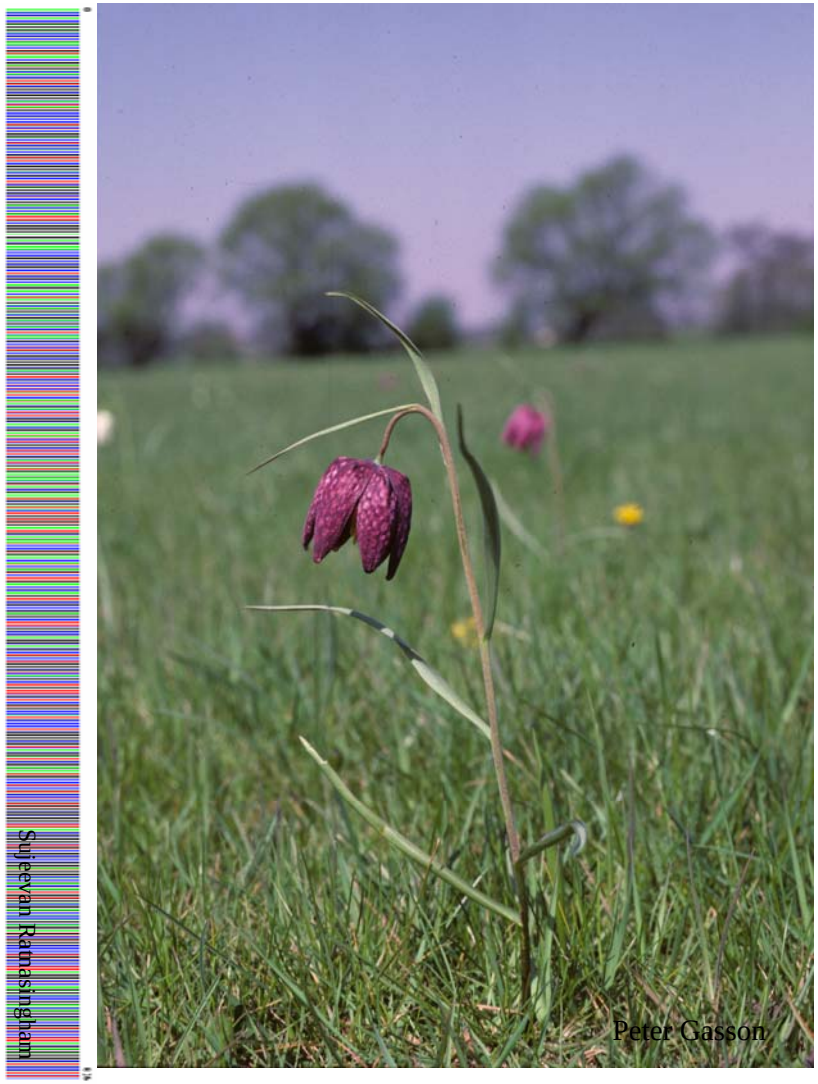
Barcoding land plants: Which loci and how many?

Kew

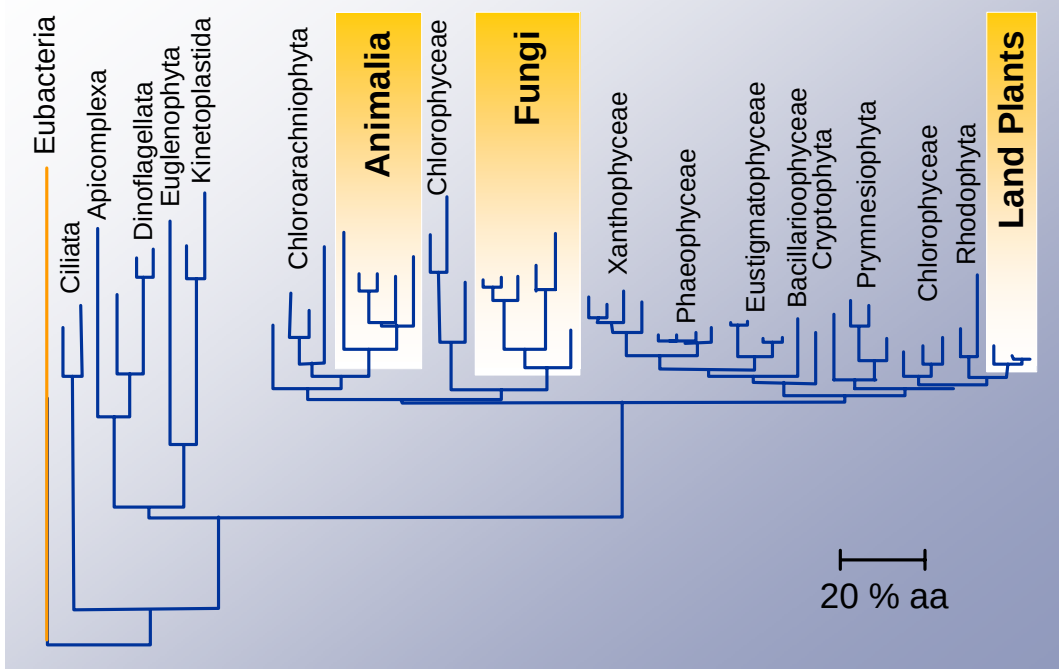
PLANTS PEOPLE POSSIBILITIES



So what are the characteristics of a good barcode?

- High inter-specific, low intra-specific sequence divergence
- Universal amplification/sequencing with standard primers
- Technically simple to sequence
- Short enough to sequence in one reaction
- Easily alignable (few insertions/deletions)
- Readily recoverable from museum or herbarium samples and other degraded samples

COI Divergence in Eukaryotes



Sloan/Moore Consortium

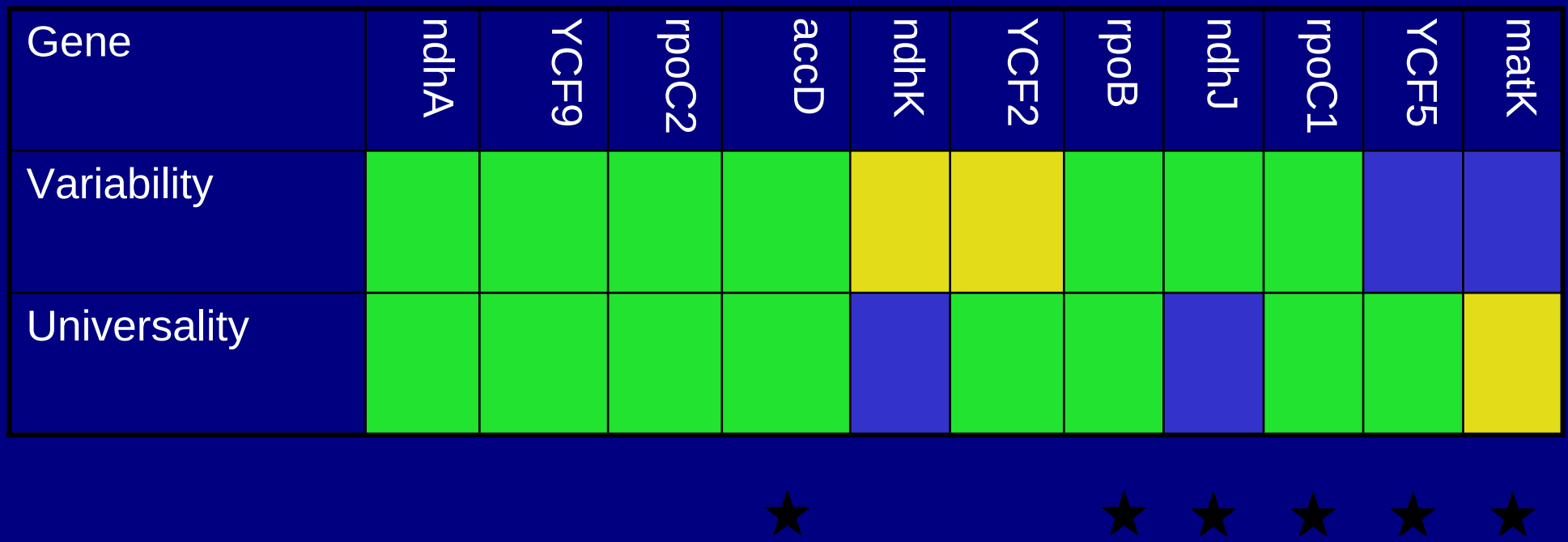
RBG Kew, RBG Edinburgh, NHM (London) NHM (Copenhagen), Imperial College, New York, BG, Mexico, Colombia, Brazil, U Capetown, SANBI

Sloan/Moore Project – to develop a universal approach to barcoding of all landplants (450 my)

- 18 months in two phases
- Phase 1: primer development (protein motifs); complete genome sequences; problems: ferns; 96 pairs of sister taxa from mosses, liverworts, hornworts, lycopods, ferns/fern allies, gymnosperms, angiosperms – percent PCR success & percent polymorphisms
- Phase 2: trials of eight markers identified in phase I, plus *psbA-trnH* (Kress eta, 2005).
- What about *rbcL*, *trnL*, and nrITS?

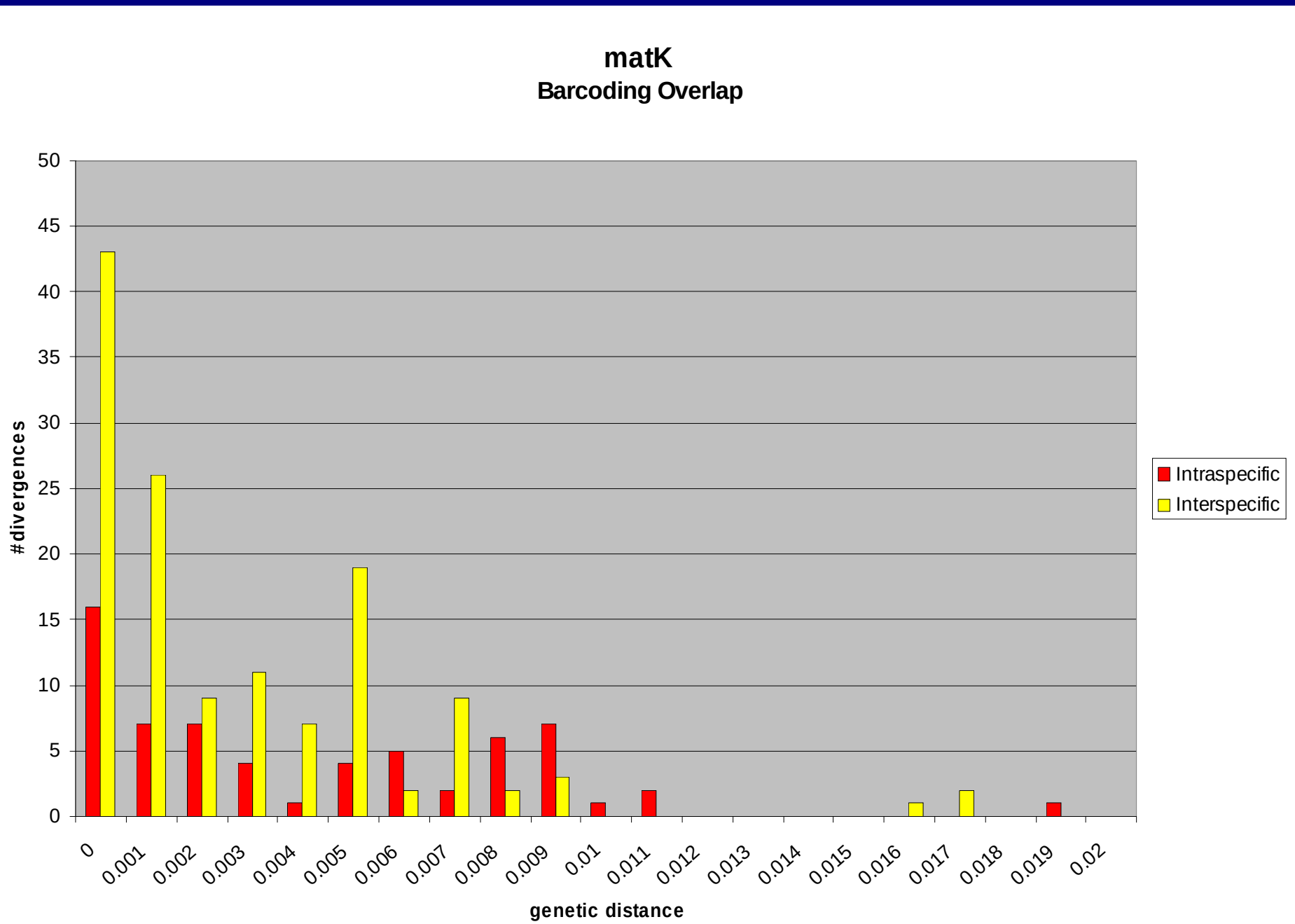
Trial regions

Selected seven genes that represent the different levels of universality and variability. Blue= high, green = medium, yellow= low.



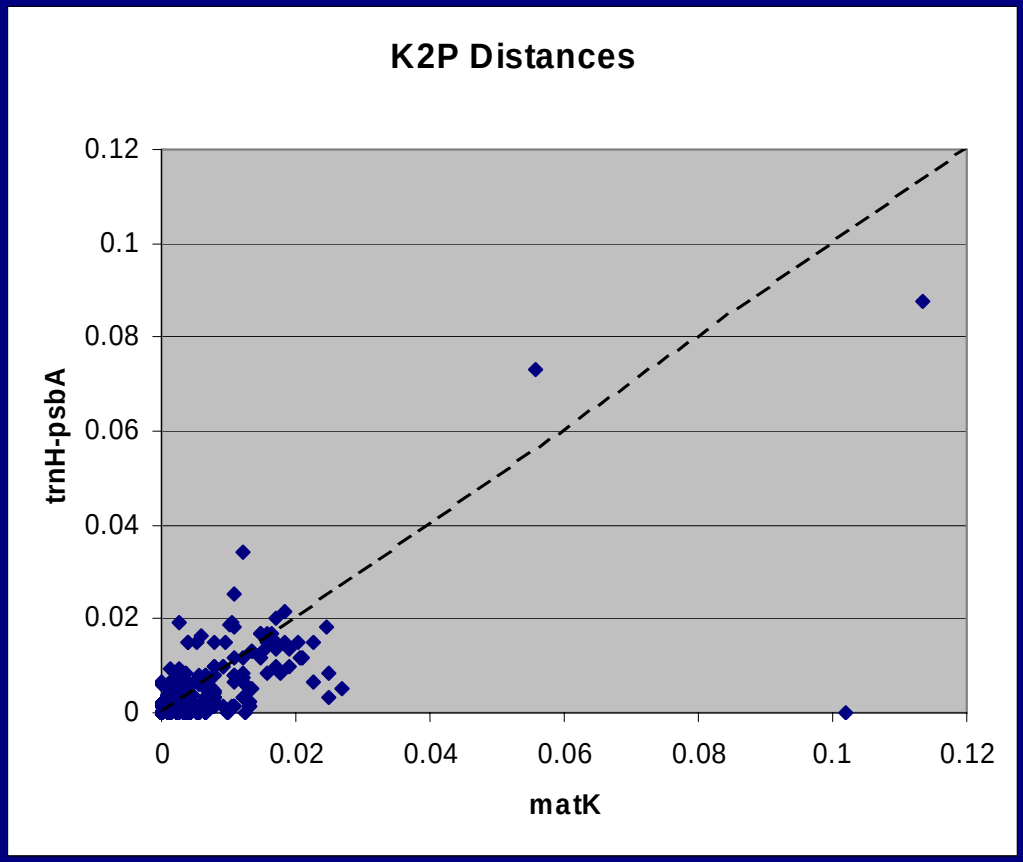
Sampled groups by region

	accD		ITS		matK		ndhJ		rpoB		rpoC1		trnH-psbA	
	Ind's	bp's	Ind's	bp's	Ind's	bp's	Ind's	bp's	Ind's	bp's	Ind's	bp's	Ind's	bp's
Agave	25	302	25	653	22	913	25	434	25	508	23	624	25	576
Anastrophyllum											57	555		
Araucaria											106	523		
Asterella					21	759								
Aulosepalum	13	302	13	695	13	924	13	435	13	515	13	622	11	836
Conostylis	64	300			61	942	64	432	62	441	64	619	3	481
Costa R Orchids	74	210			72	824	65	409	73	428	74	531	74	1191 aligned
Crocus	126	367									126	575		
Cupressus	34	294					20	429			32	467		
Dactylorhiza	22	300			22	851	22	432	22	545	22	619	3	816
Elaphoglossum	55	417									23	790		
Equisetum	22	294			22	800	22	372	22	821	22	808	3	250
Hordeum					52	797					52	575		
Inga											107	472		
Labordia	24	294			17	752	29	379	17	381	27	561		
Lauraceae	47	419			58	974	42	428	48	562	50	603		
Liverworts											77	471		
Pinus	85	300			77	890			85	863	84	619	3	606
Ptychomniales	56	303									53	621		


matK
Barcoding Overlap
#divergences
50
45
40
35
30
25
20
15
10
5
0
0
0.001
0.002
0.003
0.004
0.005
0.006
0.007
0.008
0.009
0.01
0.011
0.012
0.013
0.014
0.015
0.016
0.017
0.018
0.019
0.02
genetic distance
Intraspecific
Interspecific

Coding *versus* non-coding plastid DNA

- How much more variable is non-coding DNA?
- Length variation: *psbA-trnH*, 200-1200 bp
- Alignment
- polyA/T (plastid microsatellites)
- Molecular evolution: *psbA-trnH* contains *rpl22/rps19* in most monocots including orchids and grasses


K2P Distances
0.12
0.1
0.08
0.06
0.04
0.02
0
trnH-psbA
0
0.02
0.04
0.06
0.08
0.1
0.12
matK

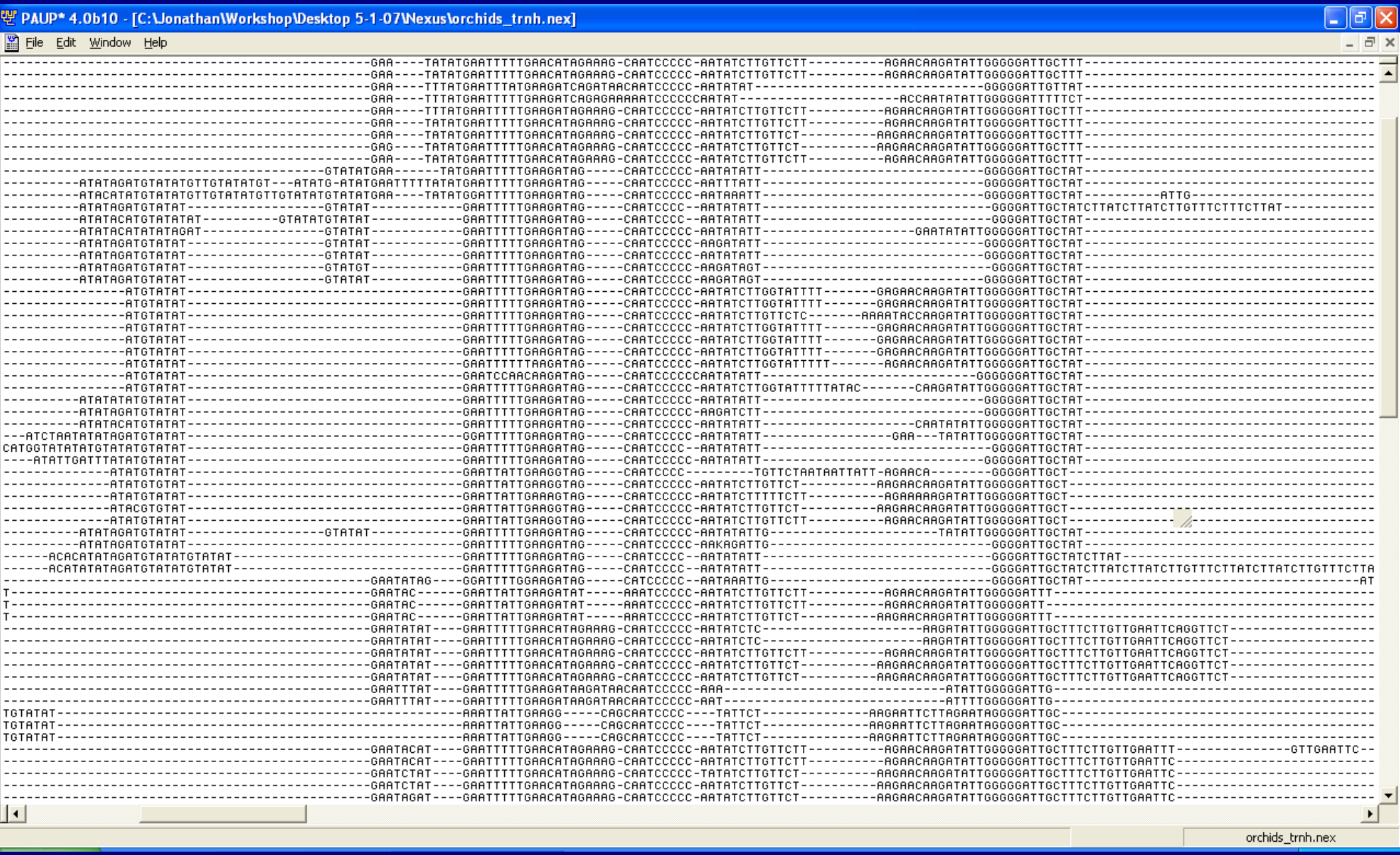
Coding *versus* non-coding plastid DNA

- How much more variable is non-coding DNA?
- Length variation: *psbA-trnH*, 200-1200 bp
- Alignment
- polyA/T (plastid microsatellites)
- Molecular evolution: *psbA-trnH* contains *rpl22/rps19* in most monocots including orchids and grasses

trnH-psbA

- Variation in sequence length
  - *Agave* – 576 bp
  - *Aulosepalum* – 836 bp
  - *Conostylis* – 481 bp
  - Orchids – 1191 bp
  - *Equisetum* – 250 bp
  - *Pinus* – 606 bp

psbA-trnH


PAUP* 4.0b10 - [C:\Jonathan\Workshop\Desktop 5-1-07\Nexus\orchids_trnh.nex]
File Edit Window Help
orchids_trnh.nex

Coding *versus* non-coding plastid DNA

- How much more variable is non-coding DNA?
- Length variation: *psbA-trnH*, 200-1200 bp
- Alignment
- polyA/T (plastid microsatellites)
- Molecular evolution: *psbA-trnH* contains *rpl22/rps19* in most monocots including orchids and grasses

trnH-psbA

	Aulosepalum	CR Orchids
species	7	50
monophyletic	3	37
singletons	0	9
%mono	**42.86%**	**90.24%**

...

matK

	Aulosepalum	Conostylis	Dactylorhiza	Equisetum	Hordeum
species	8	44	18	15	32
monophyletic	3	35	16	8	24
singletons	3	3	1	3	3
%mono	**60.00%**	**85.37%**	**94.12%**	**66.67%**	**82.76%**

	Labordia	Lauraceae	CR Orchids	Pinus	Combined
species	14	40	48	67	**117**
monophyletic	11	28	36	57	**86**
singletons	0	5	10	4	**13**
%mono	**78.57%**	**80.00%**	**94.74%**	**90.48%**	**82.69%**

Two options

- Three coding regions: *rpoC1*, *rpoB*, *matK* (*rbcL*?)
- Two coding, one non-coding: *rpoC1*, *matK* , *psbA-trnH* (*rbcL*?)