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

David Benovoy
AUTEUR DE LA THÈSE / AUTHOR OF THESIS

M.Sc. (Biology)
GRADE / DEGREE

Department of Biology
FACULTÉ, ÉCOLE, DÉPARTEMENT / FACULTY, SCHOOL, DEPARTMENT

Ectopic gene conversions in eukaryotic genomes

TITRE DE LA THÈSE / TITLE OF THESIS

G. 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

G. Carmody

M. Ekker

X. Xia

Gary W. Slater
LE DOYEN DE LA FACULTÉ DES ÉTUDES SUPÉRIEURES ET POSTDOCTORALES /
DEAN OF THE FACULTY OF GRADUATE AND POSTDOCORAL STUDIES

Ectopic gene conversions in eukaryotic genomes

David Benovoy

Thesis submitted to the
Faculty of Graduate and Postdoctoral Studies
University of Ottawa
in Partial fulfillment of the requirements for the Master's degree in the
Ottawa-Carleton Institute of Biology

Thèse soumise à la
Faculté des études supérieures et postdoctorales
Université d'Ottawa
En vue de l'obtention de la maîtrisse
L'institut de biologie d'Ottawa-Carleton

Your file Votre référence
ISBN: 0-494-14884-5
Our file Notre référence
ISBN: 0-494-14884-5

Table of Contents

List of abbreviations

ADNc	ADN complémentaire
bp	Base pairs
cDNA	Complementary DNA
CDS	Coding sequence
DNA	Deoxyribonucleic acid
DSBR	Double-strand break repair
GC-content	Guanine cytosine content
NCBI	National Center for Biotechnology Information
SDSA	Synthesis-dependent strand annealing model

Acknowledgments

I would like to thank Dr. Guy Drouin, my honors project supervisor as well as my master's advisor, whose knowledge, expertise and views have guided and helped me throughout this and other projects. He has bestowed onto me his passion and thoroughness for research which I hope to develop in my future scientific endeavours. I am also grateful to the other members of my graduate committee: Dr. George Carmody (Biology Department, Carleton University) and Dr. Xuhua Xia (Biology Department, Ottawa University) for their comments and attention at all levels of this work. I really enjoyed working in the Drouin lab and with the many students he has supervised throughout the 3 years I have been there. Particularly, Robert Morris who as been there since the beginning of my studies as an honors student up to the end of my master's. I thank him for taking the time to debug many of my Perl scripts, to discuss statistical problems and for the many conversations we had on scientific and other issues. I would like to acknowledge the NSERC Discovery grant to G.D. for the financial support of this research. Finally, I am grateful to my family and friends for their encouragement, support and patience throughout this work and in my entire life.

Abstract

We studied ectopic gene conversions, i.e., gene conversions between duplicated genes located at different chromosomal positions, in eukaryotic genomes. In the first part we examined the factors affecting ectopic gene conversions in the human genome and compared their characteristics to those observed in other eukaryotic and prokaryotic species. In the second part, we examined the effect that ectopic conversions have on the GC-content of the duplicated genes found in yeast and *Arabidopsis* genomes.

Using Stanley Sawyer's method implemented in his GENCONV program, we identified and characterized the ectopic gene conversions of the human genome. The human gene families containing 3 or more members contained 483 pairs of converted genes. The average length of conversions is 371±752 (± standard deviation) nucleotides long with the smallest conversions being 10 nucleotides long and the largest 6011 nucleotides long. Larger gene conversions are found between sequences that are more similar and the frequency of intra-chromosomal gene conversion increases as the distance between genes decreases. Pairs of intra-chromosomal genes sharing the same transcriptional orientation convert more often than intra-chromosomal genes in opposite transcriptional orientation. The excess of conversions in the 3'-end suggest incomplete cDNA molecules are often involved in gene conversions with chromosomal gene copies.

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.

Résumé

Cette étude porte sur les conversions géniques ectopiques i.e., les conversions géniques entre gènes dupliqués se trouvant à différentes positions chromosomiques. En premier lieu, nous avons examiné les facteurs affectant les conversions géniques ectopiques dans le génome humain et nous avons ensuite comparé ces caractéristiques avec celles observées dans d'autres génomes de procaryotes et d'eucaryotes. Deuxièmement, nous avons examiné les effets des conversions géniques ectopiques sur le contenu en bases GC des gènes dupliqués dans le génome d'*Arabidopsis* et de la levure (*Saccharomyces cerevisiae*).

En utilisant la méthode que Stanley Sawyer a développé dans son programme GENCONV, nous avons identifié et caractérisé les conversions géniques ectopiques du génome humain. Les familles géniques comportant un minimum de trois membres contenaient 483 paires de gènes convertis. La longueur moyenne d'une région convertie est de 371±752 (± erreur type) nucléotides tandis que la plus petite région convertie est d'une longueur de 10 nucléotides et la plus grande de 6011 nucléotides. Les plus longues conversions géniques se produisent entre gènes dont la similarité est élevée. De plus, la fréquence des conversions intra chromosomiques augmente lorsque la distance entre les gènes diminue. Les gènes partageant la même orientation transcriptionelle se convertissent plus souvent que des gènes ayant une orientation transcriptionelle opposé. Les excès de conversions observés dans la partie 3' des gènes convertis suggèrent qu'une molécule ADNc incomplète est souvent impliquée dans la conversion des gènes chromosomiques.

Plusieurs études ont démontré que la recombinaison allélique augmente le contenu en base GC dans de nombreuses espèces eucaryotes. Certaines études ont aussi démontré que la recombinaison ectopique entre gènes regroupés et répétés en tandem augmente aussi le contenu en bases GC de ces gènes. Ici, nous démontrons que les conversions entre gènes dispersés ont le même effet i.e., augmenter le contenu en bases GC des gènes qui présentent plus de 88% de similarité.

Chapter 1 - General Introduction

Gene conversions history

Gene conversion is a fundamental process that operates continually to shape and reshape the genomes of most, if not all, organisms. It rearranges genes or parts of genes both within and between chromosomes, limits the divergence of repeated DNA sequences and promotes repair of damaged DNA. It also provides a potent evolutionary force that serves both to promote genetic diversity and to preserve genetic identity (Lloyd and Brooks 1996). Observed repeatedly since 1924, these abnormal allelic segregations where originally thought to be due to technical errors because they produced expected Mendellian ratios (Esser and Kuenen 1967). It took the conclusive *Neurospora crassa* tetrad analysis presented by Mitchell (1955) to convince the scientific community of the existence of these non-reciprocal exchanges. Since the mid-seventies, most genetic and molecular studies of gene conversions have relied on *Saccharomyces cerevisiae* as a model organism (Perkins 1992). DNA sequencing later revealed that gene conversions are not limited to the alleles of genes or to fungi. The first evidence of gene conversion between mammalian multigene family members was described by Slightom, Blechl and Smithies (1980). They found that the first two thirds of the adjacent non-allelic human $^{G}\gamma$- and $^{A}\gamma$-globin genes had almost identical sequences and that the high similarity found in this region was greater than that observed between two $^{A}\gamma$-globin alleles.

Characteristics of gene conversions: from prokaryotes to eukaryotes

Gene conversion is a well documented phenomenon. One of the best known examples of a multigene family which evolves in a concerted fashion by means of gene conversion is the paralogous rRNA genes in eukaryotes (Gangloff et al. 1996), Bacteria and Archaea (Liao 2000). Other multigene families whose evolution has been influenced by gene conversion include the

visual pigment genes of primates, (Zhou and Li 1996), the silk moth chorion genes (Hibner et al. 1991) and the human globin gene families (Scott et al. 1984). Generally viewed as a process of sequence homogenization, gene conversions can also generate molecular diversity (Archibal and Roger 2002). For example, in the immunoglobulin genes of chicken, gene conversion plays a crucial role in the production of the huge diversity of peptides needed by the immune system (Reynaud et al. 1987). In addition, gene conversions in pathogenic bacteria are more frequent than in non-pathogenic strains likely because the diversity created by these gene conversions permits a better evasion of the host's immune system (Morris and Drouin 2004).

Different types of gene conversions have been observed in these many studies. A conversion event that occurs between alleles is an allelic gene conversion. Ectopic gene conversions involve paralogous genes and can further be divided into intra-chromosomal and inter-chromosomal conversions when they occur within or between chromosomes, respectively.

Allelic and ectopic gene conversion events are highly dependent on the degree of sequence similarity between the donor and recipient sequences. In fact, in *E.coli*, a 2% mismatch between the donor and the recipient sequence can decrease the frequency of recombination four-fold and 10% mismatch can decrease recombination by over 40-fold (Watt et al. 1985; Shen and Huang 1986). These events are 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 and in *E.coli*, the frequency of gene conversion has been shown to be proportional to the number of copies of donor sequences present in a cell (Morris and Drouin 2004). Interestingly, the analysis of the gene conversion events found in *E.coli* and *Saccharomyces cerevisiae* genomes showed that gene conversions where more frequent at the 3'end of yeast genes (Drouin 2002) but not in those of *E.coli* (Morris and Drouin 2004).

Molecular models that explain gene conversions

The first molecular model that explained gene conversion was initially proposed by Holliday in 1964 (Holliday 1964; Stahl 1994). His revolutionary proposal has lead to the model that bears his name (Figure 1). In this model, recombination is initiated with a single-strand nick made in the receiving strand and then follows a second single-strand nick made in the donor strand. This is then followed by the unwinding of both helices and strand exchange. A four-strand structure is thus formed (Holliday junction), which can be resolved to give two different results: crossing over or gene conversion. If the DNA molecules are not identical in sequence, some mismatches will arise in the heteroduplex DNA, and so, the repair system will recognize and correct them. The length of the gene conversion tract will depend on both the migration of the Holliday structure and the capacity to repair the mismatches. Finally, the orientation of the cuts needed to resolve the Holliday junction, either horizontal or vertical, will determine the end result: gene conversion events, crossover events, or both. Therefore, in the Holliday model, gene conversion is essentially a consequence of the DNA heteroduplex formation and the role of the repair system.

To fit in some new data from *Saccharomyces cerevisiae* Meselson and Radding 1975 (Figure 1) proposed some important modifications to the Holliday model. In this organism, little reciprocal heteroduplex DNA could be detected (Stahl 1994; Alani 1994), a feature that departs from the original Holliday model. The Meselson–Radding model is characterized by one single-strand cut that is made in only one of the chains. This chain is then transferred by the action of a DNA polymerase and invades the homolog sequence chain; ligation of the newly synthesized strand with a strand of the same polarity in the other homolog generates the Holliday junction (Figure 1). Resolution of this DNA junction occurs in the same way as in the Holliday model. In this model, the invading sequence can give origin to a non-reciprocal DNA heteroduplex, which later on becomes a reciprocal heteroduplex region, upon migration of the Holliday junction; repair in any of these heteroduplexes may generate gene conversion (Santoyo and Romero 2005).

A more recent model, which has gained wide acceptance, is the double-strand break repair model (DSBR) proposed by Szostak et al. 1983 (Figure 1). This DSBR model emerged as the canonical model because of the huge amount of genetic evidence in fungi that indicate that double-strand breaks (rather than single-strand breaks) can act as initiators of recombination (Stahl 1996; Petes 2001). In this model, recombination is initiated by a double-strand break that is continued by extensive single chain degradation in the 5′ to 3′ direction, thus generating a gap with 3′ overhangs. One of these 3′ overhangs goes on to invade the uncut homolog, thus displacing a D-loop that can pair with the remaining 3′ overhang. This paired D-loop can act as a template for DNA synthesis, primed by the 3′ overhang. This phenomenon is also found in the other invading 3′ end that acts as a primer for a DNA polymerase. These events of DNA synthesis repair the gap formed during initiation with information from the template homolog. During strand ligation, two Holliday junctions are created, which are able to migrate and extend the heteroduplex segment. As in the previous models, orientation of cutting of the Holliday junctions will result in crossover, gene conversion or both. Using this model, segments of gene conversion can be generated in two different ways. One alternative, is through repair of the gap produced during initiation, by DNA synthesis. The other is through mismatch repair of heteroduplex DNA. Thus, gene conversion segment length will depend on the extent of the gap, size of the heteroduplex region and the migration capacity of the Holliday structures.

A specific prediction of all the above models is that gene conversion is associated, half of the time, with crossover. However, a wealth of data from yeast, both from meiotic (Stahl 1996; Petes 2001; Koren et al. 2002; Wolf-Dietrich 2004; Bishop and Zickler 2004) as well as mitotic (Aguilera et al. 2000; Prado et al. 2003; Ira et al. 2003) recombination, shows that gene conversion may occur at significant proportions without an associated crossover. A recent modification of the DSBR model, termed the synthesis-dependent strand annealing model (SDSA, Figure1) (Allers and Lichten 2001) deals nicely with this result. In this model, a double-strand cut is made in one DNA duplex; this double-strand break is then processed by degradation to generate protruding 3′ ends. One of these 3′ ends invades a homologous region and starts DNA synthesis using as template the

homologous strand. So, a D-loop is formed as a consequence of strand displacing and DNA synthesis. The model then postulates that the newly synthesized strand may dislodge from the invaded duplex (conceivably by the action of helicases), making it available to pair with the other 3′ end in its original duplex. A full duplex is then restored by limited DNA synthesis (Santoyo and Romero 2005). Gene conversion may arise in this model through DNA synthesis or through mismatch repair; however, in both cases, gene conversion is not associated with crossover.

Main Objectives

The first part of this study explores the dynamics and factors dictating the occurrences of gene conversions in the human genome and attempts to compares these characteristics with those of other organisms. Using Stanley Sawyer's GENCONV method (Sawyer, 1989), we identified converted regions within genes from multigene families with three or more members. The influence of factors affecting gene conversions such as proximity of converted genes, similarity of the flanking regions, transcriptional orientation and the possible utilization of genomic or cDNA sequences as template strands was then studied. We then compared the characteristics from human gene conversions, such as length and frequency of conversions, to those found in other organisms such as yeast and bacteria.

The second part of this study explores one of the effects of gene conversions. More precisely, we investigated the effect that gene conversions have on the GC-content of the duplicated genes found in yeast and *Arabidopsis* genomes. We show that ectopic gene conversions between duplicated genes that are more than 88% similar increases their GC-content.

Literature Cited

Aguilera, A., S. Chavez, and F. Malagón. 2000. Mitotic recombination in yeast: elements controlling its incidence. Yeast **16**:731–754.

Alani, E., R.A. Reenan and R.D. Kolodner. 1994. Interaction between mismatch repair and genetic recombination in *Saccharomyces cerevisiae.* Genetics **137**:19–39.

Allers, T., and M. Lichten. 2001. Differential timing and control of noncrossover and crossover recombination during meiosis. Cell **106**:47–57.

Archibal, J.M., and A.J. Roger. 2002. Gene duplication and gene conversion shape the evolution of archaeal chaperonins. J. Mol. Evol. **316**:1041-1050.

Bishop D., and D. Zickler. 2004 Early decision: meiotic crossover interference prior to stable strand exchange and synapsis. Cell **117**:9–15.

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

Esser, k., and R. Kuenen. 1967. Genetics of fungi. Springer Verlag, New York.

Gangloff, S., H. Zou, and R. Rothstein. 1996. Gene conversion plays the major role in controlling the stability of large tandem repeats in yeast. EMBO J. **15**:1715–1725.

Hibner, B.L., W.D. Burke, and T.H. Eickbush. 1991. Sequence identity in an early chorion multigene family is the result of localized gene conversion. Genetics **128**:595–606.

Holliday, R. 1964. A mechanism for gene conversion. Genet. Res. **5**:282–304.

Ira, G., A. Malkova, G. Liberi, M. Foiani, and J.E. Haber. 2003. Src2 and Sgs1-Top3 suppress crossovers during double-strand break repair in yeast. Cell **115**:401–411.

Koren, A., S. Ben-Aroya, and M. Kupiec. 2002. Control of meiotic recombination initiation: a role for the enviroment? Curr. Genet. **42**:129–139.

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.

Lloyd, R.G., and K. Brooks Low. 1996. Escherichia coli and Salmonella. Cellular and molecular biology. ASM press, Washington, D.C.

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

Mitchell, M.B. 1955. Aberrant recombination of pyridoxine mutants of *Neurospora.* Proc.Natl. Acad. Sci. USA **41**:215-220.

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.

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

Petes, T.D. 2001. Meiotic recombination hotspots and coldspots. Nat. Rev. Gen. **2**:360–369.

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

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.

Prado,F., F. Cortes-Ledesma, P. Huertas, and A. Aguilera. 2003 Mitotic recombination in *Saccharomyces cerevisiae*. Curr. Genet. **42**:185–198.

Santoyo, G., and D. Romero. 2005. Gene conversion and concerted evolution in bacterial genomes. Bacterial Genomics. **29**:169-183.

Reynaud, C.A., V. Anquez, H. Grimal and J.C. Weill. 1987. A hyperconversion mechanism generates the chicken light chain preimmune repertoire. Cell **48**:379–388.

Sawyer, S.A. 1989. Statistical tests for detecting gene conversions. Mol. Biol. Evol. **6**:526-538.

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.

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

Slightom, J.L., A.E. Blechl, and O. Smithies. 1980. Human fetal G gamma- and A gamma-globin genes: complete nucleotide sequences suggest that DNA can be exchanged between these duplicated genes. Cell **21(3)**:627–638.

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

Stahl, F. 1994. The Holliday junction on its thirtieth anniversary. Genetics **138**:241–246.

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

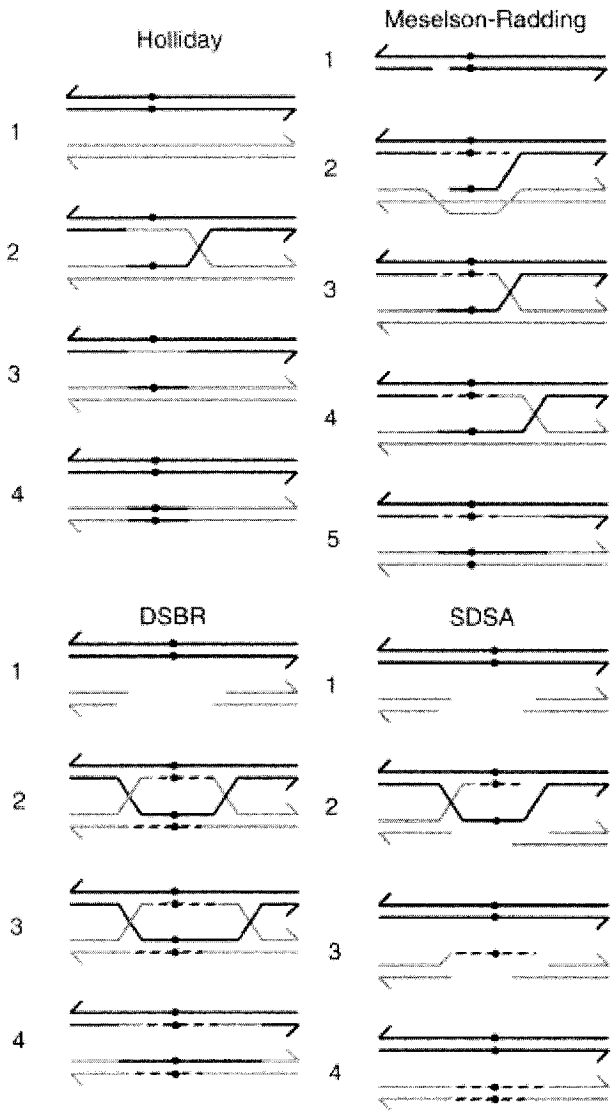
Thompson, J.D., D.G. Higgins, and T.J. Gibson. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucl. Acids Res. **22**:4673-4680.

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.

Wolf-Dietrich, H. 2004. Recombination: Holliday junction resolution and crossover formation. Curr. Biol. **14**:R56–R58.

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. Molecular models of recombination that explains gene conversion. Abbreviated versions of the Holliday, Meselson–Radding, DSBR (double-strand break repair) and SDSA (synthesis-dependent strand annealing) models are given. Different shadings indicate each interacting homologs, with the arrowhead marking the 30 end for each strand. DNA synthesis is indicated by broken lines. Black dots mark the position of sequence differences between the homologs. (Copied from Santoyo and Romer 2005)



Chapter 2

Ectopic gene conversions in the human genome

David Benovoy and Guy Drouin
Département de biologie, Université d'Ottawa, Ottawa, Ontario, Canada, K1N 6N5

Keywords: ectopic, gene conversion, recombination, human

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 in the human genome

Abstract

Using Stanley Sawyer's method implemented in his GENCONV program we identified and characterized the ectopic gene conversions of the human genome. The human gene families coding for proteins and containing 3 or more members contained 483 pairs of converted genes. The average length of conversions is 371±752 (± standard deviation) nucleotides long with the smallest conversions being 10 nucleotides long and the largest 6011 nucleotides long. Larger gene conversions are found between sequences that are more similar and the frequency of intra-chromosomal gene conversions increases as the distance between genes decreases. Pairs of intra-chromosomal genes sharing the same transcriptional orientation convert more often than intra-chromosomal genes in opposite transcriptional orientation. The excess of conversions in the 3'-end suggests that incomplete cDNA molecules are often involved in gene conversions with chromosomal gene copies.

Introduction

The sequencing of the human genome presents a scientific milestone for genomic biologists in that we are now able to apply a holistic approach to the study of the molecular and evolutionary processes that govern our lives. One such process, that continually molds our, and possibly the genomes of all organisms, is homologous recombination. It operates as a powerful evolutionary force that can either promote genetic diversity or maintain genetic identity (Lloyd and Brooks 1996). Homologous recombination is initiated by a chromosomal double stranded break and subsequent repair by the strand invasion of a homologous sequence that is then resolved through a crossover event or by a non-reciprocal transfer of genetic information between the homologous sequences, a process known as gene conversion (Bosch 2004).

Gene conversions occurring between sequences found on the same locus are referred to as allelic gene conversions whereas ectopic gene conversions occur between dispersed sequences found either on the same or on a different chromosome (Petes and Hill 1988). This latter process is fundamental in molecular evolution and has been observed in a broad range of organisms. It was first observed, and then extensively studied, in yeast (Drouin 2002; Haber et al. 1991; Petes 1988; Mitchell 1955) and then in bacteria (Santoyo and Romero 2005; Morris and Drouin 2004) and humans (Bosch 2004; Jeffreys and May 2004; Hurles 2001). Although most human studies have so far focused only on specific gene families they revealed several of the interesting characteristics of ectopic gene conversions. For instance, the paper from Slightom, Blechl and Smithies (1980) on the human γ-globin genes provided the first evidence that ectopic gene conversions occurred in the human genome. Another study showed that the average length of conversions in humans was 31 bp and ranged from 19 bp to 1365 bp (Bosch et al. 2004). In *S. cervisea*, the proximity of converted genes increases the frequency of conversion (Goldman and Lichten 1996; Drouin 2002). Furthermore, a study of the *S. cerevisiae* genome (Drouin 2002) revealed that gene conversions are more frequent in the 3' regions of genes, presumably because of recombination with incomplete

cDNA molecules.

Although gene conversions have been extensively studied, no study has yet characterized all the ectopic gene conversions in the human genome. Here, we present the analysis of gene conversions in all human protein coding gene families and compare the characteristics of the conversions found in this genome with those in other prokaryotic and eukaryotic species. We find that the majority of ectopic gene conversion characteristics are consistent with most previous studies. Surprisingly, these analyses also suggest that frequency of gene conversions is higher in human genes pairs that share the same transcriptional orientation.

Materials and Methods

Sequences

Genbank files (NCBI build 34.3) containing all known protein coding gene sequences (nucleotide and protein) and the 24 chromosomes sequences for *Homo sapiens* where downloaded from the NCBI ftp site (ftp://ftp.ncbi.nlm.nih.gov).

Multigene Families

The BLASTCLUST program was downloaded from the NCBI ftp site and used to identify multigene families from the 27 350 known protein coding gene sequences. These gene families consisted of gene family members (paralogous genes) that share at least 60% of their protein sequences over at least 50% of their lengths. The protein sequence similarity cut-off of 60% was chosen because a similar analysis found no extra gene conversions below this criterion (Morris and Drouin 2004). Redundant gene copies due to alternative gene splicing were then removed and ClustalW (Thompson et al. 1994) was used to align the protein sequences. Will Fisher's ALIGN2AA Perl script was used to align the corresponding coding sequences. (http://sunflower.bio.indiana.edu/~wfischer/Perl_Scripts/).

Gene Conversion Analyses

As in previous studies (Drouin 2002; Morris and Drouin 2004) the GENECONV 1.7 program (http://www.math.wustl.edu/~sawyer/geneconv/) was used to identify gene conversions within the human genomes. Using the same method as with other genomes (*S. cerevisiae* and *E. coli,* respectively) allows us to compare the general characteristics of their gene conversions with the ones found in our study of the human genome. Previous studies have assessed the power (type II

errors) and the rate of false positives (type I errors) of 14 different recombination detection methods using simulated and empirical data sets (Posada and Crandall 2001; Posada 2002). The GENCONV method performed well with both types of data when using gene families with at least 3 members and had a type I error rate of about 5%, when the sequence divergence was $\geq$ 5%. This method is more reliable when using gene families containing three or more members because it can then identify polymorphic sites and thus allow the differentiation between true gene conversions and mutation cold or hot spots (Sawyer 1989; Drouin 2002). Significant ($p \leq 0.05$) global inner fragments with some mutation (g-scale = 2) were used. Duplicated gene conversions were removed from the dataset using phylogenetic analyses of each gene family as previously described (Drouin 2002).

Converted regions bordering introns or spanning multiple exons were further investigated. If any gene conversions where found to be near an exon-intron boundary ($\leq$2 nucleotides away), the intron sequences of this pair of genes where aligned using ClustalW (Thompson et al. 1994). The aligned introns sequences were then inserted into the corresponding exon-intron boundary of the aligned coding sequences of this pair of genes. The same was done for gene conversions spanning multiple exons (multiple exon-intron boundaries). This new alignment was then used as input for the GENECONV program. The g-scale value was set to 2 and the "Include_monosites" option was used because it permits the analyses of only two sequences. Since analyzing only two sequences removes the control for conserved sites, any significant gene conversions detected inside an intron sequence was further investigated. These converted intron sequences were blasted against the ALU and the non-redundant database at NCBI and, if any significant hits for repeated elements were detected for a particular intron sequence, that gene conversion event was removed from the analysis.

Distance, Orientation and Location of Duplicated and Converted Genes

The chromosomal location of each converted gene and conversion event was calculated relative to the first nucleotide in the chromosome sequence file. The orientation of the pair of genes involved in the conversion event was determined by aligning, using FASTA34 (Pearson and Lipman 1988), these genes and determining if the reverse complement of the gene in question was utilized (reverse strand) or not (forward stand) to align it with its corresponding chromosome sequence. In addition, to verify if the orientation of converted genes is biased we compared them to the orientation of all paralogous genes found in the human genome. The distance between intra-chromosomal converted genes was computed by using the difference between the center positions of both converted gene regions.

Statistical analysis

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

Results

Characteristics of human gene conversions

We identified 483 ectopic gene conversion events in the data set produced by BLASTCLUST (that contained 1434 gene families with three or more members). A significant positive correlation (ρ^2 = 0.34, $p = 5.32 \times 10^{-15}$; Spearman rank correlation test) exist between the number of gene conversions and the number of genes in each gene family. Furthermore, the gene conversion frequency (number of gene conversions / total number of genes comparison) for humans is 1.12% (483/43300).

The average size of a human gene conversion is 371±752 (± standard deviation) nucleotides long. The smallest gene conversion is 10 nucleotides long and the largest is 6011 nucleotides long. Figure 1 describes the relationship between the size of the gene conversion and the maximum sequence similarity computed between both genes involved in the gene conversion event in 100 nucleotides upstream or downstream from this event. Small conversion events (< 1000 bp) occur between 23.8% and 100% sequence similarity, whereas large conversion events (> 1000 bp) only occur between 89% and 100% sequence similarity. Furthermore, there is a highly significant positive correlation between the maximum sequence similarity of the flanking region and the length of the gene conversion (ρ^2 = 0.51, $p = 2.81 \times 10^{-27}$; Spearman rank correlation test). A weaker correlation is observed when comparing the minimum sequence similarity of the flanking region with the length of the conversion event in that region (ρ^2 = 0.31, $p = 1.18 \times 10^{-12}$; Spearman rank correlation test). Therefore, maximum sequence similarity was utilized in Figure 1 because it enabled a more representative rendering of the molecular mechanism involved in gene conversion events. This is further illustrated in Figure 2 where the maximum sequence similarity presents a strong positive correlation (ρ^2 = 0.83, p = 0.002; Spearman rank correlation test) and the minimum sequence similarity present a non-significant correlation (ρ^2 = 0.44, p = 0.07; Spearman rank correlation test) with the number of converted regions.

An interesting characteristic of ectopic gene conversions in the human genome is that the converted regions are not distributed randomly along the length of the converted genes (Figure 3.). Of the 966 converted regions, 199 where found in the last 10% of the genes. This is the only significant excess ($\chi^2 = 108.5$, $p = 2.9 \times 10^{-19}$; chi square test) of converted regions detected in this distribution. All other bins show no significant deviation (chi-square results not shown) from the expected uniform distribution (96.6 converted regions per bin). Figure 3 also shows the distribution of converted regions that cover parts of exon as well as intron sequences. For these converted regions, no excess in any region of the converted genes were detected.

Figure 4 shows that the length of converted intron regions is smaller than 100 bp before 75% maximum flanking sequence similarity but can be close to 900 bp long when the maximum flanking similarity is greater than 75%. The longest and the shortest converted intron regions are 904 bp and 2 bp, respectively. The average converted intron region is 177.81 ± 185.01 bp (± standard deviation). The average length of a conversion spanning both an exon and an intron region is 318 ± 212 bp and is significantly smaller (z = -1.99, $p = 0.022$; z-test) than conversions spanning exon regions only (387 ± 852 bp). The length of converted intron regions and the maximum flanking similarity are significantly correlated ($\rho^2 = 0.56$, $p = 6.96 \times 10^{-10}$; Spearman rank correlation test).

Orientation, distance and local recombination frequency between converted genes

Interestingly, the pairs of genes involved in the gene conversion events and found on the same chromosome (intra-chromosomal gene conversions) present a bias as to their orientation on their chromosomes when compared to the distribution of all multiple family members. Of the 5394 genes from families with three members or more, 2657 (~50%) are on the Watson strand and 2737 (~50%) are on the Crick strand. However, when we compare the orientation of the pairs of genes involved in intra-chromosomal gene conversion event we find that 64.8% of these pairs of genes share the same orientation whereas 35.2% have opposite orientations. This is significantly different

from the expected distribution of 50% ($\chi^2 = 21.16$, $p = 4.21\text{x}10^{-6}$; chi square test). Furthermore, the distribution of pairs of genes involved in inter-chromosomal conversions does not differ from the expected distribution of 50% for transcriptional orientation ($\chi^2 = 0.61$, $p = 0.43$; chi square test).

Intra-chromosomal conversions are much more frequent than inter-chromosomal conversions. Indeed, they are almost 5 times more frequent than inter-chromosomal gene conversions (401 intra-chromosomal versus 82 inter-chromosomal). The frequency of intra-chromosomal gene conversion events also increases as the distance between genes implicated in the conversion event decreases. In fact, a Pearson correlation analysis performed on the ranked distances and ranked number of gene conversions shows a strong negative correlation ($\rho = -0.9$, $\rho^2 = 0.81$, $p = 0.037$; Figure 5). In addition, the median distance between intra-chromosomal converted genes is $7.80\text{x}10^4 \pm 2.88\text{x}10^7$ (± standard deviation) nucleotides. Given that the average gene density in the human genome is about 1 gene per $1\text{x}10^5$ nucleotides (International human genome sequencing consortium 2001) entails that the majority of converted genes are neighbors from one another. In fact, the majority of converted genes are $1\text{x}10^4$ (221 pairs of converted genes) and $1\text{x}10^5$ (90 pairs of converted genes) nucleotides from each other (Figure 5).

The frequency of conversions events are not uniformly distributed throughout the human chromosomes. Table 1 shows that gene conversions occur more frequently in some regions than in others. One cause of this clustering could be uneven recombination rates throughout the genome. The study of Jensen et al. (2004) estimated the local recombination rates across the human genome and has found that these rates of recombination differ between and within chromosomes. Using this information we compared the number of gene conversions and the local recombination rate in each chromosome (using windows of 5 Mb). This analysis shows that there is a significant positive correlation between these variables ($\rho^2 = 0.293$, $p = 2.65\text{x}10^{-12}$; Spearman rank correlation test).

Discussion

The ectopic gene conversions found in the human genome share several characteristics with those found in other genomes. One characteristic of ectopic gene conversions that is shared between different organisms is the correlation between gene conversion lengths and flanking region similarity. Figure 1 shows that long converted regions (> 1000 bp) are limited to regions where the flanking similarity is elevated (> 89%). This is similar to what was found in bacteria, yeast, mouse and *Drosophila* (Morrin and Drouin; Drouin 2002; Modrich and Lahue 1996). This suggests that a high level of similarity in the flanking regions is necessary to stabilize the Holliday junction (Figure 2) and permits a longer migration of the Holliday junction and, therefore, a larger converted region. In contrast, low sequence similarity (< 89 %) stops the migration of the Holliday junction and results in conversions which are smaller than 1000 bp.

A second factor influencing ectopic gene conversion events is the distance between the pairs of converted genes. In fact, the frequency of gene conversion is inversely proportional to the distance between the pairs of genes involved in conversion events (Figure 5). This observation is consistent with previous studies in *C. elegans* (Semple and Wolfe 1999) and yeast (Goldman and Litchen 1996; Drouin 2002) which also found a negative correlation between the frequency of gene conversions and the distance between gene pairs. A proposed explanation for this phenomenon is that following a double stranded break, the mechanism engaged in fixing it must recruit a DNA sequence that will serve as a template during the repair process. For this, a protein complex is engaged in sampling DNA sequences that share a high similarity with the damaged strand (Lilley and White 2001; Lilley 2000). This protein complex will seek an adequate DNA sequence that is within close proximity of the break, thus generating the observed correlation.

Our results show that gene conversions can extend into intron sequences (Figure 4). We find that the relationship between the length of converted intron sequences and the similarity of the flanking region (Figure 4) is similar to the one for all converted regions (Figure 1). In addition,

significantly smaller conversion lengths were observed in conversions overlapping intron sequences when compared to conversions in exon sequences only. Again, we find that long gene conversions occur when the flanking similarity of the converted region is high and that lower flanking similarities, has in intron sequences, tend to stop the elongation of the converted region.

Consistent with previous studies, our results also suggest that cDNA molecules are often used as templates in the repair and subsequent gene conversion process (Fink 1987; Derr et al. 1991; Melamed et al. 1992; Derr and Strathern 1993; Drouin 2002). In fact, the distribution of converted regions within converted genes (Figure 3) is uniform throughout the whole length of theses genes with the exception of the last 10%. An excess of converted regions ($\chi^2 = 108.5$, $p = 2.9\text{x}10^{-19}$; Chi-square test) is found at the 3'end. This is likely caused by gene conversions between an incompletely reverse-transcribed cDNA molecule and a genomic gene copy. This hypothesis is further corroborated by the distribution of converted intron regions (Figure 3). In fact, no significant excesses in any gene regions where detected when considering only conversions which extended into introns. This result is therefore consistent with the hypothesis that the excess of 3'converted regions are caused by the conversion between an incomplete cDNA molecule and a genomic copy. During the maturation of an mRNA molecule, its introns sequences are removed, its exon sequences are spliced together, and after a few other modifications, a mature mRNA molecule is formed. Due to the poor processivity of the reverse-transcriptase (Vanin 1985) this mature mRNA molecule is then reverse transcribed into an incomplete cDNA molecule. This incomplete cDNA molecule can then partake in the gene conversion event. Since such cDNA molecules lack intron sequences, they can only convert the 3' coding regions thus creating the bias observed here. In contrast, conversions of intron sequences require that the conversions occur between two chromosomal sequences.

So far, we only discussed the influence of the DNA repair mechanism on gene conversions. However, another factor that affects the frequency of gene conversions is the rate at which double

stranded breaks occur in a given region of a chromosome. The study of Jensen et al. (2004) found that the rate of local recombination on a chromosome is not uniform. Using these local rates of recombination, we found that there is a significant correlation ($\rho^2 = 0.293$, $p = 2.65\text{x}10^{-12}$; Spearman rank correlation test) between the frequency of gene conversion events and the rate of recombination. This suggest that genes in regions where double stranded breaks occur often (recombination hot spots) are more likely to undergo gene conversions than genes that are in regions where double stranded breaks occur less often (cold spots).

A comparative study of a trypsine gene family in *D .melanogaster* and *D. erecta* (Wang et al. 1999) found that the members of this family that where oriented in opposite transcriptional directions, converted at a higher frequency than members oriented in the same transcriptional direction. Contrary to this, pairs of human genes demonstrating intra-chromosomal conversion activity exhibit the opposite orientation bias. In fact, the frequency of gene conversions is higher for these genes when they share the same transcriptional orientation. This entails that, for an intra-chromosomal conversion event to occur between genes that share the same transcriptional orientation, the chromosome on which these pairs of genes reside must undergo a full 360° loop, which then allows the proper alignment of these pairs of genes for subsequent conversions. Pairs of genes that are in an opposite transcriptional orientation have their chromosome fold in a 180° loop to assure proper alignment between these pairs of genes and thus allow intra-chromosomal conversion events to occur. In addition, this transcriptional orientation bias is absent in pairs of gene that have undergone inter-chromosomal conversions. Indeed, ~50% of these genes share the same transcriptional orientation whereas the other ~50% share a different transcriptional orientation. A possible explanation of this orientation bias between pairs of genes that have undergone intra-chromosomal conversions is that these duplicate pairs of genes (paralogous genes) originated from a segmental duplication that usually maintains duplicated genes in the same transcriptional orientation. In addition to this, the flanking regions of these genes can also be duplicated. This should facilitate the frequency of conversion events between these duplicated sequences because of

the presence of large regions with high flanking sequence similarity.

Comparing the characteristics of human gene conversions with that of other organisms (bacteria, yeast and worm) brings us to the conclusion that the molecular processes involved in gene conversions could be inherent to most if not all organisms. For instance, the average size of a human gene conversion is 371 ± 752 (± standard deviation) nucleotides and does not differ from the average size found in *S. cerevisiae* 173 ± 220 nucleotides ($z = -0.25$, $p = 0.8$; Drouin 2002), *E-coli* K-12; 483 ± 890 nucleotides ($z = 0.10$, $p = 0.92$; Morris and Drouin 2004) and *C.elegans* 117 ± 205 nucleotides ($z = 1.96$, $p = 0.74$; Semple and Wolfe 1999). However these organisms have different frequencies of gene conversions. For humans we estimated it at 1.12% (number of gene conversions / total number of genes comparison), which is much lower than in *S.cerevisiae* (7.8%, [Drouin 2002]) and *E. coli* K-12 (7.6% [Morris and Drouin 2004]). This difference could be due to the fact that human genes have long, and often very dissimilar, intron sequences whereas both yeast and *E. coli* are essentially devoid of intron sequences. Furthermore, a similar significant positive correlation is observed for human ($r^2 = 0.34$, $p = 5.32 \times 10^{-15}$), *S.cerevisiae* ($r^2 = 0.17$ [Drouin 2002]) and in 4 different *E-coli* genomes (K-12; $r^2 = 0.0028$, CFT073; $r^2 = 0.13$, EDL933; $r^2 = 0.15$, and Sakai; $r^2 = 0.11$ [Morris and Drouin 2004]) between the number of gene conversions and the number of members in a gene family for humans. This suggests that the mechanisms employed to repair damage DNA, and thus create a gene conversion, is similar in a broad range of organisms.

In summary, this genome wide survey identified factors affecting gene conversions such as the rate of local recombination, flanking similarity, orientation and the proximity of converted human genes. Also, the characteristics of human gene conversions are similar to does found in other in other organisms. This analysis also suggests that gene conversions can occur between a genomic sequence and an incomplete cDNA molecule.

Literature Cited

Bosch, E., Hurles, M.E., Navarro, A., and Jobling, M.A. 2004. Dynamics of a human interparalog gene conversion hotspot. Genome Research **14**:835-844.

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

Derr, L.K., Strathern, J.N., and Garfinkel, D.J. 1991. RNA-mediated recombination in *S. cervisiae.* Cell **67**:355-364.

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

Fink, G.R. 1987. Pseudogenes in yeast? Cell **49:**5-6.

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

Haber, J.E., Leung, W-Y., Boris, R.H., and Lichten, M. 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.

Hurles, M.E. 2001. Gene conversion homogenizes the CMTIA paralogous repeats. BMC Genomics **2:**11-20.

International human genome sequencing consortium. 2001. Initial sequencing and analysis of the human genome. Nature **409**:860-921.

Jeffreys, A.J., and May, C.A. 2004. Intense and highly localized gene conversion activity in human meiotic crossover hot spots. 2004. Nature Genetics **36**(2):151-156

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

Lilley, D. M. J., and White, M.F. 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.

Lloyd, R.G., and Brooks Low, K. 1996. Escherichia coli and Salmonella. Cellular and molecular biology. ASM press, Washington, D.C..

Melamed, C., Nevo, Y., and Kupiec, M. 1992. Involvement of cDNA in homologous recombination between Ty elements in *Saccharomyces cerevisea.* Mol. Cell. Biol. **12**:1613-1620.

Mitchell, M.B. 1955. Aberrant recombination of pyridoxine mutants of *Neurospora.* Proc.Natl. Acad. Sci. USA **41**:215-220.

Modrich, C., and Lahue, R. 1996. Mismatch repair in replication fidelity, genetic recombination and

cancer biology. Annu. Rev. Biochem. **65**:101-133.

Morris, R.T., and Drouin, G. 2004. Ectopic Gene Conversions in Four Escherichia coli Genomes: Increased Recombination in Pathogenic Strains. J. Mol. Evol. **58**:596-605.

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

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

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

Santoyo, G., and Romero, D. 2005. Gene conversion and concerted evolution in bacterial genomes. Bacterial Genomics. **29**:169-183.

Saxonov, S., Daizadeh, I., Fedorov, A., and Gilbert, W. 2000. An exhaustive database of protein-coding intron-containing genes. Nucleic Acids Res. **28**:185-190.

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

Slightom, J.L., Blechl, A.E., and Smithies, O. 1980. Human fetal G gamma- and A gamma-globin genes: complete nucleotide sequences suggest that DNA can be exchanged between these duplicated genes. Cell **21**:627–638.

Vanin, E.F. 1985. Processed pseudogenes: Characteristics and evolution. Annu. Rev. Genet. **19:** 253-272.

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

Zhang, H., Hu, J., Recce, M., and Tian, B. 2004. PolyA_DB: a database for mammalian mRNA polyadenylation. Nucleic Acids Res. **33**:D177.

Table 1. Relative position of excess gene conversions on particular chromosomes.

Chromosome	Relative position of excess[a]	X^2 value	*p*-value
1	0.1	87.11	6.17E-15
3	0.7	28.80	7.00E-04
5	0.8	133.60	2.86E-24
6	0.2	102.72	4.42E-18
9	0.6	23.35	5.46E-03
12	0.4	112.30	5.01E-20
14	0.4	34.53	7.21E-05
15	0.3	118.35	2.90E-21
17	0.2	36.54	3.17E-05
17	0.5	142.63	2.95E-26
19	0.1	65.99	9.26E-11
19	0.3	88.50	3.25E-15
21	1	41.34	4.33E-06
X	0.4	24.01	4.28E-03

Note.[a]Relative position of excess represents the bin in which an excess of converted genes was found. These bins are defined by dividing the length of a particular chromosome into 10 regions of equal lengths. Bins showing a significant excess ($p \leq 0.05$) of converted regions when compared to the expected value for a uniform distribution are shown in this table.

Figure 1.

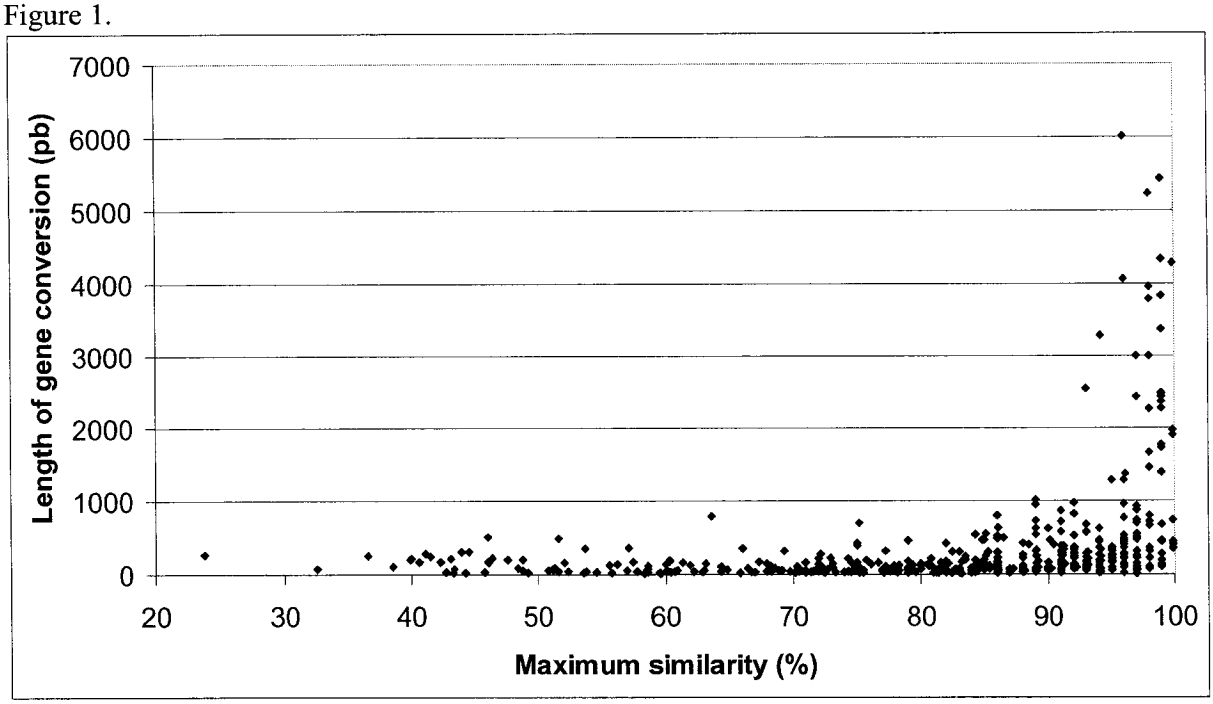


Fig. 1. Relationship between the lengths of each converted region and the maximum similarity of its flanking regions.

Figure 2.

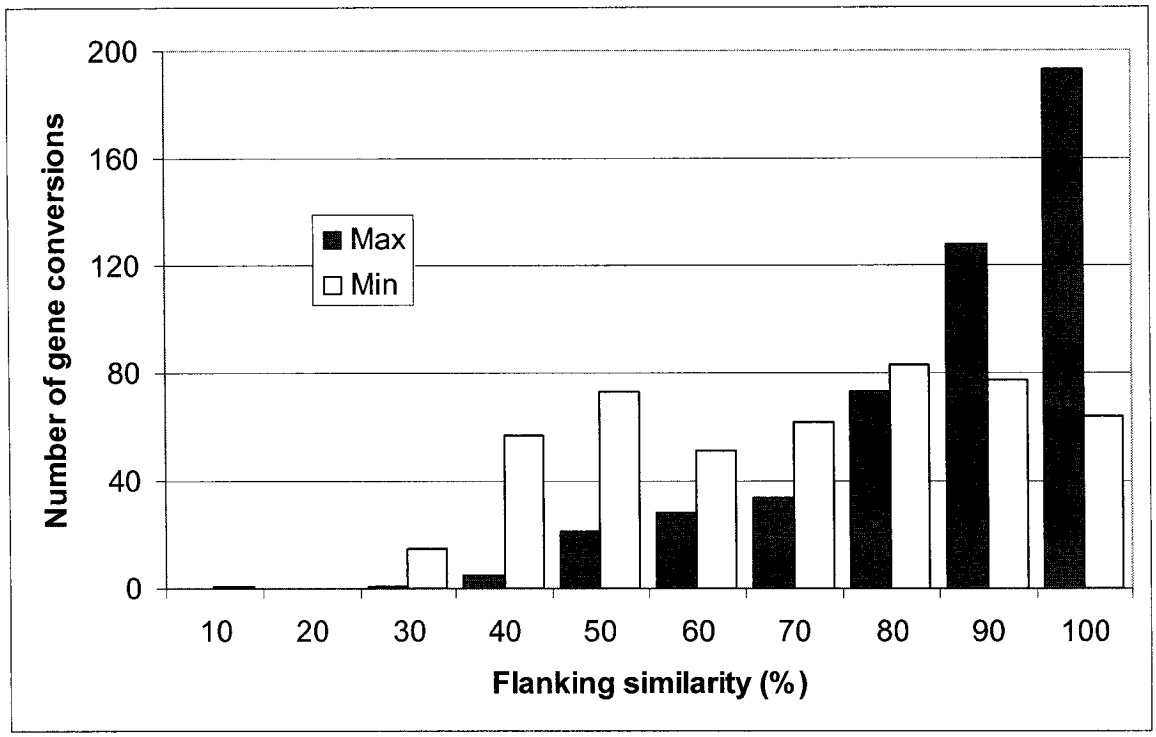


Fig. 2. Distribution of the number of gene conversions plotted against the maximum and minimum flanking similarity of the converted regions.

Figure 3.

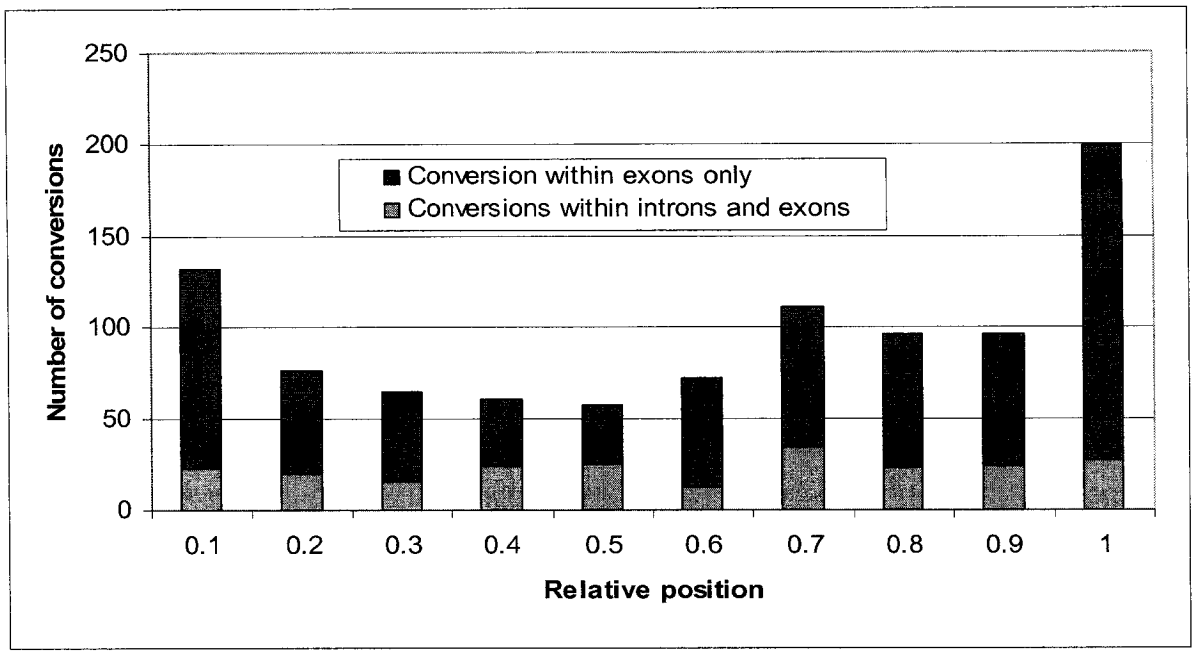


Fig. 3. Distribution of converted regions along the length of the genes (n = 966 genes). The middle coordinate of the converted region is divided by the total length of the gene and is then assigned to its relative position within that same gene. The distribution of conversions where the converted region covers only the exon sequence of a gene and the distribution of converted regions that cover both exon and intron sequences are shown.

Figure 4

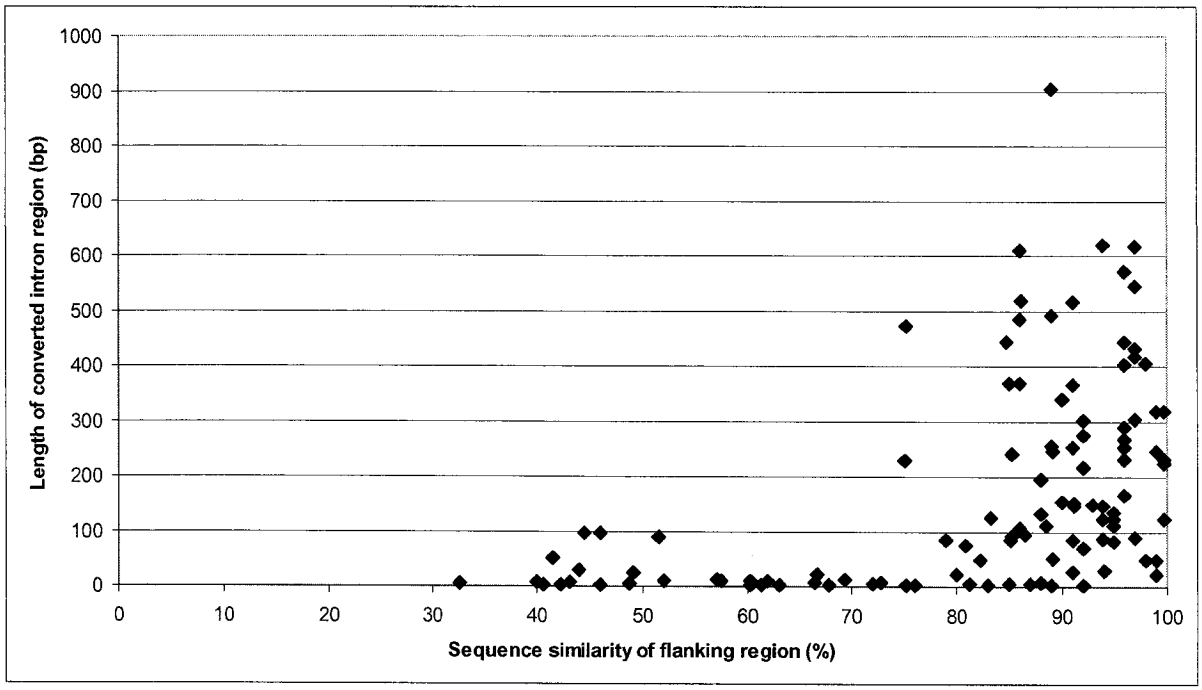


Fig 4. Relationship between the length of converted intron regions and the maximum sequence similarity of the flanking region.

Figure 5.

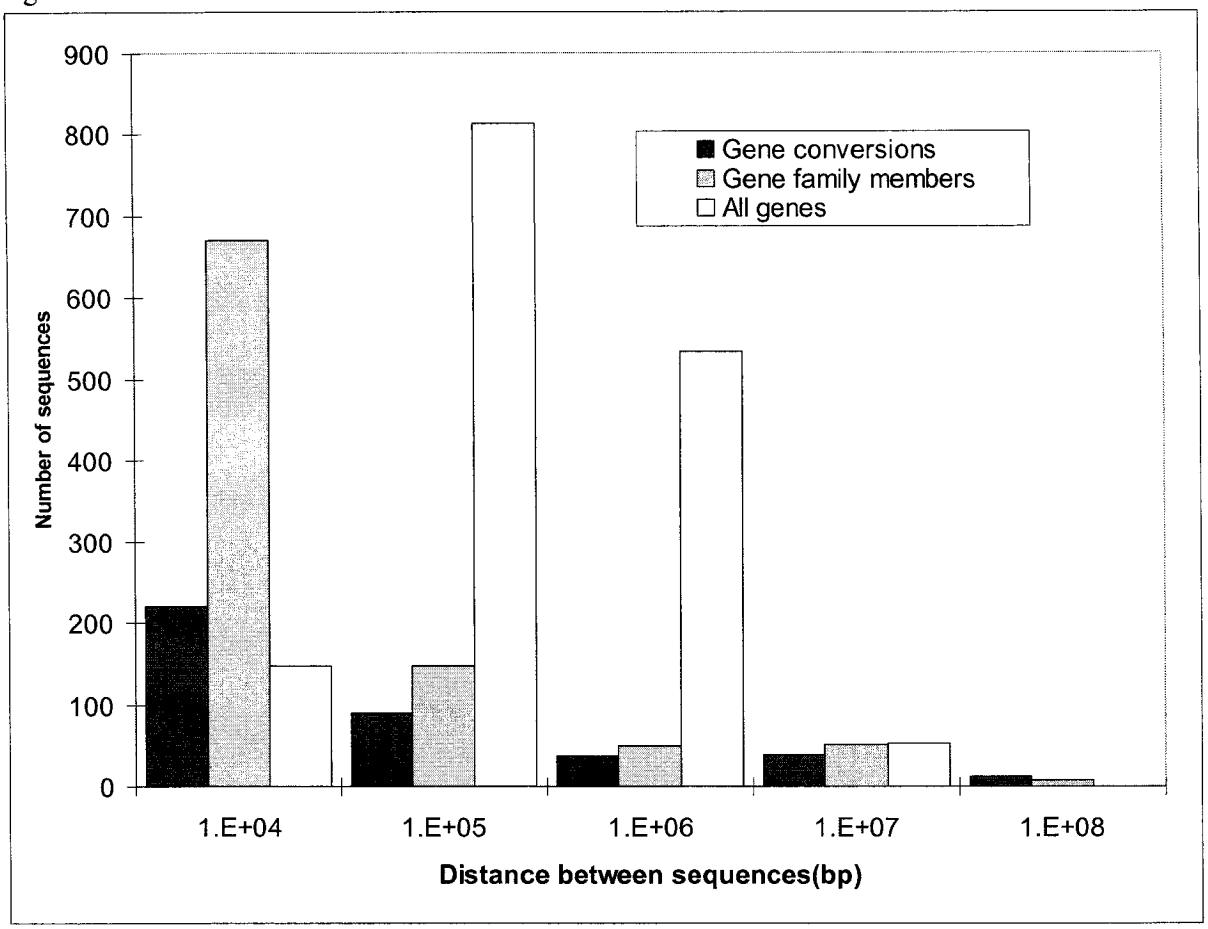


Fig. 5. Distance between genes implicated in a conversion event, adjacent gene family members and all genes from the human genome. Note that in order to fit all data points on the same Y axis, the number of the "All genes" category was divided by 10.

Chapter 3

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

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

Published in: Benovoy D., R.T. Morris, A. Morrin and G., Drouin. 2005. Ectopic gene conversions increases the GC content of duplicated yeast and Arabidopsis genes. Mol. Biol. Evol. **22**:1865-1868.

Contributions: I contributed in the analyses presented in Figures 2 and 3 of this paper.

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.

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 increase 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 a 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 found 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 x 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 conversion 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 gene having more than 88% similarity, why this correlation is limited to gene 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 small 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 ($\rho^2 = 0.156$) 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 inbreed species would be homozygote 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.

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.

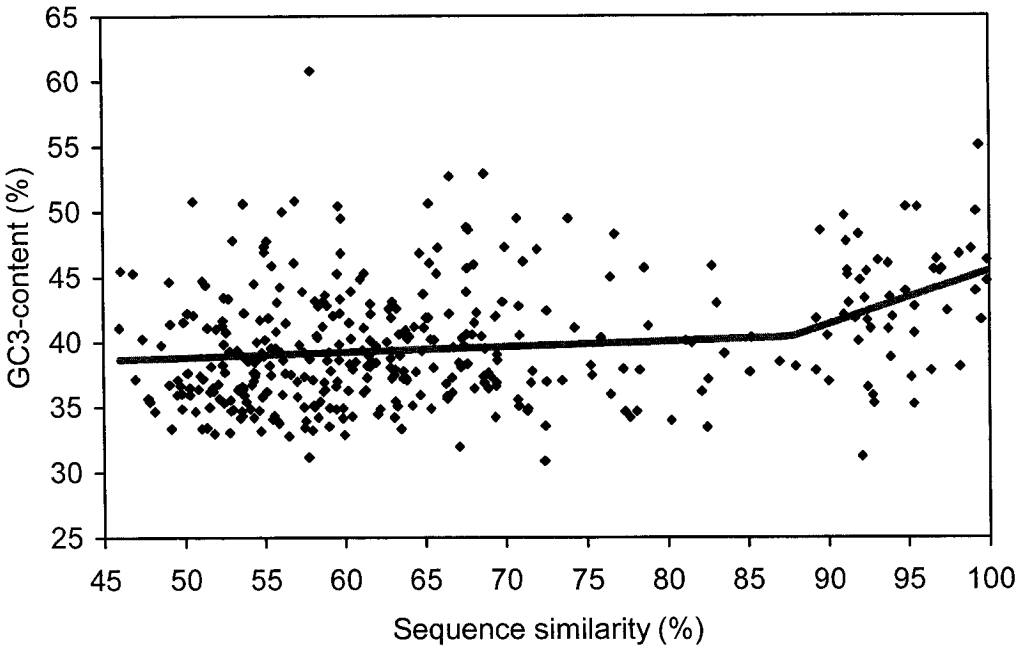

65
60
55
50
45
40
35
30
25
GC3-content (%)
45
50
55
60
65
70
75
80
85
90
95
100
Sequence similarity (%)

Figure 2.

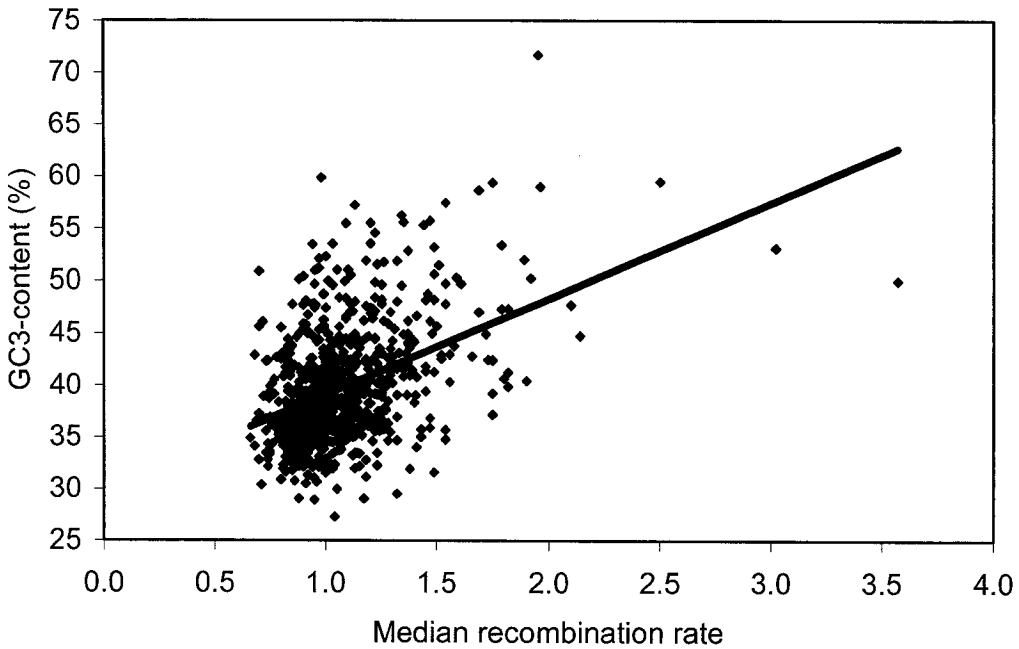
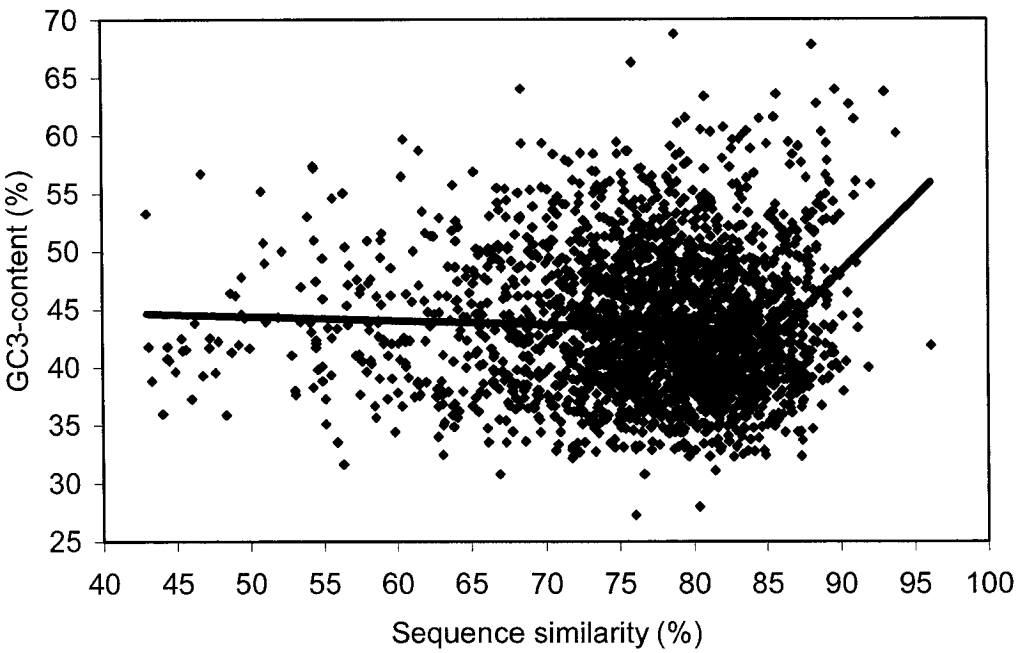

GC3-content (%)
75
70
65
60
55
50
45
40
35
30
25
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Median recombination rate

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 4

General conclusion

Motivations

The complete sequencing of the human, mouse, rat, *Arabidopsis*, chicken, chimpanzee, yeast and many bacterial genomes, to name just a few, is a remarkable achievement. What drives molecular biologist to sequence all these organisms? Taken separately, these genome sequences only forgo some information about the organisms in question but comparisons with other genomes yields tremendous insights into the mechanisms of evolution.

Evolution never makes things simple for biologists. They cannot just line up, for instance, the yeast and human genomes starting at one end of a chromosome and expect to find matching regions one after another. On the time scale of evolution, the process of recombination i.e., the genetic equivalent of cut-and-paste, is continually at work rearranging genomes. In this study, we examined one possible outcome of homolgous recombination, ectopic gene conversion.

What affects ectopic gene conversions?

In the first part of this study (Chapter 2), the human genome was surveyed for ectopic gene conversions using Sawyer's GENCONV method for the detecting gene conversions in multigene families. This allowed us to identify 483 pairs of genes that have undergone ectopic gene conversions. From this, we went on to identify factors affecting ectopic gene conversions in the human genome and compared their characteristics with ectopic gene conversions found in other organisms.

We found that the similarity in the flanking regions of the converted genes plays an important role in dictating the length of the converted regions. Presumably, it does this by stabilizing the Holliday structure, thus enabling longer ectopic conversions. Proximity between the

pairs of converted genes is another important factor affecting gene conversions. In fact, for intra-chromosomal gene conversions, we find that the closer genes are from each other, the easier it is for proteins involved in the conversion process to recruit the intact member of the pair and use it as a template sequence. Therefore, closer paralogous genes convert with a higher frequency. The rate of local recombination is another factor that influences ectopic gene conversions. The higher the recombination rate, the higher chance of there being a conversion event between the recombining sequences. Transcriptional orientation also affects the frequency of gene conversion in that pairs of genes sharing the same orientation will convert at a higher frequency than pairs of genes in opposite transcriptional orientation. This is probably due to a characteristic of segmental duplication where genes that are duplicated usually conserve the same transcriptional orientation. In addition, an excess of converted regions was found at the 3' end of genes. This suggests that an incomplete cDNA molecule is often used as a template during the conversion process. From this analysis of the factors and mechanisms affecting human ectopic gene conversions we conclude that the majority of these influences are a direct consequence of the constraints put forth by the formation and resolution of the Holliday structure.

Following this analysis of human ectopic gene conversions a second question is apropos: how does this compare to other organisms? When comparing characteristics of human ectopic gene conversions with that of others organisms we find them to be similar. Quantitative characteristics such as size of conversions and flanking region similarity are found to be similar between many different species such as yeast, bacteria and humans. In addition, we find that the same factors mentioned above affect ectopic gene conversions in many different organisms. The only difference we find is that human gene conversions are less frequent than in yeast and *E. coli*. This difference could be due to the presence of often very divergent intron sequences in human genes which decrease the frequency of gene conversions between duplicated genes. From this we can conclude that the mechanisms controlling ectopic gene conversions must be alike in a broad range of species, possibly even in every organisms.

The effect of ectopic gene conversions on GC-content

The second part of this study (Chapter 3) examined the effect that ectopic gene conversions have on yeast and *Arabidopsis* dispersed duplicated genes (ohnologs).

We showed that gene conversions between pairs of ohnologs with more than 88% sequence similarity lead to an increase in the GC-content of these genes. An explanation for this is that the repair mechanism is biased towards the incorporation of guanine and cytosine nucleotides whenever there is a mismatch between the template strand and the strand being repaired. For example, among the genes undergoing concerted evolution in mammals, the well-known ribosomal operons, transfer RNAs and histones are all GC rich as a consequence of gene conversions (Galtier et al. 2001). Although recombination has been shown to increase the GC-content in the human genome (Fullerton 2001) we did not analyze the effect of ectopic gene conversions on the GC-content of human multigene family members because these members have all originated at different times in the human lineage. This type of study was easier to perform using yeast and *Arabidopsis* sequences because all multigene family members were duplicated in a single event (i.e., a genome duplication event).

Cited Literature

Fullerton, S.M., A.B., Carvalho and A.G. Clark. 2001. Local rates of recombination are positively correlated with GC content in the human genome. Mol. Bio. 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.

Appendices

1. Published version of chapter 3.

2. CD-ROM: Contains the scripts used for the analysis of gene conversions in the human genome as well as Excel files used for statistical analyses.