uOttawa

L'Université canadienne
Canada's university

FACULTÉ DES ÉTUDES SUPÉRIEURES ET POSTOCTORALES

uOttawa

L'Université canadienne
Canada's university

FACULTY OF GRADUATE AND POSDOCTORAL STUDIES

Robert Morris

AUTEUR DE LA THÈSE / AUTHOR OF THESIS

Ph.D. (Biology)

GRADE / DEGREE

Department of Biology

FACULTÉ, ÉCOLE, DÉPARTEMENT / FACULTY, SCHOOL, DEPARTMENT

Ectopic gene conversion in prokaryotic and yeast genomes

TITRE DE LA THÈSE / TITLE OF THESIS

Guy Drouin

DIRECTEUR (DIRECTRICE) DE LA THÈSE / THESIS SUPERVISOR

CO-DIRECTEUR (CO-DIRECTRICE) DE LA THÈSE / THESIS CO-SUPERVISOR

EXAMINATEURS (EXAMINATRICES) DE LA THÈSE / THESIS EXAMINERS

Stephane Aris-Brosou

Teresa Crease

Linda Bonen

Xuhua Xia

George Carmody

Gary W. Slater

Le Doyen de la Faculté des études supérieures et postdoctorales / Dean of the Faculty of Graduate and Postdoctoral Studies

Ectopic gene conversion in prokaryotic and yeast genomes

by

Robert T. Morris

A Thesis submitted in partial fulfillment of
the requirements for the degree of
Ph. D. of Biology
The Faculty of Graduate and Postdoctoral Studies
Ottawa-Carleton Institute of Biology
University of Ottawa

Your file Votre référence
ISBN: 978-0-494-49384-7
Our file Notre référence
ISBN: 978-0-494-49384-7

Table of Contents

List of Figures 4
List of Tables 6
Acknowledgements 7
Abstract 8

Chapter 1 General Introduction 10
Literature Review 11
Review of Recombination and Gene Conversion Models 13
Gene conversion detection methods 14
Characteristics of Gene Conversions 16
Evolutionary consequences of Gene Conversions 19
Main Objectives 20
Literature Cited 23

Chapter 2 Ectopic gene conversions in four *Escherichia coli* genomes: increased recombination in pathogenic strains 32

Abstract 33
Introduction 34
Materials and Methods 36
Results 40
Discussion 46
Literature Cited 53

Chapter 3 Ectopic gene conversions in the backbone genome of *Echerichia coli* genomes 67

Abstract 68
Introduction 69
Materials and Methods 71
Results 74
Discussion 77
Literature Cited 81

Chapter 4 Ectopic gene conversions in bacterial genomes 92

Abstract 93
Introduction 94
Materials and Methods 97
Results 102

Discussion 113
References 122

Chapter 5 **Ectopic gene conversions in the genome of ten Hemiascomyctes yeast species** 144

Abstract 145
Introduction 146
Materials and Methods 150
Results 154
Discussion 163
Literature Cited 170

Chapter 6 **Ectopic gene conversions increases the G+C content of duplicated yeast and *Arabidopsis* genes** 189

Abstract 190
Introduction 191
Materials and Methods 193
Results 194
Discussion 196
Literature Cited 200

Chapter 7 **General Summary** 209

Factors affecting Prokaryote and Eukaryote ectopic gene conversion 210
Effect of gene conversion 214
Literature Cited 215

List of Figures

Chapter 1
Figure 1: Connection between gene conversions and cross-over events. 29
Figure 2: Molecular models of recombination and gene conversion and Holliday Junction resolutions. 30
Figure 3: Phylogenetic effect of gene conversion. 31

Chapter 2
Figure 1: Number of multigene families containing three or more members in the genome of the four *E. coli* strains. 58
Figure 2: Size of the gene conversions found in the genome of the four *E. coli* strains. 59
Figure 3: Relationship between the length of each converted region and the similarity of its flanking regions found within the coding region of the genes in the four E. coli strains. 60

Chapter 3
Figure 1: Multiple alignment of the Sakai, EDL933, K-12 and CFT073 genomes using the MAUVE software package with a minimum weight of 69. 90
Figure 2: Converted regions from gene families 8 and 16. 91

Chapter 4
Figure 1: Frequency of multigene family sizes within the pathogenic and nonpathogenic datasets of proteobacteria, firmicute and archaea. 133
Figure 2: Relationship between gene conversion length and maximum flanking similarity for all gene conversions. 134
Figure 3: Distribution of the distances between multigene family members and of the distances between converted genes having at least 80% maximum flanking similarity. 137
Figure 4: Relationship between gene conversion frequency and distance from the origin of replication. 140
Figure 5: Distribution of the converted regions within the converted genes having at least 80% maximum flanking similarity. 142

Chapter 5
Figure 1: Schematic indicating how ohnologous and paralogous genes are separated within a multigene family. 184
Figure 2: Distribution of the average number of multigene families within seven post duplication and three pre duplication genomes. 185
Figure 3: Gene conversion length is positively correlated with maximum flanking similarity. 186

Chapter 6
Figure 1: Relationship between the average GC-content and the average sequence similarity of 375 pairs of ohnologs in the yeast genome. 206
Figure 2: Relationship between the GC-content and the median recombination rate of 750 ohnologs found in the yeast genome. 207
Figure 3: Relationship between the average GC-content and the average sequence similarity of the 2497 pairs of recent ohnologs in the *Arabidopsis* genome. 208

List of Tables

Chapter 2
Table 1: Amino acid substitutions in the recombination genes of the three pathogenic strains. 61
Table 2: Distribution of conversions, genes and conversion frequencies. 64
Table 3: Distribution of converted genes, gene family members and conversion frequencies relative to *oriC*. 66

Chapter 3
Table 1: Gene conversions detected within the backbone of the four *E. coli* genomes. 85
Table 2: Ka/Ks ratios (± standard deviation) of converted backbone genes. 87
Table 3: Gene conversions detected between duplicated genes specific to the K-12 genome. 88

Chapter 4
Table 1: Summary of gene conversion frequencies (%) within the five groups of genomes. 126
Table 2: Summary statistics of the gene conversion lengths (bp) for each group of genomes. 127
Table 3: Summary statistics for the maximum flanking similarity (%) for each group. 129

Chapter 5
Table 1: List of the number of ohnolog and paralog multigene families in the pre- and post- WGD genomes. 174
Table 2: Percentage of gene comparisons which are between multigene family members located on the same chromosome. 175
Table 3: Intra and inter chromosomal gene conversion frequencies for pre- and post- WGD genomes. 176
Table 4: Number and frequency of gene conversions in ohnolog and paralogs of ten fungi genomes. 177
Table 5: Gene conversion length statistics of pre- and post- whole genome duplication (WGD) species. 178
Table 6: Statistics for maximum flanking similarity requirements of gene conversions in pre- and post- whole genome duplication (WGD) species. 179
Table 7: Rate of non-synonymous, synomymous and estimate of selective pressure on paralogs. 180
Table 8: Correlation between the location of the middle of converted regions and the number of gene conversions within pre- and post- WGD genomes. 181

Acknowledgements

I would like to express my gratitude to Dr. Guy Drouin for the great opportunity he gave me as my Ph. D. thesis supervisor. Under his guidance I have gained an appreciation of academic research. I deeply thank him for all his support, expertise and advice he has provided over the past four years. I would like to thank the other members of my graduate committee: Dr. George Carmody (Biology Department, Carleton University) and Dr. Xuhua Xia (Biology Department, Ottawa University) for all their constructive advice regarding my projects. I would also like to thank all the people who have worked within the Drouin Lab since I began my studies, in particular I would like to thank David Benovoy and Robert Carter for their counsel and friendship. Furthermore I would like to thank Huiling Xiong, and Dr. Antione Morin for their help with my statistical questions. Finally I would like to thank my family for their support during my studies.

Abstract

The repair of double strand DNA breaks is important to genomic stability. Gene conversions are non-reciprocal exchanges of DNA created through the repair of double strand DNA breaks. Relatively few studies have looked at examining the similarities and differences of the physical characteristics of ectopic gene conversions (conversions between non-allelic genes) within and between diverse lineages of prokaryotes and yeast.

I found that several factors affect the frequency of ectopic gene conversions in prokaryotes and yeast. The size and organization of multigene families affects how often genes are converted. In *E. coli* a positive correlation was found between gene family size and conversion frequency. In yeast, I found that intra-chromosomal gene conversions are more frequent than inter-chromosomal conversions. The amount of sequence similarity between the converted genes affects their conversion frequency. I found that long conversions tend to occur between highly similar homologs. In addition, the sequence similarity requirements are not uniform for every organism. For example, I found indications that the life-style of the organism affects the stringency of sequence similarity requirements for ectopic gene conversion. I determined that ectopic gene conversions occur more frequently and require significantly less flanking similarity in the pathogenic bacteria strains than in the non-pathogenic. The functional importance of the converted genes affects how frequently they are converted. Genes that have been functionally maintained throughout evolution are ectopically converted less frequently than genome specific paralogs. Nucleotide distance between homologous genes affects their ectopic conversion frequency. In paralogous multigene families of yeast, gene conversions occur more frequently between genes which are close together on the same chromosome than

between dispersed homologs. Converted regions are not uniformly distributed along the length of genes. Converted regions tend to be clustered near the 3' end of genes in yeast but not in prokaryotes. I also looked at the effects of gene conversions on nucleotide composition of yeast genes and found that repeated ectopic gene conversions between dispersed homologs tend to increase the converted genes GC-content.

Chapter 1

General Introduction

Literature Review

The repair of double strand DNA breaks (DSB) is a necessity for maintaining genome stability. DSB are formed spontaneously and by enzymes such as Spo11 in yeast during the cell cycle or can be caused by exogenous agents such as UV radiation and various chemical compounds (Dudas and Chovanec 2004). Two major processes exist in prokaryotes and eukaryotes to repair this type of DNA damage, non - homologous end joining (NHEJ) and homologous recombination (HR; reviewed in Aylon and Kupiec 2004, Wyman et al 2004, Bowater and Doherty 2006). The process of NHEJ involves the direct ligation of the broken strands to repair the DSB, while HR relies on the use of an undamaged template to provide a copy of the missing DNA sequence for the damaged gene (Aylon and Kupiec 2004).

Here I will focus on the study of ectopic gene conversions which are a consequence of HR repairs of DSB. Historically, the study of homologous recombination in yeast has detected deviations from the Mendelian segregation (2:2) of alleles in yeast tetrads. During meiosis this deviation was identified as a 3:1 or 1:3 ratio; this indicated that one allele was altered to resemble its homolog (reviewed in Aylon and Kupiec 2004). This process involves a non – reciprocal exchange of genetic information between two homologs (Figure 1). This process is called gene conversion. Evidence of gene conversions were also detected within mitotic yeast cells (Roman 1956). This trait was linked to the repair of double stranded breaks through the process of homologous recombination in prokaryotes and eukaryotes (Cromie et al 2001, Kobayashi and Takahashi 1988, Fogel and Hurst 1967).

Multigene families are common within eukaryote and prokaryote genomes (Rocha et al 1999, Romero et al 1999, Santoyo and Romero 2005). Gene conversion

plays a role in the concerted evolution of these gene families in both prokaryotes and eukaryotes. Some examples include the evolution of rRNA in archaea and bacteria species (Liao 2000); likewise the evolution of *tufA* and *tufB* in *Salmonella typhimurium* and *E. coli* (Hughes 2000, Lathe III and Bork 2001) has been attributed to gene conversions. In addition the evolution of primate eye pigments (Zhou and Li 1996, Balding et al 1993) and human globin (Scott et al. 1984) multigene families have been influenced by gene conversions. The end result of allelic and ectopic gene conversions is that the repaired and the template gene become more similar to each other within the converted region.

The choice of the undamaged template used to repair the damaged gene defines the type of gene conversion. An allelic gene conversion occurs when a gene is repaired using its allele from a sister chromatid (most likely during mitosis/meiosis). An ectopic gene conversion occurs when a paralog (created by a duplication event) is used as the repair template.

This thesis focuses on the detection and characterization of ectopic gene conversions in prokaryotes and yeast. This introductory section reviews the historical progression of the models developed to explain the repair mechanism of double strand DNA breaks. This is followed by a brief review of methods developed to detect gene conversions, a summary of the previous work done on the characterization of gene conversions and a discussion of the evolutionary consequences of ectopic gene conversions. Finally the main objectives of each study presented in this thesis are outlined.

Review of Recombination and Gene Conversion Models

The majority of early research into recombination and gene conversion was performed using yeast (Perkins 1992). The earliest model which attempted to explain yeast recombination and gene conversion data was proposed by Robin Holliday (Holliday 1964). This model is based on the repair of identical single strand DNA (ssDNA) breaks in two homologous chromosomes. This model proposed that there was a symmetric exchange of broken strands which resulted in the formation of a Holliday junction. This junction was free to migrate thereby facilitating the creation of heteroduplex DNA (hDNA) which was formed when ssDNA of the invading strand was base paired with ssDNA of the template strand (see Figure 2a). One of the predictions of this model was that there should be equal formation of hDNA on each chromatid. These regions of hDNA contained mismatches which must be resolved; this processes resulted in gene conversions. In 1975, Mathew Meselson and Charles Radding proposed a variation of the Holliday model to explain why some experimental recombination results did not support the symmetric formation of hDNA on each chromatid (Meselson and Radding 1975; see Figure 2a). Their model accounts for asymmetric and symmetric hDNA formation, and involved the isomerization of the Holliday junction (Meselson and Radding 1975). Previous work by Orr-Weaver and colleagues (1983) determined that double strand DNA breaks induced recombination in yeast. The first model presented to explain recombination and gene conversion via the repair of DSB was called the DSB repair model (DSBR) proposed by Szostak and colleagues (Szostak et al 1983; see Figure 2a). In this model the DSB ends are resectioned by exonucleases creating two 3' overhanging sequences, these sequences then invade the template sequence creating a D-loop. The D-

loop is enlarged by repair synthesis and the remaining free 3' end anneals to the D-loop. Repair synthesis is completed by replication from the other 3' end annealed to the D-loop. Two adjacent Holliday junctions are formed by the repair synthesis; these junctions migrate and form hDNA. The Holliday junctions are resolved by endonuclease activity. If both structures are resolved in the same direction this results in a gene conversion without a cross-over, however if the junctions are resolved in different directions then this results in gene conversion with a cross-over (Figure 2b). Significant evidence supports the DSBR model as the framework for recombination and gene conversion in prokaryotes, because the majority of the proteins implicated in this process in eukaryotes using this model have homologs in prokaryotes (Kowalczykowski et al 1994). Subsequently a variation of the DSB repair model called the strand dependent strand annealing (SDSA) model was proposed to explain variations in the results of mitotic recombination in yeast (Hastings 1988, McGill et al 1989; see Figure 2a). This model proposed that after the resectioning of the DSB ends, 3' strand invasion, formation of the D-loop structure and repair synthesis phase of the DSBR model the Holliday junctions are not resolved via endonuclease activity. Instead the two junctions migrate toward each other and are resolved by topoisomerases. Therefore this model does not involve cross-over events.

Gene conversion detection methods

Several types of recombination and gene conversion detection methods have been developed over the years. These methods can be divided tentatively into several groups including substitution, phylogenetic, distance and compatibility methods. Previous studies compared 14 recombination detection methods using simulation and empirical

data (Posada and Crandall 2001, Posada 2002). The majority of the best methods evaluated (high power; low false positive rate) were substitution methods (i.e., MaxChi^2 and GENECONV; Posada 2002). I will briefly give an overview of the GENECONV (Sawyer 1999), LIKEWIND (Archibald and Rogers 2002), PhyPro (Weiller 1998), and Reticulate (Jakobsen and Easteal 1996) programs which are examples of substitution, phylogenetic, distance and compatibility methods, respectively. The GENECONV program takes a multiple alignment of sequences and identifies all polymorphic sites; the default algorithm uses polymorphic sites to account for similarity due to evolutionarily conserved sequence motifs. Tracts of identical polymorphic sites between family members are identified. The significance of the putative gene conversions (i.e., unusually long tracts of identical nucleotides) are assessed by comparison of the observed results to a distribution of scores from randomly generated data sets (Sawyer 1999). The advantage to using the GENECONV method as opposed to other detection methods is that this method explicitly identifies the converted genes and the location of the converted regions within those genes. In addition, this method has a relatively low (< 5%) false positive rate for sequences with 5 – 20% sequence divergence (Posada and Crandall 2001). The LIKEWIND phylogenetic method identifies regions within a multiple alignment that display anomalous phylogentic signals. This method uses a maximum likelihood sliding window method to locate the converted sequences. A heuristic algorithm is used to find the most likely tree for each window and compare this with the tree for the entire dataset. The magnitude of the difference between likelihood values indicates the degree of conflict between the topology for the windows versus the entire dataset. Discrete alignment regions with discordant topologies were identified as putative gene

conversions. The significance of these regions was determined using parametric bootstrapping simulations (Archibald and Rogers 2002). The Phylogenetic profile (PhyPro) method uses a phylogenetic correlation metric to determine boundaries of recombination and gene conversion regions. A correlation matrix is calculated for each position in the alignment. A pair-wise hamming distance is calculated for each pair of sequences in the alignment for windows of sequences immediately upstream and downstream of the specific position. The correlation of the up stream and down stream distance vectors for each sequence is calculated. Low correlation of distance values for a specific gene within the alignment indicates a likely recombination participant (Weiller 1998). The final class of detection methods is compatibility method. The Reticulate algorithm was implemented by Jakobsen and Easteal (1996). This method identifies parsimoniously informative sites (i.e., those sites that have two or more nucleotides in two or more sequences each). Pair-wise comparisons of sites are used to define sites as compatible or non-compatible. If two sites support the same phylogeny then they are classified as compatible sites, whereas non-compatible sites do not support the same phylogenetic relationship. A matrix is constructed based on these pair-wise comparisons, clustering of compatible and non-compatible sites are indicative of recombination or gene conversion events. Comparison of the observed matrix against permutations of the compatibility matrix provides an indication of the probability of getting the observed results by random (Jokobsen and Easteal 1996). This provides a measure of the significance of clustered sites.

Characteristics of Gene Conversions

The frequency and size of allelic and ectopic gene conversions are dependent upon several factors. An important factor which affects the search for a template sequence during the DSB repair mechanism is the effect of sequence similarity between the template and the damaged gene. In bacteria the activity of the mismatch repair genes acts as a barrier between the damaged gene and possible templates (Zahrt et al 1994, Rayssiguier et al 1989). In effect the mismatch genes prevent very divergent related genes from being used as the repair template. This recombination barrier attempts to minimize the net sequence change induced in the damaged gene by the repair process. A significant change in the primary sequence of the repaired gene relative to its original undamaged form will likely have adverse effects on the function of the repaired gene. Analyses of gene conversions similarity requirements in *Escherichia coli* indicate a 10-fold decrease in recombination rate if there is 2% sequence divergence between the donor and recipient sequences; this recombination rate decreases by 40-fold if the sequences have 10% sequence divergence (Watt et al 1985, Shen and Huang 1986). A similar trend was found in *C. elegans* by Semple and Wolfe (1999), where a negative correlation between gene conversion frequency and sequence divergence outside the converted region was found. In addition to affecting the probability of recombination and gene conversion between the donor and recipient genes, studies have found that sequence similarity also affects the length of the converted regions. A positive correlation between flanking similarity and the length of converted regions has been found in *S. cerevisiae* and humans (Drouin 2002, Benovoy and Drouin unpublished). The length of the homologous sequences is also a factor on the rate of recombination; in the wild type *E. coli* a threshold length of 23 - 27 bp is required for efficient recombination (Shen and

Huang 1986). This minimum sequence size was found to be an intrinsic property of the DSB repair mechanism in prokaryotes. The distance between the donor and recipient genes affects the rate of gene conversion. In *S. cerevisiae*, gene conversions occurred more frequently between tandemly duplicated genes (80% conversion frequency), whereas the gene conversion frequency decreased to 60% for genes separated by one to four ORFs (Drouin 2002). A similar trend was found in human where the majority of detected ectopic gene conversions occurred between neighbouring genes (Benovoy and Drouin unpublished). Not only is intra - chromosomal distance a factor but inter - chromosomal gene conversions were less frequent than intra - chromosomal gene conversions in yeast and human (Drouin 2002, Benovoy and Drouin unpublished). The orientations of the donor and recipient genes have an effect on the intra - chromosomal gene conversion frequency. Previous work on *Drosophila* found that recombination was more frequent between genes that are in opposite orientations (i.e., inverted). The inverted homologs were able to make hairpin loops which facilitated gene conversions (Wang et al 1999). By contrast in humans and *Arabidopsis thaliana* intra - chromosomal gene conversions occur more frequently between genes which share the same transcriptional orientation (Benovoy and Drouin unpublished, Mondragon-Palomino and Gaut 2005). Previous studies have found that the converted regions within genes are not evenly distributed along the length of the gene. Ectopic gene conversions between linked genes on the same chromosome tend to be uniformly distributed along the length of converted genes; however ectopic conversions between unlinked dispersed genes tend to cluster near the 3' end of the genes (Drouin 2002). This clustering has been attributed to ectopic gene conversions via a cDNA donor. Reverse transcriptase creates cDNA by

using an mRNA molecule as a template. Reverse transcriptase has a low processivity and has a 3'→ 5' polymerase activity; this results in an excess of cDNA molecules identical to the 3' regions of the mRNA template. It is known that the size of multigene families has a positive effect on the rate of gene conversion (Melamed and Kupiec 1992). Therefore it is more likely that repair events will use the most abundant type of template. This causes an excess of gene conversions in the 3' end of genes. Recent studies have found that the recombination rate in humans and *A. thaliana* is correlated to the gene conversion frequency (Benovoy and Drouin unpublished, Mondragon-Palomino and Gaut 2005). Therefore unstable areas which tend to be involved in many recombination events also tend to have more gene conversion events.

Evolutionary consequences of gene conversions

An evolutionary consequence of gene conversions is the homogenization of duplicate genes within a species (Hughes 2000; see Figure 3). This effectively resets the molecular clock and can result in erroneous phylogenetic inferences (Ohta 1980, Ohta 1990). In addition, some species have adapted the use of gene conversion to create new sequence variability. The bacterial pathogens within the *Anaplasma* genus use ectopic gene conversion between an expressed protein coat gene and a family of pseudogenes to create complex mosaic sequences within the hyper variable regions of the expressed gene (Futse et al 2005). This is analogous to the process used by chickens to create immunoglobin variants (McCormack et al 1993) whereby point mutations are re-assorted between family members (Ohta 1991). Gene conversion has been found to influence DNA base composition in the repaired regions (Galtier et al 2001). Studies have found that regions with high amounts of recombination show an accumulation of guanine and

cytosine nucleotides (Birdsell 2002). For example, yeast genomic regions with high recombination rates also have increased GC content (Gerton et al 2000). Studies have determined that genes belonging to multigene families that share high sequence similarity (likely due to gene conversions) also have higher GC content than unique genes (Galtier 2003, Marais 2003). This phenomenon has been attributed to preferential repair of hetero-mismatched bases (AG, TG, AC or TC) to GC base pairs. This biased process may have evolved to oppose the AT mutation bias in eukaryotes (Birdsell 2002).

Main Objectives

The first part of this thesis investigates the characteristics of ectopic gene conversions in four *E. coli* strains. The gene conversions are from multigene families of at least three members using the GENECONV method implemented by Stanley Sawyer (1999). This study examines the differences between ectopic gene conversion characteristics of three pathogenic strains (*E. coli* Sakai, EDL933, and CFT073) with a single non – pathogenic strain (*E. coli* K12). The characteristics examined in this study include gene conversion frequency, sequence similarity requirements for conversion, the effect of distance on the probability of gene conversion frequency between multigene family members as well as the detection of converted region clustering within genes.

In the second study, I investigate the differences between ectopic gene conversions detected within the conserved gene set (i.e., backbone) of the four *E. coli* strains Sakai, EDL933, CFT073 and K12. Multigene families with at least 2 members were used in this analysis. GENECONV was used to detect the ectopic gene conversions. The characteristics of the backbone conversions were compared with those found in the first *E. coli* study discussed in this thesis.

In the third study, I conduct a large scale comparative analysis of ectopic gene conversion characteristics between pathogenic and non – pathogenic proteobacteria, firmicutes and archaea. Ectopic gene conversions from multigene families of at least 3 members were analyzed from 75 prokaryotic genomes included in this study. The gene conversions were detected using the GENECONV program. I looked at the same set of characteristics previously mentioned for the *E. coli* specific study.

The fourth study discussed in this thesis detects and characterizes ectopic gene conversions in ten yeast genomes. The ten yeast genomes are divided into two groups, those that diverged before and after a whole genome duplication event (pre and post whole genome duplication (WGD). The pre-WGD group includes *K. lactis*, *D. hansenii* and *Y. lipolytica*. The post-WGD includes *S. cerevisiae*, *S. paradoxus*, *S. mikatae*, *S. kudriavzevii*, *S. bayanus*, *S. castellii* and *C. glabrata*. Ectopic gene conversions were identified and characterized from species specific multigene families within the pre-WGD genomes. Ectopic gene conversions were identified within species specific and conserved backbone multigene families within the post-WGD genomes. A comparative analysis of ectopic gene conversion characteristics is performed between the groups of genomes.

The final study investigates the effects of ectopic gene conversions on nucleotide composition in yeast and *A. thaliana*. Our discussion of this paper will be based solely on the relationship between the G+C content and sequence similarity of duplicated genes in *S. cerevisiae*. Previous studies found that repeated allelic gene conversion and ectopic gene conversion between tandem duplicated genes tend to increase the G+C content of converted genes. I determine the relationship between G+C content and similarity

between dispersed ohnologs in the *S. cerevisiae* genome created by a whole genome duplication event. Using these data I investigate the relationship between the sequence similarities of dispersed ectopically converted genes and the increase of these sequences G + C content.

Literature Cited

Archibald, J. M., and A. J. Rogers. 2002. Gene conversion at the evolution of euryarcheal chaperonins: a maximum likelihood based method for detecting conflicting phylogenetic signals. J. Mol. Evol. 55:232–245.

Aylon, Y., and M. Kupiec. 2004. DSB repair: the yeast paradigm. DNA Repair 3:797-815.

Balding, D.J., R. A. Nichols, and D. M. Hunt. 1992. Detecting gene conversion: primate visual pigment genes. Proc. R. Soc. Lond. B Biol. Sci. 249:275–280.

Birdsell, J. A. 2002. Integrating genomics, bioinformatics, and classical genetics to study the effects of recombination on genome evolution. Mol. Biol. Evol. 19:1181–1197.

Bowater, R., and A. J. Doherty. 2006. Making ends meet: repairing breaks in bacterial DNA by non-homologous end-joining. PLoS Genet. 2:e8.

Cromie, G. A., J. C. Connelly, and D. R. Leach. 2001. Recombination at double-strand breaks and DNA ends: conserved mechanisms from phage to humans. Mol. Cell 8:1163–1174.

Drouin, G. 2002. Characterization of the gene conversions between the multigene family members of the yeast genome. J. Mol. Evol. 55:14-23.

Dudas, A., and M. Chovanec. 2004. DNA double-strand break repair by homologous recombination. Mutat. Res. 566:131-167.

Fogel, S. and D. D. Hurst. 1967. Meiotic gene conversion in yeast tetrads and the theory of recombination. Genetics 57:455-481.

Futse, J. E., K. A. Brayton, D. P. Knowles Jr, and G. H. Palmer. 2005. Structural basis for

segmental gene conversion in generation of Anaplasma marginale outer membrane protein variants. Mol. Microbiol. 57:212-221.

Galtier, N. 2003. Gene conversion drives GC content evolution in mammalian histones. Trends Genet. 19:65-68.

Galtier, N., G. Piganeau, D. Mouchiroud, and L. Duret. 2001. GC-content evolution in mammalian genomes: the biased gene conversion hypothesis. Genetics 159:907-911.

Gerton, J. L., J. DeRisi, R. Shroff, M. Lichten, P. O. Brown, and T. D. Petes. 2000. Global mapping of meiotic recombination hotspots and coldspots in the yeast *Saccharomyces cerevisiae*. Proc. Natl. Acad. Sci. USA 97:11383-11390.

Hastings, P. J. 1988. Recombination in the eukaryotic nucleus. Bioessays 9:61-64.

Holliday, R. 1964. A mechanism for gene conversion in fungi. Genet. Res. 5:282–290.

Hughes, D. 2000. Co-evolution of the tuf genes links gene conversion with the generation of chromosomal inversions. J. Mol. Biol. 297:355-364.

Jakobsen, I. B. and S. Easteal. 1996. A program for calculating and displaying compatibility matrices as an aid in determining reticulate evolution in molecular sequences. Comput.Appl. Biosci. 12:291-295.

Kobayashi, I., and N. Takahashi. 1988. Double-stranded gap repair of DNA by gene conversion in *Escherichia coli*. Genetics 119:751–757.

Kowalczykowski, S. C., D. A. Dixon, A. K. Eggleston, S. D. Lauder, W. M. Rehrauer. 1994. Biochemistry of homologous recombination in *Escherichia coli*. Microbiol. Rev. 58:401–465.

Lathe III, W. C., and P. Bork. 2001. Evolution of *tuf* genes: ancient duplication,

differential loss and gene conversion. FEBS Lett. 502:113–116.

Liao, D. 2000. Gene conversion drives within genic sequences: concerted evolution of ribosomal RNA genes in bacteria and archaea. J. Mol. Evol. 51:305–317.

Marais, G. 2003. Biased gene conversion: implications for genome and sex evolution. Trends Genet. 19:330-338.

McCormack, W. T., E. A. Hurley, and C. B. Thompson. 1993. Germ line maintenance of the pseudogene donor pool for somatic immunoglobulin gene conversion in chickens. Mol. Cell . Biol 13:821-830.

McGill, C., B. Shafer, and J. Strathern. 1989. Coconversion of flanking sequences with homothallic switching. Cell 57:459-467.

Melamed, C., and M. Kupiec. 1992. Effect of donor copy number on the rate of gene conversion in the yeast *Saccharomyces cerevisiae.* Mol Gen Genet 235:97–103.

Meselson, M. S., and C. M. Radding. 1975. A general model for genetic recombination. Proc Nat. Acad. Sci. USA. 72:358-361.

Mondragon-Palomino, M., and B. S. Gaut. 2005. Gene conversion and the evolution of three leucine-rich repeat gene families in *Arabidopsis thaliana.* Mol. Biol. Evol. 22:2444-2456.

Ohta, T. 1980. Amino acid diversity of immunoglobulins as a product of molecular evolution. J. Mol. Evol. 15:29–35,

Ohta, T. 1990. How gene families evolve. Theor Popul Biol 37:213 – 219.

Ohta, T. 1991. Role of diversifying selection and gene conversion in evolution of major histocompatibility loci. Proc. Natl. Acad. Sci. USA 15:6716–6720.

Orr-Weaver, T. L. and J. W. Szostak. 1983. Yeast recombination: the association

between double-strand gap repair and crossing-over, Proc. Natl. Acad. Sci. U.S.A. 80:4417–4421.

Perkins, D. D. 1992. *Neurospora*: the organism behind the molecular revolution. Genetics 130:687-701.

Posada, D., and K. A. Crandall. 2001. Evaluation of methods for detecting recombination from DNA sequences: Computer simulations. Proc. Natl. Acad. Sci. USA 98:13757-13762.

Posada, D. 2002. Evaluation of Methods for Detecting Recombination from DNA Sequences: Empirical Data. Mol. Biol. Evol. 19: 708-717.

Rayssiguier, C., D. S. Thaler, and M. Radman. 1989. The barrier to recombination between *Escherichia coli* and *Salmonella typhimurium* is disrupted in mismatch-repair mutants. Nature 342:396-401.

Rocha, E. P., A. Danchin and A. Viari. 1999. Functional and evolutionary roles of long repeats in prokaryotes. Res. Microbiol. 150:725–733.

Roman, H. 1956. Studies of gene mutation in *Saccharomyces*. Cold Spring Harb. Symp. Quant. Biol. 21:175–185.

Romero, D., J. Martínez-Salazar, E. Ortiz, C. Rodríguez, E. Valencia-Morales. 1999. Repeated sequences in bacterial chromosomes and plasmids: a glimpse from sequenced genomes, Res. Microbiol. 150:735–743.

Santoyo, G. and D. Romero. 2005. Gene conversion and concerted evolution in bacterial genomes. FEMS Micro. Rev. 29:169-183.

Sawyer, S. A. 1999. GENECONV: A computer package for the statistical detection of

gene conversion. Distributed by the author, Department of Mathematics, Washington University in St. Louis, available at http://www.math.wustl.edu/~sawyer.

Scott, A. F., P. Heath, S. Trusko, S. H. Boyer, W. Prass, M. Goodman et al. 1984. The sequence of the gorilla fetal globin genes: evidence for multiple gene conversions in human evolution. Mol. Biol. Evol. 1:371–389.

Semple, C. and K. H. Wolfe. 1999. Gene duplication and gene conversion in the Caenorhabditis elegans genome. J. Mol. Evol. 48:555-564.

Shen, P. and H. V. Huang. 1986. Homologous recombination in *Escherichia coli:* Dependence on substrate length and homology. Genetics 112:441-457.

Stahl, F. 1996. Meiotic recombination in yeast: coronation of the double-strand-break repair model. Cell 87:965-968.

Szostak, J. W., T. L. Orr-Weaver, R. J. Rothstein, and F. W. Stahl. 1983. The double-strand-break repair model for recombination. Cell 33:25–35.

Wang, S., C. Magoulas, and D. Hickey. 1999. Concerted evolution within a trypsin gene cluster in Drosophila. Mol. Biol. Evol. 16:1117-1124.

Watt, V. M., C. J. Ingles, M. S. Urdea, and W. J. Rutter. 1985. Homology requirements for recombination in *Escherichia coli.* Proc. Natl. Acad. Sci. USA 82:4768-4772.

Weiller, G.F. 1998. Phylogenetic profiles: a graphical method for detecting genetic recombinations in homologous sequences. Mol. Biol. Evol. 15:326-335.

Wyman, C., D. Ristic, and R. Kanaar. 2004. Homologous recombination-mediated double-strand break repair. DNA Repair 3:827-833.

Zahrt, T. C., G. C. Mora, and S. Maloy. 1994. Inactivation of mismatch repair overcomes

the barrier to transduction between Salmonella typhimurium and Salmonella typhi. J. Bacteriol. 176:1527-1529.

Zhou, Y. H., and W. H. Li. 1996. Gene conversion and natural selection in the evolution of X-linked color vision genes in higher primates. Mol. Biol. Evol. 13:780–783.

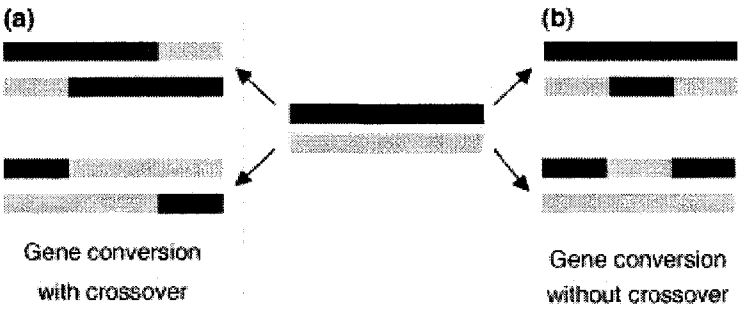


Figure 1. Connection between gene conversions and cross-over events. a) shows gene conversion with an associated cross-over. b) shows gene conversion without an associated reciprocal exchanage. This figure has been reprinted from FEMS Micro. Rev. Vol 29, G. Santoyo and D. Romero. Gene conversion and concerted evolution in bacterial genomes. Pp 169-183 2005, with permission from Blackwell Publishing.

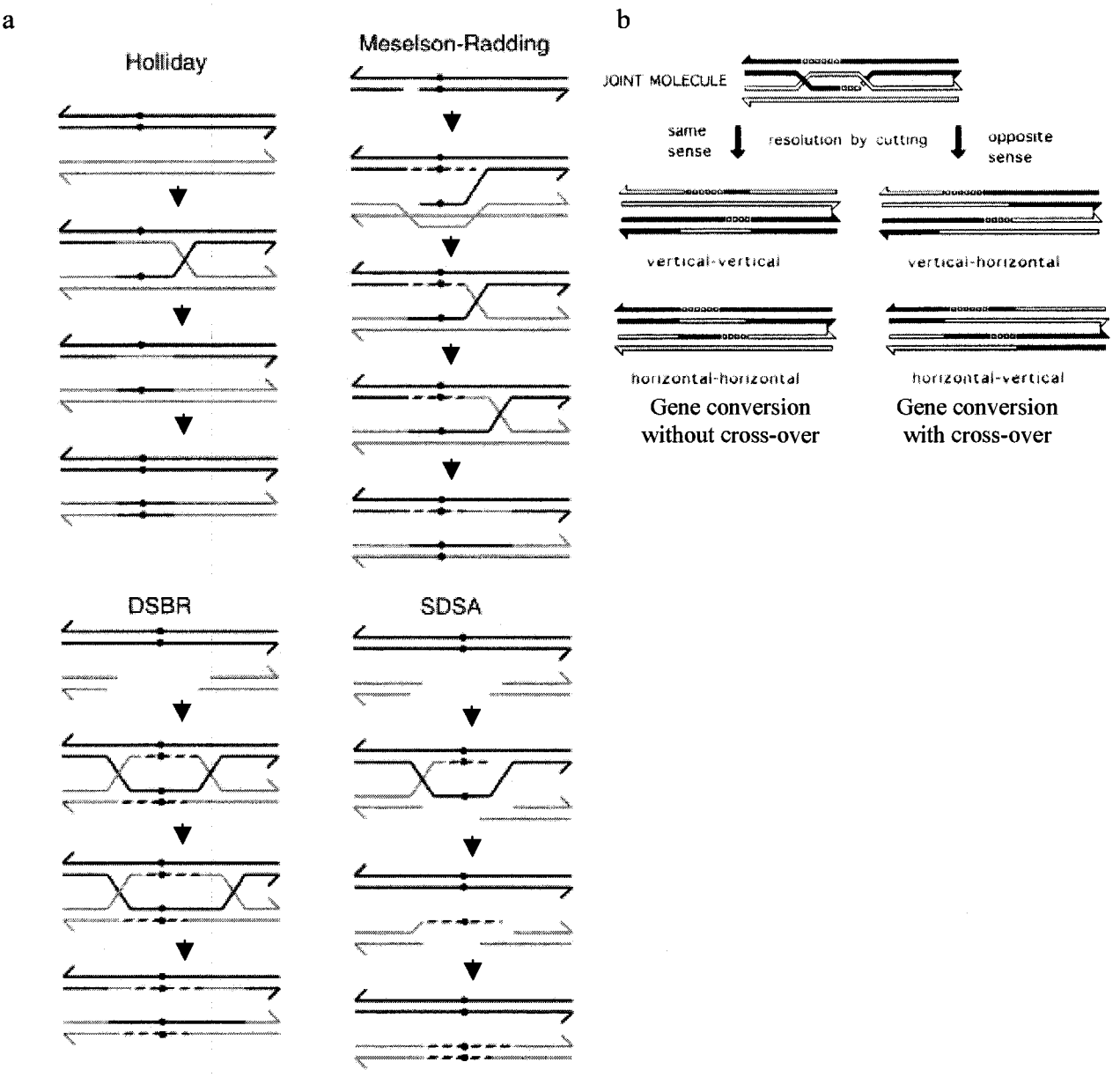


Figure 2a. Molecular models of recombination and gene conversion. Four models are summarized here, the Holliday, Menselson-Radding, double-strand break repair (DSBR), and synthesis-dependent strand annealing (SDSA). Thick and thin black lines indicate parental sequences, the dashed lines indicate newly replicated DNA. Black dots indicate positions which are different between the homologous sequences. This figure has been reprinted from FEMS Micro. Rev. Vol 29, G. Santoyo and D. Romero. Gene conversion and concerted evolution in bacterial genomes. Pp 169-183 2005, with permission from Blackwell Publishing.

Figure 2b shows four different resolutions of Holliday junctions via endonucleases during the DSBR model This figure has been reprinted from Cell Vol 87, F. Stahl.Meitotic recombination in yeast. Pp 965-968, 1996, with permission from Elsevier.

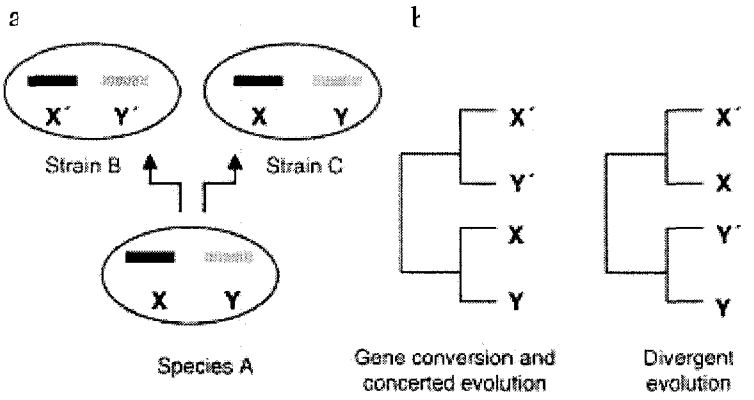


Figure 3. Phylogenetic effect of gene conversion. a) relationship between strain B and C of species A. b) phylogenetic topology effect of gene conversions. These figures have been reprinted from FEMS Micro. Rev. Vol 29, G. Santoyo and D. Romero. Gene conversion and concerted evolution in bacterial genomes. Pp 169-183 2005, with permission from Blackwell Publishing.

Chapter 2

Ectopic gene conversions in four *Escherichia coli* genomes: increased recombination in pathogenic strains.

Robert T. Morris and Guy Drouin

Département de biologie, Université d'Ottawa, Ottawa, Ontario, Canada, K1N 6N5

Research article : J. Mol. Evol. 2004, 58: 596-605

Keywords: ectopic, gene conversion, *Escherichia coli*, mismatch repair genes

Correspondence to: Guy Drouin, Département de biologie, Université d'Ottawa, 150 Louis Pasteur, Ottawa, Ontario, Canada, K1N 6N5. Tel.: (613) 562-5800 ext. 6052, FAX: (613) 562-5486, E-mail: gdrouin@science.uottawa.ca

Abstract. We characterized the ectopic gene conversions in the genomes of the K-12 MG1655, O157:H7 Sakai, O157:H7 EDL933, and CFT073 strains of *E. coli*. Compared to the three pathogenic strains, the K-12 strain has a much smaller number of gene families, its gene families contain fewer genes, and gene conversions are less frequent. Whereas the three pathogenic strains have gene conversions covering hundreds of nucleotides when their flanking regions have as little as 50% similarity, flanking regions similarity of at least 94% on both sides of the converted region is required to observe conversions of more than 87 nucleotides in the K-12 strain. Recombination is therefore more frequent and requires less sequence similarity in the three pathogenic strains than in K-12. This higher recombination level might be due to mutations in some of their mismatch-repair genes.

In contrast with the gene conversions present in the yeast genome, the gene conversions found in the *E. coli* genomes do not occur more frequently between duplicated genes that are close to one-another than between duplicated genes that are far apart and are randomly distributed along the length of the genes. In *E. coli*, gene conversions are not more frequent near the origin of replication. However, they do occur more frequently near the terminus of replication of the Sakai genome, where multigene family members are more abundant. This suggests that, in *E. coli*, gene conversions occur randomly between genes located in different chromosomal locations or located on different copies of the multiple chromosomes found in *E. coli* cells.

Introduction

Gene conversions are non-reciprocal exchanges of genetic information from one gene to another. Different types of gene conversions have been identified. Allelic conversions occur between sequences found at the same locus whereas ectopic conversions occur between sequences found at different chromosomal locations (Petes and Hill 1988). *E. coli*, together with *Saccharomyces cerevisiae*, has been an organism of choice to study the biochemistry of homologous recombination as well as the relationship between gene conversion and homologous recombination and mismatch repair mechanisms (West 1992; Kowalczykowski et al. 1994; West 1994; Lloyd and Low 1996; Pâques and Haber 1999). Allelic recombination is highly dependent on the degree of sequence similarity between the donor and recipient sequence. In *E. coli*, 2% mismatches can decrease the frequency of recombination four-fold and 10% mismatches can decrease recombination by over 40-fold (Watt et al. 1985; Shen and Huang 1986). Allelic recombination is also dependent on the length of the sequences. In *E. coli*, there is an exponential increase in the frequency of recombination when the length of the sequences increases from 20 to 74 base pairs (Watt et al. 1985). In yeast, the frequency of gene conversion has been shown to be proportional to the number of copies of donor sequences present in a cell (Melamed and Kupiec 1992). Gene conversions have also been shown to be more frequent between closely linked genes (Goldman and Lichten 1996; Drouin 2002). Furthermore, the analyses of the gene conversion events found in the *Saccharomyces cerevisiae* genome showed that gene conversions were more frequent at the 3'-end of yeast genes (Drouin 2002). It was also suggested that gene conversion events are not distributed equally along

the length of genomes. Sharp et al. (1989) observed that, in the *E. coli* genome, the substitution rates of genes near the origin of replication (*oriC*) were lower than those close to the terminus of replication. They suggested that this was the result of higher rates of recombination repair due to the higher average copy number of genes located near *oriC*.

Here, we used the complete genome sequence of four *E. coli* strains to perform genome wide analyses of the gene conversions found between their multigene family members. These analyses allowed us to verify the observations previously made on limited numbers of genes and to compare the characteristics of the gene conversions found in the K-12 strain with those of three pathogenic strains. In particular, we analyzed whether converted regions are larger between more similar genes, whether gene conversions are more frequent in larger multigene families, whether gene conversions are more frequent between closely linked genes, whether gene conversions are equally frequent in all regions of the genes and whether gene conversions are more frequent near the origin of replication. We also compared the characteristics observed for these gene conversions with those found in the yeast genome (Drouin 2002).

Materials and Methods

Genome sequences

The four strains used in this study were the K-12 MG1655 (K-12), CFT073, O157:H7 EDL933 (EDL933) and O157:H7 Sakai (Sakai) and were the only full *E. coli* genome sequences available in the fall of 2002. The genome of the K-12 laboratory strain was sequenced by Blattner et al. (1997). The CFT073 strain is an uropathogenic strain isolated at the University of Maryland Hospital and sequenced by Welch et al. (2002). The EDL933 strain was responsible for an outbreak of haemorrhagic colitis in Michigan in 1982 and was sequenced by Perna et al. (2001). The Sakai strain was responsible for a different outbreak of haemorrhagic colitis that occurred in Sakai city Japan in 1996 and its genome was sequenced by Hayashi et al. (2001). The *E. coli* genomes and the BLASTCLUST program were obtained from the NCBI ftp site (ftp://ftp.ncbi.nlm.nih.gov).

Multigene families

The BLASTCLUST program was used to identify all multigene family members within each genome (excluding rRNA and tRNA gene families). Genes were included in a family if their protein sequences were at least 60% identical over at least 50% of their lengths. We chose a protein sequence similarity of 60% because, even though BLASTCLUST analyses of the K-12 genome with 40% and 50% similarity levels did identify more multigene families containing more members, the extra gene conversions

detected were all smaller than 41 bp (results not shown). ClustalW (Thompson et al. 1994) was used to align the protein sequences and Will Fischer's ALIGN2AA Perl script was used to align the corresponding DNA sequences (http://sunflower.bio.indiana.edu/~wfischer/Perl_Scripts/).

Gene conversion analyses

The GENECONV 1.7 (http://www.math.wustl.edu/~sawyer/geneconv/) program was used to identify the gene conversions within each gene family (Sawyer 1989). Previous studies assessed the power (type II errors) and the rate of false positives (type I errors) of fourteen different recombination detection methods using simulated and empirical data sets (Posada and Crandall 2001, Posada 2002). The GENECONV method performed well with both simulated and empirical data. It was always one of the most powerful methods, and had a type I error rate of about 5%, when the sequences diverged by a minimum of 5%. Although no single method can guarantee that all gene conversion events are detected, using the same method with different genomes allowed us to compare the general characteristics of their gene conversions.

Only the families containing three or more members were analyzed because the GENECONV method is more reliable when used with more than two sequences (Sawyer 1989; Drouin 2002). The user defined g-scale parameter was set to 2 to allow for the presence of some nucleotide differences in the converted regions. Only global inner fragments with *p*-values smaller than 0.05 were included in the analysis. Duplicate gene

conversions were removed from the data sets using phylogenetic analyses of each gene family as previously described (Drouin 2002).

The percent sequence similarity flanking converted regions was calculated from the 100 bp upstream and downstream of the converted regions found in the coding regions of the genes and ignored gapped regions. Statistical analyses (data transformation, the Kolmogorov-Smirnov test of normality, correlation and linear regression) were performed using S-plus v6.1 (Insightful Corporation, Seattle, WA, USA) and Excel (Microsoft Corporation, Redmond, WA, USA).

Amino acid substitutions

We compared the protein sequences of the genes involved in DNA recombination and repair in the four *E. coli* strains. The DNA sequences of the pathogenic strains were retrieved from their corresponding genome sequence files using FASTA (Pearson and Lipman 1988) and the appropriate K-12 gene sequence as the query. These groups of DNA sequences were aligned using ClustalW, translated using GDE (Smith 1994) and visually inspected to identify differences between the K-12 sequences and the corresponding sequences in the three pathogenic strains.

Distance and location of duplicated and converted genes

The distances and chromosomal locations were calculated using the *.ptt files from the NCBI ftp site. The distances between coding regions took into account the circular nature of these genomes. Therefore, *n* distances were calculated for each family containing *n*

genes. The distance between converted genes corresponds to the smallest distance between them. The chromosomal location of each gene was calculated relative to a fixed point on the chromosome, the origin of replication (*oriC*). The origin of replication for each strain started at nucleotide 3923640, 4427045, 4788437 and 4719456 for the K-12, CFT073, EDL933 and Sakai genomes, respectively. We tested whether gene conversions were more frequent in different gene regions by calculating whether the distribution of gene conversion mid-points fitted a Poisson (random) distribution. The gene sequences were divided into bins of equal size representing percentages of the total length of the gene. The number of bins was calculated based on the total number of conversions such that each bin was predicted to contain approximately 5 conversion events; see Zar 1999 for more details on the test methodology (Zar 1999, pages 574-578).

Results

Number of gene families, genes and gene conversions

The CFT073, EDL933 and Sakai strains contain more multigene families than the K-12 strain (Fig. 1). The BLASTCLUST program identified 104, 196, 297 and 241 multigene families in the K-12, CFT073, EDL933 and Sakai genomes, respectively. These families can be divided into two classes. The first class contains 81, 143, 204 and 151 families with only 2 members. The second class contains 23, 53, 93, and 90 multigene families with more than two members. These multigene families contain a maximum of 11, 19, 21 and 18 genes in K-12, CFT073, EDL933 and Sakai respectively (Fig. 1).

The multigene families with more than two members of the K-12, CFT073, EDL933 and Sakai genomes contained 17, 52, 150 and 230 gene conversions, respectively (supplemental tables 1-4, found on CD). The gene conversion frequency (number of gene conversion / total number of gene comparisons) in K-12, CFT073, EDL933 and Sakai genomes were 7.6% (17/224), 10.4% (52/497), 10.9% (150/1381) and 16.3% (230/1409), respectively. Alternatively, gene conversion frequency based on the total number of detected paralogous genes in each species were 16.3% (17/104), 23.1% (52/225), 32.3% (150/465) and 48.5% (230/474) for K12, CFT073, EDL933 and Sakai respectively. Both sets of results indicate that pathogenic strains have a higher conversion frequency than the non pathogenic strain.

A Spearman rank correlation test was used to analyze the relationship between the size of a multigene family and the number of conversions found. Significant positive correlations ($p < 0.001$) with r^2-values of 0.0028, 0.13, 0.15, and 0.11 were obtained for

the K-12, CFT073, EDL933 and Sakai genomes, respectively. A significant positive correlation ($p < 0.001$) was also obtained from the yeast genome data set ($r^2 = 0.17$; Drouin 2002). Note that the gene conversion detection rate is likely independent of gene family size because analyses performed on larger data sets of K-12 genes, made using 40 and 50% protein sequence similarity, did not detect more conversions larger than 41 nucleotides long (results not shown).

Size of gene conversions and the similarity of their flanking regions

The distributions of the number of gene conversions with respect to the size of the converted regions for each strain analyzed are shown in Fig. 2. The majority (> 80%) of the gene conversions in all four *E. coli* strains contained converted regions less than 500 bp.

The mean converted length (± standard deviation), the median converted length, the length of the smallest conversion and the length of the largest conversion were 483 ± 890 bp, 47 bp, 12 bp and 3422 bp for the K-12 genome, 201 ± 207 bp, 155 bp, 6 bp and 917 bp for the CFT073 genome, 203 ± 240 bp, 135 bp, 4 bp and 1479 bp for the EDL933 genome, and 300 ± 477 bp, 158 bp, 9 bp and 3704 bp for the Sakai genome. The mean converted lengths are not statistically different between the different genomes. For example, the mean converted length of the most different genomes, those of K-12 and CFT073, are not statistically different ($0.2 < p < 0.5$; *t*-test with unequal variance).

Fig. 3 shows the relationship between the length of each converted region and the similarity of its respective pair of flanking regions in the four *E. coli* genomes. These figures show the data for both the mean sequence similarity found at both ends of the

converted regions and the highest flanking similarity found in the coding regions of the genes. These two sets of data are shown because several conversions in the *E. coli* genomes had a large disparity between their 5'- and 3'-flanking region similarities and that averaging these two values might mask possible correlations between conversion lengths and flanking-region similarity. In the K-12 genome, only small conversions (<87 bp) were observed within the range of 22% to 93% flanking-region similarity. Large conversions, with lengths varying from 173 to 3422 bp, are only observed between sequences having more than 93% similarity (Fig. 3a). In this genome, there is a significant correlation between the highest sequence similarity of the flanking regions and conversion lengths when the conversion lengths data is log-transformed to fit a normal distribution ($r^2 = 0.52$, $p = 0.0011$). In CFT073, the length of gene conversions shows a steep increase above 80% highest flanking-region similarity (Fig. 3b). In this genome, there is a significant correlation between the highest sequence similarity of the flanking regions and conversion lengths when the conversion lengths data is square root-transformed to fit a normal distribution ($r^2 = 0.38$, $p = 1.16 \times 10^{-6}$). In EDL933, the length of gene conversions shows a gradual increase from 28% to 100% flanking-region similarity (Fig. 3c). In this genome, there is a weak but significant correlation between the highest sequence similarity of the flanking regions and conversion lengths when the conversion lengths data is log-transformed to fit a normal distribution ($r^2 = 0.08$, $p = 0.0003$). The gene conversion lengths in the Sakai strain have a bell-shaped distribution and the largest conversions in this strain are associated with sequences having intermediate levels of similarity (Fig. 3d). In this genome, there is no correlation between the highest sequence similarity of the flanking regions and conversion lengths when the

conversion lengths data is log-transformed to fit a normal distribution ($r^2 = 0.01$, $p = 0.09$).

Mutated genes in the three pathogenic strains

We compared the sequences of the genes involved in recombination (Kowalczykowski et al. 1994; Lloyd and Low 1996) in the three pathogenic strains with the corresponding genes in the K-12 strain to determine whether these genes had obvious substitutions that might correlate with the greater amount of recombination observed in the pathogenic strains. The recE, recF, recO, recR, ruvA, ruvC, gyrB, lexA, ssb, urvD, dam, and dut protein sequences are identical in all four strains and therefore cannot explain the greater amount of recombination observed in the pathogenic strains (Table 1). At the other extreme, in the CFT073 genome, the *recT* gene is absent and the *polA* gene has a frameshift at amino acid position 226 that results in a polA protein that is only 28% of the size of the K-12 polA protein (Table 1). Note that the *polA* gene of the CFT073 strain is also listed as a pseudogene in the NC_004431.gbk GenBank file of the genome of this strain.

The recombination gene sets of other two pathogenic strains do not show any missing genes or frame-shifts (Table 1). On the other hand, many of their genes show a variety of amino acid substitutions. Some of these substitutions are likely neutral. For example, the *recA* gene of the EDL933 strain has a single conservative amino acid substitution at position 140 (asparagine in K-12 and aspartic acid in EDL933; score of 1 in the log odds BLOSUM 90 amino acid substitution scoring matrix, Henikoff and Henikoff 1992). In contrast, some genes are very different between K-12 and the three

pathogenic strains. For example, the three pathogenic strains have several amino acid substitutions in their *recB* gene (Table 1). Eight of these substitutions are common to all three pathogenic strains. They include a relatively drastic substitution of a serine to a leucine (log odds score of -3 at position 884; Table 1).

The *uvrD* (*mutU*) genes of the four *E. coli* strains are identical and the *mutH* gene of the CFT073 strain differs by a single conservative amino acid change (log odds score of 1) from that of the other three strains (Table 1). A drastic glycine to glutamic acid substitution (log odds score of -3) at position 337 of the *mutS* gene is common to all three pathogenic strains (Table 1). The *mutL* genes of the three pathogenic strains have several amino acid substitutions when compared to the K-12 gene, including a common drastic leucine to proline substitution (log odds score of -4) at position 417 (Table 1).

Effect of chromosomal location of converted genes and gene conversion distribution

An analysis of the distance between converted genes was performed to determine whether there was any relationship between their proximity and the frequency of gene conversion. Table 2 shows the distribution of the distance between converted genes, the distribution of the distance between multigene family members and the frequency of conversions given the number of duplicated genes separated by the corresponding distance (in bins of 200 kb). Spearman rank tests show that there is no significant relationship ($p > 0.5$) between the proximity of converted genes and the frequency of gene conversion in all four *E.coli* genomes (r^2 of 0.08, 0.19, 0.03, and 0.04 for the K-12, CFT073, EDL933 and Sakai genomes, respectively). Conversions are therefore not more

frequent between genes that are close to one-another than between genes that are far apart.

We also examined the distribution of multigene family members and converted genes along the chromosomes. Table 3 shows the distribution of these genes relative to the origin of replication (*oriC*). Both the number of converted genes and the frequency of conversion (i.e., the number of converted genes divided by the number of multigene family members) are randomly distributed along the chromosomes of the K-12 and CFT073 genomes (Spearman rank tests; r^2 of 0.03 ($p > 0.5$) and 0.02 ($p = 0.24$) for the K-12 genome, respectively; r^2 of 0.01 ($p > 0.5$) and 0.00 ($p = 0.41$) for the CFT073 genome, respectively). In contrast, the number of converted genes is positively correlated with the distance from *oriC* in the EDL933 and Sakai genomes (r^2 of 0.65 and 0.53, $p < 0.001$, respectively). However, the frequency of conversion is significant in the Sakai genome ($r^2 = 0.35$, $p = 0.02$) but not in the EDL933 genome ($r^2 = 0.24$, $p = 0.08$).

The distribution of the conversions within genes follows a Poisson distribution (*p*-values of 0.98, 0.08, 0.44, and 0.27 for the K-12, CFT073, EDL933, and Sakai genomes, respectively). The conversions are therefore distributed equally along the length of genes in all four *E. coli* genomes.

Discussion

The number of gene families, the maximum number of genes within families and the frequency of gene conversion are different between the K-12 strain and the three pathogenic strains. The K-12 strain has a much smaller number of gene families than the pathogenic strains (23 in K-12 versus 53, 93 and 90 in the pathogenic strains; Fig. 1). The K-12 gene families contain fewer genes than those of the pathogenic strains (a maximum of 11 genes in K-12 versus a maximum of 19, 21 and 18 genes in the pathogenic strains). Gene conversions are also less frequent in K-12 than in the pathogenic strains (a gene conversion frequency of 7.6% in K-12 versus 10.4, 10.9 and 16.3% in the pathogenic strains; Fig. 2).

The relationship between flanking sequence similarity and the length of the converted regions is also different between the K-12 strain and the three pathogenic strains (Fig. 3). In K-12, flanking regions similarity of at least 94 % on both sides of the converted region is required to observe conversions of more than 87 bp. These results are consistent with previous experimental studies which showed that mismatches within an homologous fragment can decrease recombination by at least 40-fold in K-12 (Watt *et al.* 1985; Shen and Huang 1986; Matic *et al.* 1995). In contrast, the three pathogenic strains have gene conversions covering hundreds of bp even when their flanking regions have as little as 50% similarity (Fig. 3).

The relationship between gene conversion length and flanking-region similarity of the Sakai strain is unusual. Whereas these two parameters are, as expected, correlated in all three other strains, the length of the conversions in the Sakai strain is maximal at

intermediate flanking-region similarity (Fig. 3d). This suggests that the genome of this strain might contain mutations inhibiting large conversion events between highly similar sequences.

These characteristics suggest that recombination is more frequent and requires less sequence similarity in the three pathogenic strains than in K-12. This higher recombination level in the pathogenic strains might be due to mutations in mismatch repair genes. Rayssiguier et al. (1989) showed that mutations in the *mutL, mutS, mutH* and *mutU* mismatch-repair genes increased recombination 30- to more than 1000-fold between sequences that were ~ 20% divergent. LeClerc et al. (1996) showed that such mutations are frequent among isolates of pathogenic *E. coli* and argued that these mutations could be advantageous to pathogens because they would allow them to escape immune surveillance or elude antibiotics. The study of Reid et al. (2000) is consistent with this interpretation. They showed that different pathogenic strains of *E. coli* (including O157:H7 strains) acquired the same virulence factors independently and argued that this was a result of their enhanced ability to recombine due to defective mismatch repair. Therefore the higher recombination level we observed in pathogenic strains might be due to the fact that these strains were selected because they had mutated mismatch-repairs genes which allow them to escape the immune system. Since these genes usually limit recombination, the mutant strains are also likely to experience increased recombination. However, the higher recombination level of pathogenic strains might also be due to higher expression of the *recA* gene (Matic et al. 1995).

We compared the sequences of the recombination genes in the four *E. coli* genomes to determine whether the genes of the pathogenic strains had obvious mutations

that might correlate with the greater amount of recombination observed in their genomes. The only obvious structural mutations we observed were in the CFT073 strain. This strain lacks a *recT* gene and its *polA* gene has a frame-shift that results in a polA protein that is 72% shorter than the polA protein of the other three strains (Table 1). The recT protein, like the recA protein, catalyses the pairing of complementary single-stranded DNA strands whereas DNA polymerase I fills in the gaps created during recombination and repair (Kowalczykowski et al. 1994). These mutations, in genes involved in DNA recombination and repair, could therefore be responsible for the larger amount of recombination observed in the genome of this pathogenic strain. In fact, Konrad (1997) has shown that *polA* mutants have increased recombination between chromosomal duplications. On the other hand, we cannot rule out the possibility that the elevated recombination of this strain is also due to some of the amino acid substitutions in the other recombination and repair genes of this strain relative to the K-12 strain (Table 1).

Although our analyses do not allow us to pinpoint the reasons for the elevated recombination in the other two pathogenic strains, they can be used to make testable predictions. For example, the absence of mutations in the *uvrD* gene, and the observation that the *mutH* gene of the CFT073 strain differs by a single conservative amino acid change from that of the other three strains, suggest that the elevated recombination observed in the three pathogenic strains is not due to mutations in these two genes (Table 1). On the other hand, it would be interesting to test whether the drastic *mutS* and *mutL* substitutions present in all three pathogenic strains could be responsible for part of the elevated recombination observed in these three strains (Table 1). Obviously, the elevated

amount of recombination observed in pathogenic strains might also be due to the expression level of genes involved in recombination, repair and replication.

Although the mean size of gene conversions are larger in K-12 than in the three pathogenic strains, these differences are not statistically significant (see above).

A dependence of the gene conversion frequency on the size of the multigene family was detected in all four *E. coli* genomes as well as the yeast genome. Therefore, as one would expect, an increase in the number of related copies of a gene in a genome increases the probability of gene conversion between them.

Conversions are not more frequent between genes separated by less than 200 kb when compared to genes separated by more than 200 kb (Table 2). There is therefore no relationship between the proximity of *E. coli* genes and their frequency of gene conversion. This contrasts with yeast genes where the frequency of gene conversion has been shown, both experimentally and by sequence analyses, to increase as the distance between duplicated gene decreases (Goldman and Lichten 1996; Drouin 2002). The absence of such a relationship in *E. coli* suggest that, in this species, gene conversions occur randomly between genes located in different chromosomal locations. Alternatively, it could indicate that gene conversions occur between genes located on different chromosomes. Although we often think of *E. coli* as having a single copy of its chromosome, Åkerlund et al. (1995) showed that there are 2, 4 or 8 genome equivalents per *E. coli* cell during stationary phase and as many as 18 genome equivalents per *E. coli* cell during early exponential growth (Bendich and Drlica 2000). Therefore, under these conditions, a duplicated gene is more likely to pair and recombine with another

duplicated gene on a different chromosome than with a duplicated gene on the same chromosome.

Sharp et al. (1989) showed that *E. coli* genes close to *oriC* evolved half as fast as genes located furthest away from it. They suggested that the lower substitution rate of the genes located close to *oriC* was the result of higher rates of recombination repair due to their higher average copy number during replication. The studies of Mira and Ochman (2002) and Daubin and Perrière (2003), based on completely sequenced genomes, also found that the substitution rate of genes close to the origin of replication was smaller than the substitution rate of genes near the terminus of replication. However, they concluded that this difference was due to higher rates of substitutions near the terminus of replication. It was therefore of interest to see if the frequency of ectopic gene conversions was higher near *oriC*. Our results show that converted genes are not more abundant near *oriC* (Table 3). Therefore, the lower substitution rate observed near the origin of replication cannot be explained by more frequent gene conversions.

Table 3 also shows that the number of converted genes increases with the number of multigene family members. Therefore, the location of gene conversion events is not determined by the distance from *oriC* but by the location of the multigene family members. Thus, the excess of gene conversions near the terminus of replication of the EDL933 and Sakai strains is due to the excess of multigene family members in this region. This biased distribution of multigene family members has already been observed in these two strains. The studies of Ohnishi et al. (1999) and Perna et al. (2001) showed that the increased genome size of these two strains, relative to the K-12 strain, was mainly due to insertions near the terminus of replication.

Comparison of the gene conversions observed in the K-12 and *Saccharomyces cerevisiae* genomes reveals interesting patterns. The relationship between the size of the gene conversions and the similarity between the sequences shows marked differences between these genomes. In yeast, the maximum size of gene conversions gradually increases from about 100 bp to about 1200 bp when the similarity of the sequences increases from 60% to 100% (Drouin 2002). In contrast, K-12 gene conversions larger than 87 bp all required flanking sequences that were at least 94% similar (Fig. 3a). This suggests that the mismatch-repair systems of K-12 are much more efficient than those of yeast. Gene conversions in these two genomes also differ in their relationship with the chromosomal location of genes. In yeast, the frequency of gene conversion increases as the distance between duplicated genes decreases (Drouin 2002). In contrast, in *E.coli*, nearby paralogous genes are not converted more frequently than paralogous genes located far apart (Table 2). As discussed above, the absence of such a correlation in *E.coli* might be due to the presence of multiple chromosomes in *E.coli* cells. The location of these converted regions were also different in these two species. In yeast, we previously showed that conversions are more frequent at the 3'-end of genes and we suggested that this bias was the result of recombination with incomplete cDNA molecules (Drouin 2002). In contrast, the gene conversions observed in *E.coli* are equally distributed between all gene regions. Such an equal distribution suggests that, in *E. coli*, gene conversion occur between randomly paired chromosomal DNA sequences (located either on the same chromosome or on different chromosomes).

In summary, our results clearly show that gene conversions are more frequent and much less dependent on sequence similarity in pathogenic strains than in K-12. These

differences could be the result of mutations in mismatch-repair genes in the pathogenic strains. Our results also show that, in *E. coli*, the frequency of gene conversion does not increase as the distance between duplicated genes decreases and that gene conversions occur equally in all gene regions. Both of these characteristics might be the result of recombination between genes located on different *E. coli* chromosomes. Finally, comparisons between the characteristics of the gene conversions observed in K-12 and those of yeast allowed us to identify the characteristics specific to each of these genomes.

Acknowledgements. We would like to thank Mehrdad Hajibabaei (Biology Department, University of Ottawa) and anonymous referees for their constructive comments on a previous version of this manuscript. This work was supported by a discovery grant from the Natural Sciences and Engineering Research Council of Canada to G.D.

Literature Cited

Åkerlund T, Nordström K, Bernander R (1995) Analysis of cell size and DNA content in exponentially growing and stationary-phase batch cultures of *Escherichia coli*. J Bacteriol 177:6791-6797

Bendich AJ, Drlica K (2000) Prokaryotic and eukaryotic chromosomes: what's the difference? BioEssays 22:481-486

Blattner FR, Plunkett III G, Bloch CA, Perna NT, Burland V, Riley M, Collado-Vides J, Glasner JD, Rode CK, Mayhew GF, Gregor J, Davis NW, Kirkpatrick HA, Goeden MA, Rose DJ, Mau B, Shao Y (1997) The complete genome sequence of *Escherichia co*li K-12. Science 277:1453-1462

Daubin V, Perrière G (2003) G+C3 structuring along the genome: a common feature in prokaryotes. Mol Biol Evol 20:471-483

Drouin G (2002) Characterization of the gene conversions between the multigene family members of the yeast genome. J Mol Evol 55:14-23

Goldman AS, Lichten M (1996) The efficiency of meiotic recombination between dispersed sequences in *Saccharomyces cerevisiae* depends upon their chromosomal location. Genetics 144:43-55

Hayashi T, Makino K, Ohnishi M, Kurokawa K, Ishii K, Yokoyama K, Han CG, Ohtsubo E, Nakayama K, Murata T, Tanaka M, Tobe T, Iida T, Takami H, Honda T, Sasakawa C, Ogasawara N, Yasunaga T, Kuhara S, Shiba T, Hattori M, Shinagawa H (2001) Complete genome sequence of enterohemorrhagic *Escherichia coli* O157:H7 and genomic comparison with a laboratory strain K-12. DNA Res 8:11-22

Henikoff S, Henikoff JG (1992) Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci USA 89:10915-10919

Konrad EB (1977) Method for the isolation of *Escherichia coli* mutants with enhanced recombination between chromosomal duplications. J Bacteriol 130:167-172

Kowalczykowski SC, Dixon DA, Eggleston AK, Lauder SD, Rehrauer WM (1994) Biochemistry of homologous recombination in *Escherichia coli.* Microbiol Rev 58:401-465

LeClerc JE, Li B, Payne WL, Cebula TA (1996) High mutation frequencies among *Escherichia coli* and *Salmonella* pathogens. Science 274:1208-1211

Lloyd RG, Low KB (1996) Homologous recombination. In: Neidhardt FC, Curtiss III R, Ingraham JL, Lin ECC, Low KB, Magasanik B, Reznikoff WS, Riley M, Schaechter M, Umbarger HE (eds) *Escherichia coli* and *Salmonella*: cellular and molecular biology, 2nd edition, American Society for Microbiology Press, Washington D.C., pp. 2236-2255

Mira A, Ochman H (2002) Gene location and bacterial sequence divergence. Mol Biol Evol 19:1350-1358

Matic I, Rayssiguier C, Radman M (1995) Interspecies gene exchange in bacteria: the role of SOS and mismatch repair systems in evolution of species. Cell 80:507-515

Ohnishi M, Tanaka C, Kuhara S, Ishii K, Hattori M, Kurokawa K, Yasunaga T, Makino K, Shinagawa H, Murata T, Nakayama K, Terawaki Y, Hayashi T (1999) Chromosome of the enterohemorrhagic *Escherichia coli* O157:H7; comparative analysis with K-12 MG1655 revealed the acquisition of large amount of foreign DNAs. DNA Res 6:361-368

Pâques F, Haber JE (1999) Multiple pathways of recombination induced by double-strand breaks in *Saccharomyces cerevisiae*. Microbiol. Mol Biol Rev 63:349-404

Pearson WR, Lipman DJ (1988) Improved tools for biological sequence comparison. Proc Natl Acad Sci USA 85:2444-2448

Perna NT, Plunkett III G, Burland V, Mau B, Glasner JD, Rose DJ, Mayhew GF, Evans PS, Gregor J, Kirkpatrick HA, Pósfai G, Hackett J, Klink S, Boutin A, Shao Y, Miller L, Grotbeck EJ, Davis NW, Lim A, Dimalanta ET, Potamousis KD, Apodaca J, Anantharaman TS, Lin J, Yen G, Schwartz DC, Welch RA, Blattner FR (2001) Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. Nature 409:529-533

Petes TD, Hill CW (1988) Recombination between repeated genes in microorganisms. Annu Rev Genet 22:147-168

Posada D, Crandall KA (2001) Evaluation of methods for detecting recombination from DNA sequences: Computer simulations. Proc Natl Acad Sci USA 98:13757-13762

Posada D (2002) Evaluation of Methods for Detecting Recombination from DNA Sequences: Empirical Data. Mol Biol Evol 19: 708-717

Rayssiguier C, Thaler DS, Radman M (1989) The barrier to recombination between *Escherichia coli* and *Salmonella typhimurium* is disrupted in mismatch-repair mutants. Nature 342:396-401

Reid SD, Herbelin CJ, Bumbaugh AC, Selander RK, Whittam TS (2000) Parallel evolution of virulence in pathogenic *Esherichia coli*. Nature 406:64-67

Sawyer SA (1989) Statistical tests for detecting gene conversions. Mol Biol Evol 6:526-538

Sharp PM, Shields DC, Wolfe KW, Li WH (1989) Chromosomal location and

evolutionary rate variation in enterobacterial genes. Science 246:808-810

Shen P, Huang HV (1986) Homologous recombination in *Escherichia coli*: dependence on substrate length and homology. Genetics 112:441-457

Smith SW, Overbeek R, Woese CR, Gilbert W, Gillevet P (1994) The genetic data environment: an expandable GUI for multiple sequence analysis. Compt Appl Biosci 10:671-675

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22:4673-4680

Watt VM., Ingles CJ, Urdea MS, Rutter WJ (1985) Homology requirements for recombination in *Escherichia coli*. Proc Natl Acad Sci USA 82:4768-4772

Welch RA, Burland V, Plunkett III G, Redford P, Roesch P, Rasko D, Buckles EL, Liou SR, Boutin A, Hackett J, Stroud D, Mayhew GF, Rose DJ, Zhou S, Schwartz DC, Perna NT, Mobley HLT, Donnenberg MS, Blattner FR (2002) Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*. Proc Natl Acad Sci USA 99:17020-17024

West SC (1992) Enzymes and molecular mechanisms of genetic recombination. Annu Rev Biochem 61:603-640

West SC (1994) The processing of recombination intermediates: mechanistic insights from studies of bacterial proteins. Cell 76:9-15

Zar, JH (1999) Biostatistical Analysis. Fourth edition, Prentice Hall, Upper Saddle River, New Jersey

Figure legends

Fig. 1. Number of multigene families containing three or more members in the genome of the four *E. coli* strains. These multigene families consist of genes coding for proteins that are at least 60% similar over at least 50% of their length.

Fig. 2. Size of the gene conversions found in the genome of the four *E. coli* strains.

Fig. 3. Relationship between the length of each converted region and the similarity of its flanking regions found within the coding region of the genes in the four *E. coli* genomes. a, K-12; b, CFT073; c, EDL933; d, Sakai. Circles (○) represent the mean percent similarity found 100 bp upstream and downstream of the converted regions. Plus signs (+) represent the highest flanking similarity found either 100 bp upstream or downstream of the converted regions.

Fig. 1.

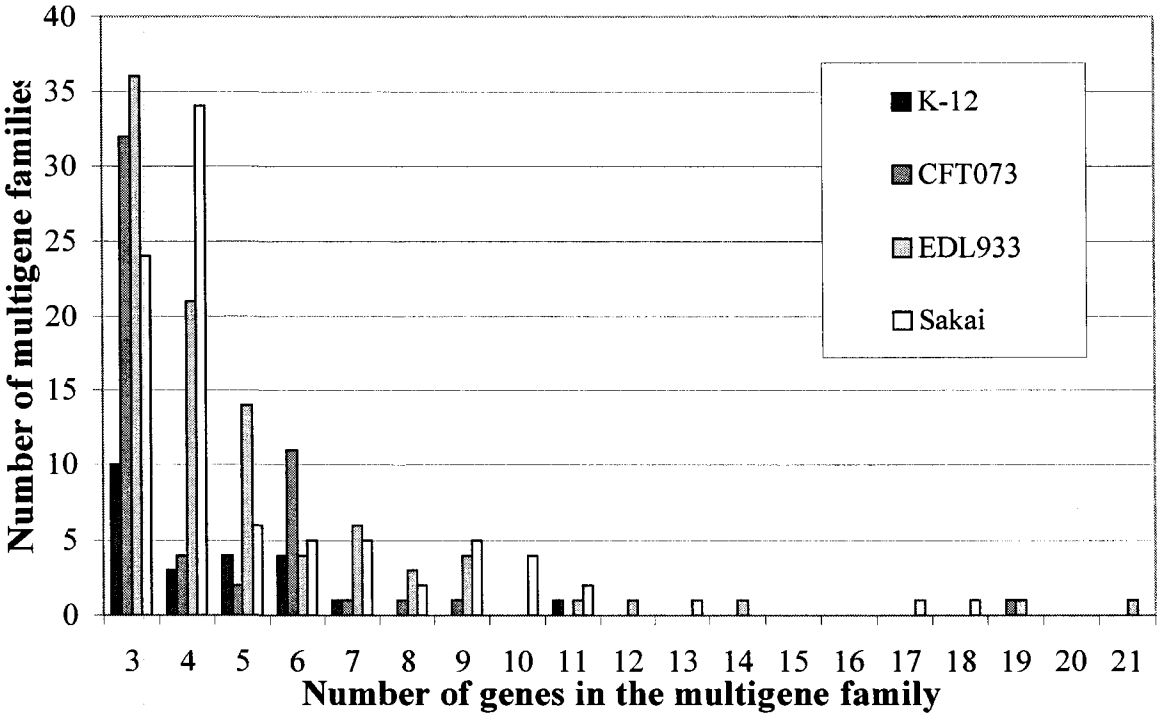

Number of multigene families
40
35
30
25
20
15
10
5
0
K-12
CFT073
EDL933
Sakai
3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21
Number of genes in the multigene family

Fig. 2.

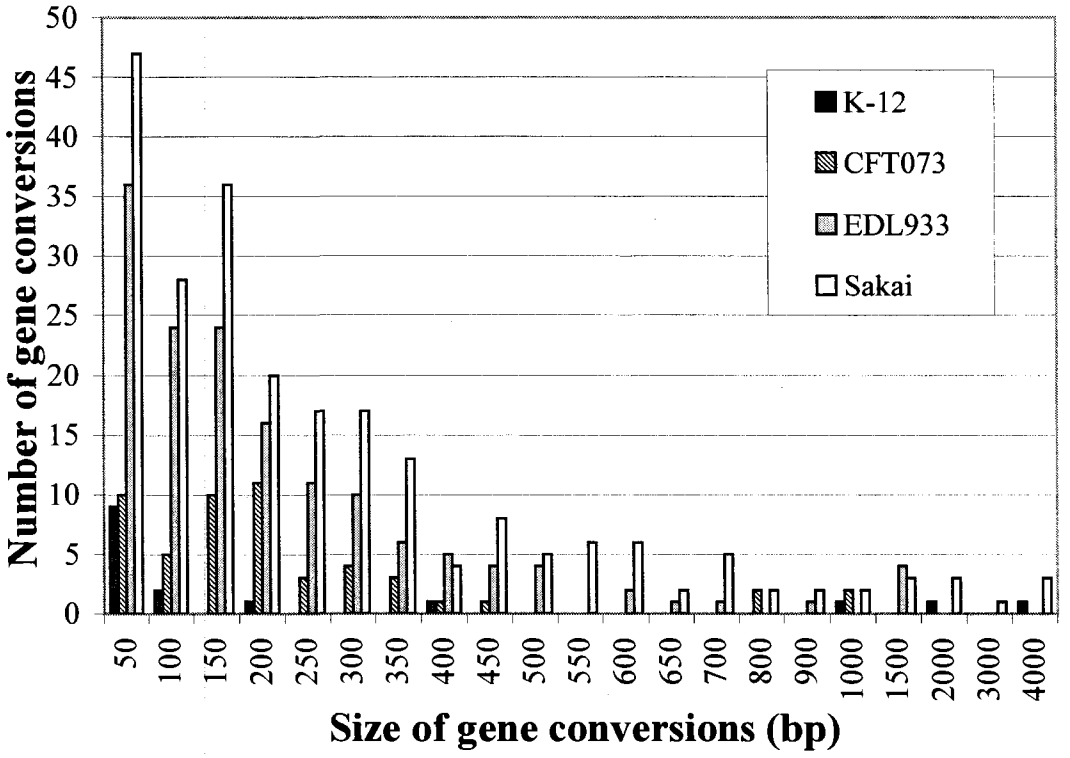

Number of gene conversions
0
5
10
15
20
25
30
35
40
45
50
K-12
CFT073
EDL933
Sakai
50
100
150
200
250
300
350
400
450
500
550
600
650
700
800
900
1000
1500
2000
3000
4000
Size of gene conversions (bp)

Fig. 3.

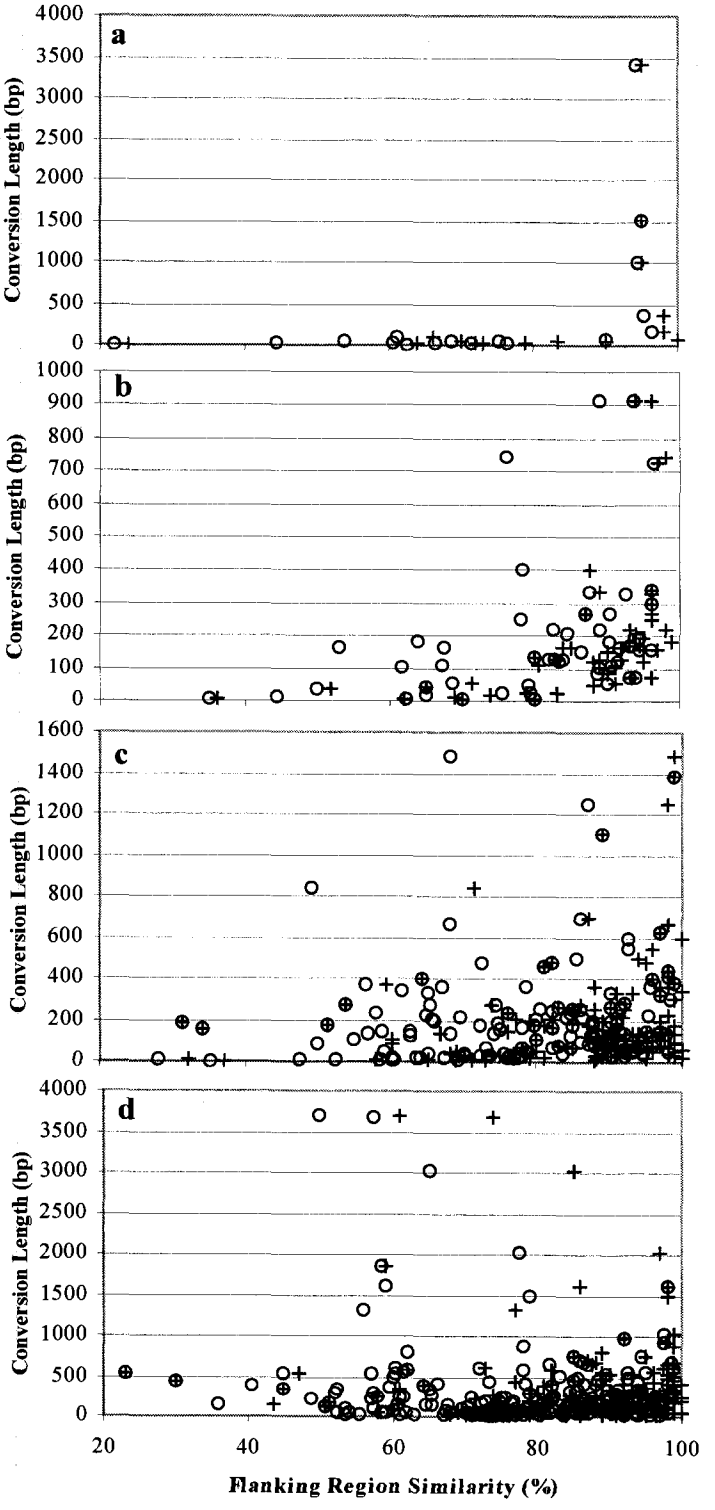

a
b
c
d
Conversion Length (bp)
Flanking Region Similarity (%)
20
40
60
80
100

Table 1. Amino acid substitutions in the recombination genes of the three pathogenic strains.

Protein	Length	Substitutions	Strains
recA	353	N_{140} - D	EDL933
recB	1180	V_{640} - A, E_{764} - D, G_{829} - S, S_{884} - L, A_{932} - V, R_{974} - Q, A_{1014} - D, N_{1016} - D	CFT073, EDL933, Sakai
		A_{792} - V	EDL933, CFT073
		L_{15} - F, Q_{112} - K, T_{227} - R, C_{522} - S, L_{661} - M, V_{883} - A, V_{1058} - A, A_{1166} - D	EDL933, Sakai
		E_{98} - A, S_{1010} - N	CFT073
recC	1122	Q_{178} - E	CFT073, EDL933, Sakai
		H_{177} - D, Q_{643} - L, A_{887} - V	CFT073
recD	608	H_{471} - N	CFT073, EDL933, Sakai
		A_{63} - E, V_{147} - T	EDL933, Sakai
recG	693	T_{256} - A	CFT073, EDL933, Sakai
		K_2 - T	EDL933, Sakai
		S_{100} - N	CFT073
recJ	577	N_{216} - H	CFT073, EDL933, Sakai
recN	553	I_{360} - T, Q_{370} - H, A_{426} - P	CFT073, EDL933, Sakai
		A_{278} - V, A_{321} - T, A_{525} - T	EDL933, Sakai
		L_{155}- Q	CFT073
recQ	610	Insertion of R_{504}, A_{505} - G	CFT073, EDL933, Sakai
recT	269	V_{24} - I, S_{27} - N, S_{149} - N, L_{200} - Q	EDL933, Sakai
		Deleted from genome	CFT073
rusA	120	V_{27} - I, N_{44} - G, A_{45} - S, R_{109} - K, M_{117} - L	EDL933, Sakai
ruvB	336	T_{322} - V	CFT073
sbcB	475	I_{115} - V, V_{451} - A	CFT073, EDL933, Sakai
		S_{38} - D	EDL933, Sakai
sbcC	1048	T_{211} - A, Q_{252} - L, I_{309} - T, P_{365} - L, L_{595} - Q, L_{645} - Q, T_{649} - A, C_{710} - S, no Asp_{817}	CFT073, EDL933, Sakai

		S_{322} - N, T_{405} - M, V_{715} - A, M_{867} - L	EDL933, Sakai
		F_{1044} - L	Sakai
		E_{84} - G, T_{350} - A, T_{413} - S, Q_{679} - H, A_{819} - S, T_{1043} - A, V_{1046} - M	CFT073
sbcD	400	T_{37} - A, A_{111} - T	CFT073
gyrA	875	A_{827} - S	CFT073
helD	684	V_{127} - M, Q_{128} - R, T_{182} - I, Q_{544} - H	CFT073
lig	671	N_{103} - S	CFT073, EDL933, Sakai
		A_{294} - E	EDL933, Sakai
		A_{584} - E, A_{662} - T	CFT073
mutH	229	Q_{134} - K	CFT073
mutL	615	N_{130} - D, A_{367} - V, L_{417} - P, A_{493} - G	CFT073, EDL933, Sakai
		A_{469} - V, H_{568} - N	EDL933 Sakai
		T_{376} - S, P_{388} - S, A_{476} - V	CFT073
MutS	853	G_{337} - E	CFT073, EDL933, Sakai
		A_{3} - T, T_{384} - N	CFT073
polA	928	frame shift at codon 226, stop codon at codon 260	CFT073
priA	732	A_{52} - V, H_{459} - Q	CFT073, EDL933, Sakai
		Q_{125} - H	EDL933, Sakai
		C_{30} - S, S_{69} - V, T_{157} - E	CFT073
topA	865	A_{43} - T	CFT073
rdgC	304	E_{169} - A	EDL933, Sakai
		A_{185} - T	CFT073
xth	268	L_{99} - P, D_{143} - E, T_{157} - G	CFT073, EDL933, Sakai
		D_{84} - E	EDl933, Sakai
		P_{123} - S	CFT073
xseA	456	T_{227} - A, T_{337} - A, I_{438} - V	CFT073, EDL933, Sakai
		R_{195} - C, N_{415} - K	EDL933, Sakai
		K_{194} - Q, Q_{346} - R, R_{347} - Q, N_{349} - V, P_{355} - S, K_{356} - R, N_{415} - A,	CFT073
		K_{419} - Q, A_{425} - V, M_{428} - T, E_{434} - G, W_{437} - V, E_{439} - I, K_{443} - S,	CFT073
		N_{444} - A, I_{445} - V, Q_{446} - T, P_{447} - K,	CFT073
		V_{448} - T, K_{449} - R, V_{455} - T, H_{456} - S, new stop codon at codon	

458	

Notes. Lengths are in number of amino acids. Amino acid substitutions are shown using the one letter code, and amino acid positions are indicated with subscript numbers. The following protein sequences were identical in all four *E. coli* strains: recE, recF, recO, recR, ruvA, ruvC, gyrB, lexA, ssb, uvrD, dam, and dut.

Table 2. Distribution of conversions, genes and conversion frequencies (in 200 kb bins).

	Distance between gene pairs (kb)													
	200	400	600	800	1000	1200	1400	1600	1800	2000	2200	2400	2600	2800
Conversions														
K-12	1	1	1	3	0	3	2	3	1	0	2	0	0	0
CFT073	6	0	4	1	6	3	0	9	7	6	7	3	0	0
EDL933	17	20	23	14	21	21	9	10	8	2	2	1	1	1
Sakai	27	38	37	37	13	24	17	10	14	3	5	3	2	0
Genes														
K-12	21	14	11	12	9	10	5	2	7	6	7	0	0	0
CFT073	46	13	32	4	34	24	4	17	31	7	10	8	1	0
EDL933	94	110	41	58	21	49	34	9	18	8	7	6	6	4
Sakai	116	133	80	31	10	15	15	15	24	9	9	3	7	4

Conversion frequencies														
K-12	0.05	0.07	0.09	0.25	0.00	0.30	0.40	1.50	0.14	0.00	0.29	0.00	0.00	0.00
CFT073	0.13	0.00	0.13	0.25	0.17	0.13	0.00	0.53	0.23	0.86	0.70	0.38	0.00	0.00
EDL933	0.18	0.18	0.56	0.24	1.00	0.43	0.26	1.11	0.44	0.25	0.29	0.17	0.17	0.25
Sakai	0.23	0.29	0.46	1.19	1.30	1.60	1.13	0.67	0.58	0.33	0.56	1.00	0.29	0.00

Table 3. Distribution of converted genes, gene family members, and conversion frequencies relative to *oriC* (in 200 kb bins).

		Distance from *oriC* (kb)													
		200	400	600	800	1000	1200	1400	1600	1800	2000	2200	2400	2600	2800
Strain	Gene type														
K-12	converted	1	1	1	1	4	2	2	1	4	2	2	1	0	0
	family	2	8	8	8	13	12	10	7	8	11	9	8	0	0
	frequency	0.50	0.13	0.13	0.13	0.31	0.17	0.20	0.14	0.50	0.18	0.22	0.13	0.00	0.00
CFT073	converted	4	0	4	1	5	10	7	1	0	2	18	10	0	0
	family	15	3	26	7	35	30	19	8	1	10	45	27	2	0
	frequency	0.27	0.00	0.15	0.14	0.14	0.33	0.37	0.13	0.00	0.20	0.40	0.37	0.00	0.00
EDL933	converted	0	2	1	2	0	4	6	5	9	13	42	11	41	43
	family	5	5	5	7	8	12	10	12	34	52	88	52	86	89
	frequency	0.00	0.40	0.20	0.29	0.00	0.33	0.60	0.42	0.26	0.25	0.48	0.21	0.48	0.48
Sakai	converted	1	3	0	3	1	4	5	3	14	28	26	24	70	24
	family	4	6	3	8	8	11	13	10	38	62	68	43	156	44
	frequency	0.25	0.50	0.00	0.37	0.13	0.36	0.38	0.30	0.37	0.45	0.38	0.56	0.45	0.55

Chapter 3

Ectopic Gene Conversions in the Backbone Genome of *E. coli*

Robert T. Morris and Guy Drouin

Département de biologie et Centre de recherche avancée en génomique environnementale, Université d'Ottawa, Ottawa, Ontario, Canada, K1N 6N5

Running head: Ectopic conversions in *E. coli*

Keywords: ectopic, gene conversion, *Escherichia coli*, pathogenic, backbone genome

Corresponding author: Guy Drouin, Département de biologie, Université d'Ottawa, 30 Marie Curie, Ottawa, Ontario, Canada, K1N 6N5. Tel.: (613) 562-5800 ext. 6052, FAX: (613) 562-5486, E-mail: gdrouin@science.uottawa.ca

ABSTRACT

Four sequenced *E. coli* genomes (the K-12 laboratory strain and the CFT073, EDL933 and Sakai pathogenic strains) share a common backbone of 3885 genes. I extracted the orthologous multigene family members found in the backbone genome of these four strains and identified the ectopic gene conversions that occurred between them. The length of the twelve conversions I detected in backbone genes range from 11 bp to 935 bp and the frequency of conversion is similar in all four genomes. This suggests that gene conversions in backbone genes are under similar purifying selection in all four strains. This suggestion is supported by four observations: the similar Ka/Ks ratio of backbone genes where I detected gene conversions in all four strains, the higher Ka/Ks ratio of K-12 specific genes where I detected gene conversions, the higher similarity of backbone genes having longer gene conversions and the fact that most nucleotides changes observed in converted regions are likely selectively neutral. Our results also show that duplicated bacterial genes can have a mosaic structure due to ectopic recombination between the duplicated genes.

INTRODUCTION

Gene conversions are non-reciprocal exchanges of DNA initiated by double stranded DNA breaks. As a result, part of the sequence of a gene is converted to that of a related gene. Several studies have documented the impact of gene conversions on the evolution of prokaryotic genes and genomes (e.g., SINGER 1988; LIAO 2000; JORDAN *et al.* 2001; PRIDE and BLASER 2002; HASHIMOTO *et al.* 2003; SANTOYO and ROMERO 2005).

Four *E. coli* genomes had been sequenced when I initiated my study of ectopic conversion in the backbone (i.e., the set of orthologs shared between each *E. coli* strain) sequences of *E. coli*: the CFT073, K-12 MG1655, O157:H7 EDL933 and O157:H7 Sakai genomes (BLATTNER *et al.* 1997; HAYASHI *et al.* 2001; PERNA *et al.* 2001; WELCH *et al.* 2002). Comparative genomics has shown that these genomes share a common backbone of genes maintained during evolution (HAYASHI *et al.* 2001; PERNA *et al.* 2001; WELCH *et al.* 2002; DARLING *et al.* 2004). Whereas the genome of the non-pathogenic K-12 strain is mainly composed of backbone genes, that of the pathogenic strains CFT073, EDL933 and Sakai also contain genes responsible for the infection of host tissues and toxin production (OCHMAN *et al.* 2000; HAYASHI *et al.* 2001; BRÜSSOW *et al.* 2004). Most of these genes are derived from horizontally acquired prophage sequences, but also from pathogenic islands, transposons and plasmids (BRÜSSOW *et al.* 2004).

This study focuses on the characterization of the ectopic gene conversions in the common backbone of these four *E. coli* genomes. Our results show that ectopic gene conversions occur between duplicated backbone genes, that most conversions are between 11 bp to 43 bp long, that the nucleotide changes they generate are likely

selectively neutral and that they are equally frequent in pathogenic and non-pathogenic strains.

MATERIALS AND METHODS

Gene Family Identification: I used the BLASTCLUST program (ftp://ftp.ncbi.nih.gov/blast/) to extract all the protein coding multigene families found in the K-12 MG1655 genome; protein coding gene sequences for each *E. coli* strain were retrieved from the NCBI ftp site (ftp://ftp.ncbi.nih.gov/genomes/Bacteria). The BLASTCLUST program performs pairwise comparisons of each protein coding gene using the BLAST algorithm. All genes which meet the user defined thresholds of sequence similarity over a percentage of the genes' length are grouped together as a multigene family. In this case a set of sequences was defined as a multigene family when at least two sequences sharing more than 60% amino acid identity over at least half their length were present in this genome. I then used these K-12 sequences to extract the orthologous sequences from the CFT073, EDL933 and Sakai genomes. Genes in a queried genome were considered to be orthologous when they were both the best FASTA match and had the same flanking neighbor genes within 10Kb (PEARSON and LIPMAN 1988). The protein sequences of each family (i.e., paralogues and orthologues) were aligned using ClustalW (THOMPSON *et al.* 1994) and the DNA sequences were fitted to the protein alignments using align2aa (http://sunflower.bio.indiana.edu/~wfischer/Perl_Scripts/).

Gene conversion detection: Gene conversions were identified using the GENECONV program developed by Sawyer (1999). This program detects gene conversions by identifying unusually long consecutive stretches of identical polymorphic sites between sequences; each conversion is flanked by a set of consecutive discordant (different)

polymorphic sites. Previous studies have shown that the method implemented in this program performs well, i.e., it is relatively free of false positives (type I error rate < 5%) and powerful (i.e., relatively free of type II errors) when the analyses are performed on at least three paralogous genes (DROUIN *et al.* 1999; SAWYER 1999; POSADA and CRANDALL 2001; DROUIN 2002; POSADA 2002). To maximize the number of informative sites (i.e., the number of sites identified as polymorphic by the GENECONV algorithm), and to insure that at least three genes were analyzed, the GENECONV input sequences contained both orthologous and paralogous sequences. However, the output of this program were filtered to consider only the conversions between paralogues (i.e., the within species conversions).

Genomes alignment, flanking identity and chromosomal location: A multiple alignment of all four genomes was computed using the Mauve application (DARLING et al 2004). This program uses the anchored alignment technique to align whole genomes (DARLING et al 2004). These anchors are identified as inexact ungapped sequence matches in each genome. These sets of matches are are defined as local collinear sequence blocks (LCB). The sets of LCBs are aligned between genomes using the CLUSTAL algorithm. To ensure that each LCB is a significant sequence match and not due to random effects each one must meet a minimum score set by the user (for this study I used 69 as the LCB weight; DARLING et al 2004). The higher the weighting the more selective (but less sensitive) the alignment algorithm will become (DARLING *et al.* 2004). The flanking identity of the converted genes was calculated using an in-house

PERL script that calculates the average nucleotide sequence identity found 100 bp upstream and downstream of the converted regions.

The position of the converted genes along the length of the genomes was analyzed to determine if converted genes were clustered along the length of the conserved backbone. The location of the converted genes was identified from the gene position file (*.ptt) obtained from the NCBI ftp site. Statistical tests were performed using S-plus v7.0 (Insightful Corporation, Seattle, WA, USA) and Excel (Microsoft Corporation, Redmond, WA, USA).

Ka/Ks ratios: The numbers of nonsynonymous substitutions per nonsynonymous site (Ka) and of synonymous substitutions per synonymous site (Ks) were estimated for each pair of converted genes using YN00 from the PAML package (version 3.13; YANG 1997; YANG and NIELSEN 2000). The evolutionary rates were calculuated on the unconverted regions of the multiple alignment to remove the bias introduced by the detected conversion events. Average Ka and Ks values were calculated between the multigene families members found in each strain.

RESULTS

Genome backbone and gene families: A multiple alignment for the conserved regions within all four genomes was created using the Mauve application (DARLING *et al.* 2004). Visual inspection of the genome alignment shows that large chromosomal segments share the same synteny in each *E. coli* strain (Figure 1). Furthermore, most of the 4.6 Mb long K-12 genome is part of this common backbone, while pathogenic strains contain numerous strain-specific regions. Statistical analysis of the location of the backbone genes shows that these genes are not clustered along the length of the genome (χ^2-tests, $p > 0.1$).

The BLASTCLUST program identified 100 multigene families with two or more members within the K-12 genome. Forty-four of those were found in the conserved backbone of the four *E. coli* genomes studied (Supplemental Table 1, found on CD). These 44 families contain a total of 95 genes for each genome and ranged in size from two to four members (Supplemental Table 1).

The 100 multigene families with two or more members I identified in the K-12 genome represent gene families where the members are at least 60% identical at the protein level over at least 50% of their length. Decreasing the percent identity and/or the length can lead to the identification of many more gene families (results not shown). Furthermore, the number of families detected is most sensitive to the percent protein identity. For example, the K-12 genome contains 166 gene families with 50% percent identity over 50% of the proteins length and 175 gene families with 50% percent identity over 25% of the proteins length. However, since I previously established that using gene

families with less than 60% protein identity did not lead to the identification of more gene conversions larger than 41 bp, the gene families I analyzed likely contains most, if not all, conversions larger than 41bp (MORRIS and DROUIN 2004).

Gene conversions in backbone genes: The frequency of conversion is similar in all four genomes (with 7, 5, 8 and 9 conversions found in the K12, CFT073, EDL933 and Sakai genomes, respectively; Table 1). To our knowledge, of these conversions, only that between the *tufA* and *tufB* elongation factor genes has previously been reported (e.g., SHARP 1991; ABDULKARIM and HUGHES 1996; LATHE and BORK 2001). Two of these twelve conversions are common to all four *E. coli* genomes, five are found only in the EDL933, Sakai and K-12 genomes, three are found only in the EDL933 and Sakai genomes, two are found only in the CFT073 genome and the final two conversions are either common to the K-12 and Sakai genomes or the K-12 and CFT073 genomes. None of the backbone conversions are specific to the K-12 strain. Apart from the large conversions present in families 48 and 59, all conversions found in more than one strain share the same position and length within their respective gene family multiple alignments (Table 1, Figure 2 and Supplemental Figure 1, found on CD). The interpretation of the data suggests that each conversion likely occurred only once, i.e., in the common ancestor of the strains in which the conversions are found. The variable lengths of the large conversions found in families 48 and 59 are likely due to post conversion nucleotide substitutions masking the original boundaries of the conversion event identified by the GENECONV detection method. For example, in the case of family 59 (*tuf* genes), the position of the converted region relative to the multiple

alignment (1230 aligned bases) varied for each genome. The converted regions for the K-12, CFT073, EDL933 and Sakai genomes were located at nucleotide positions 337-1211, 280-1085, 410-1214 and 256-630, respectively. Finally, the converted genes are uniformly distributed within the chromosome of all four *E. coli* strains ($p > 0.13$) and conversions larger than 43 bp long only occur between sequences having at least 92% sequence identity.

Ka/Ks ratios: The average Ka, Ks and Ka/Ks values of the converted backbone are similar in all four genomes (Table 2). The average (± standard deviation) Ka, Ks and Ka/Ks values for the K-12-specific genes listed in Table 3 are 0.194 ± 0.140, 1.367 ± 0.919 and 0.190 ± 0.260, respectively. The average Ka/Ks ratio of the converted K-12-specific genes is approximately twice as large as the Ka/Ks ratio of converted backbone genes.

Gene conversions in K-12 specific genes: Of the seventeen ectopic gene conversions I previously detected in the K-12 genome (MORRIS and DROUIN 2004), thirteen are specific to the K-12 genome (Table 3). In contrast with the conversions found between backbone genes (Table 1), most of these conversions are between genes known to be under relaxed selective pressure (such as insertion sequences and recombination hot spot proteins). Furthermore, conversions covering more than 43 bp occur between sequences having as little as 61% flanking sequence identity.

DISCUSSION

Gene organization: The backbones of the four *E. coli* genomes analyzed here are co-linear except for the inversion present in the EDL933 genome and some rearrangements of small regions between K-12 and the pathogenic strains (CFT073, EDL933 and Sakai; Figure 1). The EDL933 inversion is likely very recent because it must have occurred after the event that gave rise to the closely related EDL933 and Sakai strains (DARLING *et al.* 2004). Our results also show that orthologous genes are uniformly distributed along the length of the chromosome. The main difference in gene organization between the K-12 strain and the three pathogenic strains is therefore that interspersed genes have been added to the backbone of the pathogenic strains. These interspersed genes are mainly derived from prophage sequences (BRÜSSOW *et al.* 2004; results not shown). In contrast, the 4.3 Mb of conserved *E. coli* backbone sequences likely contain functionally important genes because they have been maintained in all strains.

Gene conversions in the backbone and K-12 specific genes: Most of the gene conversions I detected are small. Ten of them range from 11 bp to 43 bp long and only two of them are 375 and 935 bp long (Table 1). The fact that even the short conversions receive strong statistical support, that they are all found in more than one strain, that they are all flanked by divergent unconverted regions and that the corresponding regions of unconverted genes are clearly different from the converted regions all support the conclusion that these regions of high similarity between paralogous genes are due to conversion events (Table 1, Figure 2 and Supplemental Figure 1).

The fact that, in backbone genes, the frequency of conversion is similar in all four genomes suggests that gene conversions in these genes are under similar purifying selection in all four genomes (Table 1). This suggestion is supported by four observations. First, the Ka/Ks ratios of backbone genes where I detected gene conversions are similar in all four strains (Table 2). Second, the fact that the Ka/Ks ratio of K-12 specific genes where I detected gene conversions is twice as large as that of backbone genes where I detected gene conversions (i.e., 0.19 versus 0.09, respectively) indicates that backbone specific genes are under more selective constrains than K-12 specific genes (which are mostly composed of sequences known to be under little purifying selection such as insertions sequences and recombination hot spot elements; Table 3). Third, whereas gene conversions longer than 43 bp are limited to backbone genes that share more than 92% flanking sequence identity, such conversions occur between K-12 specific genes having as little as 61% flanking sequence identity. This is also consistent with the suggestion that backbone specific genes are under more selective constraints than K-12 specific genes. Fourth, of the 50 nucleotide changes that occurred due to conversions between the backbone genes, 43 are synonymous substitutions and the 7 nonsynonymous substitutions result in conservative amino acid changes (such as serine to threonine and isoleucine to valine changes; BETTS AND RUSSELL 2003; Table 1 and Supplementary Table 1). The nucleotide changes generated by these conversions are therefore unlikely to be under strong negative selection. This conclusion was also previously reached in the case of the conversion between the duplicated *tuf* genes of *E. coli* and *Salmonella typhimurium* (SHARP 1991; ABDULKARIM and HUGHES 1996).

The results presented here complement those of our previous study of gene conversions in *E. coli* genomes (MORRIS and DROUIN 2004). Our previous study, which analyzed gene conversions between the multigene family members found in each *E. coli* genome, showed that gene conversions were more frequent, and required less flanking sequence similarity, to occur between the multigene family members of pathogenic strains than those of K-12. However, that study suffered from the fact that most of the genes being compared between the K-12 strain and the three pathogenic strains were different. As discussed above, the K-12 genome is mainly composed of backbone genes with relatively few multigene families whereas the three pathogenic genomes contain numerous large multigene families derived from, among others, prophage and insertion sequences. Here, comparison of the same genes from all four genomes shows that backbone genes are under similar selective pressure in all four genomes irrespective of whether a strain is pathogenic or not.

Analysis of the conversions identified here also allowed us to conclusively show that conversions as short 11 bp long occur between paralogous genes (Table 1, Supplemental Figure 1). Although our previous studies did identify such short conversions, I believe that the results presented here represent the first clear demonstration of such short conversions in bacteria (DROUIN 2002; MORRIS and DROUIN 2004). As discussed above, the nucleotide changes generated by these ectopic conversions are likely selectively neutral. This is consistent with several other studies where it was observed that conversions are often limited to gene regions under little selective constraints (e.g., ZHOU and LI 1996; NOONAN *et al.* 2004).

Our results also extend previous studies on recombination in bacteria (reviewed in MAYNARD SMITH 1990). Although early molecular studies argued that the propagation of bacteria was clonal (e.g., OCHMAN and SELANDER 1984), later studies showed that the genes of these clonal organisms had a mosaic structure due to allelic recombination between alleles from the same or different species (e.g., DUBOSE *et al.* 1988; MAYNARD SMITH 1990; MAYNARD SMITH *et al.* 1991). Our results show that the mosaic structure of bacterial genes can also originate from recombination between the duplicated genes found in the genome of most bacterial species.

This research was supported by a Discovery grant from the National Science and Engineering Research Council of Canada to G. D.

LITERATURE CITED

ABDULKARIM, F., and D. HUGHES, 1996 Homologous recombination between the *tuf* genes of *Salmonella typhimurium*. J. Mol. Biol. **260**: 506-522.

BETTS, M. J., and R. B. RUSSELL, 2003 Amino acid properties and consequences of substitutions, pp.289-316 in *Bioinformatics for Geneticists*, edited by M. R. Barnes and I. C. Gray. John Wiley & Sons, New York.

BLATTNER, F. R., G. PLUNKETT 3RD, C. A. BLOCH, N. T. PERNA, V. BURLAND *et al.*, 1997 The complete genome sequence of *Escherichia coli* K-12. Science **277**: 1453-1474.

BRÜSSOW, H., C. CANCHAYA and W. D. HARDT, 2004 Phages and the evolution of bacterial pathogens: from genomic rearrangements to lysogenic conversion. Microbiol. Mol. Biol. Rev. **68**: 560-602.

DARLING, A. C., B. MAU, F. R. BLATTNER and N. T. PERNA, 2004 Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome Res. **14**: 1394-1403.

DROUIN, G., F. PRAT, M. ELL and G. D. P. CLARKE, 1999 Detecting and characterizing gene conversions between multigene family members. Mol. Biol. Evol. **16**: 1369-1390.

DROUIN, G., 2002 Characterization of the gene conversions between the multigene family members of the yeast genome. J. Mol. Evol. **55**: 14-23.

DUBOSE, R. F., D. E. DYKHUIZEN and D. L. HARTL, 1988 Genetic exchange among natural isolates of bacteria: recombination within the *phoA* gene of *Escherichia coli*. Proc. Natl. Acad. Sci. USA **85**: 7036-7040.

HASHIMOTO, J. G., B. S. STEVENSON and T. M. SCHMIDT, 2003 Rates and consequences of recombination between rRNA operons. J. Bacteriol. **185**: 966 – 972.

HAYASHI, T., K. MAKINO, M. OHNISHI, K. KUROKAWA, K. ISHII *et al.*, 2001 Complete genome sequence of enterohemorrhagic *Escherichia coli* O157:H7 and genomic comparison with a laboratory strain K-12. DNA Res. **8**: 11-22.

JORDAN, I. K., K. S. MAKAROVA, Y. I. WOLF and E. V. KOONIN, 2001 Gene conversions in genes encoding outer-membrane proteins in *H. pylori* and *C. pneumoniae*. Trends Genet. **17**: 7-10.

LATHE, W. C., and P. BORK, 2001 Evolution of *tuf* genes: ancient duplication, differential loss and gene conversion. FEBS Lett. **502**: 113-116.

LIAO, D., 2000 Gene conversion drives within genic sequences: concerted evolution of ribosomal RNA genes in bacteria and archaea. J. Mol. Evol. **51:** 305-317.

MAYNARD SMITH, J., 1990 The evolution of prokaryotes: does sex matter? Annu. Rev. Ecol. Sys. **21:** 1-12.

MAYNARD SMITH, J., C. G. DOWSON and B. G. SPRATT, 1991 Localized sex in bacteria. Nature **349:** 29-31.

MORRIS, R. T., and G. DROUIN, 2004 Ectopic gene conversions in four *Escherichia coli* genomes: increased recombination in pathogenic strains. J. Mol. Evol. **58:** 596-605.

NOONAN, J. P., J. GRIMWOOD, J. SCHMUTZ, M. DICKSON, and R. M. MYERS, 2004 Gene conversion and the evolution of protocadherin gene cluster diversity. Genome Res. **14:** 354-366.

OCHMAN, H., and R. K. SELANDER, 1984 Evidence for clonal population structure in *Escherichia coli*. Proc. Natl. Acad. Sci. USA **81:** 198-201.

OCHMAN, H., J. G. LAWRENCE and E. A. GROISMAN, 2000 Lateral gene transfer and the nature of bacterial innovation. Nature **405:** 299-304.

PEARSON, W. R., and D. J. LIPMAN, 1988 Improved tools for biological sequence comparison. Proc. Natl. Acad. Sci. USA 85: 2444-2448.

PERNA, N. T., G. PLUNKETT 3RD, V. BURLAND, B. MAU, J. D. GLASNER *et al.*, 2001 Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. Nature **409:** 529-533.

POSADA, D., 2002 Evaluation of methods for detecting recombination from DNA sequences: Empirical data. Mol. Biol. Evol. **19:** 708-717.

POSADA, D., and K. A. CRANDALL, 2001 Evaluation of methods for detecting recombination from DNA sequences: Computer simulations. Proc. Natl. Acad. Sci. USA **98:** 13757-13762.

PRIDE, D. T., and M. J. BLASER, 2002 Concerted evolution between duplicated genetic elements in *Helicobacter pylori*. J. Mol. Biol. **316:** 629-642.

SANTOYO, G., and D. ROMERO, 2005 Gene conversion and concerted evolution in bacterial genomes. FEMS Microbiol Rev. **29:** 169-183.

SAWYER, S. A., 1999 GENECONV: A computer package for the statistical detection of gene conversion. Distributed by the author, Department of Mathematics, Washington University, available at http://www.math.wustl.edu/~sawyer.

SHARP, P. M., 1991 Determinants of DNA sequence divergence between *Escherichia coli* and *Salmonella typhimurium*: codon usage, map position, and concerted evolution. J. Mol. Evol. **33**: 23-33.

SINGER, B. S., 1988 On the role of homologous sequences in chromosomal rearrangements. Genes Dev. **2:** 1800-1811.

THOMPSON, J. D., D. G. HIGGINS and T. J. GIBSON, 1994 CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. **22:** 4673-4680.

WELCH, R. A., V. BURLAND, G. PLUNKETT 3RD, P. REDFORD, P. ROESCH *et al.*, 2002 Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli.* Proc. Natl. Acad. Sci. USA **99:** 17020-17024.

YANG, Z., 1997 PAML: a program package for phylogenetic analysis by maximum likelihood. CABIOS 13: 555-556.

YANG, Z., and R. NIELSEN, 2000 Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol. Biol. Evol. **17:** 32-43.

ZHOU, Y. H., and W.-H. LI, 1996 Gene conversion and natural selection in the evolution of X-linked color vision genes in higher primates. Mol. Biol. Evol. **13:** 780-783.

Table 1. List of ectopic gene conversions detected within the common backbone of the four *E. coli* genomes. The gene names of the converted genes are provided as well as the length of the gene conversion and the maximum flanking sequence identity. The Number and type of nucleotide substitutions are also indicated (syn = synonymous; nonsyn = non-synonymous). A basic functional description is given for each pair of genes; this information was taken from the genomes' genbank file.

Multigene	Converted	Length (bp)/Flanking nucleotide identity (%)[1]				Substitutions[2]	Gene function
family	genes	K-12	Sakai	EDL933	CFT073		
8	*phoE;ompF*	16/62.9**	16/62.8**	16/62.8**	-	2 syn.	outermembrane proteins
	ompF;ompC	-	43/67.1*	43/67.1*	-	1 syn.	outermembrane proteins
12	*acrB;acrD*	27/66.4*	27/66.3*	-	-	1 syn.	acridine efflux pumps
16	*sdaA;tdcG*	23/68.7***	23/68.4***	23/68.4**	23/68.1**	6 syn., 4 nonsyn.	L-serine deaminase/dehydratase
38	*fumA; fumB*	-	21/79*	21/78.9*	-	2 syn.	fumarases
40	*narH;narY*	23/74.8**	23/75.1**	23/75.1**	-	5 syn., 1 nonsyn.	nitrate reductases
45	*araE;galP*	-	-	-	25/63**	4 syn.	arabinose/galactose symporters
48	*gadB;gadA*	935/93.9***	-	-	554/92.2***	13 syn.	glutamate decarboxylases
51	*hyp;hyp*	-	18/72.1*	18/72.1*	-	2 syn.	hypothetical proteins
59	*tufA;tufB*	875/98.5**	375/98.3*	805/98.3**	806/98.3**	n. a.	elongation factors
78	*artI;artJ*	-	-	-	27/67*	3 syn.	arginine binding protein precursors
79	*narI;narV*	11/67.8*	11/68*	11/68*	-	4 syn., 2 nonsyn.	nitrate reductases

Notes. [1] Percent identity of the regions flanking the conversions. An hyphen (-) indicates the absence of a particular conversion. The GENECONV simulated significance level is indicated for each gene conversion (*, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.001$). [2] Nucleotide substitutions are relative to orthologous genes in other strains except for the conversions found in gene family 16 where they are relative to the third (unconverted) gene and the conversions in family 59 where unconverted sequences were not available (n. a.); syn., synonymous; nonsyn., nonsynonymous. A detailed list of nucleotide substitutions is provided in Supplemental Table 2; found on CD.

Table 2. The average non-synonymous (Ka) and synonymous (Ks) nucleotide substitution rates (± standard deviation) as well as the ratio of these two rates (Ka/Ks; a.k.a. ω) for converted genes in each *E. coli* genome.

Genome	Ka	Ks	Ka/Ks
K12	0.17 ± 0.11	1.90 ± 1.40	0.09 ± 0.02
Sakai	0.17 ± 0.10	2.01 ± 0.83	0.09 ± 0.05
EDL933	0.16 ± 0.10	2.08 ± 0.83	0.08 ± 0.03
CFT073	0.17 ± 0.12	1.83 ± 1.51	0.10 ± 0.05

Table 3. List of ectopic gene conversions detected between duplicated genes specific to the *E. coli* K-12 genome. The gene names, length and maximum flanking sequence identity is provided for each conversion. In addition a basic classification of the gene function is provided as described in the genomes' genbank file.

Multigene family	Converted genes		Flanking identity	Conversion length (bp)	Gene function
1	*trs5_1*	*trs5_5*	93.8	173	IS5 transposases
3	*insB_2*	*insB_4*	98.5	368	IS1 proteins
5	*insA_5*	*insA_7*	90.4	38	IS1 proteins
6	*cspB*	*cspG*	63	47	cold shock proteins
8	*ompC*	*nmpC*	66	43	outermembrane proteins
11	*rhsA*	*rhsB*	61.2	3422	rhs proteins
	rhsB	*rhsD*	76.1	32	rhs proteins
	rhsB	*rhsD*	76.1	12	rhs proteins
	rhsB	*rhsD*	76.1	87	rhs proteins
	rhsC	*rhsD*	72	17	rhs proteins
	rhsA	*rhsD*	73.1	7	rhs proteins
13	*yfjz*	*yeeU*	62.9	21	CP4-57 and CP4-44 prophages
22	*hyp*	*hyp*	70.4	75	putative DNA repair proteins

Figure legends

Figure 1. Multiple alignment of the Sakai, EDL933, K-12 and CFT073 genomes using the Mauve software package with a minimal weight of 69. Blocks indicate regions conserved in all four genomes and lines link these orthologous regions. Within each block, white regions represent lineage-specific sequences whereas gray regions represent backbone sequences. Blocks below the line represent inverted regions relative to the reference sequence at the top of the figure.

Figure 2. Converted regions from gene families 8 (A) and 16 (B). The multiple alignments show the DNA and protein sequences of the three genes found in the four *E. coli* genomes. Converted regions are shown with black letters surrounded by a black border in both the DNA and the protein sequences whereas unconverted flanking regions are shown with grey letters. The scale shows the position of the converted regions relative to the multiple sequence alignment. The converted regions of the other gene families listed in Table 1 are shown in Supplemental Figure 1; found on CD. The labels DNA and PROT indicate the DNA and amino acid sequences of the multigene family members. For panel A: DNA/PROT 1 = *phoE*, DNA/PROT 2 = *ompF* and DNA/PROT 3 = *ompC*. For panel B: DNA/PROT 1 = *sdaA*, DNA/PROT 2 = *tdcG* and DNA/PROT 3 = *sdaB*.

Figure 1

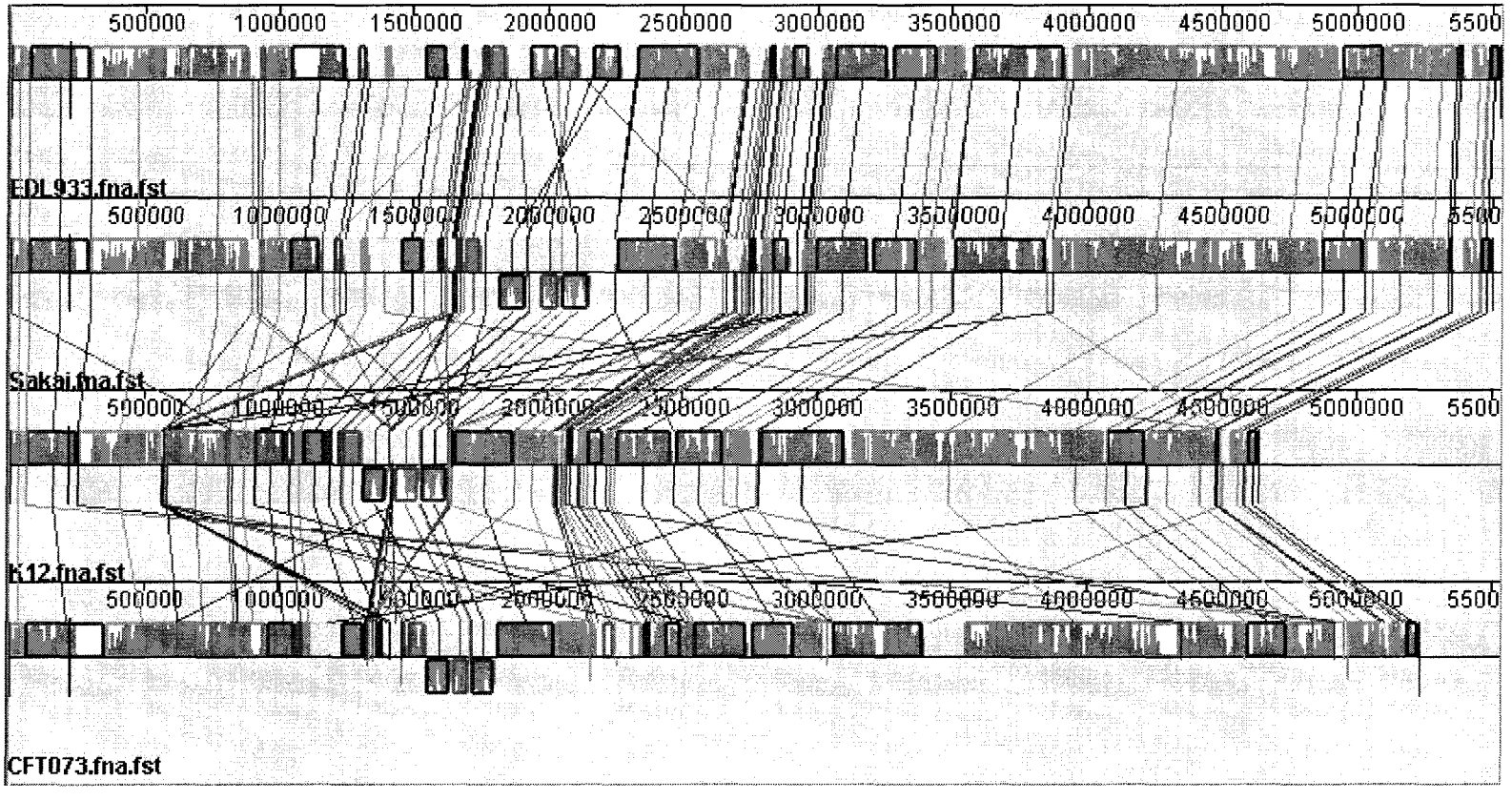
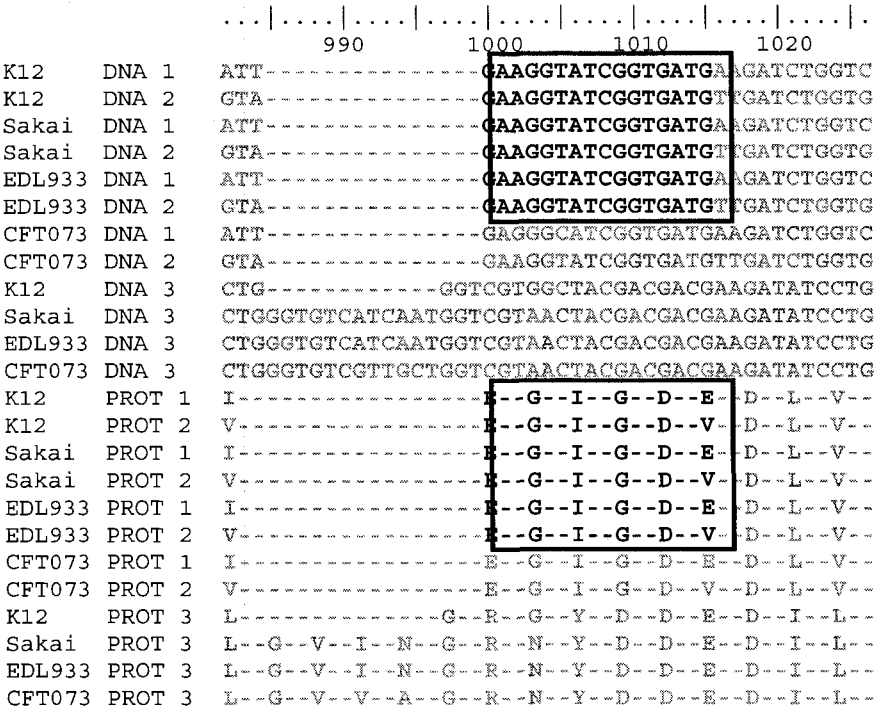
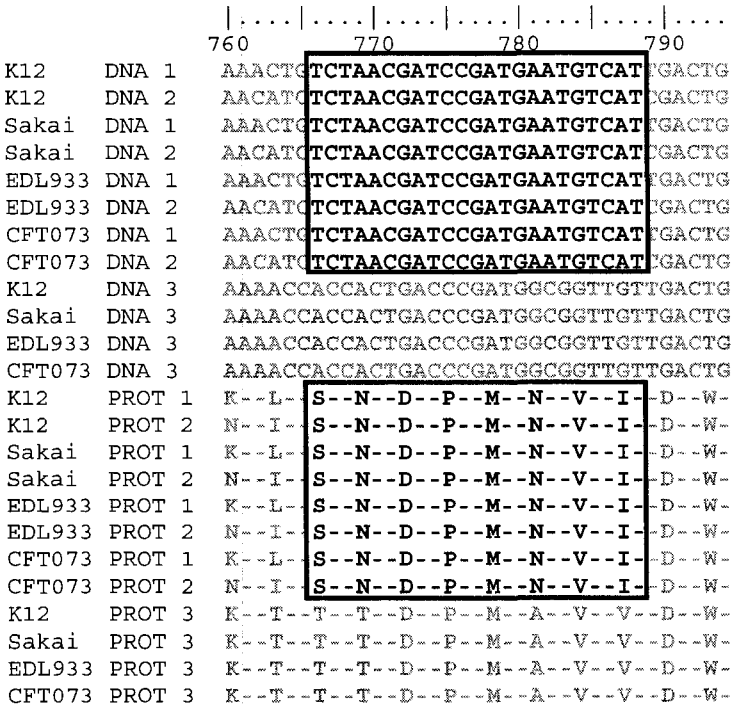

500000
1000000
1500000
2000000
2500000
3000000
3500000
4000000
4500000
5000000
5500
EDL933.fna.fst
Sakai.fna.fst
K12.fna.fst
CFT073.fna.fst

Figure 2.

A


990
1000
1010
1020
K12 DNA 1
K12 DNA 2
Sakai DNA 1
Sakai DNA 2
EDL933 DNA 1
EDL933 DNA 2
CFT073 DNA 1
CFT073 DNA 2
K12 DNA 3
Sakai DNA 3
EDL933 DNA 3
CFT073 DNA 3
K12 PROT 1
K12 PROT 2
Sakai PROT 1
Sakai PROT 2
EDL933 PROT 1
EDL933 PROT 2
CFT073 PROT 1
CFT073 PROT 2
K12 PROT 3
Sakai PROT 3
EDL933 PROT 3
CFT073 PROT 3

B


760
770
780
790
K12 DNA 1
K12 DNA 2
Sakai DNA 1
Sakai DNA 2
EDL933 DNA 1
EDL933 DNA 2
CFT073 DNA 1
CFT073 DNA 2
K12 DNA 3
Sakai DNA 3
EDL933 DNA 3
CFT073 DNA 3
K12 PROT 1
K12 PROT 2
Sakai PROT 1
Sakai PROT 2
EDL933 PROT 1
EDL933 PROT 2
CFT073 PROT 1
CFT073 PROT 2
K12 PROT 3
Sakai PROT 3
EDL933 PROT 3
CFT073 PROT 3

Chapter 4

Ectopic Gene Conversions in Prokaryotic Genomes

Robert T. Morris and Guy Drouin
Département de biologie et Centre de recherche avancée en génomique environnementale,
Université d'Ottawa, Ottawa, Ontario, Canada, K1N 6N5

Keywords: ectopic, gene conversion, recombination, proteobacteria, firmicutes, archaebacteria

Running head: Ectopic Gene Conversions in Bacteria

Correspondance to: Guy Drouin, Département de biologie, Université d'Ottawa, 30 Marie Curie, Ottawa, Ontario, Canada, K1N 6N5. Tel.: (613) 562-5800 ext. 6052, FAX: (613) 562-5486, E-mail: gdrouin@science.uottawa.ca

Abstract

I characterized the gene conversions found between the duplicated genes of 75 prokaryotic genomes from five species groups (archaea, nonpathogenic and pathogenic firmicutes, and nonpathogenic and pathogenic proteobacteria). The number of gene conversions is positively correlated with the size of multigene families but the size of multigene families is not significantly different between pathogenic and nonpathogenic taxa. Gene conversions occur twice as frequently in pathogenic species as in nonpathogenic species. Comparisons between closely related species also indicate a trend towards increased gene conversion in pathogenic species. Whereas the length of the conversions is positively correlated with flanking sequence similarity in all five groups, these correlations are weaker for pathogenic firmicutes and proteobacteria than for nonpathogenic firmicutes and proteobacteria. These results are consistent with our previous work on *E. coli* genomes and suggest that pathogenic bacteria tolerate recombination between more divergent gene sequences. The wide variation in the presence/absence of recombination genes in different taxa prohibits the detection of an associate between the frequency of gene conversion, or the size of multigene families, with the presence/absence of particular recombination genes.

Introduction

All organisms undergo recombination during their life cycle; it is an important process whereby genetic material is reordered within a genome. The reorganization can occur between homologous or non-homologous sequences, and the transfer can be reciprocal (cross-over), or non-reciprocal (gene conversion). Two main types of conversions are known, allelic and ectopic (31). Allelic conversions occur between genes located at the same locus on sister chromosomes, and ectopic conversions occur between genes located at different loci within the genome.

Gene conversions are caused by the mismatch repair of heteroduplex DNA formed by strand invasion and migration of Holliday junctions during recombination repair (45). In bacteria, the repair mechanism of double strand DNA breaks (DSB) is catalyzed by the *recBCD* pathway (44). Homologues of prokaryotic *recBCD* genes have been found in eukaryotes. Furthermore, the yeast *rad51* performs a function similar to that of *recA* in *E. coli* (46). This function is so similar that the *E. coli recA* gene can rescue a recombination deficient yeast strain (11). The resolution of Holliday junction intermediates is a necessary part of recombination repair. Homologs of recA and Holliday junction resolvases have been identified within all kingdoms of life (17, 36, 40). These observations suggest that the basic mechanism of recombination is similar in all organisms (17).

Gene conversions have been documented in numerous bacterial species (38). For example, they are thought to be responsible for the concerted evolution of the multiple ribosomal RNA genes found in the genome of numerous bacteria and archaea (15, 20). In *E. coli,* studies have found that the frequency of allelic recombination is dependent on the

size of the homologous region and the similarity between the donor and recipient sequences. The rate of recombination increases exponentially as the length of the homologous region between the damaged and the template gene increases from 20 to 74 bp (52). With respect to the effect of sequence divergence on recombination frequency, studies have found that 2% heterology will reduce the recombination rate by 4-fold and 10% sequence divergence will decrease recombination rate by 40-fold (42, 52). This recombination barrier is caused by the activity of specific repair genes affecting the recombination mechanism. Deficiencies in the repair genes *mutS* and *mutL* will permit recombination between divergent genes (34). These genes inhibit recombination between gene sequences that are very divergent, despite the fact that RecA allows the pairing of homologous sequences up to 30% divergent (8). Previous studies have proposed that mutation in some mismatch repair genes produce a hyper-recombinative state in bacterial species (7, 18, 53). Indeed, it has been shown that pathogenic strains of *E. coli* are deficient in some repair genes, thus inducing a hyper-recombinative state (19, 35). This state is thought to be selectively advantageous to pathogenic strains because the introduction of divergent gene sequences into their genome creates more genetic variation, thus allowing them to evade the immune system of their host (35). In addition, RecC has been identified as a virulence factor in *Salmonella enterica*, and *recBCD* mutants attenuate pathogens (5, 49). Thus recombination activity affects the pathogenicity of bacteria.

I previously compared the characteristics of the ectopic gene conversions found in the *E. coli* K-12 genome with those found three pathogenic *E. coli* strains (the CFT073, EDL933 and Sakai strains; 28). I found that the frequency of ectopic gene conversion

was higher in pathogenic than in nonpathogenic strains, that the similarity of the regions flanking conversions was lower in pathogenic than in nonpathogenic strains, that the frequency of conversion was dependent upon nucleotide distance between the paralogous genes along the chromosome, that the frequency of conversion was not higher near the origin of replication and that converted regions in a gene were evenly distributed along the length of that gene. I began with 75 bacterial genomes (i.e., those for which a complete genome sequence were available by May 2003) which were divided into five groups: archaea, pathogenic and nonpathogenic proteobacteria (purple bacteria and relatives), and pathogenic and nonpathogenic firmicutes (Gram-positive bacteria). Of the original 75 genomes only 58 contained ecotopic conversions therefore analysis of gene conversion characteristics is based on this subset. As with our previous work with *E. coli* strains, I find that gene conversions in pathogenic genomes tend to require less sequence similarity and to be more frequent than in nonpathogenic genomes. However, some of these differences (likely due to large variances) are not statistically significant. I also find that closely linked genes are not converted more frequently than dispersed genes, that conversions are not more frequent near the origin of replication and that the location of converted regions are either biased towards the middle of converted genes or found equally distributed along the length of converted genes.

MATERIALS AND METHODS

Sequences. Seventy-five bacterial genomes were retrieved from the NCBI FTP site (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/). The GenBank accession numbers and references for each of these genomes are listed in Supplemental Table 1; found on CD. These genomes were divided into five groups: 16 (nonpathogenic) archaea genomes, 10 nonpathogenic and 19 pathogenic firmicutes genomes, and 6 nonpathogenic and 24 pathogenic proteobacteria genomes.

Multigene families were defined using the BLASTCLUST program (available at the NCBI FTP site). As with our previous study, a gene family was defined as being composed of at least three protein sequences at least 60% identical over at least half their length (28). The protein sequences of each family were aligned using the CLUSTALW program (50). The align2aa PERL script (http://sunflower.bio.indiana.edu/~wfischer/Perl_Scripts/) was used to align the DNA sequences of each multigene family onto their aligned amino acid sequences.

Detection of gene conversions. The GENECONV v1.8 program was used to identify the gene conversions present in each genome (39). Neighbor-joining phylogenetic trees of the multigene families were constructed using the PHYLIP v3.52 software package (14). As previously described (10), these phylogenies were used to eliminate redundantly detected gene conversions.

Gene Conversions characteristics. The genes involved in conversion events and the length of the converted regions were extracted from the GENECONV output files. The sequence similarity of the region flanking the converted regions was calculated from the

percent similarity of the most similar 100 bp region upstream or downstream of the converted region. I only used one of these two values, i.e., the larger of the two. Our flanking region sequence similarities are therefore maximum flanking region sequence similarities. I used these values because gene conversion events are more likely to originate between more similar regions than between less similar ones.

The distance between converted genes found on the same chromosome were calculated using NCBI gene position files (*.ptt files). The location of converted regions was calculated as a percentile of the position of the center of the converted region relative to the distance from the start of the aligned genes (1^{st} percentile) to the end of the aligned genes (100^{th} percentile). The distance of each converted gene from the origin of replication was calculated using NCBI gene position files (*.ptt files).

Since the GENECONV method has been shown to give false positive results when sequences are more than 20% divergent (32), I present two types of results. The results presented in Figures 1 and 2 describe all gene families and gene conversions I identified, respectively, even when the genes involved had less than 80% similarity between them. The results presented in Figures 3, 4 and 5 are only for conversions between genes having at least 80% sequences similarity between them. The results presented in Tables 1, 2 and 3 include both types of results. Presenting our results this way allows us to minimize the possible effect of false positives on our conclusions regarding the characteristics of ectopic conversions in bacterial genomes.

Recombination genes. The position file (*.ptt) of each genome was searched for annotated versions of genes known to be involved in recombination. If a given annotated copy of a gene was not found in a genome, a protein sequence of an annotated gene copy

from a related species was used to search this genome using the FASTA program (30). Protein sequences having matches with E-values smaller than 1×10^{-05} were considered to be homologous genes. Each taxa group of genomes, proteobacteria, firmicute and archaea, was analyzed separately to improve reduce any effects on recombination gene ortholog detection introduced by including very divergent species.

Generalized Linear Models. The effect of the presence of recombination genes on the ectopic gene conversion frequency and the number and size of multigene families (estimated as total number of gene comparisons within each family summed over all families within a species) in each species was determined using a generalized linear model. A general representation of a linear model would be: $Y = mX + nY + \ldots$ The Y represents the dependent variable (i.e., gene conversion frequency or number of gene comparisons). The independent variables (X, Y, ...) represent the presence or absence of specific recombination genes (X = 0 or 1 for absence or presence of a specific gene respectively). The sign and magnitude of the linear coefficients (m, n, ...) indicate the type of correlation (i.e., positive or negative) and the magnitude of the relationship the presence or absence of a gene has with the dependent variable.

Genomes with fewer than 30 gene comparisons were removed from the analysis to reduce stochastic effects on the predicted models. Therefore, four proteobacteria (*Heliobacter pylori J99, Campylobacter jejuni*, *Pasteurella multicoda*, and *Rickettsia conorii*), six firmicute genomes (*Mycobacterium leprae*, *Bifidobacterium longum*, *Listeria monocytogene*, *Clostridium perfringens*, *Mycoplasma pulmonis*, and *Streptococcus pyogenes* M1 GAS), and two archaea (*Methanothermobacter thermautotrophicus*, and *Thermoplasma volcanium*) are excluded from the modeling.

A backward stepwise methodology was used to obtain the optimal set of predictors which explains the variability of the gene conversion frequency and of the number of gene comparisons. This method relies on the Akaike Information Criterion for model comparison. The AIC score is dependent upon the maximum likelihood score of the model fit to the data and on the number of parameters in the model (http://www.garfield.library.upenn.edu/classics1981/A1981MS54100001.pdf). The lower the AIC score the better the goodness of fit of the model it to the data. All terms (genes) with model coefficients that were equal to zero were removed from the models.

GLM significance test. I ran permutation tests to determine how reliable the models are in predicting the effect of the recombination gene complement on gene conversion frequency and on the number and size of multigene families. These permutations tested whether the combination of genes in the original model are better predictors of the gene conversion frequency and of the number of gene comparisons than a random set of genes. The dependent variables (gene conversion frequency or number of gene comparisons) were randomly permuted relative to the associated set of recombination genes present in the genomes. For each of the one-thousand randomly generated data sets, a generalized linear model was fit and the best model and corresponding r^2-value was calculated. If the backward stepwise model fitting algorithm found that none of the genes explained a significant amount of the variability in the dependent variable (i.e., the best model is empty), then those data set permutations were ignored in assessing the significance of the original dataset's model. Since the r^2 scores indicate the amount of variability in the data explained by the presence/absence of the genes included in the model we will reject the null hypothesis of no statistical correlation between the dependent variable (gene

converions frequency or total number of gene comparisons) and the recombination genes included in the model if fewer than 5% (out of the original 1000) of the total non-empty simulated models have r^2-values greater than the original models' fit. If the null hypothesis is rejected, this suggests that the original model fit (i.e., the presence/absence of the genes included in the model) is better than a random set recombination genes.

Statistics. Statistical tests and data management were performed using S-Plus v6.2 (Insightful Corporation, Seattle, Washington) and Excel (Microsoft Corporation, Redmond, Washington). Power tests were performed using G*Power (12).

RESULTS

Genomes without gene conversions. Gene conversions were not identified in some genomes because of three possible situations; no multigene families with at least three members were present, no gene conversions were detected in any multigene family or the maximum flanking similarity for every detected gene conversion was less than 80% (Supplemental Table 1, found on CD). Multigene families with at least three members were absent from four archaea species (*Aeropyrum pernix, Pyrococcus abyssi, Pyrococcus horikoshii* and *Thermoplasma acidophilum*), two firmicutes species (*Mycoplasma genitalium* and *Ureaplasma urealyticum*) and four proteobacteria species (*Buchnera aphidicola Bp, Buchnera aphidicola Sg, Rickettsia prowazekii* and *Haemophilus influenzae Rd*). Gene conversions were not identified in one firmicute species (*Bifidobacterium longum*) and three proteobacteria species (*Brucella suis 1330, Brucella melitensis* and *Rickettsia conorii*). All gene conversions detected within three firmicute species (*Bacillus subtilis*, *Mycobacterium leprae* TN, and *Clostridium tetani* E88) had less than 80% flanking similarity. These genomes were therefore removed from the analysis. The gene conversion characteristics described below are therefore based on 12 (nonpathogenic) archaea genomes, 8 nonpathogenic firmicute genomes, 15 pathogenic firmicute genomes, 4 nonpathogenic proteobacteria genomes and 19 pathogenic proteobacteria genomes.

Size of multigene families. The archaea genomes have families ranging from 3 to 30 genes. The multigene family sizes of nonpathogenic firmicutes range from 3 to 21 genes whereas those of pathogenic firmicutes range from 3 to 18 genes. The multigene family

sizes of nonpathogenic proteobacteria range from 3 to 15 genes whereas those of pathogenic proteobacteria range from 3 to 109 genes (Figure 1).A two sample Kolmogorov-Smirnov goodness of fit test shows that the distribution of gene family sizes are not statistically different between pathogenic and nonpathogenic species of firmicutes and proteobacteria ($p = 0.89$ and 0.68, respectively; Figure 1).

There are significant positive correlations ($p \leq 0.004$) between the sizes of the multigene family and the number of gene conversions for archaea, nonpathogenic firmicute, pathogenic firmicute, nonpathogenic proteobacteria and pathogenic proteobacteria (Spearman rank correlation tests, $r^2 = 0.36$, 0.70, 0.30, 0.54 and 0.15, respectively).

Number and frequency of gene conversions. Two hundred and eighty conversions were identified in archaea genomes. Three hundred twenty-four conversions were identified within the firmicutes, 102 and 222 from the nonpathogenic and pathogenic genomes, respectively. Four hundred and twenty seven gene conversions were identified within the proteobacteria, 86 and 341 from the nonpathogenic and pathogenic genomes, respectively (Supplemental Table 1).

The number of conversions observed in each genome can be transformed into frequency of conversion by dividing it by the total number of gene comparisons (sum total for each gene family of pairwise combinations of paralogous genes within multigene families) for each genome (Supplemental Table 1). If one considers conversion between genes at least 80% similar, conversions occur at the same frequency in all five groups even though, in both firmicutes and proteobacteria, the median frequency of the pathogenic groups are about twice those of nonpathogenic groups (Table 1). However,

the power of these tests varies from medium (0.62) to low (0.06) and may have failed to detect significant differences in gene conversion frequencies; more data is required to clarify this result. The overall low power of these tests reflects the large variance in the gene conversion frequency data as described by the 1st and 3rd quartile values (Table 1). This is supported by the fact that the power of the tests is always lower when all gene conversions are analyzed rather than only conversions between genes that are at least 80% similar in sequence (Table 1).

Length of conversions and flanking region similarity. The majority of conversions are relatively short in all genomes with a median size varying from 100 to 200 bp (Figure 2). If one considers conversion between genes at least 80% similar, conversions are significantly longer in nonpathogenic firmicutes than in pathogenic firmicutes ($p = 0$, Table 2) but the reverse is observed in proteobacteria where conversions are longer in pathogenic proteobacteria than in nonpathogenic proteobacteria ($p = 0$; Table 2). Furthermore, it is unkown whether this statistically significant difference between 125 bp and 131 bp is biologically important. A analogous difference is not found between pathogenic and nonpathogenic proteobacteria with the full dataset ($p = 0.96$; Table 2).

Larger conversions occur more frequently between genes with high flanking region similarity (Figure 2). In fact, the conversion lengths of all five groups are positively correlated with flanking sequence similarity (Spearman rank correlation tests, $p \leq 0.003$). However, the correlations for pathogenic firmicutes and proteobacteria (r^2 of 0.22 and 0.33, respectively) are smaller than the correlations for nonpathogenic firmicutes and proteobacteria (r^2 of 0.46 and 0.55, respectively). The length of

conversions found in archaea are also correlated with flanking sequence similarity and this correlation is similar to those observed in eubacteria ($r^2 = 0.28$).

In firmicutes, the median similarity of flanking regions is significantly smaller in pathogenic genomes (89%) than in nonpathogenic genomes (94.7%; Wilcoxon rank test, $p = 0$; Table 3). This difference is also reflected by the fact that conversions occur with as little as 22% flanking region similarity in pathogenic firmicutes (Figure 2a) but require at least 45% flanking region similarity in nonpathogenic firmicutes (Figure 2b). In contrast, the gene conversions in nonpathogenic and pathogenic proteobacteria have similar minimum flanking sequence similarity (35% and 38%, respectively). In addition, their median flanking similarity requirements are not statistically different (90% and 93%, respectively; Wilcoxon rank test, $p = 0.13$). However, analysis of this test showed that it had very low power (G*POWER F-Test, power = 0.05; Table 3), indicating that a difference might exist in reality but that our test cannot detect it. In archaea, gene conversions require a minimum of 37% sequence similarity and the median flanking similarity is 93% (Figure 2e).

Distance between converted genes. Closely linked genes are not converted more frequently than genes that are far apart (Figure 3). Although most converted genes are found less than 400 kb from one another, this reflects the fact that most multigene family members are close to one another and not the fact that conversions are more frequent between genes that are closer to one another. Conversions frequencies, i.e., the number of conversion per 400 kb interval divided by the number of gene family members per 400 kb interval, are not significantly correlated with distance between converted genes in all five genome groups: r = -0.36 (Spearman rank test, $p = 0.25$), r = -0.64 (Spearman rank

test, $p = 0.053$), r = 0.67 (Spearman rank test, $p = 0.08$), r = 0.44 (Spearman rank test, $p = 0.23$) and r = 0.64 (Spearman rank test, $p = 0.1$) for pathogenic firmicute, nonpathogenic firmicute, pathogenic proteobacteria, nonpathogenic proteobacteria and archaea, respectively (Figure 3).

Proximity to the origin of replication. Conversions are not more frequent in genes that are close to the origin of replication (Figure 4). Although the distance of converted genes from the origin of replication is highly variable, the frequency of converted genes within 200 kb of the origin of replication is roughly equal to the average frequency of converted genes located further away. There are also no obvious differences between pathogenic and nonpathogenic strains (Figures 4a and 4b). Analysis of the relationship between distance and gene conversion frequency indicates that for pathogenic firmicute and proteobacteria, nonpathogenic proteobacteria and archaea genomes no significant correlation exists (r = -0.27, r = 0.22, r = -0.16, and r = -0.08 and Spearman rank correlation test $p = 0.24$, $p = 0.53$, $p = 0.23$, and $p = 0.56$, respectively). Gene conversions occur more frequently between genes which are far from the origin of replication in nonpathogenic firmicute genomes (r = 0.59, Spearman rank correlation test $p = 0.02$).

Location of conversion within genes. The distributions of gene conversion positions (positions are measured as the percentage along the genes' length of the converted regions centre) within genes reach maxima near the middle of converted genes in all groups except nonpathogenic proteobacteria. In archaea, conversions occur most frequently between 30% and 70% of the genes length, and less frequently in the first 10% of the length of the converted genes (χ^2-test, $p = 0.0003$). The genes of nonpathogenic firmicutes have an excess of conversion at 40% of the converted genes length, and fewer

conversion than expected at their 5'- and 3'-ends (χ^2-test, $p = 0.006$). Pathogenic firmicute genomes have an excess of conversions at 70% of the genes length (χ^2-test, $p = 0.008$). In pathogenic proteobacteria, gene conversions occur most frequently between 60% and 90% of the genes length, and conversions are less frequent than expected in the first 30% and last 10% of the converted genes length (χ^2-test, $p = 9.5 \times 10^{-9}$). In contrast, gene conversions are uniformly distributed along the length of converted genes found in nonpathogenic proteobacteria (χ^2-test, $p = 0.49$).

Genus level analyses. I compared gene conversion characteristics between pathogenic and nonpathogenic species from two firmicutes genera: *Bacillus* and *Clostridium*. In the *Bacillus* genus, the pathogenic species *B. anthracis* has a higher gene conversion frequency than the nonpathogenic species *B. haloduran* (6.5%, and 5.3% respectively; Table 1). The mean flanking similarity (*B. anthracis* = 86 ± 5.6 %; *B. haloduran* = 97 ± 2 %) and mean gene conversion lengths (*B. anthracis* = 44 ± 39 bp; *B. haloduran* = 451 ± 222 bp) are significantly smaller in the *B. anthracis* genome than in the *B. haloduran* genome (flanking similarity, two-tailed Z-test, $p = 0.0002$; conversion length, two-tailed Z-test, $p = 0$). In both species, gene conversions are more frequent near the centre of the converted genes (χ^2-test, $p = 0.001$) and gene conversions do not occur more frequently between closely linked genes (χ^2-tests; *B. anthracis*, $p = 0.72$; *B. haloduran*, $p = 0.99$).

The *Clostridium* genus contains one pathogenic (*C. perfringens*) and one nonpathogenic (*C. acetobutylicum*) species. The *C. perfringens* genome has a conversion frequency of 22%, whereas that of the *C. acetobutylicum* genome is 5%. Analysis of the mean flanking similarity (*C. perfringens* = 92.5 ± 3.4 %; *C. acetobutylicum* = 95 ± 6 %) and gene conversion length distributions (*C. perfringens* = 65.8 ± 12.3 bp; *C.*

acetobutylicum = 173 ± 166 bp) show that there are no statistical differences between the flanking similarity or gene conversion length from pathogenic and nonpathogenic species (flanking similarity, two-tailed Z-test, $p = 0.94$; gene conversion length, two-tailed Z-test, $p = 0.13$). Converted regions within genes are uniformly distributed along the length of converted genes (χ^2-test, $p = 0.87$) and gene conversions do not occur more frequently between closely linked genes (χ^2-tests; *C. perfringens*, $p = 1$; *C. acetobutylicum*, $p = 0.06$). Finally, the converted genes of the species from these two genera are also not significantly close to the origin of replication (results not shown).

Recombination genes. Supplemental Table 2 (found on CD) reports the presence or absence of several of the genes known to be involved in recombination in the genomes I analyzed. The absence of gene families in the two *Buchnera* genomes correlates with their lack of a *recA* gene (*radA* in archaea). However, these two genomes also lack many other recombination genes, including *ruvA*, *ruvB* and *ruvC* genes. The *mutL* gene is only lacking from three proteobacteria (*Helicobacter pylori* 26695, *H. pylori* J99 and *Campylobacter jejuni*) and these three proteobacteria have elevated gene conversion frequencies (30%, 33% and 33%, respectively). MutL is involved in the branch migration of Holliday junctions. I therefore compared the gene conversion tract lengths of *H. pylori* 26695 and J99 and *C. jejuni* to determine whether they were affected by the lack of *mutL*. The median gene conversion length in *H. pylori* 26695 (288 bp, mean = 387 ± 288 bp) is significantly larger than the overall median size for proteobacteria (131 bp, mean = 221 ± 246 bp, Wilcoxon two sample test, $p = 0.01$) whereas that of *C. jejuni* (160 bp, mean = 177 ± 41 bp) is not significantly different from the overall proteobacteria median

(Wilcoxon two sample test, $p = 0.14$). *H. pylori* J99 contains a single (35 bp long) conversion. I therefore did not make statistical inferences for this species.

Of the two *Sulfolobus* genomes, that of *S. solfataricus* has a low gene conversion frequency (3%) whereas that of *S. tokodaii* has a much higher gene conversion frequency (18%). These two species differ by the absence of functional *rad2* and *rad54* genes in the *S. tokodaii* genome. Similarly, gene conversion frequencies are at least four times higher in the two strains of *Streptococcus pyogenes* (24 and 33%) than they are in the two strains of *Streptococcus pneumoniae* (3 and 8%). These two species differ from *S. pyogenes* by having a functional *recT* gene and a nonfunctional a*ddB* gene, whereas *S. pneumoniae* has a nonfunctional *recT* gene and a functional *addB* gene.

Despite the examples above, the group of recombination genes found in each genome is most often a poor predictor of gene conversion frequency. For example, the genomes of the three *Staphylococcus aureus* subspecies have gene conversions frequencies varying from 10 to 22% despite the fact that they all lack the same set of functional recombination genes (Supplemental Table 2). Similarly, whereas the *Neisseria meningitidis* strain Z2491has a gene conversion frequency three times higher than that of the MC58 strain (27 versus 9.8%, respectively), both genomes contain the same complement of recombination genes. Furthermore, the two *Salmonella* species, *S. enterica* and *S. typhimurium*, also have widely different gene conversion frequencies (1 and 18%, respectively), but both their genomes contain the same complement of recombination genes.

The very low gene conversion frequency observed in *Shigella flexneri* is surprising given that it contains an almost full complement of recombination pathway

genes as well as a large set of multigene families (its largest family has 109 members). In the *Mycobacterium* genus, the *M. leprae* genome has only one gene family with three members whereas the two *M. tuberculosis* genomes have dozens of multigene family members (Supplemental Table 1). These two species differ by the absence of *recC* and *recD* genes in the *M. leprae* genome.

Generalized Linear Models. I used generalized linear models to determine whether the presence or absence of certain genes could predict the gene conversion frequency and/or the size of the multigene families of different genomes. Note that, for convenience, I used the number of (pairwise) gene comparisons as our measure of the size of multigene families. Furthermore, the gene conversion frequency results presented here are based of the conversion between genes that are at least 80% similar.

The majority of the proteobacteria gene conversion frequency variability is due to the *recC, recG, mutH* and *mutL* genes. The model has a significant fit to the data (multiple $r^2 = 0.5214$, $p = 0.002$). This model indicates that the presence of *recC* and *recG* has a positive effect on the amount of gene conversion in a genome (model coefficients: 0.2600 and 0.4022, respectively). Conversely, the presence of *mutH* and *mutL* has a negative effect on gene conversion (model coefficients: -0.2809 and –0.4743, respectively). The results of the permutation test indicates that 301 simulated models have multiple r^2-values greater than 0.5214 (i.e., $p = 301/870 = 0.34$; 870 indicates the number of non-empty model produced by the permutation test out of a possible 1000). The predictive power of these genes is therefore not any better than a random set of genes.

The total number of gene comparisons in proteobacteria is best modeled by *sbcD* (multiple $r^2 = 0.338$, $p = 0.002$) with the presence of the *sbcD* gene causing an increase in the number of gene comparisons (model coefficient: 30.94). The permutation test found that 718 models had larger multiple r^2-values than 0.338 ($p = 718/1000 = 0.718$). The presence or absence of this gene therefore does not have a significant effect on the size of multigene families.

The firmicute gene conversion frequency is best modeled by the *addB, recC, recD, recG, recJ, recQ, recT, ruvC*, and *sbcD* genes. This model has a statistically significant fit for the data (multiple $r^2 = 0.8724$, $p = 0.0001$). The presence of *addB, recC, recT* and *ruvC* has a negative effect on the amount of gene conversion (model coefficients: -0.4494, -0.1789, -0.3515, -0.1861, and -0.2822, respectively). The presence of *recD, recG, recJ* and *sbcD* increased the amount of gene conversion in firmicute genomes (model coefficients: 0.1836, 0.3098, 0.1606, and 0.1575, respectively). The permutation test results indicate that the null hypothesis should be rejected ($p = 3/816 = 0.004$). This indicates that this model is a better predictor than a random set of genes. The presence or absence of these genes therefore has a significant effect on gene conversion frequency.

The total number of gene comparisons in firmicutes is best modeled by *recD, recF, recJ, recQ* and *ruvC* genes (multiple $r^2 = 0.5844$, $p = 0.01$). The presence of *recD, recJ, recQ*, and *ruvC* induces a negative effect on the number of total gene comparisons (model coefficients: -1.5689, -3.8879, -4.2677, and -2.7191, respectively). In contrast, the presence of *recF* had a positive effect (model coefficient: 4.7880). However, the

permutation test results show that the presence or absence of these genes does not have a significant effect on the size of multigene families ($p = 302/1000 = 0.302$).

The archaea gene conversion frequency is best modeled by the presence or absence of the *mre11, Hje, Hjc, mutS, gyrA*, and *rad54* genes (multiple $r^2 = 0.9678$, $p = 0.002$). The presence of the *mre11, Hje, Hjc, mutS*, and *gyrA* genes is predicted to have a positive effect on the amount of gene conversion in archaea genomes (model coefficient: 3.9295, 1.1676, 1.1813, 1.9207 and 1.1109, respectively). The presence of the *rad54* gene predicts a decrease in the amount of gene conversion (model coefficient: -1.7566). The permutation test results indicate that the null hypothesis should be rejected, therefore these six genes are better predictors of gene conversion frequency than a random set of recombination genes ($p = 6/611 = 0.009$).

The total number of gene comparisons in archaea multigene families **is correlated with** the presence of *gyrA* and *rad54* genes (multiple $r^2 = 0.6111$, $p = 0.005$). The presence of the *rad54* gene has a positive effect on the number of gene comparisons (model coefficients: 1472.4342 respectively). The presence of the *gyrA* gene is predicted to have a negative effect on the number of gene comparisons in a genome (model coefficient: -824.5658). The simulation results indicate that the null hypothesis should not be rejected ($p = 683/822 = 0.83$). The genes in the observed model are therefore not better predictors of the variability in the total number of gene comparisons than a random set of recombination genes.

DISCUSSION

Our analyses allowed us to compare several gene conversion characteristics in five taxonomic groups (archaea, nonpathogenic firmicutes, pathogenic firmicutes, nonpathogenic proteobacteria and pathogenic proteobacteria) and to compare them between pathogenic and nonpathogenic taxa. Although the size of multigene families is not significantly different between pathogenic and nonpathogenic taxa, the number of gene conversions in each family is positively correlated with the multigene family size in the five species groups I studied (r^2-values ranging from 0.15 to 0.70). The fact that the size of genes families are not statistically different between pathogenic and nonpathogenic species of firmicutes and proteobacteria contrasts with the results of our previous study where I had observed that the frequency of large families were greater in pathogenic than nonpathogenic *E. coli* genomes (28). On the other hand, the positive correlation between the number of gene conversions and the size of multigene families is consistent with our previous *E. coli* study where I observed r^2-values up to 0.15 (28). The fact that no difference was found between the numbers of large multigene families in pathogenic and non-pathogenic taxa may be a result of genome sampling. Some types of pathogenic strains undergo genome reduction as a result of co-oping the host cells machinery. This reductive evolution would decrease the size of multigene families in the genomes.

Even though, in both firmicutes and proteobacteria, the median gene conversion frequencies of the pathogenic taxa are about twice those of nonpathogenic taxa, these differences are not significant (Table 1). However, this lack of significant differences is

likely due to the small sample sizes of genomes. A significant difference in gene conversion frequency may therefore exist between pathogenic and nonpathogenic taxa but the dataset was not large enough to detect it. Another way to address this question is to compare pathogenic and nonpathogenic species from the same genus. Gene conversion frequencies of pathogenic species from the genus *Bacillus* and *Clostridium* are greater than those of nonpathogenic species (6.5 versus 5.3% and 22 versus 5%, respectively; Supplemental Table 1). Higher gene conversion frequencies are also observed in more pathogenic strains of the same species. For example, in *Neisseria meningitides*, serogroup A strains are responsible for the majority of epidemics and most of the mortality associated with meningitis (29). One such strain, strain Z2491, has a higher gene conversion frequency (26.8%) than the closely related, but less pathogenic, *N. meningitides* NC58 serogroup B strain (9.8%; Supplemental Table 1). Overall, these results are consistent with our previous *E. coli* study where I observed conversion frequencies of 10.4%, 10.9% and 16.3% for CFT073, EDL933, and Sakai pathogenic genomes, respectively, as opposed to 7.6% for the nonpathogenic K12 genome (28). These results are also consistent with previous work that showed that ectopic recombination repair is important to the survival of pathogenic bacteria and may reflect the selective pressure for higher recombination rates in more virulent genomes (4, 5, 22). This could indicate selection for mismatch repair gene mutations that relax the sequence similar requirements between damaged and template genes and facilitate the creation of greater sequence variability in these genomes in order to aid in immune system response evasion (34). Mutations in mismatch repair genes are common in pathogenic genomes

and are likely advantageous to these strains because such mutations allow them to evade the immune system of their hosts (19, 33).

Within proteobacteria, the median length of gene conversions in pathogenic species (131 bp) is statistically larger than that in nonpathogenic species (125 bp); it is unkown whether this difference has biologically importance. In firmicutes, the median gene conversion length of pathogenic species (100 bp) is significantly smaller than in nonpathogenic genomes (200 bp). This is opposite to what I previously observed in *E. coli* were the median gene conversion length observed in the nonpathogenic K-12 genome (47 bp) was smaller than in the three pathogenic genomes (155 bp, 135 bp and 158 bp for the CFT073, EDL933 and Sakai genomes, respectively). The reasons for this difference are not known.

The length of conversions is positively correlated with flanking sequence similarity in all five groups and these correlations are smaller for pathogenic species than the correlations for nonpathogenic species. This is consistent with our previous observation that gene conversions occur frequenctly between less similar sequences in pathogenic genomes (28). As discussed above, it is also consistent with the hypothesis of selection for greater sequence variability in pathogenic genomes. As a result pathogenic bacteria may allow ectopic gene conversions between more divergent gene copies in order to generate more sequence variability.

In all five genome groups, as well as in *Bacillus* and *Clostridium* pathogenic and nonpathogenic species, our results show that conversions are not more frequent between genes closely linked than between dispersed genes (Figure 3). This agrees with our previous *E. coli* results where I argued that the presence of multiple chromosome bodies

in *E. coli* cells "unlinked" all genes (28). Interestingly, our results imply that multiple chromosome copies are present in most bacterial cells. This suggestion is supported by reports of the presence of multiple chromosomes in *E. coli*, *Neisseria gonorrhoeae* and diverse archaea species (2, 3, 24, 51).

Gene conversions are not more frequent near the origin of replication of pathogenic firmicute, all proteobacteria, and archaea genomes (41; Figure 4). These results are consistent with our previous observations in *E. coli* genomes (28) as well as those made by previous studies (9, 27). In nonpathogenic firmicutes, gene conversions tend to occur more frequently close to the terminus of replication. This result is similar to what I previously observed in the genome of the Sakai strain of *E. coli* (28).

Gene conversions are biased toward the middle of the converted genes in most species except for nonpathogenic proteobacteria where they are uniformly distributed along the length of converted genes (Figure 5). This overall bias towards the middle of converted genes may be the result of the fact that gene conversions occur more frequently when nucleotide sequence similarity extends on both sides of converted regions.

Previous work done on human and yeast gene conversions found that their location was biased towards the 3' end of converted genes (10). This bias is thought to be the result of gene conversions between chromosomal gene copies and cDNA molecules. In agreement with our previous *E. coli* results, our results show that a similar bias is not present in prokaryotic genomes (28). To confirm that no bias exists due to gene conversions via a cDNA intermediate created by reverse transcriptase, I looked for a 3' bias in genomes which have putative copies of reverse transcriptase. A search of the pathogenic proteobacteria annotated genes indicated that two Salmonella strains have

putative copies of reverse transcriptase: *Salmonella enterica* serovar Typhi CT18 and *S. enterica* serovar Typhimurium LT2. Previous studies confirmed the presence of reverse transcriptase in *Salmonella* serovar Typhimurium (1, 25). I did not detect any 3' bias in the pooled data from the *Salmonella* serovar Typhi CT18 and *Salmonella* serovar Typhimurium LT2 genomes (χ^2 test, $p = 0.41$). This confirms that gene conversion involving cDNA intermediates is likely not an important factor in bacterial genomes.

I studied genes involved in recombination to determine if there are any obvious correlations between the presence of specific recombination genes and the gene conversion frequency or the number of duplicated genes. Unfortunately, the wide variation in the presence/absence of many recombination genes makes it difficult to associate the frequency of gene conversion, or the presence of large multigene families, with the presence/absence of particular recombination genes (Supplemental Table 2). For example, the absence of *recA* genes, coding for a protein essential for DNA pairing and strand exchange (23, 48), in two *Buchnera* species correlates with the absence of gene families in these two species. However, even this intuitive conclusion suffers from the fact that several other recombination genes are also missing from these two genomes. The fact that the only three proteobacteria lacking a *mutL* gene (*Helicobacter pylori* 26695, *H. pylori* J99 and *Campylobacter jejuni*) have elevated gene conversion frequencies (31%, 33% and 33%, respectively) may also indicate a causal relationship. However, here again, these three taxa are also lacking other recombination genes. Furthermore, the genes that these three taxa lack are almost the same as those absent from the genome of the two *Rickettsia* species that have neither gene conversions nor multigene family members. Similarly, gene conversion frequencies are at least four times higher in the two strains of

Streptococcus pyogenes (33 and 24%) than they are in the two strains of *Streptococcus pneumoniae* (7 and 3%). Since *S. pyogenes* has a functional *recT* gene and a nonfunctional a*ddB* gene, whereas *S. pneumoniae* has a nonfunctional *recT* gene and a functional *addB* gene, these differences may be responsible for the different gene conversion frequencies of these species. The RecT protein catalyzes the renaturation of complementary ssDNA strands whereas the AddAB enzyme (encoded by the *addA* and *addB* genes) is the counterpart of the *E. coli* RecBCD enzyme (6, 16). However, again, this interpretation does not take into account the fact that these genomes lack functional copies of other genes.

Of the two *Sulfolobus* genomes, that of *S. solfataricus* has a low gene conversion frequency (3%) whereas that of *S. tokodaii* has a much higher gene conversion frequency (17%). Paradoxically, only the *S. tokodaii* genome is missing a functional *rad2* gene (coding for an endonuclease necessary for excision repair; 26) and a functional *rad54* gene (coding for a protein necessary for the strand exchange of recombination repair; 47). If recombination gene content was the only factor affecting gene conversion frequency, one would therefore have expected a lower gene conversion frequency in the *S. tokodaii* genome than in the *S. solfataricus* genome because, in yeast, loss of the Rad54 function decreases the amount of inter and intrachromosomal gene conversions (43). On the other hand, the absence of these two genes in the *S. tokodaii* genome might explain why it has less multigene family members.

The group of recombination genes in a genome is also a poor predictor of gene conversion frequency between closely related species. For example, the genomes of the three *Staphylococcus aureus* subspecies have gene conversions frequencies varying from

10 to 22 % despite the fact that they all lack the same set of functional recombination genes. Similarly, whereas the *Neisseria meningitidis* strain Z2491has a gene conversion frequency three times higher than that of the MC58 strain (26.8 versus 9.79%, respectively), both genomes contain the same complement of recombination genes. Furthermore, the two *Salmonella* species, *S. enterica* and *S. typhimurium*, also have widely different gene conversion frequencies (1 and 18%, respectively), but both their genomes contain the same complement of recombination genes. These differences in gene conversion frequency are therefore either caused by the presence or absence of genes I did not analyze or by mutations in some recombination genes.

The permutation tests show that the gene conversion frequency of proteobacteria species cannot be predicted based solely on the presence or absence of recombination genes. In contrast, within firmicutes the simulations indicate that gene conversion frequency variability could be explained by recombination gene complement. The model specified that the presence of *addB, recC, recT* and *ruvC* is negatively correlated with the amount of gene conversion, whereas the presence of *recD, recG, recJ* and *sbcD* is correlated with an increase in the amount of gene conversion in firmicute genomes. Unfortunately this model does not agree with accepted biological knowledge. The AddA/AddB complex is analogous to recBCD, which is an integral component of dsDNA recombination repair. In the absence of AddB the rate of gene conversion in firmicutes should be low. This indicates that the biological relevance of the fitted model should be questioned. Similarly, the gene conversion frequencies in archaea genomes are predictable using the presence or absence *of mre11, Hje, Hjc, mutS, gyrA*, and *rad54*. The model indicates that the presence of the Rad54 protein is negatively correlated with the

gene conversion frequency in archaea. This is contrary to previous work which showed that ectopic recombination rate was decreased 30-fold in yeast lacking *rad54* (21). I therefore conclude that the gene complements studied here do not explain the variability of gene conversion frequency or size and number of multigene families.

Other observations are also puzzling. For example, the very low gene conversion frequency observed in *Shigella flexneri* is surprising given that it contains an almost full complement of recombination pathway genes as well as a large set of multigene families (its largest family has 109 members). Since the frequency of gene conversions is correlated with the size of multigene families I would expect to find a high rate of gene conversion in this genome. Furthermore, its recombination pathway genes are most likely functional because this strain is the most prevalent cause of widespread bacillary dysentery and those recombination pathway genes are necessary for virulence (5, 37).

Recombination gene complement is also a poor predictor of the number of multigene family members present in a genome. For example, the *Mycobacterium leprae* genome has only one gene family with three members whereas the two *Mycobacterium tuberculosis* strains have dozens of multigene family members (Supplemental Table 1). Since these two species have very similar recombination gene contents, except for the absence of *recC* and *recD* genes in the *M. leprae* genome. The absence of these two genes might be responsible for the near absence of multigene family members in the *M. leprae* genome because the loss of *recC* and *recD* function would inhibit the recombination mechanism resulting in fewer duplication events. However, these two genes are also missing from numerous other species and many of these species (e.g., *Bacillus haloduran*, *Mycoplasma penetrans* and *Mesorhizobium loti*) have dozens of

multigene family members (Supplemental Tables 1 and 2). Finally, the permutation tests show that in firmicutes, proteobacteria, and archaea genomes, gene complement is not a useful predictor of the size and number of multigene families.

In conclusion, our results show that a trend between high gene conversion frequency and pathogenicity (although not statistically significant) does exist. Furthermore, differences between closely related pathogenic and nonpathogenic species do indicate that conversion occurs more frequently in pathogenic strains. In addition, the flanking sequence similarity requirements for pathogenic bacteria are lower than for nonpathogenic bacteria. This suggests that relaxed recombination requirements is an important characteristic in some pathogenic bacteria. In bacteria, gene conversions do not occur more frequently between closely linked genes, do not occur more frequently in genes close to the origin of replication and are usually more frequent in the middle of genes. The fact that these results are common to very diverse bacterial species suggests that they likely apply to most, if not all, bacterial genomes.

ACKNOWLEDGEMENTS

This work was supported by a Discovery Grant from the National Science and Engineering Research Council of Canada.

REFERENCES

1. **Ahmed, A. M., and T. Shimamoto. 2003.** msDNA-St85, a multicopy single-stranded DNA isolated from *Salmonella enterica* serovar Typhimurium LT2 with the genomic analysis of its retron. FEMS Microbiol. Lett. 224:291-297.
2. **Akerlund, T., K. Nordstrom, and R. Bernander.** 1995. Analysis of cell size and DNA content in exponentially growing and stationary-phase batch cultures of *Escherichia coli*. J. Bacteriol. **177:** 6791–6797.
3. **Bernander, R., and A. Poplawski. 1997.** Cell cycle characteristics of thermophilic archaea. J. Bacteriol. **179:**4963-4969.
4. **Buchmeier, N. A., C. J. Lipps, M. Y. So, and F. Heffron.** 1993. Recombination-deficient mutants of Salmonella typhimurium are avirulent and sensitive to the oxidative burst of macrophages. Mol. Microbiol. **7:**933-936.
5. **Cano, D. A., M. G. Pucciarelli, F. Garcia-del Portillo, and J. Casadesus.** 2002. Role of the RecBCD recombination pathway in Salmonella virulence. J. Bacteriol. **184:**592-595.
6. **Chédin, F., P. Noirot, V. Biaudet, and S. D. Ehrlich.** 1998. A five-nucleotide sequence protects DNA from exonucleolytic degradation by AddAB, the RecBCD analogue of *Bacillus subtilis*. Mol. Microbiol. 29:1369-1377.
7. **Clyman, J., and R. P. Cunningham.** 1987. *Escherichia coli* K-12 mutants in which viability is dependent on *recA* function. J. Bacteriol. **169:**4203-4210.
8. **Das Gupta, C., and C. M. Radding.** 1982. Lower fidelity of RecA protein catalysed homologous pairing with a superhelical substrate. Nature **295:**71-73.
9. **Daubin, V., and G. Perrière.** 2003. G+C3 structuring along the genome: a common feature in prokaryotes. Mol. Biol. Evol. **20:**471-483.
10. **Drouin, G.** 2002. Characterization of the gene conversions between the multigene family members of the yeast genome. J. Mol. Evol. **55:**14-23.
11. **Dudas, A., E. Markova, D. Vlasakova, A. Kolman, Z. Bartosova, J. Brozmanova, and M. Chovanec.** 2003. The *Escherichia coli* RecA protein complements recombination defective phenotype of the *Saccharomyces cerevisiae rad52* mutant cells. Yeast **20:**389-396.
12. **Erdfelder, E., F. Faul, and A. Buchner.** 1996. GPOWER: A general power analysis program. Behav. Res. Methods. Instrum. Comput. **28:**1-11.
13. **Essers, J., R. W. Hendriks, S. M. Swagemakers, C. Troelstra, J. de Wit, D. Bootsma, J. H. Hoejmakers, and R. Kanaar.** 1997. Disruption of mouse RAD54 reduces ionizing radiation resistance and homologous recombination. Cell **89:**195-204.
14. **Felsenstein, J.** 1989. PHYLIP – Phylogeny Inference Package. Cladistics **5:**164-166.
15. **Hashimoto, J. G., B. S. Stevenson, and T. M. Schmidt.** 2003. Rates and consequences of recombination between rRNA operons. J. Bacteriol. 185: 966-972.
16. **Kooistra, J., B. J. Haijema, and G. Venema.** 1993. The *Bacillus subtilis* addAB genes are fully functional in *Escherichia coli*. Mol. Microbiol. **7:**9159-9123.
17. **Komori, K., S. Sakae, H. Shinagawa, K. Morikawa, and Y. Ishino.** 1999. A Holliday junction resolvase form *Pyrococcus furiosus*: functional similarity to *Escherichia coli* RuvC provides evidence for conserved mechanism of homologous

recombination in Bacteria, Eukarya, and Archaea. Proc. Natl. Acad. Sci. USA **96:**8873-8878.

18. **Konrad, E. B.** 1977. Method for the isolation of *Escherichia coli* mutants with enhanced recombination between chromosomal duplications. J. Bacteriol. **130:**167-172.
19. **LeClerc, J. E., B. Li, W. L. Payne, and T. A. Cebula.** 1996. High mutation frequencies among *Escherichia coli* and Salmonella pathogens. Science **274:**1208-1211.
20. **Liao, D.** 2000. Gene conversion drives within genic sequences: concerted evolution of ribosomal RNA genes in bacteria and archaea. J. Mol. Evol. **51:**305-317.
21. **Liefshitz, B., A. Parket, R. Maya, and M. Kupiec.** 1995. The role of DNA repair genes in recombination between repeated sequences in yeast. Genetics **140:**1199-1211.
22. **Loughlin, M. F., F. M. Barnard, D. Jenkins, G. J. Sharples, and P. J. Jenks.** 2003. *Helicobacter pylori* mutants defective in RuvC Holliday junction resolvase display reduced macrophage survival and spontaneous clearance from the murine gastric mucosa. Infect. Immun. **71:**2022-2031.
23. **Lloyd, R. G., and K. B. Low.** 1996. Homologous recombination, pp. 2236-2255. *In* F. C. Neidhardt, R. Curtiss III, J. L. Ingraham, E. C. C. Lin, K. B. Low, B. Magasanik, W. S. Reznikoff, M. Riley, M. Schaechter, and H. E. Umbarger (ed.) *Escherichia coli* and *Salmonella*: cellular and molecular biology, 2nd edition. American Society for Microbiology, Washington, D.C.
24. **Malandrin, L., H. Huber, and R. Bernander.** 1999. Nucleoid structure and partition in *Methanococcus jannaschii*: an archeon with multiple copies of the chromosome. Genetics **152:**1315-1323.
25. **Matiasovicova, J., M. Faldynova, M. Pravcova, R. Karpiskova, I. Kolackova, J. Damborsky, and I. Rychlik.** 2003. Retron reverse transcriptase rrtT is ubiquitous in strains of *Salmonella enterica* serovar Typhimurium. FEMS Microbiol. Lett. **223:**281-286.
26. **McCready, S., and L. Marcello.** 2003. Repair of UV damage in *Halobacterium salinarum*. Biochem. Soc. Trans. **31:**694-698.
27. **Mira, A., and H. Ochman.** 2002. Gene location and bacterial sequence divergence. Mol. Biol. Evol. **19:**1350-1358.
28. **Morris, R. T., and G. Drouin.** 2004. Ectopic gene conversions in four *Escherichia coli* genomes: increased recombination in pathogenic strains. J. Mol. Evol. **58:**596-605.
29. **Parkhill, J., M. Achtman, K. D. James, S. D. Bentley, C. Churcher, S. R. Klee, G. Morelli, D. Basham, D. Brown, T. Chillingworth, R. M. Davies, P. Davis, K. Devlin, T. Feltwell, N. Hamlin, S. Holroyd, K. Jagels, S. Leather, S. Moule, K. Mungall, M. A. Quail, M. A. Rajandream, K. M. Rutherford, M. Simmonds, J. Skelton, S. Whitehead, B. G. Spratt, and B. G. Barrell.** 2003. Complete DNA sequence of a serogroup A strain of *Neisseria meningitidis* Z2491. Nature **404:**502-506.
30. **Pearson, W. R., and D. J. Lipman.** 1988. Improved tools for biological sequence comparison. Proc. Natl. Acad. Sci. USA **85:**2444-2448.

31. **Petes, T. D., and C. W. Hill.** 1988. Recombination between repeated genes in microorganisms. Annu. Rev. Genet. **22:**147-168.
32. **Posada, D., and K. A. Crandall.** 2001. Evaluation of methods for detecting recombination from DNA sequences: computer simulations. Proc. Natl. Acad. Sci. USA **98:**13757-13762.
33. **Rattray, A. J., and J. N. Strathern.** 2003. Error-pone DNA polymerases: when making a mistake is the only way to get ahead. Annu. Rev. Genet. **37:**31-66.
34. **Rayssiguier, C., D. S. Thaler, M. Radman.** 1989. The barrier to recombination between *Escherichia coli* and *Salmonella typhimurium* is disrupted in mismatch-repair mutants. Nature **342:**396-401.
35. **Reid, S. D., C. J. Herbelin, A. C. Bumbaugh, R. K. Selander, and T. S. Whittam.** 2000. Parallel evolution of virulence in pathogenic *Esherichia coli.* Nature **406:**64-67.
36. **Sandler, S. J., L. H. Satin, H. S. Samra, and A. J. Clark.** 1996. *recA*-like genes from three archaean species with putative protein products similar to Rad51 and Dmc1 proteins of the yeast Saccharomyces cerevisiae. Nucleic Acid Res. **24:**2125-2132.
37. **Sansonetti, P. J.** 2001. Microbes and microbial toxins:paradigms for microbial-mucosal interatctions III. Shigellosis: from symptoms to molecular pathogenesis. Am. J. Physiol. Gastrointest. Liver Physiol. 280:G319-G323.
38. **Santoyo, G., and D. Romero.** 2005. Gene conversion and concerted evolution in bacterial genomes. FEMS Microbiol. Rev. **29:** 169-183.
39. **Sawyer, S. A.** 1999. GENECONV: A computer package for the statistical detection of gene conversion. [Online.] http://www.math.wustl.edu/~sawyer.
40. **Seitz, E. M., J. P. Brockman, S. J. Sandler, A. J. Clark, and S. C. Kowalczykowski.** 1998. RadA protein is an archaeal RecA protein homolog that catalyzes DNA strand exchange. Genes Dev. **12:**1248-1253.
41. **Sharp, P. M., D. C. Shields, K. W. Wolfe, and W.-H. Li.** 1989. Chromosomal location and evolutionary rate variation in enterobacterial genes. Science **246:**808-810.
42. **Shen, P., and H. V. Huang.** 1986. Homologous recombination in *Escherichia coli*: dependence on substrate length and homology. Genetics **112:**441-457.
43. **Shinohara, M., E. Shita-Yamaguchi, J. M. Buerstedde, H. Shinagawa, H. Ogawa, and A. Shinohara.** 1997. Characterization of the roles of the Saccharomyces cerevisiae RAD54 gene and a homologue of RAD54, RDH54/TUD1, in mitosis and meiosis. Genetics **147:**1545-1556.
44. **Smith, G. R.** 2001. Homologous recombination near and far from DNA breaks: alternative roles and contrasting views. Ann. Rev. Genet. **35:**243-274.
45. **Snyder, L., and W. Champness.** 2002. Molecular basis of recombination. pp. 343-368. *In* L. Snyder, and W. Champness (ed.), Molecular genetics of bacteria, 2nd edition. American Society for Microbiology,Washington, D.C.
46. **Sung, P., L. Krejci, S. Van Komen, and M. G. Sehorn.** 2003. Rad51 recombinase and recombination mediators. J. Biol. Chem. **278:**42729-42732.
47. **Symington, L. S.** 2002. Role of RAD52 epistasis group genes in homologous recombination and double-strand break repair. Microbiol. Mol. Biol. Rev. **66:**630-

670.
48. **Tamas, I., L. Klasson, B. Canback, A. K. Naslund, A. S. Eriksson, J. J. Wernegreen, J. P. Sandstrom, N. A. Moran, S. G. Andersson.** 2002. 50 million years of genomic stasis in endosymbiotic bacteria. Science **296:**2376-2379.
49. **Tenor, J. L., B. A. McCormick, F. M. Ausubel, and A. Aballay.** 2004. *Caenorhabditis elegans*-based screen identifies Salmonella virulence factors required for conserved host-pathogen interactions. Curr. Biol. **14:**1018-1024.
50. **Thompson, J. D., D. G. Higgins, and T. J. Gibson**. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. **22:**4673-4680.
51. **Tobiason, D. M., and H. S. Seifert. 2006**. The obligate human pathogen, *Neisseria gonorrhoeae*, is polyploid. PLoS Biol. **4:**e185.
52. **Watt, V. M., C. J. Ingles, M. S. Urdea, and W. J. Rutter.** 1985. Homology requirements for recombination in *Escherichia coli*. Proc. Natl. Acad. Sci. USA **82:**4768-4772.
53. **Zieg, J., V. F. Maples, and S. R. Kushner.** 1978. Recombination levels of *Escherichia coli* K-12 mutants deficient in various replication, recombination and repair genes. J. Bacteriol. **134:**958-966.

Table 1. The list of gene conversion frequencies (%) calculated for each of the five groups of genomes as well as statistical comparison of the median values between each group was performed using the Wilcoxon rank sum test.

Group	Median	1st quartile	3rd quartile	Minimum frequency	Maximum frequency
1. Pathogenic firmicute	10 (14.3)	6 (10.5)	22.2 (33.3)	2 (0.009)	33 (46.7)
2. Nonpathogenic firmicute	5 (5.6)	3.5 (4.7)	5.8 (8.8)	3 (3.5)	14 (14)
3. Pathogenic proteobacteria	9.8 (11.3)	2.3 (1.8)	25.3 (36.5)	0.002	45 (167)
4. Nonpathogenic proteobacteria	4.7 (7)	4.3 (6.2)	5.7 (8.9)	3.2 (6)	8.5 (12)
5. Archaea	6 (10)	4.5 (5.6)	14.6 (21.7)	1 (1)	67 (67)

Group comparisons	P-value	Power
1. vs 2.	0.06 (0.04)	0.62 (0.05)
1. vs 3.	0.68 (0.19)	0.12 (0.05)
2. vs 4.	0.93 (0.63)	0.06 (0.08)
3. vs 4.	0.65 (0.82)	0.47 (0.05)
2. vs 5.	0.30 (0.33)	0.42 (0.05)
4. vs 5.	0.38 (0.68)	0.24 (0.05)

Notes. The median, 1st and 3rd quartile, minimum and maximum frequencies of gene conversions with at least 80% flanking similarity and all conversions (bracketed values) are presented. P-value and power statistics for the Wilcoxon two sample tests between medians of groups are provided.

Table 2. Summary statistics of the gene conversion lengths (bp) for each group of genomes. Pairwise comparison of median gene conversion lengths between each group of genomes was performed using the Wilcoxon rank sum test.

Group	Median	1st quartile	3rd quartile	Minimum length	Maximum length
1. Pathogenic firmicute	100 (83)	65 (44)	163 (152)	9 (5)	1486 (1576)
2. Nonpathogenic firmicute	200 (170)	84 (56)	369 (362)	14 (12)	1442 (1442)
3. Pathogenic protoebacteria	131 (98)	65 (41)	288 (235)	9 (5)	1515 (1929)
4. Nonpathogenic proteobacteria	125 (90)	40 (32)	317 (350)	5 (5)	1260 (1386)
5. Archaea	115 (94)	56 (48)	224 (199)	4 (4)	1392 (1392)

Group Comparisons	P-value	Power
1. vs. 2.	0 ($1.4x10^{-6}$)	NA (NA)
3. vs. 4.	0 (0.96)	NA (0.50)
2. vs. 5.	0.0005 (0.0006)	NA (NA)
4. vs. 5.	0.93 (0.99)	0.11 (0.42)

Notes. The median, 1st and 3rd quartile, minimum and maximum gene conversion tract lengths of gene conversions with at least 80% flanking similarity and all conversions (bracketed values) are presented. P-value and power statistics for the Wilcoxon two sample tests between medians of groups are provided. NA, not applicable.

Table 3. Summary statistics for the maximum flanking similarity (%) for each group. Pairwise Wilcoxon rank sum tests were performed to determine whether maximum flanking similarity medians are statistically different between groups of genomes.

Group	Median	1st quartile	3rd quartile	Minimum similarity	Maximum similarity
1. Pathogenic firmicute	89 (83)	84 (74)	94 (91)	80 (22)	100 (100)
2. Nonpathogenic firmicute	94.7 (93)	90 (87.3)	97 (97)	80 (45)	100 (100)
3. Pathogenic protoebacteria	93 (88)	87 (78)	96 (95)	80 (38)	100 (100)
4. Nonpathogenic proteobacteria	90 (84)	86 (69)	96 (92)	80 (35)	100 (100)
5. Archaea	93 (90)	88 (81)	97.7 (97)	80 (37)	100 (100)

Group Comparisons	P-value	Power
1. vs. 2.	0 (4.8×10^{-15})	NA (NA)
3. vs. 4.	0.13 (0.47)	0.32 (0.05)
2. vs. 5.	0 (0.02)	NA (NA)
4. vs. 5.	0.01 (0)	NA (NA)

Notes. The median, 1st and 3rd quartile, minimum and maximum of the maximum flanking similarity for conversions with at least 80% flanking similarity and all conversions (bracketed values) are presented. P-value and power statistics for the Wilcoxon two sample tests between medians of groups are provided. NA, not applicable.

Figure legends

Figure 1 Frequency of multigene family sizes within the pathogenic and nonpathogenic datasets of proteobacteria, firmicute and archaea.

Figure 2. Relationship between gene conversion length and maximum flanking similarity for all gene conversions. A) Pathogenic firmicutes (n = 360), B) Nonpathogenic firmicutes (n = 117), C) Pathogenic proteobacteria (n = 469), D) Nonpathogenic proteobacteria (n= 138), E) Archaea (n = 360). An outlier (99%, 7578 bp) was removed from panel C.

Figure 3. Distribution of the distances between multigene family members and of the distances between converted genes having at least 80% maximum flanking similarity. The conversion frequency is displayed by a grey line. A) Pathogenic firmicute, B) Nonpathogenic firmicute, C) Pathogenic proteobacteria, D) Nonpathogenic proteobacteria, E) Archaea. The legend for all panels of figure 3 is provided below panel E.

Figure 4. Relationship between gene conversion frequency and distance from the origin of replication. Gene conversion frequencies are those of gene pairs with at least 80% maximum flanking similarity. A) Firmicutes, B) Proteobacteria, C) Archaea. The legend for the three figure panels is supplied in panel A.

Figure 5. Distrubution of the converted regions within the converted genes having at least 80% maximum flanking similarity. A) Firmicutes (n = 324), B) Proteobacteria (n = 427), C) Archaea (n = 280). The relative position of each conversion is calculated as the location of the centre of converted regions with respect to the length of the converted gene. The legend for the three figure panels is provided in panel A.

Figure 1.

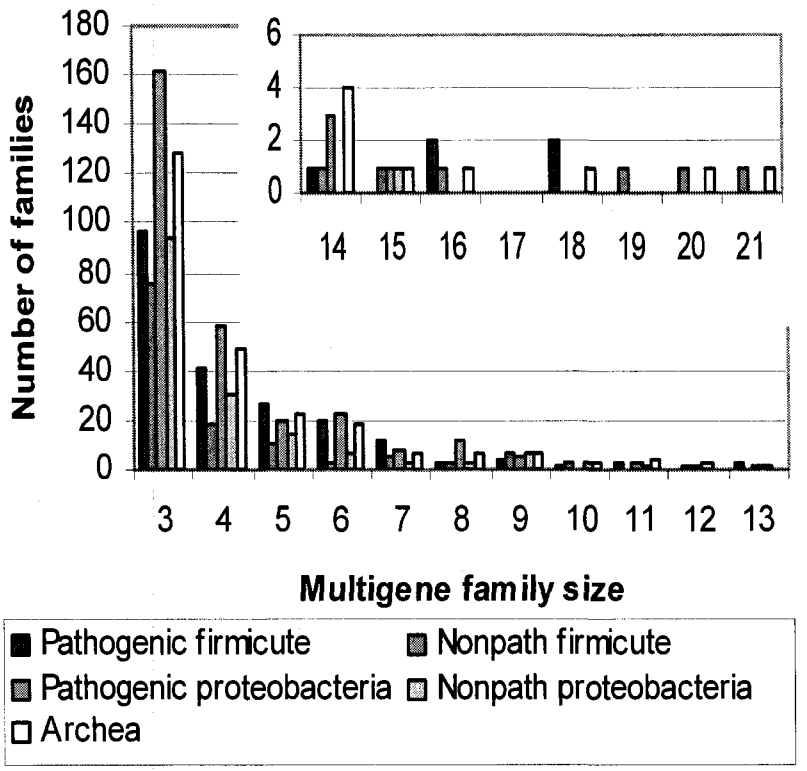
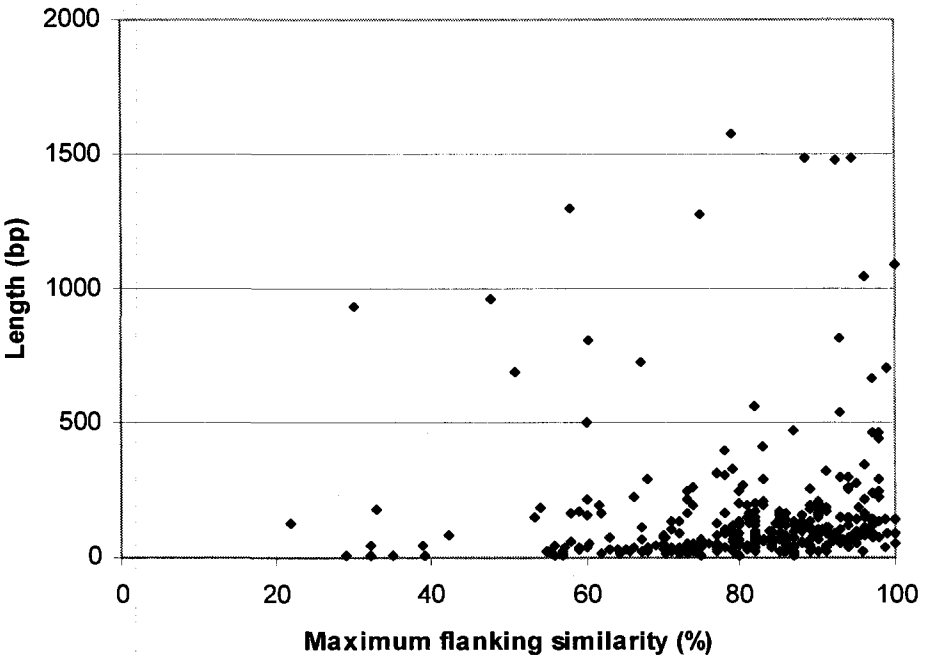
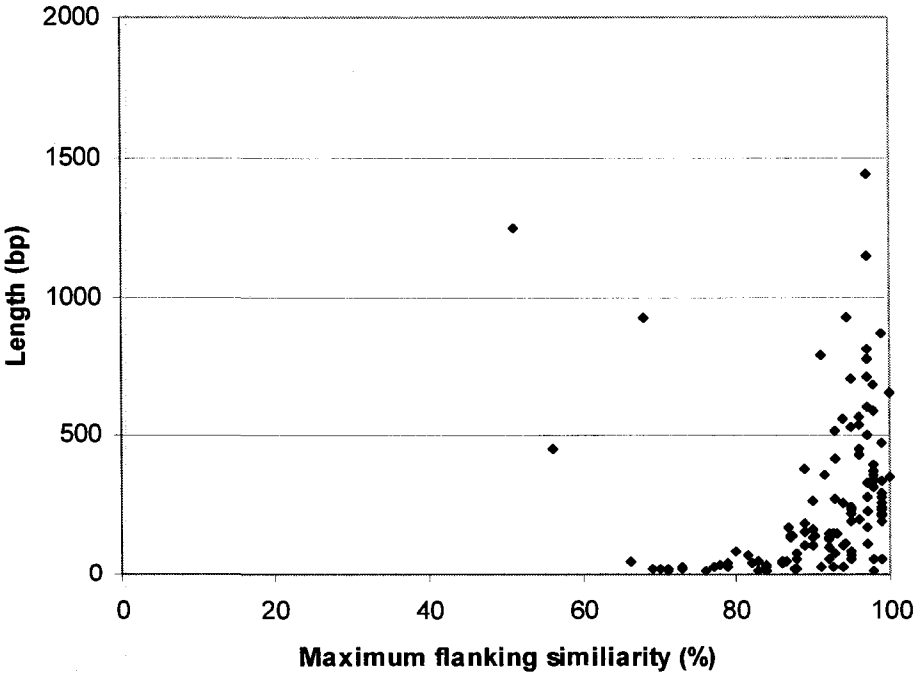

Number of families
180
160
140
120
100
80
60
40
20
0
6
4
2
0
14 15 16 17 18 19 20 21
3 4 5 6 7 8 9 10 11 12 13
Multigene family size
Pathogenic firmicute
Nonpath firmicute
Pathogenic proteobacteria
Nonpath proteobacteria
Archea

Figure 2.

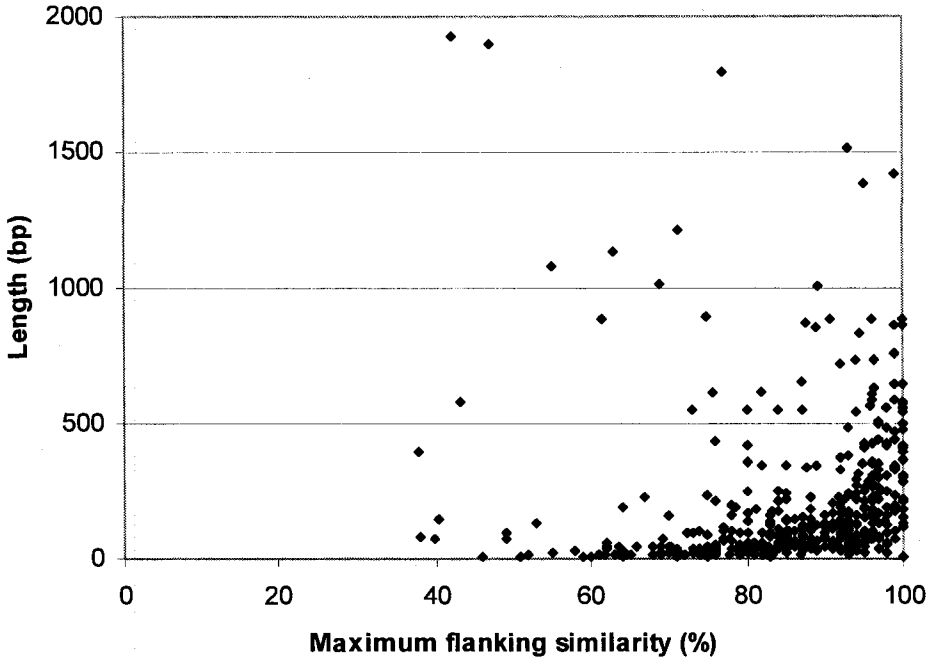
A


2000
1500
1000
500
0
Length (bp)
0
20
40
60
80
100
Maximum flanking similarity (%)

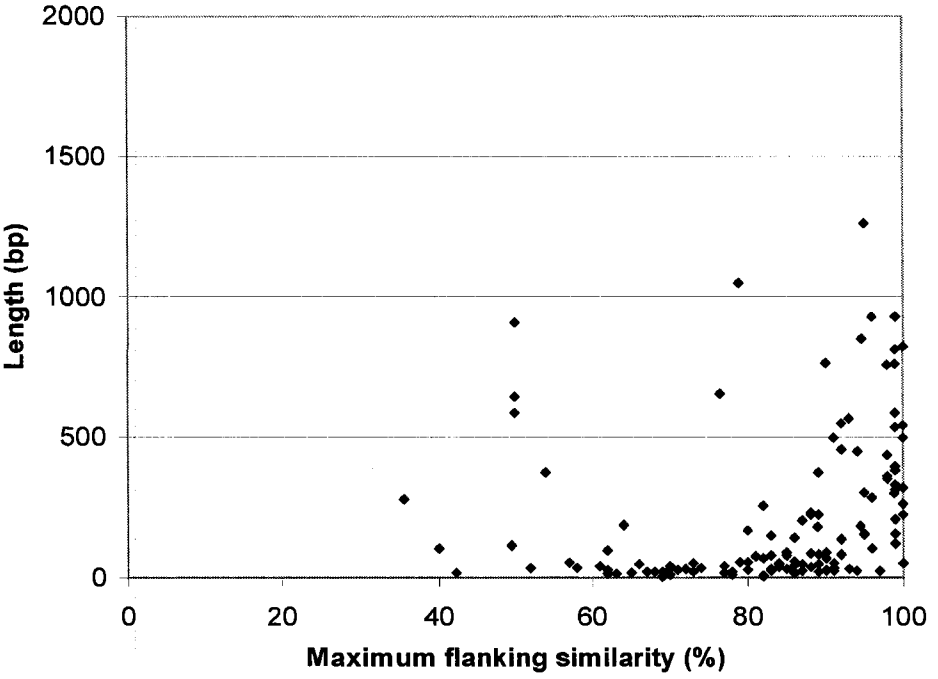
B


2000
1500
1000
500
0
Length (bp)
0
20
40
60
80
100
Maximum flanking similiarity (%)

C


2000
1500
1000
500
0
Length (bp)
0
20
40
60
80
100
Maximum flanking similarity (%)

D


2000
1500
1000
500
0
Length (bp)
0
20
40
60
80
100
Maximum flanking similarity (%)

E

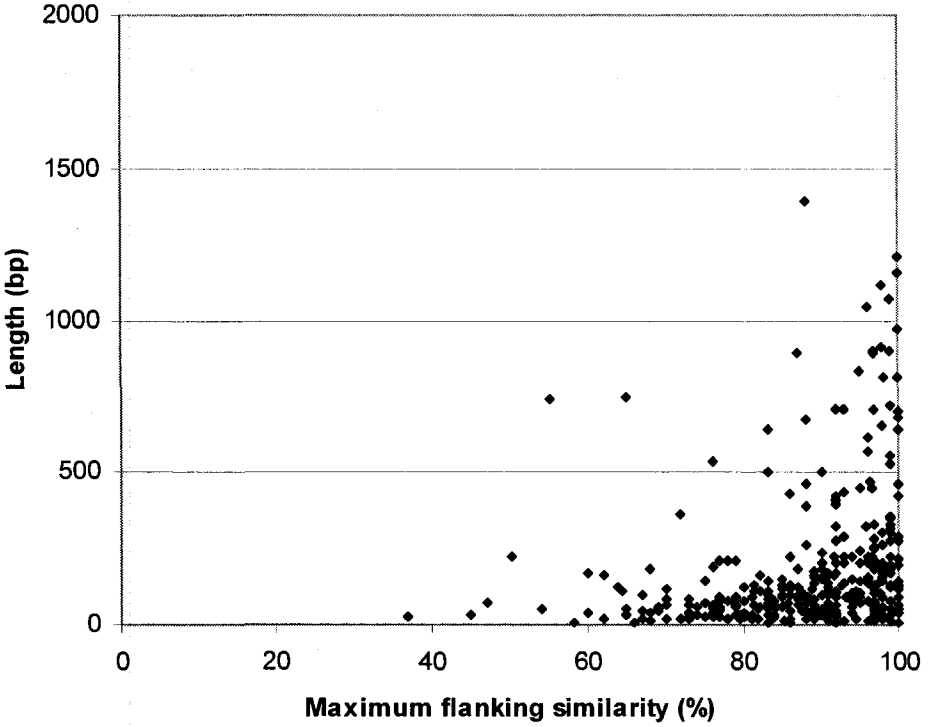
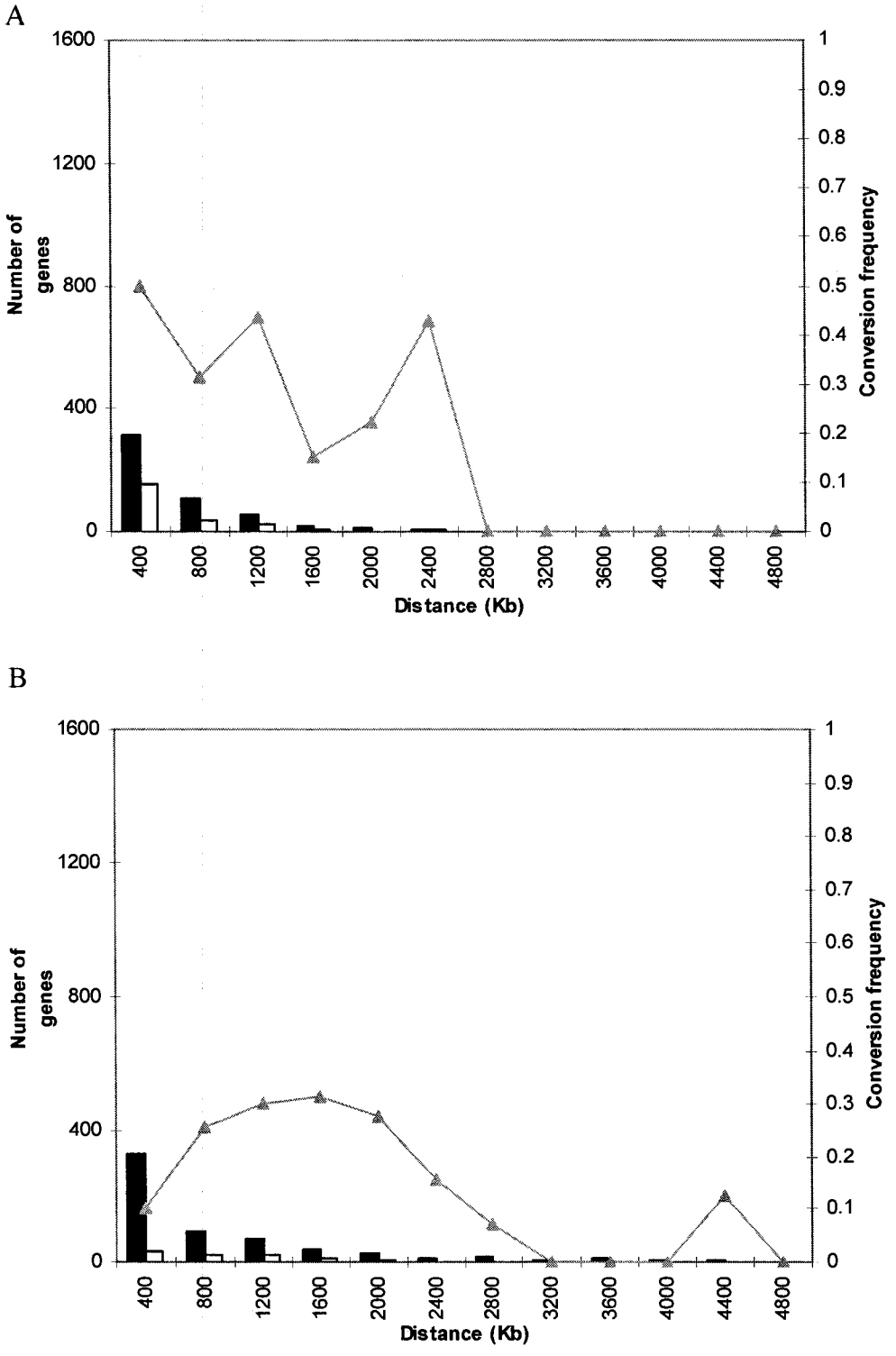
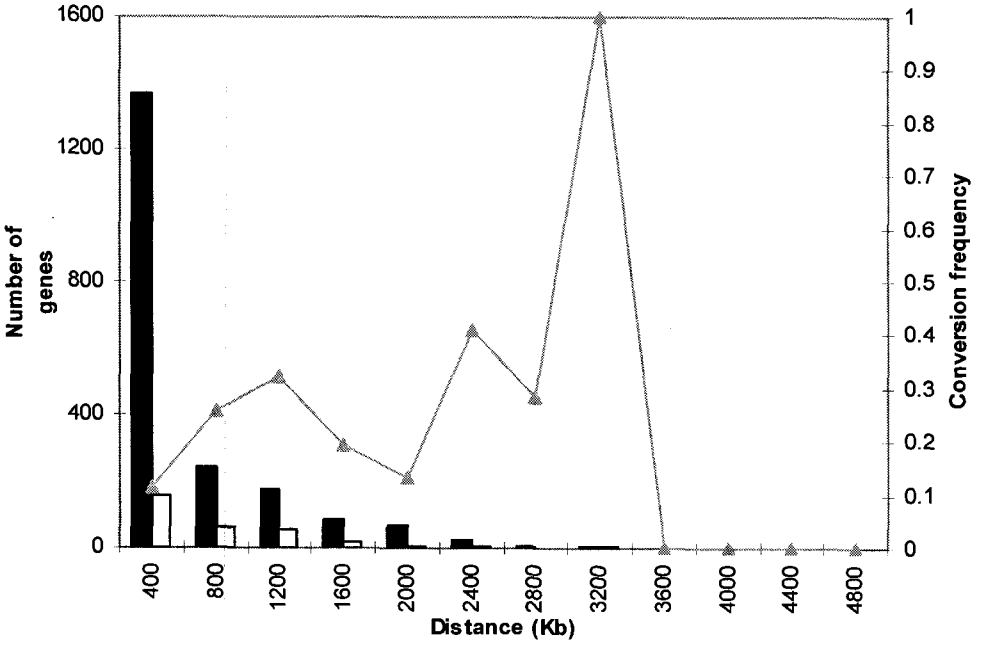

2000
1500
1000
500
0
Length (bp)
0
20
40
60
80
100
Maximum flanking similarity (%)

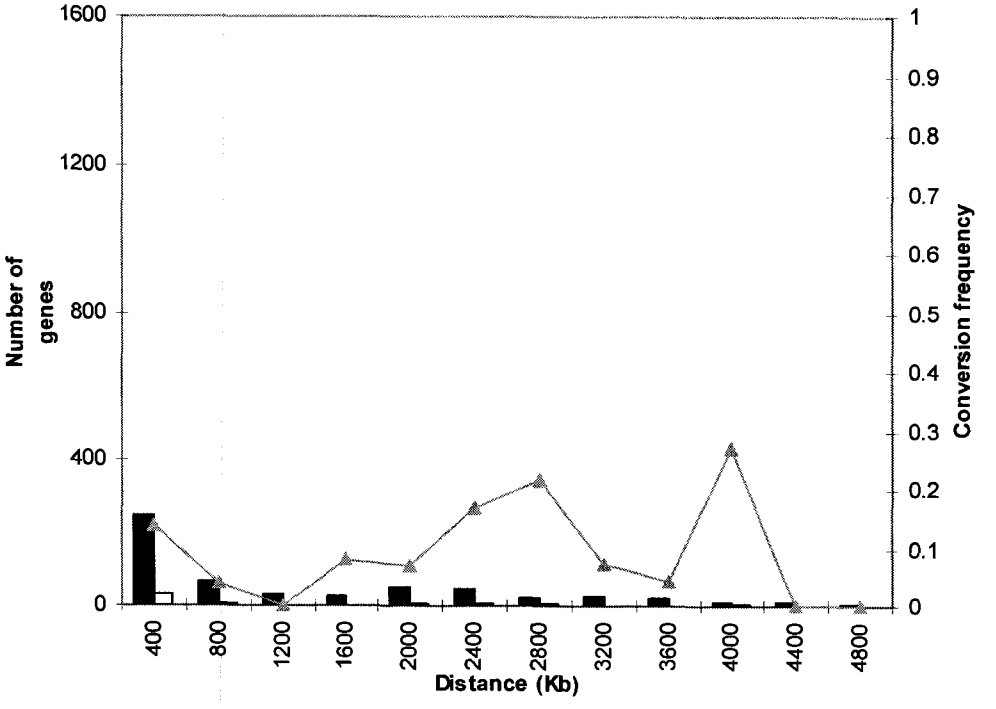
Figure 3


A
1600
1200
800
400
0
Number of genes
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Conversion frequency
400
800
1200
1600
2000
2400
2800
3200
3600
4000
4400
4800
Distance (Kb)
B
1600
1200
800
400
0
Number of genes
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Conversion frequency
400
800
1200
1600
2000
2400
2800
3200
3600
4000
4400
4800
Distance (Kb)

C


1600
1200
800
400
0
Number of genes
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Conversion frequency
400
800
1200
1600
2000
2400
2800
3200
3600
4000
4400
4800
Distance (Kb)

D


1600
1200
800
400
0
Number of genes
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Conversion frequency
400
800
1200
1600
2000
2400
2800
3200
3600
4000
4400
4800
Distance (Kb)

E

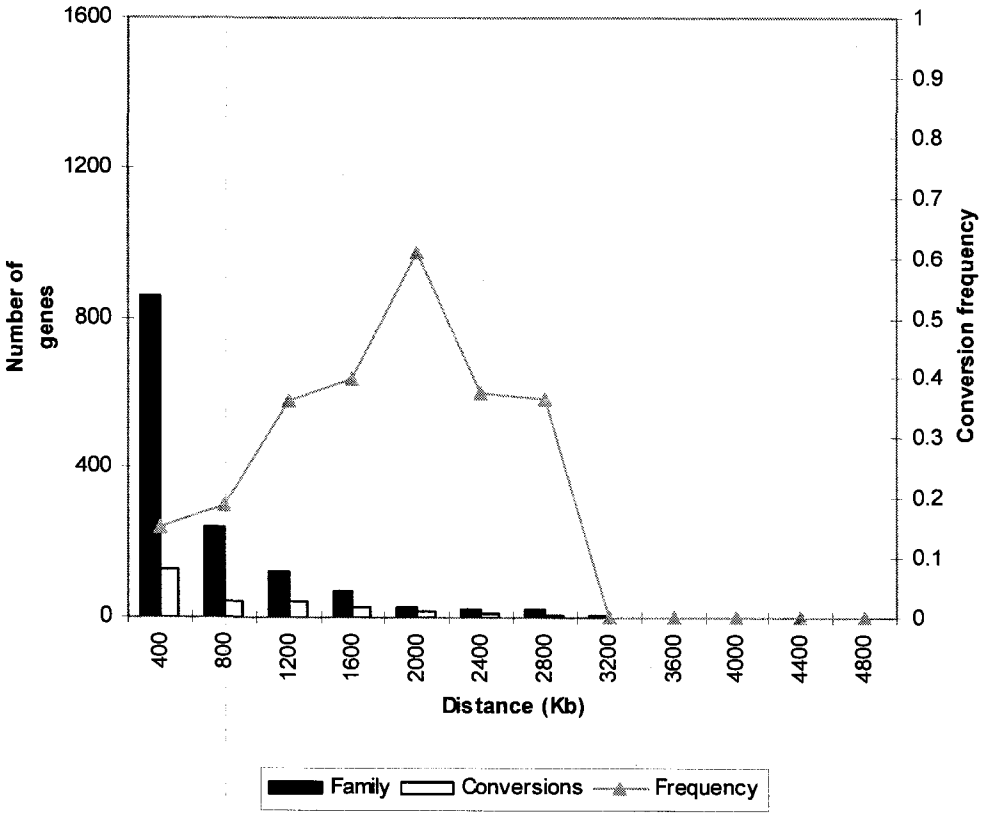
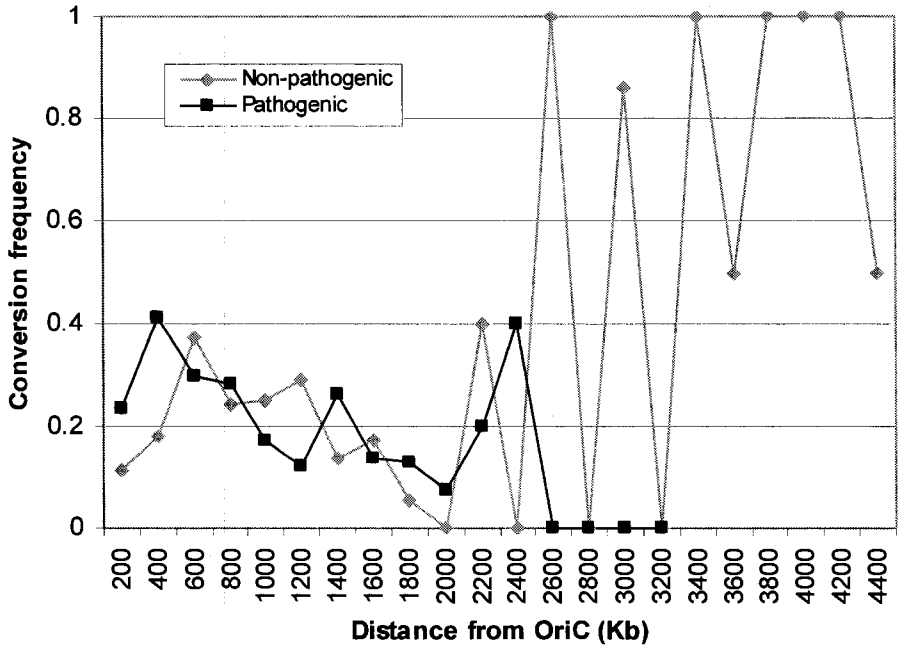
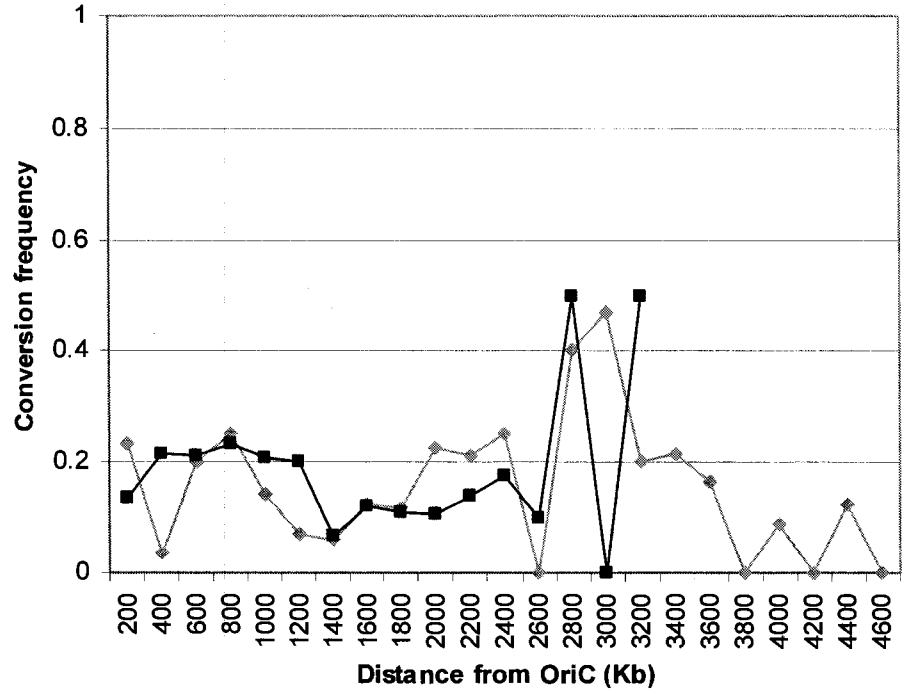

1600
1200
800
400
0
Number of genes
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Conversion frequency
400
800
1200
1600
2000
2400
2800
3200
3600
4000
4400
4800
Distance (Kb)
Family
Conversions
Frequency

Figure 4.

A


Non-pathogenic
Pathogenic
Conversion frequency
0
0.2
0.4
0.6
0.8
1
200
400
600
800
1000
1200
1400
1600
1800
2000
2200
2400
2600
2800
3000
3200
3400
3600
3800
4000
4200
4400
Distance from OriC (Kb)

B


Conversion frequency
0
0.2
0.4
0.6
0.8
1
200
400
600
800
1000
1200
1400
1600
1800
2000
2200
2400
2600
2800
3000
3200
3400
3600
3800
4000
4200
4400
4600
Distance from OriC (Kb)

C

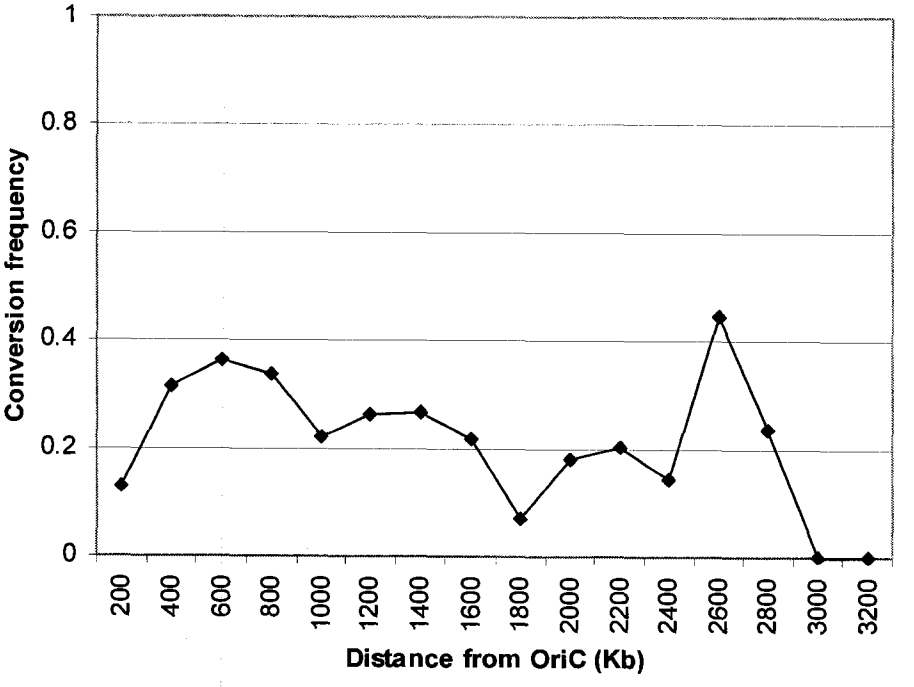
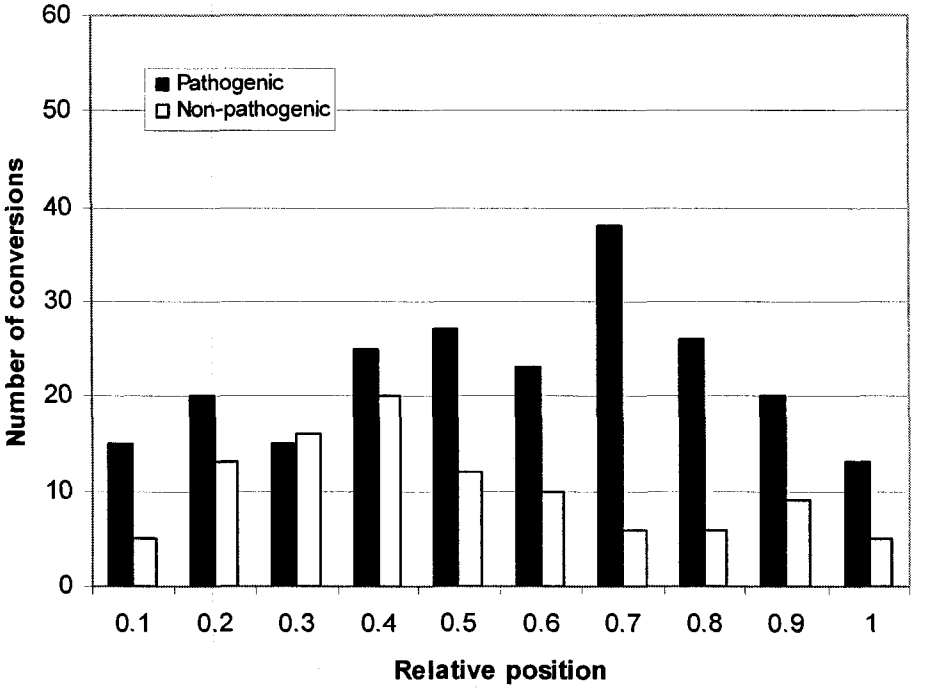
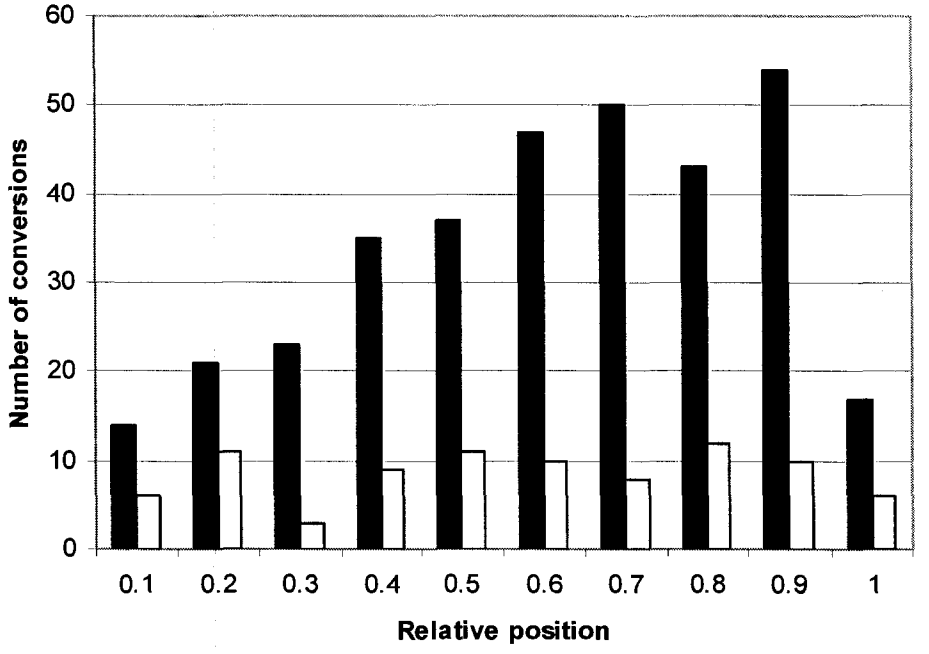

1
0.8
0.6
0.4
0.2
0
Conversion frequency
200
400
600
800
1000
1200
1400
1600
1800
2000
2200
2400
2600
2800
3000
3200
Distance from OriC (Kb)

Figure 5.

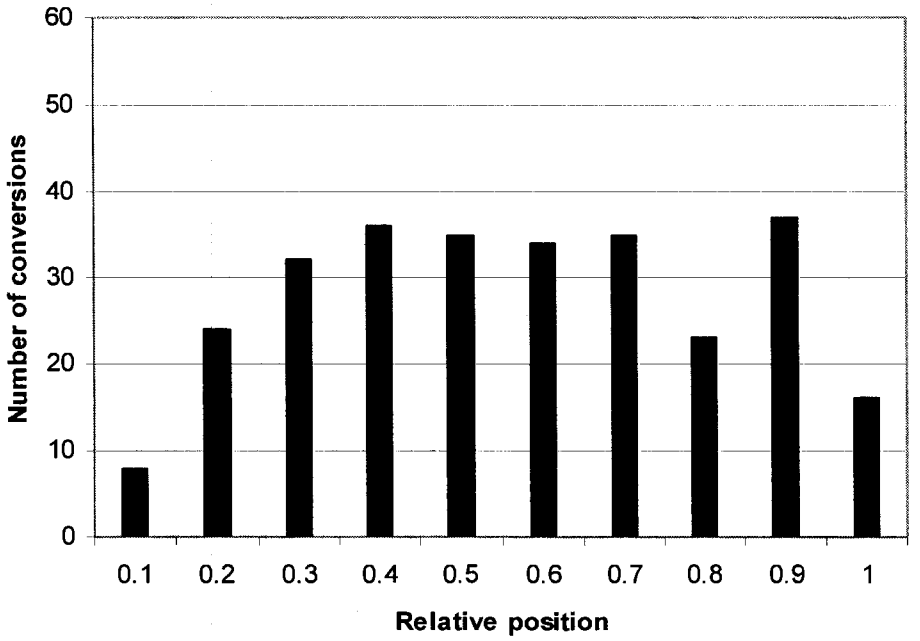
A


Pathogenic
Non-pathogenic
Number of conversions
60
50
40
30
20
10
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Relative position

B


Number of conversions
60
50
40
30
20
10
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Relative position

C


60
50
40
30
20
10
0
Number of conversions
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Relative position

Chapter 5

Ectopic Gene Conversions in the Genome of Ten Hemiascomycete Yeast Species

Robert T. Morris and Guy Drouin

Département de biologie et Centre de recherche avancée en génomique environnementale, Université d'Ottawa, Ottawa, Ontario, Canada, K1N 6N5

Running head: Ectopic conversions in ten yeast species

Keywords: ectopic, gene conversion, Hemiascomyctes, yeast, backbone, genome

Corresponding author: Guy Drouin, Département de biologie, Université d'Ottawa, 30 Marie Curie, Ottawa, Ontario, Canada, K1N 6N5. Tel.: (613) 562-5800 ext. 6052, FAX: (613) 562-5486, E-mail: gdrouin@science.uottawa.ca

Abstract

I characterized ectopic gene conversions in ten hemiascomycete yeast species. Of the ten species, seven diverged after a whole genome duplication event (WGD) and three diverged prior to the genome duplication. I analyzed gene conversions from three separate datasets: ohnologous gene pairs of seven post-WGD species, paralogous gene families from seven post-WGD species, and paralogous gene families from three pre-WGD species. I found that evolutionarily conserved (ohnologs) genes tend to be less frequently ectopically converted than species specific paralogs. Ectopic gene conversions occur more frequently between chromosomally linked genes and between genes which are situated close together on a chromosome. Converted regions are not uniformly distributed along the length of genes. Gene conversions occur more frequently near the 3' end of converted genes.

Introduction

The repair of double strand DNA breaks (DSB) is a critical biological process which maintains genome stability. The primary process whereby DSB are repaired is via homologous recombination (HR); this process requires the use of a repair template gene which provides a copy of the missing information caused by the DSB. The repair template can either be an allele (allelic recombination) or a paralog (ectopic recombination). An end product of the HR pathway is the replacement of the broken part of the damaged gene by a homologous portion of the repair template gene; therefore the damaged gene has been converted by the template gene (reviewed in Alyon and Kupiec 2004).

The factors affecting, and the characteristics of, ectopic and allelic gene conversions have been the focus of many studies. It has been found that sequence similarity has a profound effect on gene conversion propensity between paralogs. In *E. coli* a 2 - 4% decrease in sequence similarity between a damaged gene and its repair template can cause a 10 to 40 fold decrease in recombination frequency (Watt et al. 1985, Shen and Huang 1986). Chromosomally linked genes are converted more frequently than dispersed genes in Drosophila and humans (Engels et al. 1994, Benovoy and Drouin, unpublished). In *S. cerevisiae* it has been found that increasing nucleotide distance between chromosomally linked paralogs tends to decrease their conversion frequency (Goldman and Lichten 1996, Achaz et al. 2000, Drouin 2002). In some genomes, different regions of genes are converted at different rates. For example, in *S. cerevisiae*, genes conversions between dispersed paralogs are more frequent at their 3' ends (Drouin

2002). This bias is likely a product of using cDNA intermediates as repair templates (Drouin 2002; Derr and Strathern 1993).

I recently characterized ectopic gene conversions in the conserved backbone genes in four *E. coli* strains (K12, EDL933, Sakai and CFT073; Morris and Drouin, unpublished results). The fact that these functional genes are maintained throughout the evolution of these bacteria suggests they are important for survival. Our results show that conserved backbone genes in *E. coli K12* tend to be converted less frequently than the K12 species specific genes. This lower conversion frequency is likely the result that conserved genes are under more selective pressure than the species specific genes (Morris and Drouin, unpublished results). Similar levels of purifying selection were found in each genome's backbone genes.

The recent sequencing of ten hemiascomycete genomes provides the opportunity to study ectopic gene conversions within a clade with as much sequence divergence as the entire Chordate phylum (Dujon et al. 2004). The evolution of several hemiascomycetes species was affected by a whole genome duplication event (WGD) occurring approximately 150 MYA (Wolfe and Shields 1997; Dietrich et al. 2004; and Kellis et al. 2004; Herrero 2005; Scannell et al. 2006). The *Saccharomyces cerevisiae*, *S. paradoxus*, *S. mikatae*, *S. bayanus*, *S. kudriavzevii*, *S. castellii* and *Candida glabrata* genomes diverged from a common ancestor after the duplication event (post-WGD species; Goffeau et al. 1996; Kellis et al. 2003; Cliften et al. 2003). The genomes *Kluyveromyces lactis*, *Debaryomyces hansenii* and *Yarrowia lipolytica* all diverged from the last common ancestor with *S. cerevisiae* before the duplication event (pre-WGD species; Dujon et al. 2004).

The advantage of separating these genomes into two groups is that I am able to perform two comparisons. The first compares the characteristics of ectopically converted ohnologs and paralogs between the post-WGD species. The post-WGD ohnologs are composed of the duplicated gene pairs that are still found in the genome of post-WGD species (i.e., they are ohnologous gene pairs). These duplicated blocks of genes are the remnants of the yeast genome duplication of the post-WGD common ancestor (Wolfe and Shields 1997, Wolfe 2001). The post-WGD paralogs data set is composed of the genes from multigene families containing at least three members in the genome of the 7 post-WGD species but excluding any ohnologous genes that are grouped into a paralogous gene family. The second comparison involves the contrast of the characteristics of ectopically converted paralog genes between pre- and post-WGD species. The pre-WGD paralogs data set is composed of the genes from multigene families containing at least three members in the genome of the 3 pre-WGD species.

Evolutionarily conserved genes like ohnologs are likely to be important for the survival of a group of organisms. I expect that most ectopic gene conversions between ohnolog genes will be deleterious and therefore removed by selection. I predict that ectopic gene conversions between ohnologs will be less frequent and have more stringent sequence similarity requirements than for the paralog genes. In addition, gene conversion frequency should decrease as the nucleotide distance between related genes increases and be less frequent between dispersed genes. High sequence similarity between converted genes will permit long conversion tracts. Furthermore, I expect to find that converted regions within genes will be clustered near the 3' end of genes in each genome because a

previous report of *S. cerevisiae* detected a 3' bias of converted regions within converted genes (Drouin 2002).

Many of the results presented here are consistent with previous work done on yeast. For instance, the positive correlation between sequence similarity on gene conversion tract lengths, the fact that dispersed related genes are converted less frequently than closely linked genes and the observation that the gene conversions are clustered near the 3' end of genes have all been found previously in *S. cerevisiae*. However the lack of data resulting in low statistical power has prevented the universal detection of some other characteristic relationships such as the negative correlation between nucleotide distance and gene conversion frequency.

Materials and Methods

Genome sequences:

The *S. cerevisiae*, *S. paradoxus*, *S. mikatae*, *S. bayanus*, *S. kudriavzevii*, and *S. castellii* genome sequences were retrieved from the Saccharomyces Genome Database (SGD; ftp://genome-ftp.stanford.edu/pub/yeast/sequence/). The *C. glabrata*, *K. lactis*, *D. hansenii*, and *Y. lipolytica* genome sequences and distance files (*.ptt files) were retrieved from the NCBI ftp website (ftp://ftp.ncbi.nih.gov/).

Gene Family Data Sets

I used three different data sets of protein coding genes. To retrieve the post-WGD ohnologs from the 7 post-WGD species, I used the 551 *S. cerevisiae* duplicated gene pairs (1102 ohnologs) identified by Byrne and Wolfe (2005) as queries. Sequences from *C. glabrata* and *S. castellii* were retrieved using the Yeast Gene Order Browser (http://wolfe.gen.tcd.ie/ygob/) and those from the other 4 species were retrieved from the Saccharomyces Genome Database (ftp://genome-ftp.stanford.edu/pub/yeast/sequence/fungal_genomes/Multiple_species_align/other/fungalAlignCorrespondance.txt).

The post-WGD paralog data set was constructed using the BLASTCLUST program available at the NCBI FTP site. Gene families were defined as being composed of sequences having at least 60% amino acid identity over at least 50% of their length. If genes previously identified as ohnologs were grouped into paralogous multigene families,

then these genes were removed from the family to ensure that there are no redundancies between the ohnolog and paralog dataset (see Figure 1).

The pre-WGD paralog data set was also constructed using the BLASTCLUST program and gene families were also defined as being composed of sequences having at least 60% amino acid identity over at least 50% of their length.

Sequence Alignments and Gene Conversion Detection:

ClustalW was used to align the protein sequences of multigene families' members (Thompson et al. 1994). DNA sequences were then fitted to the protein alignments using the align2aa PERL script (http://sunflower.bio.indiana.edu/~wfischer/Perl_Scripts/).

Gene conversions were detected using the GENECONV method (Sawyer 1999). Redundant gene conversions within a multigene family were detected by examining the phylogenetic tree of each family and removed from the analysis (Drouin 2002). If the same gene conversion was detected at the same location in the multigene family alignment in closely related decendents of a common ancestor then the most parsimonous explanation is that the conversion event occurred within the common ancestor, therefore only one pf the conversions detected in the set of decendents are retained for further analysis. To control for false positives, gene conversions between sequences having less than 80% maximum flanking similarity were removed from the analysis (Posada and Crandall 2001).

Gene Conversion characteristics:

The gene conversion frequency for each species was calculated using two different methods. The first method calculates the conversion frequency as the ratio of the number of species specific conversions divided by the total number of gene comparisons between multigene family members. The second method calculates the frequency as the ratio of the number of gene conversions divided by the total number of multigene family members. Intra and inter-chromosomal gene conversion frequencies were calculated for the *S. cerevisiae*, *C. glabrata* ohnolog and paralog multigene families. In addition intra and inter chromosomal conversion frequencies were calculated for the paralog multigene families of *K. lactis*, *D. hansenii* and *Y. lipolytica* genomes. These frequencies are calculated as the ratio of intra (or inter) chromosomal conversions divided by the total number of intra (or inter) chromosomal gene comparisons. The gene conversion length was obtained from the GENECONV output. The maximum similarity for the flanking 100 nucleotides was calculated for each converted gene pair using an in-house PERL script. The locations of the converted regions were calculated as the relative position of the centre of each conversion with respect to the length of the converted genes. The distance between converted genes was calculated only for conversions detected within *S. cerevisiae*, *C. glabrata*, *K. lactis, D. hansenii*, and *Y. lipolytica* because position data for the other five species was not available. Data tabulation and analysis was done using Microsoft Excel (Microsoft, Redmond, WA, USA) and S-plus v 7.0 (Insightful, Seattle WA USA). The G-Power program was used to calculate the power of the ANOVA tests which were used to compare gene conversion characteristic result distributions like gene conversion frequency between groups of paralog or ohnolog genes in pre or post WGD

genomes (Erdfelder et al. 1996). Power calculations for correlation tests were done using an online application (http://calculators.stat.ucla.edu/powercalc/correlation/) and SAS 9.1.3 (SAS Institute Inc., Cary NC USA).

Results

Ohnolog and Paralog Multigene families:

The collections of ohnolog and paralog multigene families were examined to quantify differences between the number and size of families of the yeast genomes (Table 1). *S. kudriazevii* and *C. glabrata* have fewer ohnolog families than other post-WGD genomes. Similarly, with respect to the paralog families, *S. kudriavzevii* and *C. glabrata* have the smallest number of families, of the post-WGD species. *C. glabrata* paralog families tend to be smaller than gene families in the remaining genomes. The small amount of gene redundancy in *C. glabrata* is likely the result of reductive evolution in this genome (Dujon et al. 2004). The *K. lactis*, *D. hansenii* and *Y. lipolytica* genomes seem to have a similar range of numbers of paralog multigene families as the post-WGD species (Table 1).

Given that I observed differences between the numbers of multigene families found within pre- and post-WGD species, I applied statistical tests to determine whether these differences were significant. Pairwise chi-square tests using a Bonferroni correction were used to determine the significance of the observed differences in multigene family number between yeast genomes. Most of the comparisons indicate that the post-WGD species have similar numbers of paralogous multigene families. However I found that *S. bayanus* has significantly more multigene families than *S. kudriavzevii*, *C. glabrata*, and *S. castellii* (χ^2 p = $1.52\text{x}10^{-05}$, p = 0.003, and p = 0.0001, respectively). In addition, *S. paradoxus* has more multigene families than *C. glabrata* and *S. castellii* (χ^2 p = $1.95\text{x}10^{-05}$ and p = 0.0004 respectively). Using a chi-square test I found that the numbers of multigene families within the post-WGD species *D. hansenii* and *Y. lipolytica* are

identical. However, both have significantly more paralogous families than *K. lactis* (χ2 p = 0.003 and p = $1.95x10^{-05}$ respectively). Finally, the average distributions of pre- and post-WGD paralogous families are identical (χ^2, p = 0.66, Figure 2).

Organization of gene families

The organization of the multigene families is measured as the proportion of multigene family members located on the same chromosome. A high proportion of contiguous family members suggest that these gene families are likely a result of tandem duplication events. In addition, the large percentage of contiguous family members could reflect high stability of genomic regions, suggesting that frequent chromosomal rearrangements in these genomes may be detrimental. I find that paralogous genes of the pre- and post-WGD genomes tend to be located on the same chromosome more frequently whereas ohnologous genes are more likely to be dispersed (Table 2). Given the fact that the ohnologs are remnants of an ancient genome duplication and subsequent diploidization I expect to find fewer ohnologs than paralogs located on the same chromosome. The paralog families are relatively young compared to the ohnolog (Ohnolog Ks > Paralog Ks). Therefore fewer genome rearrangements would have dispersed these genes throughout the genome as opposed to the ohnologous genes. The inspection of the percentage of paralogous multigene family members found on the same chromosomes between pre and post-WGD genomes suggests that they have similar organizations (Table 2).

Gene conversion number and frequency

I looked at the differences in conversion frequency between genes located on the same chromosomes and between genes dispersed in the genome. I found that intra-chromosomal gene conversions tend to occur more frequently than inter-chromosomal conversions. Genes located on the same chromosome are converted 2 – 10 times more frequently than inter-chromosomal genes in *S. cerevisiae* and *C. glabrata* paralogous families (Table 3). Similarily in ohnolog families, intra-chromosomal genes are converted 4 times more frequently in *S. cerevisiae* (Table 3). By contrast, there is an absence of intra-chromsomally linked gene conversions within the *C. glabrata* ohnolog families (Table 3). As found in the post-WGD, intra-chromosomal genes are converted more frequently (approximately 3 times more frequently) than inter-chromosomal genes in the *D. hansenii* paralog families (Table 3). In contrast, no difference was observed between intra and inter-chromosomal conversion frequencies within the pre-WGD paralog gene families of *K. lactis* and *Y. lipolytica* (Table 3).

On average, more gene conversions are detected within the paralogous multigene familes of the pre- and post-WGD genomes than in the ohnologous families of the post-WGD genomes (Table 4). The mean number (± S.D.) of conversions detected within the paralog gene families of the pre- and post-WGD genomes are 38 ± 33 and 30 ± 16 respectively. By contrast, only 7 ± 5 conversions are found in the post-WGD ohnolog families.

Except for the *S. paradoxus* and *S. mikatae* genomes, gene conversions are more frequent in the post-WGD paralog families than in the post-WGD ohnolog families

(Table 4). On average, the paralogous gene conversion frequency is larger than the ohnologous gene conversions frequqency. The mean difference (± S.D.) between these two frequencies is 3.65 ± 3.56% (Table 4). The previous result was based on the conversion frequency calculated relative to the number of gene comparisons, by looking at the other frequency values calculated (i.e., relative to the total number of multigene family members; see Table 4) I found that the mean conversion frequency for paralogs (19.03 ± 16.29%) is statistically larger than for ohnologs (0.74 ± 0.46; Wilcoxon two sample test, $p = 0.0006$).

Ectopic gene conversions are equally frequent in both pre- and post- WGD genomes. Median gene conversion frequencies relative to both total number of comparisons and number of multigene family members are equal between pre- (12.09%, 21.3%) and post-WGD (5.06%, 19.03%) paralogous gene families (Table 4; Wilcoxon two sample test, $p = 0.26$ with respect to gene comparisons and $p = 0.83$ with respect to the number of multigene family members).

In addition to the effect of chromosomal position of homologous genes, the physical nucleotide distance between related genes on the same chromosome influences the frequency at which they are converted. Gene conversions were more frequent between contiguous genes in the paralogous genes of S. *cerevisiae* ($r = -0.54$), *C. glabrata* ($r = -0.74$) and *D. hansenii* ($r = -0.45$). Each correlation was significant at $\alpha = 5\%$ (Spearman rank correlation test, $p = 0.008$, 0.048, and 0.008 for *S. cerevisiae*, *C. glabrata* and *D. hansenii*, respectively). Non-significant correlations using the Spearman rank correlation test were found in the paralogous genes of *K. lactis* ($r = -0.14$ and $p = 0.28$) and *Y. lipolytica* ($r = -0.53$ and $p = 0.28$). Significant correlations were not found in

pre-WGD *K. lactis* and *Y. lipolytica* paralogous familes because of the small amount of data available. An analysis of power was not possible for *K.lactis* and *Y. lipolytica* due to the lack of data (the power test is confounded by small data sets, as found in the two aforementioned genomes which contained fewer than 4 data points each).

A significant correlation was not found in the ohnologs of *S. cerevisiae* because only two conversions between chromosomally contiguous genes were identified and no conversions between ohnolog genes located on the same chromosome were found in *C. glabrata.*

Gene conversion length and flanking similarity

Median lengths of post-WGD ohnolog gene conversions are identical between each genome (Table 5). Multiple ANOVA tests of the gene conversion length distributions of the seven post-WGD genomes indicate that there is no statistical difference between them (ANOVA, $p = 0.84$, $\alpha = 0.05$).

A comparison of the median lengths of post-WGD paralog gene conversions indicates that *S. cerevisiae* conversion lengths are larger than any of the other six genomes (Table 5). Multiple ANOVA tests of gene conversion lengths indicate that the *S. cerevisiae* median conversion lengths are significantly longer than any other post-WGD genome (ANOVA, $p = 9.0 \times 10^{-09}$, $\alpha = 0.05$).

The median lengths of paralog conversions between pre-WGD genomes are equal. Multiple comparison ANOVA result indicates that median gene conversion lengths of all pre-WGD genomes are not significantly different (ANOVA, $p = 0.34$, $\alpha = 0.05$).

The median length of Post-WGD paralog and ohnolog conversions are not significantly different. Pair-wise Wilcoxon rank tests between backbone and species specific gene conversion median lengths found no difference between each type of conversion (Table 5).

Gene conversion tract lengths for pre- and post-WGD paralogous genes are not significantly different from each other. Comparison of the median gene conversion length between pre- and post-duplication genomes indicates that they are not statistically different (median gene conversion length, pooled over all post-duplication genomes, is 167 bp and the median gene conversion length, pooled over all pre-duplication genomes, is 150 bp; Wilcoxon two sample test $p = 0.09$, $\alpha = 0.05$).

Median flanking similarity of ohnologous genes between post-WGD genomes are not significantly different. Multiple ANOVA tests revealed that the median flanking similarities for each post-WGD genome are not significantly different (ANOVA, $p = 0.98$, $\alpha = 0.05$, Table 6).

Median flanking similarity of paralogous genes between post-WGD genomes are not identical. Multiple ANOVA tests indicates that median flanking similarity of *S. cerevisiae* is greater than the median similarity of *S. paradoxus* and *S. mikatae* (ANOVA $p = 0.02$, $\alpha = 0.05$; Table 6).

In *S. cerevisiae* the median flanking similarity for converted ohnologs is significantly less than for converted paralogs (Table 6). However, an analogous difference was not detected in any other post-WGD species due to the relatively low power of the statistical test (the power for each test was $\leq 55\%$; see Table 6).

The flanking similarity requirements paralog gene conversions in all three pre-WGD genomes are equal. Multiple ANOVA tests results indicate that the median flanking similarities of each genome are statistically equal (ANOVA, $p = 0.20$, $\alpha = 0.05$, Table 6).

Converted genes within pre – WGD paralogous genes have significantly less flanking similarity (pooled median 90.3%) than paralogous converted genes in post – WGD genomes (pooled median 93%; Wilcoxon two sample test $p = 0.018$, $\alpha = 0.05$, Table 6). It is unkown whether this difference has any biological significance.

Analysis of the relationship between the length of gene conversions and the flanking similarity indicates a significant positive correlation within the ohnolog gene families of the post-WGD genomes (Spearman rank correlation test $p = 0.0049$, $r = 0.44$ [pooled data for all seven genomes]; Figure 3A). The flanking similarity and gene conversion length between paralogous genes in post-WGD genomes are also correlated (Spearman rank correlation test $p = 0$, $r = 0.36$ [pooled data for all seven genomes]; Figure 3B).

A comparison of the correlations of conversion length and flanking similarity between ohnologous and paralogous genes for the post-WGD genomes indicates that similarity requirements between paralog genes are less stringent than for ohnolog genes. The correlation coefficients between conversion lengths and flanking similiarity for the ohnologous genes is slightly stronger than for the paralogous genes ($r = 0.44$ and $r = 0.36$, respectively).

Gene conversion lengths increase with flanking similarity in the pre-WGD paralogous gene families (Figure 3C, D, and E). Spearman rank correlation tests indicate

that the correlation between maximum flanking similarity and gene conversion length are significant in *D. hansenii* ($r = 0.40$, $p = 0.001$), and *Y. lipolytica* ($r = 0.17$, $p = 0.0052$). However, a significant correlation was not found in *K. lactis* ($r = 0.42$, $p = 0.15$), likely because of the small sample size of the data set resulting in low statistical power (SAS correlation power test, power = 0.34).

Ka, Ks, and Ka/Ks Ratios of ohnolog and paralog converted genes

Non-synonymous substitutions are more frequent in the paralogous genes than in the ohnologs of post-WGD genomes (Table 7). In contrast, synonymous substitutions (Ks) are more frequent in the ohnologs than in the paralogous genes (Table 7), suggesting that the ohnologs are ancient relative to the paralogs. Paralogous genes are under less negative selection than the ohnolog genes because the Ka/Ks ratios for the paralogs are consistently greater than those for the ohnologs (Table 7). The Ka, Ks, and Ka/Ks ratios of pre- and post-WGD converted genes are very similar (Table 7).

Location of converted regions and distance between converted genes.

Ectopic gene conversions are not biased toward the 3' end of genes within the seven post-WGD ohnolog gene family data sets (Table 8). Weak power results suggest that a bias may exist in these genomes but that our data are not sufficient to detect it (Table 8).

Similarly, gene conversions are not biased toward the 3' end of paralogous genes in each post-WGD genomes except for *S. cerevisiae*, where gene conversions are more frequent near the 3' end of genes (Table 8). The significant bias found in *S. cerevisiae*

supports previous work which indicated a similar bias (Drouin 2002). The small sample size resulted in low power of these correlation tests (power $\leq$ 7.6% for each non significant test). Furthermore, ectopic gene conversions are not biased toward the 3' end of the paralogous genes in the three pre-WGD genomes (Table 8). More data are required to determine whether a gene conversion location bias exists in these species because the power of the statistical test was also low (power $\leq$ 14%).

Discussion

On average, the genome duplication event in the post-WGD genome common ancestor did not significantly increase the number of paralogous multigene families (Figure 2, Table 1). Three of the genomes that have among the smallest complement of large multigene families (and fewest ohnolog pairs) were *S. kudriazevii*, *S. castelli* and *C. glabrata*. The small number and size of gene families in *C. glabrata* are likely the result of reductive evolution and gene loss through relatively high genome instability (Kellis et al. 2004, Fischer et al. 2006). It is unknown whether similar phenomena are happening in *S. kudriavzevii* and *S. castellii*. However, these factors could explain why these three genomes have significantly fewer multigene families than *S. bayanus*.

The percentage of contiguous genes on the same chromosome within paralog families of pre- and post-WGD families are similar, however the post-WGD ohnolog families tend to be more dispersed than the paralog families (Table 2). A likely explanation for this difference is that the paralog family members are created by tandem duplication which would predispose them to be more frequently located on the same chromosome. The ohnolog genes are more dispersed because it is likely they have been subjected to greater numbers of genome rearrangements than the relatively young paralog genes. Furthermore the fact that the ohnologous genes originate from a whole genome duplication event suggests that by their nature they should be dispersed onto different chromosomes.

The relative location of ohnologous and paralogous genes in the post-WGD species is an important factor influencing the frequency of gene conversion. Our results indicate that ohnologous and paralogous genes of *S. cerevisiae* and the paralogous genes

of *C. glabrata* and *D. hansenii* that are located together on the same chromosome tend to convert more frequently than dispersed genes located on different chromosomes (Table 3). Previous work on Drosophila and humans has also found that intra-chromosomal gene conversions are more frequent than inter-chromosomal (Engels et al. 1994, Benovoy and Drouin, unpublished). A possible explanation for the high number of intra-chromosomal conversions in *D. hansenii* is the fact that multiple tandem duplication events have been identified within this genome and therefore most paralogs are still located on the same chromosomes (Dujon et al. 2004). Some notable exceptions to this relationship between intra-chromosomal and inter-chromosomal conversions were observed. Unlike *S. cerevisiae* ohnologs, intra-chromosomal conversions between the ohnologs of *C. glabrata* were not detected. Since the proportion of ohnolog genes locaed on the same chromosome in of *S. cerevisiae* and *C. glabrata* is equal it is unlikely that multigene family organization explains this difference in the proportion of intra vs. inter-chromosomal conversions (Table 2). Recent work done by Fischer et al. (2006) found that the *C. glabrata* genome is more unstable than *S. cerevisiae*, and has a high propensity of genome reorganization and synteny loss. This could cause the number of intra-chromosomal gene conversions in *C. glabrata* to be underestimated if the detected conversion events occurred prior to the genome rearrangements. Similarly, in *K. lactis* and *Y. lipolytica* paralog family's intra- and inter-chromosomal conversions were equally frequent events (Table 3). This result may be attributed to the differences in multigene family organization between the *K. lactis* and *Y. lipolytica* and the other yeast genomes. Previous studies have documented that the highly redundant *Y. lipolytica* genome has undergone a high degree of map dispersion evident from the lack of numerous duplicated

blocks despite the fact that the genome contains a large number of paralogous genes. Therefore we expect to find fewer contiguous homologous genes, thereby reducing the probability of intra-chromosomal conversions (Dujon et al. 2004). *K. lactis* has a comparable percentage of paralogs located on the same chromosome as *Y.lipolytica*; therefore a dispersive effect may be evident in this genome as well. It is unlikely that these exceptions are due to mechanistic differences in the repair of DSB between pre and post-WGD species because the majority of repair genes have been maintained throughout the evolution of the hemiascomycetes (Richard et al. 2005).

In addition to chromosomal location of homologous genes, the nucleotide separation of related genes on the same chromosome has an effect on the frequency of gene conversion. A negative correlation between chromosomal distance between genes and conversion frequency of paralogous genes was detected in *S. cerevisiae*, *C. glabrata* and *D. hansenii*. Other studies of yeast support our finding that increased distance between related genes negatively affects their conversion frequency (Goldman and Lichten 1996, Drouin 2002). A lack of data in *K. lactis* and *Y. lipolytica* paralogous families and the ohnologous families of *S. cerevisiae* and *C. glabrata* prevented the detection of a similar relationship between chromosomal distance between homologous genes and conversion frequency. An explanation for the relationship between location and conversion frequency is that the repair mechanism preferentially searches for a suitable repair template close to the damaged gene first because a delay in DNA double strand break repair may have an adverse effect on cell growth. This explanation also suggests a reason for the small number of ohnologous conversions detected. The fact that few ohnologs are located close together suggests that it is unlikely the repair mechanism

will be able to use them as repair templates; therefore detection of ohnologs converting each other would be rare.

Sequence similarity requirements for ectopic conversions and the amount of negative selection are very similar between pre- and post-WGD paralogs. Several pieces of information support these conclusions. The fact that the frequency (Table 4), length (Table 5), and flanking sequence similarities (Table 6) of gene conversion of the paralog genes within pre- and post-WGD species are similar indicates that mechanistic similarities are present between these groups of genomes. Furthermore the fact that the pre- and post-WGD species share the majority of DNA repair genes (Richard et al. 2005) suggests they have common repair mechanisms. In addition, the fact that the mean Ka/Ks values for the paralog families of pre and post-WGD species are alike suggests that these genes are under similar selective pressures and that these genes have similar gene conversion constraints. This suggests that despite the different ecological niches of the yeast species, these paralogs evolve in similar ways.

Conversely, I find that post-WGD ohnologs tend to have more stringent sequence similarity requirements and are under more negative selection than paralog genes. Converted ohnologs tend to have a stronger correlation between flanking similarity and conversion length than converted paralogs (Figure 3). In addition, I found that ectopic gene conversions occur more frequently between paralogs than ohnologs, which suggests differences in selective constraints. The fact that Ka/Ks ratios of ohnologs tend to be smaller than those in paralogs genes suggests that there is weaker negative selection acting on paralogs than on ohnologs (Table 7). Ohnologs have larger Ks values than paralogous genes which reflect their ancient origin (Table 7). Despite the fact that the

ohnologs are ancient, almost all Ka values are smaller for ohnologs than paralogs (Table 7). This indicates that the paralogous genes have been able to accumulate more amino acid changes in a relatively short time compared to the ohnologous genes. This may be a result of weak negative selection on the paralogous genes. A possible alternative explanation is that short regions of the the paralogous genes where under positive selection despite the fact that the omega ratios for the entire gene was less than 1 (Table 7). Taken together these results indicate that the gene conversion similarity requirements are less stringent for paralogs than for ohnologs. A similar result was found in *E. coli K12*; the number of gene conversions in the species specific regions was greater than in the conserved backbone regions of the genome (Morris and Drouin, unpublished). The fact that these functional ohnologs have been conserved for over 150 million years indicates that they are vital for the survival of each yeast species. The ectopic conversion of any of these genes would introduce sequences differences that could adversely affect their function, and by extension the survival of the organism. This suggestion is supported by our results which indicate that stringent sequence similarity requirements and greater negative selection minimize the number of ectopic conversions involving ohnologs. In addition, the fact that the non-synonymous substitutions are more frequent in the paralogs genes than in the ohnolog genes suggests that amino acid changes are not as detrimental in the paralogous genes.

One of the obvious effects of repeated gene conversion due to less negative selective pressure on paralogous genes is that the sequence similarity between related genes will increase. As support, I found that *S. cerevisiae*, the ohnolog genes tend to have less flanking sequence similarity than paralog genes (Table 6). In addition, I found that

synonymous substitutions are less frequent in the paralog genes than in ohnologs reflecting the homogenization of the genes by numerous ectopic conversions. These observations are likely due to increased ectopic gene conversions between the multigene family paralogs. The small sample sizes of ohnolog conversions in post-WGD genomes may prevent the detection of significant differences as found in *S. cerevisiae*.

Previous studies on *S. cerevisiae* have found that gene conversions are biased toward the 3' end of converted genes which has been attributed to ectopic gene conversion via cDNA intermediates (Drouin 2002). Our results support the previous work by indicating that conversions are biased toward the 3' end of genes within the *S. cerevisiae* paralog dataset (Table 8). The fact that no significant bias was detected within any other species is likely a result of the small amount of data and resulted in low statistical power (Table 8).

In conclusion this study has provided a comparison of ectopic gene conversion characteristics both between ohnologs and paralogs genes within post –WGD species and between paralogs of pre- and post-WGD yeast species. In general, our results indicate that the characteristics of ecoptic gene conversion between paralogs of pre and post-WGD species are similar. The only differences were observed between the ohnolog and paralog gene conversions within genomes. These are likely the result of differences in negative selection between the evolutionarily conserved ohnologs and the relatively recent paralogs. It is obvious from our work here that more data are required to conclusively prove several trends like a 3' conversion bias in all pre- and post-WGD species and the fact that intra- chromosomal conversions are more frequent than inter-chromosomal conversions. In addition, other studies have found that expression levels of

duplicated genes influence the rate of sequence divergence (Pyne et al. 2005). It would be interesting to determine whether the increased ectopic gene conversion frequency in *C. glabrata*, *D. hansenii*, and *K .lactis* is due to conversions between highly expressed genes.

Literature Cited

Achaz G, Coissac E, Viari A, Netter P (2000) Analysis of intrachromosomal duplications in yeast Saccharomyces cerevisiae: a possible model for their origin. Mol Biol Evol 17:1268-1275

Aylon Y, Kupiec M (2004) DSB repair: the yeast paradigm. DNA Repair 3:797-815.

Byrne KP, Wolfe KH (2005) The Yeast Gene Order Browser: combining curated homology and syntenic context reveals gene fate in polyploid species. Genome Res 15:1456-1461

Cliften P, Sudarsanam P, Desikan A, Fulton L, Fulton B, Majors J, Waterston R, Cohen BA, Johnston M (2003) Finding functional features in Saccharomyces genomes by phylogenetic footprinting. Science 301:71-76

Derr LK, Strathern JN (1993) A role of reverse transcripts in gene conversion. Nature 361:170-173

Dietrich FS, Voegeli S, Brachat S, Lerch A, Gates K, Steiner S, Mohr C, Pohlmann R, Luedi P, Choi S, Wing RA, Flavier A, Gaffney TD, Philippsen P (2004) The *Ashbya gossypii* genome as a tool for mapping the ancient *Saccharomyces cerevisiae* genome. Science 204:304-307

Drouin G (2002) Characterization of the gene conversions between the multigene family members of the yeast genome. J Mol Evol 55:14-23

Dujon B, Sherman D, Fischer G, Durrens P, Casaregola S, Lafontaine I, De Montigny J, Marck C, Neuveglise C, Talla E, Goffard N, Frangeul L, Aigle M, Anthouard V, Babour A, Barbe V, Barnay S, Blanchin S, Beckerich JM, Beyne E, Bleykasten C, Boisrame A, Boyer J, Cattolico L, Confanioleri F, De Daruvar A, Despons L,

Fabre E, Fairhead C, Ferry-Dumazet H, Groppi A, Hantraye F, Hennequin C, Jauniaux N, Joyet P, Kachouri R, Kerrest A, Koszul R, Lemaire M, Lesur I, Ma L, Muller H, Nicaud JM, Nikolski M, Oztas S, Ozier-Kalogeropoulos O, Pellenz S, Potier S, Richard GF, Straub ML, Suleau A, Swennen D, Tekaia F, Wesolowski-Louvel M, Westhof E, Wirth B, Zeniou-Meyer M, Zivanovic I, Bolotin-Fukuhara M, Thierry A, Bouchier C, Caudron B, Scarpelli C, Gaillardin C, Weissenbach J, Wincker P, Souciet JL (2004) Genome evolution in yeasts. Nature 430:35-44

Erdfelder E, Faul F, Buchner A (1996) GPOWER: A general power analysis program. Behav Res Methods Instrum Comput 28:1-11

Engels WR, Preston CR, Johnson-Schlitz DM (1994) Long-range cis preference in DNA homology search over the length of a Drosophila chromosome. Science 263:1623-1625

Fischer G, Rocha E, Brunet F, Vergassola M, Dujon B (2006) Highly variable rates of genome rearrangements between hemiascomycetous yeast lineages. PloS Genetics 2:253-261

Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M, Louis EJ, Mewes HW, Murakami Y, Philippsen P, Tettelin H, Oliver SG (1996) Life with 6000 genes. Science 274:563-567

Goldman AS, Lichten M (1996) The efficiency of meiotic recombination between dispersed sequences in Saccharomyces cerevisiae depends upon their chromosomal location. Genetics 144:43-55

Herrero E (2005) Evolutionary relationship between between *Saccharomyces cerevisiae*

and other fungal species as determined from genome comparisons. Rev Iberoam Micol 22:217-222

Kellis M, Birren BW, Lander ES (2004) Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. Nature 428:617-624

Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. Nature 423:241-254

Posada D, Crandall KA (2001) Evaluation of methods for detecting recombination from DNA sequences: Computer simulations. Proc Natl Acad Sci USA 98:13757-13762

Pyne S, Skiena S, Futcher B (2005) Copy correction and concerted evolution in the conservation of yeast genes. Genetics 170:1501-1513

Richard GF, Kerrest A, Lafontaine I, Dujon B (2005) Comparative genomics of hemiascomycete yeasts: genes involved in DNA replication, repair, and recombination. Mol Biol Evol 22:1011-1023

Sawyer S (1999) GENECONV molecular biology computer program. Available online from http://www.math.wustl.edu/~sawyer/geneconv/

Scannell DR, Byrne KP, Gordon JL, Wong S, Wolfe KH (2006) Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts. Nature 440:341-345

Shen P, Huang HV (1986) Homologous recombination in *Escherichia coli*: dependence on substrate length and homology. Genetics 112:441-457

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22:4673-4680

Watt VM, Ingles CJ, Urdea MS, Rutter WJ (1985) Homology requirements for recombination in *Escherichia coli*. Proc Natl Acad Sci USA 82:4768-4772

Wolfe KH (2001) Yesterday's polyploids and the mystery of diploidization. Nat Rev Genet 2:333-341

Wolfe KH, Shields DC (1997) Molecular evidence for an acient duplication of the entire yeast genome. Nature 387:708–713

Table 1: List of the number of ohnolog and paralog multigene families in the pre- and post- whole genome duplication (WGD) genomes. The range of multigene family size is provided in brackets. Ohnolog families were not identified in the pre-WGD genomes (NA)

Genome	Number of Ohnolog Families	Number of Paralog Families
Post WGD		
S. cerevisiae	551 (2)	30 (3-40)
S. paradoxus	436 (2)	80 (3-68)
S. mikatae	412 (2)	86 (3-37)
S. kudriavzevii	226 (2)	13 (3-20)
S. bayanus	462 (2)	75 (3-23)
C. glabrata	300 (2)	16 (3-7)
S. castellii	398 (2)	17 (3-10)
Pre WGD		
K. lactis	NA	15 (3-9)
D. hansenii	NA	43 (3-9)
Y. lipolytica	NA	60 (3-26)

Table 2: Percentage of gene comparisons which are between multigene family members located on the same chromosome. The percentage of coupled multigene family members is used as a measure of the dispersion of ohnolog and paralog gene families. Ratio in brackets indicates the number of gene comparisons between linked genes divided by the total number of gene comparisons between multigene family members within each genome. NA entries indicate genomes in which ohnologs were not identified.

Genome	Ohnologs	Paralogs
Post-WGD		
S. cerevisiae	4.0% (22/551)	8.4% (163/1930)
C. glabrata	5.0% (15/300)	38.6% (29/75)
Pre-WGD		
K. lactis	NA	21% (26/124)
D. hansenii	NA	31% (86/270)
Y.lipolytica	NA	18% (158/884)

Table 3: Intra and inter chromosomal gene conversion frequencies for pre- and post-WGD genomes. Values in brackets indicate the ratio of the number of gene conversions divided by the number of gene comparisons. Data for S. paradoxus, S. mikatae, S. kudriavzevii, S. bayanus and S. castellii were not provided because position data were not available for the genes of these genomes. NA values indicate that backbone data was unavailable for the pre WGD genomes.

Genome	Ohnolog Families		Paralog Families	
Post WGD	Intrachromosomal Frequency	Interchromosomal Frequency	Intrachromosomal Frequency	Interchromosomal Frequency
S. cerevisiae	9.1% (2/22)	2.1% (11/529)	9.2% (15/163)	5.4% (95/1767)
C. glabrata	0% (0/15)	0.007% (2/285)	24.1% (7/29)	2.2% (1/46)
Pre WGD				
K. lactis	NA	NA	11.5% (3/26)	12.2% (12/98)
D. hansenii	NA	NA	36% (31/86)	11.4% (21/184)
Y. lipolytica	NA	NA	1.9% (3/158)	2.9% (21/726)

Table 4. Number and frequency of gene conversions in ohnologs and paralogs of ten fungi genomes. The gene conversion frequencies for each dataset (i.e., ohnolog and paralog) are calculated as two different ratios; the first is calculated relative to the total number of gene comparisons summed over every family and the second is calculated relative to the total number of multigene family members identified in all multigene families.

Genomes	Ohnolog Gene Conversions			Paralog Gene Conversions		
	Number	Frequency with respect to total # of comparisons (%)	Frequency with respect to total # of multigene family members (%)	Number	Frequency with respect to total # of comparisons (%)	Frequency with respect to total # of multigene family members (%)
S. cerevisiae	13	2.35	1.17	110	5.71	51.40
S. paradoxus	7	1.60	0.80	44	1.54	9.20
S. mikatae	6	1.45	0.73	26	1.50	4.80
S. kudriavzevii	2	0.88	0.44	20	7.96	29.80
S. bayanus	14	3.03	1.51	50	3.60	12.40
C. glabrata	2	0.67	0.33	8	10.67	14.80
S. castellii	2	0.50	0.25	8	5.06	10.80
K. lactis	-	-	-	15	12.09	23.80
D. hansenii	-	-	-	52	19.25	31.70
Y. lipolytica	-	-	-	24	2.71	8.20

Table 5. Gene conversion length statistics of pre- and post – whole genome duplication (WGD) species. Data for the post-WGD species are sub-divided into two categories, the backbone data relates to the conversions detected within the conserved backbone genes. The species specific category relates to the conversions detected between the genes outside the conserved backbone. The Wilcoxon two-sample test was used to detect differences between the conserved backbone and genome specific median gene conversion lengths.

Genome	Backbone gene conversion length (bp)					Species specific gene conversion length (bp)					Wilcoxon test
	Median	1st quartile	3rd quartile	Min	Max	Median	1st quartile	3rd quartile	Min	Max	p-value
Post –WGD											
S. cerevisiae	272	107	465	60	773	315	122.5	729.8	8	2642	0.52
S. paradoxus	235	102.5	335.5	50	531	106	52	223	14	1060	0.15
S. mikatae	165	101	376.3	68	568	112	75	346.5	8	535	0.42
S. kudriavzevii	270.5	208.3	332.8	146	395	270.5	138.5	82.3	192.3	11	0.26
S. bayanus	149.5	71.8	307.5	45	905	122	76	190	21	724	0.45
C. glabrata	83.5	55.3	111.8	27	140	130	68	296.8	15	817	0.40
S. castellii	144	131	157	118	170	226	75.3	456.3	44	862	0.71
Pre – WGD											
K. lactis	-	-	-	-	-	99	51	233	32	1127	-
D. hansenii	-	-	-	-	-	183	109	308	18	1309	-
Y. lipolytica	-	-	-	-	-	83	28	189	16	1770	-

Table 6. Statistics for maximum flanking similarity requirements of gene conversions in pre and post – whole genome duplication (WGD) species. The Wilcoxon two sample test was used to detected differences between the conserved backbone and genome specific median flanking similarity.

Genome	Backbone maximum flanking similarity (%)					Species specific maximum flanking similarity (%)					Wilcoxon test
	Median	1st quartile	3rd quartile	Min	Max	Median	1st quartile	3rd quartile	Min	Max	p-value
Post –WGD											
S. cerevisiae	88	84	94	80	97	94	89	99	80	100	0.0067
S. paradoxus	89	83.5	92.9	82	97	90.2	87	97	80	100	0.26
S. mikatae	87.5	83	91.3	81	96	91	86.3	95.8	80	100	0.27
S. kudriavzevii	86.8	86.3	87.4	85.7	88	93.5	89.7	99	80	100	0.23
S. bayanus	87.6	85	91.5	80	98	92.8	86	98	80	100	0.0507
C. glabrata	84.5	83.8	85.3	83	86	90.5	86.5	93.8	80	100	0.06
S. castellii	87	86.5	87.5	86	88	93	87.5	95.5	85	100	0.35
Pre – WGD											
K. lactis	-	-	-	-	-	90	86	95	81	98	-
D. hansenii	-	-	-	-	-	93	87	97	80	100	-
Y. lipolytica	-	-	-	-	-	86	84	93	80	100	-

Table 7: Rate of non-synonymous (Ka), synonymous substitutions (Ks) and estimate of selective pressure on paralogs within pre and post WGD species and on ohnologs within post WGD genomes. The (±) standard deviations are provided for each entry.

Genome	Ka		Ks		Ka/Ks	
Post-WGD	Ohnolog	Paralog	Ohnolog	Paralog	Ohnolog	Paralog
S. cerevisiae	0.04±0.03	0.09±0.08	0.96±0.49	0.37±0.44	0.04±0.02	0.38±0.27
S. paradoxus	0.09±0.11	0.18±0.20	0.91±0.76	0.56±0.40	0.10±0.05	0.46±0.57
S. mikatae	0.09±0.11	0.17±0.19	1.87±1.06	0.56±0.31	0.04±0.04	0.34±0.45
S. kudriavzevii	0.06±0.04	0.08±0.04	0.95±0.46	0.47±0.59	0.06±0.01	0.38±0.34
S. bayanus	0.11±0.09	0.13±0.12	1.91±1.68	0.40±0.46	0.07±0.05	0.40±0.28
C. glabrata	0.32±0.25	0.04±0.04	1.14±0.09	0.36±0.55	0.30±0.24	0.37±0.45
S. castellii	0.18±0.09	0.13±0.07	2.8±1.18	0.29±0.12	0.06±0.01	0.61±0.46
Pre WGD						
K. lactis	NA	0.20±0.26	NA	0.61±0.40	NA	0.49±0.58
D. hansenii	NA	0.10±0.07	NA	0.50±0.40	NA	0.31±0.17
Y. lipolytica	NA	0.25±0.19	NA	1.12±0.38	NA	0.46±0.38

Table 8: Correlation between location of the middle of converted regions and the number of gene conversions within pre- and post-WGD genomes. The R-values indicate correlation values; significant correlations are labeled with (*). The power of each correlation test is provided except for *S. cerevisiae* paralogs where the null hypothesis was rejected (power = NA). NA values in the ohnolog columns indicate that backbone data was unavailable for the pre - WGD genomes.

Genome	Ohnolog		Paralog	
Post WGD	R value	Power	R	Power
S. cerevisiae	-0.07	0.036	0.73 *	NA
S. paradoxus	0.12	0.049	-0.19	0.072
S. mikatae	2.8×10^{-17}	0.025	-0.19	0.076
S. kudriavzevii	-0.17	0.065	-0.09	0.043
S. bayanus	0.24	0.095	0.11	0.047
C. glabrata	-1.0×10^{-17}	0.025	0.06	0.034
S. castellii	0.17	0.066	-0.09	0.043
Pre WGD				
K. lactis	NA	NA	-0.32	0.14
D. hansenii	NA	NA	0.02	0.028
Y. lipolytica	NA	NA	0.14	0.055

* Spearman rank correlation test $p < 0.05$

Figure Legends

Figure 1. Schematic indicating how ohnologous and paralogous genes are separated within a multigene family. Gene A and A' represent ohnologous genes (i.e., duplicated genes created by a whole genome duplication), while the grey genes B and C are paralogous genes created by tandem duplications of gene A. If all four genes (A, A', B and C) are grouped together during the paralogous family definition then genes A and A' are removed from the family leaving only genes B and C.

Figure 2. The distribution of the average number of multigene families (mean + S. D.) within the seven post duplication genomes and three pre duplication genomes are shown. Five outlier families including two families of size 63 and 68 from *S. paradoxus*, two families of size 32 and 40 from *S. cerevisiae*, and a single family of 38 genes from *S. mikatae* are not shown on the figure to improve the visual clarity of the data.

Figure 3. Gene conversion length is positively correlated with maximum flanking similarity. Panel A.) Conversions detected between the ohnologs of six Saccharyomyces species and *C. glabrata*, 107 total conversions; 46 conversions have ≥80% flanking similarity, panel B.) Conversions detected between paralogs of six Saccharomyces species and *C. glabrata*, 401 total conversions; 311 conversions have ≥80% flanking similarity. Conversions within three pre-WGD genomes between multigene family paralogs; panel C) *K. lactis* has 17 gene conversions; 15 conversions have ≥80% flanking similariy, panel D) *D. hansenii* has 78 gene conversions; 52 conversions have ≥ 80%

flanking similarity, and panel E) *Y. lipolytica* has 52 gene conversions; 24 conversions have ≥80% flanking similarity.

Figure 1.

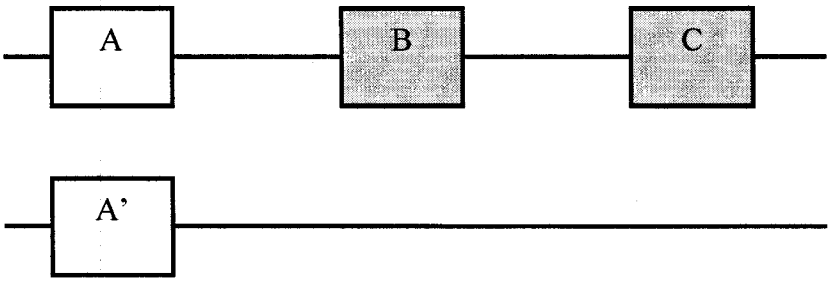


Figure 2.

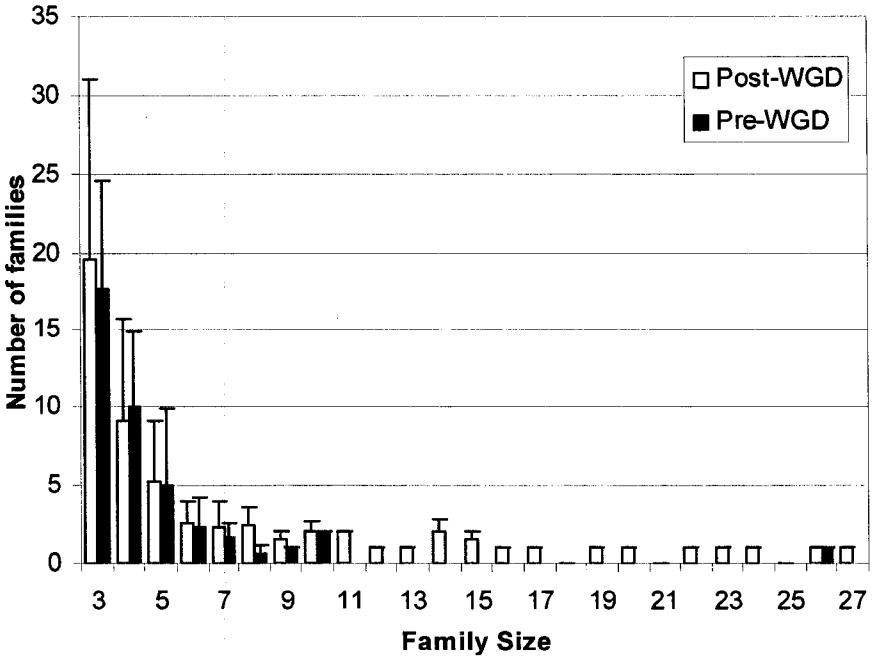
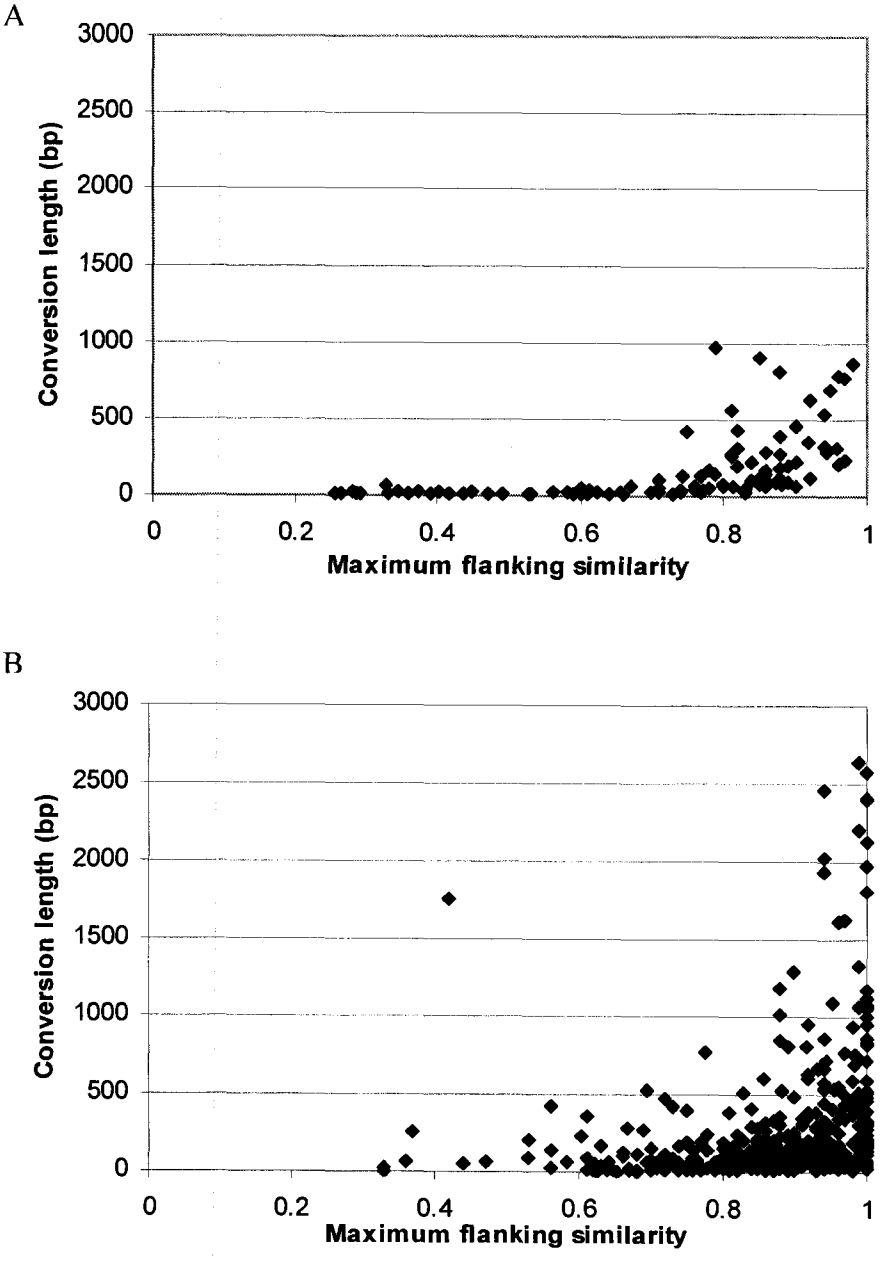
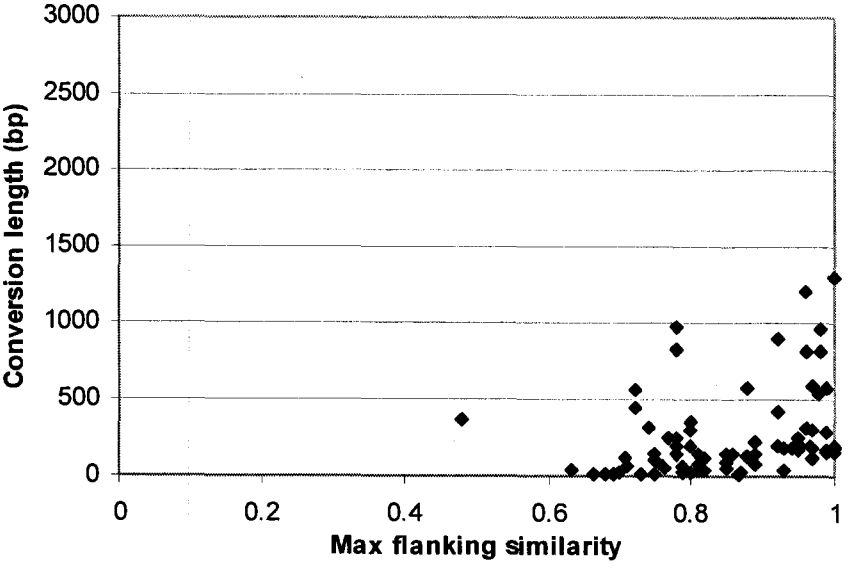

35
30
25
20
15
10
5
0
Number of families
Post-WGD
Pre-WGD
3
5
7
9
11
13
15
17
19
21
23
25
27
Family Size

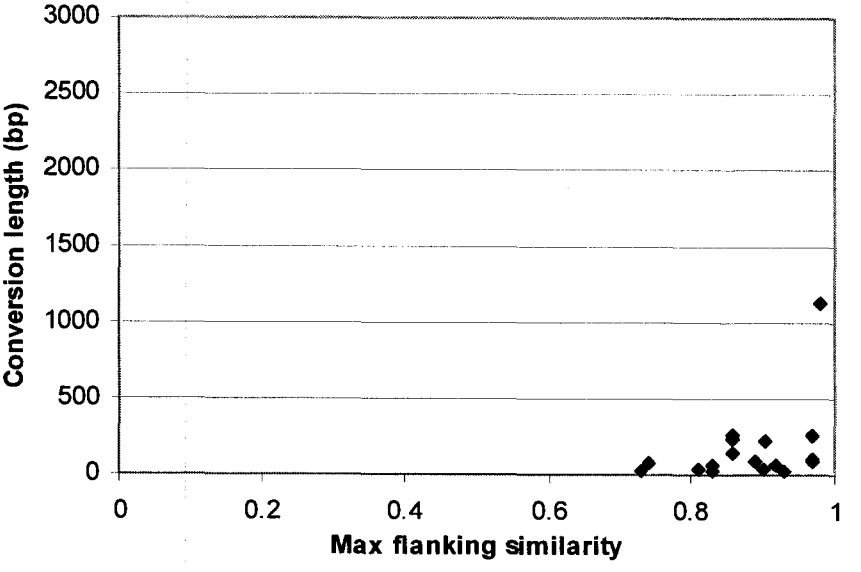
Figure 3.


A
3000
2500
2000
1500
1000
500
0
Conversion length (bp)
0
0.2
0.4
0.6
0.8
1
Maximum flanking similarity
B
3000
2500
2000
1500
1000
500
0
Conversion length (bp)
0
0.2
0.4
0.6
0.8
1
Maximum flanking similarity

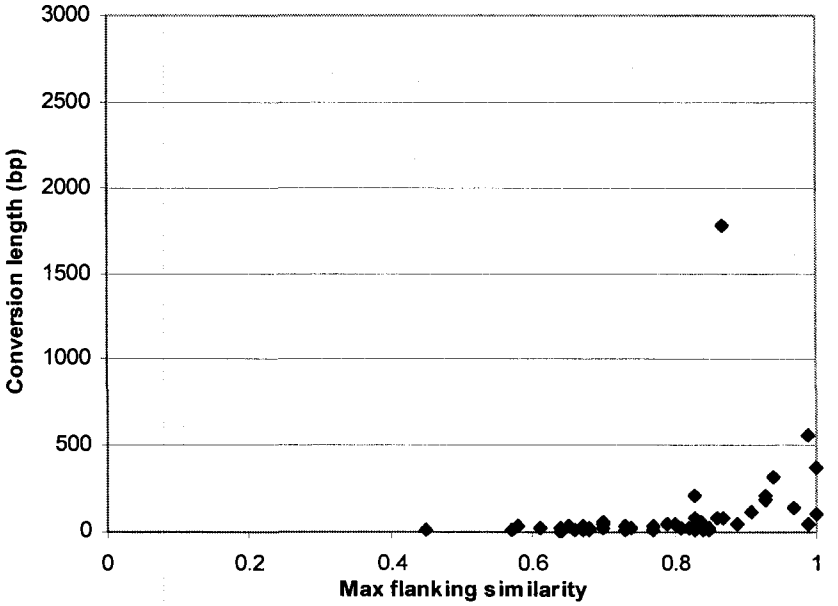
C


3000
2500
2000
1500
1000
500
0
Conversion length (bp)
0
0.2
0.4
0.6
0.8
1
Max flanking similarity

D


3000
2500
2000
1500
1000
500
0
Conversion length (bp)
0
0.2
0.4
0.6
0.8
1
Max flanking similarity

E


3000
2500
2000
1500
1000
500
0
Conversion length (bp)
0
0.2
0.4
0.6
0.8
1
Max flanking similarity

Chapter 6

Ectopic gene conversions increase the G+C content of duplicated yeast and *Arabidopsis* genes.

David Benovoy, Robert T. Morris, Antoine Morin and Guy Drouin
Département de biologie, Université d'Ottawa, Ottawa, Ontario, Canada, K1N 6N5

Research article : Mol. Biol. Evol. 2005, 22:1865-1868.

Keywords: ectopic, gene conversion, GC-content, recombination, *Saccharomyces cerevisiae*, *Arabidopsis thaliana*

Correspondence to: Guy Drouin, Département de biologie, Université d'Ottawa, 150 Louis Pasteur, Ottawa, Ontario, Canada, K1N 6N5. Tel.: (613) 562-5800 ext. 6052, FAX: (613) 562-5486, E-mail: gdrouin@science.uottawa.ca

Running head: Ectopic gene conversions increase GC-content

Abstract

Allelic recombination has previously been shown to increase the GC-content of the sequences of a wide variety of eukaryotic species. Ectopic recombination between clustered tandemly repeated genes has also been shown to increase their GC-content. Here we show that gene conversions between the dispersed genes found in the duplicated regions of the yeast and *Arabidopsis* genomes also increases their GC-content when these genes are more than 88% similar.

My contribution to this published manuscript involved data collection solely for the relationship between GC-content and sequence similarity of *S. cerevisiae* gene pairs (see Figure 1).

Introduction

The nucleotide content of genes and genomes changes during evolution. Processes that increase AT-content are well known. They include the deamination of 5-methylcytosine into thymine and oxidative damage to cytosine or guanine (Lindahl 1993, Birdsell 2002). Processes that increase GC-content are not as well known (Sueoka 2002). However, many studies have shown that DNA repair mechanisms are biased towards GC nucleotides (Brown and Jiricny 1988, 1989; Bill et al. 1998). Frequent DNA repair, such as the DNA repair associated with recombination, is therefore expected to increase GC-content during evolution. These predictions have been confirmed by several studies that showed that allelic recombination does increase the GC-content of yeast, *Caenorhabditis elegans*, *Drosophila*, *Xenopus*, bird and mammalian DNA sequences (Gerton et al. 2000; Fullerton, Bernardo Carvalho, and Clark 2001; Galtier et al. 2001; Marais, Mouchiroud, and Duret 2001; Takano-Shimizu 2001; Birdsell 2002; Duret 2002; Kong et al. 2002; Galtier 2003; Marais 2003; Jensen-Seaman et al. 2004, Meunier and Duret 2004). One would also expect that ectopic gene conversions, i.e., gene conversions between duplicated genes located at different loci, would also increase the GC-content of the genes involved. In fact, some studies have shown that ectopic recombination between clustered tandemly repeated genes also increases their GC-content (Hickey, Wang, and Magoulas 1994; Galtier 2003; Kudla, Helwak, and Lipinski 2004; Noonan et al. 2004). Here we use ohnologs, i.e., duplicated genes produced by genome duplications (Wolfe 2001), to show that gene conversions between dispersed duplicated genes also increase their GC-content.

The ohnologs found in the yeast (*Saccharomyces cerevisiae*) and *Arabidopsis thaliana* genomes are particularly well suited to test the effect of ectopic gene conversions on the GC-content of genes because they consist of pairs of duplicated genes which were all created at the same time. The yeast genome duplication occurred some 150 million years ago (Langkjaer et al. 2003). As a result of this duplication, 54 duplicated gene blocks can still be found in the yeast genome and all but two of these duplicated gene blocks are found on different chromosomes (Wolfe and Shields 1997). The *Arabidopsis thaliana* genome contains ohnologs derived from at least two complete genome duplications, the last of which occurred some 24-40 million years ago (Blanc, Hokamp, and Wolfe 2003). Here, we only analyzed the *Arabidopsis* ohnologs from the most recent duplication in order to use genes that were duplicated at the same time. These recently duplicated genes represent 85% of the onhologs found in the *Arabidopsis* genome and most of them are located on different chromosomes (Blanc, Hokamp, and Wolfe 2003).

Materials and Methods

The sequences of the 750 yeast ohnologs (375 pairs of genes), and of the 4994 *Arabidopsis* recent ohnologs (2497 pairs of genes), were downloaded from the NCBI web site (http://www.ncbi.nlm.nih.gov/) using the lists of duplicated genes generated by the studies of Wolfe and Shields (1997) and Blanc, Hokamp, and Wolfe (2003) (http://wolfe.gen.tcd.ie/). Each pair of duplicated genes was aligned using ClustalW (Thompson, Higgins, and Gibson 1994). The average GC-content (%) at the third position of codons and the average uncorrected sequence similarity of each aligned gene pair were then computed using an in-house PERL script.

The yeast recombination data of the Gerton et al. (2000) study was obtained from http://derisilab.ucsf.edu/hotspots/. The median recombination rate was computed from the seven replicates of red:green ratios for each of the 750 yeast ohnologs. Our yeast recombination values are therefore median recombination rates. Because of the low density of genetic markers, the recombination map of *Arabidopsis* still does not allow to measure local recombination rates (Wright, Agrawal, and Bureau 2003; Marais, Charlesworth, and Wright 2004). We therefore did not attempt to measure the effect of recombination on the GC3-content of *Arabidopsis* ohnologs.

All statistical analyses (Kolmogorov-Smirnov tests of normality, linear and non-linear regression analyses, etc.) were performed using S-plus v6.2 (Insightful Corporation, Seattle, WA) and Excel (Microsoft Corporation, Redmond, WA).

Results

Figure 1 clearly shows that the genes found in the duplicated regions of the yeast genome are divided into two groups. The first group is composed of sequences less than 87.7% similar and there is no correlation between sequence similarity and GC-content at third positions of codons ($r^2 = 1 \times 10^{-6}$, $p = 0.98$). The second group is composed of sequences more than 87.7% similar and there is a significant correlation between sequence similarity and GC-content at third positions of codons ($r^2 = 0.085$, $p = 0.036$). This division into two groups (i.e., with two regressions) is significantly better than a less complex model with a single regression ($F = 2.93$, $p = 0$) and the inflection point is at 87.7% similarity (95% CI of 79.1 - 94.3%). The mean GC3-content of sequences less than 87.7% similar is 39.3% and is significantly lower (Wilcoxon rank-sum test, $Z = 5.42$, $p = 0$) than that of sequences more than 87.7% similar (with a GC3-content of 43.0%). In contrast, the mean median recombination rate (and standard error) of sequences less than 87.7% similar (1.07 ± 0.01) is not significantly different ($Z = 0.07$, $p = 0.95$) from that of sequences more than 87.7% similar (1.09 ± 0.02).

Figure 2 does not show a clear division of yeast ohnologs into two groups based on their median recombination rates. However, it shows that lower recombination rates are more frequent than higher recombination rates and that recombination rates are positively correlated with GC3-content ($r^2 = 0.16$, $p = 0$). We also performed a multiple non-linear regression analysis of the effect of similarity and recombination on GC3-content. We found that recombination rate has no effect on GC3-content. In fact, for

recombination rate, both the slopes before and after the inflection point are not significantly different from zero ($p = 0.24$ and 0.16, respectively).

Figure 3 shows that the ohnologs found in the *Arabidopsis* genome are also divided into two groups. The first group is composed of sequences less than 86.6% similar and there is no correlation between sequence similarity and GC-content at third positions of codons ($r^2 = 0.001$, $p = 0.08$). The second group is composed of sequence more than 86.6% similar and there is a significant correlation between sequence similarity and GC-content at third positions of codons ($r^2 = 0.10$, $p = 2 \times 10^{-5}$). This division into two groups (i.e., with two regressions) is significantly better than a less complex model with a single regression ($F = 20.70$, $p = 0$) and the inflection point is at 86.6% similarity (95% CI of 85.6 - 87.6%). The mean GC3-content of sequences less than 86.6% similar is 43.49% and is significantly lower (Wilcoxon rank-sum test, $Z = 3.40$, $p = 0.0007$) than that of sequences more than 86.6% similar (45.65%).

Discussion

In both yeast and *Arabidopsis*, the GC-content of the third codon positions of sequences less than 88% similar shows no correlation with sequence similarity whereas that of sequences more than 88% similar shows a significant correlation with sequence similarity (Figures 1 and 3). Since this division into two groups is not due to differences in recombination (Figure 2), our results suggest that ectopic gene conversions increase the CG-content of dispersed duplicated yeast and *Arabidopsis* genes. Some of the genes which were duplicated 150 MYA in the yeast genome and 24-40 MYA in the *Arabidopsis* genome have not only retained a high level of similarity through gene conversions but these conversions have also increased their GC-content.

Both experimental and sequence analyses studies have shown that gene conversions are more frequent between more similar sequences (Borts and Haber 1987; Modrich and Lahue 1996; Drouin 2002) and the study of Gao and Innan (2004) has shown that many yeast ohnologs have been subject to numerous gene conversions. One therefore expects similar sequences to become even more similar due to gene conversions whereas less similar sequence will gradually diverge from one another and thus escape gene conversions. In fact, the clear division of the yeast ohnologs into two groups (below and above 87.7% similarity; Figure 1) likely represents genes that escaped and genes still undergoing gene conversions, respectively. Furthermore, this division into two groups is not due to different recombination rates of the genes found into two groups because the average recombination rate is the same in both groups. The absence of such visually obvious groups in *Arabidopsis* (Figure 3) could be the result of the lower level of

recombination in *Arabidopsis* and the fact that the ohnologs of this species diverged more recently. In fact, the shape of the distribution observed in *Arabidopsis*, and the excess of data points between 70 and 85% similarity, is what would be expected under the hypothesis of recently duplicated genes undergoing a continuous rate of escape from gene conversion (Figure 3). The fact that a similarity of at least 88% is necessary to observe an effect in both species suggests that the mechanisms responsible for ectopic gene conversions are similar in fungi and plants.

Another hypothesis which could explain our results would be that they reflect differences in codon usage. Under this hypothesis, more conserved genes would use more codons containing G or C in third codon positions. However, this hypothesis would not explain why the correlation between GC-content and similarity is limited to genes having more than 88% similarity, why this correlation is limited to genes having more than 88% similarity in two very different species, and why this correlation is of the same magnitude in both species (r^2 of 0.085 and 0.10 for yeast and *Arabidopsis*, respectively). Since optimal codons are known to be species specific, and that there is strong selection for optimal codons in yeast but not in Arabidopsis, one would not expect selection for optimal codons to lead to similar increases in GC3 in these two very different species (Sharp et al. 1988; Duret and Mouchiroud 1999). In contrast, the GC-biased gene conversion hypothesis explains both the fact that conversions are limited to very similar sequences and the fact that GC-content increases with similarity.

The fact that the correlation between GC-content and similarity is relatively low is consistent with the previous yeast study of Gerton et al. (2000) where frequent allelic recombination only resulted in GC-content increases of a few percent. Similarly, the

correlation between the GC-content of the third codon positions of 6,143 yeast open reading frames and their mean allelic recombination rate is also relatively low (ρ^2 = 0.156) but is highly significant ($p = 3.7 \times 10^{-211}$, Birdsell 2002). Since gene conversions between unlinked repeated sequences are less frequent than between alleles (Petes and Hill 1988; Haber et al. 1991; Goldman and Lichten 1996), one expects ectopic gene conversions to have a smaller effect than allelic gene conversions. Interestingly, the correlation we observed between the recombination rate and GC3-content of yeast ohnologs ($r^2 = 0.16$; Figure 2) is very similar to that of Birdsell (2002). This suggests that yeast ohnologs are a representative sample of yeast genes.

The effect of biased gene conversion on GC-content requires that the gene being converted and its template be different (Galtier et al. 2001). Since highly inbred species would be homozygous for most of their genes, one would not expect biased gene conversion to affect the GC-content of their genes. This prediction is supported by the absence of correlation between the rate of crossing over and the GC-content of the genes found in *Arabidopsis*, a species with a selfing rate of about 99% (Marais, Charlesworth, and Wright 2004). The presence of a positive correlation between recombination rate and GC-content in yeast (see above), another species with a selfing rate of about 99% (Johnson et al. 2004), might be due to the very high level of recombination of this species. The fact that we observed significant correlations between the similarity of ohnologs more than 88% similar and GC-content in both *Arabidopsis* and yeast is therefore likely due to the relatively high level of mismatches between these duplicated genes relative to those of alleles and that ectopic conversions occur even in self-fertilizing species (Haubold et al. 2002).

Acknowledgments. We would like to thank the two anonymous reviewers for their constructive comments on previous versions of this manuscript. This research was supported by NSERC Discovery grants to A.M. and G.D.

Literature Cited

Bill, C. A., W. A. Duran, N. R. Miselis, and J. A. Nickoloff. 1998. Efficient repair of all types of single-base mismatches in recombination intermediates in Chinese hamster ovary cells. Competition between long-patch and G-T glycosylase-mediated repair of G-T mismatches. Genetics **149**:1935-1943.

Birdsell, J. A. 2002. Integrating genomics, bioinformatics, and classical genetics to study the effects of recombination on genome evolution. Mol. Biol. Evol. **19**:1181-1197.

Blanc, G., K. Hokamp, and Wolfe, K. H. (2003). A recent polyploidy superimposed on older large-scale duplications in the *Arabidopsis* genome. Genome Res. 13: 137-144.

Borts, R. H., and J. E. Haber. 1987. Meiotic recombination in yeast: alteration by multiple heterozygosities. Science **237**:1459-1465.

Brown, T. C, and J. Jiricny. 1988. Different base/base mispairs are corrected with different efficiencies and specificities in monkey kidney cells. Cell **54**:705-711.

Brown, T. C., and J. Jiricny. 1989. Repair of base-base mismatches in simian and human cells. Genome **31**:578-583.

Drouin, G. 2002. Characterization of the gene conversions between the multigene family members of the yeast genome. J. Mol. Evol. **55**:14-23.

Duret, L. 2002. Evolution of synonymous codon usage in metazoans. Curr. Opin. Genet. Dev. **12**:640-649.

Duret, L, and D. Mouchiroud. 1999. Expression pattern and, surprisingly, gene length shape codon usage in *Caenorhabditis*, *Drosophila* and *Arabidopsis*. Proc. Natl.

Acad. Sci. USA **96**:4482-4487.

Fullerton, S. M., A. Bernardo Carvalho, and A. G. Clark. 2001. Local rates of recombination are positively correlated with GC content in the human genome. Mol. Biol. Evol. **18**:1139-1142.

Galtier, N., G. Piganeau, D. Mouchiroud, and L. Duret. 2001. GC-content evolution in mammalian genomes: the biased gene conversion hypothesis. Genetics **159**:907-911.

Galtier, N. 2003. Gene conversion drives GC content evolution in mammalian histones. Trends Genet. **19**:65-68.

Gao, L. Z., and H. Innan. 2004. Very low gene duplication rate in the yeast genome. Science **306**:1367-1370.

Gerton, J. L., J. DeRisi, R. Shroff, M. Lichten, P. O. Brown, and T. D. Petes. 2000. Global mapping of meiotic recombination hotspots and coldspots in the yeast *Saccharomyces cerevisiae*. Proc. Natl. Acad. Sci. USA **97**:11383-11390.

Goldman, A. S., and M. Lichten. 1996. The efficiency of meiotic recombination between dispersed sequences in *Saccharomyces cerevisiae* depends upon their chromosomal location. Genetics **144**:43-55.

Haber, J. E., W. Y. Leung, R. H. Borts, and M. Lichten. 1991. The frequency of meiotic recombination in yeast is independent of the number and position of homologous donor sequences: implications for chromosome pairing. Proc. Natl. Acad. Sci. USA **88**:1120-1124.

Haubold, B., J. Kroymann, A. Ratzka, T. Mitchell-Olds, and T. Wiehe. 2002.

Recombination and gene conversion in a 170-kb genomic region of *Arabidopsis thaliana*. Genetics **161**:1269-1278.

Hickey, D. A., S. Wang, and C. Magoulas. 1994. Gene duplication, gene conversion and codon bias. Pp 199-207 *in* G. B. Golding, ed. Non-neutral evolution: Theories and Molecular Data. Chapman and Hall, Inc., NY.

Jensen-Seaman,, M. I., T. S. Furey, B. A. Payseur, Y. Lu, K. M. Roskin, C. F. Chen, M. A. Thomas, D. Haussler, and H. J. Jacob. 2004. Comparative recombination rates in the rat, mouse, and human genomes. Genome Res. **14**:528-538.

Johnson, L. J., V. Koufopanou, M. R. Goddard, R. Hetherington, S. M. Schäfer, and A. Burt. 2004. Population genetics of the wild yeast *Saccharomyces paradoxus*. Genetics **166**:43-52.

Kudla, G., A. Helwak, and L. Lipinski. 2004. Gene conversion and GC-content evolution in mammalian Hsp70. Mol. Biol. Evol. **21**:1438-1444.

Kong, A., D. F. Gudbjartsson, J. Sainz, et al. (16 co-authors). 2002. A high-resolution recombination map of the human genome. Nat. Genet. **31**:241-247.

Langkjaer, R. B., P. F. Cliften, M. Johnston, and J. Piskur. 2003. Yeast genome duplication was followed by asynchronous differentiation of duplicated genes. Nature **421**:848-852.

Lilley, D. M. J., and M. F., White. 2001. The junction-resolving enzyme. Nat. Rev. Mol. Cell. Biol. **2**:433–443.

Lilley, D. M. J. 2000. Structures of helical junctions in nucleic acids. Q. Rev. Biochem. **33**:109–159.

Lindahl, T. 1993. Instability and decay of the primary structure of DNA. Nature **362**:709-715.

Marais, G., D. Mouchiroud, and L. Duret. 2001. Does recombination improve selection on codon usage? Lessons from nematode and fly complete genomes. Proc. Natl. Acad. Sci. USA **98**:5688-5692.

Marais, G. 2003. Biased gene conversion: implications for genome and sex evolution. Trends Genet. **19**:330-338.

Marais, G., B. Charlesworth, and S. I. Wright. 2004. Recombination and base composition: the case of the highly self-fertilizing plant *Arabidopsis thaliana.* Genome Biol. **5**:R45.

Meunier, J., and L. Duret. 2004. Recombination drives the evolution of GC-Content in the human genome. Mol. Biol. Evol. **21**:984-990.

Modrich, P., and R. Lahue. 1996. Mismatch repair in replication fidelity, genetic recombination, and cancer biology. Annu. Rev. Biochem. **65**:101-133.

Noonan, J. P., J. Grimwood, J. Schmutz, M. Dickson, and R. M. Myers. 2004. Gene conversion and the evolution of protocadherin gene cluster diversity. Genome Res. **14**:354-366.

Petes, T. D., and C. W. Hill. 1988. Recombination between repeated genes in microorganisms. Annu. Rev. Genet. **22**:147-168.

Sharp, P. M., E. Cowe, D. G. Higgins, D. C. Shields, K. H. Wolfe, and F. Wright. 1988. Codon usage patterns in *Escherichia coli, Bacillus subtilis, Saccharomyces cerevisiae, Schizosaccharomyces pombe, Drosophila melanogaster* and *Homo*

sapiens; a review of the considerable within-species diversity. Nucleic Acids Res. **16**:8207-8211.

Sueoka, N. 2002. Wide intra-genomic G+C heterogeneity in human and chicken is mainly due to strand-symmetric directional mutation pressures: dGTP-oxidation and symmetric cytosine-deamination hypotheses. Gene **300**:141-154.

Takano-Shimizu, T. 2001. Local changes in GC/AT substitution biases and in crossover frequencies on *Drosophila* chromosomes. Mol. Biol. Evol. **18**:606-619.

Thompson, J. D., Higgins, D. G., and T. J. Gibson. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. **22**:4673-4680.

Wolfe, K. H., and D. C. Shields. 1997. Molecular evidence for an ancient duplication of the entire yeast genome. Nature **387**:708-713.

Wolfe, K. H. 2001. Yesterday's polyploids and the mystery of diploidization. Nat. Rev. Genet. **2**:333-341.

Wright, S. I., N. Agrawal, and T. E. Bureau. 2003. Effects of recombination rate and gene density on transposable element distributions in *Arabidopsis thaliana*. Genome Res. **13**:1897-1903.

Figure 1. Relationship between the average GC-content of third codon positions (GC3) and the average sequence similarity of the 375 pairs of ohnologs in the yeast genome.

Figure 2. Relationship between the GC-content of third codon positions (GC3) and the median recombination rate of the 750 ohnologs found in the yeast genome.

Figure 3. Relationship between the average GC-content of third codon positions (GC3) and the average sequence similarity of the 2497 pairs of recent ohnologs in the *Arabidopsis* genome.

Figure 1.

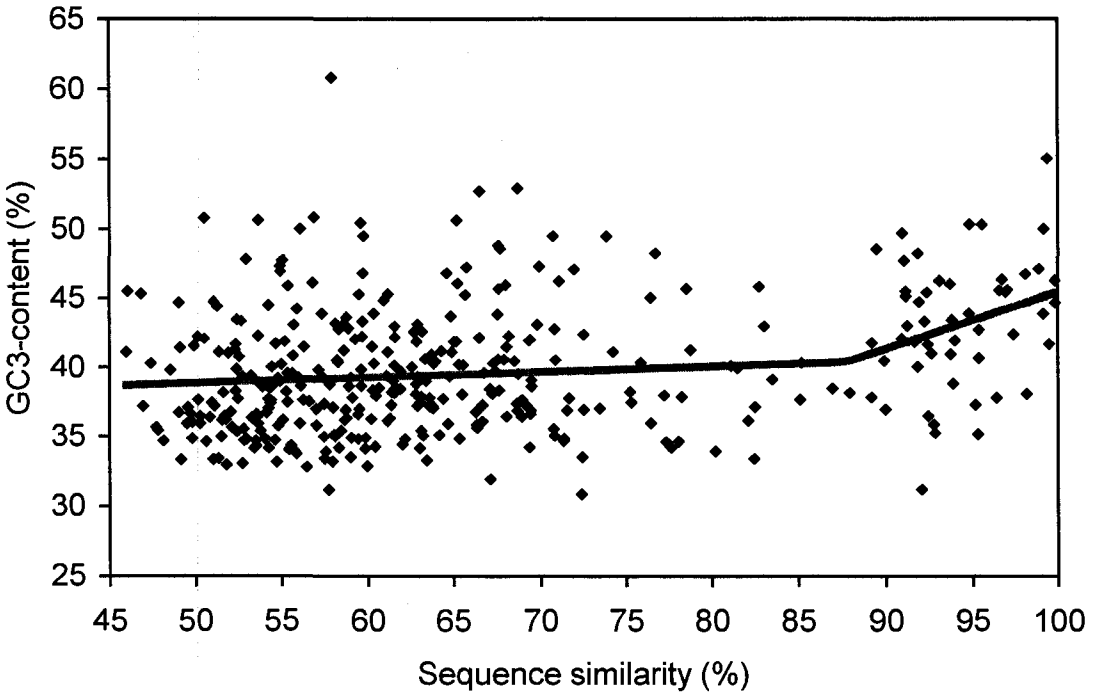

65
60
55
50
45
40
35
30
25
GC3-content (%)
45
50
55
60
65
70
75
80
85
90
95
100
Sequence similarity (%)

Figure 2.

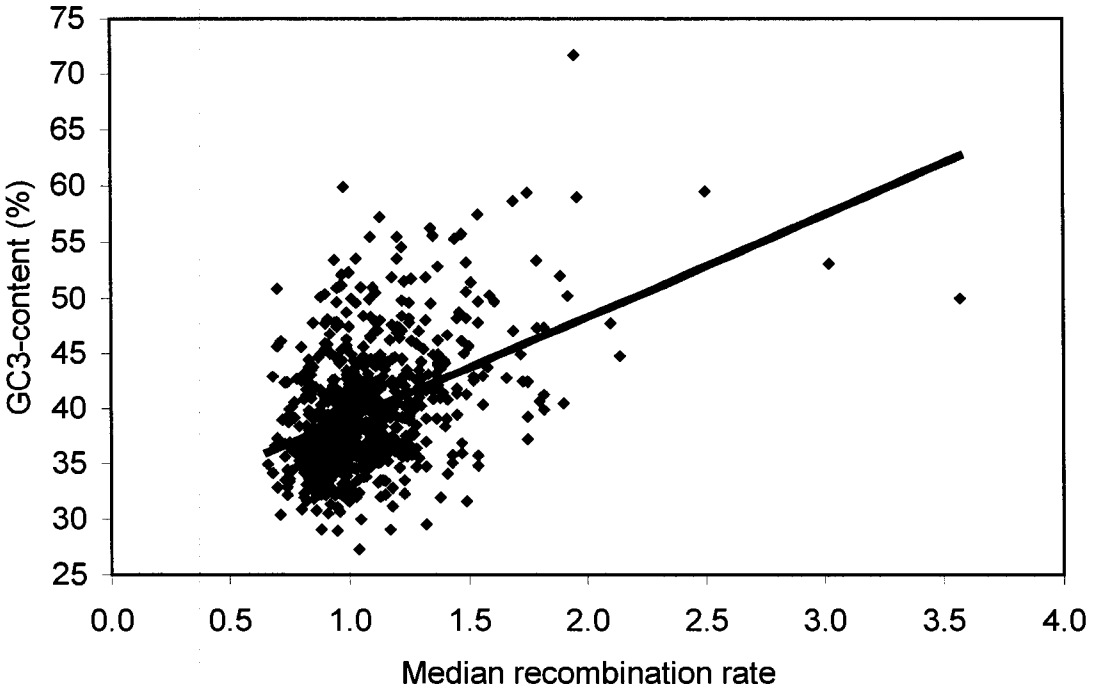
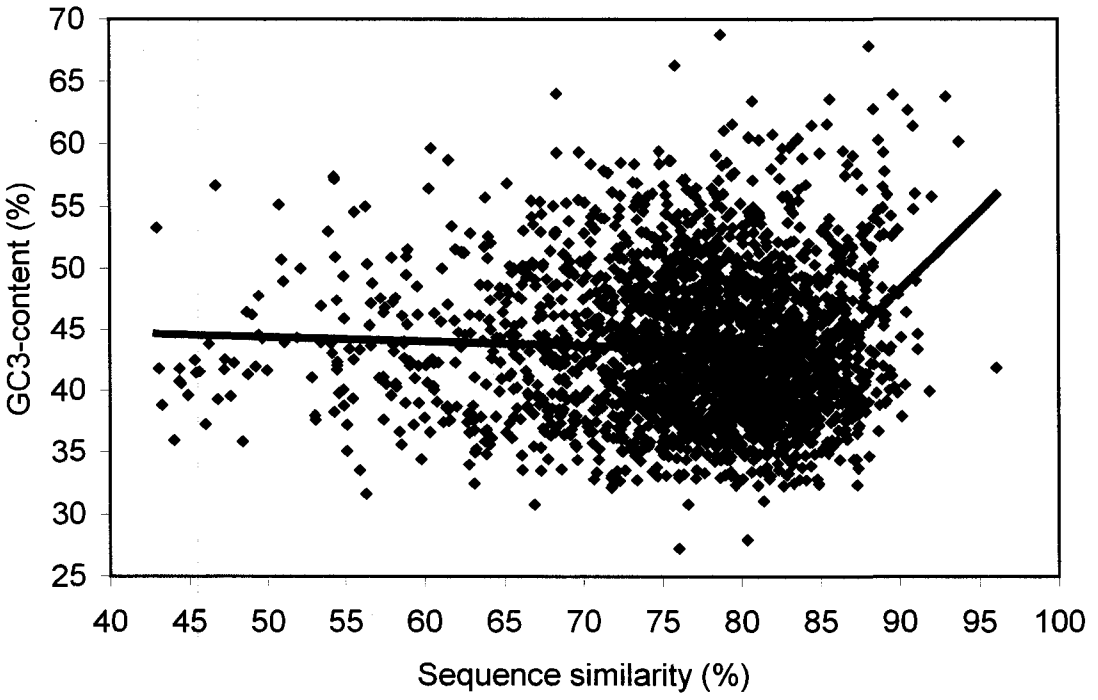

GC3-content (%)
75
70
65
60
55
50
45
40
35
30
25
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Median recombination rate

Figure 3.


GC3-content (%)
70
65
60
55
50
45
40
35
30
25
40
45
50
55
60
65
70
75
80
85
90
95
100
Sequence similarity (%)

Chapter 7

General Summary

The repair of double strand DNA breaks is critically important to genomic stability, and a by-product of this process is gene conversion. *E. coli* and *S. cerevisiae* are historically important model organisms and numerous studies have looked at homologous recombination and mismatch repair mechanisms of DNA. Relatively few studies have examined the similarities and differences of the physical characteristics of ectopic gene conversions within and between diverse lineages of prokaryotes and yeast. I have identified and provided plausible explanations for these similarities and differences.

Factors affecting Prokaryote and Eukaryote ectopic gene conversion

Double strand DNA breaks are commonly repaired via homologous recombination in both prokaryotes and eukaryotes. Despite the differences in complexity of this process in eukaryote and prokaryotes, these systems of DNA repair rely on similar basic principles in both classes of organisms.

The amount of sequence similarity between the converted genes affects the efficiency of the repair mechanism. A 2 – 4% decrease of sequence similarity reduces recombination repair efficiency by 10 to 40 fold (Watt et al. 1985 and Shen and Huang 1986). Inbar and colleagues found that efficient repair of DSB by gene conversion required at least 250 bp of sequence homology between the damaged gene and repair template in *S. cerevisiae* (Inbar et al. 2000). I found that the length of gene conversions is correlated with the amount of flanking similarity between the converted genes in both prokaryotes and yeast. Therefore, long conversions tend to occur between highly similar homologs (Figure 3, Chapter 2; Figure 2, Chapter 4; Figure 3, Chapter 5). This suggests that there is a molecular mechanism regulating the propensity of recombination between

divergent genes. Studies have found that the mismatch repair system (MMR), most notably the MutS protein in *E. coli* and its homolog in yeast the Msh2 – Msh6 complex have the role of regulating the similarity requirements between the damaged gene and the repair template (Rayssiguier et al. 1989, Selva et al. 1995). In addition, the sequence similarity requirements are not uniform for every organism. For example, I found indications that the life-style of the organism affects the stringency of sequence similarity requirements for ectopic gene conversion. In *E. coli* strains, and in the firmicute (Gram-positive bacteria) phylum, I determined that ectopic gene conversions occur more frequently and require significantly less flanking similarity in the pathogenic strains than in the non-pathogenic (Chapters 2 and 4). Previous studies have found that pathogenic species tend to develop deficiencies in their mismatch repair system, particularly with respect to preventing recombination between divergent sequences (LeClerc et al. 1996). The loss of the recombination barrier explains the more frequent and less stringent similarity requirements for ectopic gene conversions for pathogenic genomes and may be advantageous for these species by allowing them to avoid the immune response.

Ectopic gene conversions can cause changes in the amino acid sequence of the converted genes. This may introduce substitutions in the converted genes which may be disadvantageous; therefore the importance of the functional homologous genes affects the frequency of gene conversions. I found that genes which have been functionally maintained throughout evolution (the ohnologs of hemiascomycetes genomes) are ectopically converted less frequently than genome specific paralogs (Table 4, Chapter 5). In addition, the conserved backbone and ohnolog genes tend to be under stronger negative selection than the genome specific paralogs in *E. coli* and yeast, respectively

(Table 2 Chapter 3; Table 7 Chapter 5). The stonger negative selection in conversed genes and ohnologs suggests that these genes are unable to tolerate frequenct amino acid changes cause by ectopic recombination.

As mentioned previously, ectopic gene conversion frequency is affected by the function of MMR proteins which prevent divergent genes from recombining. I looked at the gene complement in proteobacteria firmicutes and archaea to determine the dependency of gene conversions on presence or absence of DNA double strand break repair genes. I was unable to identify unifying trends between the conversion frequency and gene complement which completely agreed with previous experimental studies. However, I did find isolated instances where the gene conversion frequency could be interpreted based on gene complement. For example, ectopic gene conversions were very rare in two *Buchnera* species that also lacked a homolog of *recA* (Supplementary Tables 1 and 2, Chapter 3). The lack of a *recA* homolog would affect the homology search phase during pre-synapsis which is a requirement for gene conversion. Therefore, a low conversion frequency and the absence of *recA* may be biologically significant (Clark 1973, Wyman et al 2004).

In *E. coli* and yeast, there is a positive correlation between the size of the multigene families and the conversion frequency of those genes (Chapter 2, Drouin 2002). This suggests that the more abundant the homologous genes are in the genome the more likely they will be used as a repair template.

The relative organization of multigene families also affects how often genes are converted. In yeast, I found that intra-chromosomal gene conversions are more frequent than inter-chromosomal conversions (Table 3 Chapter 5). A similar relationship between

chromosomally linked genes and the frequency of their conversion was found in Drosophila and humans (Engels et al. 1994, Benovoy and Drouin, unpublished).

I found that, in paralogous multigene families of yeast, gene conversions occur more frequently between homologous genes which are close together on the same chromosome than between dispersed homologs. This result concurred with other studies where a similar effect was observed in yeast and human (Drouin 2002, Benovoy and Drouin, unpublished). In contrast, within the paralogous families of prokaryotes, I do not find that gene conversions occur more frequently between closely linked genes on the chromosome (Chapters 2 and 4). Previous work suggests that multiple copies of the same chromosome are present within the stationary phase of the *E. coli* cell cycle (i.e., not solely during exponential growth phases) and in archaea genomes (Åkerlund et al. 1995; Bernander and Poplawski 1997; Malandrin et al. 1999). I proposed that the presence of multiple copies of the chromosome effectively unlinked the family members allowing any pairwise associations to occur thereby eliminating the effect of nucleotide distance between converted genes. This is contrasted with diploid yeast where only two genome copies are present and the eukaryotic genomes are several times larger than the prokaryotic chromosomes. Therefore the probability of associating dispersed homologs from different copies of the genome is unlikely.

Converted regions tend to be clustered near the 3' end of genes in yeasts but not in prokaryotes. I found that gene conversions occur more frequently near the 3' end of genes in *S. cerevisiae*. This agrees with previous analyses done on *S. cerevisiae* and human (Drouin 2002; Benovoy and Drouin, unpublished). This bias may be caused by gene conversion via cDNA intermediates (Derr and Strathern 1993; Drouin 2002). These

cDNA intermediates are created by reverse transcriptase (RT) activity on native mRNA in the nucleus. I was unable to detect a similar significant bias in the *E. coli* strains or in the proteobacteria, firmicute or archaea datasets (Chapter 2 and 4). Searches of the annotated prokaryotic genomes found that only two genomes contained copies of functional reverse transcriptase. Further analysis of these individual genomes indicated they did not have gene conversions biased toward the 3' end of genes (Chapter 4). This suggests that cDNA intermediates are rarely used as repair templates in these prokaryotes.

Effect of gene conversion

Studies of alleles and tandem duplicated genes in eukaryotes have found that repeated gene conversions can cause an increase in GC-content. This increase has been attributed to preferential replacement of mismatched bases found in heteroduplex DNA by a guanine or cytosine. Therefore this GC-biased repair model predicts that gene conversions result in a net increase of GC-content of the converted genes. I found that homologous pairs of genes with less than 88% sequence similarity did not show an increase in their GC-content. However those gene pairs that shared at least 88% sequence similarity showed a significant positive correlation between GC-content and sequence similarity. The interpretation of these data suggests that genes which share less than 88% similarity diverged over time and escaped ectopic gene conversion. Those genes that share at least 88% similarity are continually converted and the effect of this is an increase in their GC-content. A similar effect was also detected in Arabidopsis (Benovoy et al. 2005).

Literature Cited

Åkerlund, T., Nordström, K., and R. Bernander. 1995. Analysis of cell size and DNA content in exponentially growing and stationary-phase batch cultures of *Escherichia coli.* J Bacteriol 177:6791-6797.

Benovoy, D., Morris, R. T., Morin, A., and G. Drouin. 2005. Ectopic gene conversions increase the G+C content of duplication yeast and *Arabidopsis* genes. Mol. Biol. Evol. 22:1865-1868.

Bernander, R., and A. Poplawski. 1997. Cell cycle characteristics of thermophilic archaea. J. Bacteriol. 179:4963-4969.

Clark, A. J. 1973. Recombination deficient mutants of *E. coli* and other bacteria. Annual Reviews Genetics 7:67-86.

Derr, L. K., and J. N. Strathern. 1993. A role of reverse transcripts in gene conversion. Nature 361:170-173.

Drouin, G. 2002. Characterization of the gene conversions between the multigene family members of the yeast genome. J. Mol. Evol. 55:14-23.

Engels, W. R., Preston, C. R., and D. M. Johnson-Schlitz. 1994. Long-range cis preference in DNA homology search over the length of a Drosophila chromosome. Science 263:1623-1625.

Inbar, O., Liefshitz, B., Bitan, G., and M. Kupiec. 2000. The relationship between homology length and crossing over during the repair of a broken chromosome. Journal of Biological Chemistry 275:30833-30838.

LeClerc, J. E., Li, B., Payne, W. L., and T. A. Cebula. 1996. High mutation frequencies among *Escherichia coli* and *Salmonella* pathogens. Science 274:1208-1211.

Malandrin, L., H. Huber, and R. Bernander. 1999. Nucleoid structure and partition in *Methanococcus jannaschii*: an archeon with multiple copies of the chromosome. Genetics 152:1315-1323.

Rayssiguier, C., Thaler, D. S., and M. Radman. 1989. The barrier to recombination between *Escherichia coli* and *Salmonella typhimurium* is disrupted in mismatch-repair mutants. *Nature* 342:396-401.

Selva, E. M., New, L., Crouse, G. F., and R. S. Lahue. 1995. Mismatch correction acts as a barrier to homeologous recombination in Saccharomyces cerevisiae . Genetics 139:1175-1188.

Shen, P. and H. V. Huang. 1986. Homologous recombination in *Escherichia coli:* Dependence on substrate length and homology. Genetics 112:441-457.

Watt, V. M., Ingles, C. J., Urdea, M. S., and W. J. Rutter. 1985. Homology requirements for recombination in *Escherichia coli.* Proc. Natl. Acad. Sci. USA 82:4768-4772.

Wyman, C., Ristic, D., and R. Kanaar. 2004. Homologous recombination-mediated double-strand break repair. DNA Repair 3:827-833.