Reference-free comparative genomics of tropical Fagaceae using shallow *Illumina* sequencing

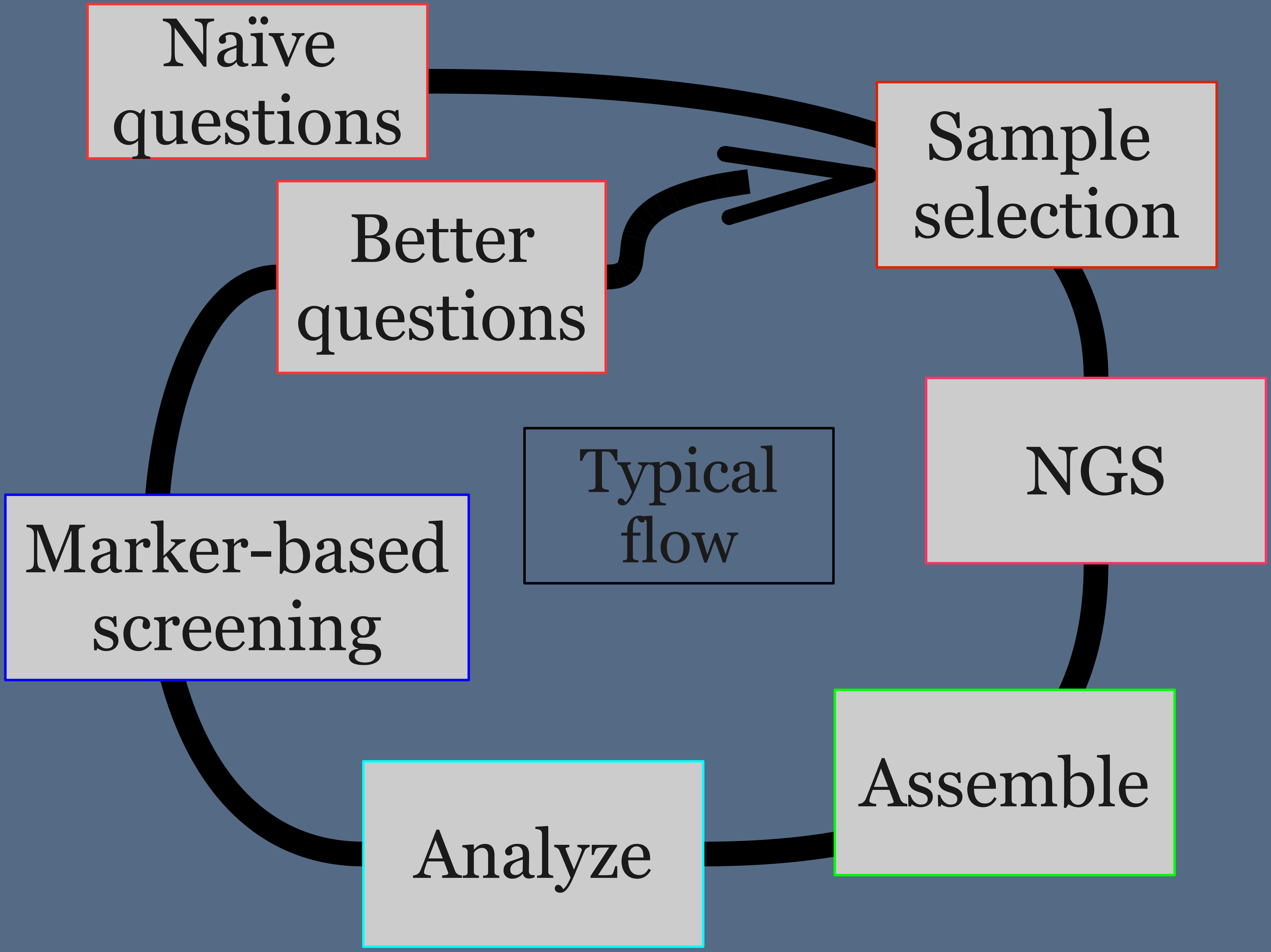
Chuck Cannon
Associate Professor, TTU
Professor, XTBG/CAS

~ E-SQUARED ~ XTBG-CAS ~ ECOLOGICAL EVOLUTION ~

XTBG

reference-free (?) comparative genomics
174 chloroplasts – proof of concept

tropical Fagaceae?



Naïve questions

Sample selection

NGS

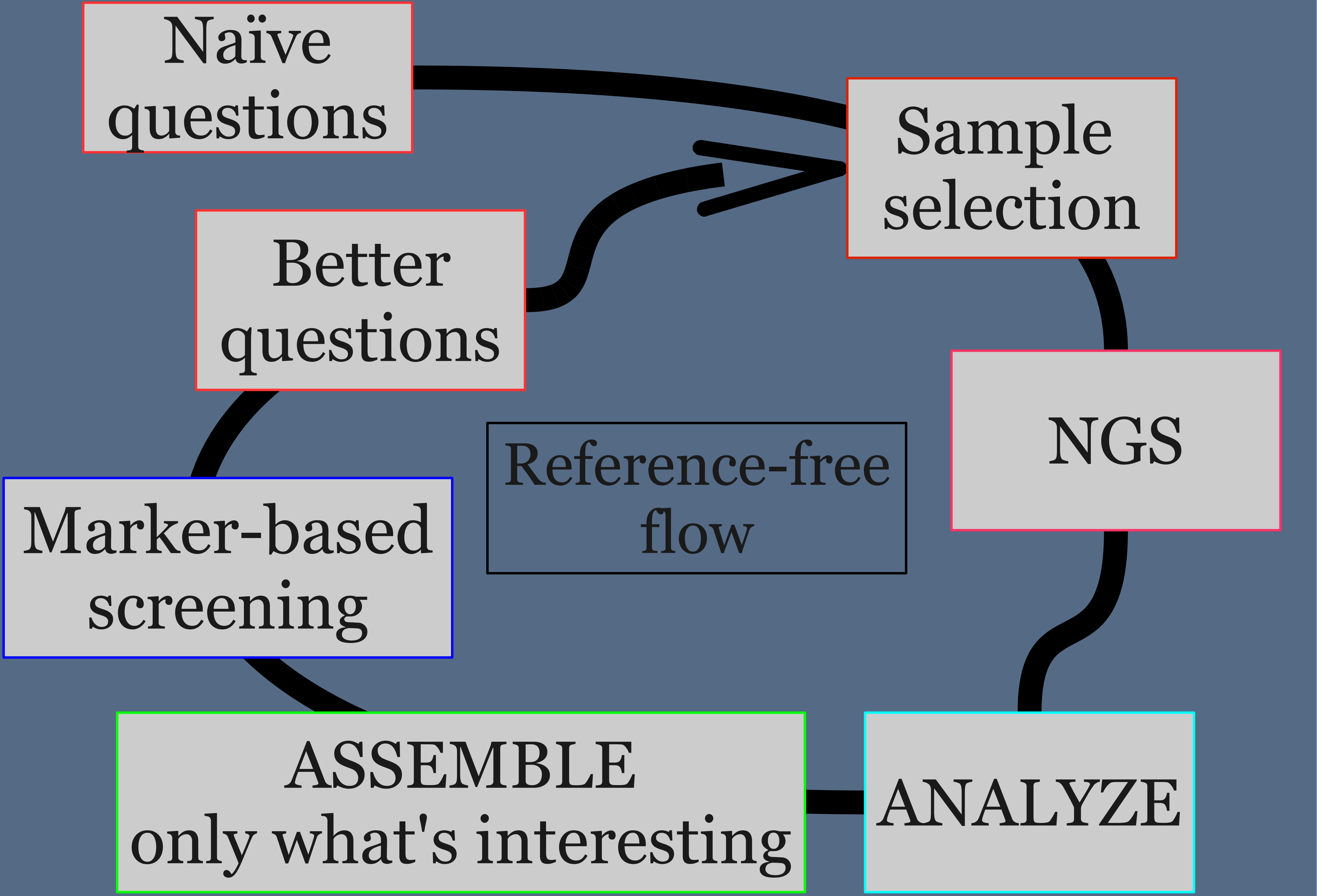
Assemble

Analyze

Marker-based screening

Better questions

Typical flow



Naïve questions

Sample selection

Better questions

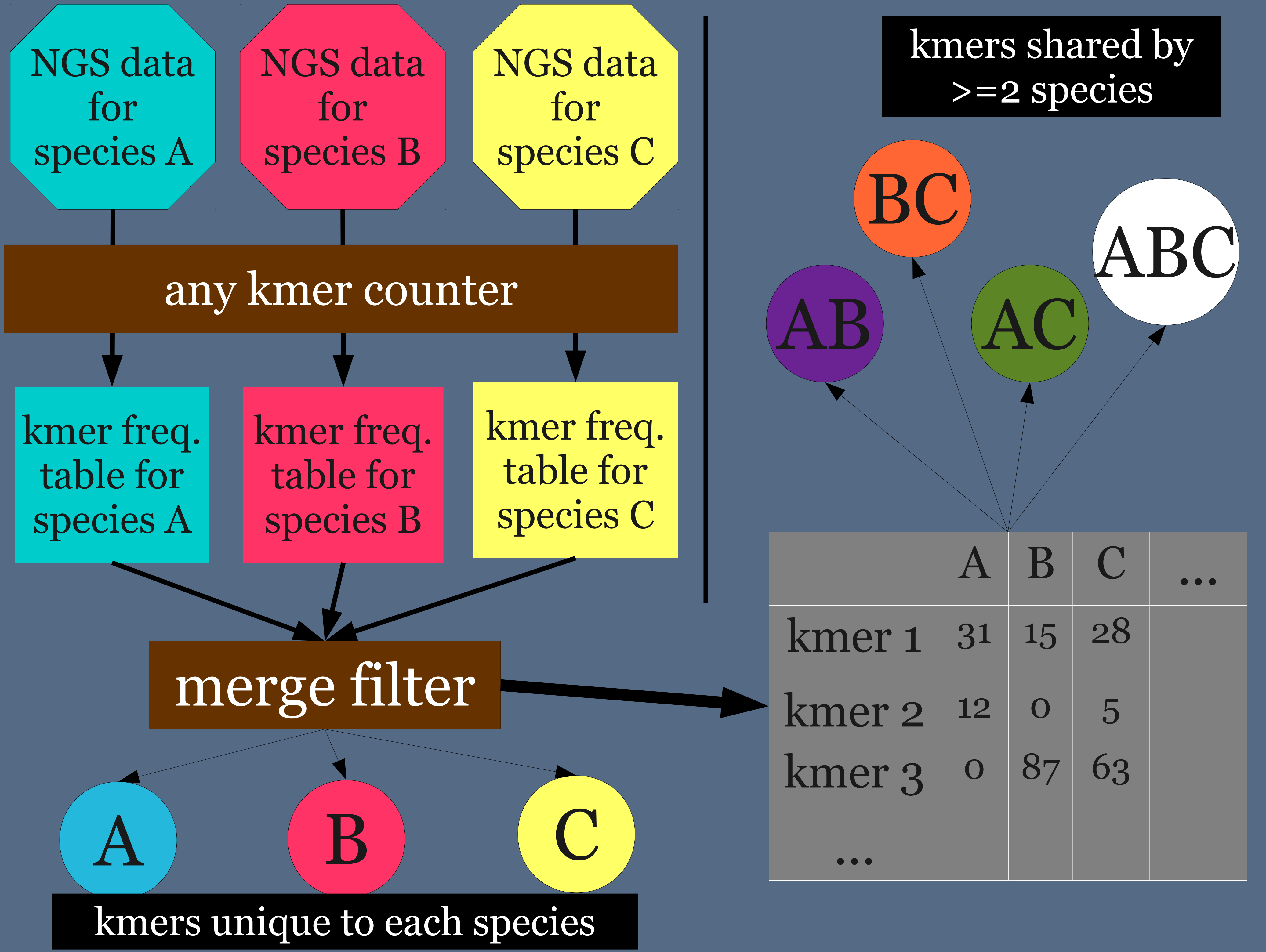
NGS

Reference-free flow

Marker-based screening

ASSEMBLE only what's interesting

ANALYZE



NGS data for species A

NGS data for species B

NGS data for species C

any kmer counter

kmer freq. table for species A

kmer freq. table for species B

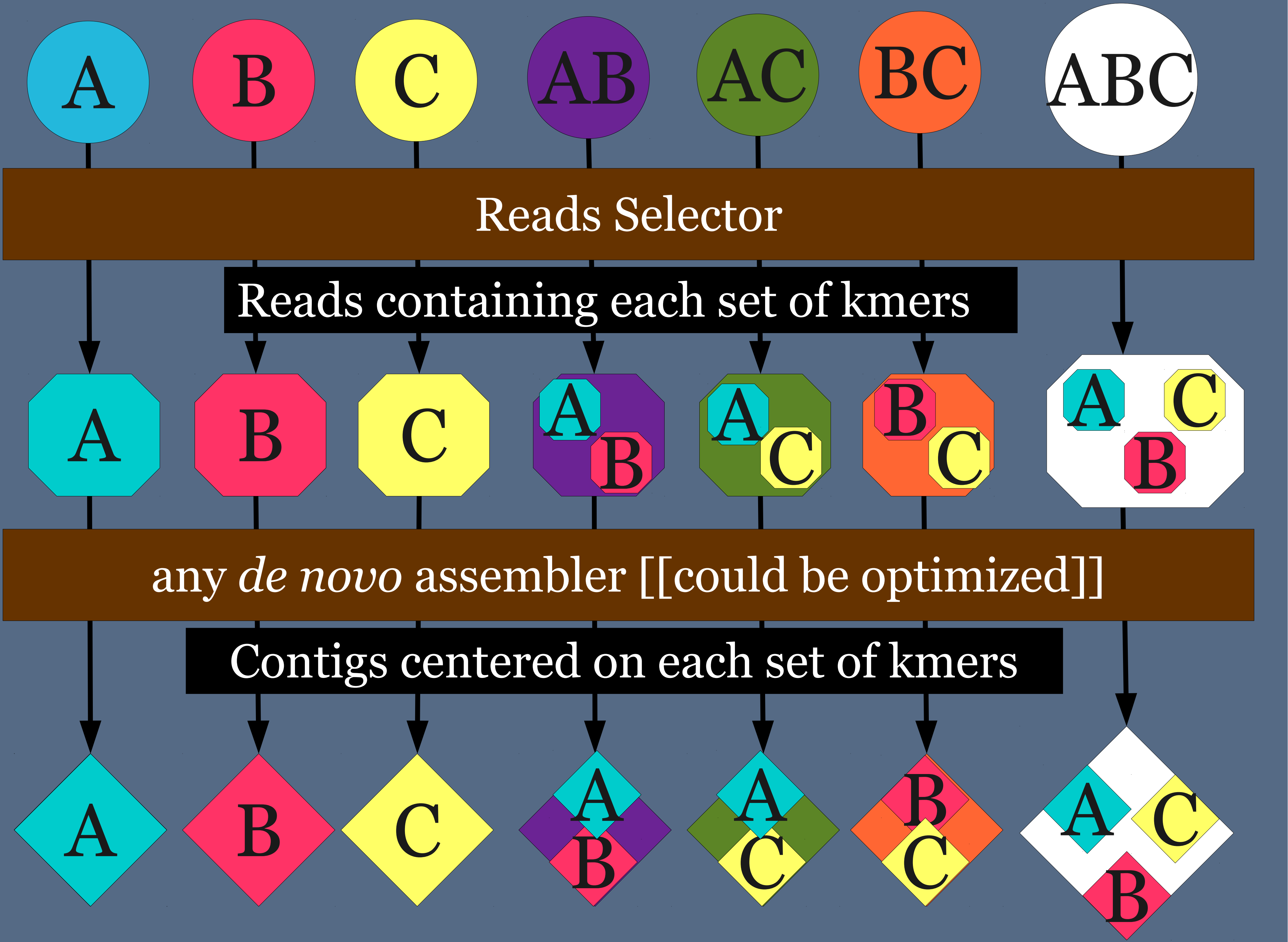
kmer freq. table for species C

merge filter

kmers unique to each species

kmers shared by >=2 species

	A	B	C	...
kmer 1	31	15	28	
kmer 2	12	0	5	
kmer 3	0	87	63	
...				



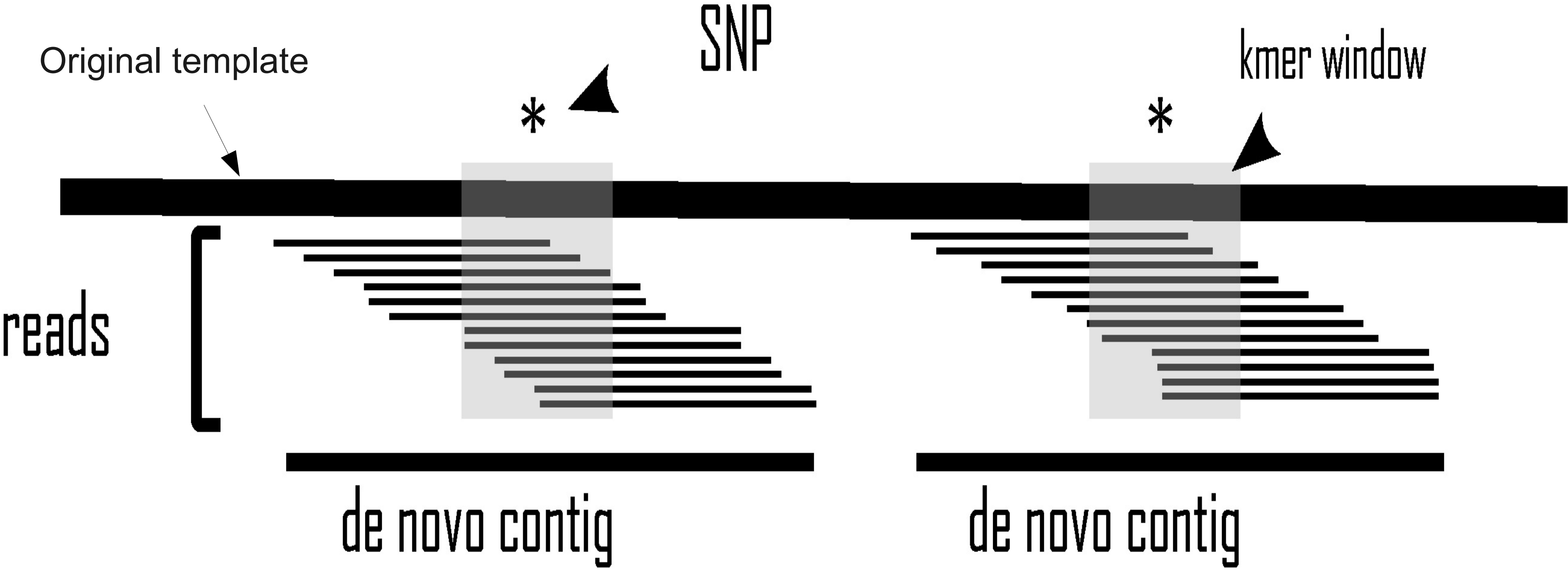
Reads Selector

Reads containing each set of kmers

any *de novo* assembler [[could be optimized]]

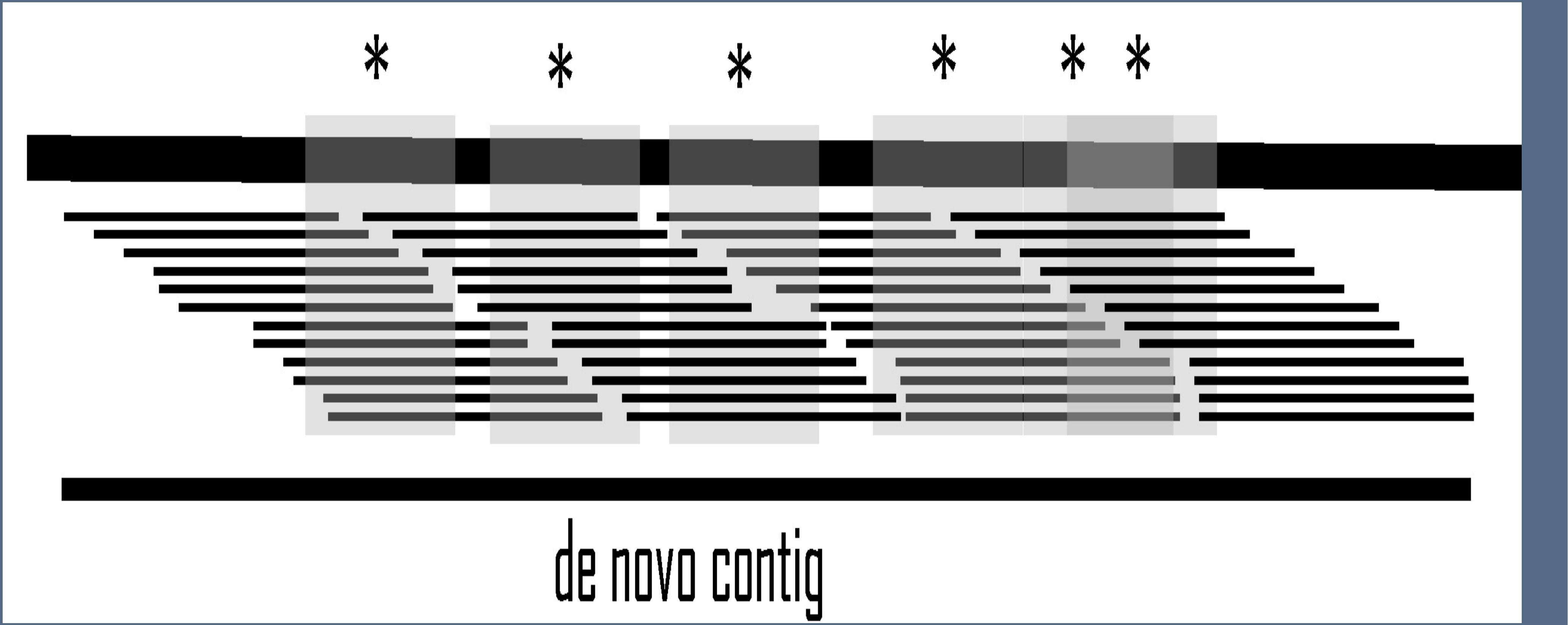
Contigs centered on each set of kmers

Properties of local *de novo* assemblies correspond to the density of variants in relation to read length.



When the distance between variants exceeds read length, separate short contigs are constructed, each containing a SNP

Properties of local *de novo* assemblies correspond to the density of variants in relation to read length.

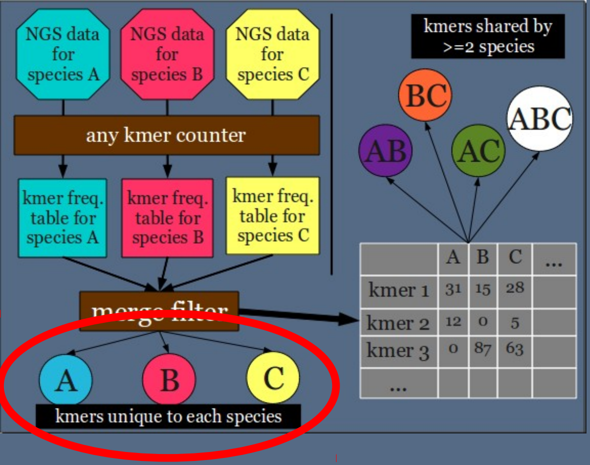


When the distance between variants is smaller than read length, then a single long contig is constructed, containing many variants.

These local de novo assemblies, centered on group-specific kmers, can discover:

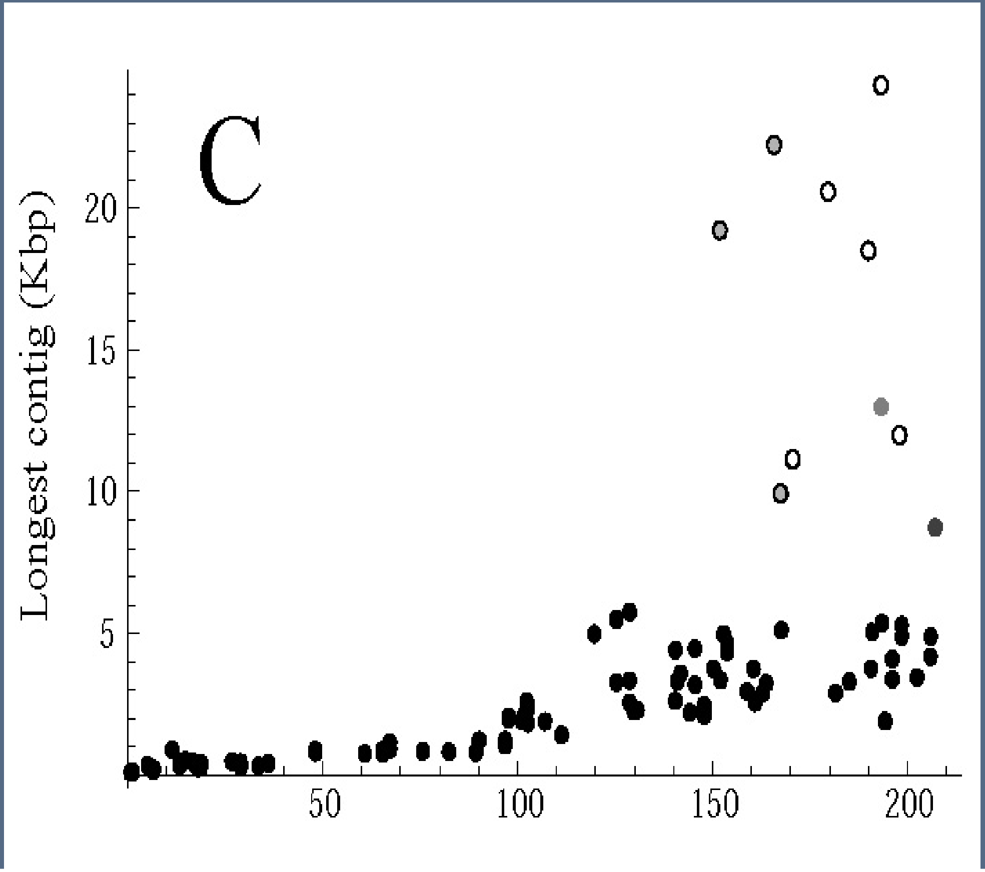
- regions with a high-density genetic change
- group-specific SNPs
- recalcitrant regions that do not assemble but have high levels of informative polymorphism

  [[and will need other approaches to characterize]]


NGS data for species A
NGS data for species B
NGS data for species C
any kmer counter
kmer freq. table for species A
kmer freq. table for species B
kmer freq. table for species C
merge filter
A
B
C
kmers unique to each species
kmers shared by >=2 species
BC
ABC
AB
AC

	A	B	C	...
kmer 1	31	15	28	
kmer 2	12	0	5	
kmer 3	0	87	63	
...				

Reference-free flowchart



Phylodistance to nearest relative in analysis

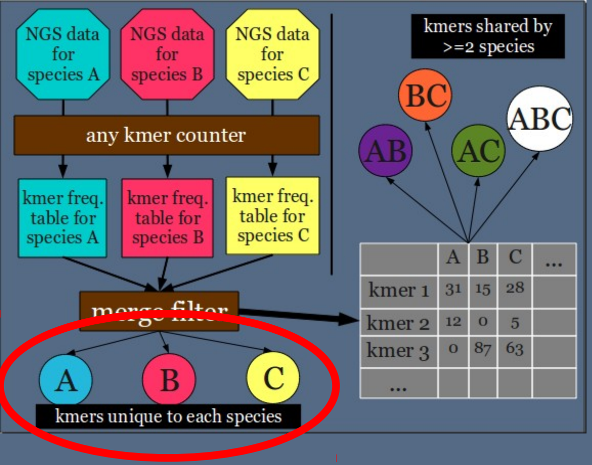
Looking at the 'private' kmers first

======

Length of local assemblies is directly related to the phylogenetic distance from the nearest relative in the analysis.

White and gray circles are lower plants and distance is under-estimated.

Very long contigs assemble for these largely unique genomes.


NGS data for species A
NGS data for species B
NGS data for species C
any kmer counter
kmer freq. table for species A
kmer freq. table for species B
kmer freq. table for species C
merge filter
A
B
C
kmers unique to each species
kmers shared by >=2 species
BC
ABC
AB
AC

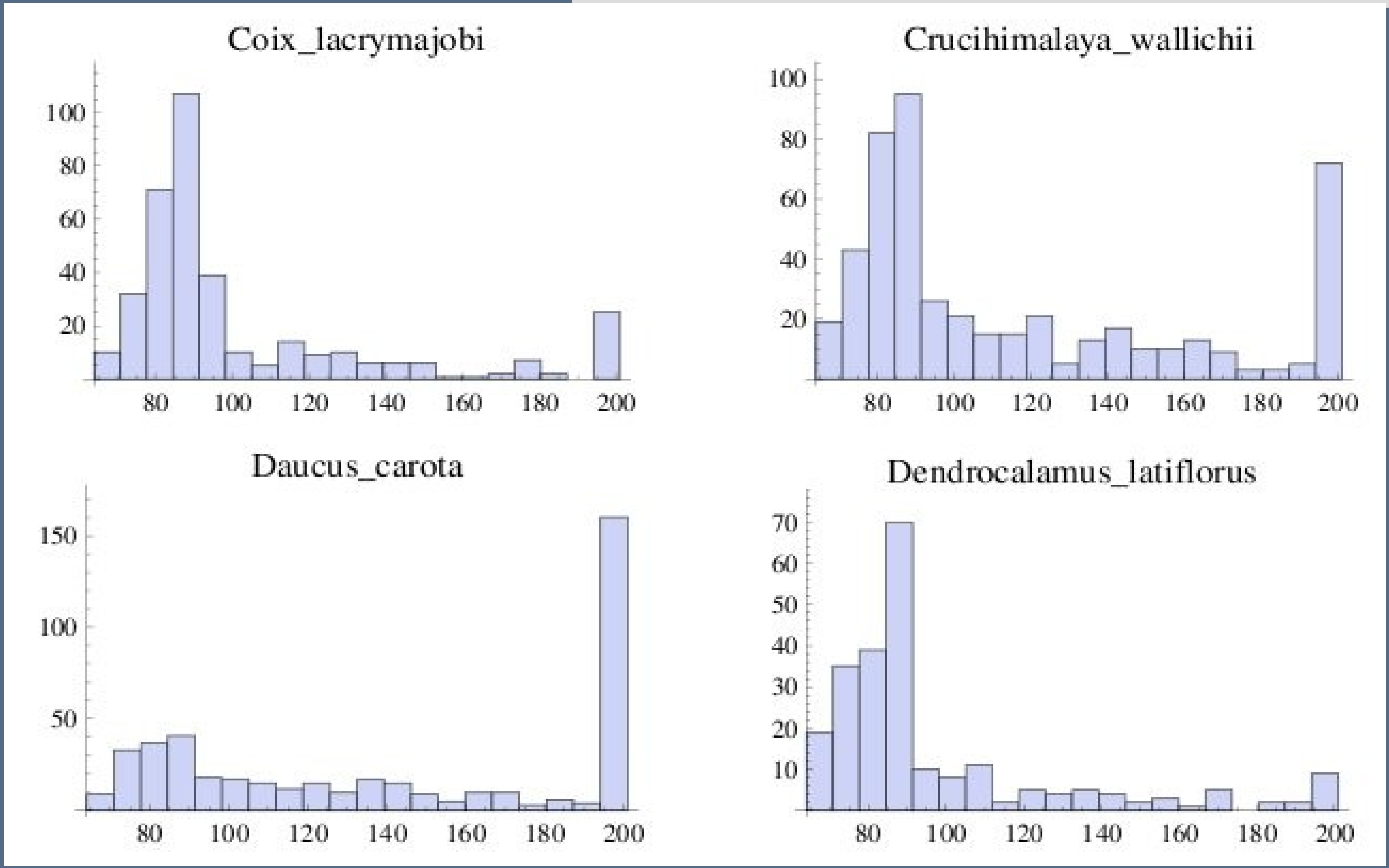
	A	B	C	...
kmer 1	31	15	28	
kmer 2	12	0	5	
kmer 3	0	87	63	
...				

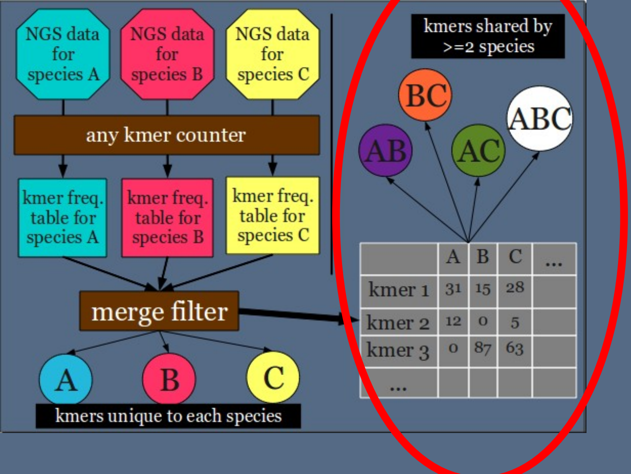
Reference-free flowchart

More on the 'private' kmers

======

Size distribution of local de novo contigs depends on relatedness of target genome to other genomes in analysis and the level of 'hotspot' diversity

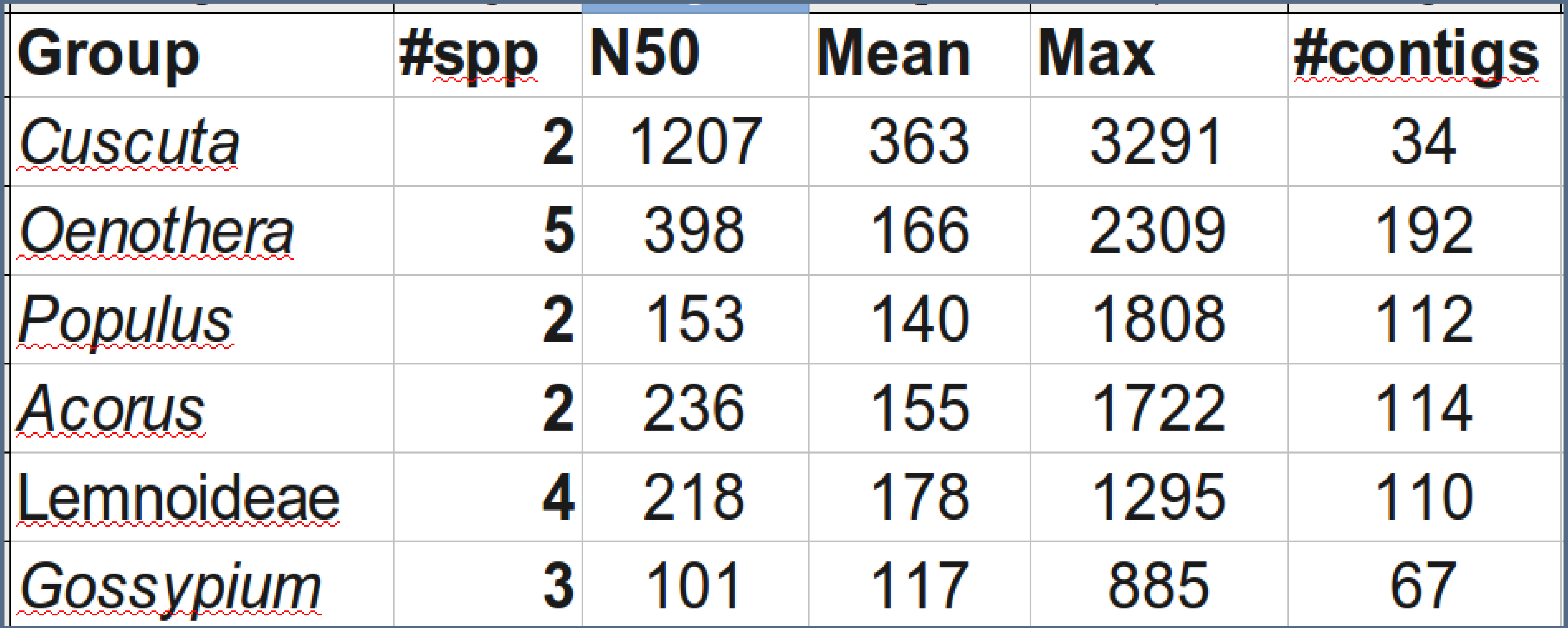




Reference-free flowchart

The "group contigs" assembled can be explored in many ways, including basic properties of N50 and length.

These are the most unusual sets of group contigs generated, given 4 basic properties.



Group	#spp	N50	Mean	Max	#contigs
Cuscuta	2	1207	363	3291	34
Oenothera	5	398	166	2309	192
Populus	2	153	140	1808	112
Acorus	2	236	155	1722	114
Lemnoideae	4	218	178	1295	110
Gossypium	3	101	117	885	67

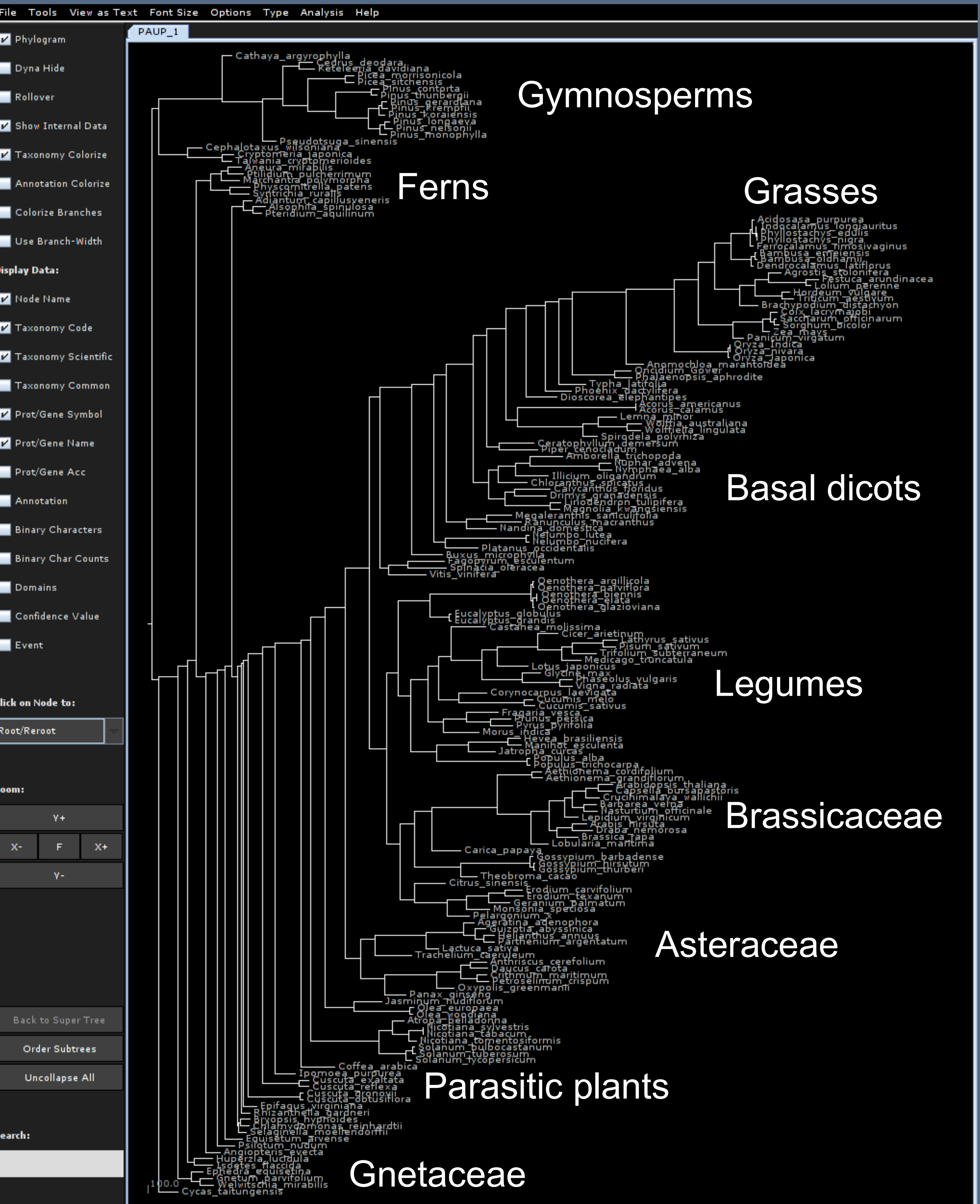
	A	B	C	...
kmer 1	31	15	28	
kmer 2	12	0	5	
kmer 3	0	87	63	
kmer 4	12	21	17	
kmer 5	0	45	23	
kmer 6	0	27	76	
...				

The shared kmer frequency table can also be converted into phylogenomic data. Several ways of doing this could be developed.

Straight binary presence-absence is the simplest way.

Transpose and convert to binary states.

	k1	k2	k3	k4	k5	k6	...
A	1	1	0	1	0	0	
B	1	0	1	1	1	1	
C	1	1	1	1	1	1	
...							



Parsimony tree based on presence-absence data of 25bp kmers in 174 whole chloroplast genomes

----

235974 characters
10% subsample

This approach does not get the deep branches correct but at the ordinal level and above, the results are congruent with the APG tree and more detailed studies at the family level.

Reference-free comparative genomics

Long local contigs assemble when a genomic region has a high density of variants peculiar to a genome or set of genomes.

The results from our combined analysis of 174 chloroplast genomes discovered many of the same results found in many separate analyses.

We also discovered a number of novel features, both conserved and divergent, not previously found.

Strong phylogenetic signal in data, although the reconstruction model certainly needs to be improved.

Comparative genomics of 15 Fagaceae species

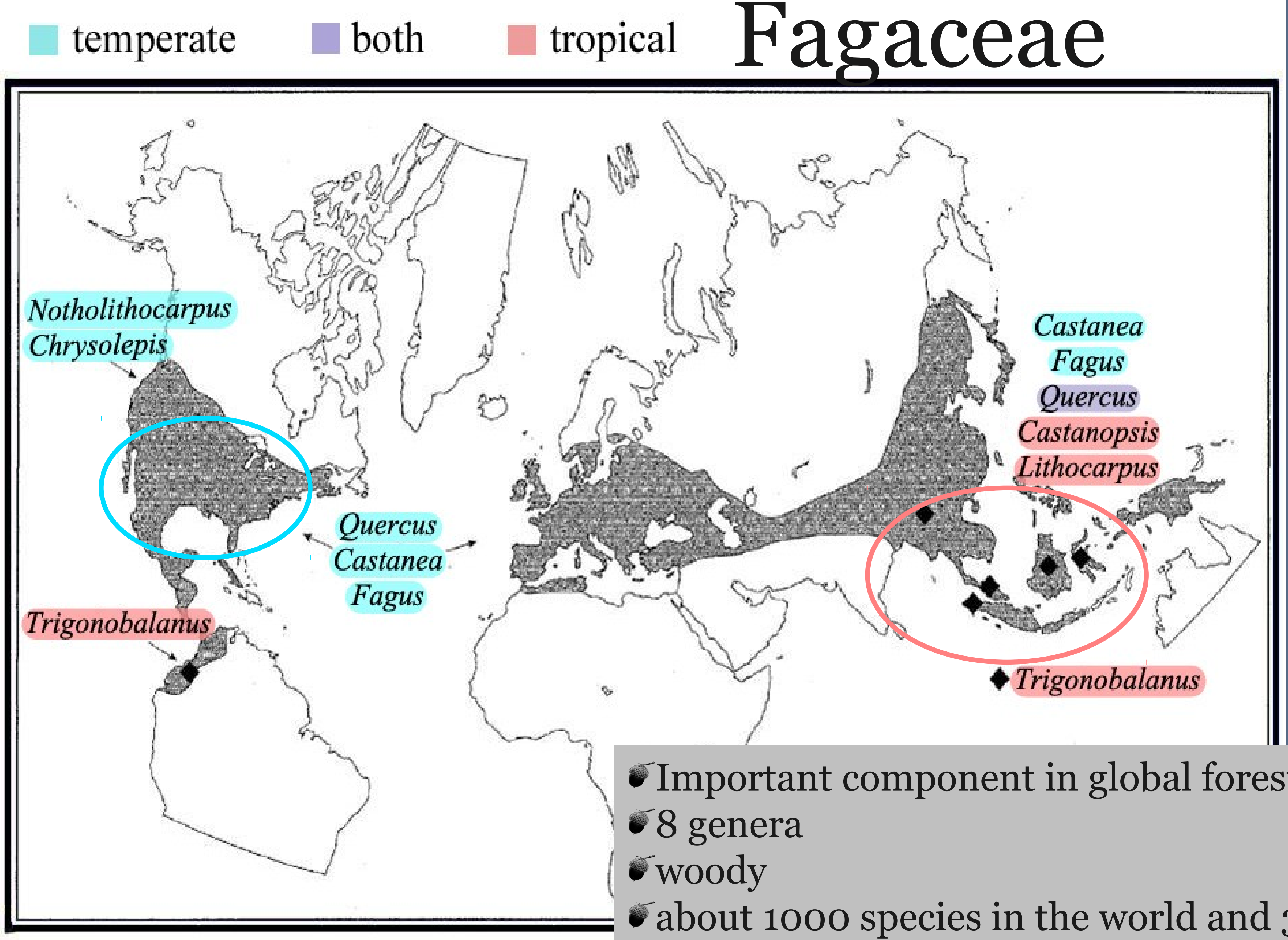
- Whole genome sequencing at low coverage.
  (~0.5x-10x coverage on Illumina)
- Range of read sizes from 36 to 76 base pairs.
- "Completing" a genome has never been the objective.
  (for many purposes, a reference is not necessary)
- Developing a high-density and direct marker panel for wide-scale screening is a main objective.

Sample selection

·Sequencing exemplar species representing interesting phenotypic and geographic variation.

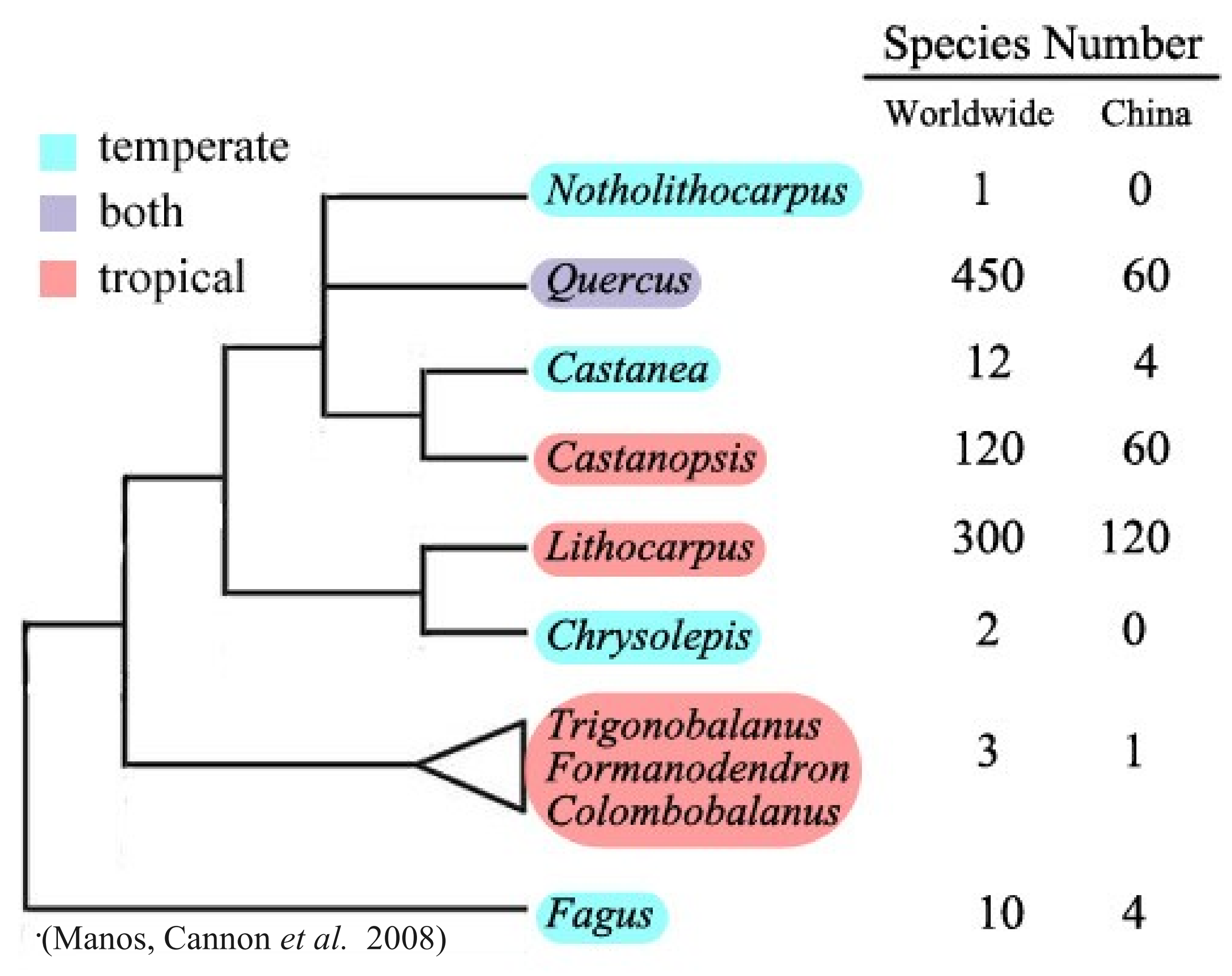
·Geographic samples includes Borneo and China.

Fagaceae



- Important component in global forests
- 8 genera
- woody
- about 1000 species in the world and 350 species in China
- tropical to boreal
- centered in tropical Asia and North America

Phylogeny and Species Richness



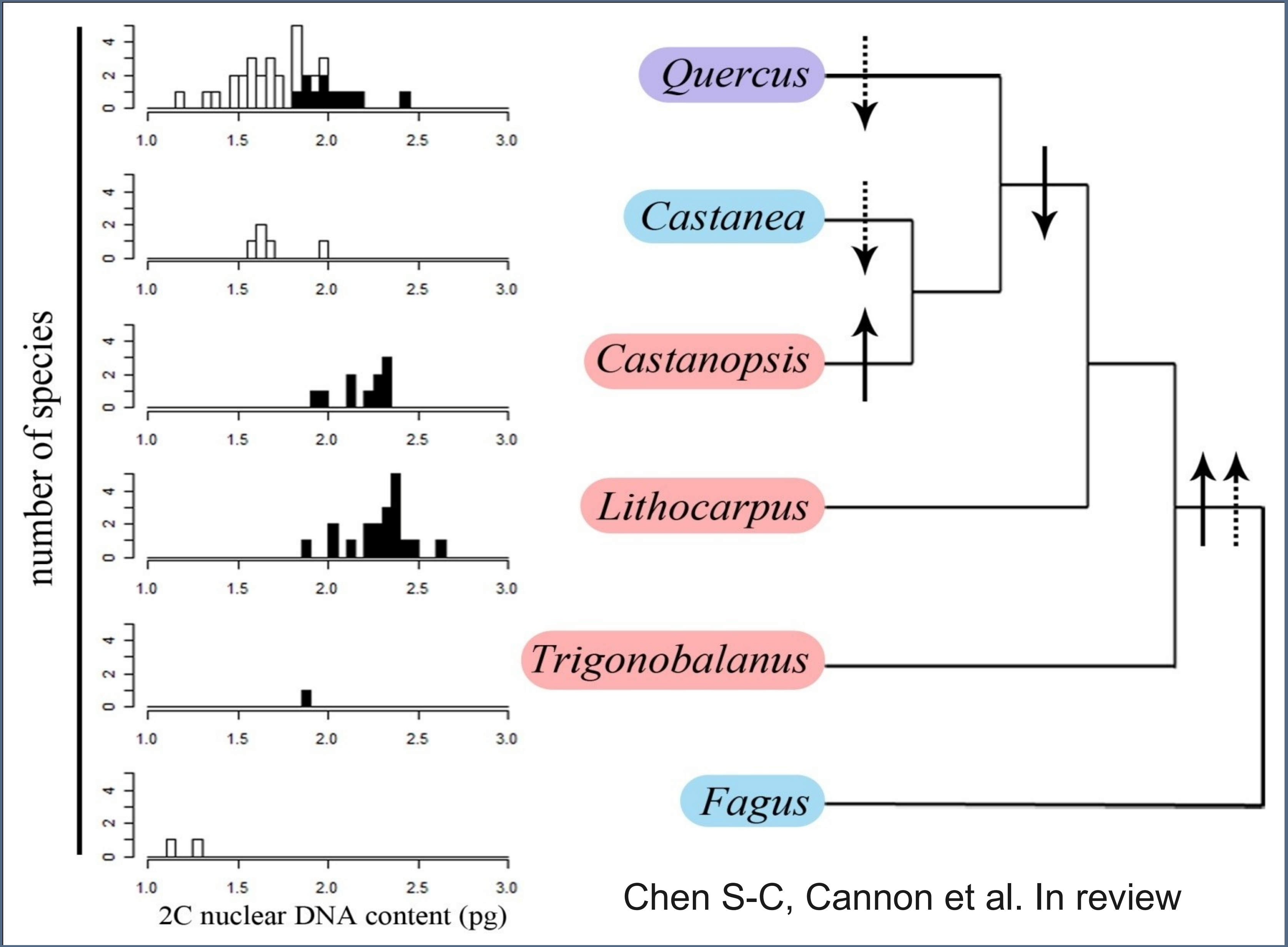
	Species Number	
	Worldwide	China
Notholithocarpus	1	0
Quercus	450	60
Castanea	12	4
Castanopsis	120	60
Lithocarpus	300	120
Chrysolepis	2	0
Trigonobalanus *Formanodendron* *Colombobalanus*	3	1
Fagus	10	4

·(Manos, Cannon *et al.* 2008)

·Ploidy number is fixed in most genera (n=12), except the relictual Trigonobalanus (n=7), with little evidence of polyploids.

·(Demerico et al. 1995; Chen et al. 2007; Chen and Sun 2010; Armstron and Wylie 1965)

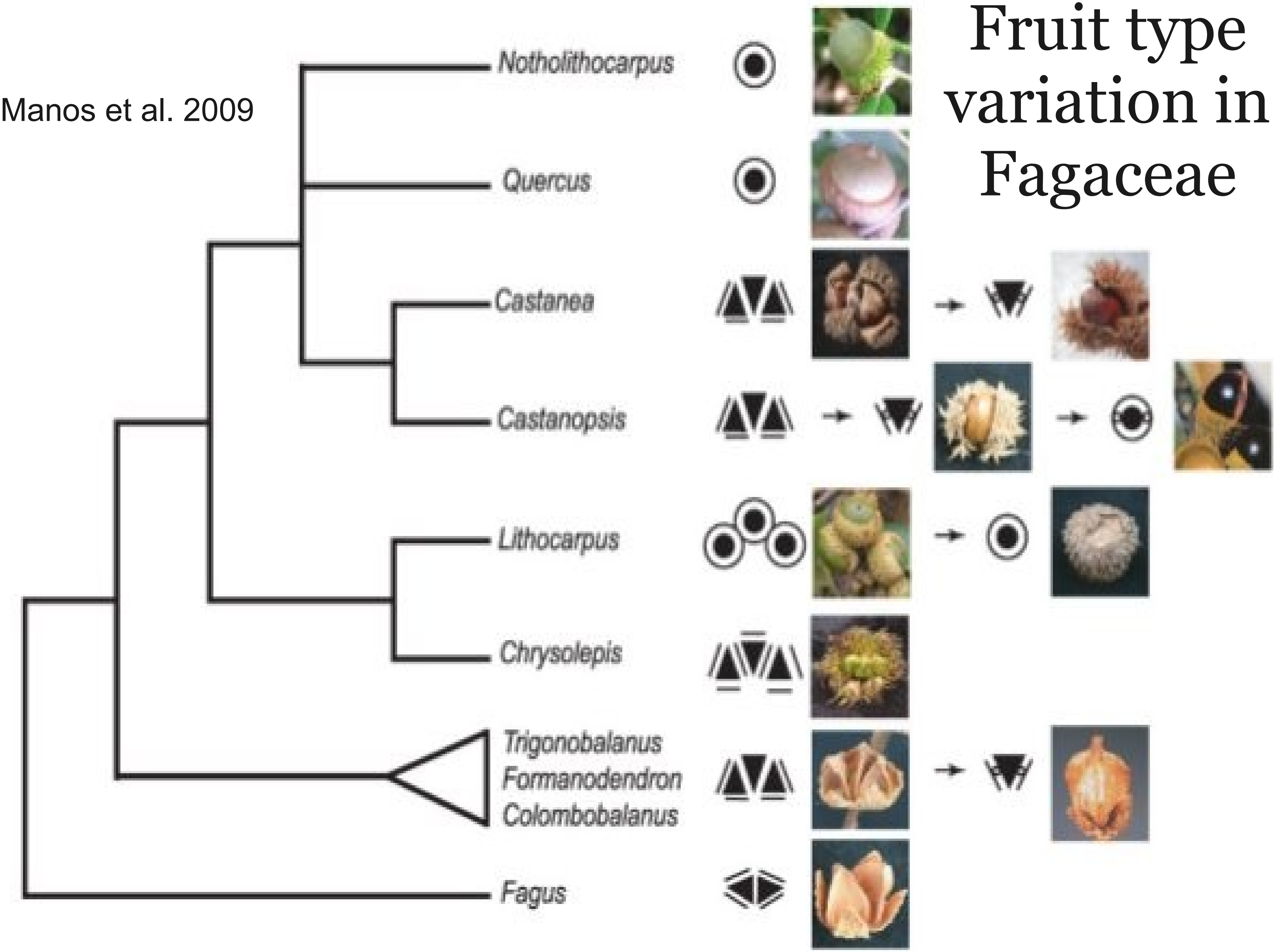
Tropical species have slightly larger genomes



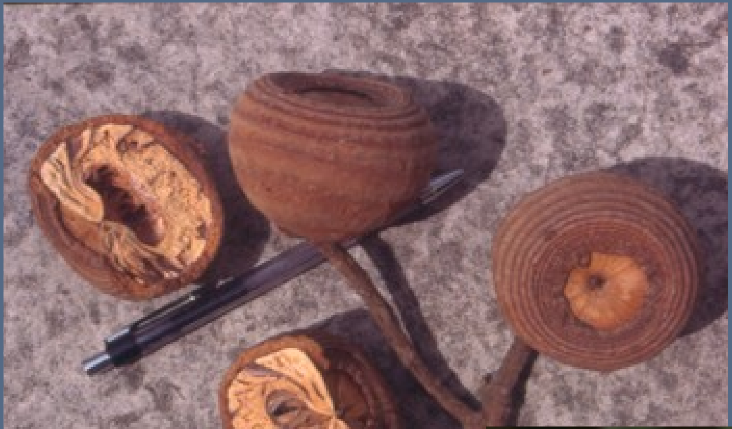
Chen S-C, Cannon et al. In review

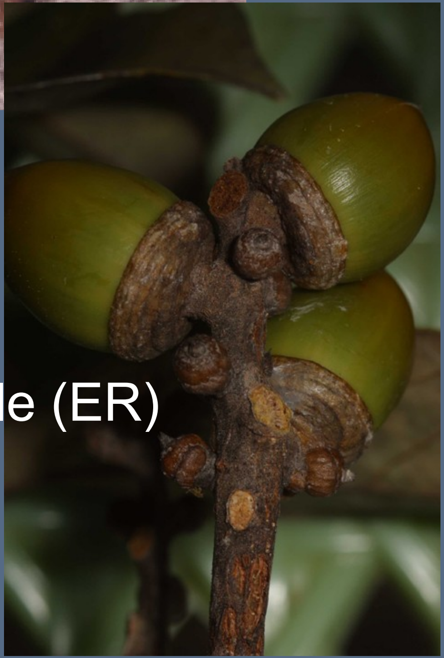
Fruit type variation in Fagaceae

Manos et al. 2009

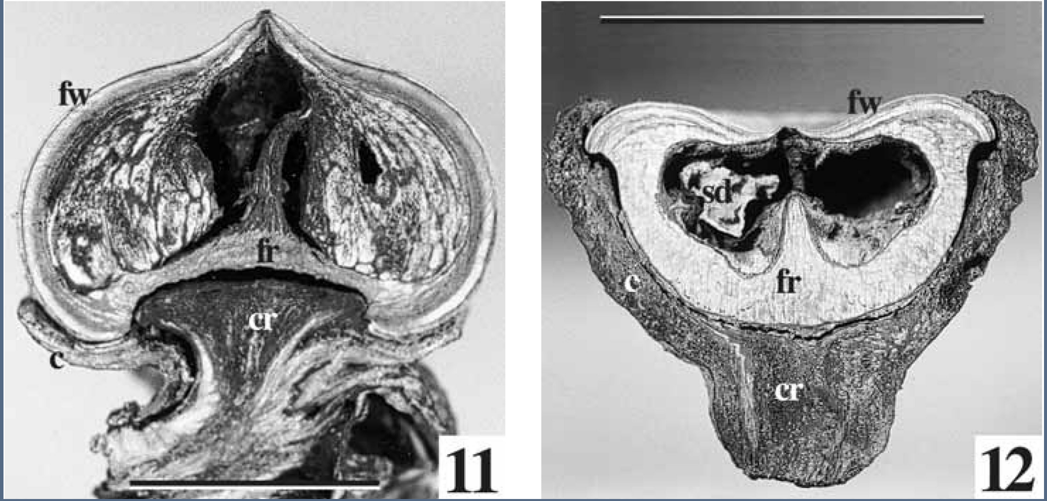


Lithocarpus fruits





Cannon and Manos 00



Two main types

Acorn

Enclosed receptacle (ER)

Typically found living sympatrically in mixed communities.

Two extremes of fruit type

Evidence of a trade-off between chemical and physical protection of seeds (Chen et al. in revision)

Lots of other potential life history correlates.

Our naïve questions

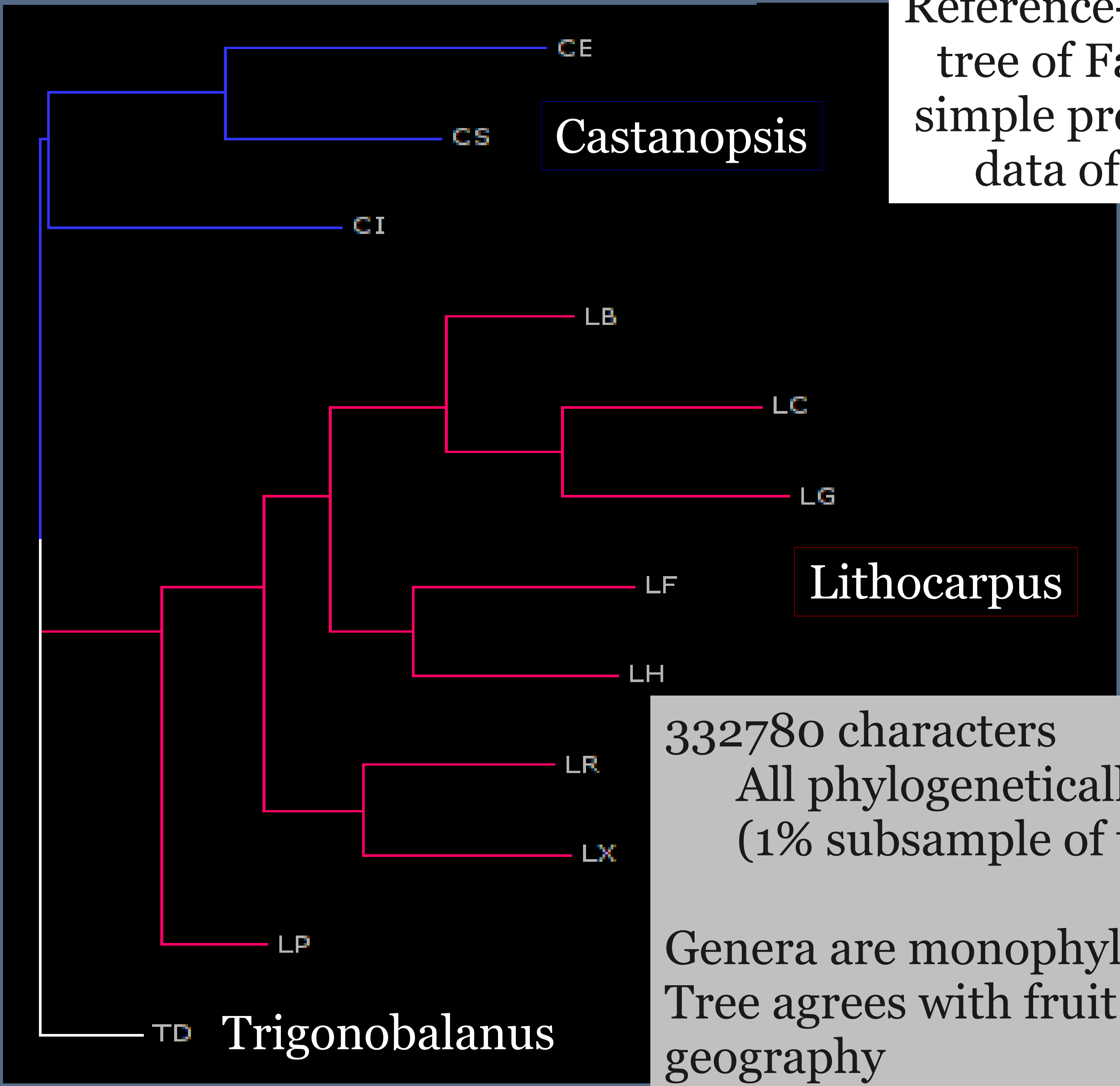
How do tropical forest tree species diversify, from a genomic perspective?

- big novel innovations or small trivial changes?
- participation in a syngameon?
- copy number variants?
- regulatory elements
  - (if no apparent differences found → RNAseq)

What genetic elements are associated with phenotypic diversity?

- how does fruit type affect genomic diversity?
- can we discover functional elements?
- how big a role do repetitive elements play?
- is there no obvious association?

Reference-free parsimony tree of Fagaceae, using simple presence-absence data of 25bp kmers



332780 characters

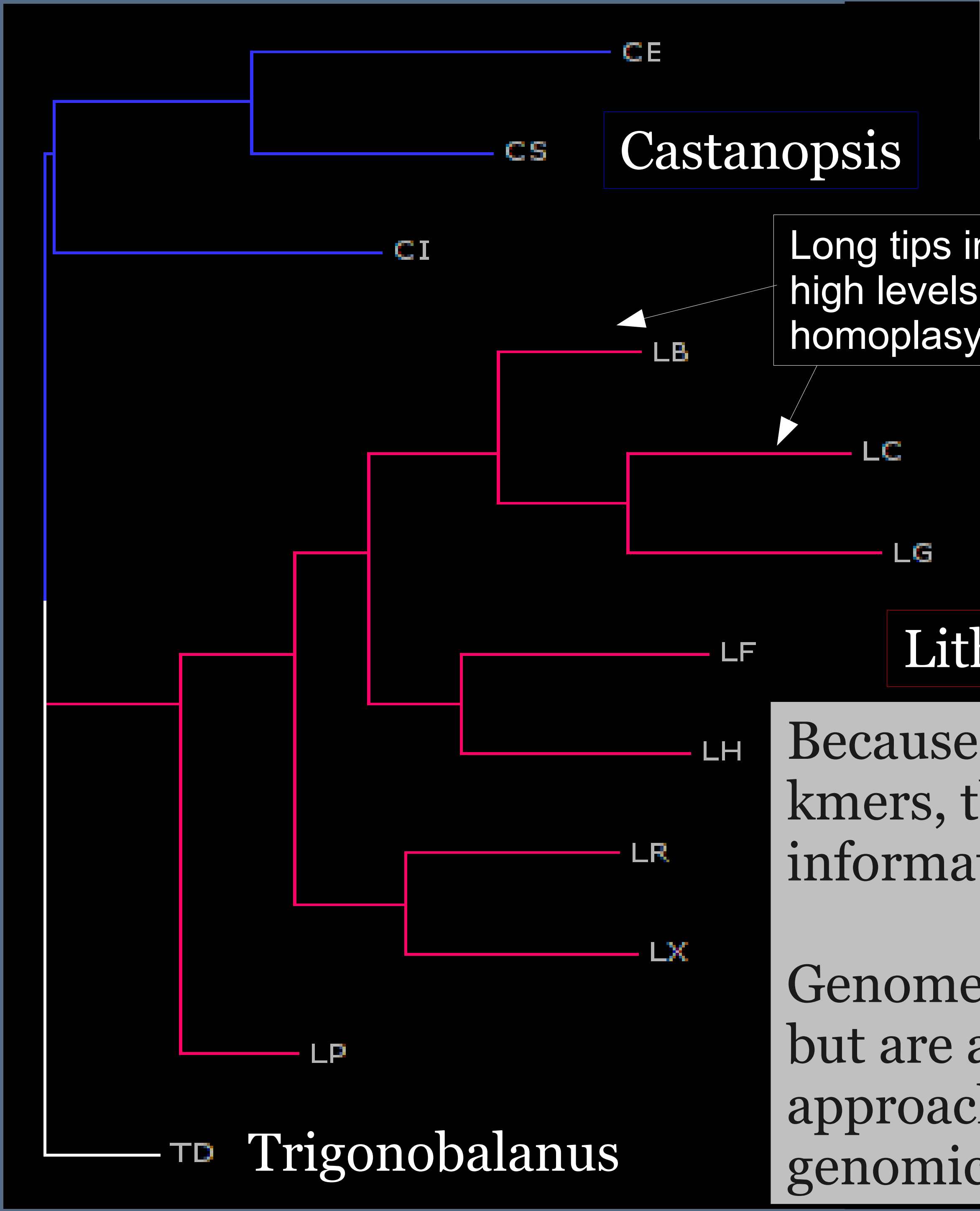
All phylogenetically informative

(1% subsample of total shared kmers)

Genera are monophyletic

Tree agrees with fruit type and geography

Reference-free parsimony tree of Fagaceae, using simple presence-absence data

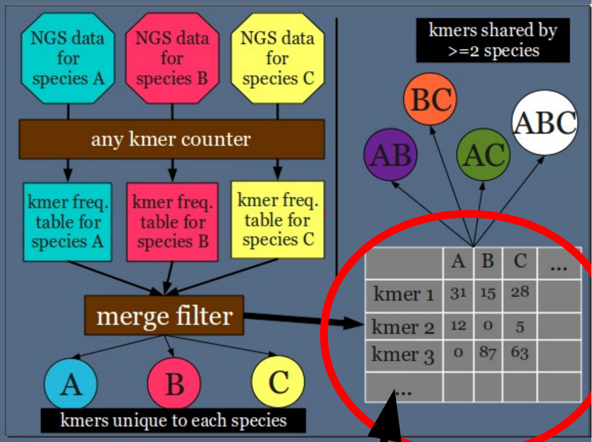


Castanopsis

Long tips indicate high levels of homoplasy

Lithocarpus

Trigonobalanus

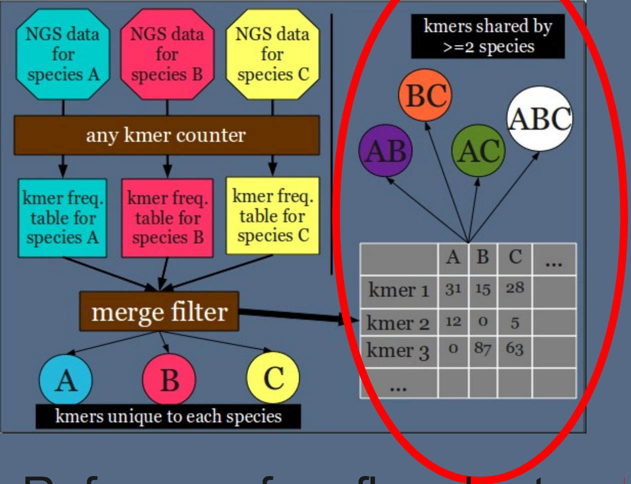

NGS data for species A
NGS data for species B
NGS data for species C
any kmer counter
kmer freq. table for species A
kmer freq. table for species B
kmer freq. table for species C
merge filter
A
B
C
kmers unique to each species
kmers shared by >=2 species
BC
ABC
AB
AC

	A	B	C	...
kmer 1	31	15	28	
kmer 2	12	0	5	
kmer 3	0	87	63	
...				

Reference-free flowchart

Because it only contains shared kmers, there such be no “tip” information in this table.

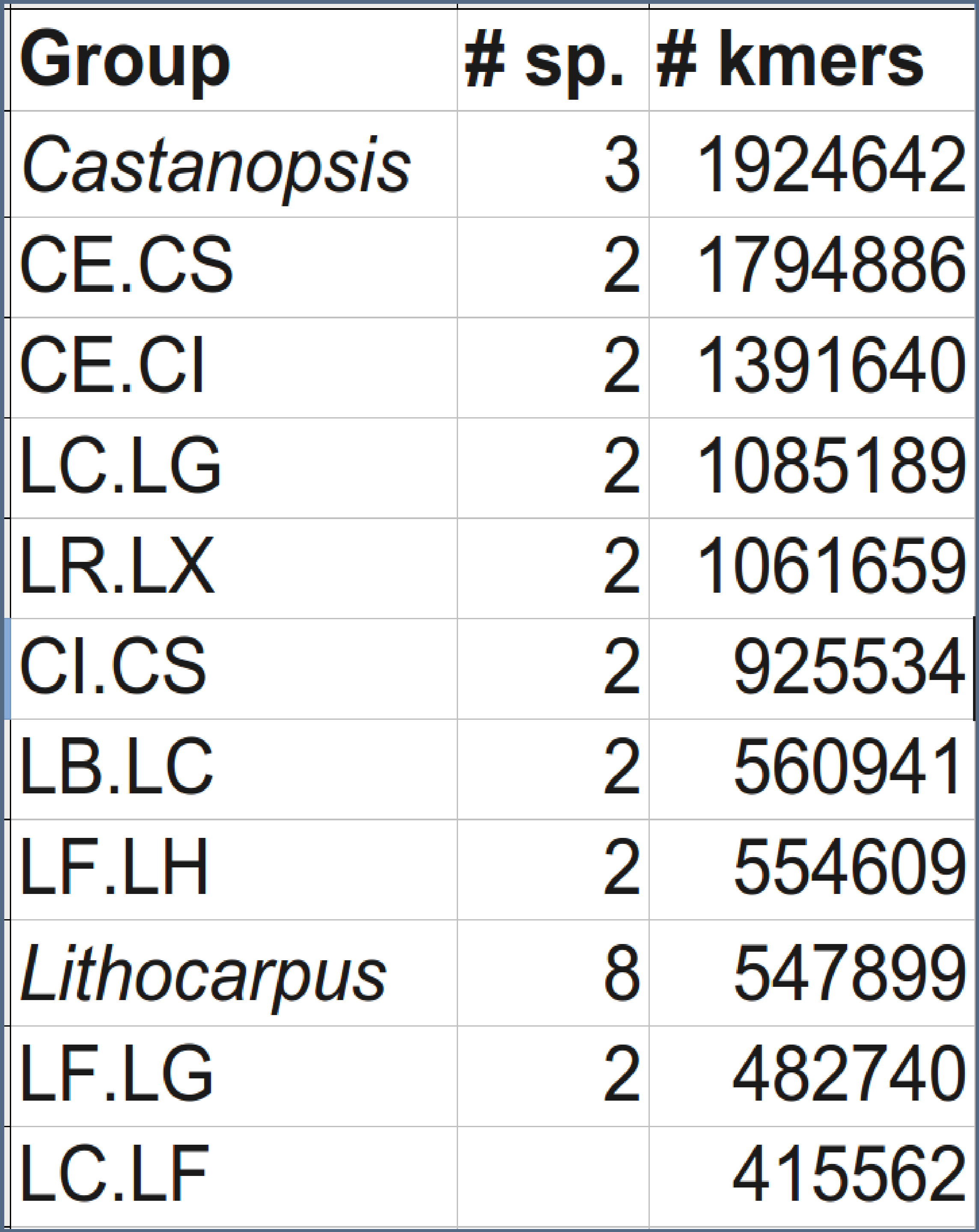
Genomes do not have “ONE” history but are a mixture of histories. This approach can be used to disentangle genomic partitions.



Reference-free flowchart

The Castanopsis species are quite distinct from other Fagaceae but are NOT terribly distinct from one another.

Instead many species appear to be strongly "mixed", indicating interspecific hybridization or retention of ancestral polymorphisms.

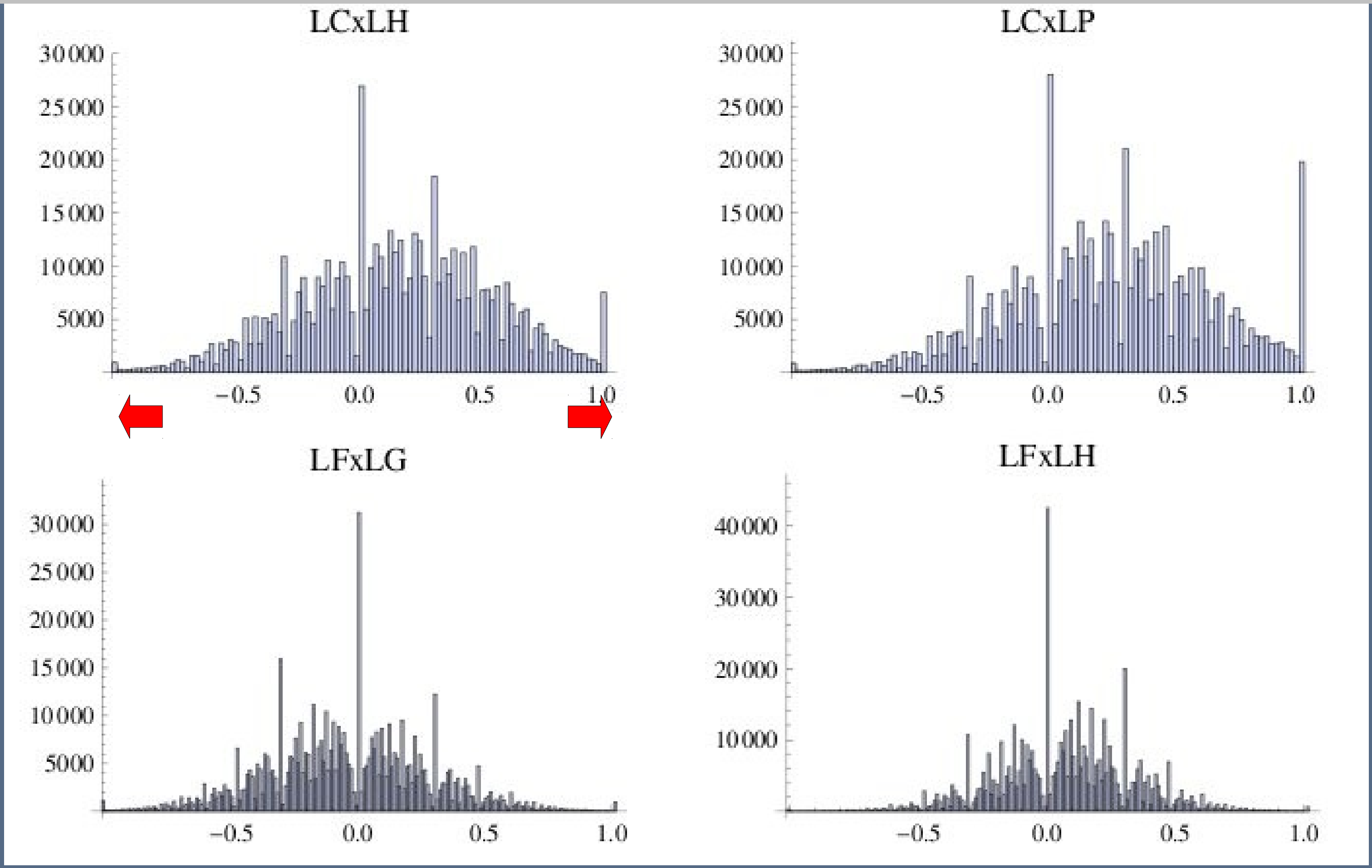


Group	# sp.	# kmers
Castanopsis	3	1924642
CE.CS	2	1794886
CE.CI	2	1391640
LC.LG	2	1085189
LR.LX	2	1061659
CI.CS	2	925534
LB.LC	2	560941
LF.LH	2	554609
Lithocarpus	8	547899
LF.LG	2	482740
LC.LF		415562

Now, looking at *Lithocarpus* specific kmers

=========

Plots of Log10 ratio of standardized frequencies fit normal distributions and outliers can be identified for more detailed studies.



While most of these are uncharacterized by BLAST, numerous examples of retrotransposons and ubiquitins are identified.

More on *Lithocarpus* specific kmers

=========

Local assembly is poor and almost completely exists of SNP-like contigs (~2x read length). [[this is not surprising, given the depth of genus]]

Roughly 25 contigs longer than 300 bp do assemble, the vast majority of which are previously unknown, although several are on the mitochondrion genome.

These could be a good starting point, if one was interested in discovering the conserved genetic elements associated with *Lithocarpus* species.

Advantages of reference-free comparative genomics

- Allows a quick analysis of NGS data, prior to assembly.
- Makes few assumptions about underlying process of divergence.
- Provides an simple estimate of phylogenomic relationships

  [[reconstruction model can be improved]]
- Greatly reduces the complexity of the data, given a specific comparative question.
- No reference needed.

Acknowledgements

- Xishuangbanna Tropical Botanical Garden, Yunnan, China
  - Particularly its Director (Prof. CHEN Jin)
- Chinese Academy of Sciences
- Yunnan Provincial Government (2009-2012)
  - Funds to study the genomics of biodiversity
- Texas Tech University
- Conservation Trust Fund, National Geographic Society
- Canada’s Michael Smith Genome Sciences Center, especially ZHAO Yong-Jun

- Chai-Shian KUA, RUAN Jue, ZHANG Di, CHEN Sichong, Matt HELMUS, YANG FengXi, YU Song, and FAN Huan.