Genome analysis

Recombinational DSBs-intersected genes converge on specific disease- and adaptability-related pathways

Zhi-Kai Yang\textsuperscript{1,2,3,4} Hao Luo\textsuperscript{1,2,3} Yanming Zhang\textsuperscript{4} Baijing Wang\textsuperscript{4} and Feng Gao\textsuperscript{1,2,3,*}

\textsuperscript{1}Department of Physics, School of Science, Tianjin University, Tianjin, China, \textsuperscript{2}Key Laboratory of Systems Bioengineering (Ministry of Education), Tianjin University, Tianjin, China, \textsuperscript{3}SynBio Research Platform, Collaborative Innovation Center of Chemical Science and Engineering (Tianjin), Tianjin, China, \textsuperscript{4}SinoGenoMax Co., Ltd./Chinese National Human Genome Center, Beijing, China

*To whom correspondence should be addressed.

Received on XXXX; revised on XXXX; accepted on XXXX

Abstract

Motivation:
The budding yeast \textit{Saccharomyces cerevisiae} is a model species powerful for studying the recombination of eukaryotes. Although many recombination studies have been performed for this species by experimental methods, the population genomic study based on bioinformatics analyses is urgently needed to greatly increase the range and accuracy of recombination detection. Here, we carry out the population genomic analysis of recombination in \textit{S. cerevisiae} to reveal the potential rules between recombination and evolution in eukaryotes.

Results:
By population genomic analysis, we discover significantly more and longer recombination events in clinical strains, which indicates that adverse environmental conditions create an obviously wider range of genetic combination in response to the selective pressure. Based on the analysis of recombinational DSBs-intersected genes (RDIGs), we find that RDIGs significantly converge on specific disease- and adaptability-related pathways, indicating that recombination plays a biologically key role in the repair of DSBs related to diseases and environmental adaptability, especially the human neurological disorders (NDs). By evolutionary analysis of RDIGs, we find that the RDIGs highly prevailing in populations of yeast tend to be more evolutionarily conserved, indicating the accurate repair of DSBs in these RDIGs is critical to ensure the eukaryotic survival or fitness.

Contact: fgao@tju.edu.cn

Supplementary information: Supplementary data are available at Bioinformatics online.

1 Introduction
The budding yeast \textit{Saccharomyces cerevisiae} has long served as an important model organism for higher eukaryotes as it could promote the combination of classical genetics and biochemistry with recombinant genetics (Amberg \textit{et al.}, 2016), and its cells have striking similarities to mammalian cells at both the organelle and macromolecular level (Bracio\textit{ni} \textit{et al.}, 2011). Just recently, a systematic study has shown that most of the genes conserved between yeast and human play the same exact roles in both organisms, and human genes can completely substitute for yeast orthologous genes, raising that humanizing entire cellular processes or pathways in yeast should be feasible (Kachroo \textit{et al.}, 2015). Due to the high biological relevance between yeast and human, biological findings in the budding yeast can be efficiently translated into new insights about human biology (Skrzypek \textit{et al.}, 2017).

It has been known that genomic instability is an important factor for human diseases. Though endogenous or exogenous genotoxic assaults constantly destroy the genome integrity and lead to DNA damages, cells usually repair them by various mechanisms (Jeggo \textit{et al.}, 2016). Here we
focus on the most hazardous damage, namely the double strand breaks (DSBs), which are generally repaired by two main mechanisms, the error-free homologous recombination (HR) and the fast but error-prone non-homologous end-joining (NHEJ), respectively. These mechanisms are shared by eukaryotes (Carroll et al., 2014), but HR is usually less frequently used than NHEJ (Suhane et al., 2015). Considering HR is an accurate DNA repair mechanism, we supposed that recombinational DSBs-intersected genes (RDIGs) were associated with special biological significances.

As the most lethal form of DNA damage, DSBs pose a great potential risk to human health. For example, defects in DNA repair pathways are the causal factors for atherosclerosis, and led to higher amounts of DNA damages after exposure to reactive oxygen species (ROS) (Mahmoudi et al., 2006). When HR repair was defective, smooth muscle cells became increasingly senescent in patients with Hutchinson–Gilford progeria syndrome (Zhang et al., 2014), indicating DNA damages were related to aging (Gorbunova et al., 2007). Differentiation of osteoblasts was inhibited obviously by DNA damages, which contributed to the weakened bone structure in mouse model (Schmidt et al., 2012). These examples suggest that DSBs are associated with human diseases, but it is still unclear which diseases significantly rely on the error-free HR repair mechanism. Hence, we undertake this study aiming at exploring the biological feature and function of RDIGs, which reflects what potential roles the recombination may play in human.

2 Materials and methods

2.1 Strain and data

Based on the phylogenetic relationship previously built by us (Yang et al., 2018), we totally selected 42 strains with consideration of the geographic and environmental origins (Supplementary Table S1), and downloaded the corresponding genome sequences from the NCBI FTP site (ftp://ftp.ncbi.nih.gov/genomes/all/) in August 2016. These 42 strains represent an extensive geographic distribution, and all these genomes are well annotated and highly complete with more than 99% integrity. In order to facilitate the following analyses, each sequence header of every strain genome was unified according to the strain name, chromosome and corresponding identifier. In addition, the protein sequences and GBFF files from NCBI by a BioPerl Script.

2.2 Identification of homologous sequences

Firstly, all the genomes were split into individual chromosome sequence files. Then, all the sequences of each chromosome were aligned by Mugsy, which was a robust whole-genome alignment program suitable for accurately aligning the closely related whole genome sequences (Angiuoli et al., 2011). And then, we extracted all the local aligned blocks of every chromosome, and generated a multi FASTA format file for each of local aligned blocks. However, these blocks with the length < 1kbp or the number of homologous sequences < 3 would be viewed as ineligible blocks.

2.3 Phylogenetic network analysis

We adopted these aligned blocks containing 42 homologous sequences to construct the phylogenetic network. Firstly, eligible aligned blocks were realigned separately to produce more accurate alignments by using the program MAFFT v7.271 with default parameters (Katoh et al., 2013). Then, all these realigned blocks were concatenated to a single alignment. And then, all the SNP loci were extracted to generate the resulting alignment. Finally, this alignment was used as the input of NeighborNet program in SplitsTree v4.14.5 (Huson et al., 2006) to construct the phylogenetic network with default settings. In addition, a statistical analysis was carried out by the pair-wise homoplasy (PHI) test (Bruen et al., 2006) as implemented by SplitsTree program to estimate the situation of recombination among these strains.

2.4 Detection and analysis of recombination events

After we constructed the phylogenetic network and carried out the pair-wise homoplasy index (PHI) statistical test for recombination, we employed seven different methods, the RDP (Martin et al., 2000), GENECONV (Padidam, 1999), 3SEQ (Boni et al., 2007), BOOTSCAN (Salminen et al., 1995), MAXCHI (Smith, 1992), SISCAN (Gibbs et al., 2000) and CHIMAERA (Posada et al., 2001) respectively, to detect the recombination signal, and all these methods were implemented in Recombination Detection Program (RDP4) Beta v. 4.91 (Martin et al., 2015). The breakpoint position of recombination events was determined by RDP4 using a hidden Markov model. In this process, we identified the recombination events separately for each of eligible multiple alignments. And then, all the outputs of RDP4 were parsed to get the simplified and reliable results, but we only considered those recombination events that at least five out of seven methods showed the significant P value. Finally, we performed the statistical analysis, as well as the association analysis of recombination with environmental factors.

2.5 Identification of recombinational DSBs-intersected genes (RDIGs)

In this study, recombinational DSBs-intersected genes (RDIGs) refer to the genes including recombinational DSBs. The identification and functional analysis of RDIGs may be important for exploring the biological significance of recombination because recombinational DSBs were previously reported to be an extraordinarily nonrandom distribution along the whole genome (Lam et al., 2015). So we identified all the RDIGs and calculated the average frequency of recombinational DSBs by a custom PERL script. In addition, we also examined the distribution of recombinational DSBs in each chromosome by 2 kbp non-overlapping windows, and plotted them by Circos program (Krzywinski et al., 2009).

2.6 Functional enrichment analysis

Here we used the S. cerevisiae strain S288c as reference, and reannotated all its protein-coding genes based on the latest functional data (Please see the Supplementary Methods for details). Then, a standard procedure EnrichPipeline was adopted to carry out the enrichment analysis of GO term/KEGG pathway (Huang et al., 2009; Beissbarth et al., 2004). In this process, the procedure firstly identified the significantly enriched results according to the P-values by Chi-square test and Fisher exact test. And then, the P-values of gene set enrichment were further adjusted by the default method fdr to filter out those false-positive results. Finally, we took the cutoff of 0.05 for the adjusted P-values as statistical significance.

2.7 Estimation of dN/dS value

Downloaded from https://academic.oup.com/bioinformatics/advance-article-abstract/doi/10.1093/bioinformatics/bty376/4992144 by guest on 22 May 2018
Population genomic analysis of recombination

3 Results

3.1 Phylogenetic network

Originally, we constructed a reticulate phylogenetic network based on the alignment consisted of SNP loci (358 kbp) which were extracted from 240 eligible aligned blocks by a custom Perl script. As shown in Supplementary Fig. S1, a complex and reticulate tree topology structure was presented, which indicated the presence of significantly conflicting phylogenetic signals. Generally, conflicting signals were caused by genetic recombination. By the PHI test analysis, we detected the statistically significant evidence of recombination in the 42 budding yeast strains. Thus, a more comprehensive assessment of recombination events was necessary for the insight into the underlying biological significance.

3.2 Overview of recombination events

There were totally 762 blocks after filtering the ineligible blocks. By PHI test analysis, we detected the statistically significant evidence of recombination in the 42 budding yeast strains. Then, we identified all possible recombination events (Supplementary Table S2) by the RDP4 program and compared these results predicted by seven different methods implemented in this program, finding that the number of recombination events predicted by different methods ranged from 12,935 to 16,107 (Fig. 1a). According to the overlapping relationships constructed by Venn diagram, we found very few recombination events with significant signals detected by only one method (Fig. 1a), while the number of recombination events with significant signals detected by all seven methods was far greater than the others (Fig. 1b), indicating that these seven methods had a good consistency in the detection of recombination signals. Please note that the two sets of data presented in Fig. 1b have a good correspondence. For example, the case with number of methods = 5 indicated by red dot includes three cases, namely 5, 6 and 7, indicated by bar plot.
In order to assess whether the number and length of recombination events in clinical environments differed significantly from those in other environments, we carried out the Wilcoxon-Wilcox test to detect the statistical significance of difference between any two environments. As a result, we found the number of recombination events in clinical environments was obviously more than that in the other two environments (wine and other, respectively) (Fig. 3a), but the total length of recombination events in clinical environments was obviously more than that in the other two environments (Fig. 3b). These findings indicated that strains from harmful environments had significantly more genetic combination than those from benign environments, which was in accordance with what had been observed in the fungus *Arthrobotrys oligospora* (Zhang et al., 2013).

### 3.4 DSBs and RDIGs

Based on all the recombination events detected above, we identified 23,356 recombinational DSBs from genomes of 42 budding yeast strains, and plotted the distribution of DSBs along each chromosome. As a result, we found that DSBs were obviously distributed non-randomly along chromosomes (Fig. 2), which was consistent with previous findings (Szekvolygi et al., 2015). Then, we mapped the positions of all these DSBs to the genome of strain s288c, and totally identified 3060 RDIGs. Actually, RDIGs were such genes spliced together by different portions of two homologous genes, therefore they also belonged to the mosaic genes. We calculated the average frequency of recombinational DSBs occurring in RDIGs by a custom Perl script. We focused more on these RDIGs with high frequency (≥ 0.9) of recombinational DSBs (referred as HRDIGs here), because these genes were highly dependent on error-free HR. Here, we totally identified 60 HRDIGs. Considering that perfect complementary pairings of invasive strand with template were crucial for the initiation of HR (Maher et al., 2011), we speculated that HRDIGs were likely to be evolutionarily conserved and biologically significant genes.
Population genomic analysis of recombination

By KEGG enrichment analysis, we found 13 pathways related to human diseases significantly enriched in HRDIGs (Fig. 4; Supplementary Table S3). Specifically, there were three NDs (PD, HD and AD), which indicated that organisms relied on HR to thwart the occurrence of NDs. This was supported by previous findings that accumulation of DNA lesions was tightly related to NDs (Maynard et al., 2015), and organisms may resist not only the normal aging but also age-related NDs by reliable DNA repair mechanism (Leandro et al., 2015). And there were two cancer pathways (prostate cancer and chemical carcinogenesis, respectively), which was in accordance with previous findings that inaccurate repair of DSBs was related to the development of certain cancer (Dexeheiner, 2013), and defects of recombination predisposed to cancer (Martino et al., 2016). There were also three other disease pathways, the non-alcoholic fatty liver disease, diseases fluid shear stress and atherosclerosis (Jackson et al., 2013), the drug resistance not only the normal aging but also age-related NDs by reliable DNA repair mechanism (Leandro et al., 2015). And there were two cancer pathways (prostate cancer and chemical carcinogenesis, respectively), which was in accordance with previous findings that inaccurate repair of DSBs was related to the development of certain cancer (Dexeheiner, 2013), and defects of recombination predisposed to cancer (Martino et al., 2016). There were also three other disease pathways, the non-alcoholic fatty liver disease, diseases fluid shear stress and atherosclerosis (Jackson et al., 2013); Gray et al., 2009; Gray et al., 2009; Gray et al., 2009; Gray et al., 2009). These results together with the results of KOG and GO analysis suggested that organisms relied on HR to thwart the occurrence of NDs. This was supported by previous findings that accumulation of DNA lesions was tightly related to NDs (Maynard et al., 2015), and organisms may resist not only the normal aging but also age-related NDs by reliable DNA repair mechanism (Leandro et al., 2015). And there were two cancer pathways (prostate cancer and chemical carcinogenesis, respectively), which was in accordance with previous findings that inaccurate repair of DSBs was related to the development of certain cancer (Dexeheiner, 2013), and defects of recombination predisposed to cancer (Martino et al., 2016). There were also three other disease pathways, the non-alcoholic fatty liver disease, diseases fluid shear stress and atherosclerosis (Jackson et al., 2013).

3.5 Results of KEGG enrichment analysis in HRDIGs

As to the enrichment of measles pathway, previous experimental study observed a similar result that all the measles cases had an obviously positive correlation with the length of chromosomes, hence DSBs occur more frequently in longer chromosomes. This was supported by previous findings that accumulation of DNA lesions was tightly related to NDs (Maynard et al., 2015), and organisms may resist not only the normal aging but also age-related NDs by reliable DNA repair mechanism (Leandro et al., 2015). Moreover, there were also five disease-related pathways, namely the IL-17 signaling pathway, oxidative phosphorylation, Antigen processing and presentation, Th17 cell differentiation, and cardiac muscle contraction respectively. These results together with the results of KOG and GO analysis (Supplementary Results, Supplementary Figs. S6-S7 and Table S4) suggested that HRDIGs were tightly associated with certain human diseases. Moreover, there were also some pathways related to specific adaptability. For instance, estrogen signaling pathway regulated a plethora of physiological processes in mammals (including cardiovascular protection, reproduction, and behavior) (Rottroff et al., 2013), the drug metabolism and metabolism of xenobiotics by cytochrome P450 were responsible for xenobiotics biodegradation and metabolism (Kirchmair et al., 2015), and the two-component system acted as a basic stimulus-response mechanism to permit organisms to sense and respond to the changes of environmental conditions (Kreamer et al., 2015). These results indicated that functional impairment of HRDIGs was likely to result in obvious decrease of environmental adaptability, hence DSBs occurring in HRDIGs were highly dependent on the error-free HR repair mechanism.

3.6 DSBs frequency and dN/dS of RDIGs

By analyzing the relationship of recombinational DSBs frequency and dN/dS value in RDIGs, we found that those RDIGs prevailing more widely in populations tended to keep a lower dN/dS value, and almost all of positively selective genes (PSGs) held an extremely low frequency of DSBs (Fig. 5a). This was consistent with previous studies that high genetic similarity corresponded more significantly to the distribution of DSBs (Magiorkinis et al., 2003). In HR repair, a long stretch of single strand DNA would be generated by the end resection of DSBs (Traong et al., 2013), and this single strand DNA might be the most critical factor for determining if a homolog should be used as the recombination partner (Joshi et al., 2015). In this analysis, we found that the frequency of recombinational DSBs in PSGs was significantly lower than that in negatively selective genes (NSGs) (Fig. 5b), and none of HRDIGs were PSGs. Hence, we speculated that HR preferred to repair these DNA damages in highly homologous regions, which in turn contributed to higher evolutionary conservation in these regions.

4 Discussions

As discussed above, HR is a biologically significant mechanism of repairing DSBs that occur spontaneously or are induced by biological, chemical or physical agents. In the HR process, 5′ end of DSBs is firstly resected to generate an ssDNA tail of ~800 nt. And then, this ssDNA tail initiates the homologous pairing after invasion, and serves as the primer for DNA biosynthesis. Ultimately, an intact homologous duplex is used as template to repair the broken strands. However, HR is only responsible for a small proportion of DSBs (Sundararajan et al., 2016). Its operating frequency is usually regulated homeostatically, and does not change with alteration of DSB frequency (Modliszewski et al., 2017). Though HR is an error-free DNA repair mechanism, it is also a non-random and biased process. While HR was found to prefer to repair these DSBs in evolutionarily conserved regions as observed above (Fig. 5), there may be some other factors influencing this priority of HR. We supposed that HR would be used preferentially to repair those biologically significant DSBs.

In present study, we found that the number of recombination events had an obviously positive correlation with the length of chromosomes, and also found that the total length of recombination events in clinical strains were significantly longer than that in wine or other strains. A reasonable interpretation for this is that clinical strains are subjected to other factors influencing this priority of HR. We supposed that HR would be used preferentially to repair those biologically significant DSBs.

In present study, we found that the number of recombination events had an obviously positive correlation with the length of chromosomes, and also found that the total length of recombination events in clinical strains were significantly longer than that in wine or other strains. A reasonable interpretation for this is that clinical strains are subjected to other factors influencing this priority of HR. We supposed that HR would be used preferentially to repair those biologically significant DSBs.

In present study, we found that the number of recombination events had an obviously positive correlation with the length of chromosomes, and also found that the total length of recombination events in clinical strains were significantly longer than that in wine or other strains. A reasonable interpretation for this is that clinical strains are subjected to other factors influencing this priority of HR. We supposed that HR would be used preferentially to repair those biologically significant DSBs.
the development of new therapeutic strategies for specific human diseases.

Acknowledgements

The authors would like to thank Prof. Chun-Ting Zhang for the invaluable assistance and inspiring discussions.

Funding

This work was supported by the National Natural Science Foundation of China (Grant Nos. 31571358, 21621004, 31712138 and 91746119).

Conflict of Interest: none declared.

References


