Supplementary material

Is optimal gene order impossible?

Juan F. Poyatos[1] and Laurence D. Hurst[2]

[1]Evolutionary Systems Biology Initiative, Structural and Computational Biology Programme, Spanish National Cancer Centre (CNIO), Melchor Fernández Almagro 3, 28029 Madrid, Spain
[2]Department of Biology and Biochemistry, University of Bath, Somerset BA2 7AY, UK
Corresponding authors: Poyatos, J.F. (jpoyatos@cnio.es); Hurst, L.D. (l.d.hurst@bath.ac.uk).

This supplementary material accompanies the article by Poyatos and Hurst published in the August issue of *Trends in Genetics*.

Data sources

Our analysis used mainly the database of interacting proteins (DIP) [1]. This data set had 4741 interacting proteins combining a variety of sources and it is curated both manually and automatically. From this set, we considered only those proteins for which we also had expression data (see below), and filtered out proteins interacting only with themselves. To assess the reliability of the patterns found, we also used a data set of protein interactions assembled by Yu *et al* [2], a merge of several protein-protein interaction data sets (we termed it TopNet data set). We used public expression data compiled by Stuart Kim laboratory (a data set of 6209 yeast genes and 639 experiments [3]). We normalized the raw data set by subtracting the median value for each experiment and then dividing by the interquartile range (difference between the 75th percentile and the 25th percentile). We used genomic data available in the Munich Information Center for Protein Sequences [4].

Stable complexes

We considered a data set of complexes annotated by the Munich Information Center for Protein Sequences (MIPS, complexcat-data-28102004.htm [4]). It includes manually annotated complexes and complexes obtained by systematic analyses representing high-throughput experiments. We used complexes with sizes ranging from 2 to 35 components for our calculations. We divided these complexes into stable (permanent), having a particularly strong co-expression pattern, and transient (following the method by Jansen R. *et al.* [5], see also Teichmann [6]). In short, we generated null distributions of average pairwise co-expression for all constituents of a complex of a given size and determined whether a MIPS complex is permanent or transient according to such null distributions. Permanent complexes would clearly contribute to increase the co-expression signal in the clustering analysis so we control for these type of complexes specifically. We further clarify this by computing the mean genomic distance of constituents of stable (permanent) complexes when they are found on the same chromosome versus that of the transient complexes. Complex 'stability' clearly contributes to its genomic linkage ($d_{\text{permanent}}$ = 271 450 bp, P <0.0001; $d_{\text{transient}}$ = 335 020 bp, P = 0.39) and supports the classification used to control complexes. In the genome linkage study, we split the stable set into two (small complexes, <15 components, and large complexes, ≥15 components). To examine the effects of complexes in the proximity analyses, we randomly set the connectivity of one of an adjacent pair to 0 when both proteins were part of at least one common complex. We computed all related measures with this new set as before.

Double-strand break data

We used the recombination data set by Gerton *et al* [7], an estimate of recombination rate by using double strand break analysis. This data provides very high resolution coverage of recombination rates. However, is this data reliable? For example, can it adequately find regions of low and high recombination? And how does it compare with observed recombination rates? To address the first issue, we asked whether the double strand break data reports reduced recombination rates in the vicinity of centromeres, these being known to be regions of reduced recombination rates. To this end, we constructed sliding window plots of mean double strand break rates averaged over ten successive genes and plotted the midpoint of each window (defined as the mean of the start of the first gene and the end of the last gene) against the mean double strand break rate. We then examined the plots to determine whether the centromeres

were repeatedly reported to be at dips in the recombination rate. The centromere positions were obtained from: http://db.yeastgenome.org/cgi-bin/locus.pl?locus=cenN, where N is the chromosome number. In all cases the centromeres are in regions of reduced double strand break density (see Figures S.1–4). Moreover, in nearly all instances the centromere has a double strand break rate more than one standard deviation under the mean. The one possible exception is chromosome 8 where the centromere appears to be positioned on the edge of very sharp declines in recombination. This confirms that the double strand break measure captures domains of reduced recombination.

To ask whether the double strand break data accurately recovers observed recombination rates, we compared the observed recombination rate between two markers per kilobase, with the double strand break rate per kilobase. We obtained the data on recombination rates from the yeast Motimer maps via ftp://genome-ftp.stanford.edu/pub/yeast/data_download/chromosomal_feature/. This provides measures of the recombination rates between any two markers but does not provide the same high level of resolution as double strand break data. For all pairs of markers in the genetic distance set, if both are present in the double strand break data we compare the two rates (normalised to per kilobase rates). When a given pair occurs more than once in the data set, we consider the average recombination rate. Using pairs that are ≥10 kb apart (so as to reduce the level of noise owing to inaccuracy between very close markers) this then resolves to 843 comparisons between genetic distance and double strand break distance. The two data sets robustly agree ($r = 0.35$, $P < 0.0001$; Figure S.5).

Randomization

We identified the location of all interacting proteins in the genome. We assessed significance to the different measures relating genome proximity with network proximity by generating 10 000 new data sets in which these locations are randomly associated to the interacting proteins. In this way, we avoided any genomic regionality associated to the interacting proteins (e.g. the fact that essential genes cluster could provide one possible force for such regionality) if essential genes tend to be interacting genes. The generalized clustering coefficient is defined as:

$$C_{ij} = | Adj(i) \cap Adj(j) | / \min(|Adj(i)|, | Adj(j)|).$$

Here, $|\ldots|$ denotes the size of the set, $\cap$ the intersection and $Adj(i)$ the adjacency matrix (i.e. the set of proteins interacting with protein i). Alternative definitions did not change the significance of the results. We computed the Pearson correlation of co-expression and the mean of recombination rates between all pairs of adjacent genes. To determine the significance of the instances of positive co-expression and mean recombination rate we randomly sample from this set a subset of size that of the number of adjacent pairs with network distance $d < 3$.

Module types

We introduced three types of modular structures to probe the network intermediate level of organization, motifs, SPC modules and overlap cores. Motifs have been recently introduced as the fundamental building blocks of biological networks [8]. We considered here only those motifs that are fully connected graphs, or cliques, so we can extend the search to large motifs. The search starts with the smallest one, size 3, and increases in size until cliques of all sizes are found. Note that we considered all available subgraphs not only those that are overrepresented [9]. Redundant cliques (i.e. those whose constituents are all constituents of the clique next in size) were removed. As a second module type, we considered super-paramagnetic modules. Super-paramagnetic clustering (SPC) has been recently introduced as new approach for clustering data, based on the physical properties of an inhomogeneous ferromagnetic model [10]. More recently, this algorithm has been applied to extract modules in protein interaction networks [9]. We used this algorithm to define SPC modules (see Spirin and Mirny [9] for details). As a third module type, we considered overlap cores. These structures are extracted by using a graph-based algorithm that we have recently introduced [11]. To validate the biological relevance of these structures the algorithm uses information from the phylogenetic conservation of the network components [11].

Tandem duplication

We used BLASTP with standard parameters and searched all yeast proteome against itself. Duplicates are those pairs with *E*-values <*E*-threshold. Tandem duplicates are those adjacent pairs in the genome with *E*-values <*E*-threshold (we considered different BLASTP *E*-values

thresholds: 10^{-20}, 10^{-10}, 10^{-5}; we used 10^{-5} in all data shown). In the proximity analyses, for all pairs of interacting proteins which are tandem duplicates, we randomly set the connectivity of one of them to 0. We then generated a new interacting matrix by considering the interactions among the selected set. We computed the graph distances again, clustering coefficients and the corresponding mean measures.

Conservation of adjacent pairs

To examine the conservation of adjacent pairs, we compared gene order in *Saccharomyces cerevisae* with that of *Kluyveromyces lactis* (available at http://www.ebi.ac.uk/genomes/eukaryota.html). We carried out similarity searches of both proteomes with BLASTP using standard parameters. Non-overlapping adjacent pairs in *S. cerevisae* with homologs with reciprocal-best hits and *E*-value < 10^{-10} in *K. lactis* were retained.

Table S1. Tandem duplicate analysis

Measure	Significance
Shortest distance	$P = 0.0199$
Adjacent genes with $d<3$	$P < 0.0001$
Generalized clustering coefficient	$P < 0.0001$

We eliminated one of the two genes from a tandem pair and re-assembled a new network. We then computed the following tests: (i) shortest distance between pairs of proteins in the interaction network whose genes are adjacent in the genome; (ii) number of close proteins in the network (d <3) whose genes are adjacent in the genome; and (iii) generalized clustering coefficient between pairs of proteins in the interaction network whose genes are adjacent in the genome. All three tests remained significant.

Table S2. Protein complex analysis

Measure	Significance
Shortest distance	$P = 0.0293$
Adjacent genes with d <3	$P < 0.0001$
Generalized clustering coefficient	$P < 0.0001$

We eliminated one of the two genes from a genomically neighouring pair if both proteins belonged to the same stable complex and re-assembled a new network. We then computed the following tests: (i) shortest distance between pairs of proteins in the interaction network whose genes are adjacent in the genome; (ii) number of close proteins in the network (d <3) whose genes are adjacent in the genome; and (iii) generalized clustering coefficient between pairs of proteins in the interaction network whose genes are adjacent in the genome. All three tests remained significant.

Table S3. Protein complex and tandem duplicate analysis

Measure	Significance
Shortest distance	$P = 0.1093$
Adjacent genes with $d<3$	$P < 0.0001$
Generalized clustering coefficient	$P < 0.0001$

We eliminated one of the two genes from a genomically neighouring pair if both proteins belonged to the same stable complex and one of the two genes from a tandem pair. We re-assembled a new network. We then computed the following tests: (i) shortest distance between pairs of proteins in the interaction network whose genes are adjacent in the genome; (ii) number of close proteins in the network (d <3) whose genes are adjacent in the genome; and (iii) generalized clustering coefficient between pairs of proteins in the interaction network whose genes are adjacent in the genome. Two out of the three tests remained significant. The third one is close to be significant.

Reliability analysis

Table S4. Examination of a subset of the DIP interactions

Measure	Significance
Shortest distance	$P = 0.0037$
Adjacent genes with $d<3$	$P < 0.0001$
Generalized clustering coefficient	$P < 0.0001$
We examined a subset of the DIP interactions believed to be correct. This core data set (downloaded from http://dip.doe-mbi.ucla.edu/) is identified by merging several interacting sets assessed computationally. We then computed the following tests: (i) shortest distance between pairs of proteins in the interaction network whose genes are adjacent in the genome; (ii) number of close proteins in the network (d <3) whose genes are adjacent in the genome; and (iii) generalized clustering coefficient between pairs of proteins in the interaction network whose genes are adjacent in the genome. All three tests remained significant.	

We analysed the contribution of tandem pairs and stable complexes for the DIP-core network similarly to the full DIP network. Note that we include data from randomization as random values, in the case of adjacent genes with d <3, or Z-scores (negative for the distance and positive for the clustering coefficient). The decrease of significance in Table S4c might be partially associated with the low statistical power owing to the smaller network size with respect to the corresponding networks associated with Table S4a and Table S4b.

Table S4a. Tandem duplicate analysis

Measure	Observed	Random	Significance
Shortest distance	4.2697	-2.4	$P = 0.0079$
Adjacent genes with d <3	27	18.4 ± 4.2	$P = 0.0325$
Generalized clustering coefficient	0.0138	2.7	$P = 0.0070$
Equivalent to Table S1 with DIP-core data.			

Table S4b. Protein complex analysis

Measure	Observed	Random	Significance
Shortest distance	4.2774	-2.1	$P = 0.0179$
Adjacent genes with $d<3$	29	18.3 ± 4.2	$P = 0.0112$
Generalized clustering coefficient	0.0165	3.9	$P = 0.0008$
Equivalent to Table S2 with DIP-core data.			

Table S4c. Protein complex and tandem duplicate analysis

Measure	Observed	Random	Significance
Shortest distance	4.2906	-1.5	$P = 0.0691$
Adjacent genes with $d<3$	21	18.1 ± 4.2	$P = 0.2762$
Generalized clustering coefficient	0.0111	1.5	$P = 0.0741$
Equivalent to Table S3 with DIP-core data.			

Table S5. Robustness of co-expression and recombination rates against tandem duplication and stable protein complexes

	Co-expression			Recombination		
	Observed	**Random**	***P***	**Observed**	**Random**	***P***
Duplication	124	110.9 ± 5.7	0.0071	1.0880	1.05± 0.01	0.0203
Complexes	126	113.6 ± 5.8	0.0112	1.0864	1.05± 0.02	0.0230

We examined whether the patterns of co-expression and recombination rates were sensitive to exclusion of tandem duplicates or complex sharing (see main text for details).

We used four different data sets to further assess the high co-expression/high mean recombination pattern: Dip1: DIP data set; Dip2: DIP-core (see previous description); Int1: the so-called TopNet data set [4]; and Int2: a subset of the DIP proteins that are close(far) and adjacent in the genome and are also found as close(far) and adjacent using the TopNet (Int1) data set.

Table S6a. High co-expression pattern of close (*d* <3) and adjacent genes [corresponding to Figure 1a (main text)]

		Co-expression		
Database	**Observed**	**Random**	***P*-value**	**# close/far**
Dip1: DIP	126	113.5 ± 5.9	0.0226	167/3304
Dip2: DIP-core	30	25.2 ± 2.7	0.0518	37/935
Int1: TopNet4	318	295.7 ± 8.4	0.0048	441/1123
Int2: DIP/TopNet	79	70.5 ± 4.6	0.0393	107/1024

Observed values correspond to the number of close and adjacent genes with positive co-expression. The second and third columns denote the corresponding null and P-values. The final column is the number of close and far (d >2) adjacent genes.

Table S6b. High recombination pattern of close (*d* <3) and adjacent genes [corresponding to Figure 1b (main text)]

		Recombination		
Database	**Observed**	**Random**	**P-value**	**# close/far**
Dip1: DIP	126	113.5 ± 5.9	0.0226	167/3304
Dip2: DIP-core	30	25.2 ± 2.7	0.0518	37/935
Int1: TopNet[4]	318	295.7 ± 8.4	0.0048	441/1123
Int2: DIP/TopNet	79	70.5 ± 4.6	0.0393	107/1024

Observed values correspond to mean recombination rates of close and adjacent genes. The second and third columns denote the corresponding null and *P*-values. The final column is the same as the previous table.

Table S7. A list of protein pairs found close (*d* <3) in the network and adjacent in the genome using the DIP-based network

YAL041W	YAL040C	YGL049C	YGL048C	*YLL051C	YLL050C
YAR018C	YAR019C	*YGL048C	YGL047W	YLL040C	YLL039C
*YAR030C	YAR031W	*YGL025C	YGL024W	*YLR050C	YLR051C
YAR042W	YAR044W	YGR074W	YGR075C	YLR058C	YLR059C
YBL039C	YBL038W	YGR086C	YGR087C	YLR077W	YLR078C
YBL017C	YBL016W	YGR090W	YGR091W	YLR116W	YLR117C
YBL004W	YBL003C	YGR119C	YGR120C	*YLR124W	YLR125W
YBL003C	YBL002W	YGR154C	YGR155W	YLR182W	YLR183C
YBR009C	YBR010W	YGR192C	YGR193C	YLR196W	YLR197W
*YBR066C	YBR067C	YGR233C	YGR234W	*YLR292C	YLR293C
*YBR074W	YBR076W	*YGR262C	YGR263C	*YLR294C	YLR295C
YBR083W	YBR084W	YGR266W	YGR267C	YLR321C	YLR322W
YBR087W	YBR088C	YGR267C	YGR268C	YLR328W	YLR329W
*YBR106W	YBR107C	YHL045W	YHL044W	YLR423C	YLR424W
YBR125C	YBR126C	YHL006C	YHL004W	YLR438C-A	YLR439W
YBR127C	YBR128C	YHR018C	YHR019C	YLR446W	YLR447C
YBR142W	YBR143C	*YHR034C	YHR035W	YLR452C	YLR453C
YBR144C	YBR145W	YHR043C	YHR044C	*YLR453C	YLR454W
*YBR160W	YBR161W	YHR051W	YHR052W	YML042W	YML041C
*YBR216C	YBR217W	YHR057C	YHR058C	YMR042W	YMR043W
*YBR240C	YBR241C	*YHR058C	YHR059W	YMR095C	YMR096W
YBR253W	YBR254C	YHR060W	YHR061C	*YMR152W	YMR153W
YBR288C	YBR289W	YHR068W	YHR069C	YNL334C	YNL333W
YCL028W	YCL027W	*YHR079C	YHR079C-A	YNL308C	YNL307C
*YCR066W	YCR067C	YHR088W	YHR089C	YNL282W	YNL281W
YCR076C	YCR077C	*YHR111W	YHR112C	YNL244C	YNL243W
*YCR087W	YCR087C-A	YHR112C	YHR113W	*YNL215W	YNL214W
*YCR096C	YCR097W	YHR113W	YHR114W	*YNL211C	YNL210W
*YDL066W	YDL065C	*YHR127W	YHR128W	*YNL088W	YNL087W
YDL060W	YDL059C	YHR129C	YHR130C	YNL085W	YNL084C
YDL044C	YDL043C	*YHR158C	YHR159W	YNL007C	YNL006W
YDR075W	YDR076W	YHR200W	YHR201C	*YNR005C	YNR006W
YDR115W	YDR116C	YHR206W	YHR207C	*YNR051C	YNR052C
YDR127W	YDR128W	YIL111W	YIL110W	YNR068C	YNR069C
YDR174W	YDR175C	*YIL051C	YIL050W	*YOL148C	YOL147C
YDR192C	YDR194C	*YIL005W	YIL004C	YOR158W	YOR159C
YDR224C	YDR225W	YJL152W	YJL151C	*YOR177C	YOR178C
YDR237W	YDR238C	YJL107C	YJL106W	YOR178C	YOR179C
*YDR268W	YDR269C	*YJL064W	YJL065C	YOR229W	YOR230W
*YDR306C	YDR307W	YJL041W	YJL039C	YOR302W	YOR303W
YDR308C	YDR309C	YJL014W	YJL013C	YOR340C	YOR341W
YDR342C	YDR343C	YJR009C	YJR010W	YOR361C	YOR362C
YDR438W	YDR439W	*YJR022W	YJR023C	*YOR387C	YOR388C
*YDR441C	YDR442W	*YJR027W	YJR028W	*YPL271W	YPL270W
*YDR526C	YDR527W	YJR064W	YJR065C	YPL256C	YPL255W
*YER009W	YER010C	YJR089W	YJR090C	*YPL234C	YPL233W
YER021W	YER022W	YJR090C	YJR091C	*YPL159C	YPL158C
YER043C	YER044C	*YJR091C	YJR092W	YPL128C	YPL127C
*YER112W	YER113C	YJR104C	YJR105W	*YPR010C	YPR011C

YER114C	YER115C	*YJR120W	YJR121W	*YPR045C	YPR046W
YER156C	YER157W	YKL104C	YKL103C	*YPR084W	YPR085C
YER172C	YER173W	YKL081W	YKL080W	YPR086W	YPR088C
YFL060C	YFL059W	YKL068W	YKL067W	YPR110C	YPR111W
YFL007W	YFL006W	*YKR077W	YKR078W	YPR119W	YPR120C
YFR024C	YFR024C-A	*YKR089C	YKR090W	*YPR174C	YPR175W
*YFR037C	YFR038W	*YKR090W	YKR091W		

Pairs preceded by an asterisk denote those obtained using DIP data only. Pairs not preceded by an asterisk denote those ones found using both the DIP (Dip1) and the DIP/TopNet (Int2) data sets.

References

1 Xenarios, I. *et al.* (2002) DIP, the database of interacting proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res.* 30, 303–305

2 Yu H, *et al.* (2004) TopNet: a tool for comparing biological sub-networks, correlating protein properties with topological statistics. *Nucleic Acids Res.* 32, 328-37

3 Stuart, J. M., *et al.* (2003) A gene co-expression network for global discovery of conserved genetic modules. *Science*, 302, 249–255

4 Mewes, H *et al.* (2002) Mips: a database for genomes and protein sequences. *Nucleic Acids Res.* 30, 31–34

5 Jansen R., *et al.* (2002) Relating whole-genome expression data with protein-protein interaction. *Genome Re*s. 12, 37–46

6 Teichmann S.A. (2002) The constraints protein-protein interactions place on sequence divergence. *J. Mol Biol.* 324, 399–407

7 Gerton J. L., *et al.* (2000) Inaugural article: global mapping of meiotic recombination hotspots and coldspots in the yeast *Saccharomyces cerevisiae. Proc Natl Acad Sci U S A,* 97, 11383–11390

8 Barabasi A. L. and Oltvai Z. N., (2004) Network biology: understanding the cell's functional organization, *Nat Rev Genet.* 5, 101–113

9 Spirin V. and Mirny L. A. (2003) Protein complexes and functional modules in molecular networks. *Proc Natl Acad Sci U S A* 100, 12123–12128

10 Blatt M., *et al.* (1996) Superparamagnetic clustering of data. *Phys Rev Letters,* 76, 3251–3254

11 Poyatos J. F. and Hurst L. D., (2004) How biologically relevant are interaction-based modules in protein networks? *Genome Biol.* 5, R93

12 Deane C.M. *et al.* (2002) Proteins interactions: two methods for assessment of the reliability of high throughput observations. *Mol Cell Proteomics,* 1, 349–356

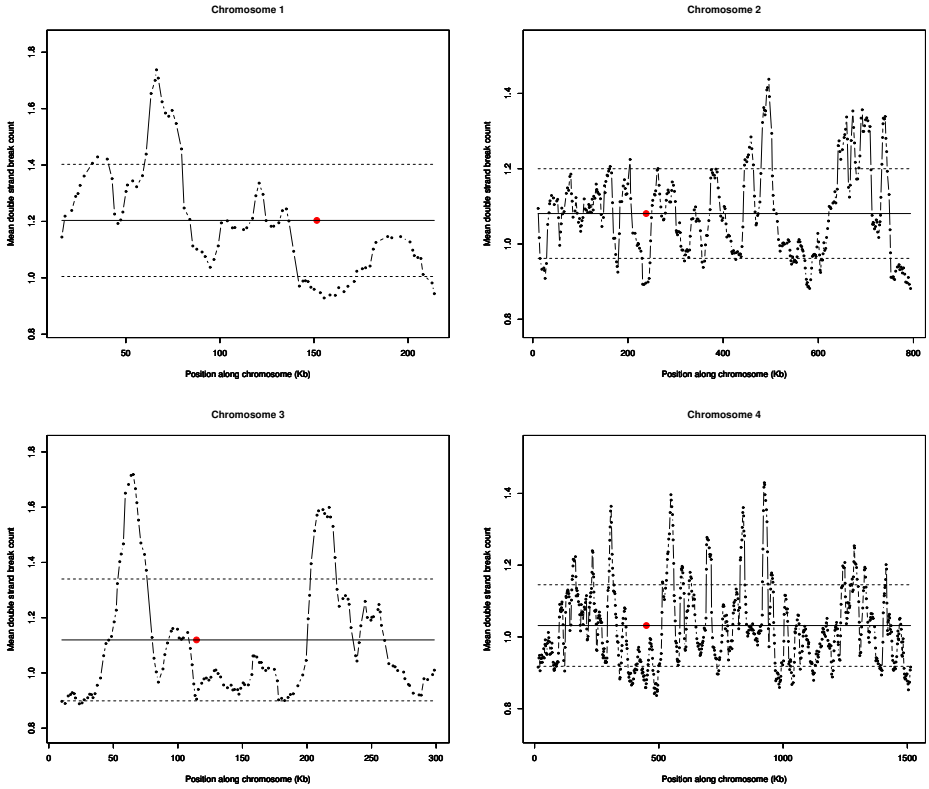

Figure 1: Variation in double strand break density across yeast chromosomes 1 to 4. The red spot indicates the position of the centromere. The horizontal line running through the centrome indicates the mean double strand break level in the chromosome. Dashed lines indicates plus and minus one standard deviation.

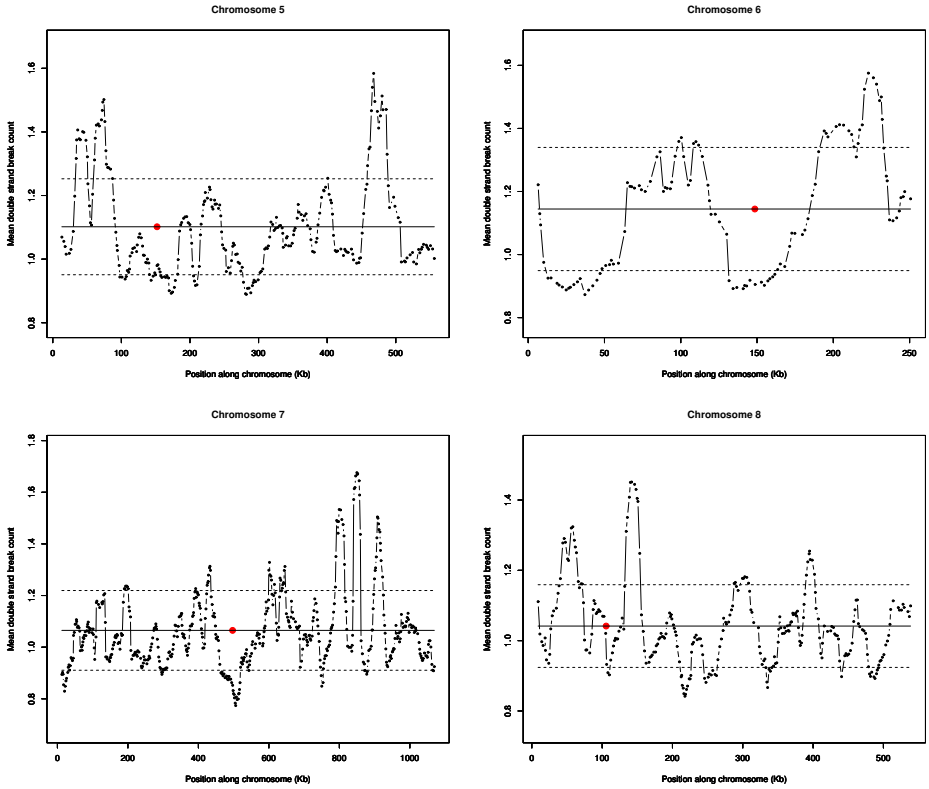

Figure 2: Variation in double strand break density across yeast chromosomes 5 to 8. The red spot indicates the position of the centromere. The horizontal line running through the centrome indicates the mean double strand break level in the chromosome. Dashed lines indicates plus and minus one standard deviation.

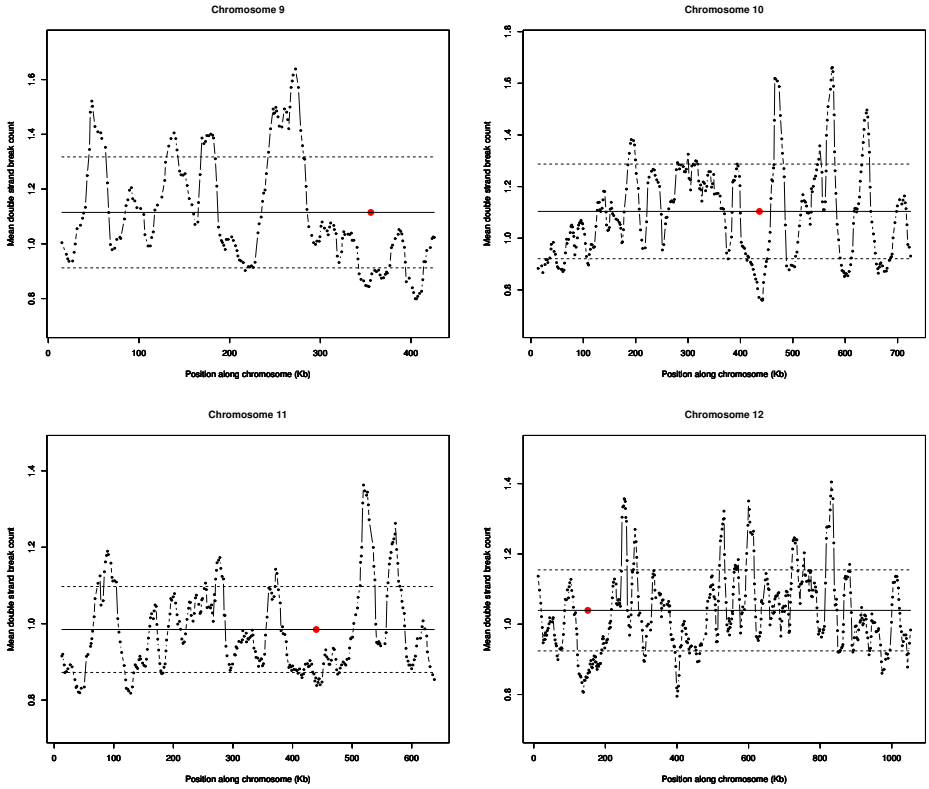

Figure 3: Variation in double strand break density across yeast chromosomes 9 to 12. The red spot indicates the position of the centromere. The horizontal line running through the centrome indicates the mean double strand break level in the chromosome. Dashed lines indicates plus and minus one standard deviation.

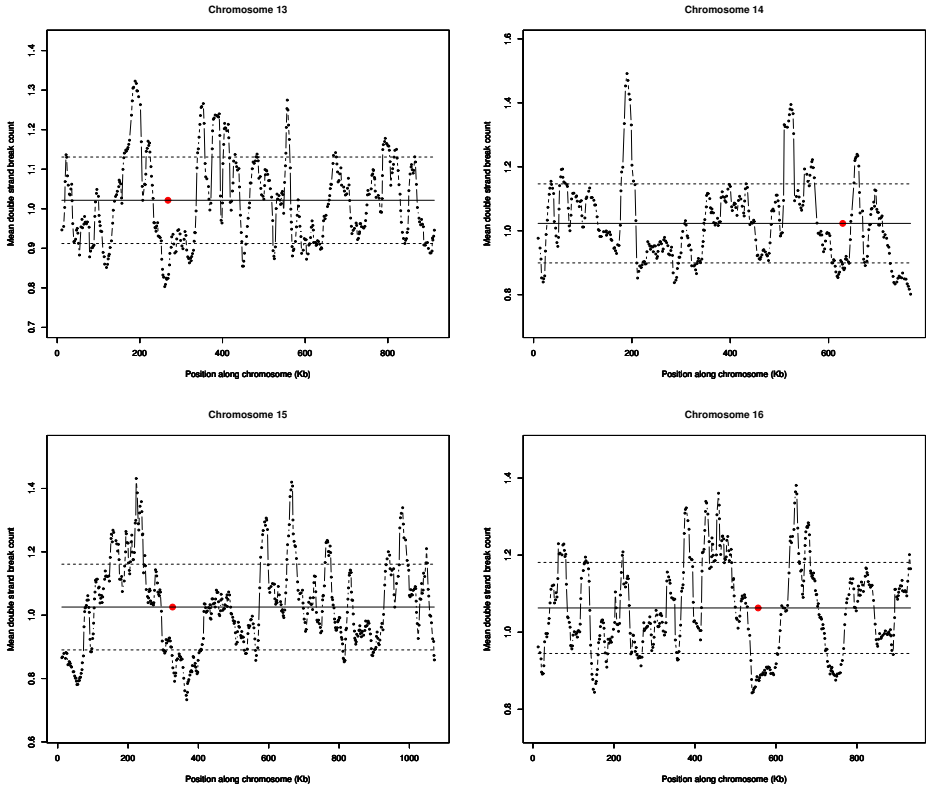

Figure 4: Variation in double strand break density across yeast chromosomes 13 to 16. The red spot indicates the position of the centromere. The horizontal line running through the centrome indicates the mean double strand break level in the chromosome. Dashed lines indicates plus and minus one standard deviation.

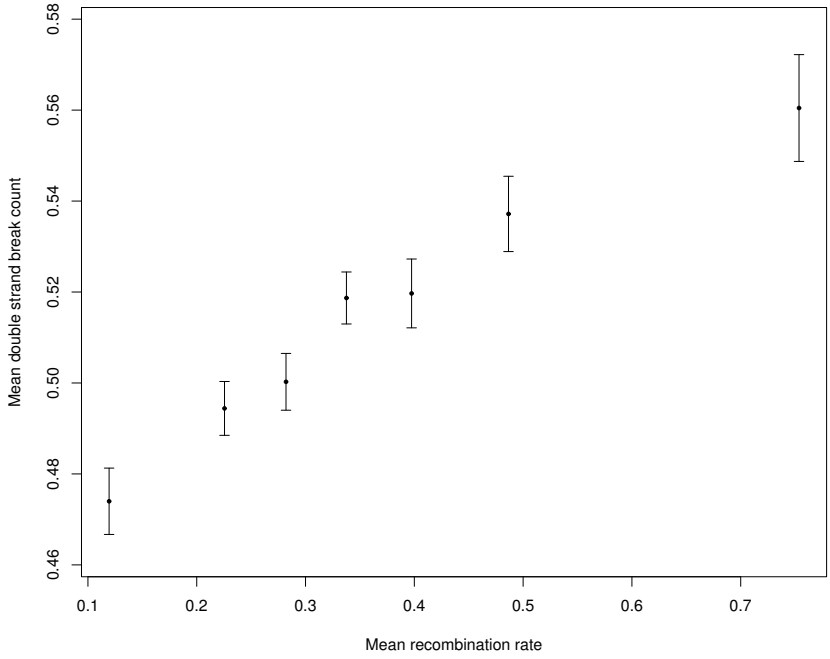

Figure 5: Comparison of double strand break data with recombination rate data. The data was assorted into equal sized bins ($N = 120$) after rank ordering by recombination rate. The X axis indicates the mean recombination rate per kb betweem pairs of genes derived from crossing data. The Y axis gives the mean for the double strand break rate per kb for the corresponding gene pairs.