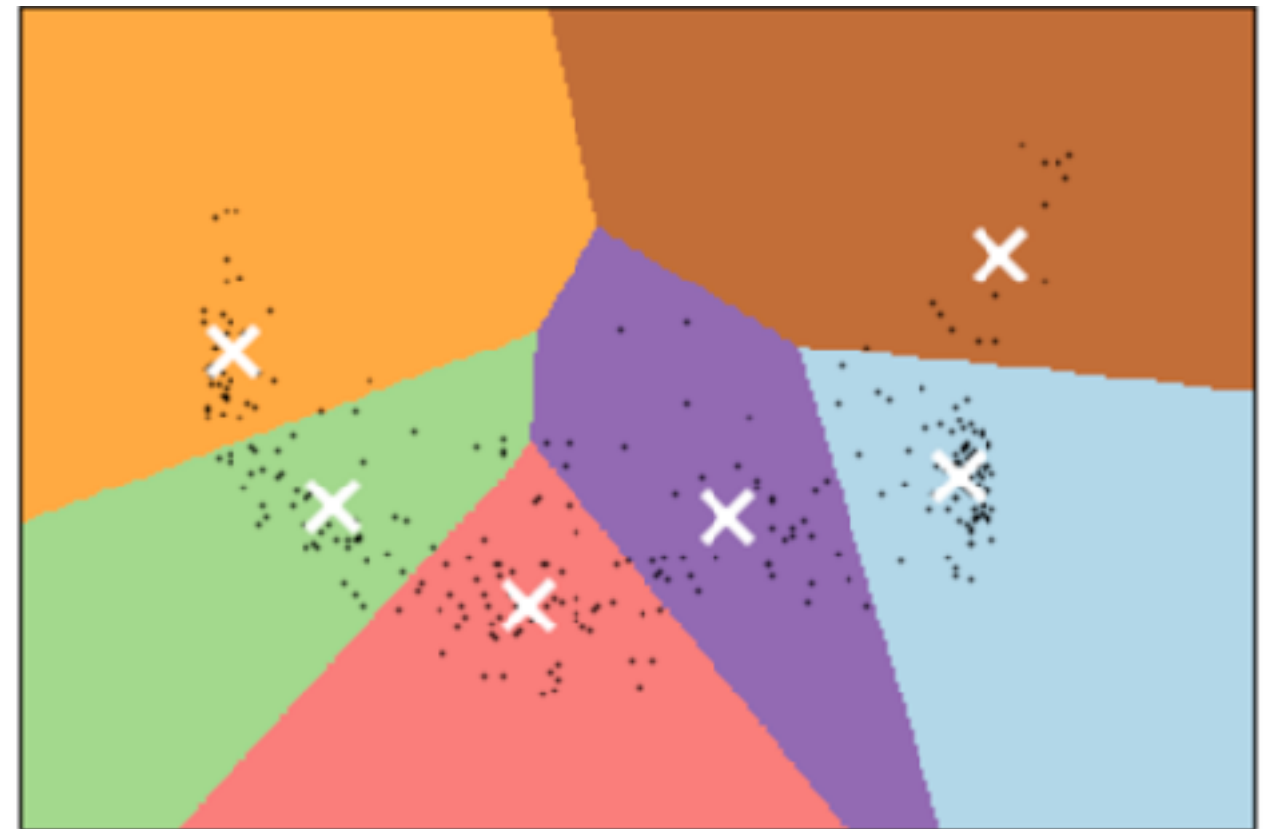


MATT JOHNSON, CHICAGO BOTANIC GARDEN

mjohnson@chicagobotanic.org



@mossmatters



MarkerMiner
Locus development made easy

OUTLINE

Locus selection: two strategies

Examples at different phylogenetic scales

PAFTOL probe design from 1KP

Hands On With MarkerMiner

PROBE DESIGN: TWO STRATEGIES

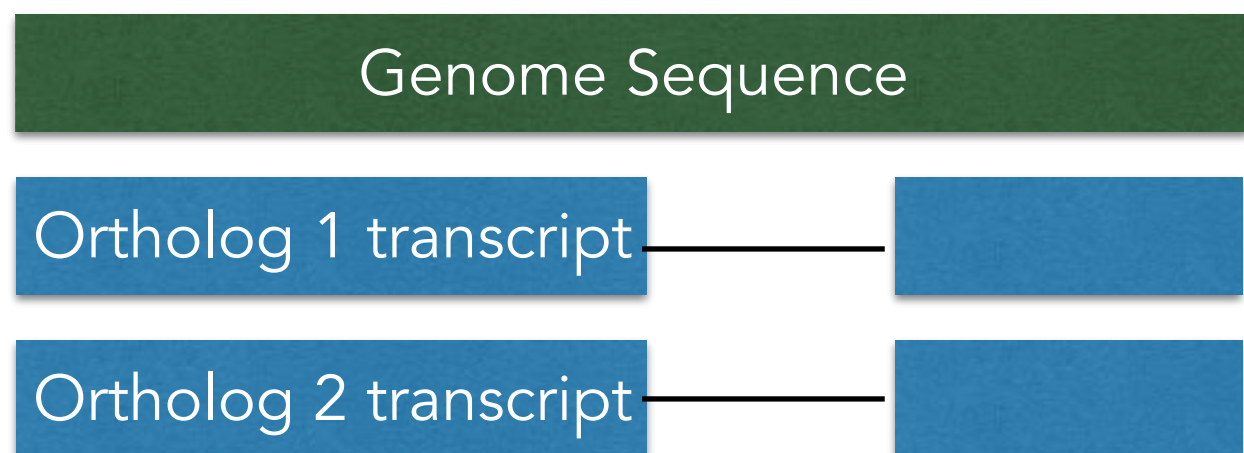
Genome Skimming

- 10x - 20x genome coverage
- Intron location information
- May not be suitable for large genomes
- Organellar genome assembly

Transcriptome Sequencing

- 20-30 million RNASeq reads
- Reduced genome representation: exons only
- Tissue and time dependent
- Can multiplex several species

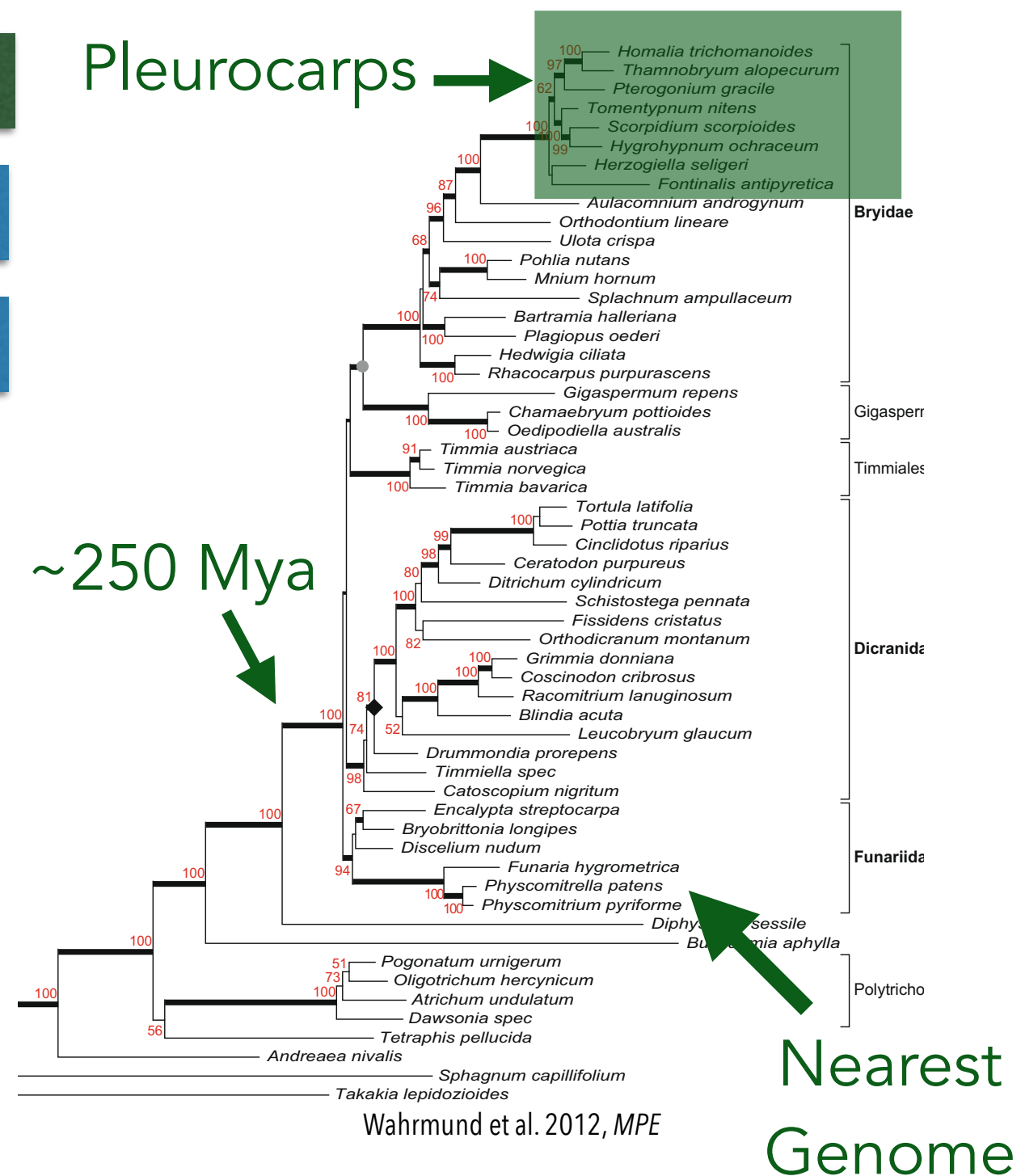
TARGET DESIGN EXAMPLES: MOSS TREE OF LIFE



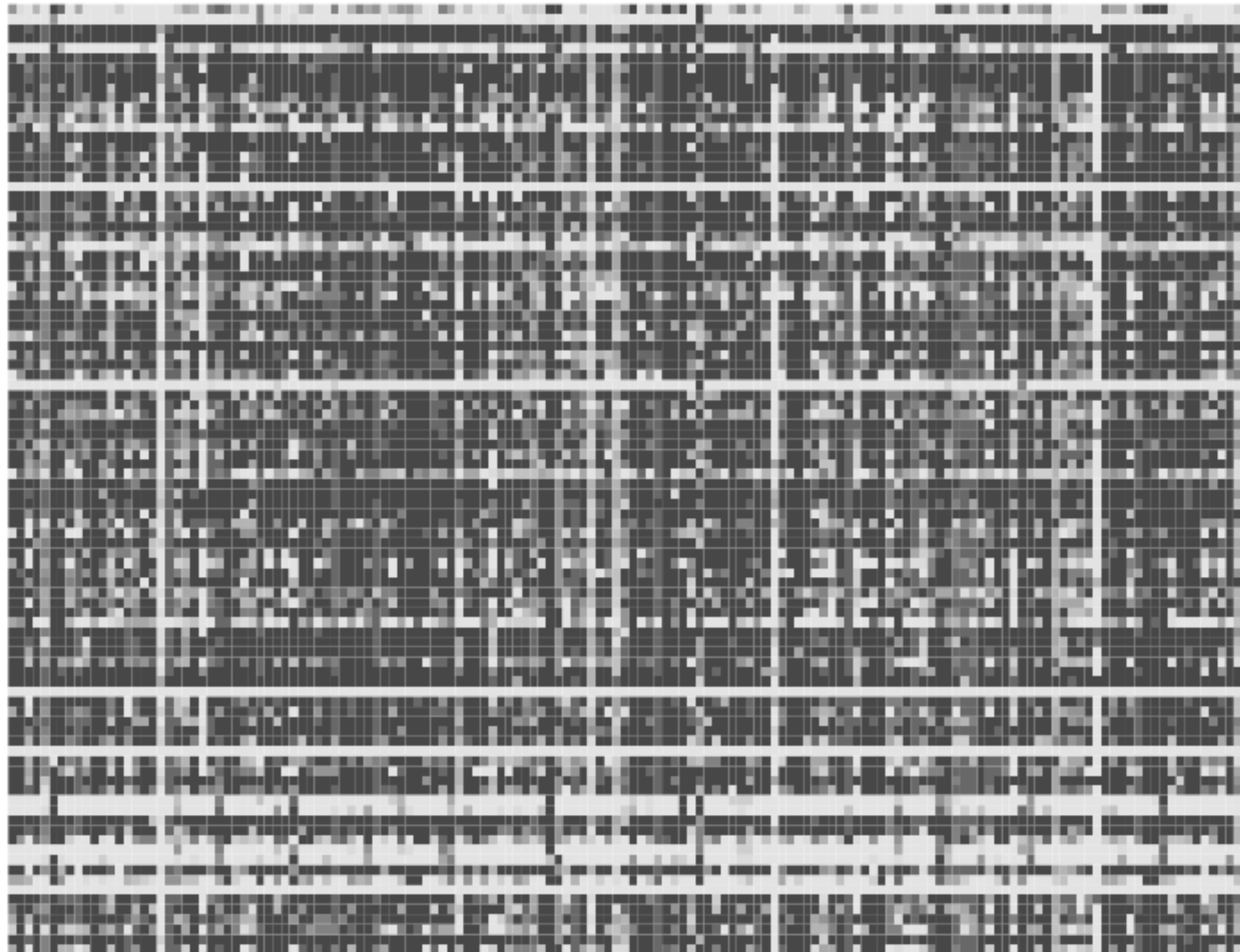
Align moss transcripts (1KP) to
Physcomitrella genome

Select genes expressed in at least
two pleurocarpous mosses

Design probes from multiple
sequences



TARGET DESIGN EXAMPLES: MOSS TREE OF LIFE

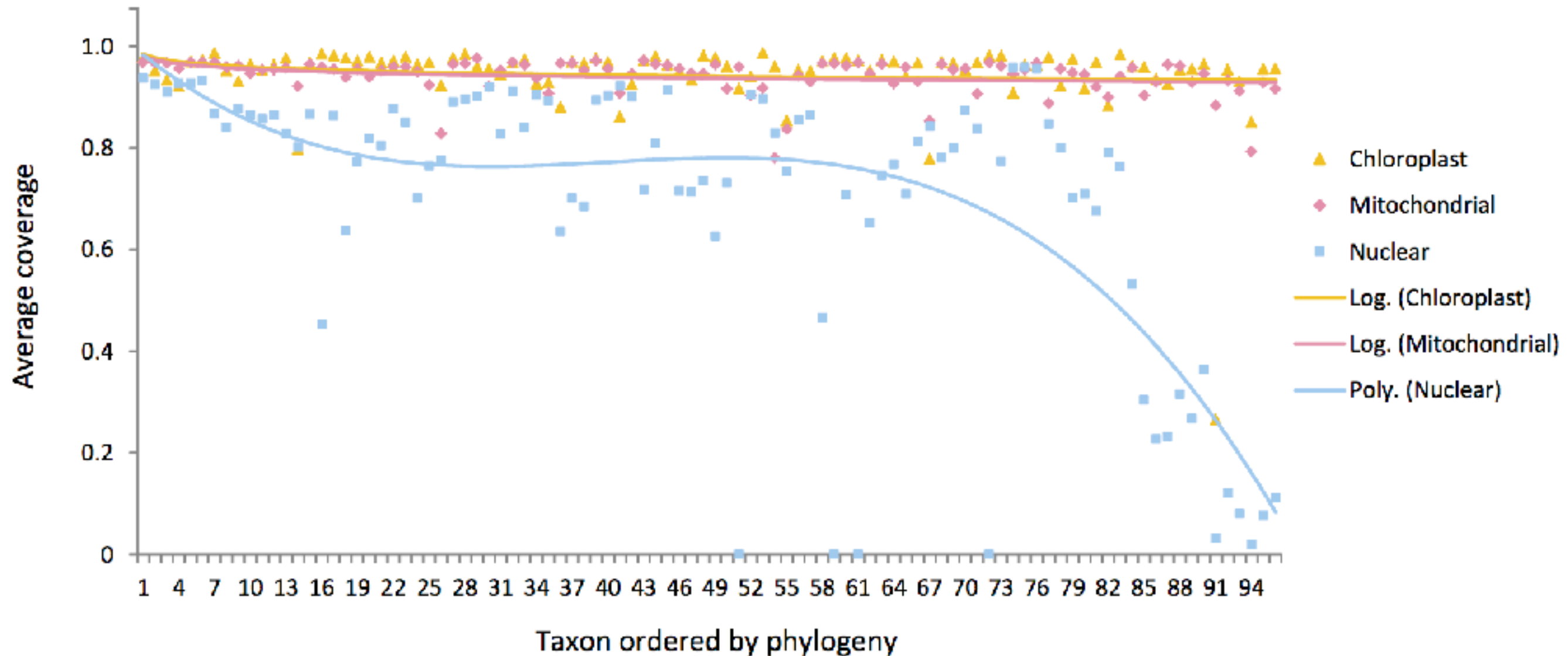


96 Samples

150 Genes

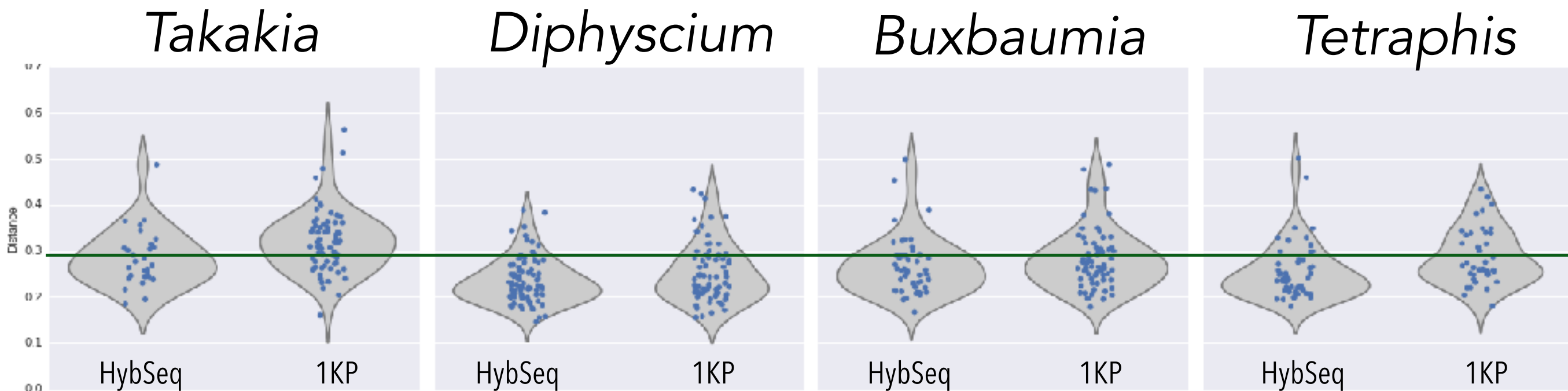
Liu, Johnson, et al., in prep

TARGET DESIGN EXAMPLES: MOSS TREE OF LIFE



Liu, Johnson, et al., in prep

TARGET DESIGN EXAMPLES: MOSS TREE OF LIFE



Comparing divergence between probe sequences and:

Sequences recovered by HybPiper (left)

Transcripts from 1KP (right)

In mosses far diverged from *Physcomitrella*

Liu, Johnson, et al., in prep

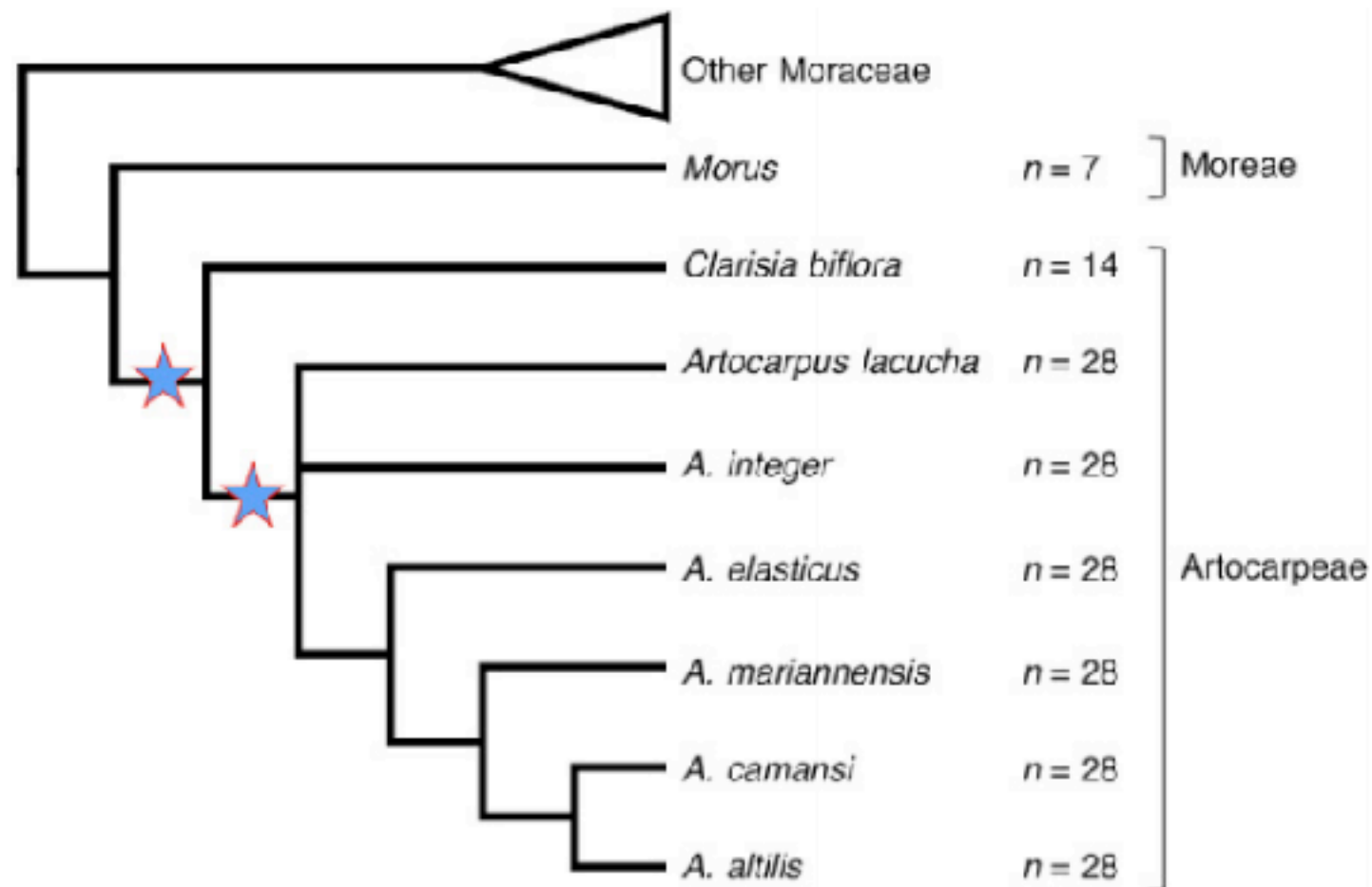
THE IMPORTANCE OF MULTIPLE ORTHOLOGS PER GENE

More accurate alignment

Provides redundancy (free extra tiling)

Expands phylogenetic breadth

TARGET DESIGN EXAMPLES: ARTOCARPUS GENOME SKIM



17x Whole-Genome Sequencing (*Artocarpus camansi*)

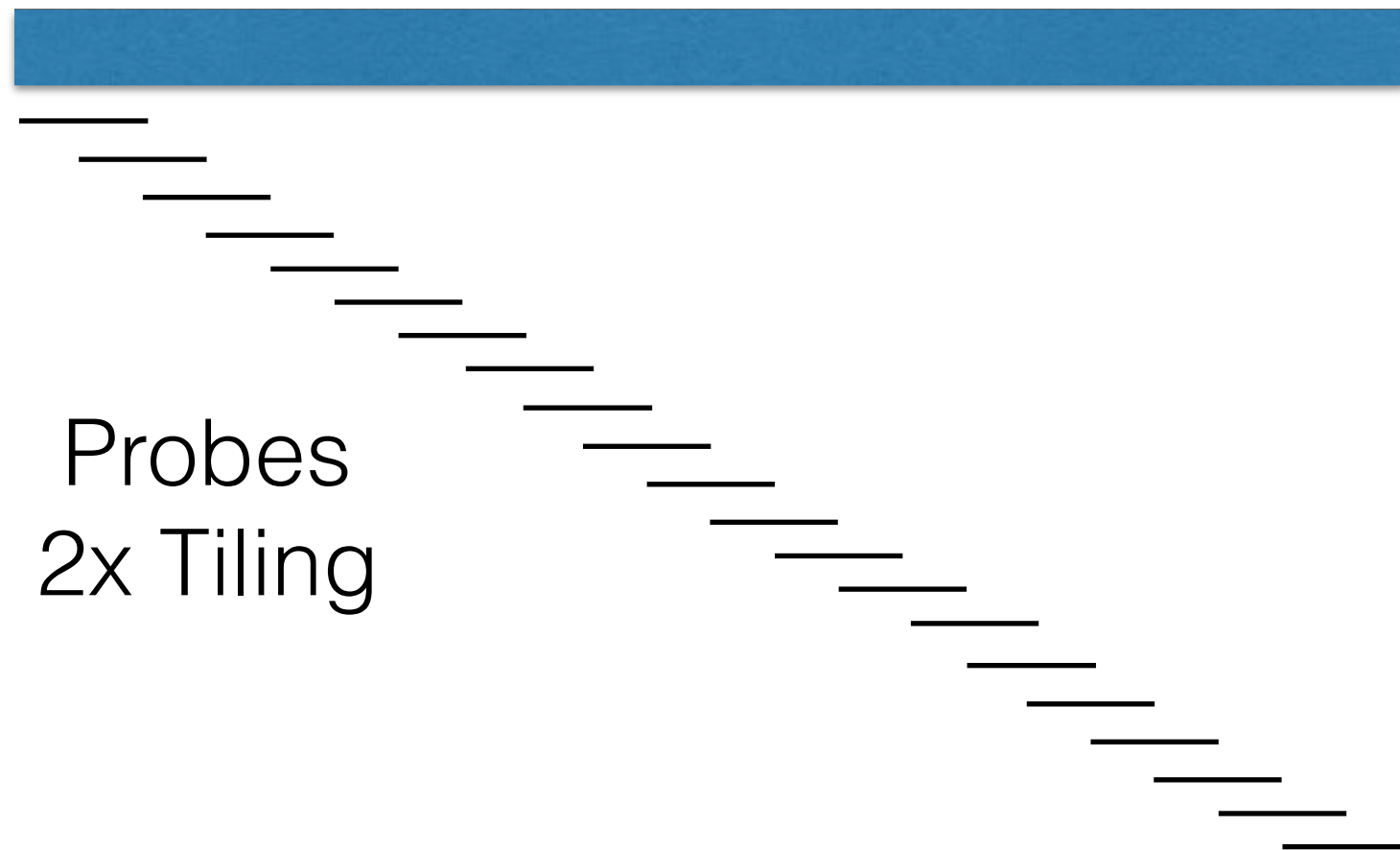
Determined orthology with *Morus*

333 "phylogeny" genes, plus MADS-Box and volatiles

Gardner et al., APPS, 2016

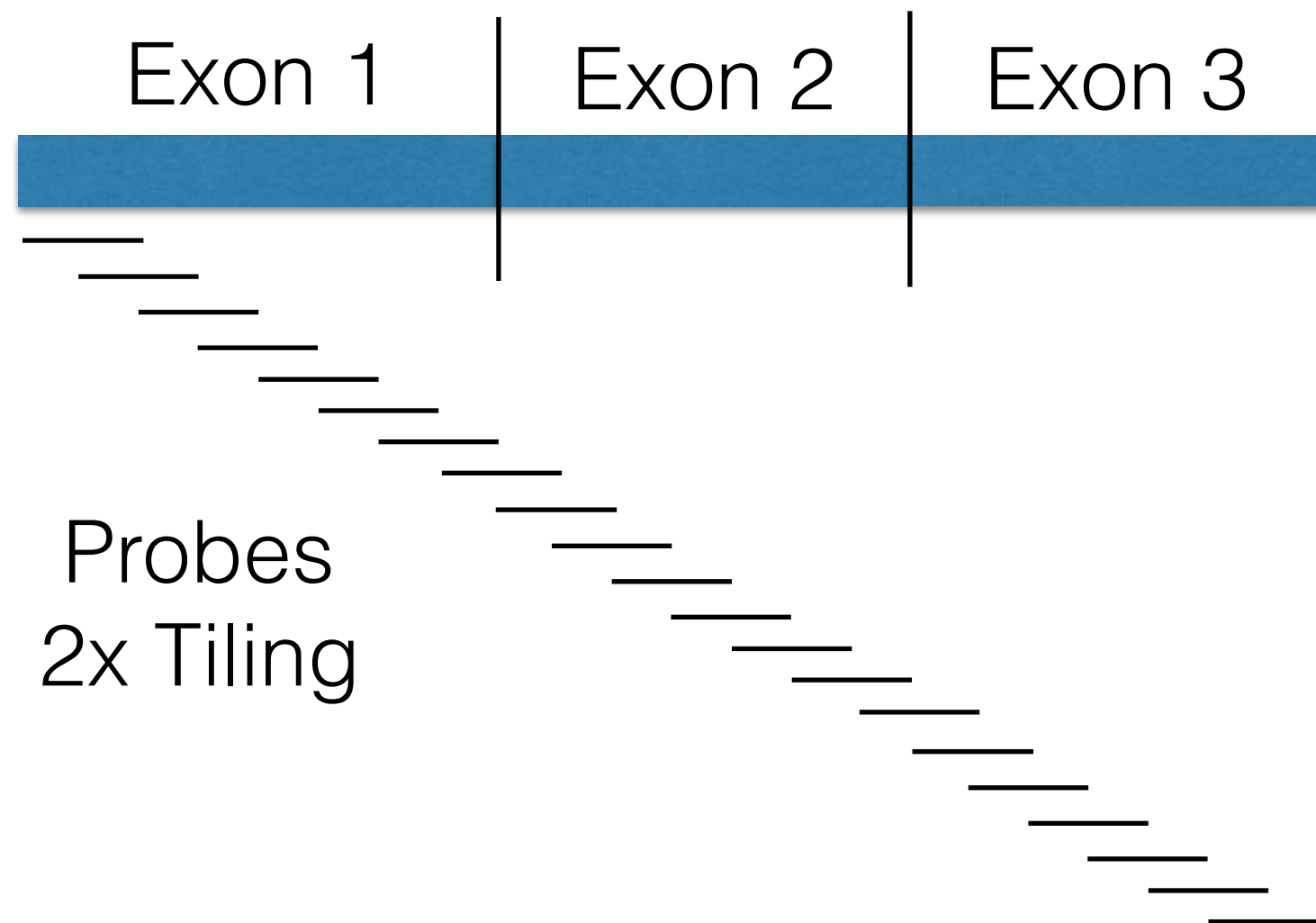
TARGET DESIGN EXAMPLES: USING GENOME SEQUENCE

Transcript



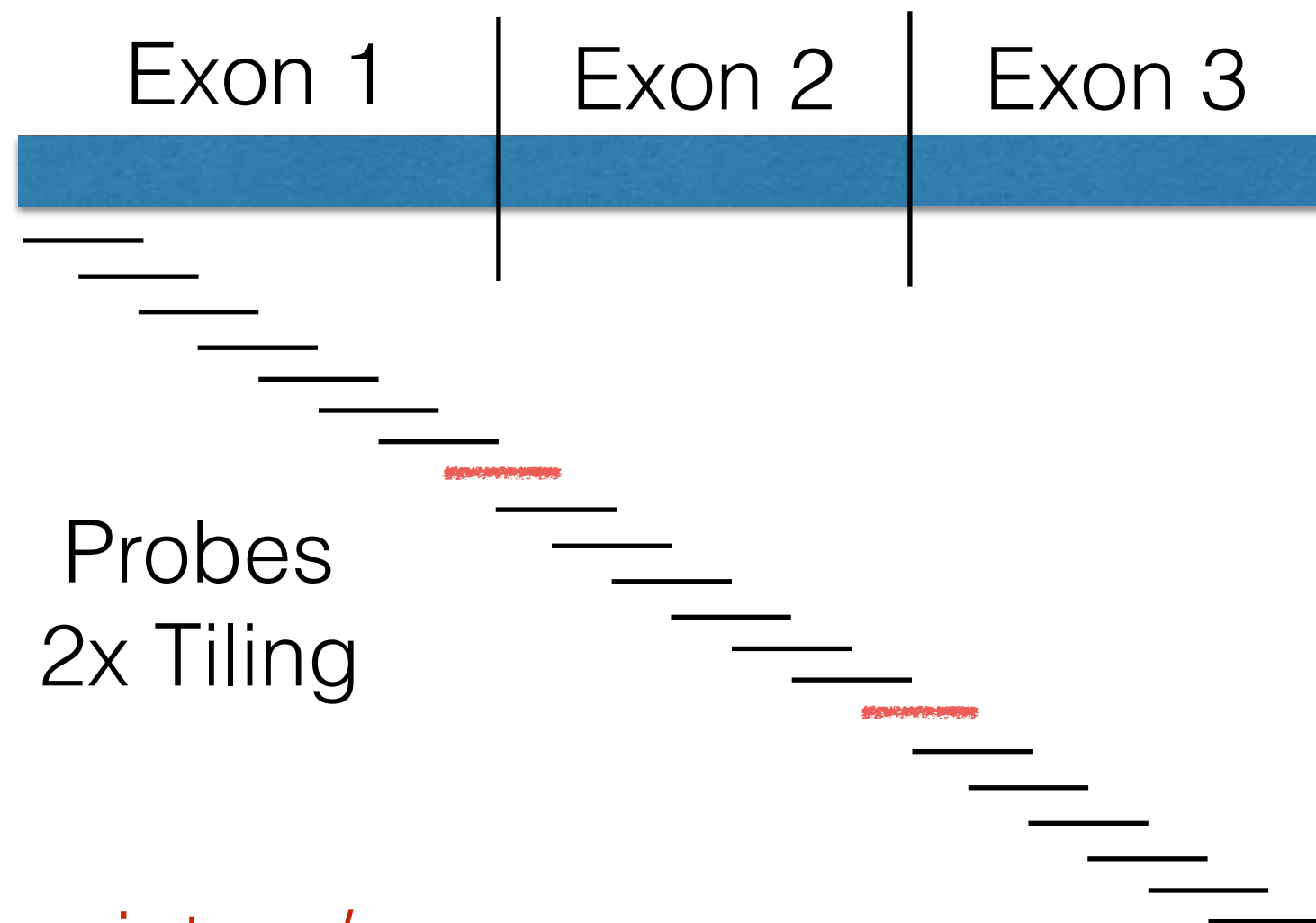
TARGET DESIGN EXAMPLES: USING GENOME SEQUENCE

Transcript



TARGET DESIGN EXAMPLES: USING GENOME SEQUENCE

Transcript



Probes spanning intron/exon
boundaries may not work

TARGET DESIGN EXAMPLES: USING GENOME SEQUENCE

Genome Scaffold



Exons

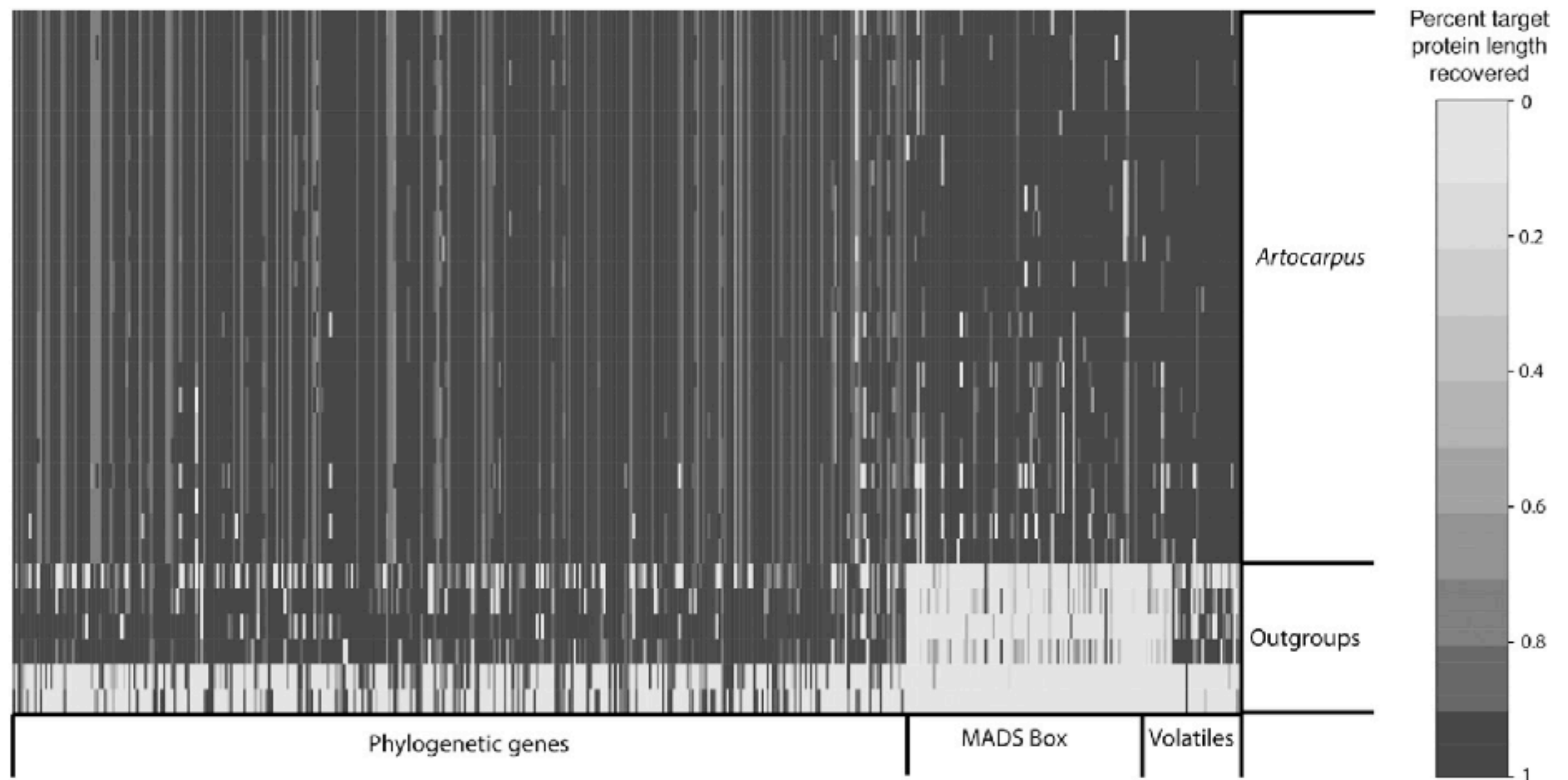


```
DAT1 DAT1 sample 2001962
DAT2 DAT2 sample 2016938
DAT3 DAT3 sample 2097319
DAT4 DAT4 sample 2007085
AT3G07700
```



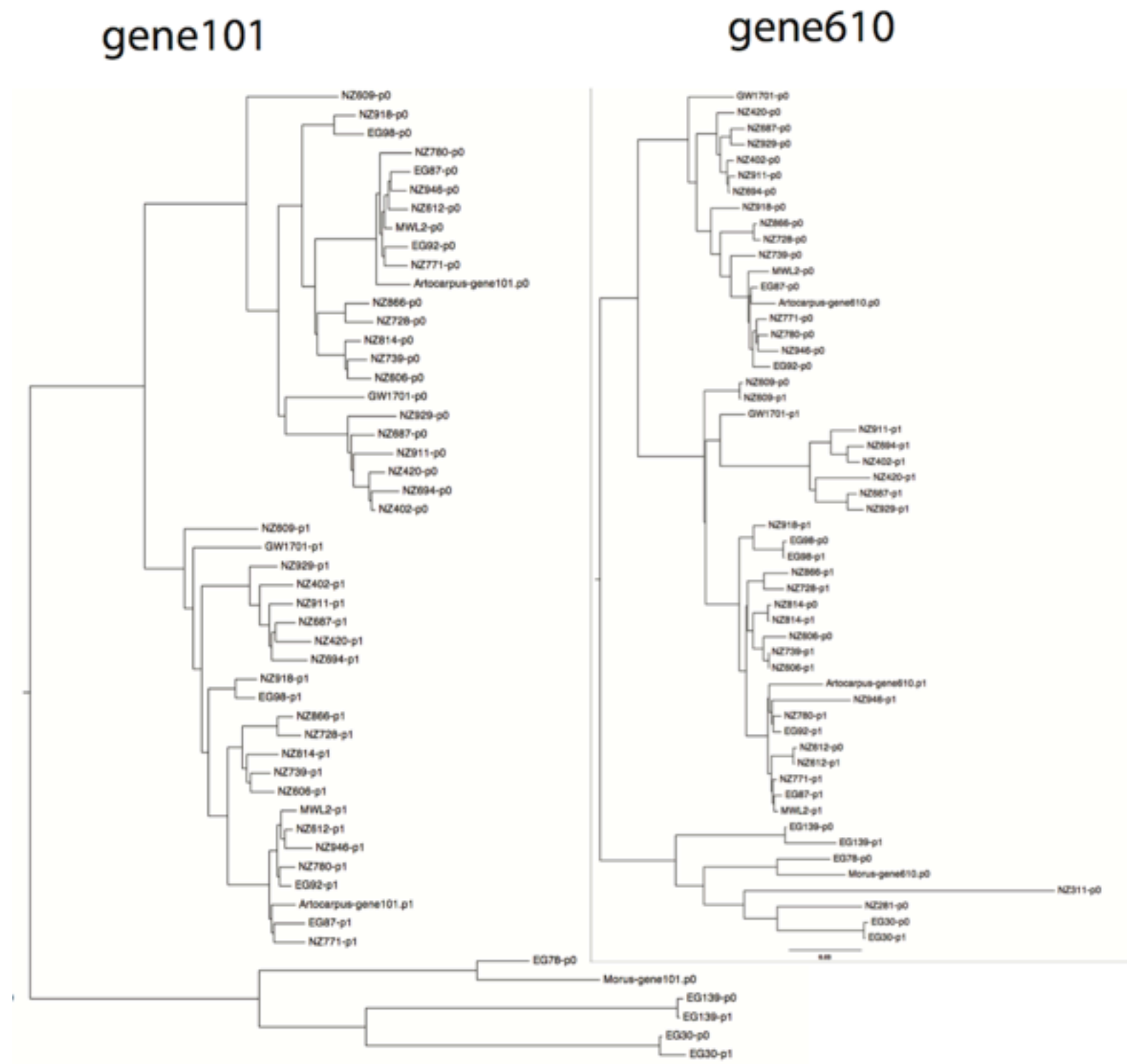
With genome, probes extend to intron/exon boundaries
Remainder of probe can be filled with T

TARGET DESIGN EXAMPLES: ARTOCARPUS GENOME SKIM



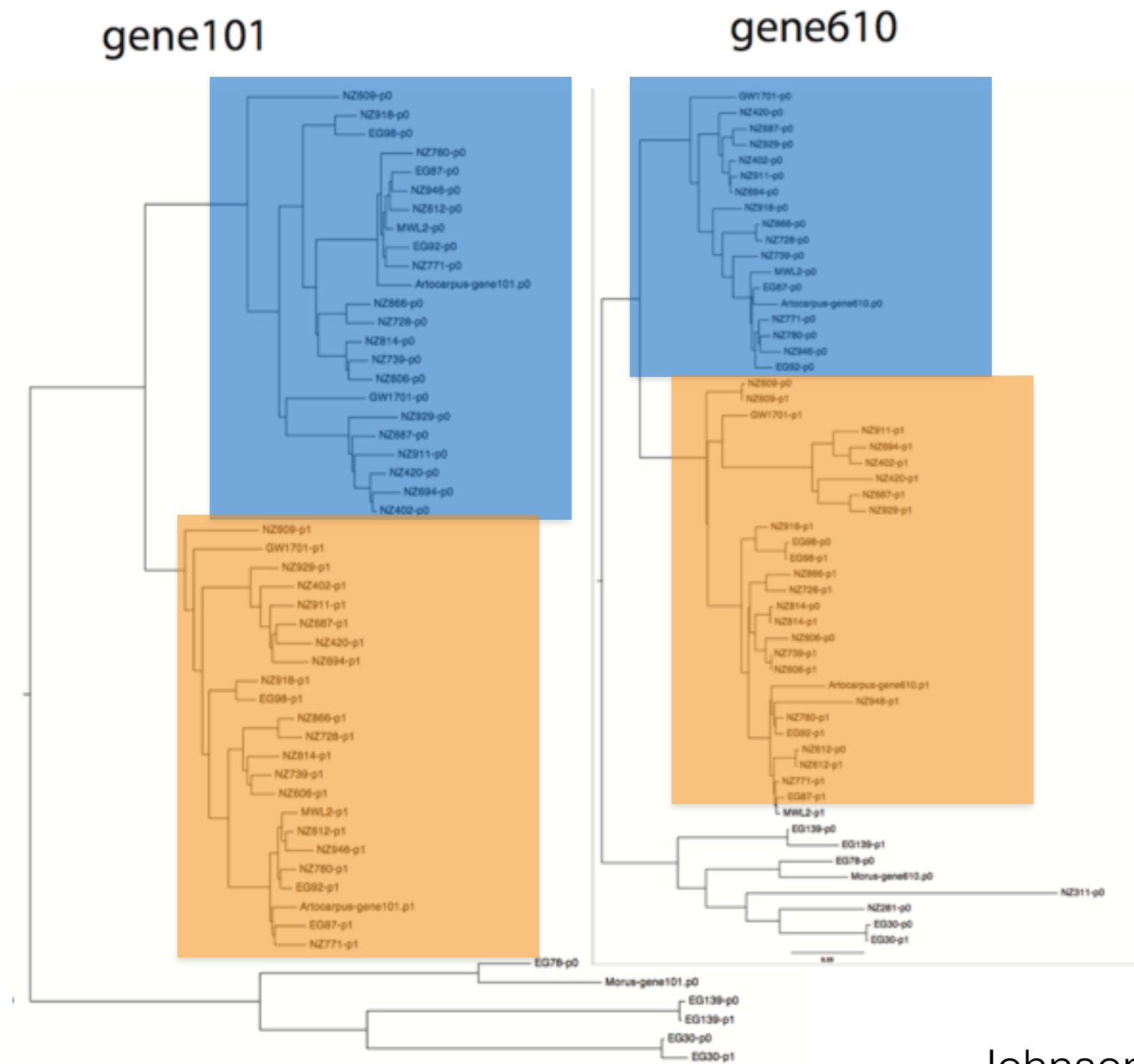
Johnson et al., APPS, 2016

TARGET DESIGN EXAMPLES: ARTOCARPUS GENOME SKIM



Johnson et al., APPS, 2016

TARGET DESIGN EXAMPLES: ARTOCARPUS GENOME SKIM



Johnson et al., APPS, 2016

PAFTOL PROBE DESIGN

Build a genus level
phylogeny of flowering
plants

One Kit to Rule Them All?

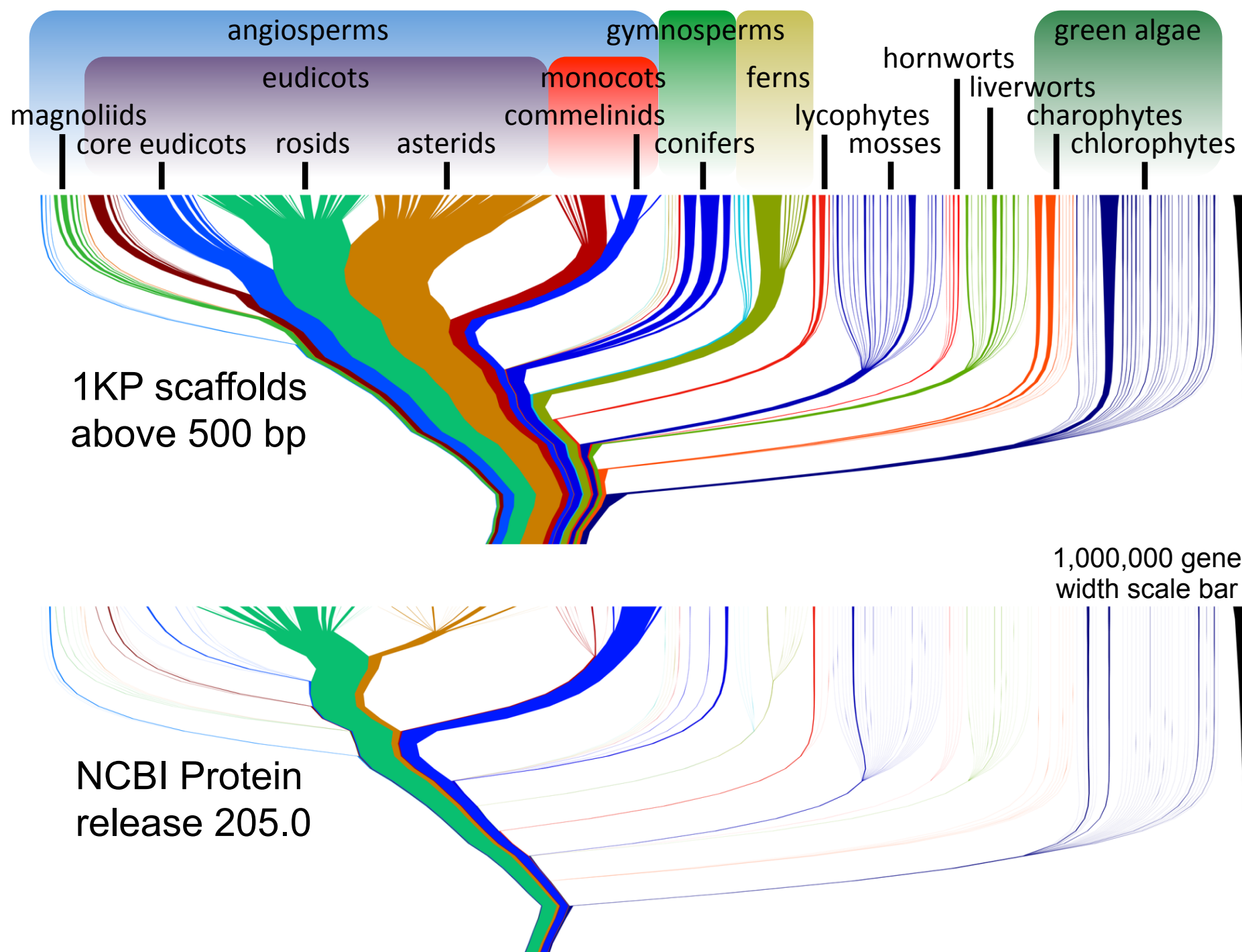
Which genes?

Which orthologs?



Data: 400 genes from 1000
Plant Transcriptome Project
(1KP)

PAFTOL HYBSEQ PROJECT: COVERING ALL ANGIOSPERMS



onekp.com

Phylogenetic Tree of Green Plants (Viridiplantae) Genes

Phylogenetic tree of Green Plants (Viridiplantae) based on gene data. The tree is rooted at the bottom left and branches upwards. Major clades are labeled on the left, and specific species names are listed on the right. Bootstrap values are indicated at the nodes.

Major Clades and Species:

- Streptophytes**
 - Chlorophyta**
 - Nephroselmis pyriformis*
 - Pyramonas parakeae*
 - Monomastix opisthostigma*
 - Uronema* sp.
 - Mesostigmatales & Chlorokybales**
 - Mesostigma viride*
 - Spirotaenia minuta*
 - Chlorokybus atmophyicus*
 - Klebsormidium subtile*
 - Entransia fimbriata*
 - Chara vulgaris*
 - Chaetosphaeridium globosum*
 - Coleochaete irregularis*
 - Coleochaete scutata*
 - Mesotaenium endlicherianum*
 - Cylindrocapsa breibsonii*
 - Cylindrocapsa cushleakeae*
 - Spirogyra* sp.
 - Netrium digitus*
 - Roya obtusa*
 - Penium margaritaceum*
 - Cosmarium ochthodes*
 - Nothoceros vincentianus*
 - Nothoceros aenigmaticus*
 - Metzgeria crassipilis*
 - Bazzania trilobata*
 - Sphaerocarpos texanus*
 - Ricciocarpos natas*
 - Marchantia polymorpha*
 - Marchantia emarginata*
 - Sphagnum lescurei*
 - Polytrichum commune*
 - Physcomitrella patens*
 - Ceratodon purpureus*
 - Hedwigia ciliata*
 - Bryum argenteum*
 - Rosulabryum cf. capillare*
 - Anomodon attenuatus*
 - Leucodon brachypus*
 - Thuidium delicatulum*
 - Rhynchostegium serrulatum*
 - Selaginella moellendorffii*
 - Selaginella stauntoniana*
 - Huperzia squarrosa*
 - Pseudolycopodiella caroliniana*
 - Dendrolycopodium obscurum*
 - Equisetum diffusum*
 - Pteridium aquilinum*
 - Alsophila spinulosa*
 - Angiopteris evecta*
 - Ophioglossum petiolatum*
 - Psilotum nudum*
 - Ginkgo biloba*
 - Zamia vaxezii*
 - Cycas rumphii*
 - Cycas micholitzii*
 - Pinus taeda*
 - Cedrus libani*
 - Ephedra sinica*
 - Gnetum montanum*
 - Welwitschia mirabilis*
 - Prumnopitys andina*
 - Sciadopitys verticillata*
 - Taxus baccata*
 - Juniperus scopulorum*
 - Cunninghamia lanceolata*
 - Amborella trichopoda*
 - Nuphar advena*
 - Kadsura heteroclita*
 - Acorus americanus*
 - Colchicum autumnale*
 - Smilax bona*
 - Yucca filamentosa*
 - Sabal bermudana*
 - Sorghum bicolor*
 - Zea mays*
 - Oryza sativa*
 - Brachypodium distachyon*
 - Monocots**
 - Persea americana*
 - Houttuynia cordata*
 - Saruma henryi*
 - Sarcandra glabra*
 - Eschscholzia californica*
 - Podophyllum peltatum*
 - Aquilegia formosa*
 - Vitis vinifera*
 - Larrea tridentata*
 - Medicago truncatula*
 - Boehmeria nivea*
 - Populus trichocarpa*
 - Hibiscus cannabinus*
 - Arabidopsis thaliana*
 - Carica papaya*
 - Kochia scoparia*
 - Diospyros malabarica*
 - Tanacetum parthenium*
 - Inula helenium*
 - Rosmarinus officinalis*
 - Ipomoea purpurea*
 - Allamanda cathartica*
 - Catharanthus roseus*
 - Eudicots**
- Euphyllophytes (Vascular Plants)**
 - Tracheophytes (Vascular Plants)**
 - Monilophytes**
 - Angiosperms (Flowering Plants)
 - Spermatophytes (Seed Plants)**
 - Gymnosperms**
 - Ephedra sinica*
 - Gnetum montanum*
 - Welwitschia mirabilis*
 - Prumnopitys andina*
 - Sciadopitys verticillata*
 - Taxus baccata*
 - Juniperus scopulorum*
 - Cunninghamia lanceolata*
 - Angiosperms (Flowering Plants)**
 - Spermatophytes (Seed Plants)**
 - Gymnosperms**
 - Ephedra sinica*
 - Gnetum montanum*
 - Welwitschia mirabilis*
 - Prumnopitys andina*
 - Sciadopitys verticillata*
 - Taxus baccata*
 - Juniperus scopulorum*
 - Cunninghamia lanceolata*
 - Angiosperms (Flowering Plants)**
 - Spermatophytes (Seed Plants)**
 - Gymnosperms**
 - Ephedra sinica*
 - Gnetum montanum*
 - Welwitschia mirabilis*
 - Prumnopitys andina*
 - Sciadopitys verticillata*
 - Taxus baccata*
 - Juniperus scopulorum*
 - Cunninghamia lanceolata*
 - Angiosperms (Flowering Plants)**
 - Spermatophytes (Seed Plants)**
 - Gymnosperms**
 - Ephedra sinica*
 - Gnetum montanum*
 - Welwitschia mirabilis*
 - Prumnopitys andina*
 - Sciadopitys verticillata*
 - Taxus baccata*
 - Juniperus scopulorum*
 - Cunninghamia lanceolata*
 - Angiosperms (Flowering Plants)**
 - Spermatophytes (Seed Plants)**
 - Gymnosperms**
 - Ephedra sinica*
 - Gnetum montanum*
 - Welwitschia mirabilis*
 - Prumnopitys andina*
 - Sciadopitys verticillata*
 - Taxus baccata*
 - Juniperus scopulorum*
 - Cunninghamia lanceolata*
 - Angiosperms (Flowering Plants)**
 - Spermatophytes (Seed Plants)**
 - Gymnosperms**
 - Ephedra sinica*
 - Gnetum montanum*
 - Welwitschia mirabilis*
 - Prumnopitys andina*
 - Sciadopitys verticillata*
 - Taxus baccata*
 - Juniperus scopulorum*
 - Cunninghamia lanceolata*
 - Angiosperms (Flowering Plants)**
 - Spermatophytes (Seed Plants)**
 - Gymnosperms**
 - Ephedra sinica*
 - Gnetum montanum*
 - Welwitschia mirabilis*
 - Prumnopitys andina*
 - Sciadopitys verticillata*
 - Taxus baccata*
 - Juniperus scopulorum*
 - Cunninghamia lanceolata*
 - Angiosperms (Flowering Plants)**
 - Spermatophytes (Seed Plants)**
 - Gymnosperms**
 - Ephedra sinica*
 - Gnetum montanum*
 - Welwitschia mirabilis*
 - Prumnopitys andina*
 - Sciadopitys verticillata*
 - Taxus baccata*
 - Juniperus scopulorum*
 - Cunninghamia lanceolata*
 - Angiosperms (Flowering Plants)**
 - Spermatophytes (Seed Plants)**
 - Gymnosperms**
 - Ephedra sinica*
 - Gnetum montanum*
 - Welwitschia mirabilis*
 - Prumnopitys andina*
 - Sciadopitys verticillata*
 - Taxus baccata*
 - Juniperus scopulorum*
 - Cunninghamia lanceolata*
 - Angiosperms (Flowering Plants)**
 - Spermatophytes (Seed Plants)**
 - Gymnosperms**
 - Ephedra sinica*
 - Gnetum montanum*
 - Welwitschia mirabilis*
 - Prumnopitys andina*
 - Sciadopitys verticillata*
 - Taxus baccata*
 - Juniperus scopulorum*
 - Cunninghamia lanceolata*
 - Angiosperms (Flowering Plants)**
 - Spermatophytes (Seed Plants)**
 - Gymnosperms**
 - Ephedra sinica*
 - Gnetum montanum*
 - Welwitschia mirabilis*
 - Prumnopitys andina*
 - Sciadopitys verticillata*
 - Taxus baccata*
 - Juniperus scopulorum*
 - Cunninghamia lanceolata*
 - Angiosperms (Flowering Plants)**
 - Spermatophytes (Seed Plants)**
 - Gymnosperms**
 - Ephedra sinica*
 - Gnetum montanum*
 - Welwitschia mirabilis*
 - Prumnopitys andina*
 - Sciadopitys verticillata*
 - Taxus baccata*
 - Juniperus scopulorum*
 - Cunninghamia lanceolata*
 - Angiosperms (Flowering Plants)**
 - Spermatophytes (Seed Plants)**
 - Gymnosperms**
 - Ephedra sinica*
 - Gnetum montanum*
 - Welwitschia mirabilis*
 - Prumnopitys andina*
 - Sciadopitys verticillata*
 - Taxus baccata*
 - Juniperus scopulorum*
 - Cunninghamia lanceolata*
 - Angiosperms (Flowering Plants)**
 - Spermatophytes (Seed Plants)**
 - Gymnosperms**
 - Ephedra sinica*
 - Gnetum montanum*
 - Welwitschia mirabilis*
 - Prumnopitys andina*
 - Sciadopitys verticillata*
 - Taxus baccata*
 - Juniperus scopulorum*
 - Cunninghamia lanceolata*
 - Angiosperms (Flowering Plants)**
 - Spermatophytes (Seed Plants)**
 - Gymnosperms**
 - Ephedra sinica*
 - G**

PAFTOL HYBSEQ PROJECT: COVERING ALL ANGIOSPERMS

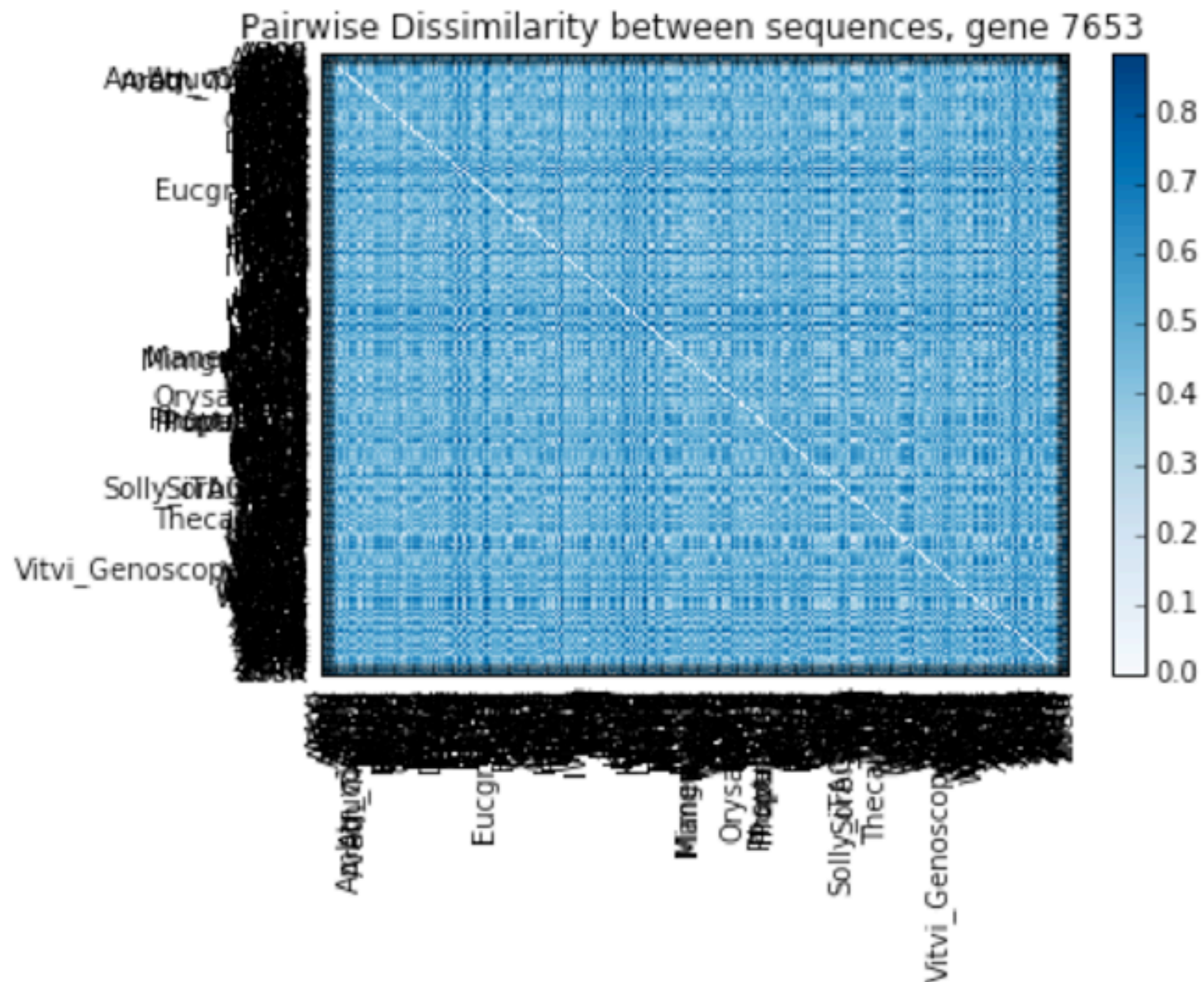
Data: 400 loci used for phylogenetics

Orthology already determined (using genomes)

Pre-screened for low-copy in Viridiplantae

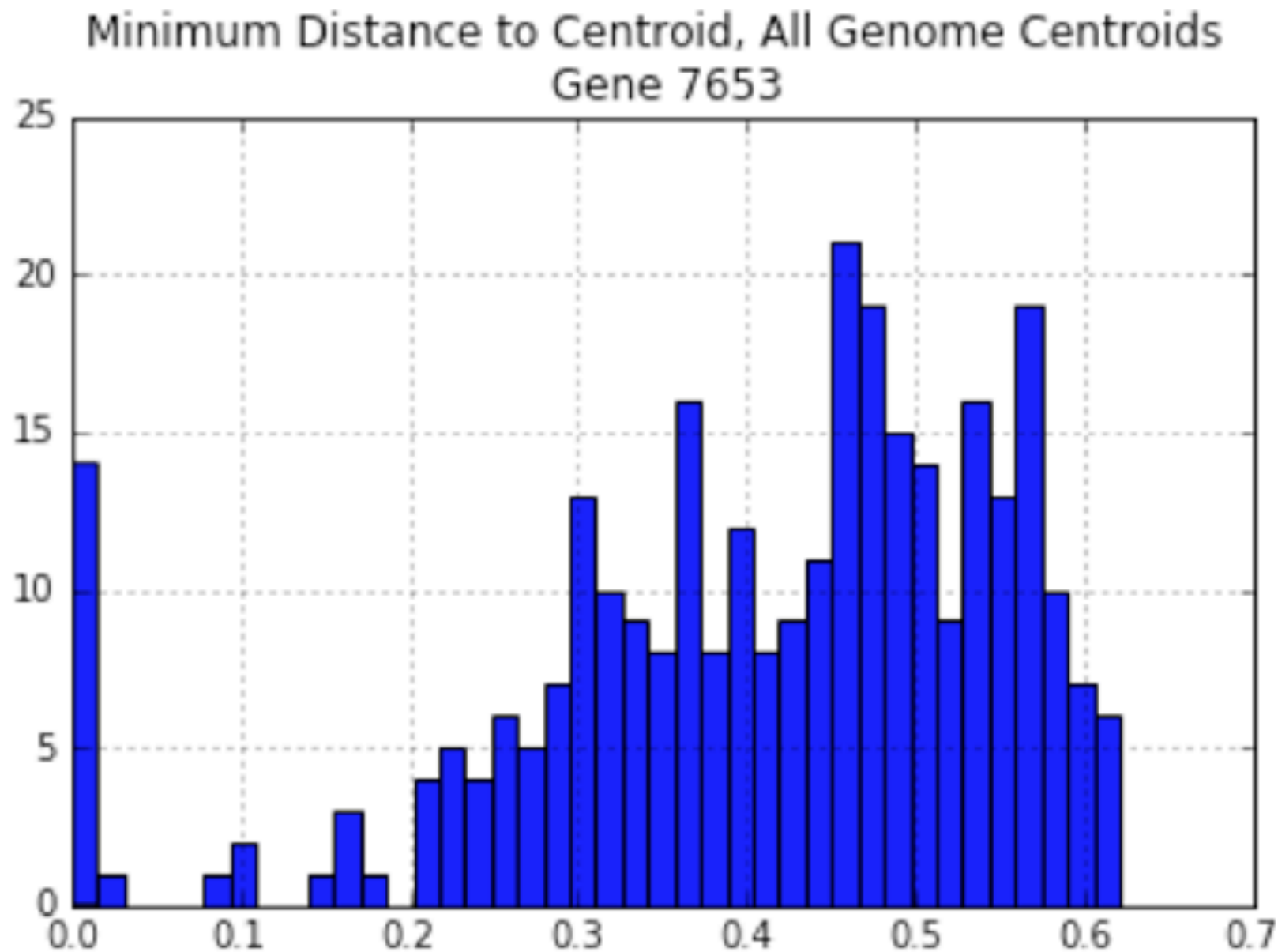
Too many sequences to start (600+ Angiosperms)

SELECTING SEQUENCES: MACHINE LEARNING



SELECTING SEQUENCES: MACHINE LEARNING

What if we only select from genomes?

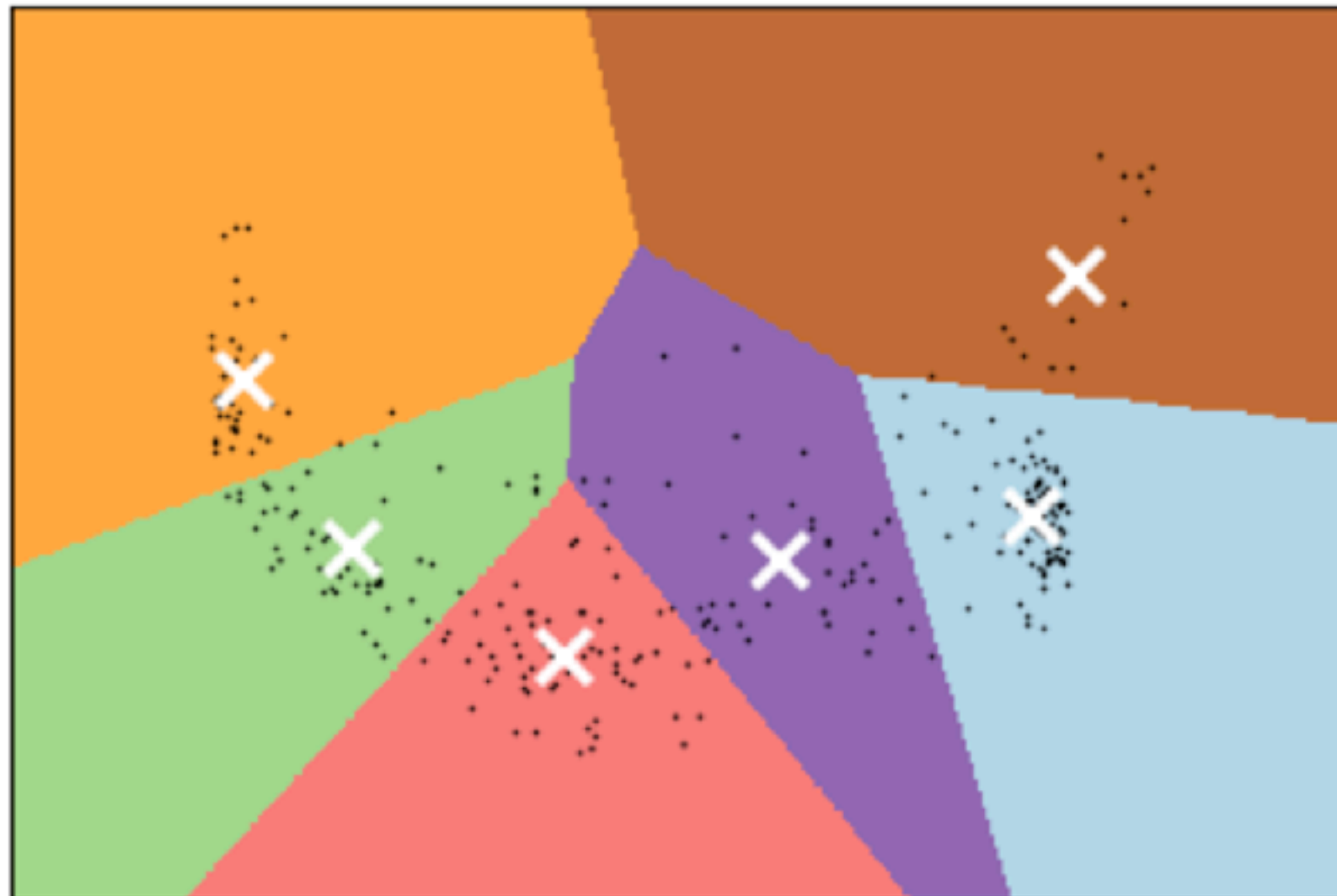


Too far for HybSeq!

SELECTING SEQUENCES: MACHINE LEARNING

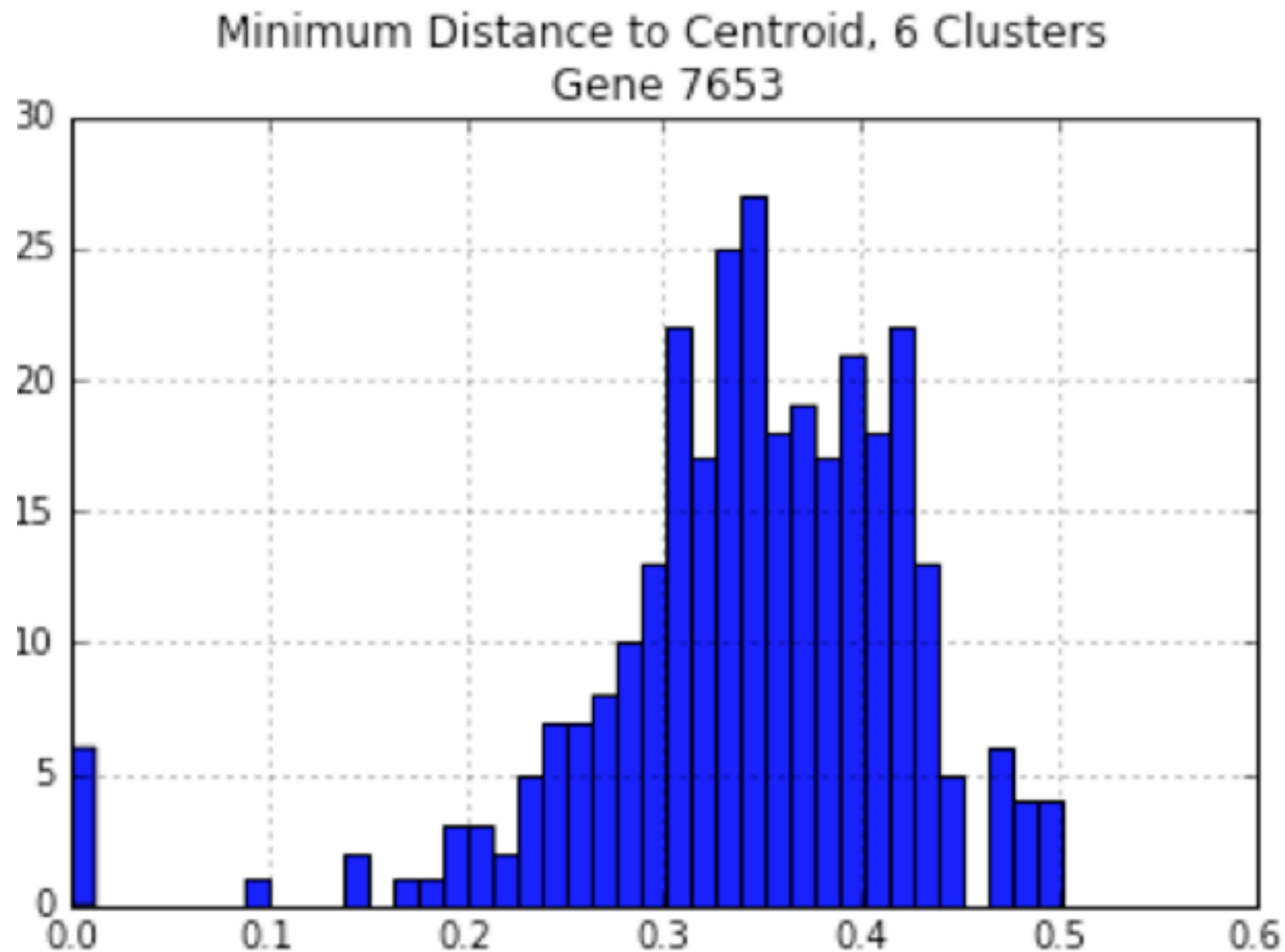
Reducing the dimensionality with K-means clustering

K-means clustering on the DNA sequence dataset
(PCA-reduced distance matrix)
Centroids are marked with white cross



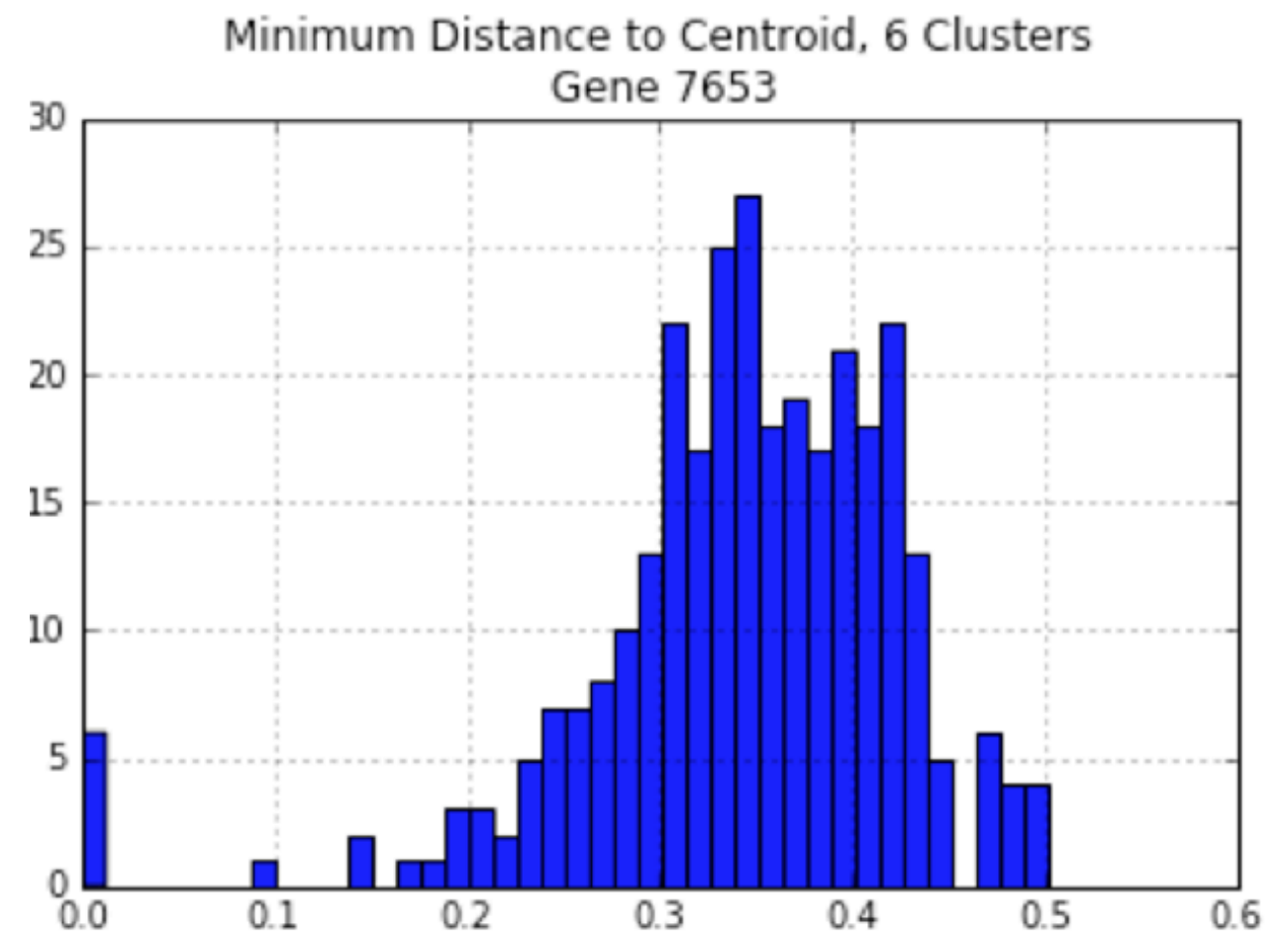
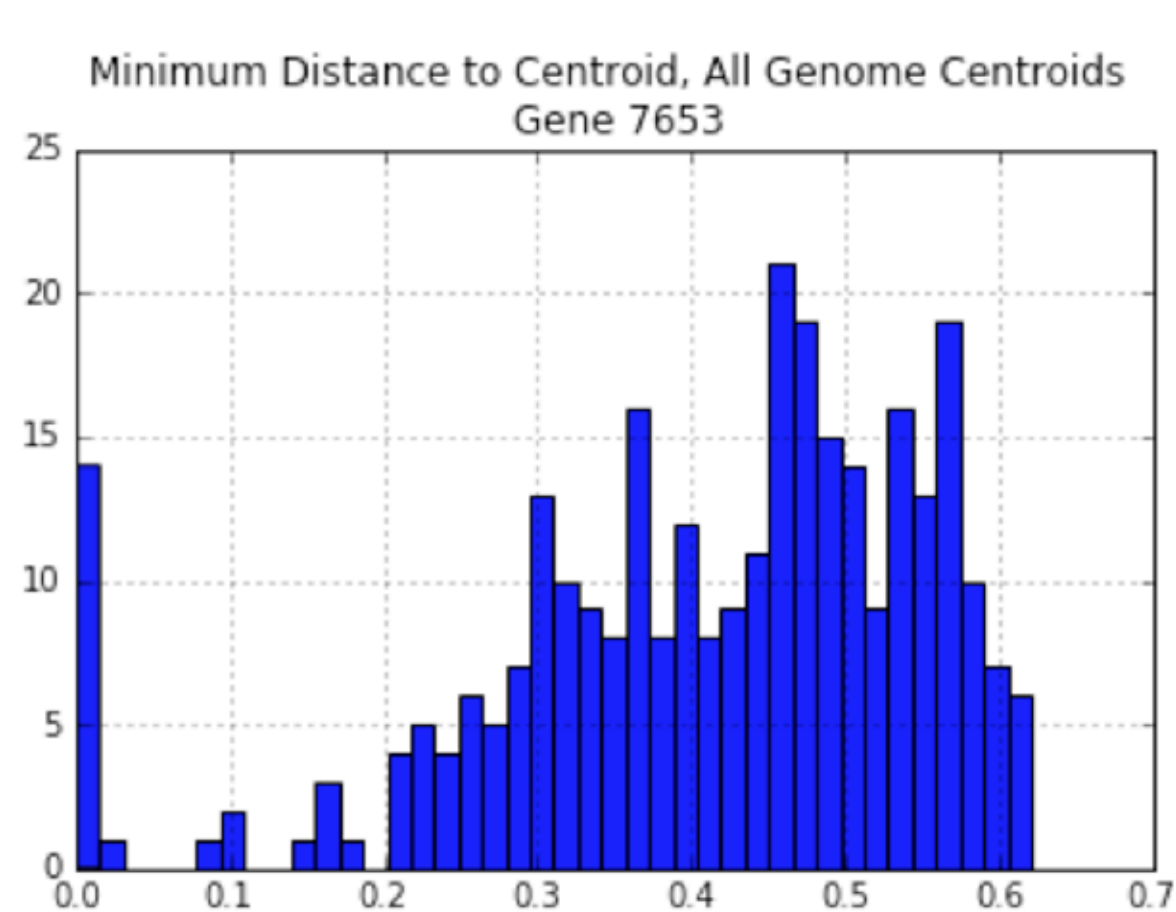
SELECTING SEQUENCES: MACHINE LEARNING

Reducing the dimensionality with K-means clustering

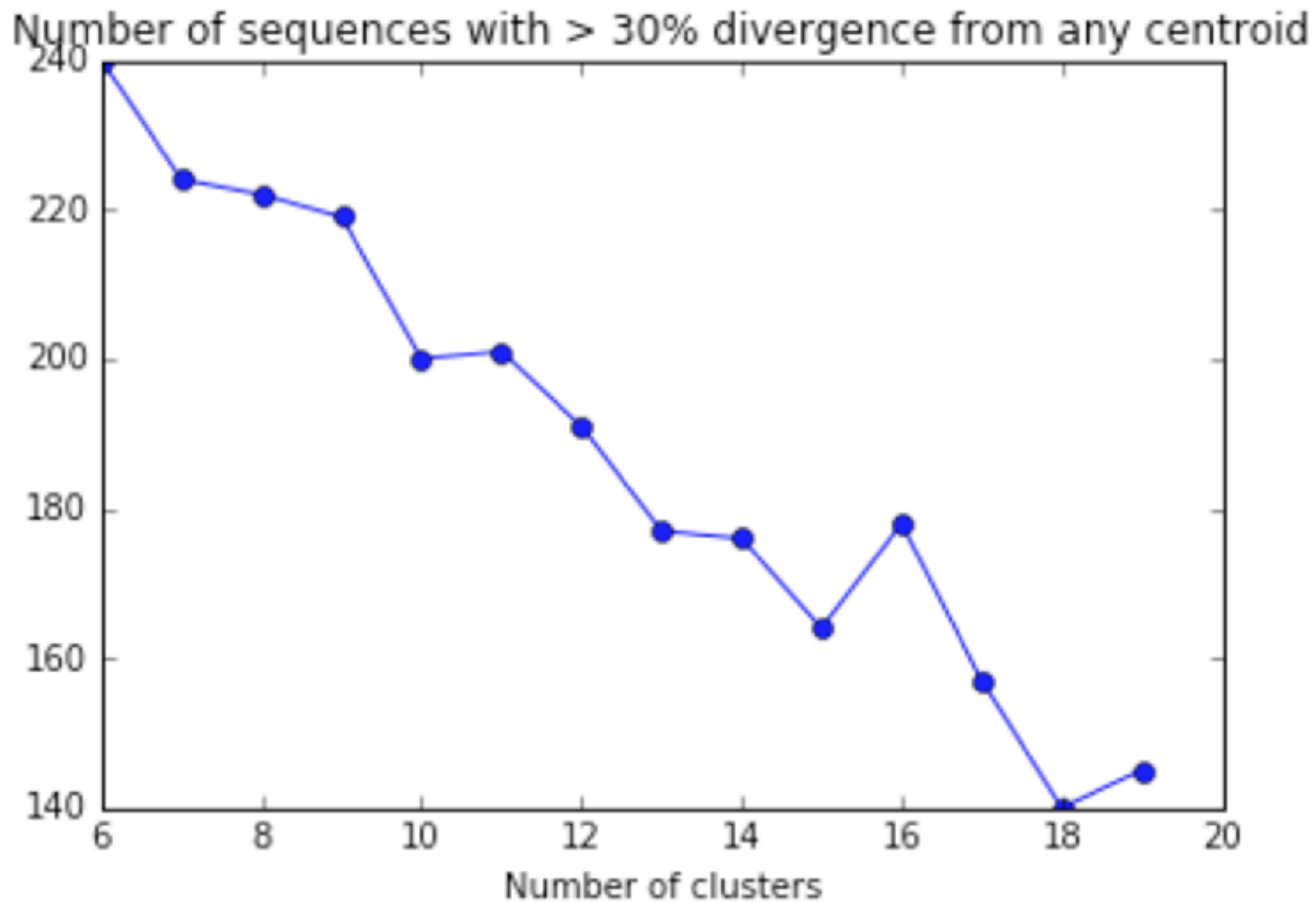


SELECTING SEQUENCES: MACHINE LEARNING

Reducing the dimensionality with K-means clustering



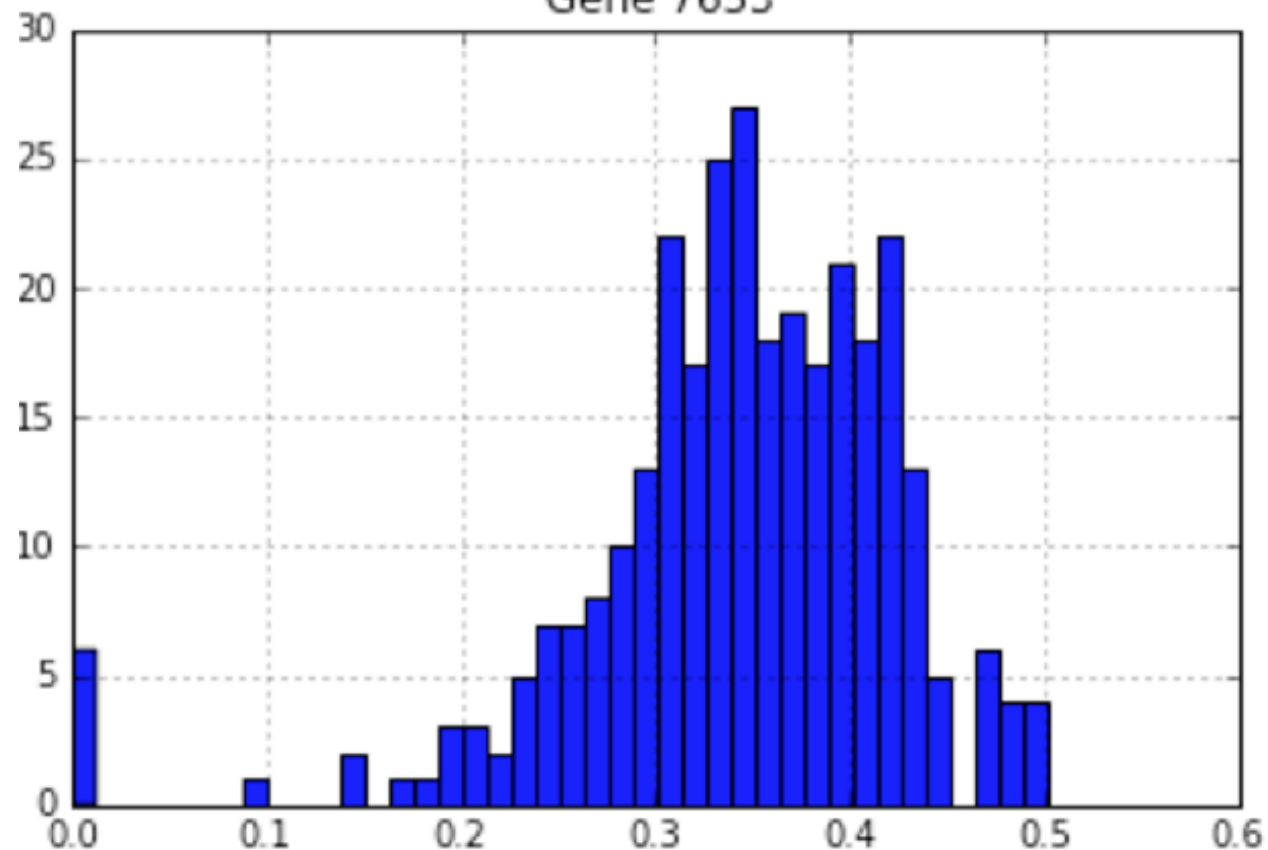
SELECTING SEQUENCES: MACHINE LEARNING



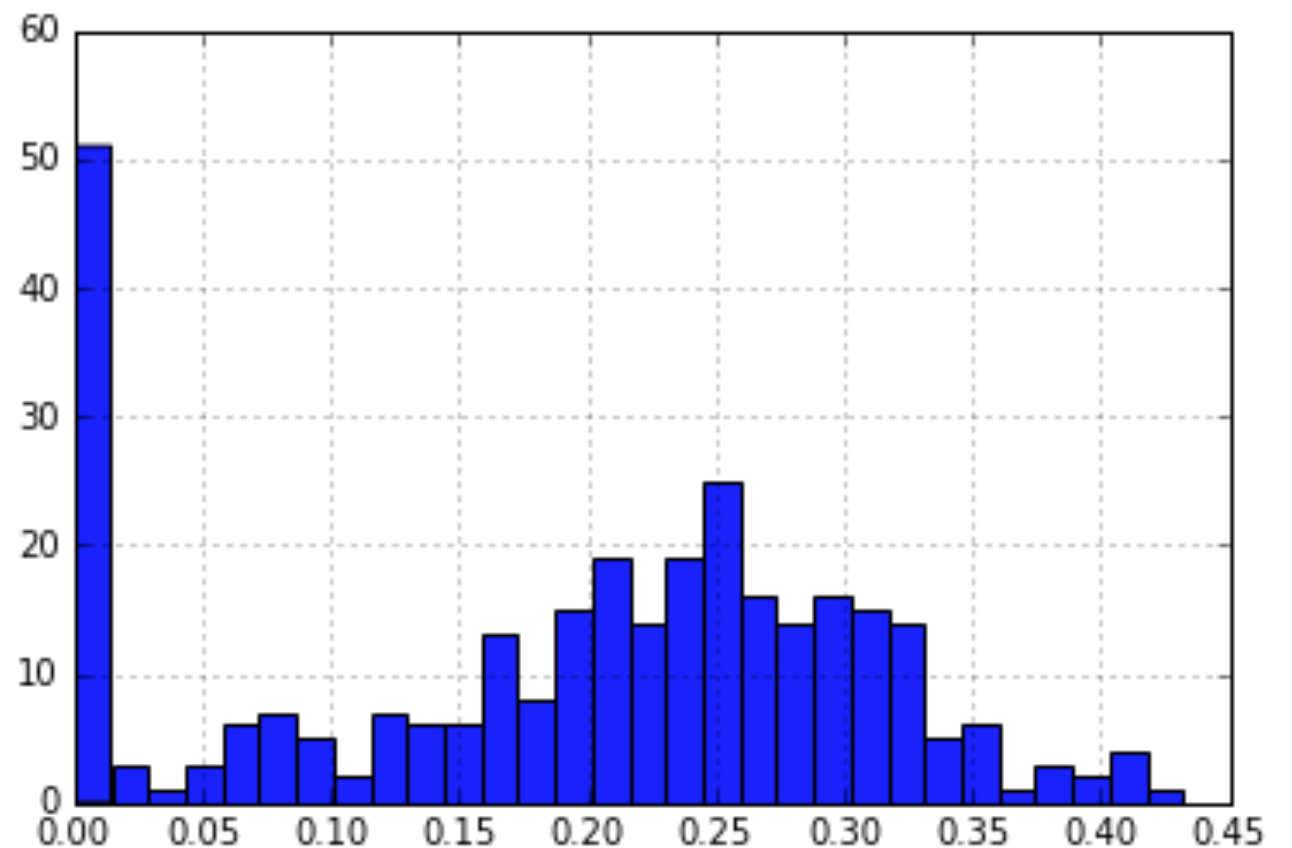
603 total angiosperm sequences

SELECTING SEQUENCES: MACHINE LEARNING

Minimum Distance to Centroid, 6 Clusters
Gene 7653



10 K-medoids



603 total angiosperm sequences

PAFTOL ANGIOSPERM PROBE DESIGN

Using k-medoids method, 354 loci selected

Between 6 and 15 medoids represent $> 95\%$ of all 1KP angiosperms

80,000 probes, designed by MycroArray (Michigan, USA)

Pilot project: Sequencing 288 Angiosperms

Stay tuned...

HANDS ON EXERCISE



MarkerMiner
Locus development made easy

Chamala et al., APPS, 2015

Align transcriptomes to existing genomic resources

Select single-copy loci

Generate alignments ready to submit for probe design

Command line or web interface

IMPORTANT WEBSITES

atmo.cyverse.org

github.com/mossmatters/KewHybSeqWorkshop

mossmatters.com