Computational approaches to explore variation and dynamics in ribosomal DNA sequences

Rob Davey

NCYC

2008

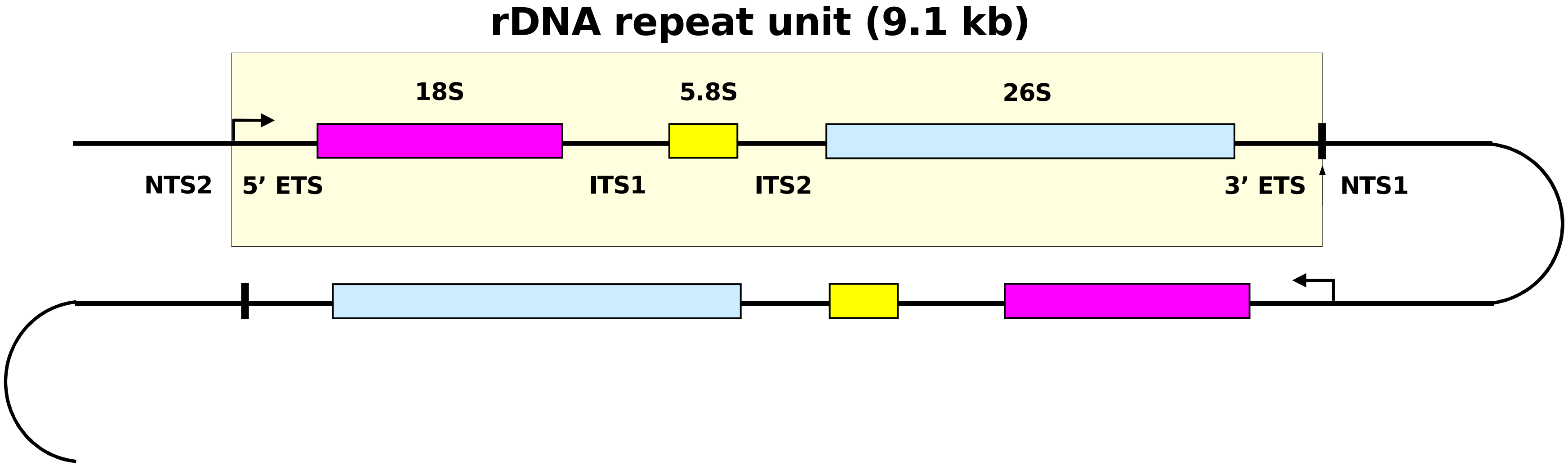
Introduction

- 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)

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%)
- 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 with associated quality per base (FASTQ format)
- Cannot assemble repeats due to high similarity (*Ganley 2007*)
- Single rDNA 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)
- Perl

TURNIP

consensus

```
..agcaaactgtccgggcaaatcctttcacgctcgggaagctttgtgaaagcccttctctttcaa..

            ccgggcaaatcctttcacactcgggaagctttgtgaaagcccttctctttcaa..
..agcaaactgtccgggcaaatcctttcacactcgggaagctttgtgaaagcccttctcttt
        ctgtccgggcatatcctttcacactcgggaagctttgtgaaagcccttctctttcaa.
..agcaaactgtccgggcaaatcctttcacactcgggaagctttgtgaaaaagccct
..agcaaactgtccgggcatatcctttcacactcgggaagc---gtgaaagcccttctctttcaa..
..agcaaactgtccgggcatatcctttcacactcgggaagctttgtgaaagc
   gcaaactgtccgggcatatcctttcacactcgggaagctttgtgaaagcccttctctttc
..agcaaactgtccgggcaaatcctttcacactcgggaagctttgtgaaagcccttctctttcaa..
```

pSNP
4/8
(50%)

SNP

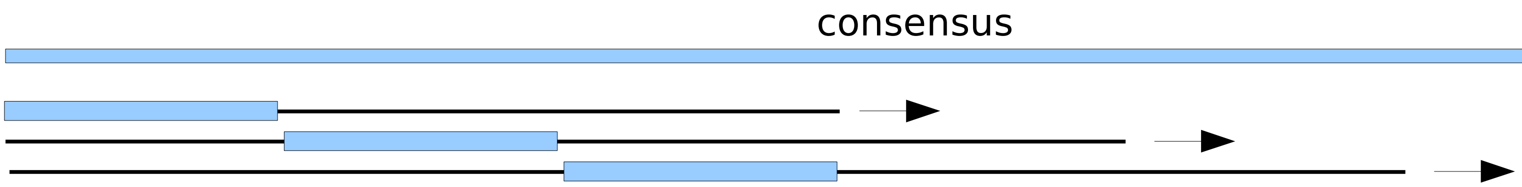
DEL

INS

- 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

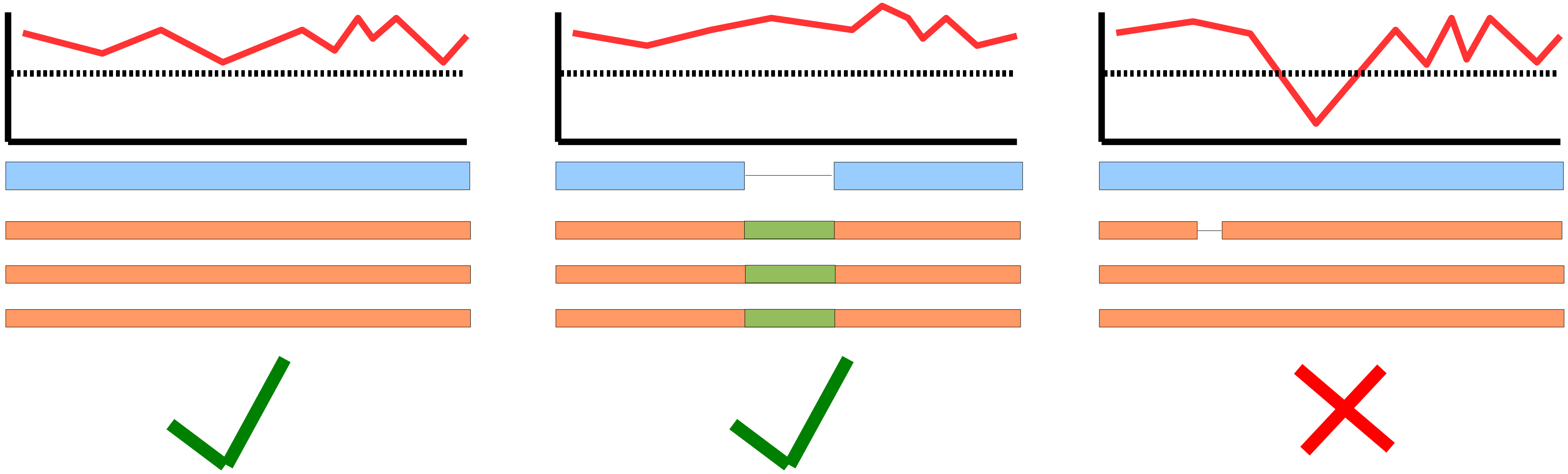
- Take 20bp slices of consensus (query sequence)
- Anchored on each side by 40bp flanking sequence to give a more accurate alignment



- '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 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 PHRED quality score > threshold
  - Variation less likely to be sequencing error
- For each accepted 20bp slice, check for insertions, i.e. gaps introduced into BLAST query sequence by MUSCLE



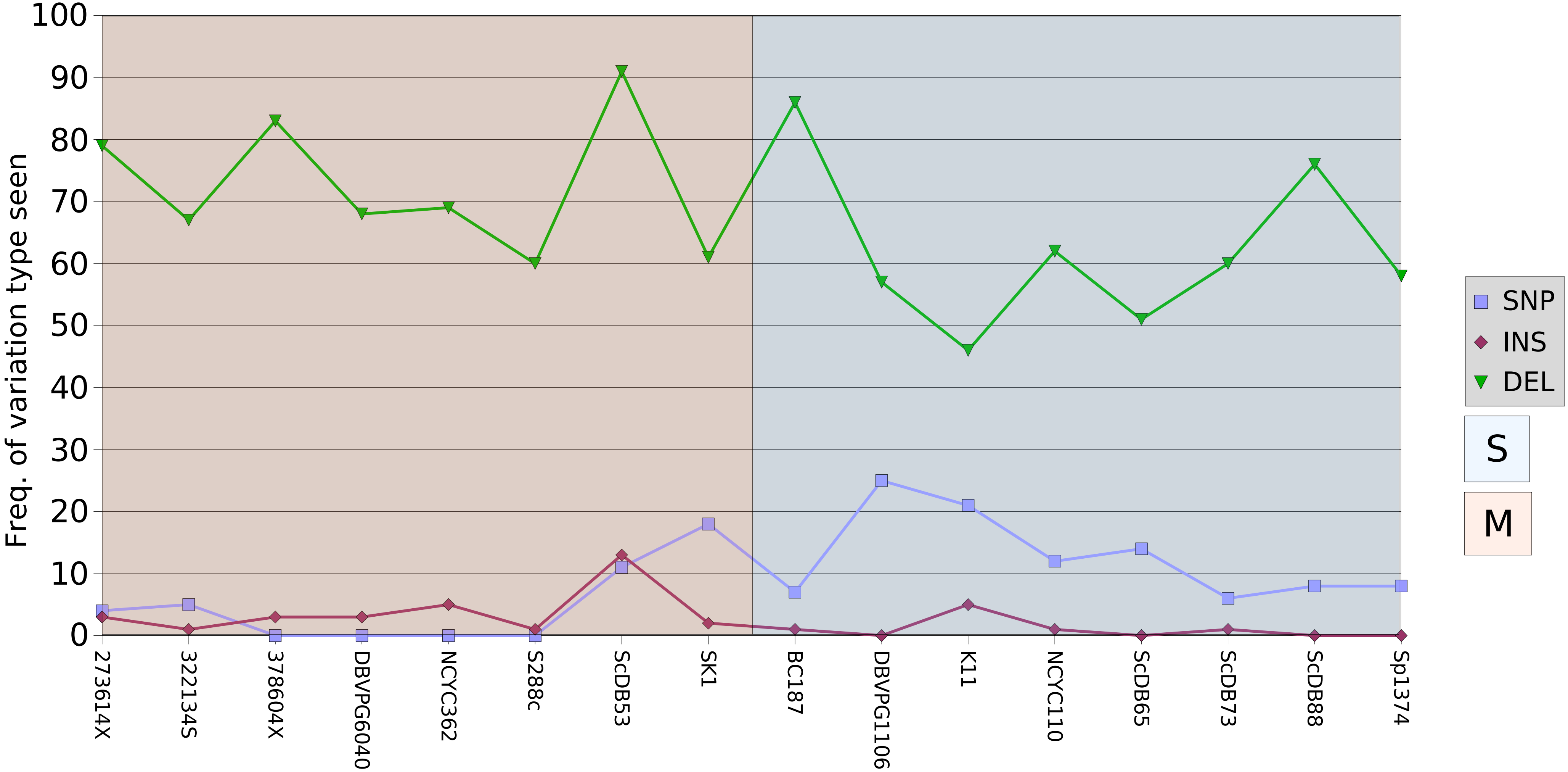
TURNIP

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

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

- Outputs
  - Raw text, Excel, SQL, GFF
  - Use GFF to import data into GBrowse

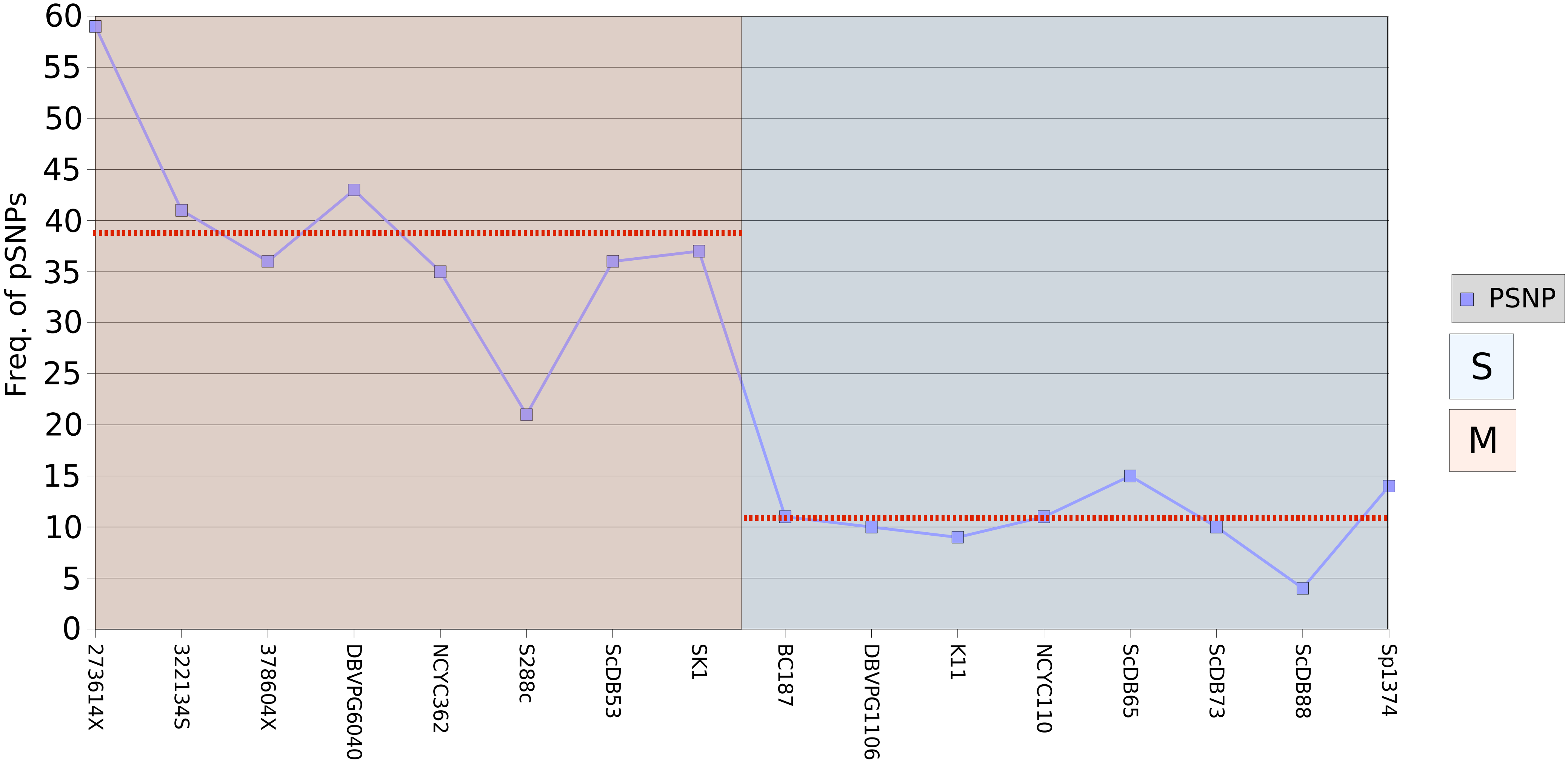
14 S. cerevisiae strains - Mosaic vs Structured



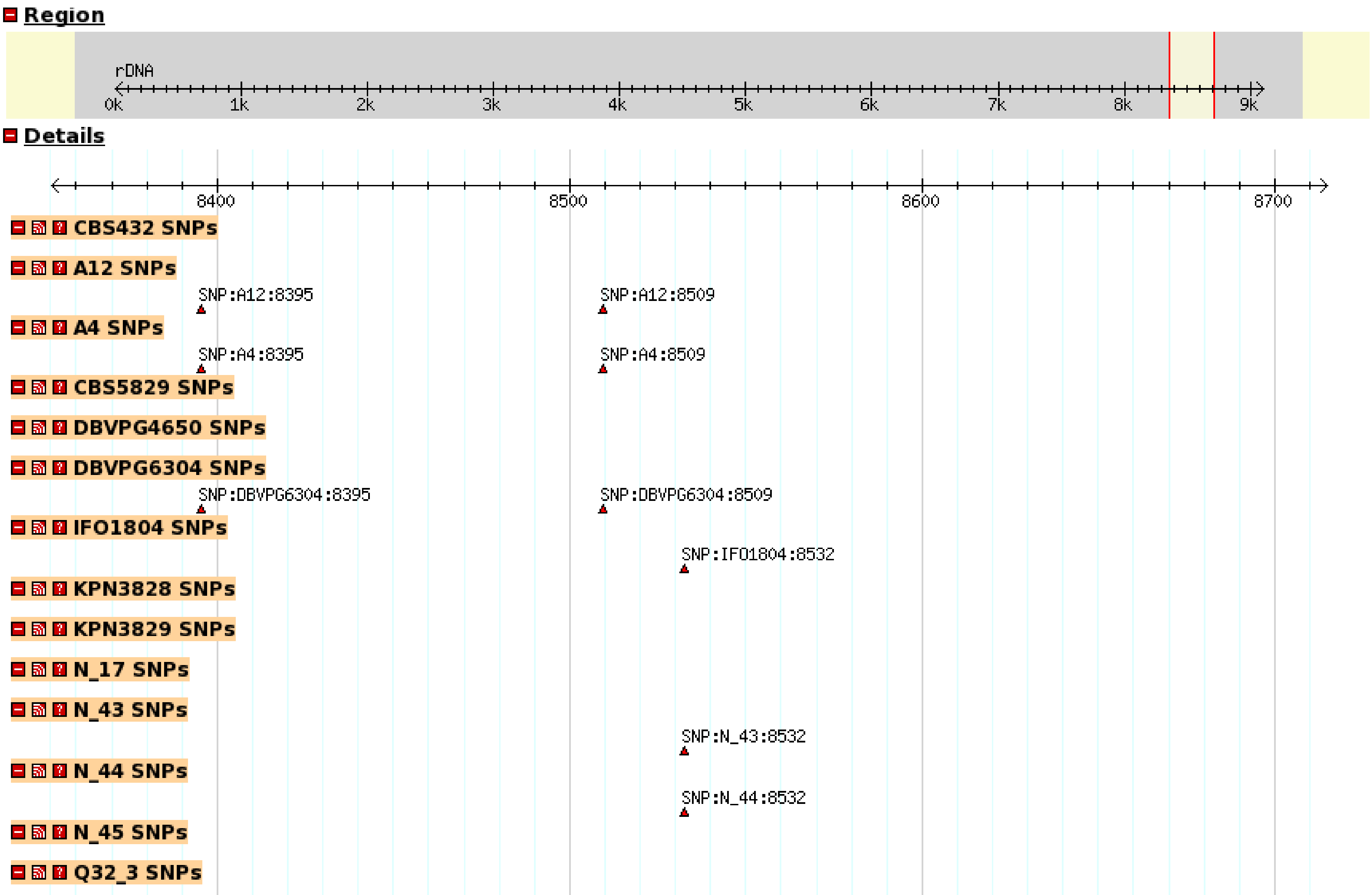
- Two genome types, structured and mosaic (*Carter 2008*)
- Structured – 'clean' genome, assumed pure lineage
- Mosaic - genetically different cell lines from a single zygote (hybrid)

Preliminary Results

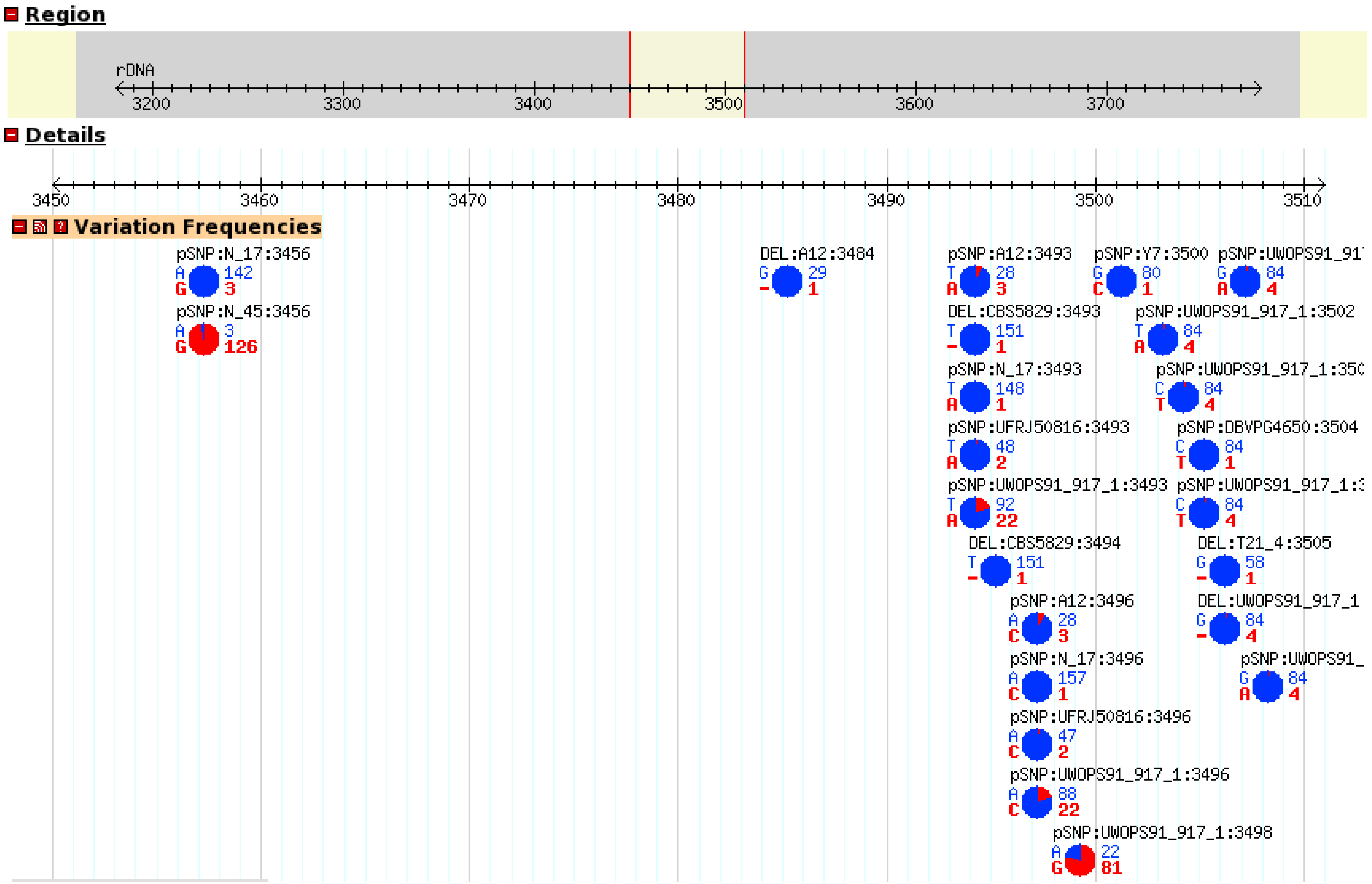
14 S. cerevisiae strains - Mosaic vs Structured



GBrowse


Region
rDNA
0k
1k
2k
3k
4k
5k
6k
7k
8k
9k
Details
8400
8500
8600
8700
CBS432 SNPs
A12 SNPs
SNP:A12:8395
SNP:A12:8509
A4 SNPs
SNP:A4:8395
SNP:A4:8509
CBS5829 SNPs
DBVPG4650 SNPs
DBVPG6304 SNPs
SNP:DBVPG6304:8395
SNP:DBVPG6304:8509
IFO1804 SNPs
SNP:IFO1804:8532
KPN3828 SNPs
KPN3829 SNPs
N_17 SNPs
N_43 SNPs
SNP:N_43:8532
N_44 SNPs
SNP:N_44:8532
N_45 SNPs
Q32_3 SNPs

GBrowse



Region

Details

Variation Frequencies

pSNP:N_17:3456
A 142
G 3

pSNP:N_45:3456
A 3
G 126

DEL:A12:3484
G 29
- 1

pSNP:A12:3493
T 28
A 3

DEL:CBS5829:3493
T 151
- 1

pSNP:N_17:3493
T 148
A 1

pSNP:UFRJ50816:3493
T 48
A 2

pSNP:UWOPS91_917_1:3493
T 92
A 22

DEL:CBS5829:3494
T 151
- 1

pSNP:A12:3496
A 28
C 3

pSNP:N_17:3496
A 157
C 1

pSNP:UFRJ50816:3496
A 47
C 2

pSNP:UWOPS91_917_1:3496
A 88
C 22

pSNP:UWOPS91_917_1:3498
A 22
G 81

pSNP:Y7:3500
G 80
C 1

pSNP:UWOPS91_91
G 84
A 4

pSNP:UWOPS91_917_1:3502
T 84
A 4

pSNP:UWOPS91_917_1:350
C 84
T 4

pSNP:DBVPG4650:3504
C 84
T 1

pSNP:UWOPS91_917_1:3
C 84
T 4

DEL:T21_4:3505
G 58
- 1

DEL:UWOPS91_917_1
G 84
- 4

pSNP:UWOPS91_
G 84
A 4

Conclusions

- Variation within individual *S. cerevisiae* rDNA repeats to be remarkably high
- 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
- On average structured genomes have fewer pSNPs, hybrids tend to have more
- pSNPs may provide simple measure of genome mosaicism
- Shared pSNPs between different lineages may provide novel measure of recombination rates and gene conversion
- A new way to aid strain identification? Supply of probiotic *S. boulardii* across EU requires precise quality control

Acknowledgements

NCYC

Ian Roberts

Steve James

MIT

Michael O'Kelly

Sanger

David Carter

BBSRC TRDF Project

JIC

John Walshaw

Jo Dicks

http://www.ncyc.co.uk