Resolution of fine-scale ribosomal DNA variation in *Saccharomyces* yeast

Rob Davey

NCYC

2009

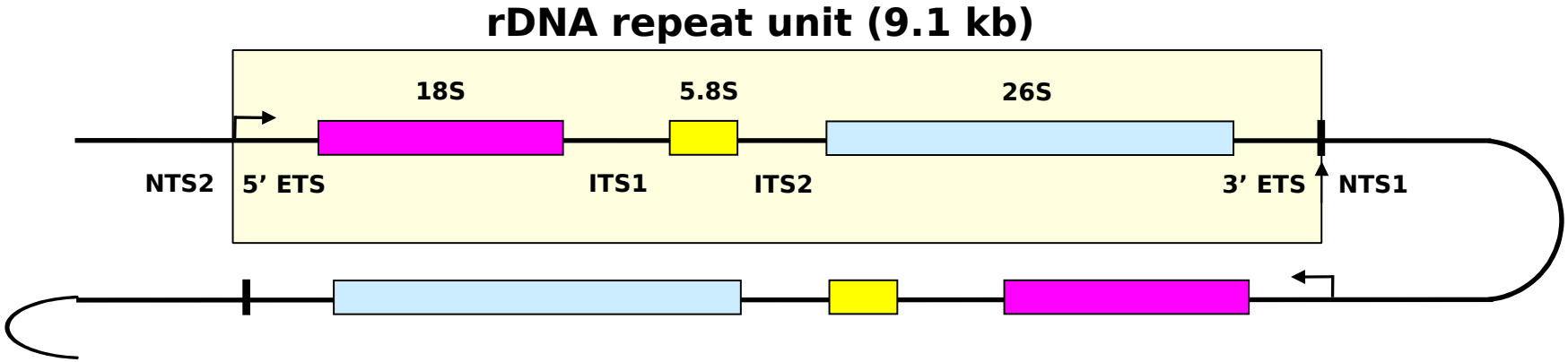
Introduction

- SGRP project
- Ribosomal DNA and variation
- Computational methods
- Preliminary Results
- Conclusions

SGRP

- *Saccharomyces* Genome Re-sequencing Project
- Ed Louis, Nottingham and Richard Durbin, Sanger
- Whole genome shotgun sequence (WGSS) for
  - 34 haploid *S. cerevisiae*
  - 36 *S. paradoxus*
  - 1-3x coverage (>1,000 Mb)
  - So approx 35Mb for *S. cerevisiae* alone

Ribosomal DNA

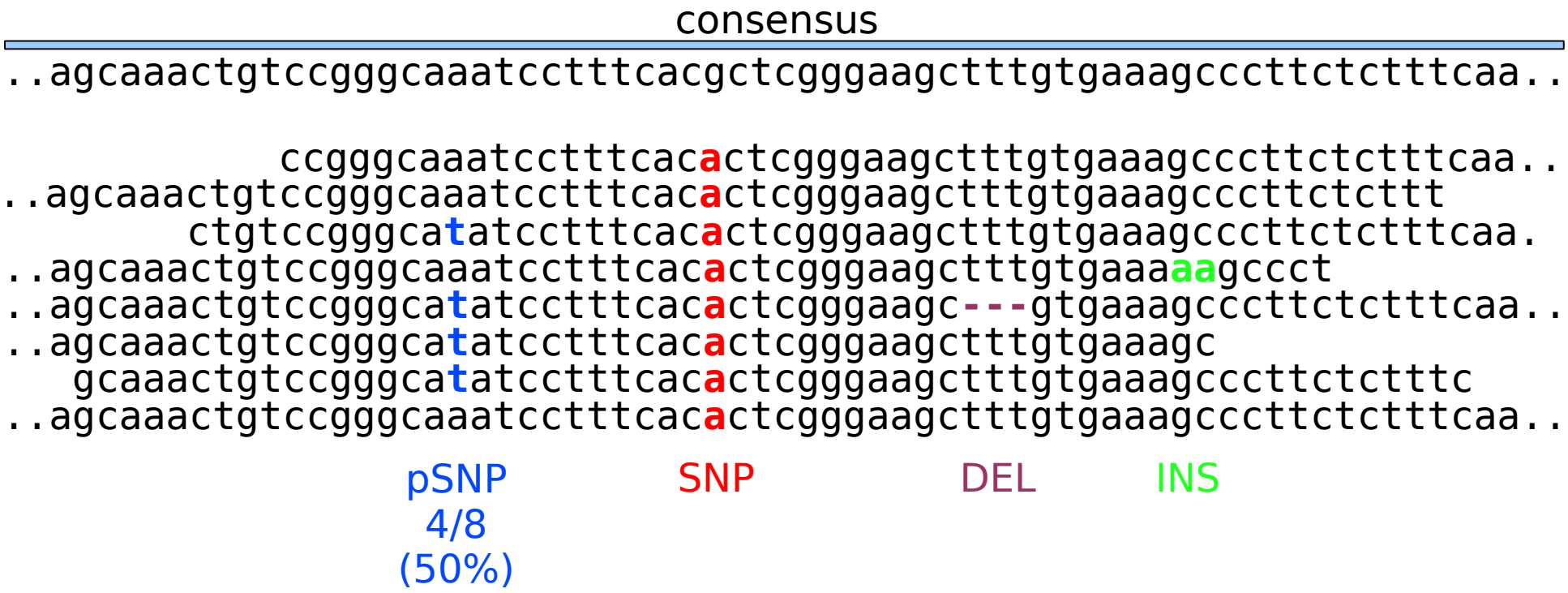


- rDNA provides 'roadmap' of species diversity (26S)
- Drill down to fine-scale sub-species diversity (ITS)
- Tandem array of 100-200 copies on Chromosome XII (~60%)
- Approx 1-2 Mb per tandem array
- YGD lists two identical copies (left- and rightmost copies)
- All other copies assumed identical (evolutionary theory predicts rapid homogenisation by gene conversion)
- SGRP dataset enables us to test this prediction

TURNIP

- WGSS produces reads (assumed to cover whole genome equally) with associated quality per base (FASTQ format)
- Cannot assemble rDNA repeats due to high similarity of region in tandem array (*Ganley 2007*)
- Single repeat consensus alignment for each strain
- Need a way of computing:
  - reads that align to the rDNA repeat consensus
  - reads that are of sufficient sequence quality to be accurate
  - quantifiable differences between consensus and read
    - SNPs = 100% read variance compared to consensus
    - pSNPs = 'partial SNPs' 0% < x < 100% read variance
- TURNIP (Tracking Unresolved rDNA Nucleotide Polymorphisms)

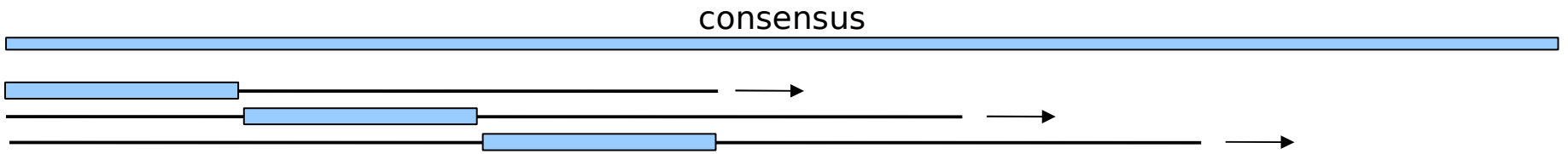
TURNIP



- Assume that there is an **equal probability** that a read sequence is obtained from any of the repeat units
- Quantifiable microheterogeneity would provide a phylogenetic signal for comparative genomics and test for mathematical models of gene conversion

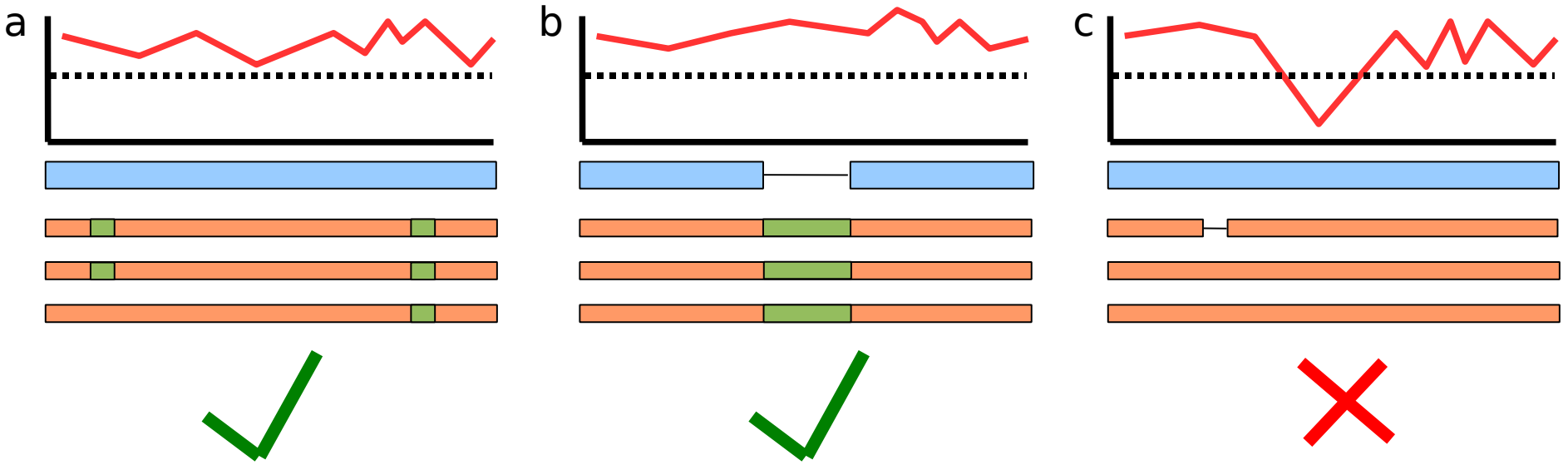
TURNIP

- Take 20bp slices of S288c (*Goffeau 1996*) 'query' consensus
- Anchored on each side by 40bp flanking sequence to give a more accurate alignment



- 'sliding window' of 100bp segments
- Gapped BLAST against FASTA database of shotgun reads
- For each hit above threshold, take highest scoring pair (HSP)
- Store template consensus query sequence and each *distinct* HSP subject sequence at each sequential window position for alignment
- Run multi-alignment (MUSCLE) on subject sequence dataset against template segment

TURNIP



- For each 20bp slice, check quality for each associated read
  - Span introduced gaps with surrounding quality scores
  - Ensure all 20 bases have quality score > threshold
- For each accepted 20bp slice, check for SNPs/pSNPs (fig a), insertions (fig b) or deletions (fig c)
  - Keep those seen in >= 2 reads
- Variation less likely to be sequencing error

TURNIP

- At each position, record the query letter(s), subject letter(s), quality and read name ('ascriptions')
- Compare each position to the original consensus

- Parallelised for use on NBI cluster (Parallel::ForkManager)
- Allows MUSCLE processes and ascription generation to occur in limited memory space

TURNIP

- Outputs

```
3640: t (32) -> a (1) pSNP [ X ]
4810: a (0)  -> g (41) SNP
5680: c (13) -> - (27) DEL
6700: ----- (3) -> actgg (42) INS
```

  - Raw text, Excel, SQL, GFF
  - Use GFF3 to import data into GBrowse
  - Use SQL for storage and preliminary phylogenetic analysis

Results

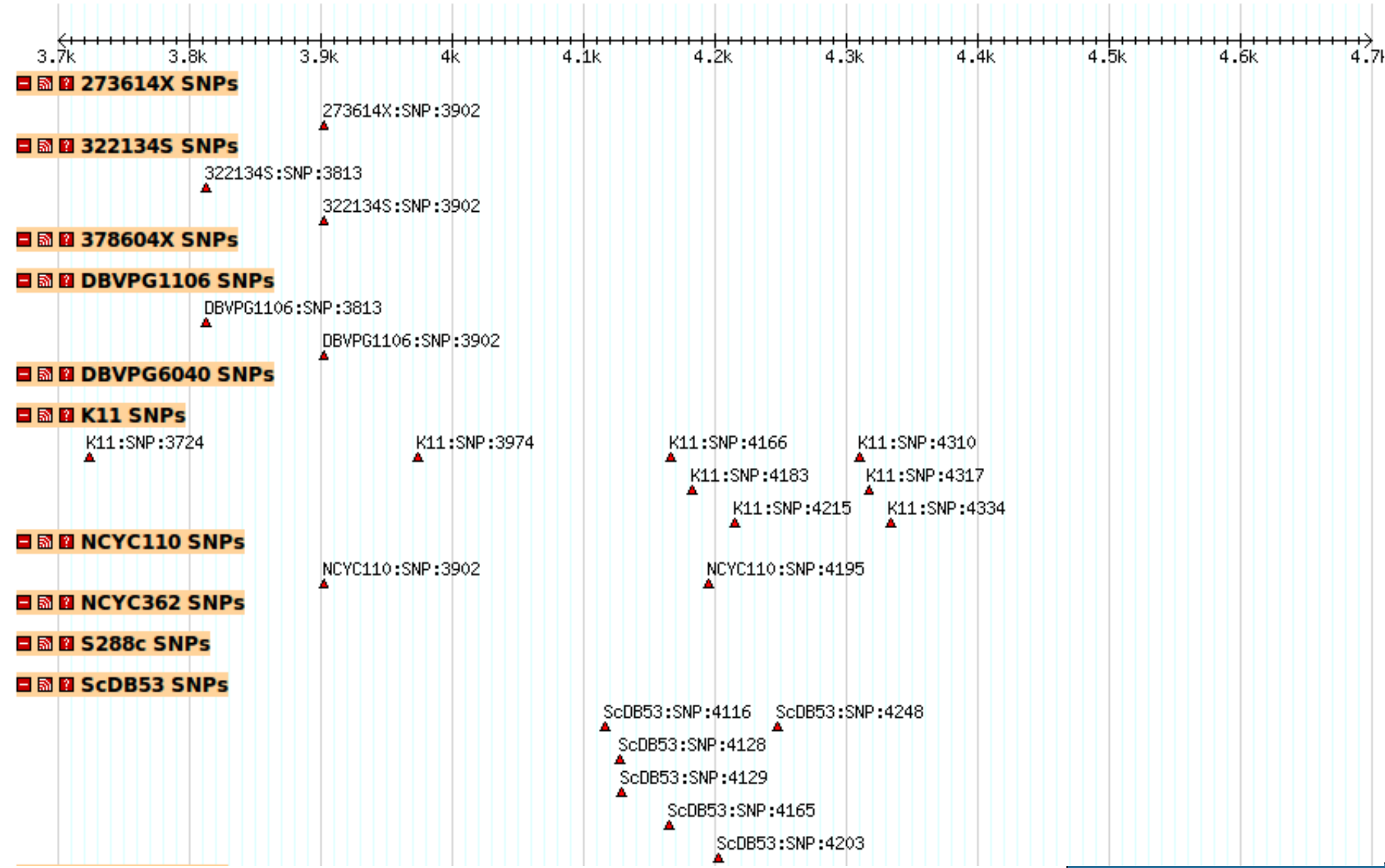
Table 1. *S. cerevisiae* rDNA array polymorphism totals

Strain[a]	Type[b]	Polymorphism total[c]
273614N	Clinical	43
322134S	Clinical	33
378604X	Clinical	40
BC187	Fermentation	17
DBVPG 1106	Fermentation	16
DBVPG 1373	Wild	16
DBVPG 1788	**Wild**	**10**
DBVPG 1853	Fermentation	40
DBVPG 6040	Spoilage	44
DBVPG 6044	Fermentation	29
DBVPG 6765	Pro-biotic	27
K11	Fermentation	31
L_1374	Fermentation	18
NCYC 110	Fermentation	23
NCYC 361	Spoilage	31
S288c	Laboratory	23
SK1	Laboratory	35
UWOPS03-461-4	Wild	33
UWOPS05-217-3	Wild	36
UWOPS05-227-2	Wild	33
UWOPS83-787-3	**Wild**	**76**
UWOPS87-2421	Wild	21
W303	Laboratory	23
Y12	Fermentation	27
Y55	Laboratory	40
Y9	Fermentation	29
YIIc17_E5	Fermentation	39
YJM975	Clinical	12
YJM978	Clinical	17
YJM981	Clinical	12
YPS128	Wild	16
YPS606	Wild	21
YS4	Baking	41
YS9	Baking	38

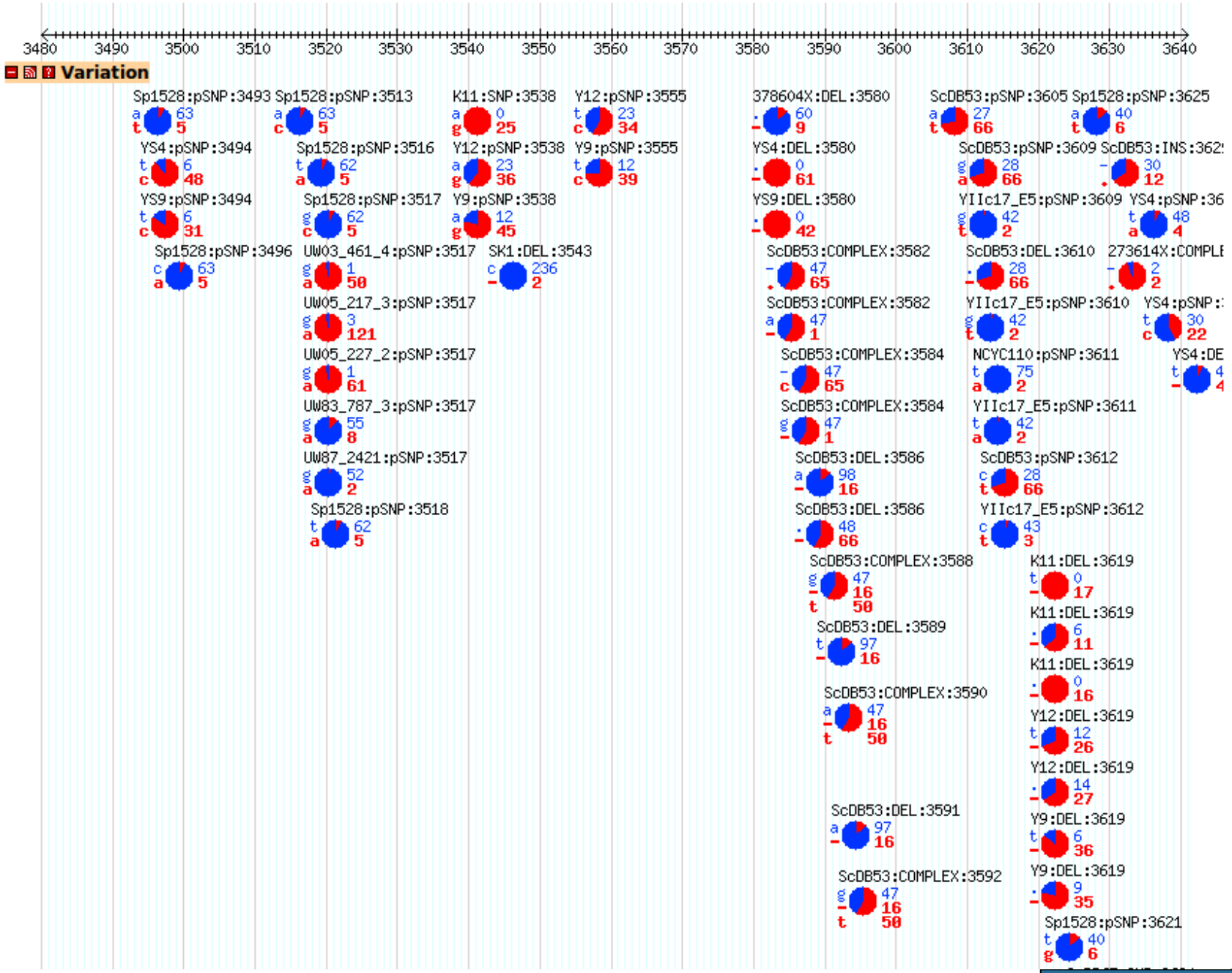
- Base substitution only
- Contrasts greatly with previous studies (*Ganley, 2007*)
- Discovered 4 polymorphic sites
- Could be explained by use of different strain (RM11-1a)

James SA. et al, Gen Res. 2009

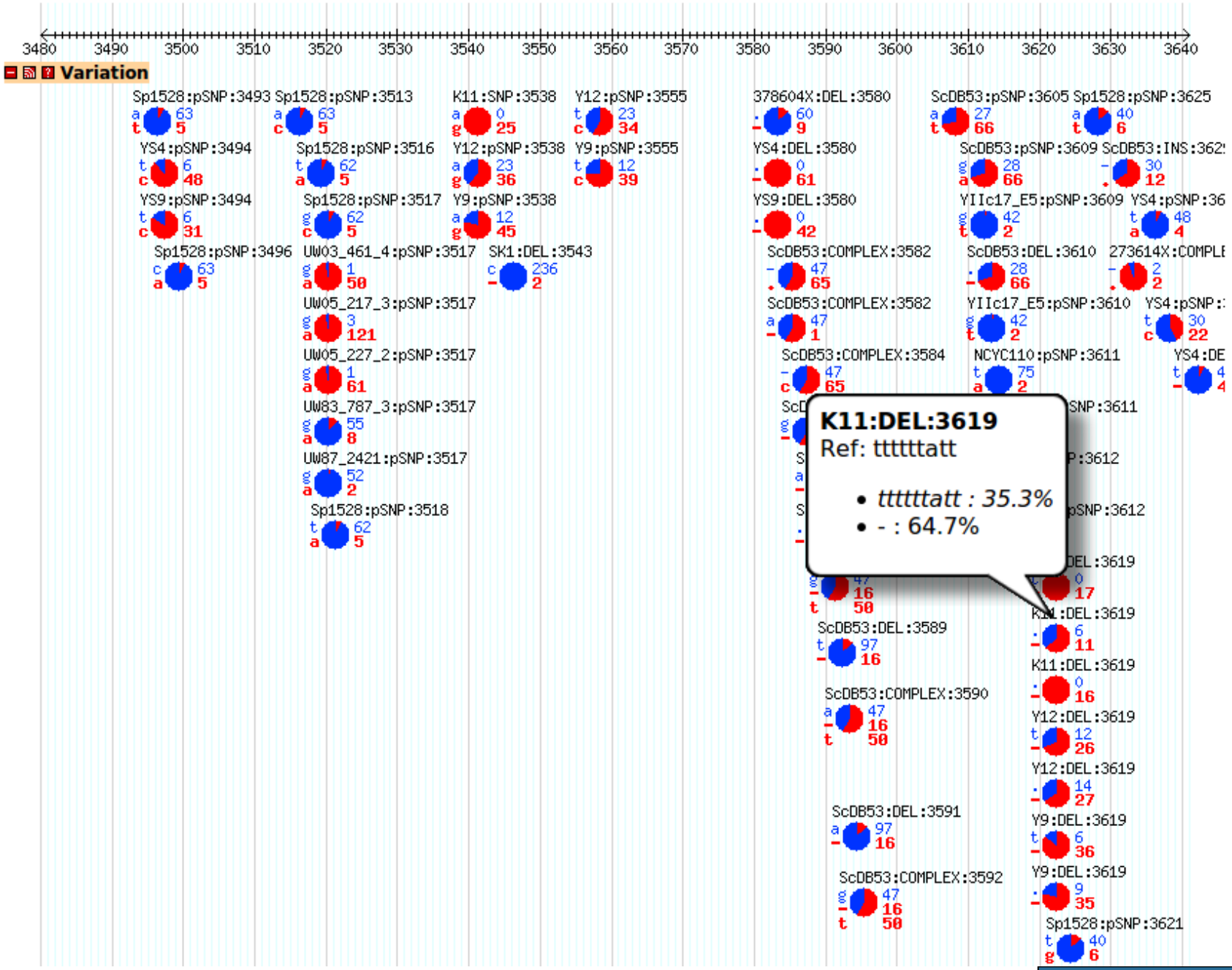
GBrowse


3.7k
3.8k
3.9k
4k
4.1k
4.2k
4.3k
4.4k
4.5k
4.6k
4.7k
273614X SNPs
273614X:SNP:3902
322134S SNPs
322134S:SNP:3813
322134S:SNP:3902
378604X SNPs
DBVPG1106 SNPs
DBVPG1106:SNP:3813
DBVPG1106:SNP:3902
DBVPG6040 SNPs
K11 SNPs
K11:SNP:3724
K11:SNP:3974
K11:SNP:4166
K11:SNP:4310
K11:SNP:4183
K11:SNP:4317
K11:SNP:4215
K11:SNP:4334
NCYC110 SNPs
NCYC110:SNP:3902
NCYC110:SNP:4195
NCYC362 SNPs
S288c SNPs
ScDB53 SNPs
ScDB53:SNP:4116
ScDB53:SNP:4248
ScDB53:SNP:4128
ScDB53:SNP:4129
ScDB53:SNP:4165
ScDB53:SNP:4203

GBrowse


3480 3490 3500 3510 3520 3530 3540 3550 3560 3570 3580 3590 3600 3610 3620 3630 3640
Variation
Sp1528:pSNP:3493
Sp1528:pSNP:3513
K11:SNP:3538
Y12:pSNP:3555
378604X:DEL:3580
ScDB53:pSNP:3605
Sp1528:pSNP:3625
YS4:pSNP:3494
Sp1528:pSNP:3516
Y12:pSNP:3538
Y9:pSNP:3555
YS4:DEL:3580
ScDB53:pSNP:3609
ScDB53:INS:362
YS9:pSNP:3494
Sp1528:pSNP:3517
Y9:pSNP:3538
YS9:DEL:3580
YIIc17_E5:pSNP:3609
YS4:pSNP:36
Sp1528:pSNP:3496
UW03_461_4:pSNP:3517
SK1:DEL:3543
ScDB53:COMPLEX:3582
ScDB53:DEL:3610
273614X:COMPLE
UW05_217_3:pSNP:3517
ScDB53:COMPLEX:3582
YIIc17_E5:pSNP:3610
YS4:pSNP:
UW05_227_2:pSNP:3517
ScDB53:COMPLEX:3584
NCYC110:pSNP:3611
YS4:DE
UW83_787_3:pSNP:3517
ScDB53:COMPLEX:3584
YIIc17_E5:pSNP:3611
UW87_2421:pSNP:3517
ScDB53:DEL:3586
ScDB53:pSNP:3612
Sp1528:pSNP:3518
ScDB53:DEL:3586
YIIc17_E5:pSNP:3612
ScDB53:COMPLEX:3588
K11:DEL:3619
ScDB53:DEL:3589
K11:DEL:3619
ScDB53:COMPLEX:3590
K11:DEL:3619
Y12:DEL:3619
ScDB53:DEL:3591
Y12:DEL:3619
Y9:DEL:3619
ScDB53:COMPLEX:3592
Y9:DEL:3619
Sp1528:pSNP:3621

GBrowse


3480 3490 3500 3510 3520 3530 3540 3550 3560 3570 3580 3590 3600 3610 3620 3630 3640
Variation
Sp1528:pSNP:3493
Sp1528:pSNP:3513
K11:SNP:3538
Y12:pSNP:3555
378604X:DEL:3580
ScDB53:pSNP:3605
Sp1528:pSNP:3625
YS4:pSNP:3494
Sp1528:pSNP:3516
Y12:pSNP:3538
Y9:pSNP:3555
YS4:DEL:3580
ScDB53:pSNP:3609
ScDB53:INS:362
YS9:pSNP:3494
Sp1528:pSNP:3517
Y9:pSNP:3538
YS9:DEL:3580
YIIc17_E5:pSNP:3609
YS4:pSNP:36
Sp1528:pSNP:3496
UW03_461_4:pSNP:3517
SK1:DEL:3543
ScDB53:COMPLEX:3582
ScDB53:DEL:3610
273614X:COMPLE
UW05_217_3:pSNP:3517
ScDB53:COMPLEX:3582
YIIc17_E5:pSNP:3610
YS4:pSNP:
UW05_227_2:pSNP:3517
ScDB53:COMPLEX:3584
NCYC110:pSNP:3611
YS4:DE
UW83_787_3:pSNP:3517
UW87_2421:pSNP:3517
Sp1528:pSNP:3518
K11:DEL:3619
Ref: ttttttatt
ttttttatt : 35.3%
- : 64.7%
ScDB53:DEL:3589
ScDB53:COMPLEX:3590
ScDB53:DEL:3591
ScDB53:COMPLEX:3592
K11:DEL:3619
K11:DEL:3619
Y12:DEL:3619
Y12:DEL:3619
Y9:DEL:3619
Y9:DEL:3619
Sp1528:pSNP:3621

GBrowse

Name:	**YS9:pSNP:3494**
Type:	pSNP
Description:	
Source:	TURNIP_YS9
Position:	rDNA:3494..3494
Length:	1
acounts:	TUR:t 0.162162162162162 6 c 0.837837837837838 31 37
alleles:	t c
load_id:	YS9:pSNP:3494
parent_id:	gnl|ti|1750954050 gnl|ti|1750954237 gnl|ti|1750956120 gnl|ti|1750956247 gnl|ti|1750957739 gnl|ti|1750959489 gnl|ti|1750959522 gnl|ti|1750959628 gnl|ti|1750959900 gnl|ti|1750959917 gnl|ti|1750960749 gnl|ti|1750962204 gnl|ti|1750963051 gnl|ti|1750963163 gnl|ti|1750963195 gnl|ti|1750963337 gnl|ti|1750963590 gnl|ti|1750964720 gnl|ti|1750965368 gnl|ti|1750966172

GBrowse

Name:	**YS9:pSNP:3494**
Type:	pSNP
Description:	
Source:	TURNIP_YS9
Position:	rDNA:3494..3494
Length:	1
acounts:	TUR:t 0.1621621
alleles:	t
	c
load_id:	YS9:pSNP:3494
parent_id:	gnl|ti|175095405
	gnl|ti|175095423
	gnl|ti|175095612
	gnl|ti|175095624
	gnl|ti|175095773
	gnl|ti|175095948
	gnl|ti|175095952
	gnl|ti|175095962
	gnl|ti|175095990
	gnl|ti|175095991
	gnl|ti|175096074
	gnl|ti|175096220
	gnl|ti|175096305
	gnl|ti|175096316
	gnl|ti|175096319
	gnl|ti|175096333
	gnl|ti|175096359
	gnl|ti|175096472
	gnl|ti|175096536
	gnl|ti|175096617

Main | Obtaining Data | Statistics | Tracking | Documentation | Trace Assembly | SRA | Trace Home | Trace BLAST

Search | Searching Tips | Searchable Fields | Registered Species | Submitting Centers | FTP

Enter a *query string* (use Query Builder) or *TI number*

Submit 1750959489

Search result: found 1 item

Your request is: *1750959489*

Save result of search as trace **.tar** **.gz file.**

Save query's result **All** FASTA Quality SCF Mate Pair

Retrieve

Show as FASTA in color

>gnl|ti|1750959489 name:*YS9-19n06.q1k* Mate pair:*1750959488* **Send to BLAST**

Quality score: not available >=0 - <20 >=20 - <40 >=40 - <60 >=60 - <80 >=80 - <100

```
TTTTTACAATACTATAGAATCGAGCTCGGTCCCGGGGATCCTCTAGAGTCTGTCTGATTTGTTTTTTTTAT
TTCTTTCTAAGTGGGTACTGGCAGGAGCCGGGGCCTAGTTTAGAGAGAAGTAGACTGAACAAGTCTCTAT
AAATTTTATTTGTCTTAAGAACTCTATGATCCGGGTAAAAACATGTATTGTATATATCTATTATAATATA
CGATGAGGATGATAGTGTGTAAGAGTGTACCATTTACTATTTGGTCTTTTTTATTTTTTTATTTTTTATTTTT
CTTTTTTTTTTTTTTTTCGTTGCAAAGATGGGTTAAAAGAGAAGGGCTTTCACGAAGCTTCCCGAGCGTGA
AAGGATTTGCCCGGACAGTTTGCTTCATGGAGCAGTTTTTTCCGCACCATCAGAGCGGCAAACATGAGTG
CTTTTATAAGTTTAGAGAATTGAGAAAAGCTCATTTCCTATAGTTAACAGGACATGCCTTTGATATGAAA
AAAAATACTACGAACTACGATTTTACCAAGAAAGATGTAAGAGACAAGTGAACAGTGAACAGTGATAGTG
GGGACATTTTTTTTTTTTTTAAGTAAATGGCAGTTTCTAGGGAATGATGATGGCAAGTTCCAGAGAGGCAG
CGTAAAAGGATGAGGCTACTGGGAAGAAGAAAAGAGGAAAAGTGCAAGATGAATAGCCAGTGCAATATATA
TACATGTATACTTAACAGATATGGAATGGTTGGCGAAGTAAATTTTTGGCGACGCGGTATGCGGAGTTGT
AAGATGTACTACGATCGTATAGTGTACTGCGGCAAAATGTTAGTGCAGGAAAGCGGGAAGGAAAAAGAAG
CAACTAAACGAGGGTGTAGAAAAAGACGAAGAGGAAAGGATAACCGTAGAAGAGAGGAAATGGAGGGAAG
AGAACAAAACGGGAGTGTTTTTTTTTTTTTGTAGGATATCGGAAGAAATATTGTTTGTACTATCAGGTAT
AGCAACCCATGGGCTTAATGGAGGGCTAACCTCTCAAAGAA
```

Basic Phylogenetic Analysis

- Take each base position and set 1 for variation type seen, 0 if not

```
0000111001000100000000000010100000010000001
00001010010001000000100000101000000000000000
```

- Produce a distance matrix
- Pairwise using simple Euclidean distance (histogram intersection)

$$\sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}.$$

- quicktree → tree file in Newick format
- newicktops → image file

Basic Phylogenetic Analysis

- Culture Search
- Photomicrograph Search
- SGRP Strains
- Services
- About NCYC
- NCYC Cultures
- Catalogue
- Research
- Software
- Yeast Help

TURNIP Variation Maps

You can view the output from the TURNIP scripts by selecting the species, and then the strains in the box below.

(•) S. cerevisiae ( ) S. paradoxus

Y55
Y9
YIIc17_E5
YJM975
YJM978
YJM981
YPS128
YPS606
YS4
YS9

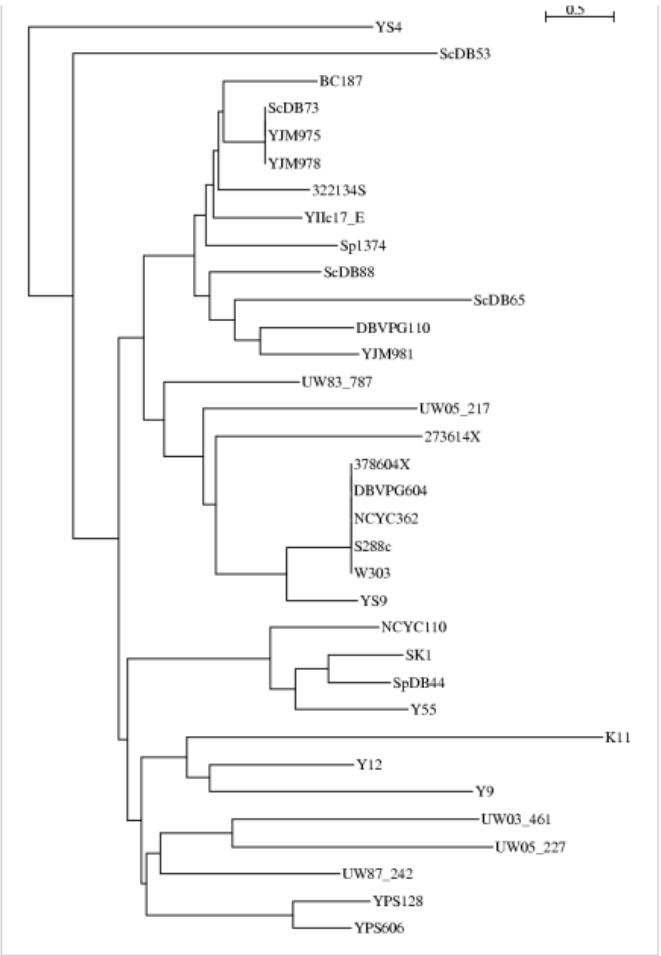
Select All

Select the type of variation you want to see:

( ) All (•) SNP ( ) pSNP ( ) Indel [Go]

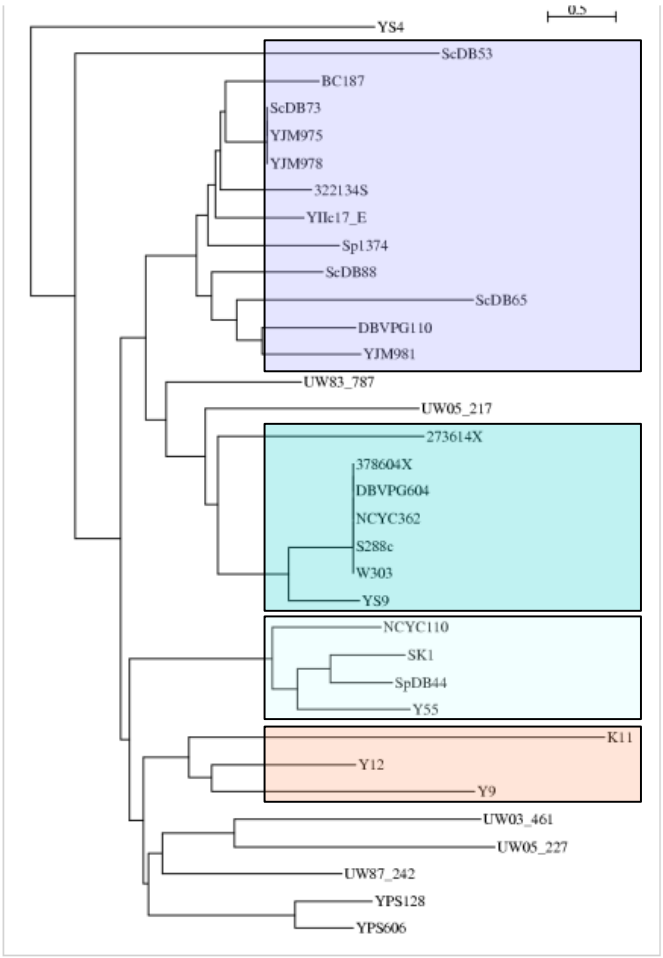
Basic Phylogenetic Analysis

S. cerevisiae, by SNP



Basic Phylogenetic Analysis

S. cerevisiae, by SNP



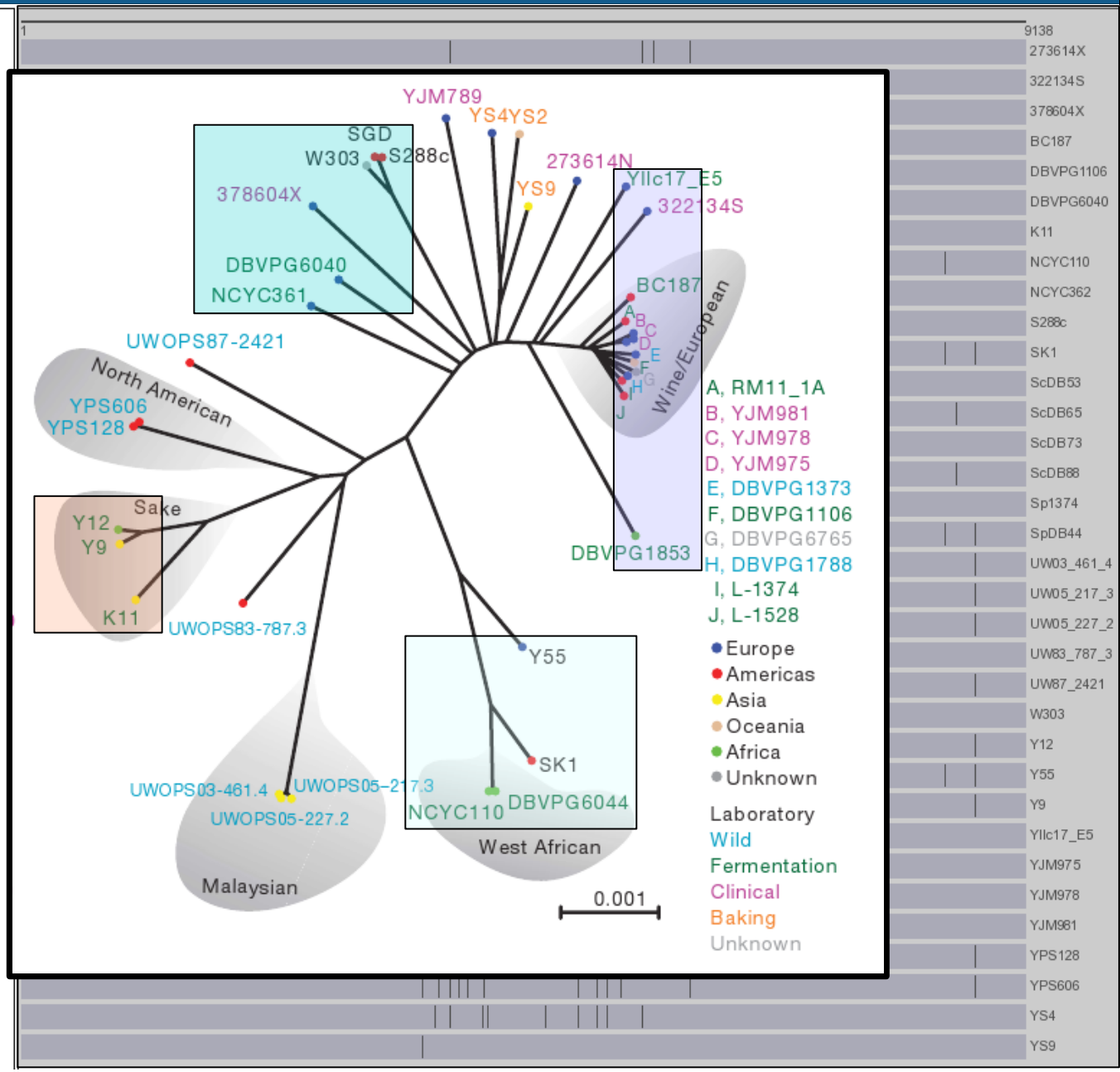
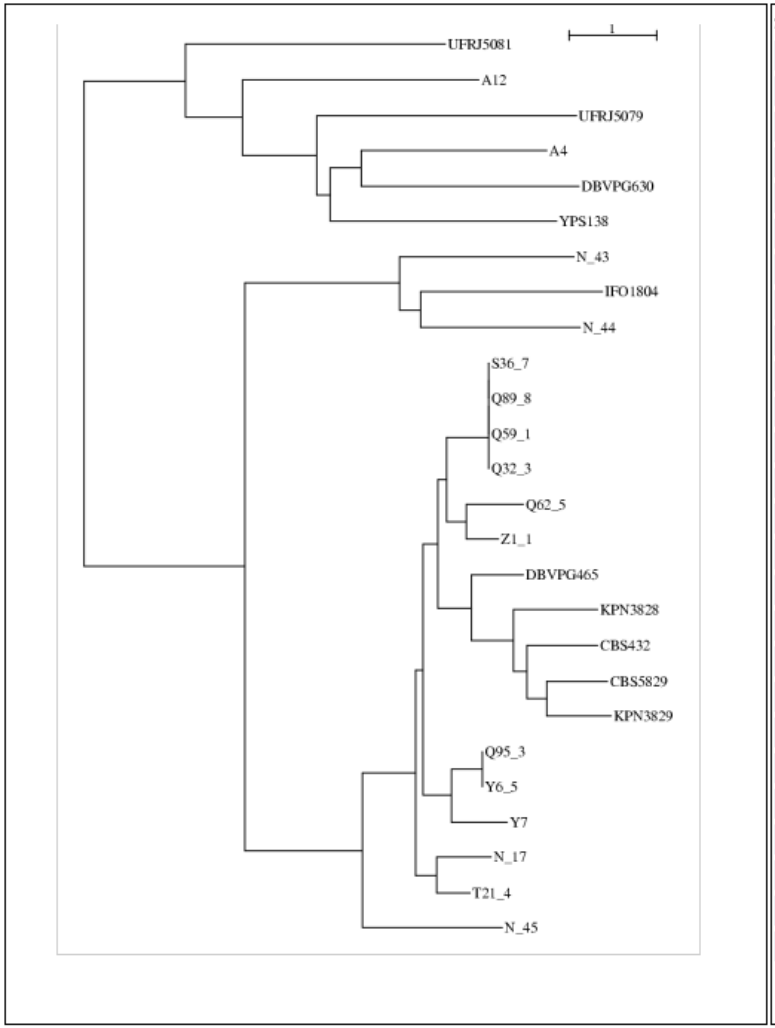
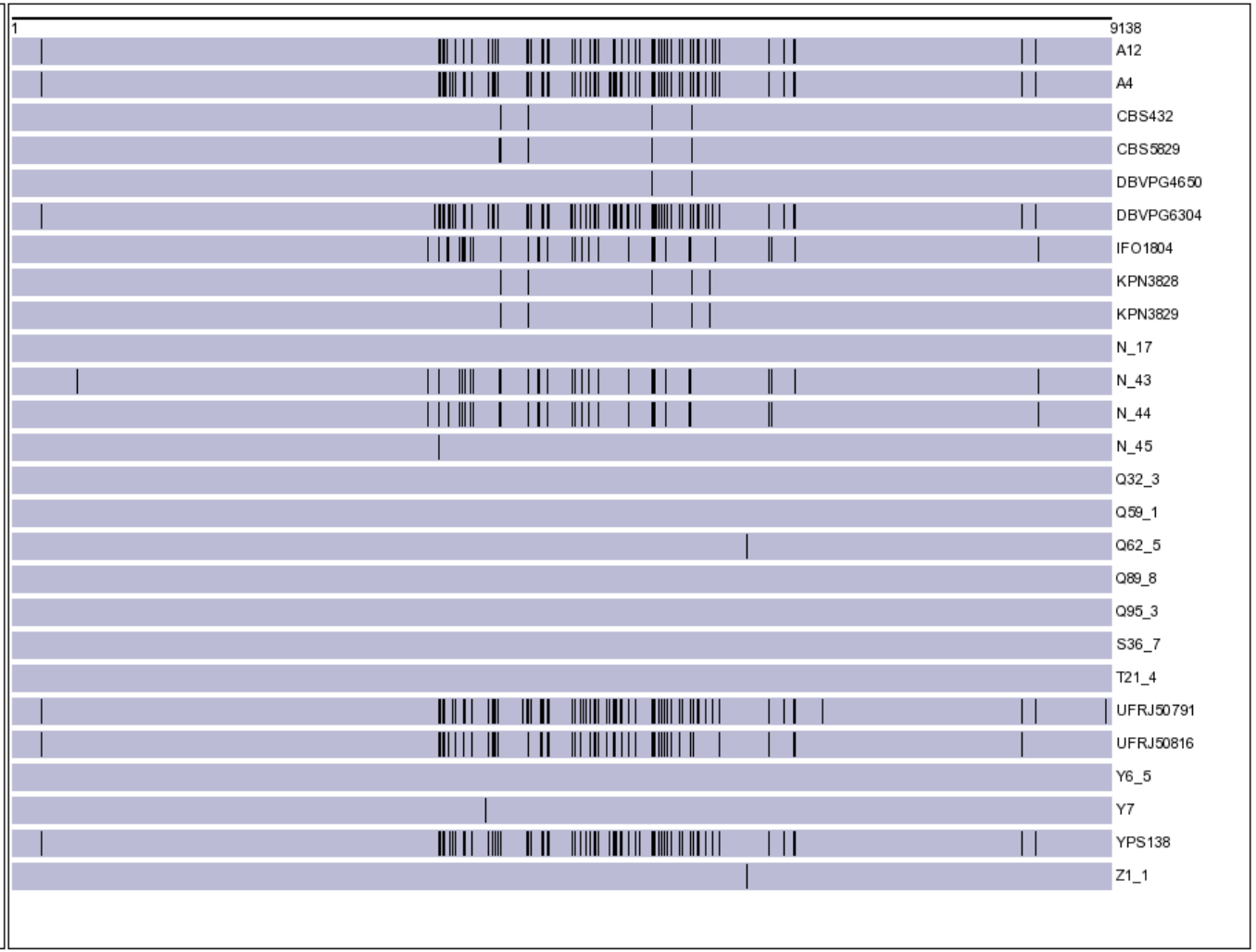


Image courtesy of Liti G. et al, Nature, 2009

Basic Phylogenetic Analysis

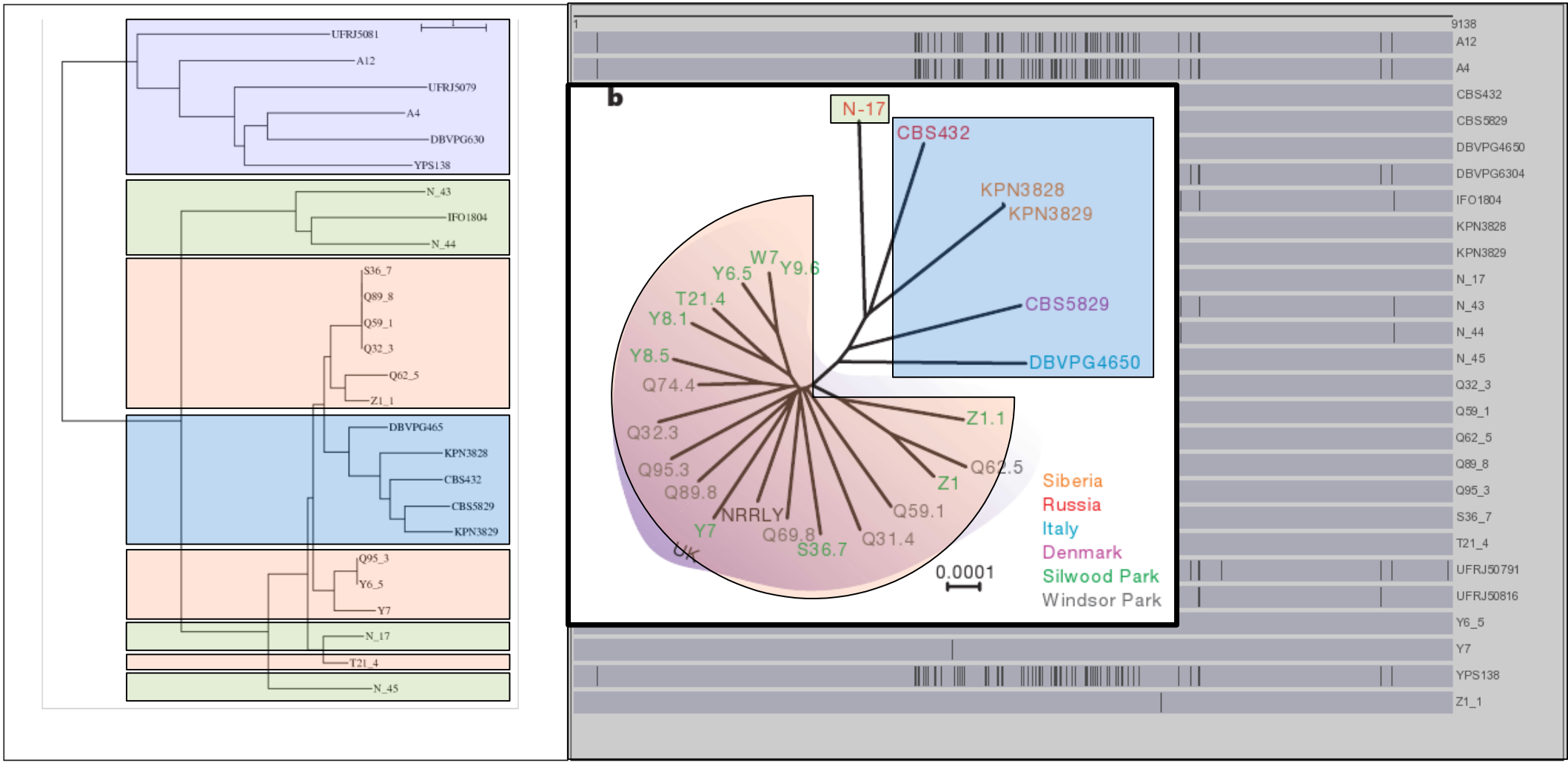
S. paradoxus, by SNP





Basic Phylogenetic Analysis

S. paradoxus, by SNP



Americas

Asia

Image courtesy of Liti G. et al, Nature, 2009

Conclusions

- Variation within individual *Saccharomyces* rDNA repeats remarkably high, and differs markedly between strains
- Some pSNPs strain specific, others shared between a number of strains, potentially at variable frequencies
- Correlation between genome type and pSNP number, i.e.
  - structured (*clean lineage*) genomes have fewer pSNPs
  - mosaic (*hybrid lineage*) tend to have more
  - pSNPs may provide measure of genome mosaicism
- Shared pSNPs between different lineages may provide novel measure of recombination rates and gene conversion
- Hope to employ better distance algorithms in the near future
- Better understand mechanisms of rDNA evolution
- Mathematical simulations and qPCR

Acknowledgements

NCYC

Ian Roberts

Steve James

JIC

John Walshaw

Jo Dicks

MIT

Michael O'Kelly

http://www.ncyc.co.uk

Sanger

David Carter

BBSRC TRDF Project