# Supplemental Results

## Detection of aneuploidy in clinical and laboratory strains

Clinical strains of *C. albicans* are predominantly diploid (Hickman *et al.*, 2013; Hirakawa *et al.*, 2015), yet aneuploidy is often observed and may be an important mechanism for adaptation (Li *et al.*, 2015; Todd *et al.*, 2017). In addition, clinical strains undergo loss of heterozygosity (LOH) with even greater frequency (Ropars *et al.*, 2018) We compared oak and clinical strain ploidy by comparing the short-read genome sequence data generated here to published data for laboratory and clinical strains. More specifically, we applied a standard base calling approach for estimating ploidy from sequences to data for oak strains. We showed that our methods would be able to detect aneuploidy by also applying them to the laboratory reference strain (SC5314), a related mutant (1AA) (Muzzey *et al.*, 2013), and a panel of 20 clinical strains that includes several well-characterized aneuploids (Hirakawa *et al.*, 2015). In addition, we generated short-read genome data for the clinical type strain of *C. albicans* (NCYC 597) for a direct comparison between oak strains and a clinical strain from the same sequencing batch.

In contrast to the oak strains, the type strain of *C. albicans* that we sequenced is predominantly triploid (Figure S1). Therefore, we could have detected ploidy variation in the oak strains had it been present. Indeed, our data suggest that the type strain has probably undergone large scale chromosomal rearrangements because most of its chromosomes show a mixture of ploidy states along their sequence (chromosomes 1, 4, 5, 6 and R, Figure S1). We were also able to detect six out of the seven cases of trisomy identified by Hirakawa *et al.* (2015), but missed tetrasomy for chromosome 5 in strain P75010 which was was fully homozygous (Figure S1). Therefore we could miss aneuploidy for homozygous chromosomes, but there are few cases of this for the oak strains (only chromosomes 5 and 7 of strain NCYC 4144 from oak, Figure 1a).

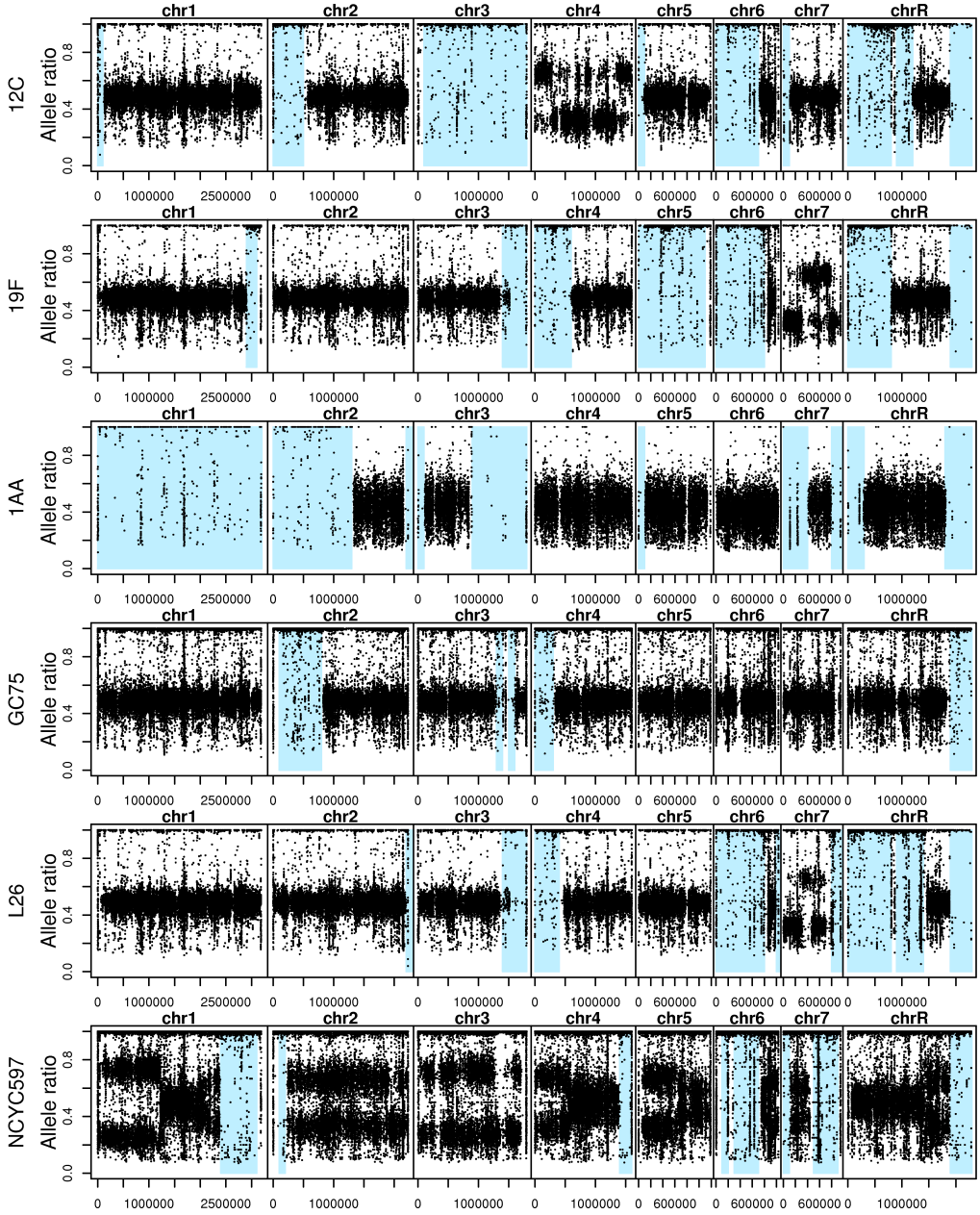
Using our methods, we also detected the previously known LOH events for clinical strains reported by (Hirakawa *et al.*, 2015, Figure S1) and the artificially induced whole-chromosome LOH of chromosome 1 for the mutant strain 1AA (Figure S1). Read depth in regions with low heterozygosity is similar to that in the rest of the genome (Figure S3), therefore these regions LOH events and not deletions or monosomy.

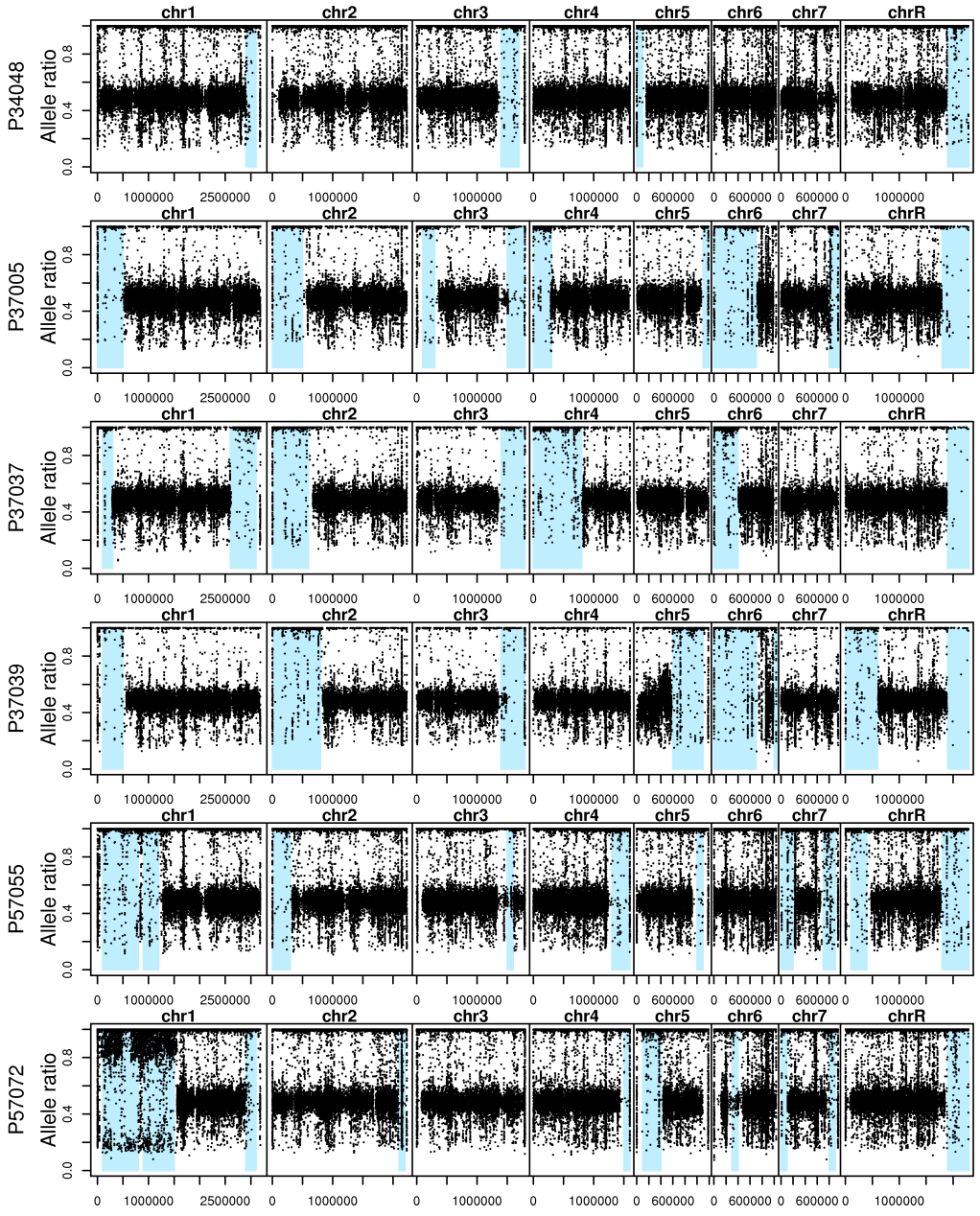
Table S 1: ***C. albicans*** **from oak show higher heterozygosity than the clinical strains in Hirakawa *et al.* (2015).**

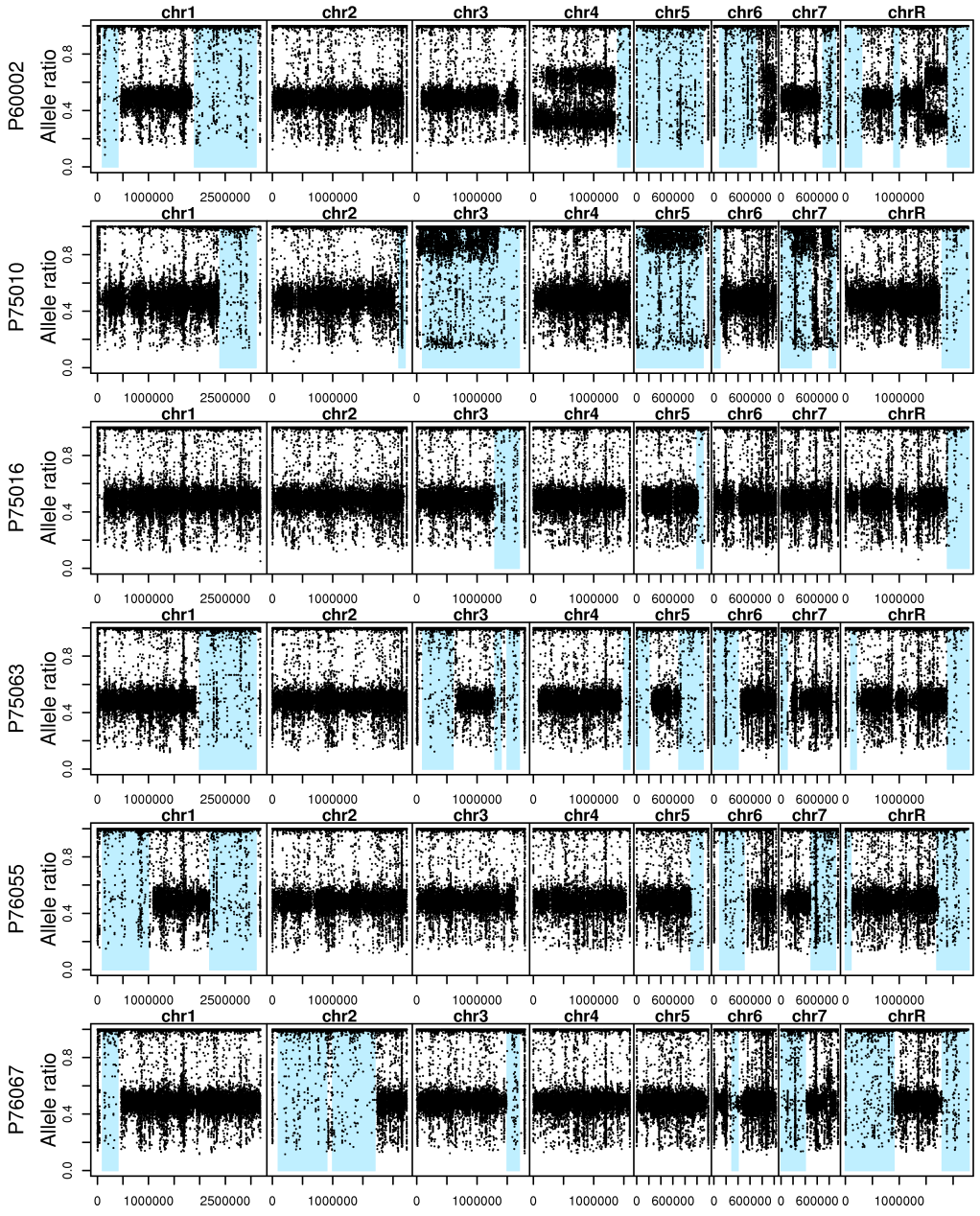
| Strain | MTL | MLST Clade (FP[a]) | Heterozygosity[b] | LOH length (Mbp)[c] | Filtered length (Mbp)[d] | Filtered heterozygosity[e] | Heterozygosity in 950 kb[f] |
|---|---|---|---|---|---|---|---|
| NCYC 4145 | $a/\alpha$ | ? | 0.0077 | 0.7 | 11.4 | 0.0078 | 0.0075 |
| NCYC 4146 | $a/\alpha$ | 4 (SA) | 0.0062 | 1.1 | 11.1 | 0.0066 | 0.0068 |
| NCYC 4144 | $a/a$ | 18 | 0.0061 | 2.3 | 9.7 | 0.0068 | 0.0074 |
| **3 oak strains** | | Mean: | **0.0066** | **1.4** | | **0.0070** | **0.0072** |
| P34048 | $a/\alpha$ | 3 (III) | 0.0060 | 1.0 | 10.3 | 0.0064 | 0.0070 |
| P75016 | $a/\alpha$ | 4 (SA) | 0.0059 | 0.9 | 10.5 | 0.0064 | 0.0066 |
| P78042 | $\alpha/\alpha$ | 3 (III) | 0.0057 | 1.3 | 10.1 | 0.0061 | 0.0065 |
| GC75 | $\alpha/\alpha$ | 4 (SA) | 0.0057 | 1.6 | 9.4 | 0.0067 | 0.0067 |
| P78048 | $\alpha/\alpha$ | 1 (I) | 0.0054 | 2.1 | 9.5 | 0.0061 | 0.0054 |
| P57055 | $a/\alpha$ | 3 (III) | 0.0052 | 3.0 | 8.9 | 0.0065 | 0.0068 |
| P37037 | $a/\alpha$ | 1 (I) | 0.0048 | 3.3 | 8.3 | 0.0061 | 0.0058 |
| NCYC597 | $a/\alpha$ | ? | 0.0048 | 2.4 | 9.3 | 0.0054 | 0.0061 |
| P37005 | $a/a$ | 1 (I) | 0.0047 | 3.3 | 8.4 | 0.0059 | 0.0057 |
| P57072 | $\alpha/\alpha$ | 2 (II) | 0.0047 | 2.7 | 8.4 | 0.0057 | 0.0058 |
| P75010 | $\alpha/\alpha$ | 11 (E) | 0.0046 | 4.7 | 6.7 | 0.0067 | 0.0068 |
| SC5314 | $a/\alpha$ | 1 (I) | 0.0046 | 2.8 | 10.0 | 0.0055 | 0.0053 |
| P76067 | $a/\alpha$ | 2 (II) | 0.0046 | 3.9 | 8.6 | 0.0061 | 0.0055 |
| P37039 | $a/\alpha$ | 1 (I) | 0.0045 | 3.9 | 8.1 | 0.0060 | 0.0057 |
| P75063 | $a/\alpha$ | 4 (SA) | 0.0045 | 3.6 | 8.1 | 0.0059 | 0.0067 |
| L26 | $a/a$ | 1 (I) | 0.0044 | 3.7 | 9.2 | 0.0058 | 0.0056 |
| 19F | $\alpha/\alpha$ | 1 (I) | 0.0042 | 4.3 | 8.1 | 0.0059 | 0.0057 |
| P76055 | $a/\alpha$ | 2 (II) | 0.0042 | 3.5 | 8.3 | 0.0052 | 0.0052 |
| P60002 | $a/a$ | 8 (SA) | 0.0041 | 4.4 | 7.7 | 0.0056 | 0.0070 |
| 12C | $a/a$ | 1 (I) | 0.0041 | 4.7 | 7.7 | 0.0061 | 0.0057 |
| P87 | $a/a$ | 4 (SA) | 0.0038 | 5.7 | 6.8 | 0.0062 | 0.0060 |
| P94015 | $a/a$ | 6 (I) | 0.0035 | 6.5 | 6.8 | 0.0062 | 0.0068 |
| 1AA | $a/\alpha$ | 1 (I) | 0.0029 | 7.1 | 5.6 | 0.0056 | 0.0039 |
| **22 clinical strains**[g] | | Mean: | **0.0047** | **3.3** | | **0.0060** | **0.0061** |

[a] Clade assignments are as summarized from past MLST and fingerprinting (FP) studies in Hirakawa *et al.* (2015). [b] Heterozygosity was estimated as the proportion of high quality sites (with phred-scaled quality over 40) where 20-80% of reads differed from the reference sequence. For all strains, this was estimated from approximately 14 Mbp of high quality sequence. [c] Length of sequence showing loss of heterozygosity (LOH). LOH was assumed where the proportion of heterozygous sites in a 100 kb window was lower than 0.001. [d] The length of genome sequence after excluding LOH regions, known repeats, putatively repetitive regions (positions with over double the mean genome-wide read depth) and centromeres. [e] The proportion of heterozygous sites after excluding LOH regions, repeats and centromeres. [f] the proportion of heterozygous sites in at 948,860 nucleotide sites with high quality, unrepetitive, non-LOH sequence for all 25 oak and clinical strains. [g] Means for clinical strains exclude data for strain 1AA because this strain was derived from SC5314.

Figure S 1: **Clinical *C. albicans* are mostly diploid and the type strain is mostly triploid.** The proportion of base calls differing from the reference strain (allele ratios or B allele frequencies) are mostly 1.0 or 0.5 for clinical strains, as expected for diploid strains, whereas allele ratios are mostly 0.33, 0.66 and 1.0 for the type strain (NCYC 597) suggesting triploidy. As expected, SC5314 differs from the SC5314_A22 reference at heterozygous sites, and the laboratory mutant (1AA) is homozygous on chromosome 1. Allele ratios also confirm 6 cases of trisomy identified by Hirakawa *et al.* (2015): 12C chr4, 19F chr7, L26 chr7, P60002 chr4 and chr6, P78042 chr4. However, a B allele frequency approach misses aneuploidy of chromosome 5 of P75010 which was homozygous. Regions that recently underwent Loss of Heterozygosity (LOH) are shaded light blue. Some clinical strains (19F, P60002 and P75010) are homozygous at the mating locus as a result of whole-chromosome LOH for chromosome 5.




chr1
chr2
chr3
chr4
chr5
chr6
chr7
chrR
P34048
P37005
P37037
P37039
P57055
P57072
Allele ratio
0.0
0.4
0.8
0
1000000
2500000
600000


chr1
chr2
chr3
chr4
chr5
chr6
chr7
chrR
P60002
P75010
P75016
P75063
P76055
P76067
Allele ratio
0.0
0.4
0.8
0
1000000
2500000
600000

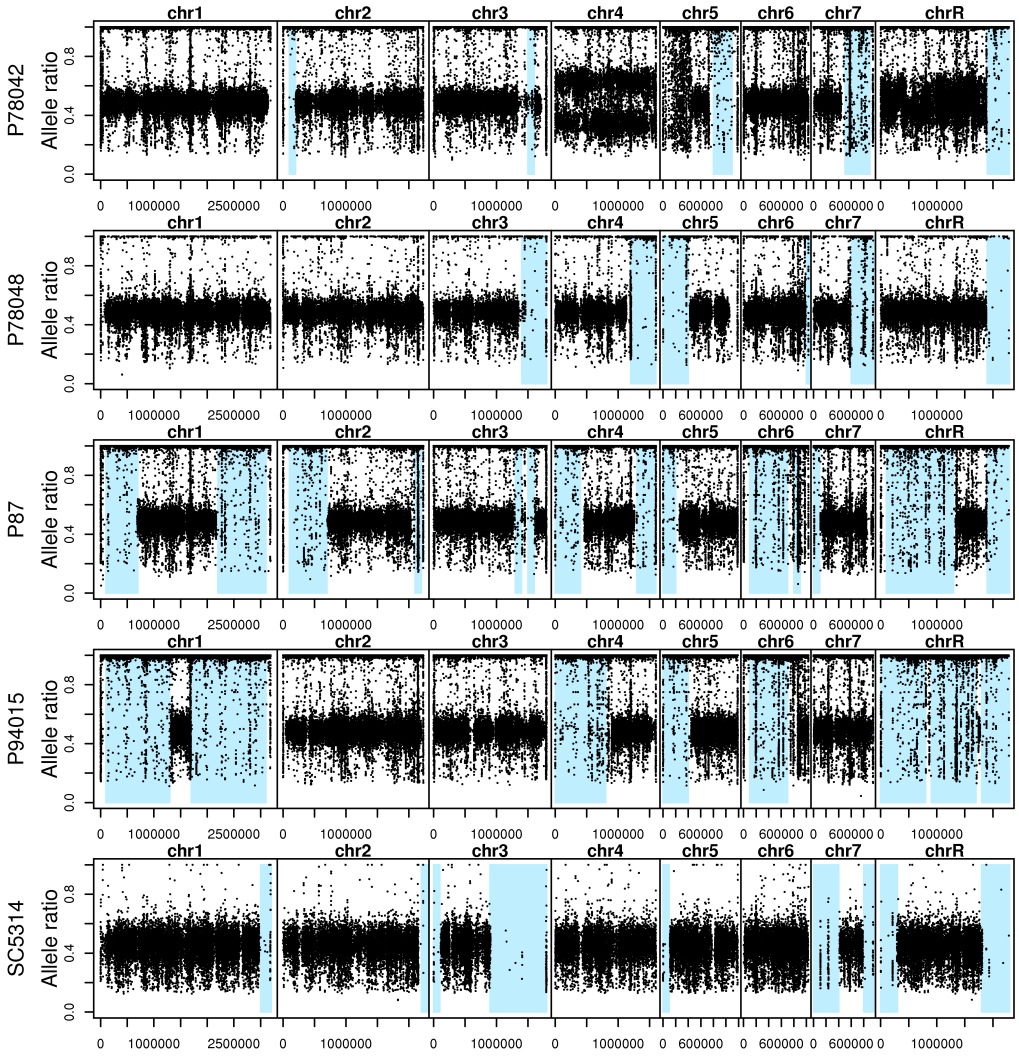
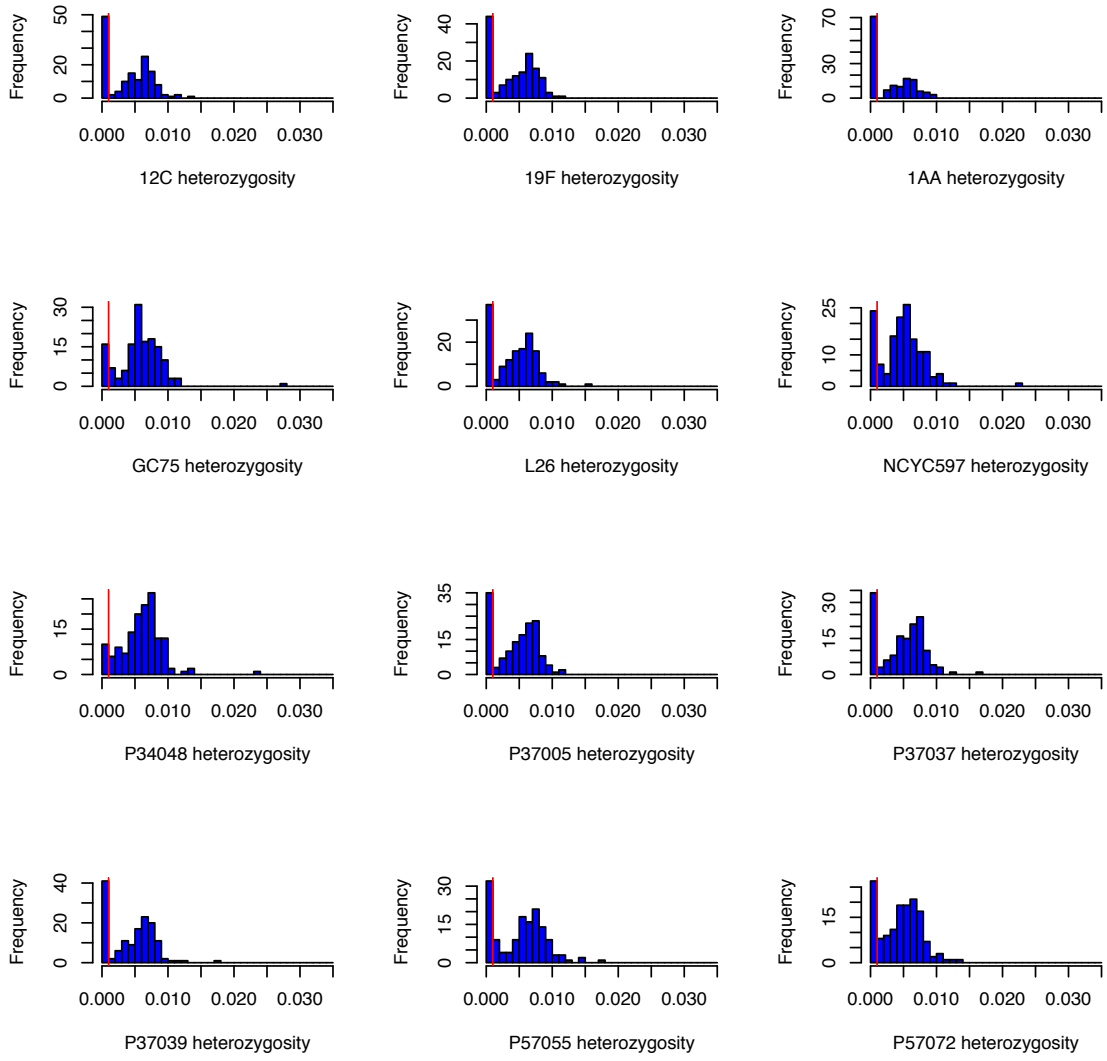
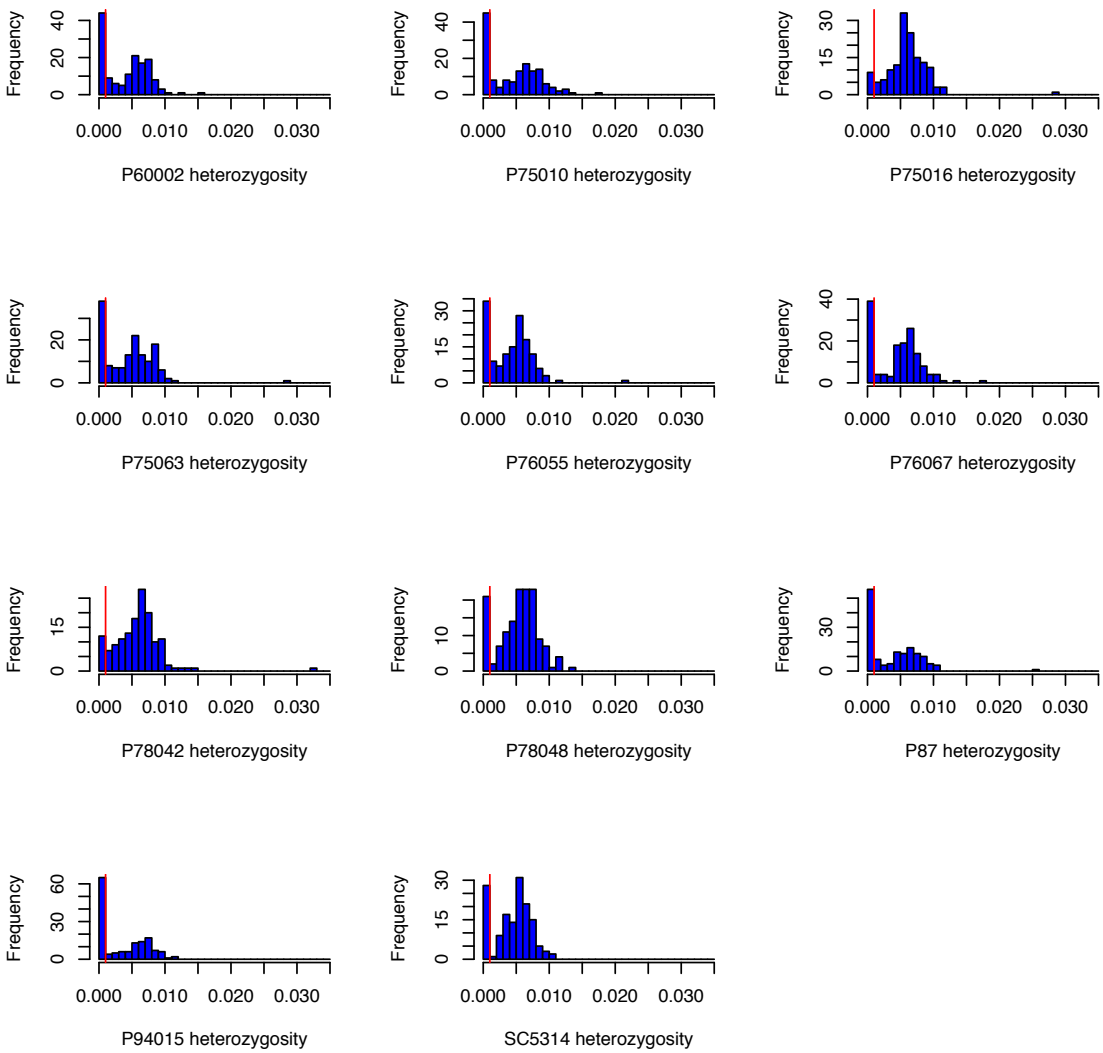

P78042
P78048
P87
P94015
SC5314
Allele ratio
0.0
0.4
0.8
chr1
chr2
chr3
chr4
chr5
chr6
chr7
chrR
0
1000000
2500000
600000

Figure S 2: **the distribution of levels of heterozygosity (LOH) estimated in 100 kb non-overlapping windows across the genomes of clinical strains.** The proportion of high quality heterozygous sites estimated from 100 kb windows. Levels of heterozygosity are bimodal, with one mode below 0.1% and another above 0.4%. By using a threshold of 0.1% to define LOH, we exclude 100 kb windows with fewer than 100 heterozygous sites. A large window size (100 kb) minimizes the chances of excluding a region of low heterozygosity due to sampling error, but could mean that short LOH regions are missed.




Frequency
0 20 40
0.000 0.010 0.020 0.030
P60002 heterozygosity
Frequency
0 20 40
0.000 0.010 0.020 0.030
P75010 heterozygosity
Frequency
0 15 30
0.000 0.010 0.020 0.030
P75016 heterozygosity
Frequency
0 20
0.000 0.010 0.020 0.030
P75063 heterozygosity
Frequency
0 15 30
0.000 0.010 0.020 0.030
P76055 heterozygosity
Frequency
0 20 40
0.000 0.010 0.020 0.030
P76067 heterozygosity
Frequency
0 15
0.000 0.010 0.020 0.030
P78042 heterozygosity
Frequency
0 10
0.000 0.010 0.020 0.030
P78048 heterozygosity
Frequency
0 30
0.000 0.010 0.020 0.030
P87 heterozygosity
Frequency
0 30 60
0.000 0.010 0.020 0.030
P94015 heterozygosity
Frequency
0 15 30
0.000 0.010 0.020 0.030
SC5314 heterozygosity

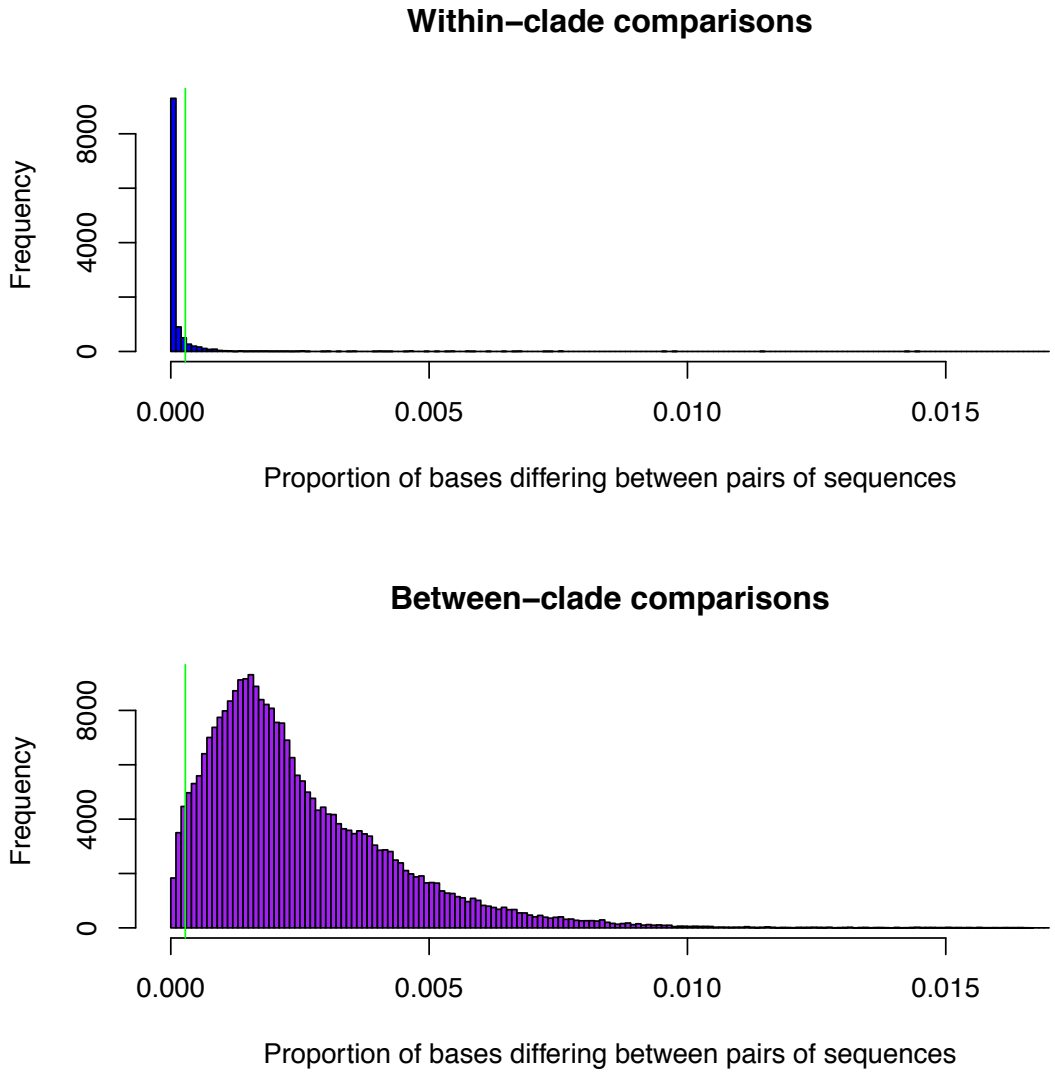
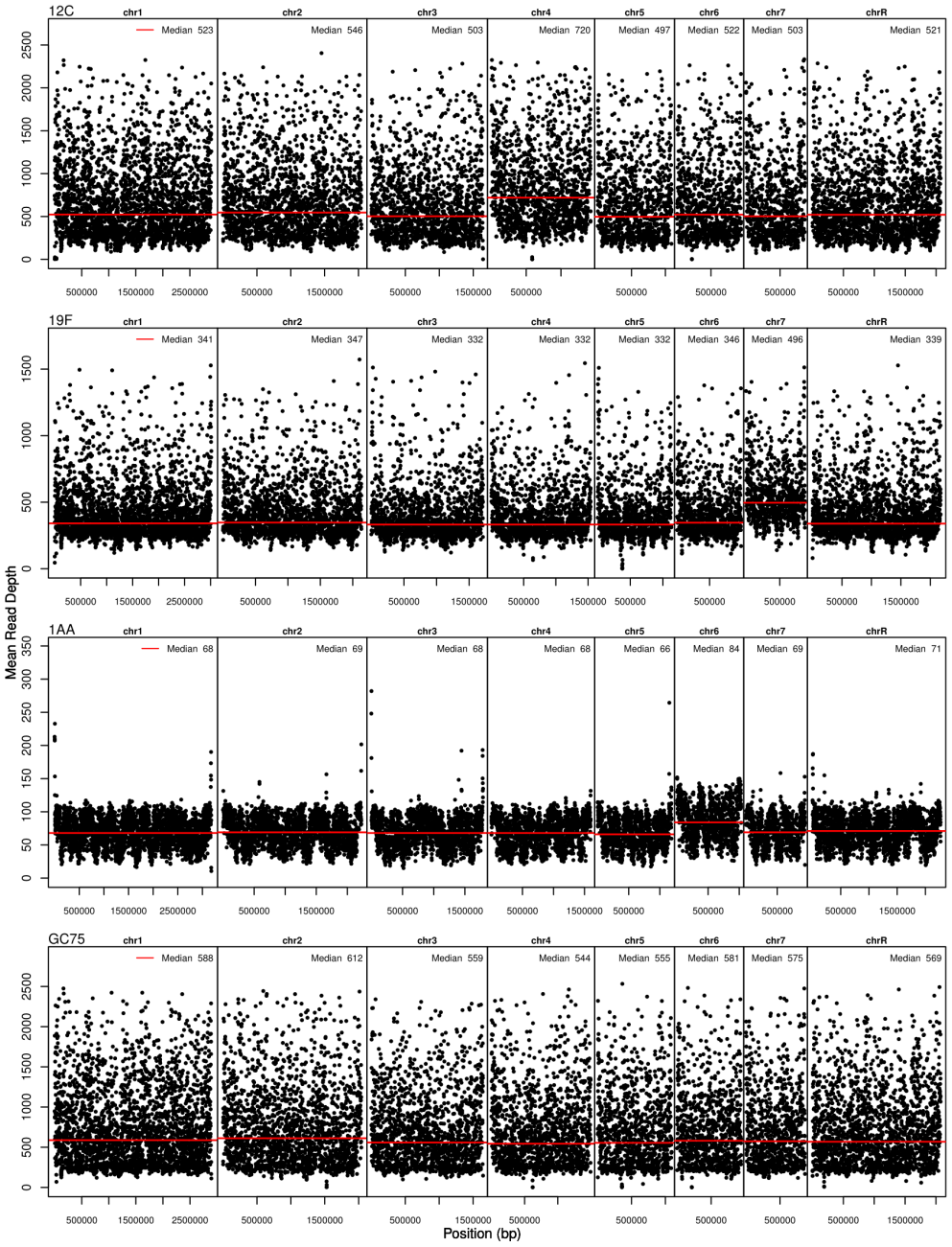
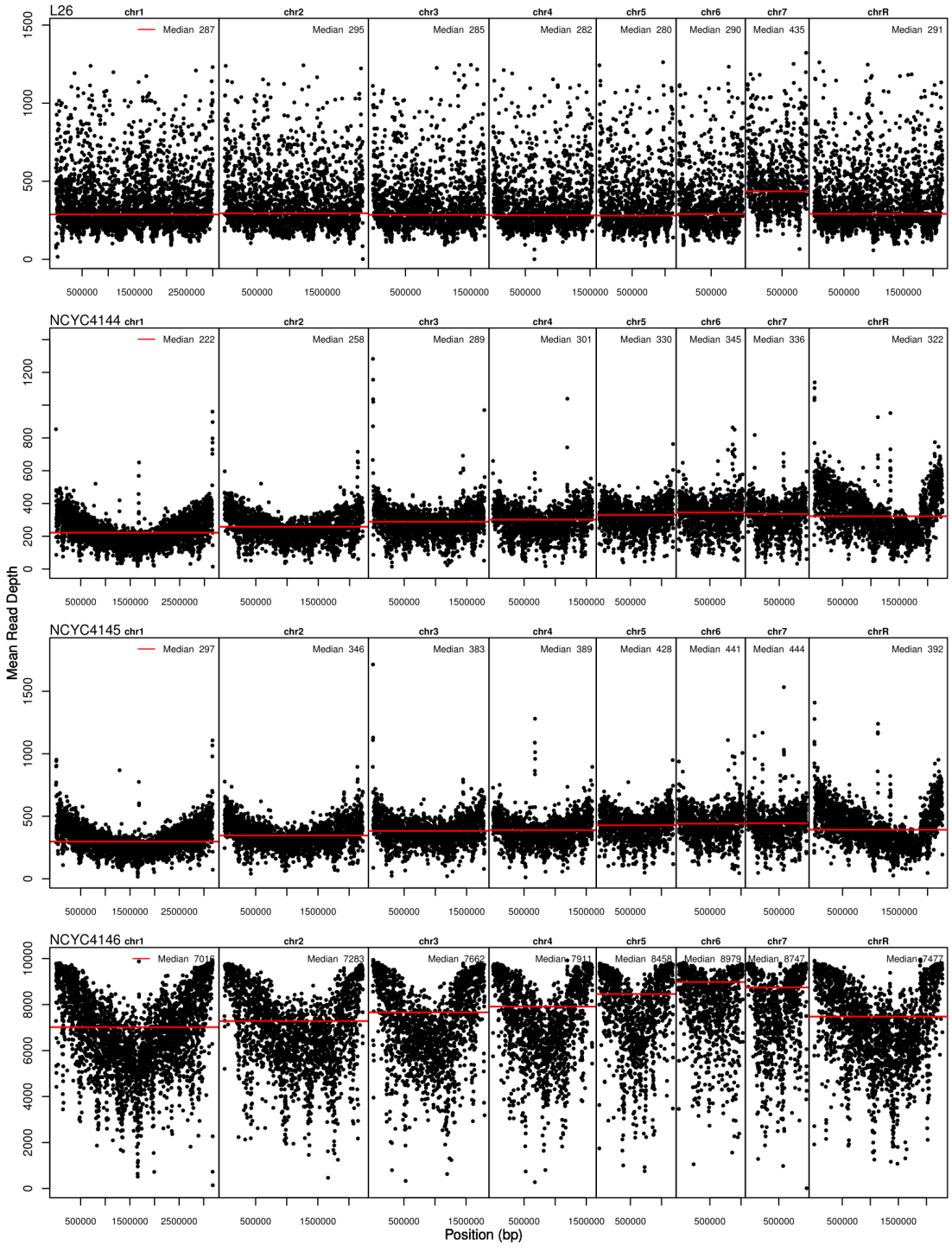
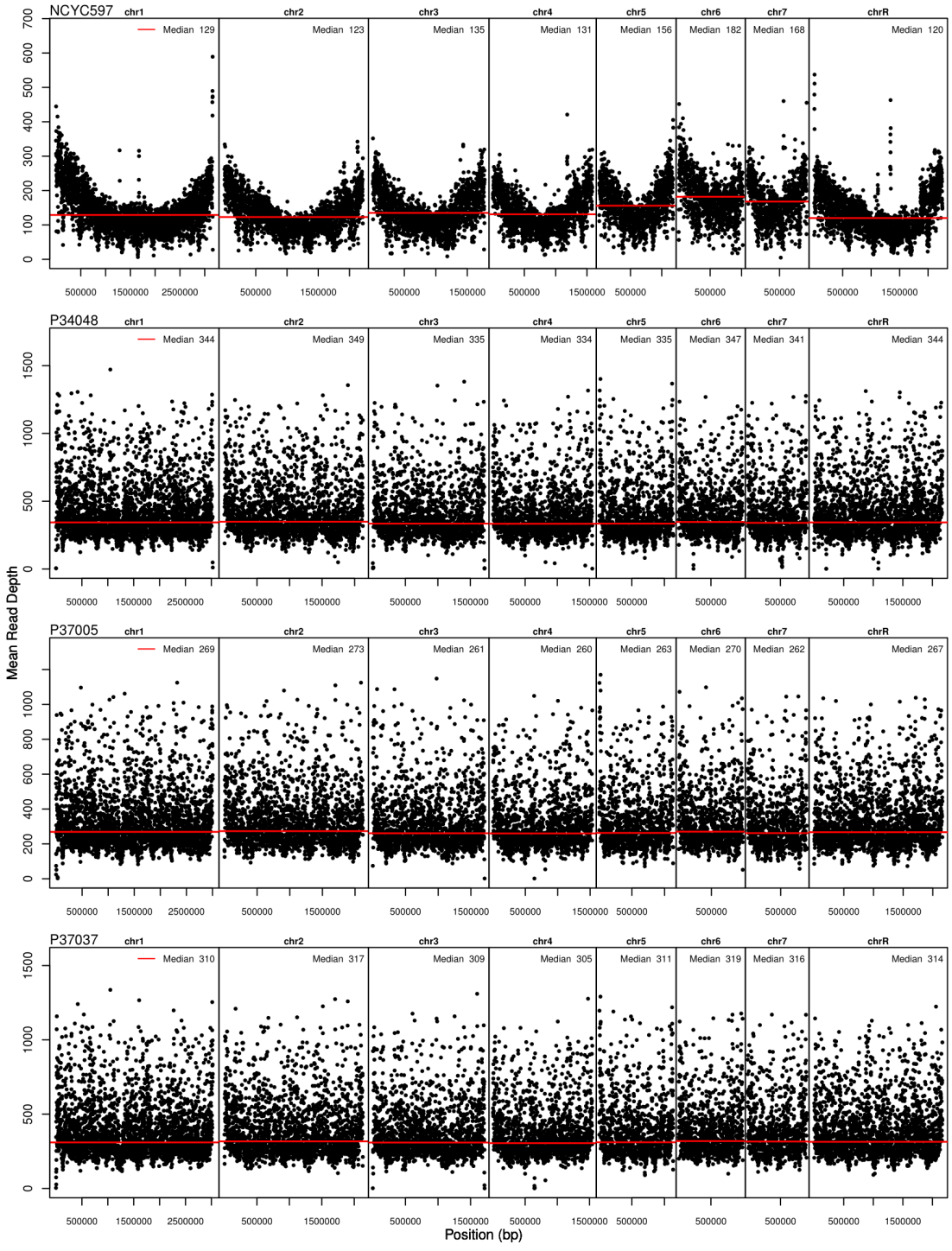


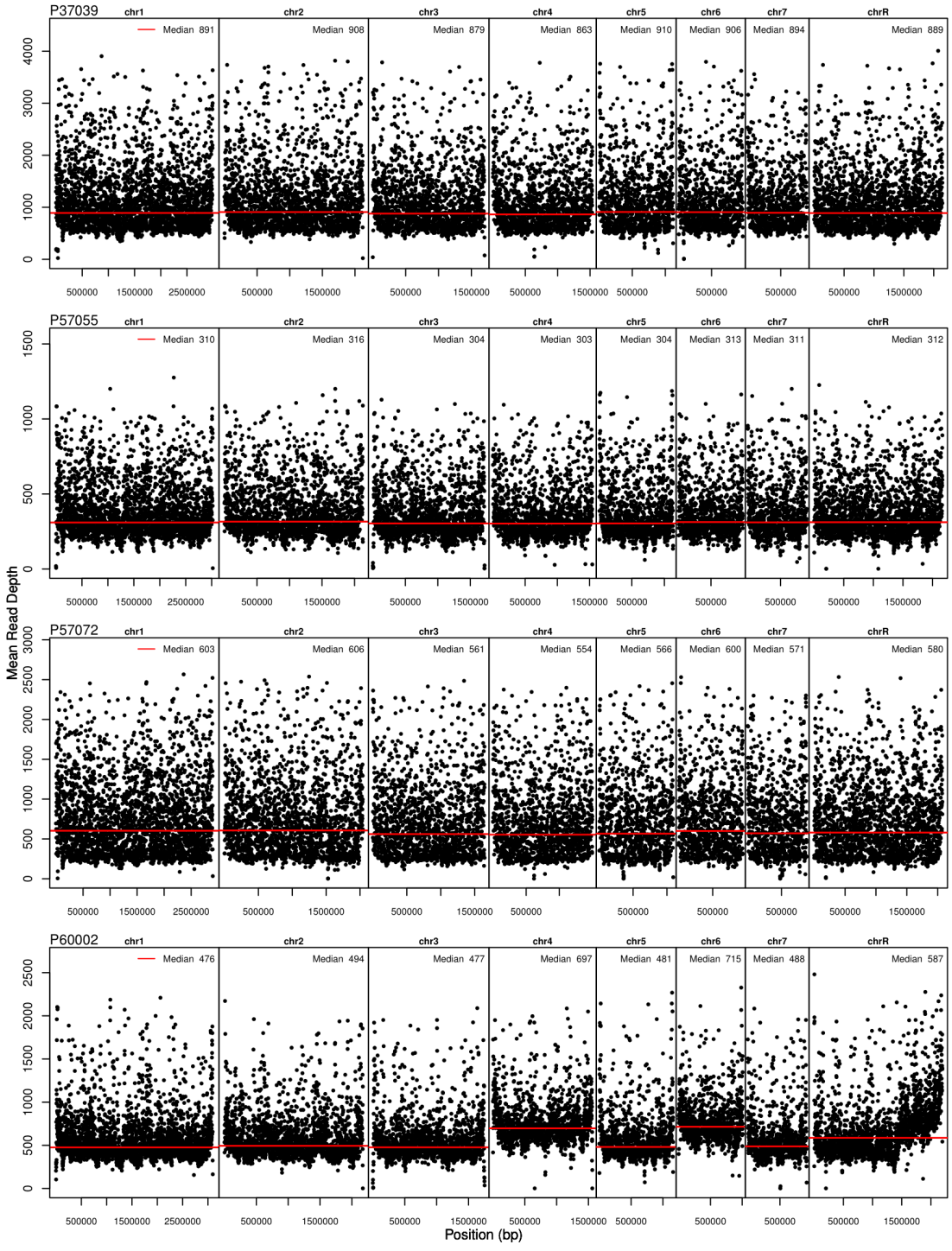

Figure S 3: **The proportion of bases differing for within-clade pairwise comparisons is lower than the pairwise divergences between clades.** The DNA sequence of each strain was compared to every other in 100 kb windows and divergence was estimated as the proportion of sites that differ in high quality sequence (phred-scaled quality over 40). Within-clade divergences are shown in blue, between-clade divergences are in purple. Most within-clade divergences (90%) are below 0.028% (green line) while most between clade divergences are above it. In cases where sequence divergence between sequences is above a threshold of 0.028% chromosomes were painted white in Figures 3b and S5 to show that they were too diverged from other sequences for clade assignment.

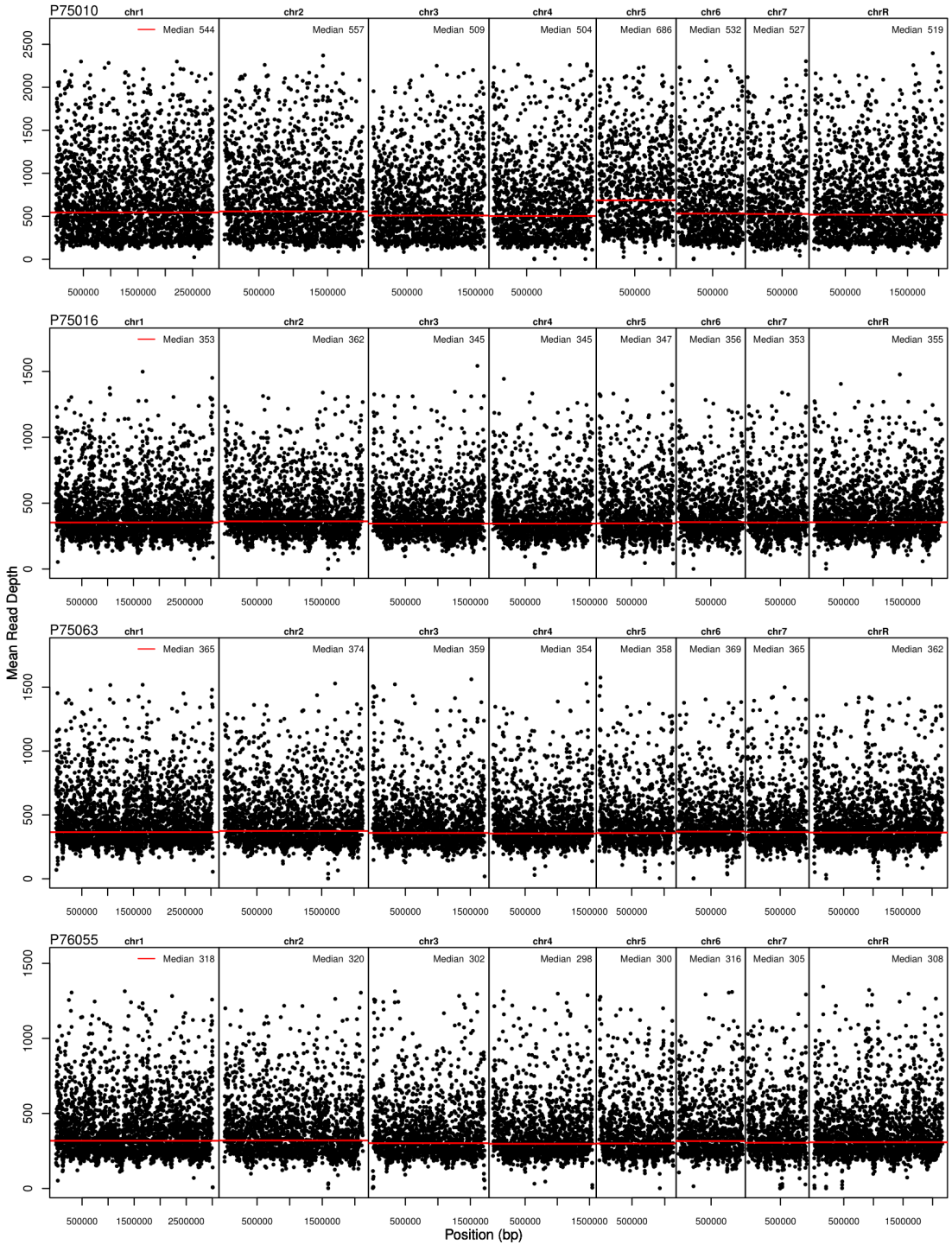
Figure S 4: **Read depth analysis comparison between chromosomes for assessing aneuploidy.** Mean read depth was calculated using SAMtools depth (version 1.3.1; Li *et al.*, 2009), then R was used to estimate mean read depth in non-overlapping 1 kb windows. Mean read depth values over 5 times the median are not shown. For strains where DNA was fragmented using physical methods, aneuploidy (12C, 19F, L26, P60002, P75010, P78042) could be distinguished from euploidy (GC75, P34048, P37005, P37037, P37039, P57055, P57072, P75016, P75063, P76055, P78048, P87, P94015, SC5314, 1AA) as discrete jumps in pairwise comparisons of median values between chromosomes (with differences over 35%). Chromosomes that are homozygous in analyses of base calls (the B-allele frequency approach in Figures 1a and S1) do not show lower read depth and are therefore not monosomic, but have probably undergone LOH. Read depth was uneven within and between chromosomes for the data generated as part of this study (NCYC 597, NCYC 4144, NCYC 4145, NCYC 4146) as expected for a transposase-based DNA fragmentation protocol (Marine *et al.*, 2011; Quail *et al.*, 2012; Teo *et al.*, 2012). We therefore do not rely on read depth analysis for determining aneuploidy in these strains.

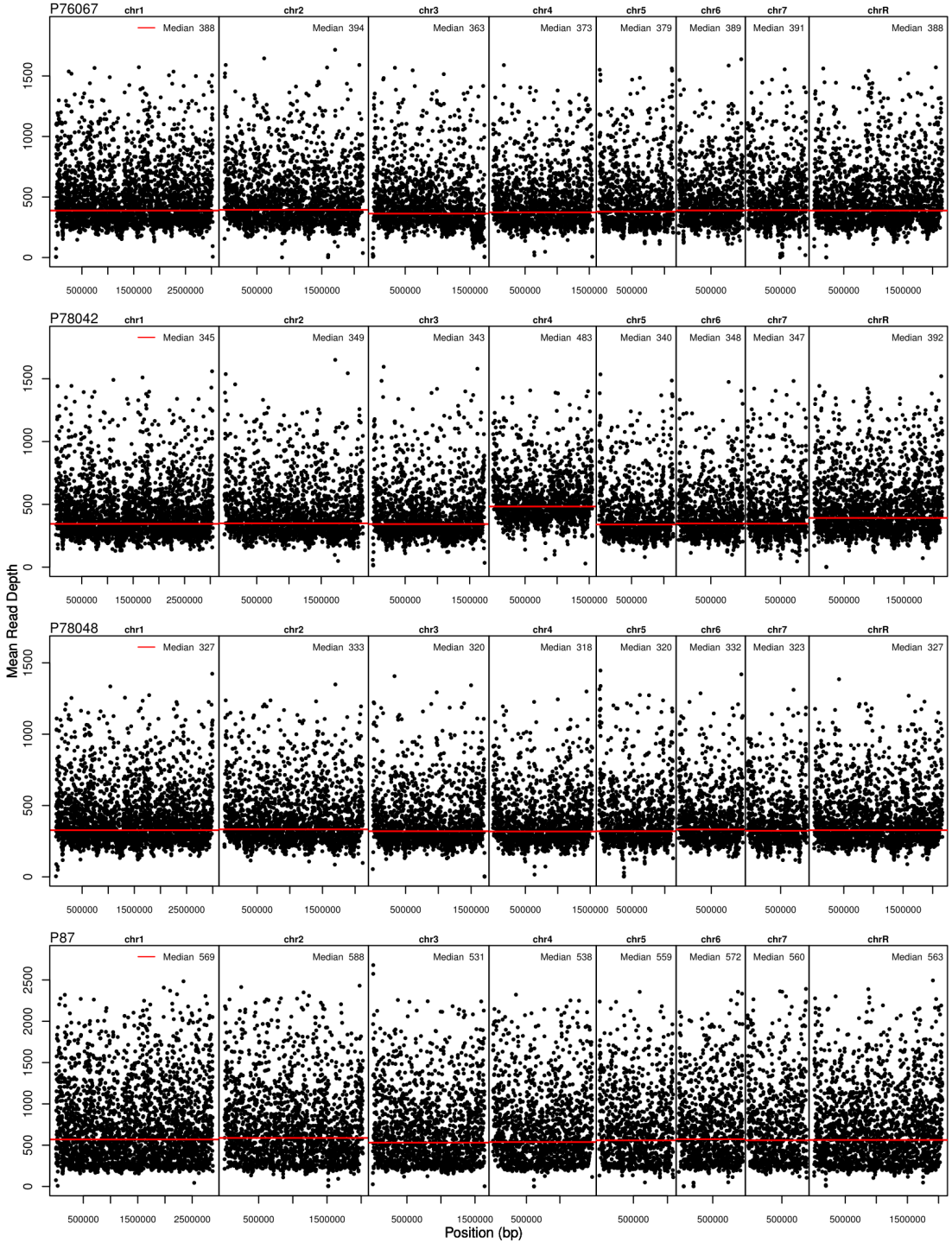



L26
chr1 chr2 chr3 chr4 chr5 chr6 chr7 chrR
Median 287
Median 295
Median 285
Median 282
Median 280
Median 290
Median 435
Median 291
NCYC4144
Median 222
Median 258
Median 289
Median 301
Median 330
Median 345
Median 336
Median 322
NCYC4145
Median 297
Median 346
Median 383
Median 389
Median 428
Median 441
Median 444
Median 392
NCYC4146
Median 7016
Median 7283
Median 7662
Median 7911
Median 8458
Median 8979
Median 8747
Median 7477
Mean Read Depth
Position (bp)


NCYC597
chr1 chr2 chr3 chr4 chr5 chr6 chr7 chrR
Median 129
Median 123
Median 135
Median 131
Median 156
Median 182
Median 168
Median 120
P34048
Median 344
Median 349
Median 335
Median 334
Median 335
Median 347
Median 341
Median 344
P37005
Median 269
Median 273
Median 261
Median 260
Median 263
Median 270
Median 262
Median 267
P37037
Median 310
Median 317
Median 309
Median 305
Median 311
Median 319
Median 316
Median 314
Mean Read Depth
Position (bp)


P37039
chr1
chr2
chr3
chr4
chr5
chr6
chr7
chrR
Median 891
Median 908
Median 879
Median 863
Median 910
Median 906
Median 894
Median 889
P57055
Median 310
Median 316
Median 304
Median 303
Median 304
Median 313
Median 311
Median 312
P57072
Median 603
Median 606
Median 561
Median 554
Median 566
Median 600
Median 571
Median 580
P60002
Median 476
Median 494
Median 477
Median 697
Median 481
Median 715
Median 488
Median 587
Mean Read Depth
Position (bp)


P75010
chr1
chr2
chr3
chr4
chr5
chr6
chr7
chrR
Median 544
Median 557
Median 509
Median 504
Median 686
Median 532
Median 527
Median 519
P75016
Median 353
Median 362
Median 345
Median 345
Median 347
Median 356
Median 353
Median 355
P75063
Median 365
Median 374
Median 359
Median 354
Median 358
Median 369
Median 365
Median 362
P76055
Median 318
Median 320
Median 302
Median 298
Median 300
Median 316
Median 305
Median 308
Mean Read Depth
Position (bp)


P76067
chr1 chr2 chr3 chr4 chr5 chr6 chr7 chrR
Median 388
Median 394
Median 363
Median 373
Median 379
Median 389
Median 391
Median 388
P78042
Median 345
Median 349
Median 343
Median 483
Median 340
Median 348
Median 347
Median 392
P78048
Median 327
Median 333
Median 320
Median 318
Median 320
Median 332
Median 323
Median 327
P87
Median 569
Median 588
Median 531
Median 538
Median 559
Median 572
Median 560
Median 563
Mean Read Depth
Position (bp)

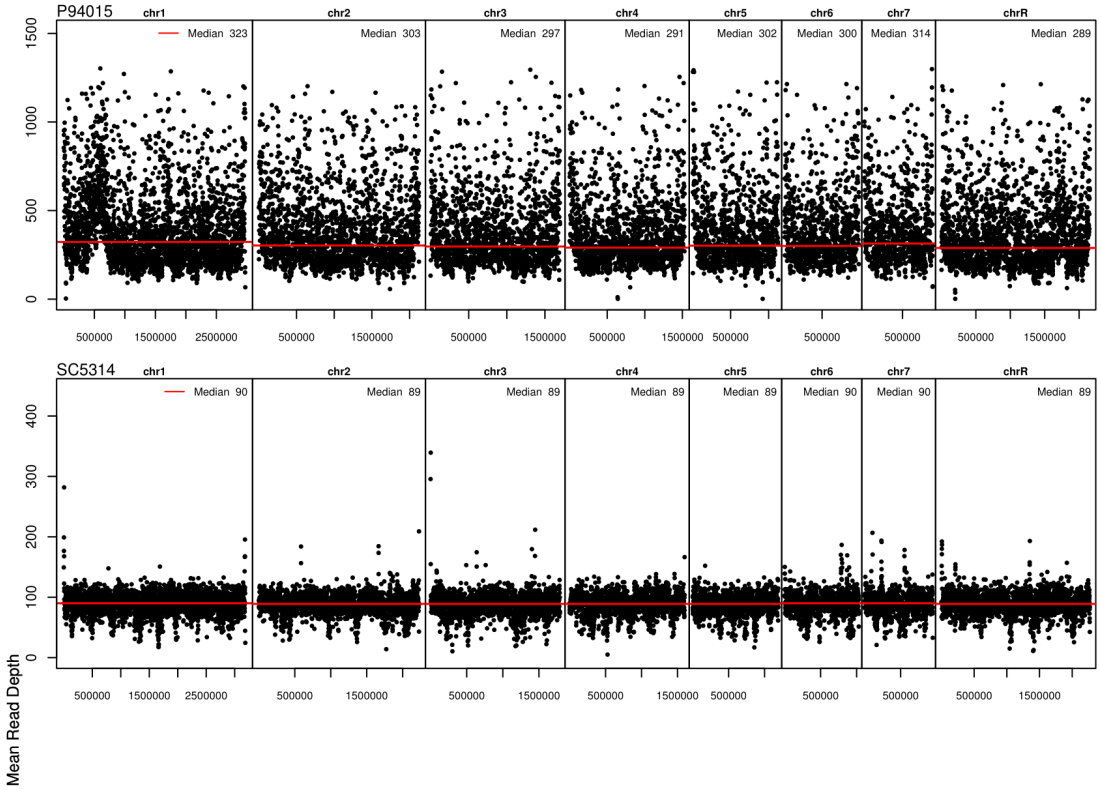

P94015
chr1
chr2
chr3
chr4
chr5
chr6
chr7
chrR
Median 323
Median 303
Median 297
Median 291
Median 302
Median 300
Median 314
Median 289
SC5314
Median 90
Median 89
Median 89
Median 89
Median 89
Median 90
Median 90
Median 89
Mean Read Depth
Position (bp)

Figure S 5: **Oak strains are no more similar to animal strains from Illinois than they are to clinical strains.** Phylogenetic analysis of MLST sequences shows the relationships between clinical strains, oak strains and animal strains. Strains from animals are prefixed with "ST". Sequence types and clade assignments for domestic and wild animals were determined by Wrobel *et al.* (2008) and sequences were downloaded from http://pubmlst.org/calbicans/.

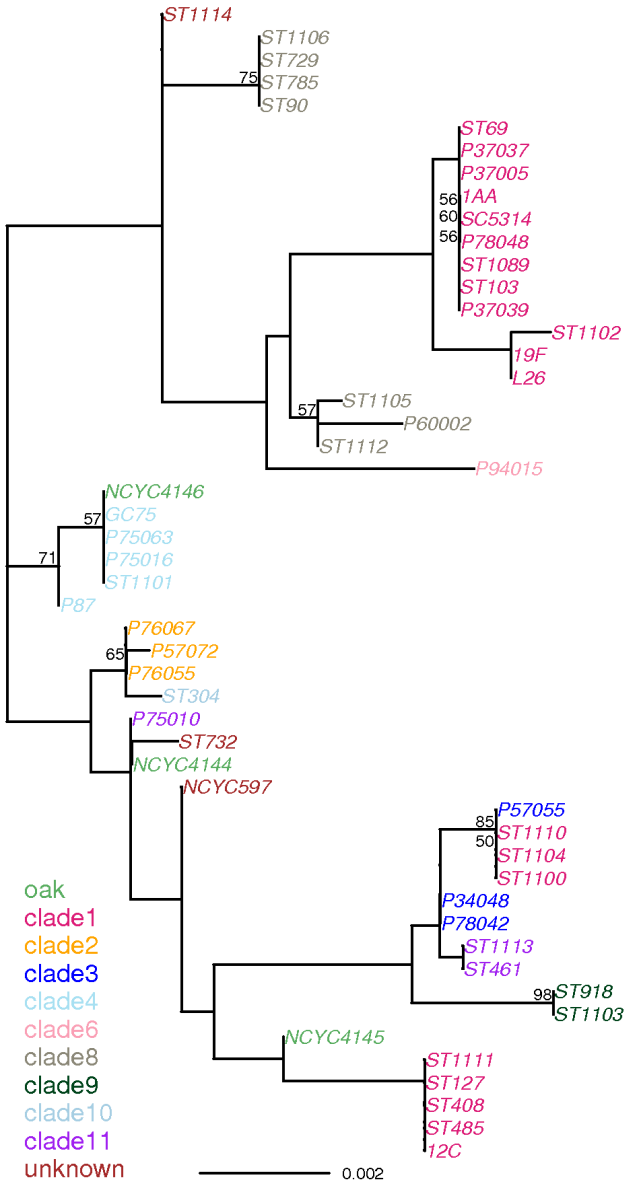
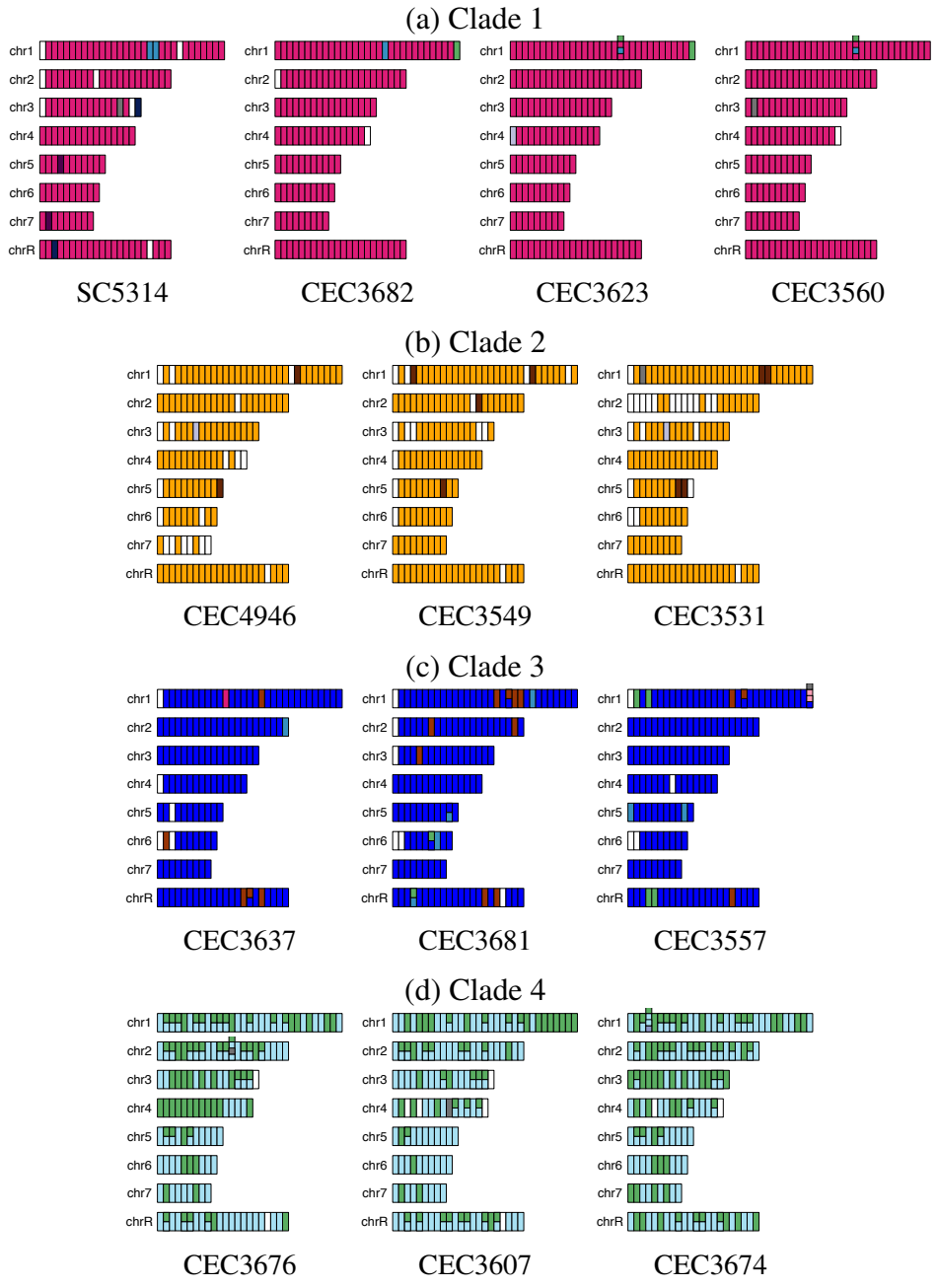
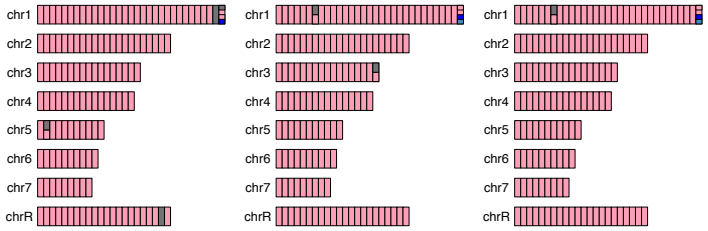


Figure S 6: **Chromosome painting for 59 clinical strain genomes from Ropars *et al.* (2018)**. The genome of each strain was subdivided into 100 kb non-overlapping sliding windows. Each window was coloured according to the clade assignment of the most similar strain. Clade assignments were made for clinical strains by Ropars *et al.* (2018), and in cases where the most similar strain was an oak strain the window was coloured green. In the case of ties, windows were split and coloured with the colour representing the equally similar strains (e.g. light blue and green for clade 4). Regions are coloured white if a strain sequence is diverged from all the other oak or clinical strains that we sampled (the proportion of sites differing is over 0.028%). Similarity to the type strain was not used for clade assignment because this strain is from an unknown clade.

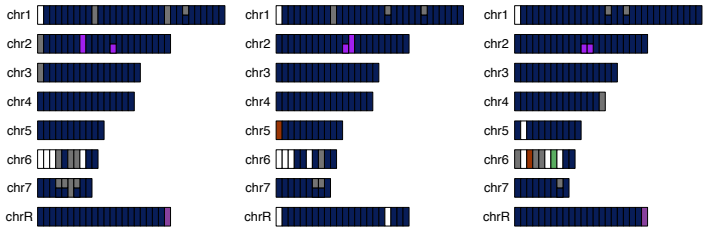


(e) Clade 8



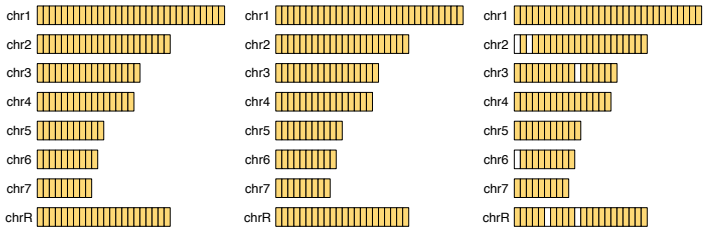

CEC2023 CEC3622 CEC3634

(f) Clade 9



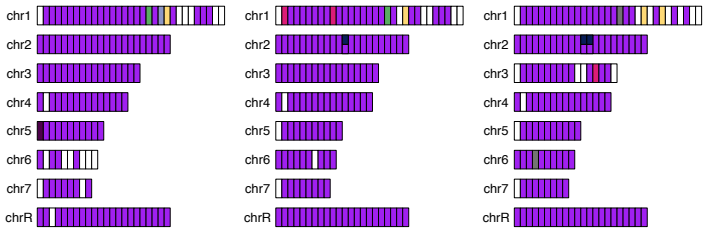

CEC2018 CEC3533 CEC5026

(g) Clade 10



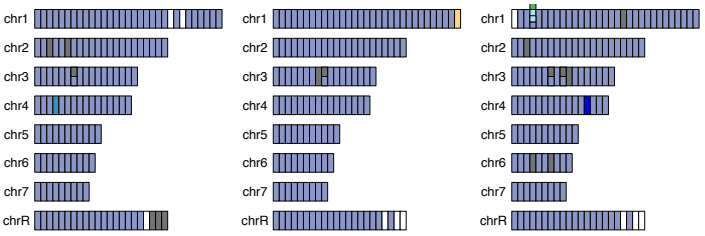

CEC3616 CEC3706 CEC3711

(h) Clade 11




CEC4495 CEC3618 CEC3704

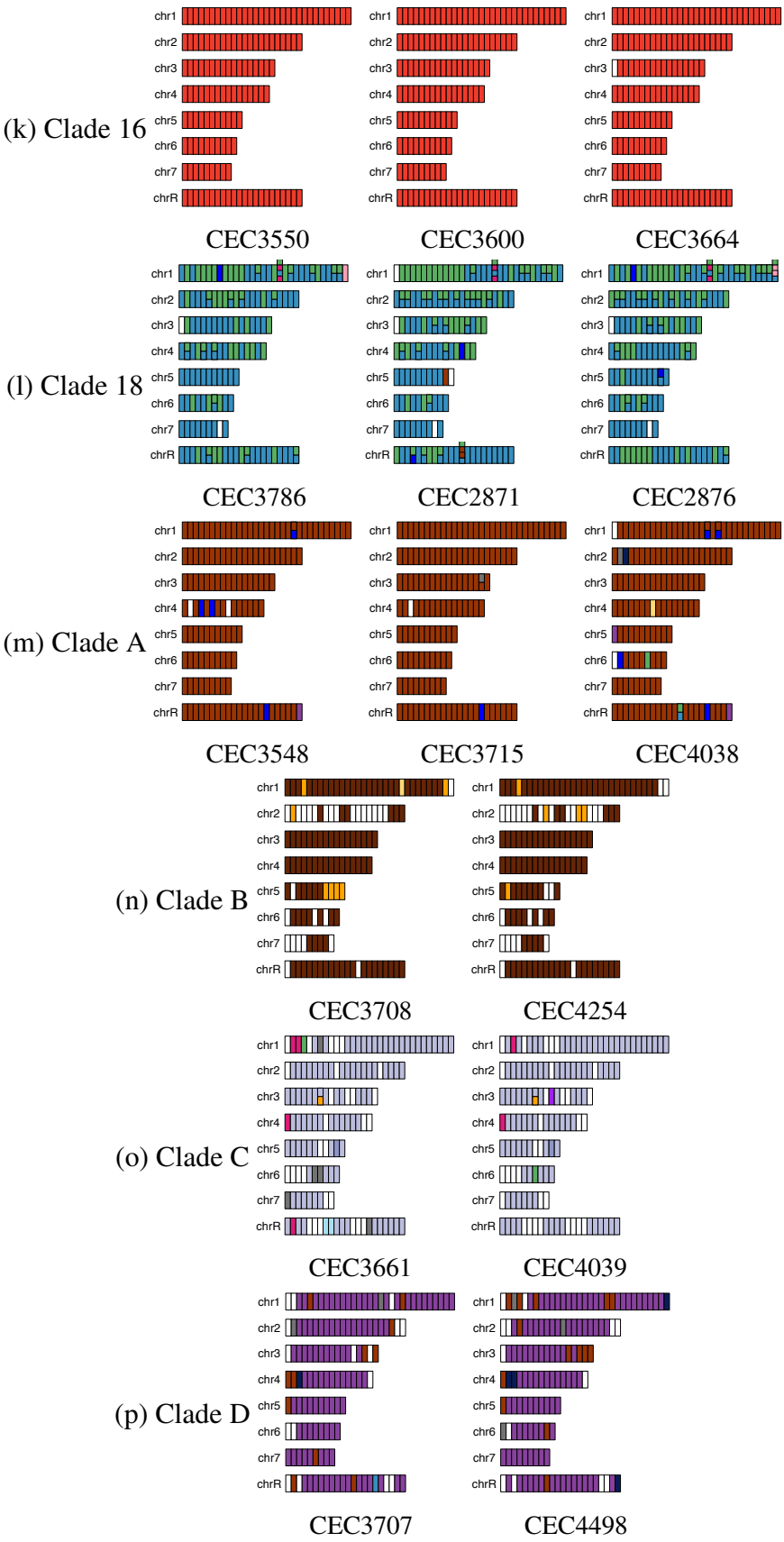
(i) Clade 12



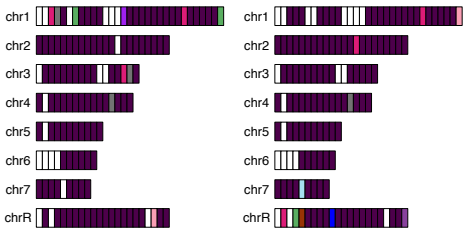

CEC3680 CEC2019 CEC3537

(j) Clade 13


CEC4877 CEC5020 CEC5019

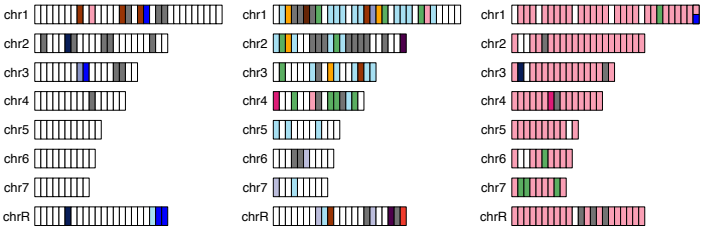

(k) Clade 16
CEC3550
CEC3600
CEC3664
(l) Clade 18
CEC3786
CEC2871
CEC2876
(m) Clade A
CEC3548
CEC3715
CEC4038
(n) Clade B
CEC3708
CEC4254
(o) Clade C
CEC3661
CEC4039
(p) Clade D
CEC3707
CEC4498
chr1
chr2
chr3
chr4
chr5
chr6
chr7
chrR

(q) Clade E

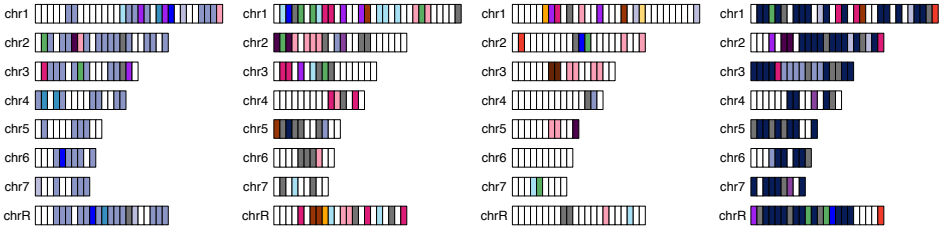



CEC3712 CEC4486

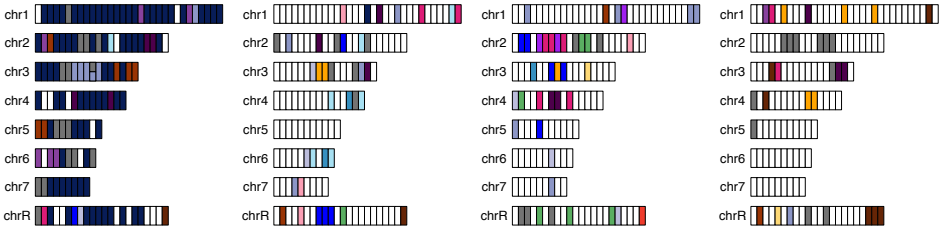
(r) Unknown clades




NCYC597 CEC3555 CEC3613




CEC3619 CEC3663 CEC3665 CEC3671




CEC4261 CEC4481 CEC4485 CEC723

# References


Hickman, M.A., Zeng, G., Forche, A., Hirakawa, M.P., Abbey, D., Harrison, B.D., Wang, Y.M., Su, C.h., Bennett, R.J., Wang, Y., and Berman, J. (2013). "The 'obligate diploid' *Candida albicans* forms mating-competent haploids." *Nature*, **494**(7435): 55.

Hirakawa, M.P., Martinez, D.A., Sakthikumar, S., Anderson, M.Z., Berlin, A., Gujja, S., Zeng, Q., Zisson, E., Wang, J.M., Greenberg, J.M., Berman, J., Bennett, R.J., and Cuomo, C.A. (2015). "Genetic and phenotypic intra-species variation in *Candida albicans*." *Genome Research*, **25**(3): 413–425.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). "The Sequence Alignment/Map format and SAMtools." *Bioinformatics*, **25**(16): 2078–9.

Li, X., Yang, F., Li, D., Zhou, M., Wang, X., Xu, Q., Zhang, Y., Yan, L., and Jiang, Y. (2015). "Trisomy of chromosome R confers resistance to triazoles in *Candida albicans*." *Medical Mycology*, **53**(3): 302–309.

Marine, R., Polson, S.W., Ravel, J., Hatfull, G., Russell, D., Sullivan, M., Syed, F., Dumas, M., and Wommack, K.E. (2011). "Evaluation of a Transposase Protocol for Rapid Generation of Shotgun High-Throughput Sequencing Libraries from Nanogram Quantities of DNA." *Applied and Environmental Microbiology*, **77**(22): 8071–8079.

Muzzey, D., Schwartz, K., Weissman, J.S., and Sherlock, G. (2013). "Assembly of a phased diploid *Candida albicans* genome facilitates allele-specific measurements and provides a simple model for repeat and indel structure." *Genome Biology*, **14**(9): R97.

Quail, M.A., Smith, M., Coupland, P., Otto, T.D., Harris, S.R., Connor, T.R., Bertoni, A., Swerdlow, H.P., and Gu, Y. (2012). "A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers." *BMC Genomics*, **13**: 341.

Ropars, J., Maufrais, C., Diogo, D., Marcet-Houben, M., Perin, A., Sertour, N., Mosca, K., Permal, E., Laval, G., Bouchier, C., Ma, L., Schwartz, K., Voelz, K., May, R.C., Poulain, J., Battail, C., Wincker, P., Borman, A.M., Chowdhary, A., Fan, S., *et al.* (2018). "Gene flow contributes to diversification of the major fungal pathogen Candida albicans." *Nature Communications*, **9**(1): 2253.

Teo, S.M., Pawitan, Y., Ku, C.S., Chia, K.S., and Salim, A. (2012). "Statistical challenges associated with detecting copy number variations with next-generation sequencing." *Bioinformatics*, **28**(21): 2711–2718.

Todd, R.T., Forche, A., and Selmecki, A. (2017). "Ploidy Variation in Fungi: Polyploidy, Aneuploidy, and Genome Evolution." *Microbiology Spectrum*, **5**(4).

Wrobel, L., Whittington, J.K., Pujol, C., Oh, S.H., Ruiz, M.O., Pfaller, M.A., Diekema, D.J., Soll, D.R., and Hoyer, L.L. (2008). "Molecular Phylogenetic Analysis of a Geographically and Temporally Matched Set of *Candida albicans* Isolates from Humans and Nonmigratory Wildlife in Central Illinois." *Eukaryotic Cell*, **7**(9): 1475–1486.