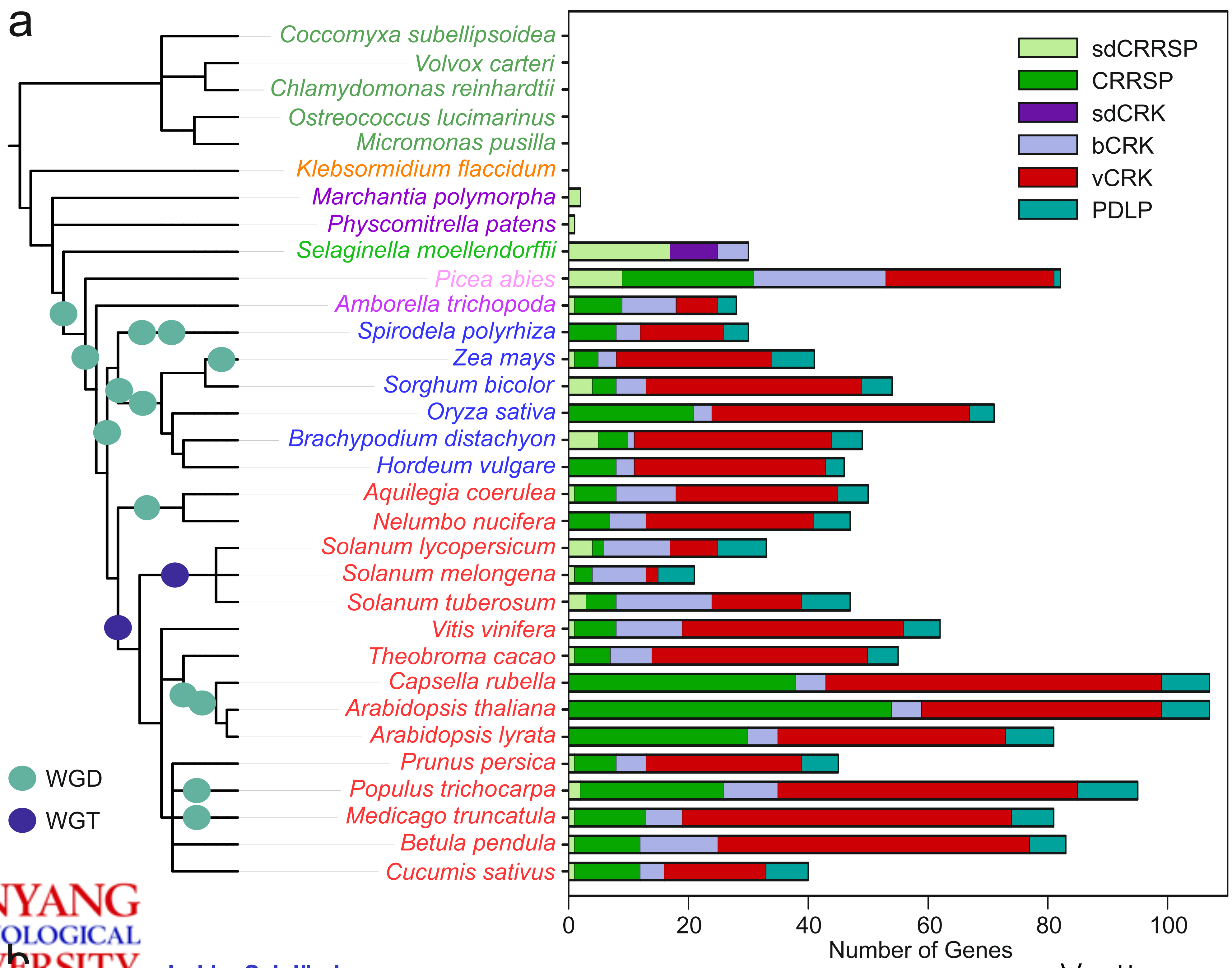
DUF26: Stress-antifung domain

- Analysis on a protein in *Ginko biloba*: A mannose-binding lectin that exhibits antifungal activity.



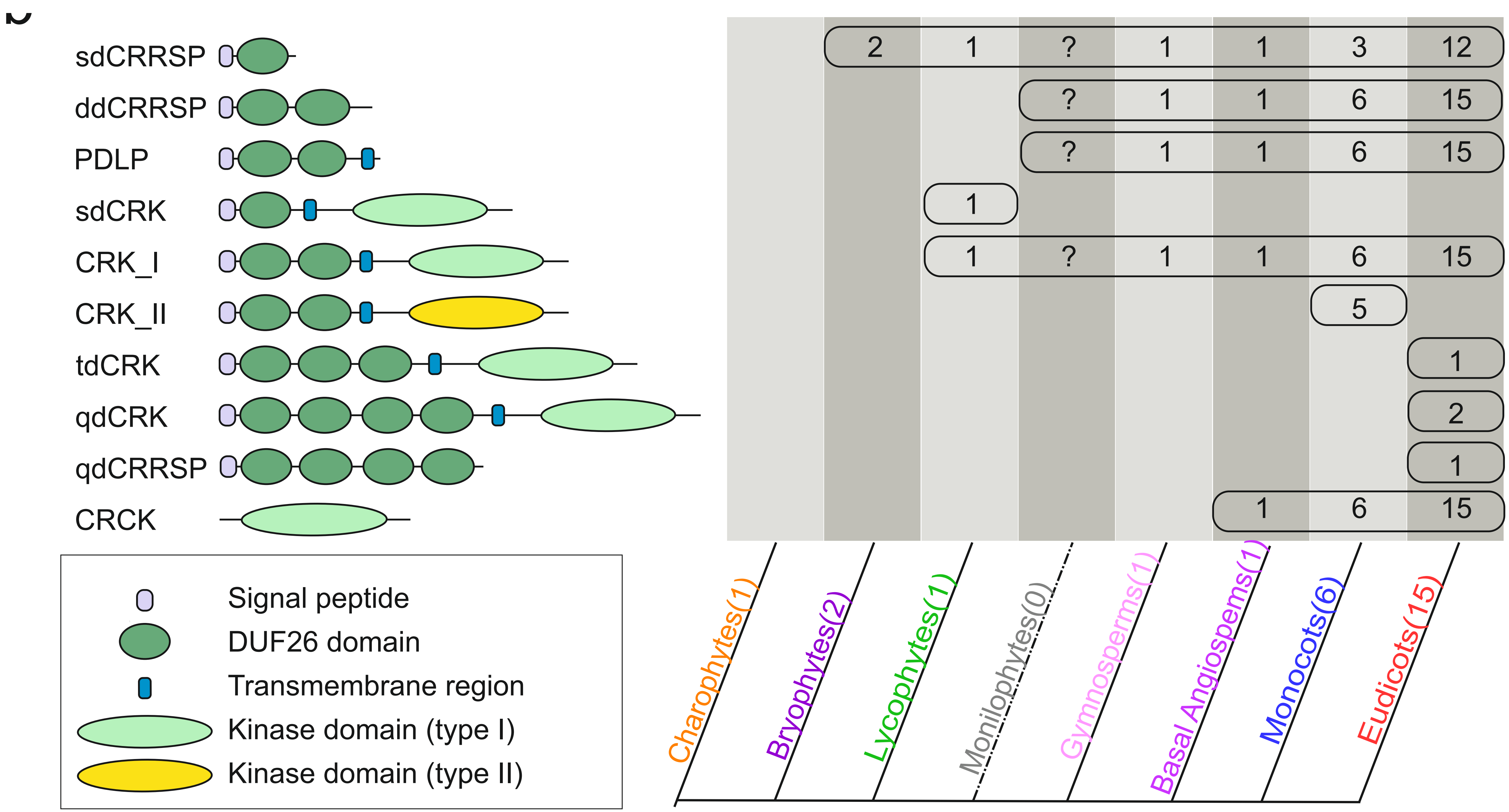
Conserved cystein motif:
C-8X-C-2X-C

DUF26 is widely present throughout the plant kingdom



Vaattovaara et al. submitted

…and mostly in 2xDUF26 configuration



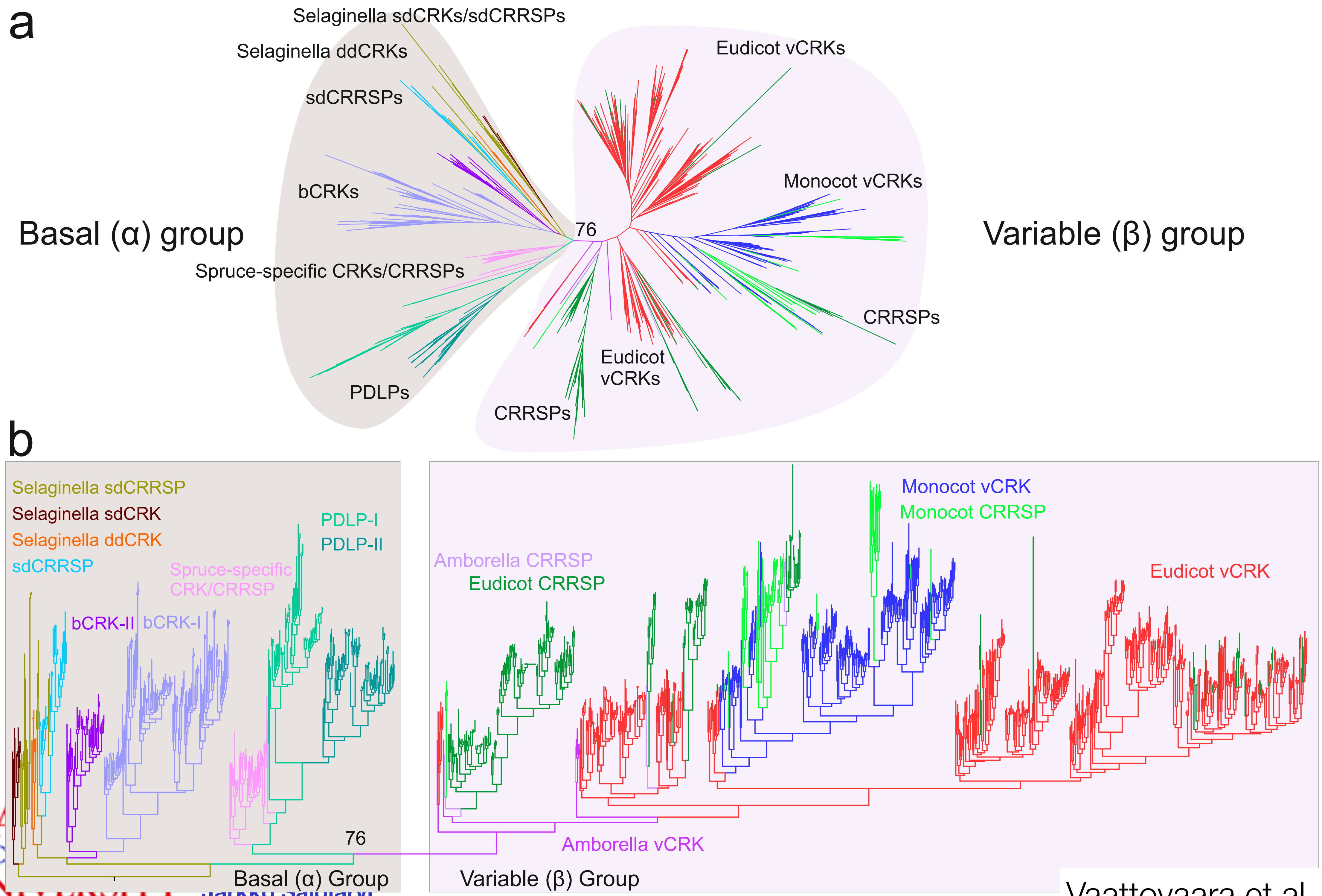
	Charophytes(1)	Bryophytes(2)	Lycophytes(1)	Monilophytes(0)	Gymnosperms(1)	Basal Angiosperms(1)	Monocots(6)	Eudicots(15)
sdCRRSP		2	1	?	1	1	3	12
ddCRRSP				?	1	1	6	15
PDLP				?	1	1	6	15
sdCRK			1					
CRK_I			1	?	1	1	6	15
CRK_II							5	
tdCRK								1
qdCRK								2
qdCRRSP								1
CRCK						1	6	15

NANYANG TECHNOLOGICAL UNIVERSITY

Jarkko Salojärvi

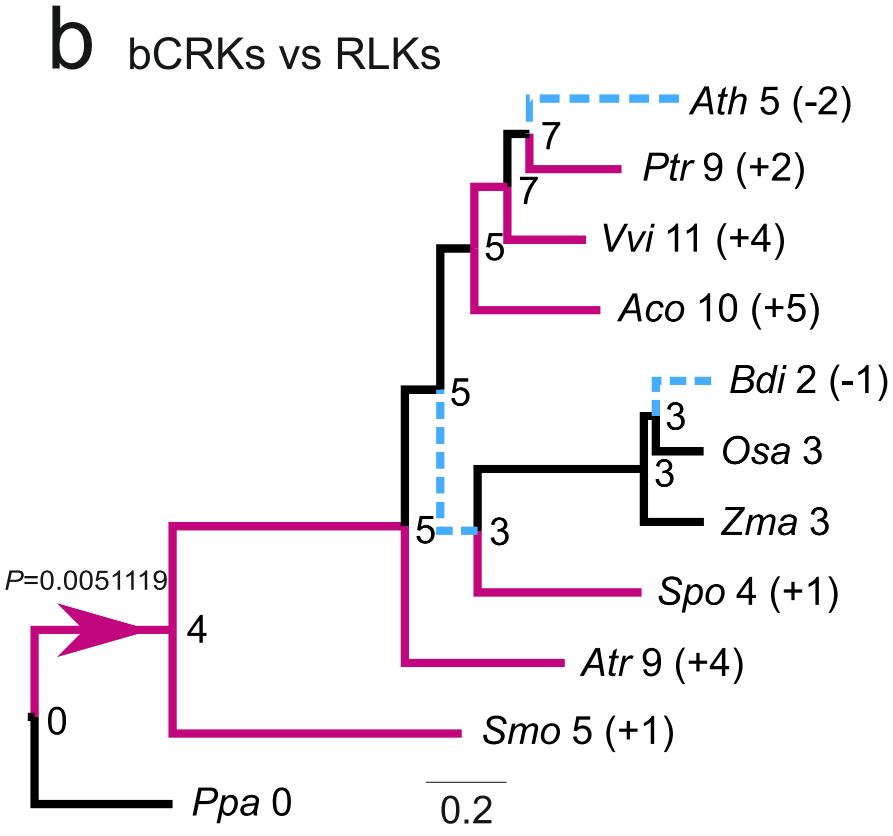
Vaattovaara et al. submitted

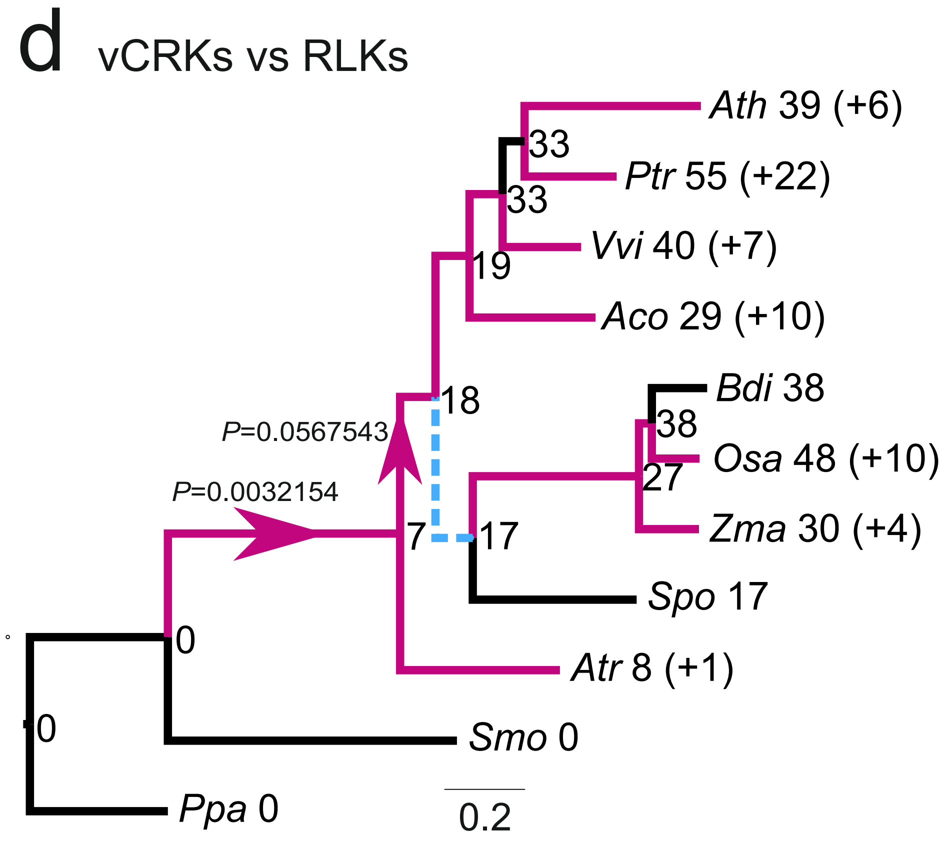
CRK radiation in angiosperms, many lineage-specific tandem expansions



Vaattovaara et al. submitted

Significant expansions in branches with WGD events





Vaattovaara et al. submitted

Evolution to harsh environmental conditions

Detecting early signs of heat and drought stress in *Phoenix dactylifera* (date palm)

1. Very important crop.
2. Resistant to extreme drought, heat and UV conditions.

Safronov et al. PLoS ONE, 2017

Effects of heat and drought on date palm seedlings under controlled conditions

Day/Night: 16h/8h:
20°C/15°C

Well-watered
Low temperature

Day/Night: 16h/8h:
35°C/15°C

Well-watered
High temperature

Transcriptomics

Day/Night: 16h/8h: **20°C/15°C**

Drought (9 days no water, 2 week acclimation).
Low temperature

Day/Night: 16h/8h: **35°C/15°C**

Drought (5 days no water, 2 week acclimation).
High temperature

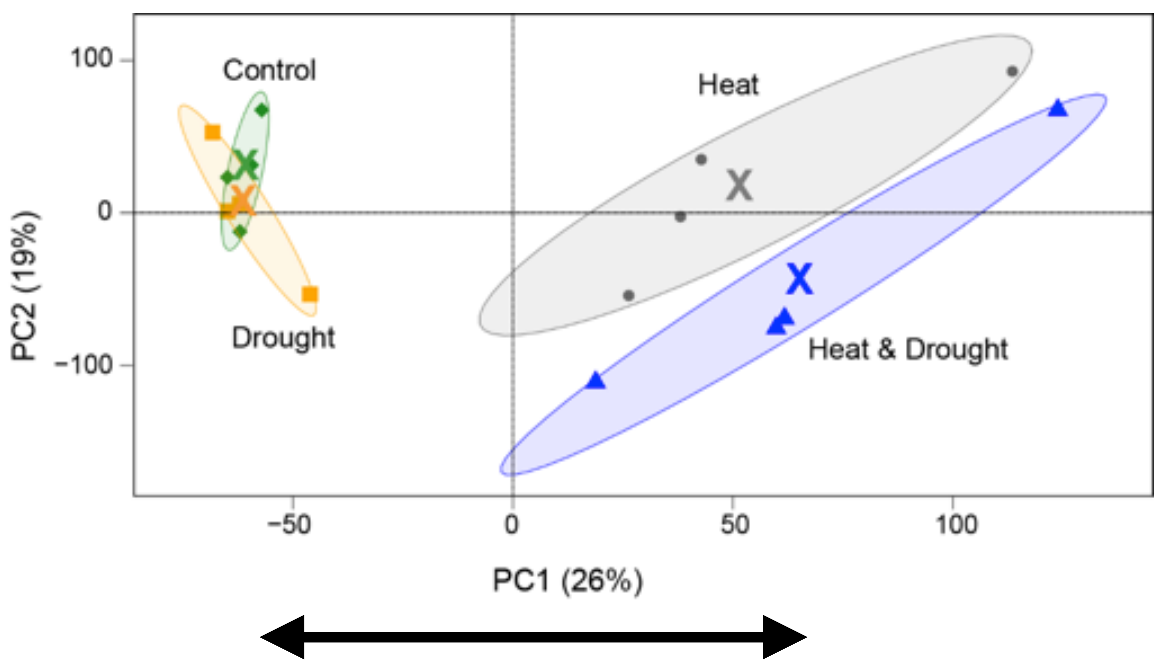
Metabolomics

Safronov et al. PLoS ONE, 2017

PCA of transcriptome and metabolome tell different stories

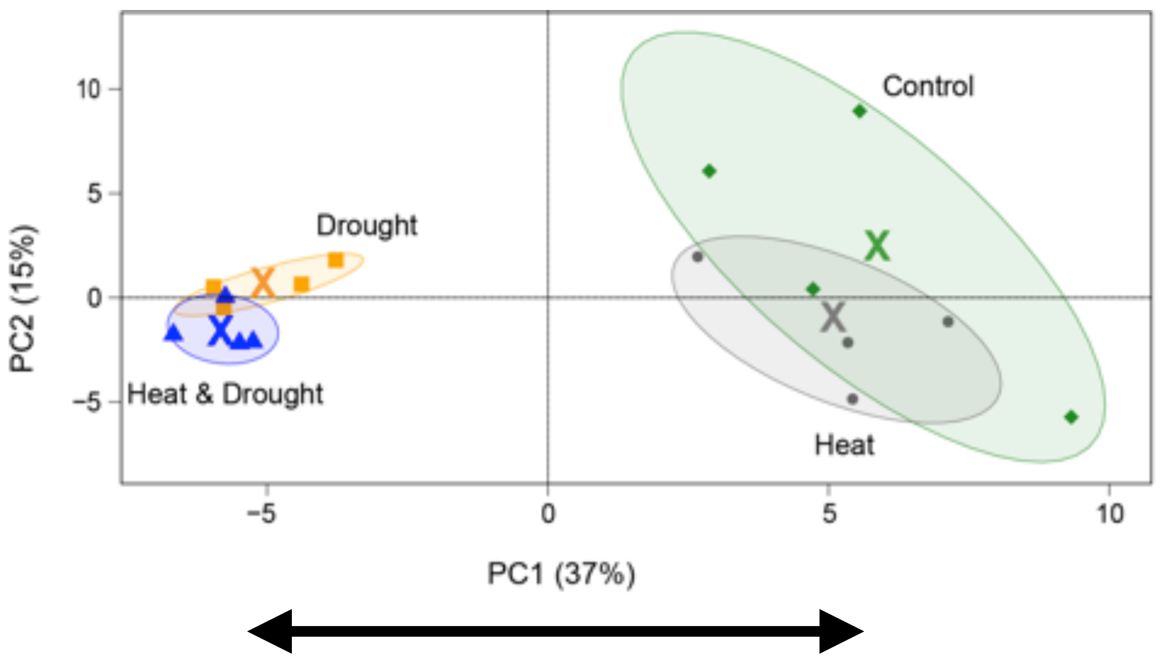
- Metabolome: immediate responses, transcriptome: slower, anticipatory response

Transcriptome



Heat

Metabolome



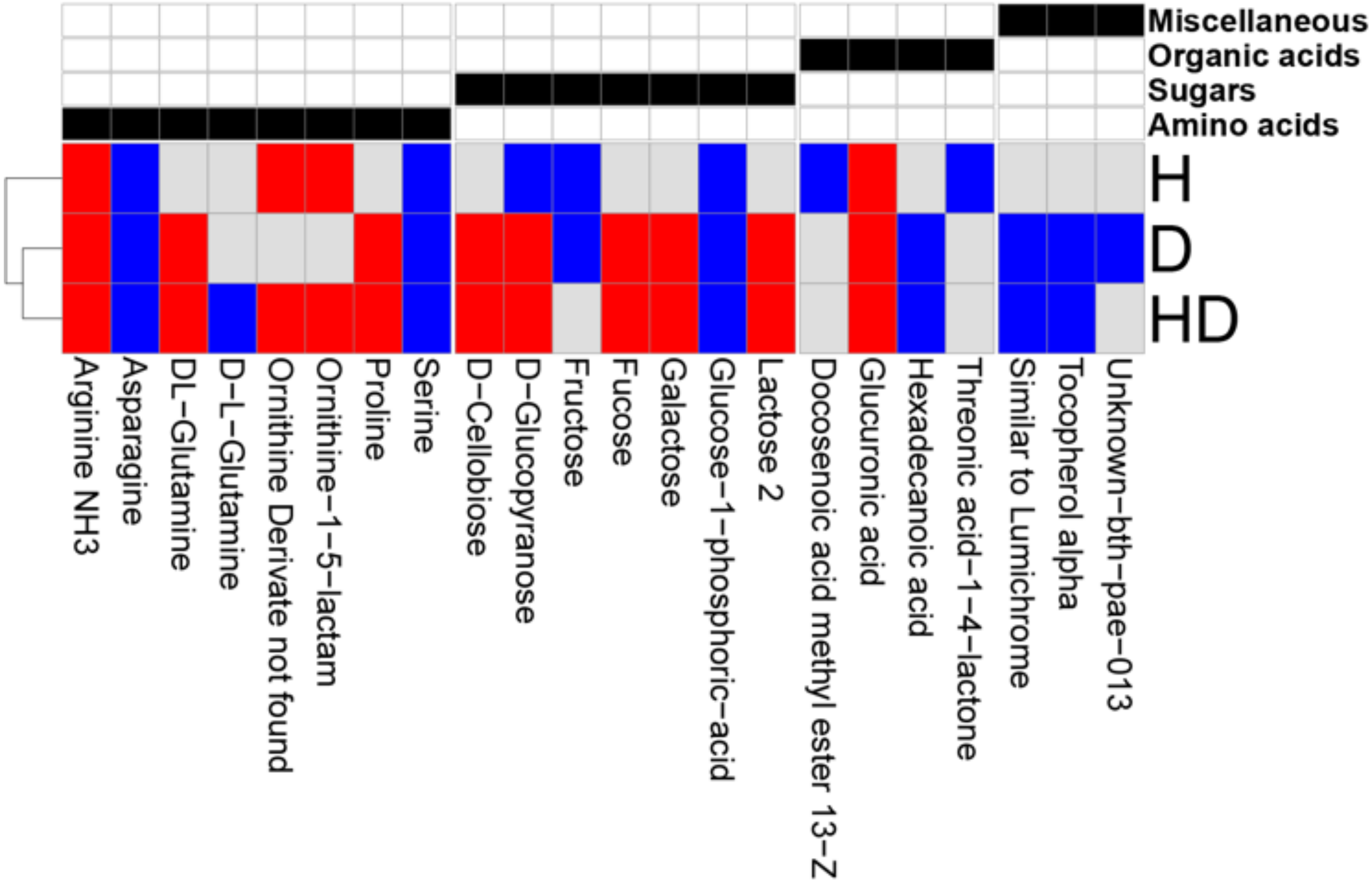
Drought

NANYANG TECHNOLOGICAL UNIVERSITY

Jarkko Salojärvi

Safronov et al. PLoS ONE, 2017

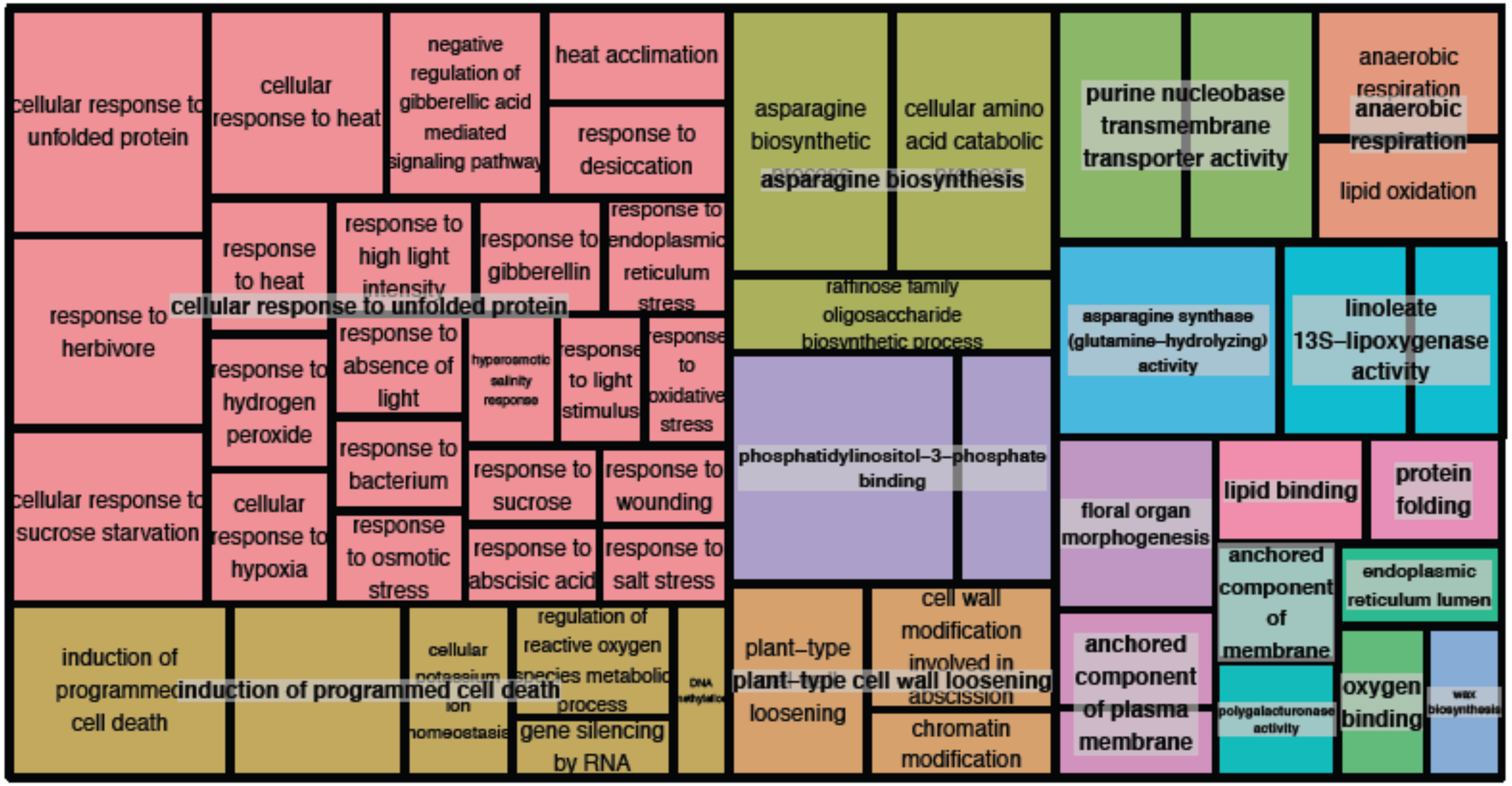
Carbohydrate metabolism activated by drought



Safronov et al. PLoS ONE, 2017

Gene expression: preparation for heat

- Expression of heat shock proteins and chaperones.

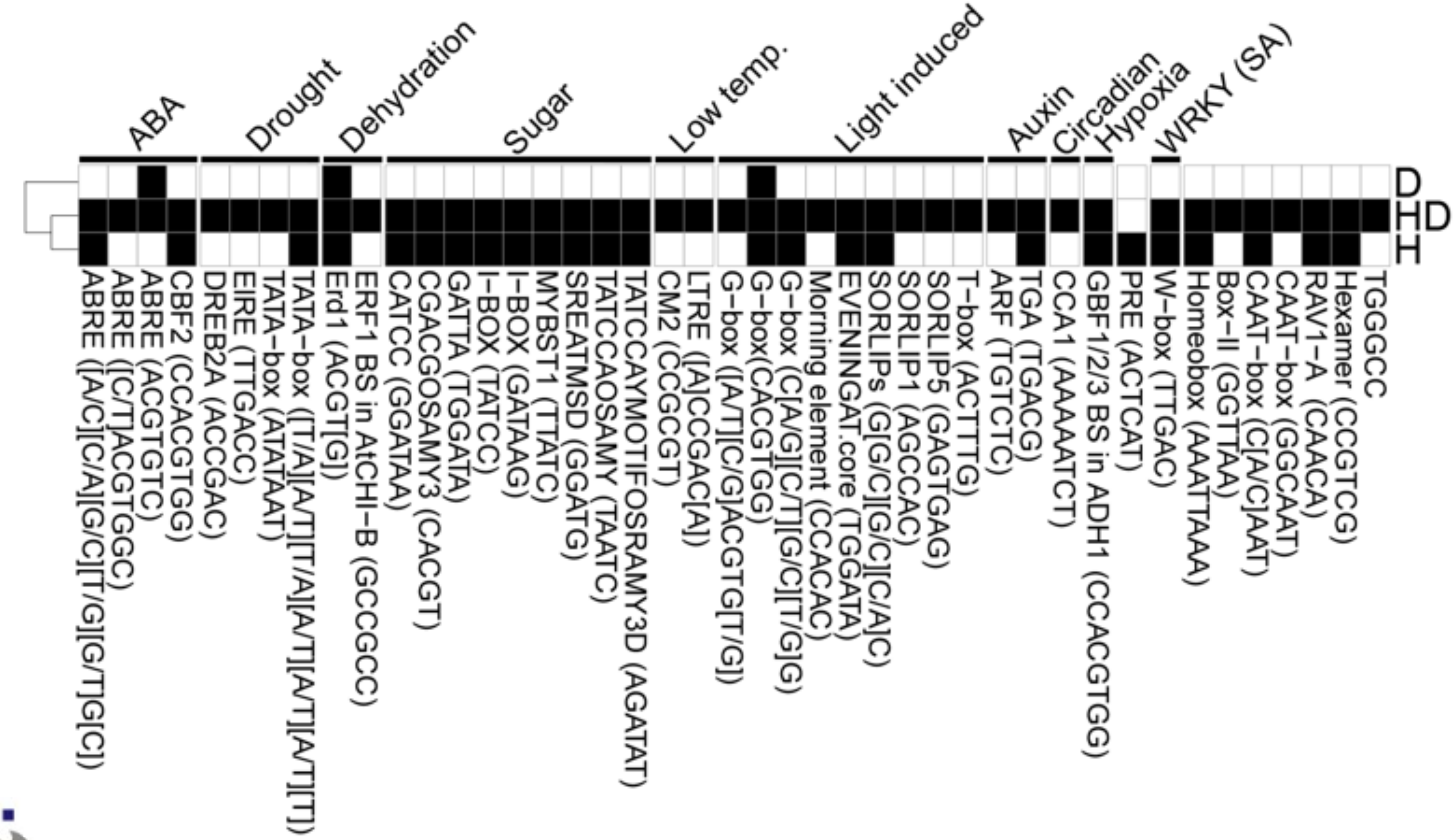


Safronov et al. PLoS ONE, 2017

Jarkko Salojärvi

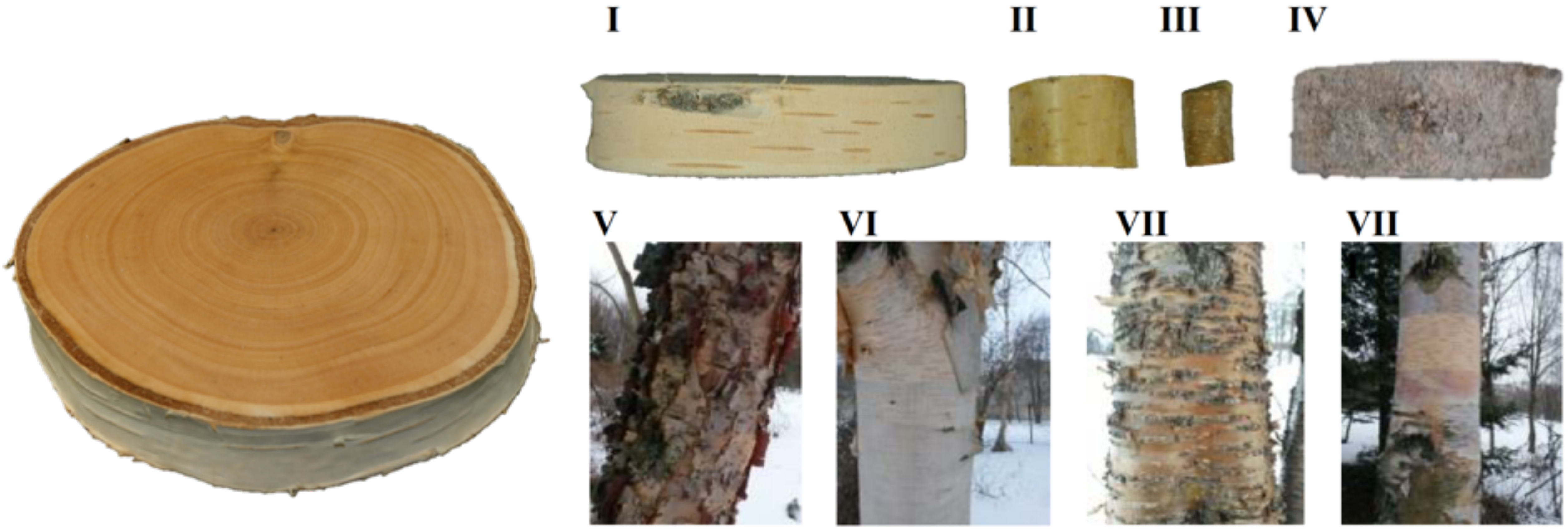
Circadian&light motifs and sugar signaling motifs enriched in promoters of diff exp genes

- Perhaps new adaptation?



Safronov et al. PLoS ONE, 2017

Evolution of environmental defenses



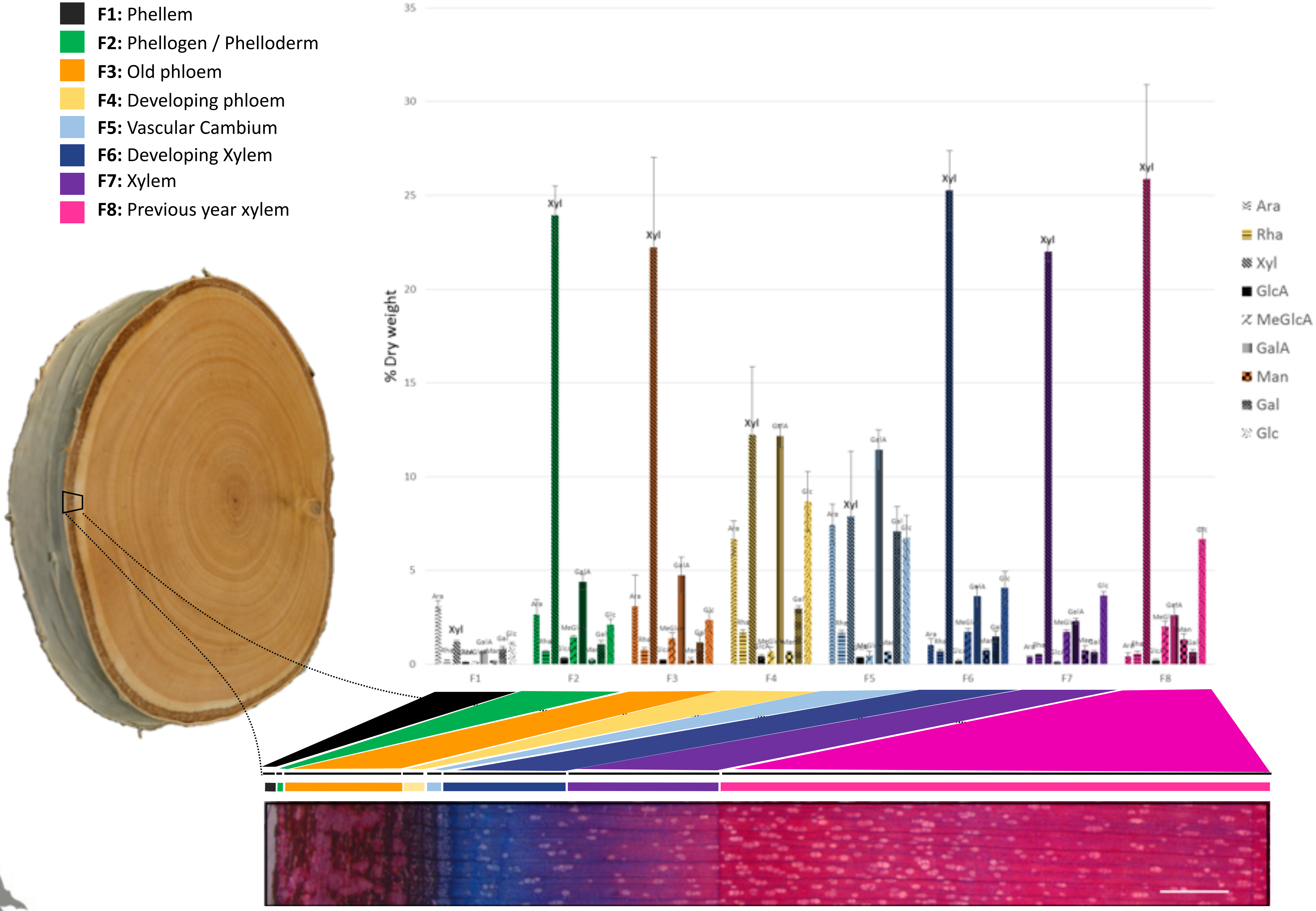
Tissue-specific study across the stem reveals the transcriptome dynamics of bark.

Juan Alonso-Serra*[1,2,3], Omid Safronov*[1,2], Kean-Jin Lim*[2,4,5], Sara J. Fraser-Miller**[6,7], Olga B. Blokhina[1,2], Ana Campilho[8], Sun-Li Chong[9], Kurt Fagerstedt[1,2], Raisa Haavikko[6], Ykä Helariutta[1,2,3,10], Juha Immanen[1,2,3], Jaakko Kangasjärvi[1,2], Tiina J Kauppila[6], Mari Lehtonen[11], Laura Ragni[12], Sitaram Rajaraman[1,2], Pia Runeberg-Roos[3], Riikka-Marjaana Räsänen[6], Pezhman Safdari[1,2], Maija Tenkanen[9], Jari T Yli-Kauhaluoma[6], Teemu H. Teeri***[2,4], Clare Strachan***[6], Kaisa Nieminen***[13], Jarkko Salojärvi***[14,1,2,15].

New Phytologist

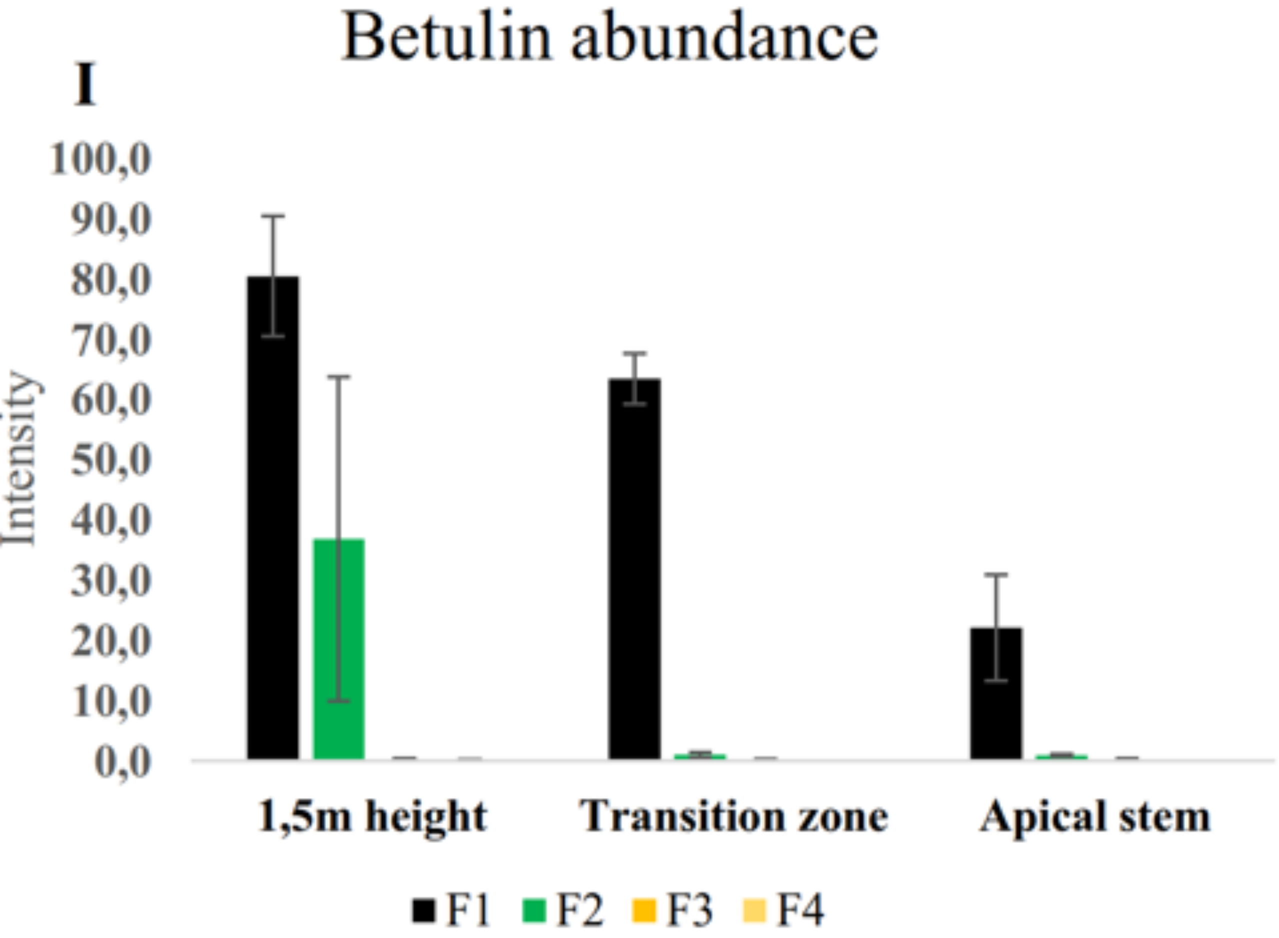
Cryosectioning of birch bark tissues

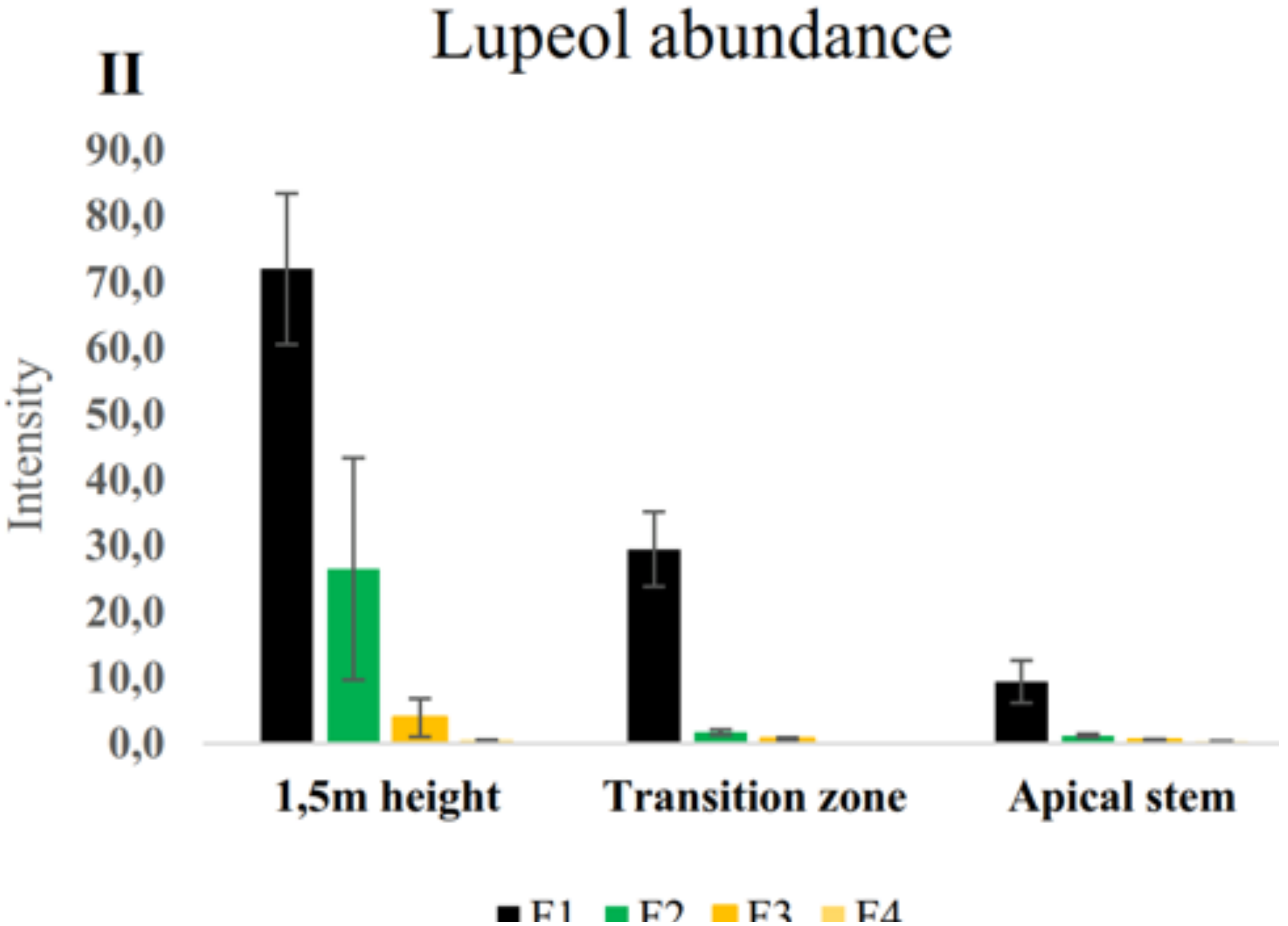
- ■ **F1:** Phellem
- 🟩 **F2:** Phellogen / Phelloderm
- 🟧 **F3:** Old phloem
- 🟨 **F4:** Developing phloem
- 🟦 **F5:** Vascular Cambium
- 🟦 **F6:** Developing Xylem
- 🟪 **F7:** Xylem
- 🟥 **F8:** Previous year xylem



Serra, Safronov, Lim et al. accepted

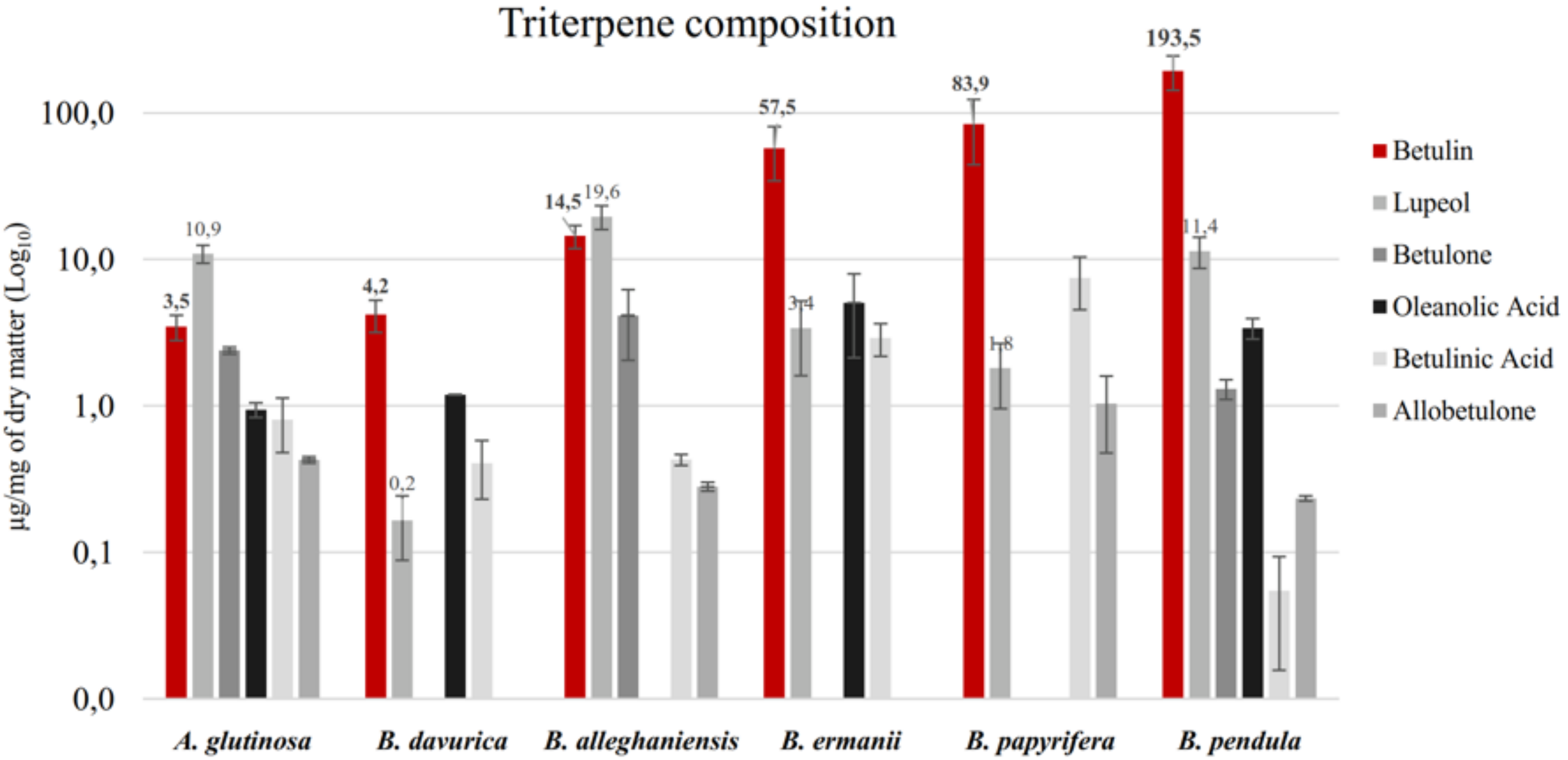
Birch phellem contains high amounts of betulin and its precursor lupeol





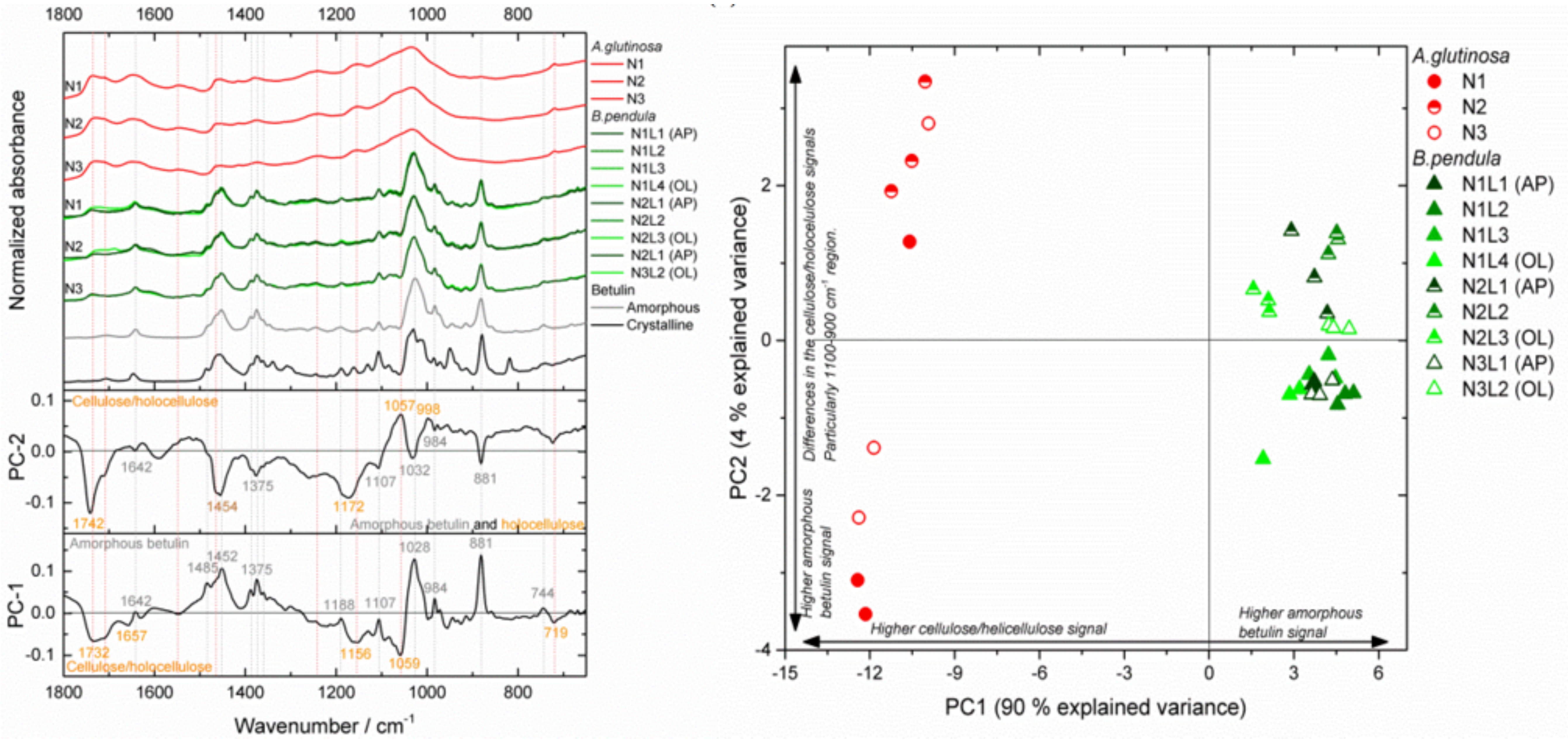
Serra, Safronov, Lim et al. accepted

Highest betulin content in *B. pendula*



Serra, Safronov, Lim et al. accepted

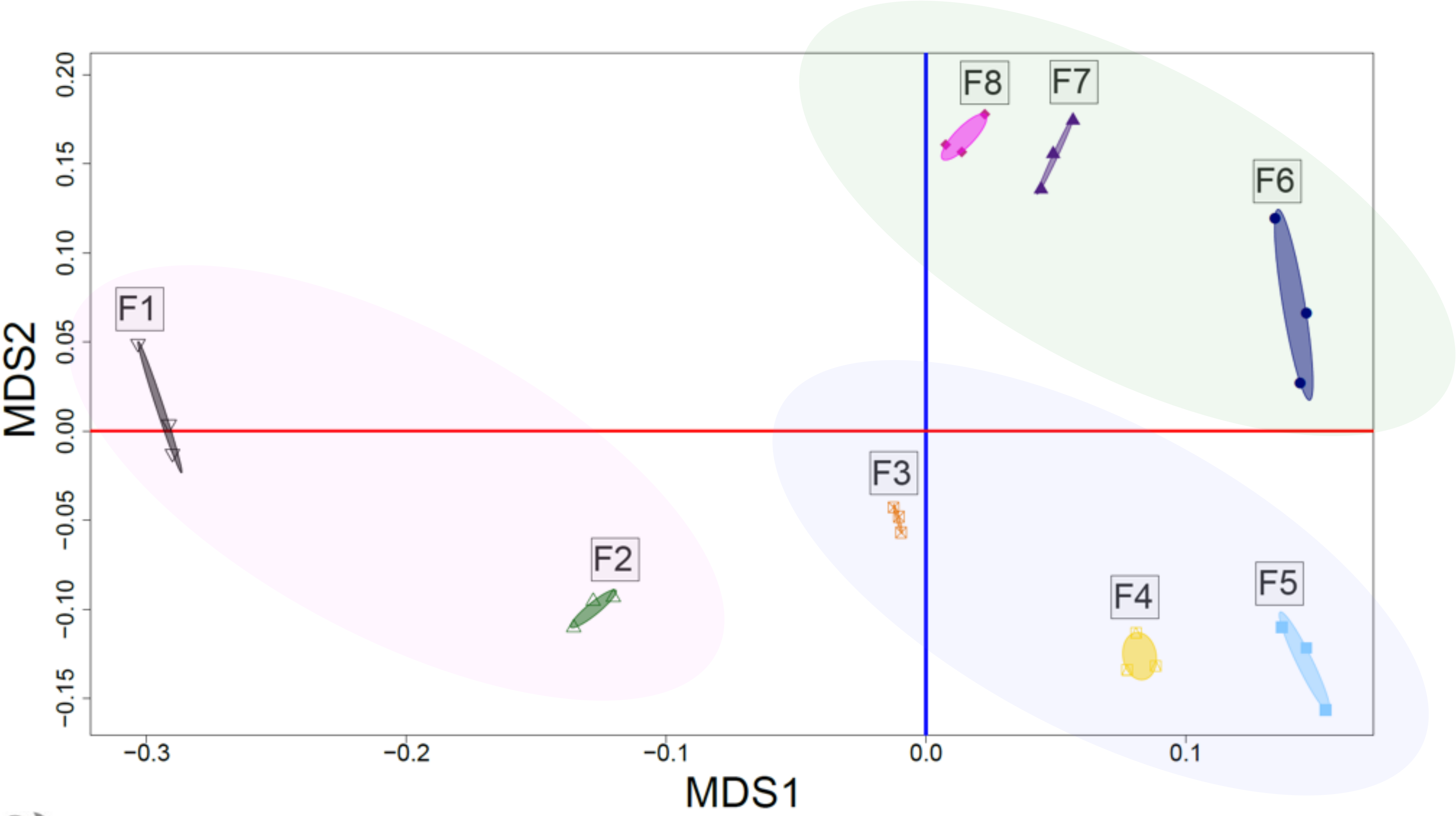
Betulin is the main distictive feature between *B. pendula* and *A. glutinosa* phellem



Jarkko Salojärvi

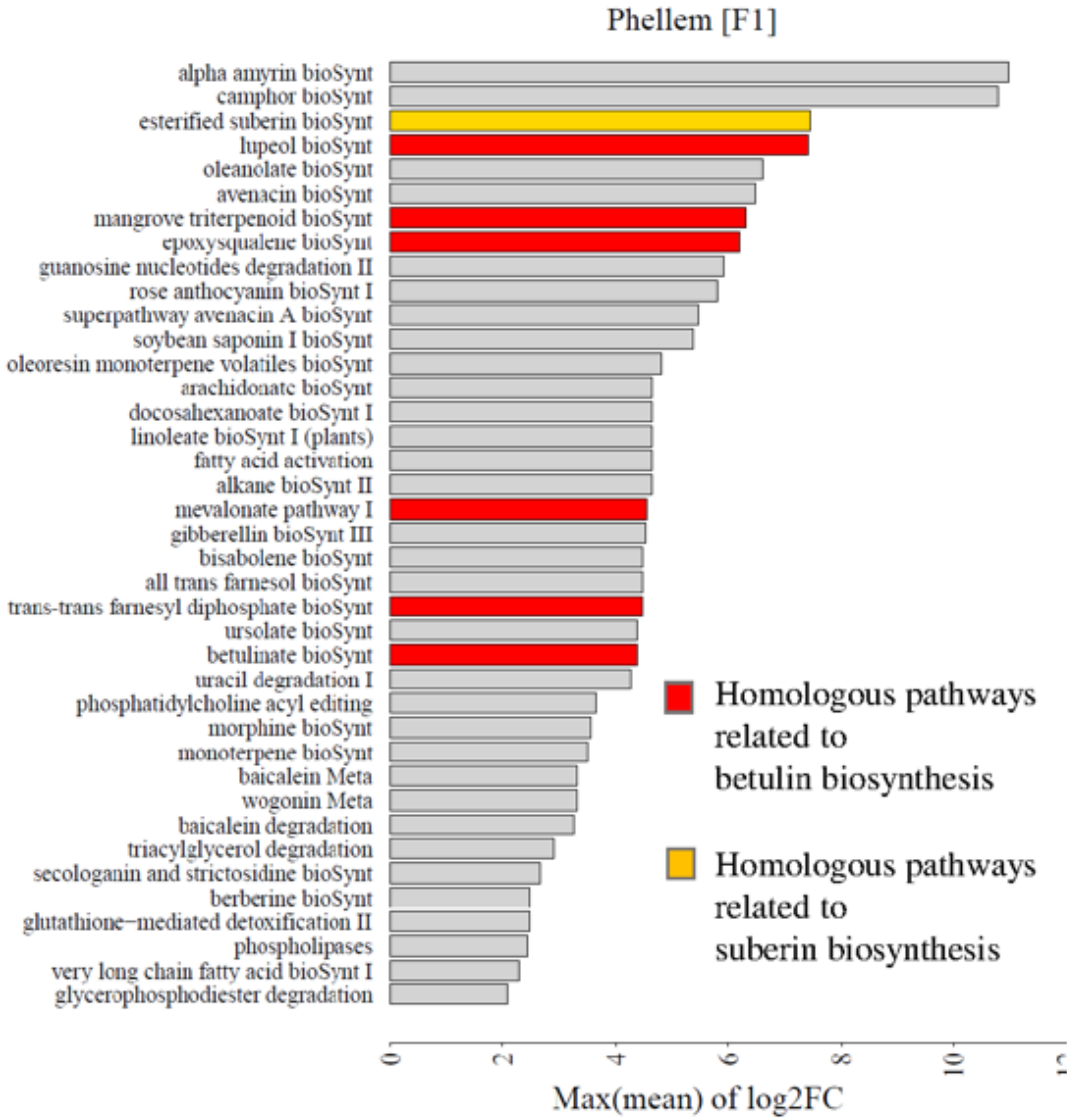
Serra, Safronov, Lim et al. accepted

Bark tissue transcriptomes are stratified by the tissue from which they originate



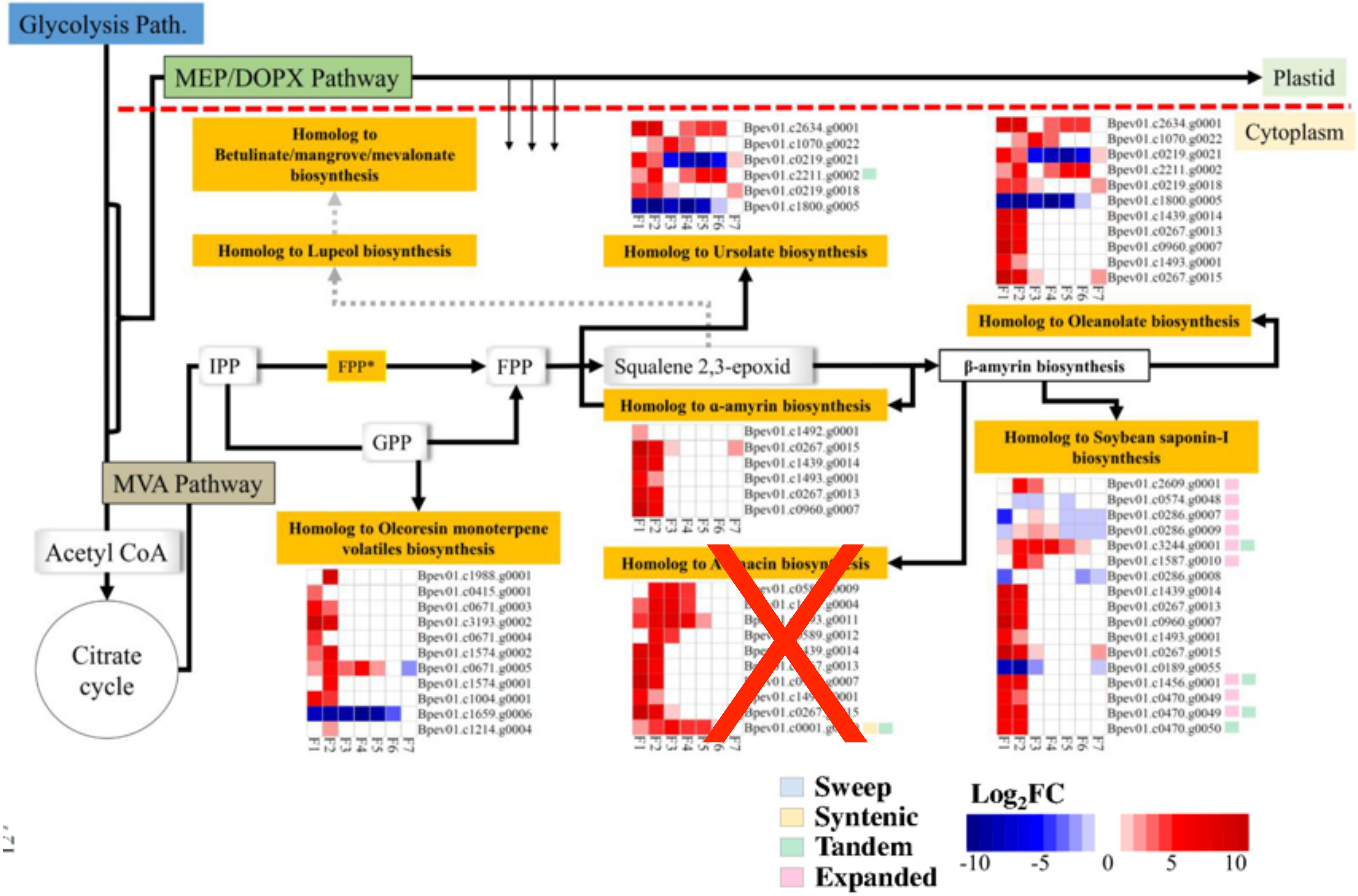
Serra, Safronov, Lim et al. accepted

Secondary metabolism pathways enriched in phellem (F1)



Serra, Safronov, Lim et al. accepted

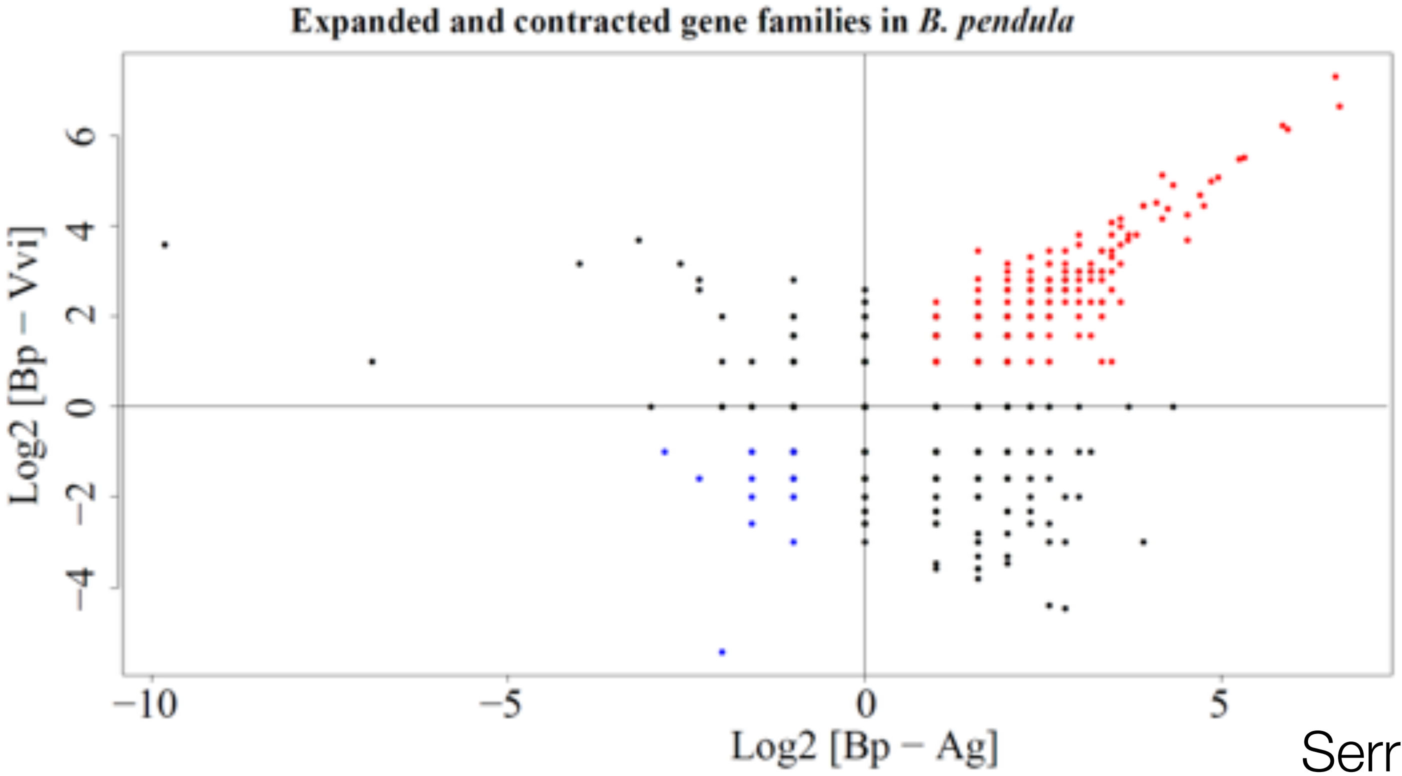
…especially triterpenoids



Serra, Safronov, Lim et al. accepted

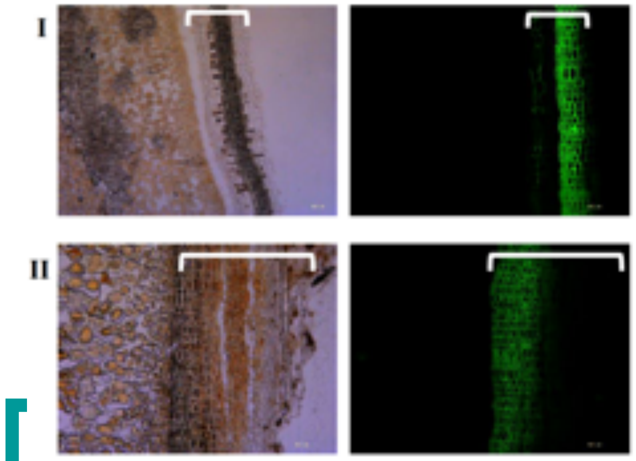
Comparative transcriptomics of B. pendula vs. A.glutinosa

- To compare the transcriptomes, we extracted **phellogen (F2)** and **old phloem (F3)** from black alder (*A.glutinosa*) and carried out RNAseq.
- RNAseq mapped against a draft assembly of *A.glutinosa.*
- Birch-specific gene family expansions detected by running orthofinder using *B.pendula*, *A.glutinosa*, *V. vinifera*, *A. thaliana* and *P.trichocarpa* gene models.
    - Birch-specific expansions were identified from among orthogroups.

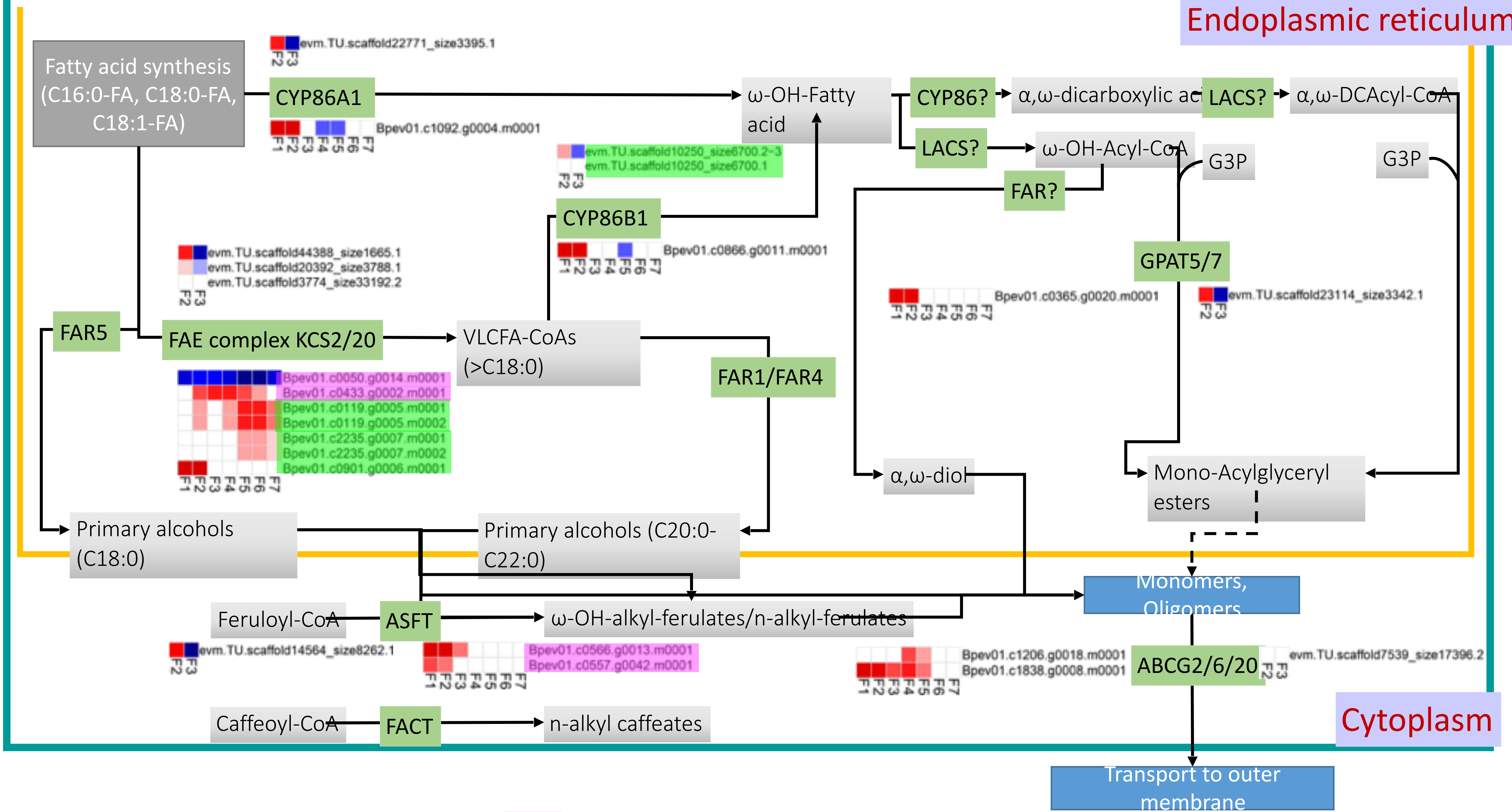


Jarkko Salojärvi

Serra, Safronov, Lim et al. accepted



Suberin biosynthesis

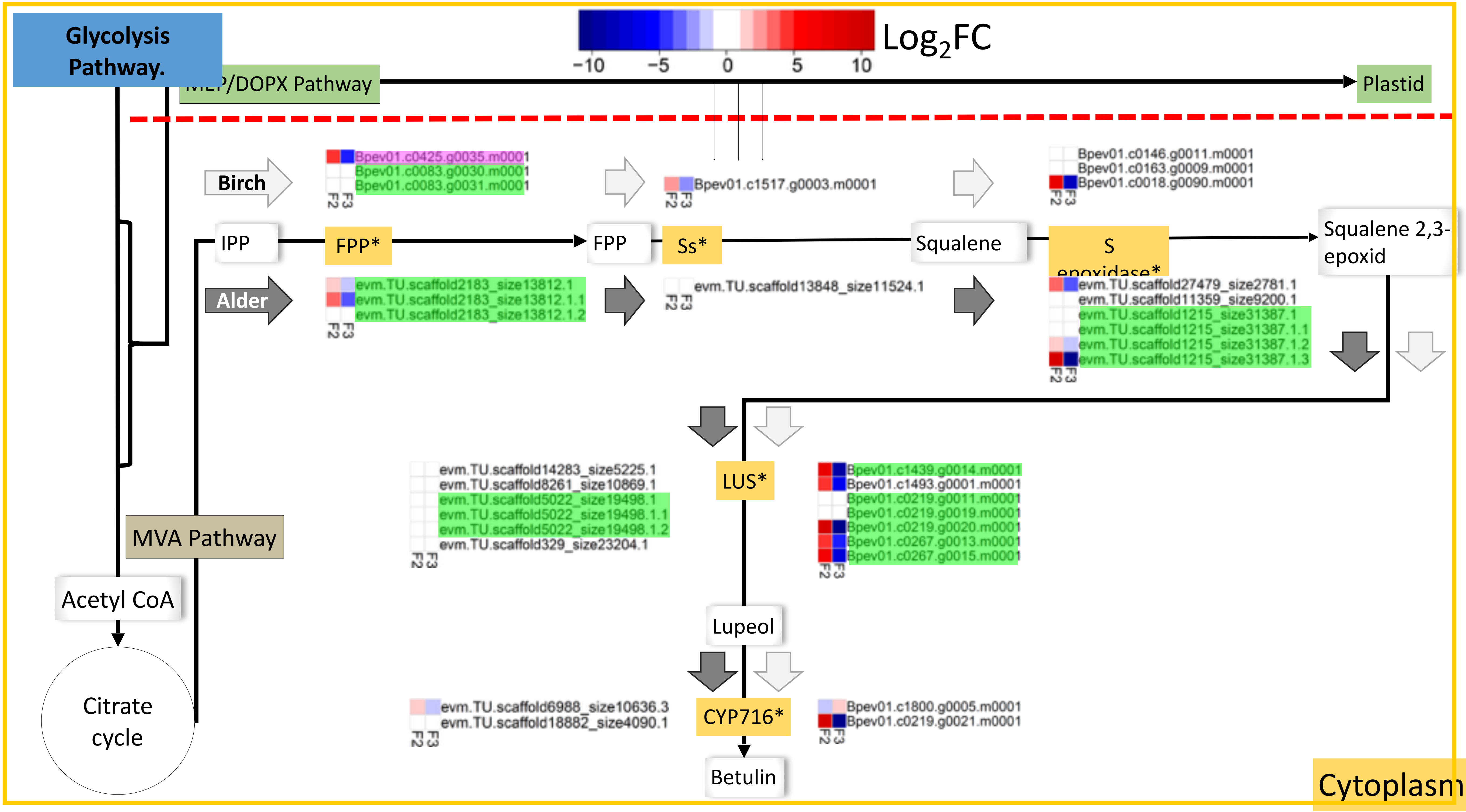


🟩 TANDEM

🟪 BIRCH-SPECIFIC EXPANSION

Serra, Safronov, Lim et al. accepted

Betulin biosynthesis

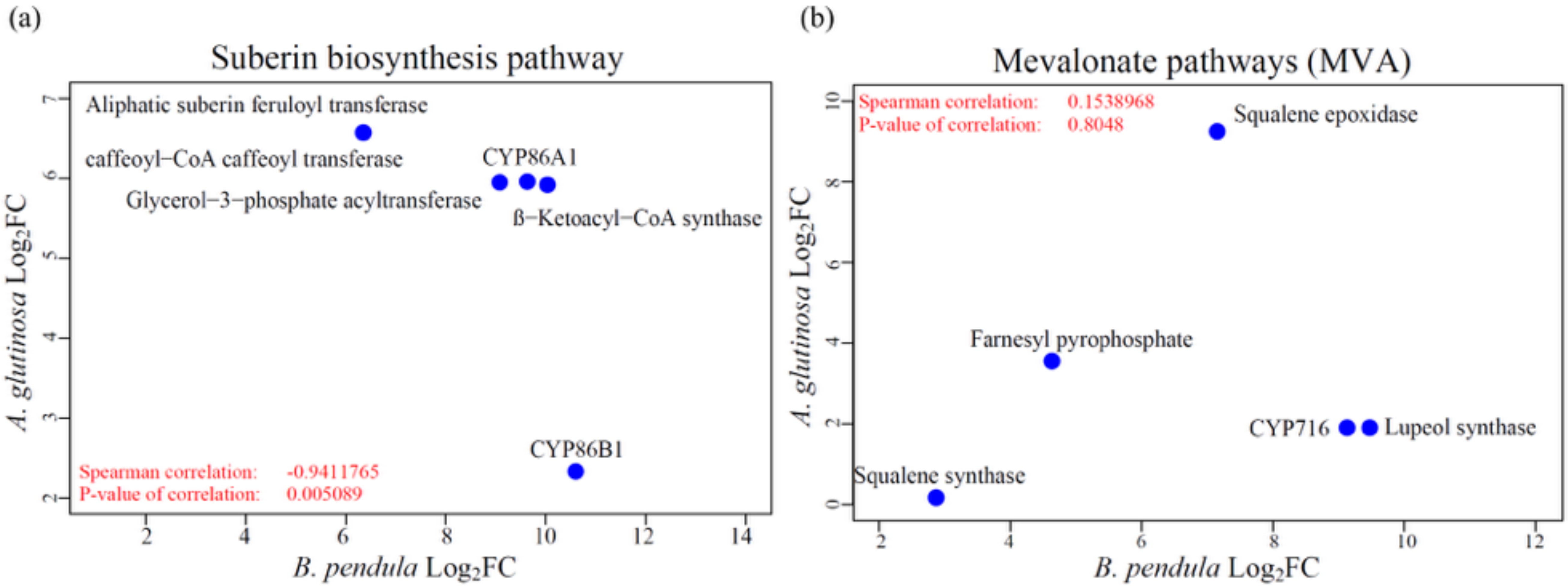


🟩 TANDEM

🟪 BIRCH-SPECIFIC EXPANSION

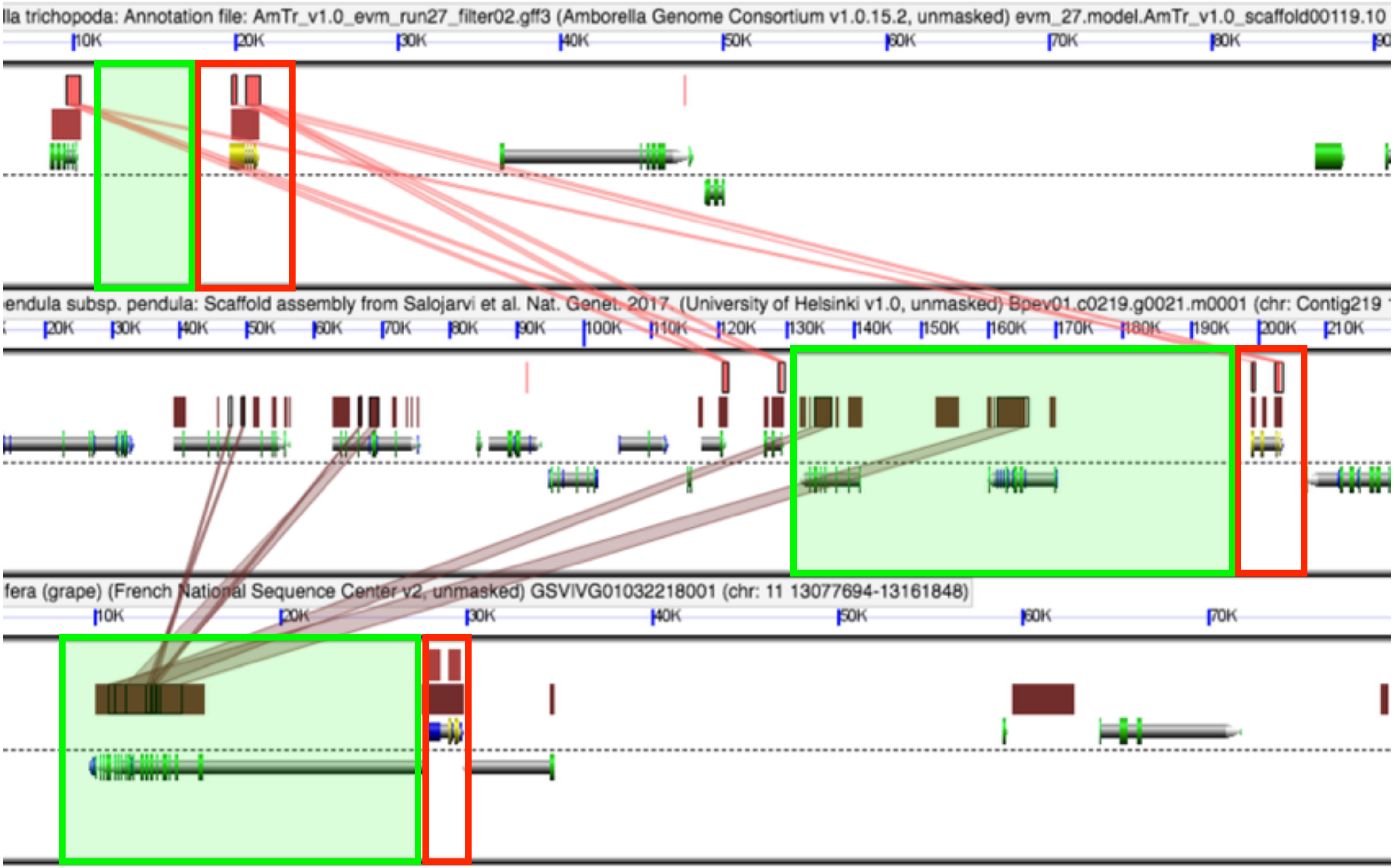
Serra, Safronov, Lim et al. accepted

Betulin biosynthesis is less active in alder



Serra, Safronov, Lim et al. accepted

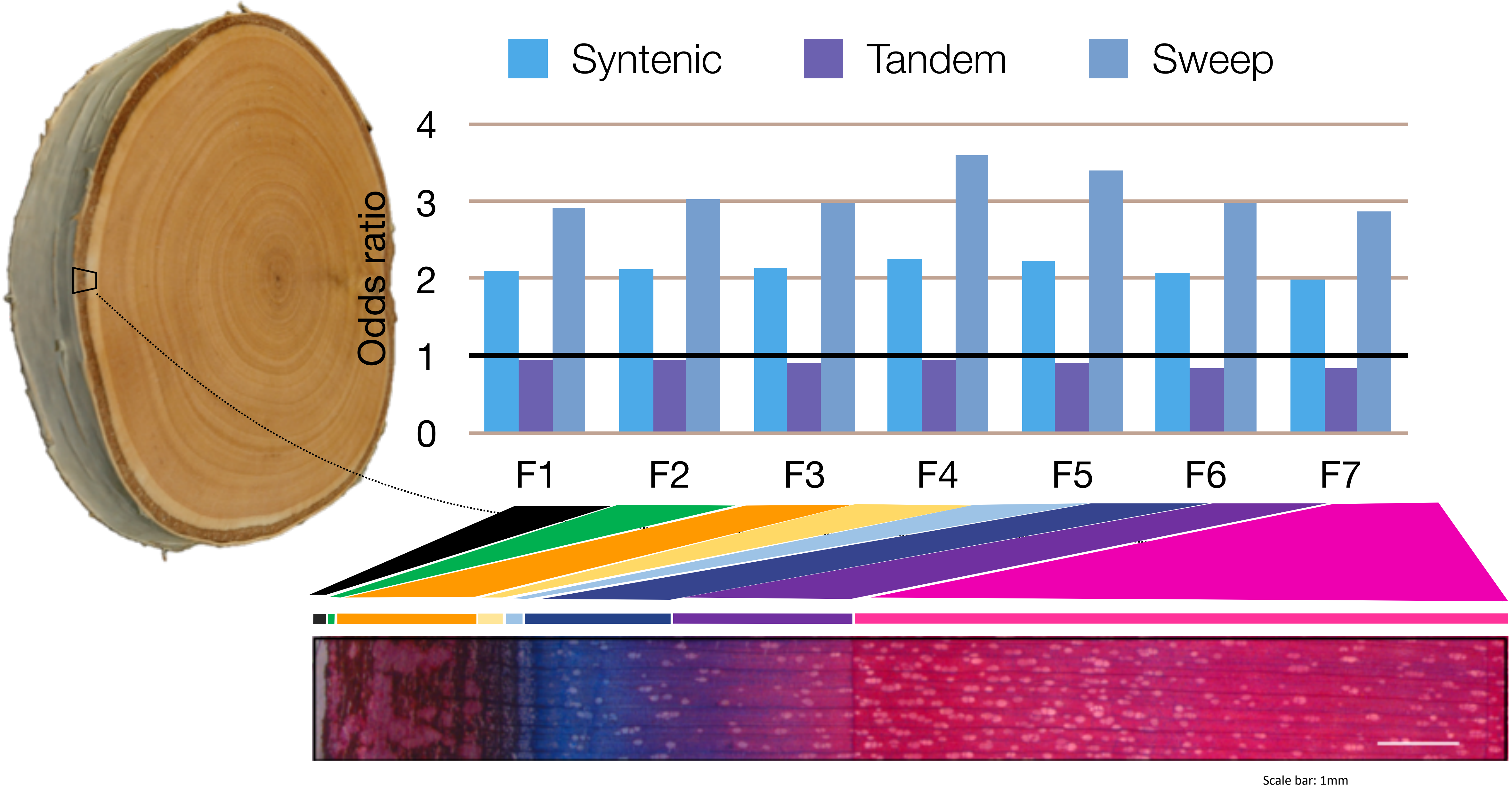
Birch-specific genome structure: Tandem LUS expansion next to CYP716



Serra, Safronov, Lim et al. accepted

- ⬛ **F1:** Phellem
- 🟩 **F2:** Phellogen / Phelloderm
- 🟧 **F3:** Old phloem
- 🟨 **F4:** Developing phloem
- 🟦 **F5:** Vascular Cambium
- 🟦 **F6:** Developing Xylem
- 🟪 **F7:** Xylem
- 🟥 **F8:** Previous year xylem

Syntenic genes and genes under selective sweeps are overrepresented in expressed genes



Serra, Safronov, Lim et al. accepted

Conclusions

- Climate change will dramatically alter the living conditions of the earth within the next decades.
- Competition in ecological niches will change, other species will get an advantage.
- Genomes contain information about the past selection and adaptation.
    - Can be used to derive hypotheses about gene family evolution and selection.

Bioinformatics for Molecular Biology & Genomics

UNIVERSITY OF HELSINKI

Maja Ilievska, Cui Wang (postdoc)

NANYANG TECHNOLOGICAL UNIVERSITY

Postdoc needed..

? ?

Nicholas Cho

Centre of Excellence of Academy of Finland
Molecular Biology of Primary Producers

ViPS

Acknowledgments

University of Helsinki

- **Jaakko Kangasjärvi**
- **Ykä Helariutta**
- **Petri Auvinen**
- Airi Lamminmäki
- Olli-Pekka Smolander
- Lars Paulin
- Pasi Rastas
- Juan Alonso Serra
- Juha Immanen
- **Teemu Teeri**
- Kean-jin Lim
- **Clare Strachan**
- Jari Yli-Kauhaluoma
- Maija Tenkanen

University at Buffalo

- **Victor Albert**
- Tianying Lan

Bioinformatics for Molecular Biology and Genomics

- **Sitaram Rajaraman**
- **Omid Safronov**
- **Pezhman Safdari**
- **Aleksia Vaattovaara**
- **Maja Ilievska**
- **Meharji Arumilli**

- **Uppsala universitet**
  - **Martin Lascoux**
  - Luis Leal

Natural Resources Institute Finland

- Kaisa Nieminen
- Risto Hagqvist
- Matti Rousi

UNIVERSITY OF HELSINKI

Tekes

ACADEMY OF FINLAND