The evolution of recombination and genomic structures : a modelling approach

Alexandra-Mariela Dumitru Popa

N° d'ordre: 87-2011 Année 2011

THESE

présentée devant

l'UNIVERSITE CLAUDE BERNARD - LYON I

pour l'obtention

du DIPLOME DE DOCTORAT

(arrêté du 7 août 2006)

par

Alexandra Mariela POPA

The evolution of recombination and genomic structures: a modeling approach.

Directeur de thèse: **Christian GAUTIER**

Co-directrice de thèse: **Dominique MOUCHIROUD**

JURY:	**Laurent DURET**	Président du jury
	Christian GAUTIER	Directeur
	Sylvain GLEMIN	Rapporteur
	Christine MEZARD	Rapporteur
	Dominique MOUCHIROUD	Directrice
	Matthew WEBSTER	Rapporteur

UNIVERSITE CLAUDE BERNARD - LYON 1

Président de l'Université	**M. A. Bonmartin**
Vice-président du Conseil d'Administration	M. le Professeur G. Annat
Vice-président du Conseil des Etudes et de la Vie Universitaire	M. le Professeur D. Simon
Vice-président du Conseil Scientifique	M. le Professeur J-F. Mornex
Secrétaire Général	M. G. Gay

COMPOSANTES SANTE

Faculté de Médecine Lyon Est – Claude Bernard	Directeur : M. le Professeur J. Etienne
Faculté de Médecine et de Maïeutique Lyon Sud – Charles Mérieux	Directeur : M. le Professeur F-N. Gilly
UFR d'Odontologie	Directeur : M. le Professeur D. Bourgeois
Institut des Sciences Pharmaceutiques et Biologiques	Directeur : M. le Professeur F. Locher
Institut des Sciences et Techniques de la Réadaptation	Directeur : M. le Professeur Y. Matillon
Département de formation et Centre de Recherche en Biologie Humaine	Directeur : M. le Professeur P. Farge

COMPOSANTES ET DEPARTEMENTS DE SCIENCES ET TECHNOLOGIE

Faculté des Sciences et Technologies	Directeur : M. le Professeur F. Gieres
Département Biologie	Directeur : M. le Professeur F. Fleury
Département Chimie Biochimie	Directeur : Mme le Professeur H. Parrot
Département GEP	Directeur : M. N. Siauve
Département Informatique	Directeur : M. le Professeur S. Akkouche
Département Mathématiques	Directeur : M. le Professeur A. Goldman
Département Mécanique	Directeur : M. le Professeur H. Ben Hadid
Département Physique	Directeur : Mme S. Fleck
Département Sciences de la Terre	Directeur : Mme le Professeur I. Daniel
UFR Sciences et Techniques des Activités Physiques et Sportives	Directeur : M. C. Collignon
Observatoire de Lyon	Directeur : M. B. Guiderdoni
Ecole Polytechnique Universitaire de Lyon 1	Directeur : M. P. Fournier
Ecole Supérieure de Chimie Physique Electronique	Directeur : M. G. Pignault
Institut Universitaire de Technologie de Lyon 1	Directeur : M. le Professeur C. Coulet
Institut de Science Financière et d'Assurances	Directeur : M. le Professeur J-C. Augros
Institut Universitaire de Formation des Maîtres	Directeur : M. R. Bernard

UNIVERSITE CLAUDE BERNARD - LYON 1

Président de l'Université	**M. A. Bonmartin**
Vice-président du Conseil d'Administration	M. le Professeur G. Annat
Vice-président du Conseil des Etudes et de la Vie Universitaire	M. le Professeur D. Simon
Vice-président du Conseil Scientifique	M. le Professeur J-F. Mornex
Secrétaire Général	M. G. Gay

COMPOSANTES SANTE

Faculté de Médecine Lyon Est – Claude Bernard	Directeur : M. le Professeur J. Etienne
Faculté de Médecine et de Maïeutique Lyon Sud – Charles Mérieux	Directeur : M. le Professeur F-N. Gilly
UFR d'Odontologie	Directeur : M. le Professeur D. Bourgeois
Institut des Sciences Pharmaceutiques et Biologiques	Directeur : M. le Professeur F. Locher
Institut des Sciences et Techniques de la Réadaptation	Directeur : M. le Professeur Y. Matillon
Département de formation et Centre de Recherche en Biologie Humaine	Directeur : M. le Professeur P. Farge

COMPOSANTES ET DEPARTEMENTS DE SCIENCES ET TECHNOLOGIE

Faculté des Sciences et Technologies	Directeur : M. le Professeur F. Gieres
Département Biologie	Directeur : M. le Professeur F. Fleury
Département Chimie Biochimie	Directeur : Mme le Professeur H. Parrot
Département GEP	Directeur : M. N. Siauve
Département Informatique	Directeur : M. le Professeur S. Akkouche
Département Mathématiques	Directeur : M. le Professeur A. Goldman
Département Mécanique	Directeur : M. le Professeur H. Ben Hadid
Département Physique	Directeur : Mme S. Fleck
Département Sciences de la Terre	Directeur : Mme le Professeur I. Daniel
UFR Sciences et Techniques des Activités Physiques et Sportives	Directeur : M. C. Collignon
Observatoire de Lyon	Directeur : M. B. Guiderdoni
Ecole Polytechnique Universitaire de Lyon 1	Directeur : M. P. Fournier
Ecole Supérieure de Chimie Physique Electronique	Directeur : M. G. Pignault
Institut Universitaire de Technologie de Lyon 1	Directeur : M. le Professeur C. Coulet
Institut de Science Financière et d'Assurances	Directeur : M. le Professeur J-C. Augros
Institut Universitaire de Formation des Maîtres	Directeur : M. R. Bernard

Abstract

Meiotic recombination plays several critical roles in molecular evolution. First, recombination represents a key step in the production and transmission of gametes during meiosis. Second, recombination facilitates the impact of natural selection by shuffling genomic sequences. Furthermore, the action of certain repair mechanisms during recombination affects the frequencies of alleles in populations via biased gene conversion. Lately, the numerous advancements in the study of recombination have unraveled the complexity of this process regarding both its mechanisms and evolution.

The main aim of this thesis is to analyze the relationships between the different causes, characteristics, and effects of recombination from an evolutionary perspective. First, we developed a model based on the control mechanisms of meiosis and inter-crossover interference. We further used this model to compare the recombination strategies in multiple vertebrates and invertebrates, as well as between sexes. Second, we studied the impact of the sex-specific localization of recombination hotspots on the evolution of the GC content for several vertebrates. Last, we built a population genetics model to analyze the impact of recombination on the frequency of deleterious mutation in the human population.

Résumé

La recombinaison méiotique joue un double rôle de moteur évolutif en participant à la création d'une diversité génétique soumise à la sélection naturelle et de contrôle dans la fabrication des gamètes lors de la méiose. De plus, en association avec certains mécanismes de réparation, la recombinaison, au travers de la conversion génique biaisée manipule les fréquences alléliques au sein des populations. Les connaissances sur le fonctionnement même de ce processus ont considérablement augmenté ces dernières années faisant découvrir un processus complexe, autant dans son fonctionnement que dans son évolution.

Le thème général de la thèse est l'analyse, dans un contexte évolutif, des relations entre les différents rôles et caractéristiques fonctionnelles de la recombinaison. Un modèle de la recombinaison prenant en compte des contraintes liées au contrôle de la méiose et le phénomène d'interférence a permis une comparaison entre espèces au sein des vertébrés et des non-vertébrés de même qu'une comparaison entre sexes. Par ailleurs, nous avons montré l'impact de la localisation spécifique aux sexes des points chauds de recombinaison sur l'évolution du contenu en GC des génomes de plusieurs vertébrés. Finalement, nous proposons un modèle à l'échelle de la génétique des populations, permettant d'analyser l'impact de la recombinaison sur la fréquences de mutations délétères dans les populations humaines. Cette thèse, nous l'espérons, apportera sa pierre à l'étude interdisciplinaire de la recombinaison, à la fois au sein de la biologie et par ses relations au travers de la modélisation avec l'informatique et les mathématiques.

Acknowledgments

Doing a PhD was not my dream as a little girl, neither was it my intention at the end of the 5 years of engineer studies. And to sum-up I don't regret it. Maybe there are many things I would do differently if I could but I am not sure. And if I am here it's because of all those people that have helped me overcome all the difficulties and lack of self-confidence. I would have to write another 200 pages to thank them all and I fear I don't have the energy anymore. But I would like to mention groups of people, rather than names (of course with a few exceptions) that helped me, through scientific, administrative and personal challenges, and to whom I am greatly indebted.

I am heartily thankful to my supervisors, Christian Gautier and Dominique Mouchiroud. Even today I fail to understand the confidence Christian had to take me as a Master student. He had numerous reasons not to but he gave me a chance. When I first came to see you, Christian, was just to test my luck and get some information. Never did I imagine you would give me an internship. Five years later, I thank you sincerely for all the things that you taught me. It wasn't easy, and I think I wasn't very easy to deal with either, but your scientific rigor challenged me every time and taught me one first important thing: "Don't take anything for granted!" Also your voluntary and involuntary encouragement of my independence had two outcomes: the improving of my literary skills through numerous reports (essential to communicate) and scientific autonomy (there is still place for improvement here). An equally important part, for both my scientific and personal development, was played by Dominique. I don't think the subject of this thesis would have evolved in this direction without your help Dominique. While you offered me the same scientific rigor as Christian, your approaches were very often complementary. Finding the middle way was not always easy, but it was most surely enriching. I want to thank you, Dominique, in particular for the many interesting ideas that you encouraged me to pursue. You have also taught me how to organize and present my results clearly which was of essential help for the writing of the thesis and its presentation.

I owe my deepest gratitude to the members of the defense jury, who have honored me by reading and evaluating this work. Their invaluable remarks, as well as their critical perspectives have improved this work and shed new light on many issues.

I said I wouldn't mention names, but there are still a few persons that I have to thank in particular. I would like to thank Laurent Duret. I thank you, Laurent, for the time you spent explaining to me the biased gene conversion and for the opportunity you gave me to collaborate on your article with Anuk. I must admit that there is a mixture of fear and admiration that I have every time I speak to you. And talking about Anuk, well to be honest I think there was a moment during my first year of PhD when I was sincerely considering quitting. If it wasn't for you dear Anuk, I don't know if I would have found

the energy to go on. You are a friend as no other, and furthermore you are a role model for me.

Special thanks go to AnneSoso for all the time we spent working together on our manuscripts. And I would also like to thank all my office colleagues for the good time and scientific and technical support. I am indebted to all the members of the LBBE laboratory for offering such a strong scientific and human environment. The numerous seminars as well as the availability of all the members of the lab for scientific discussions offer continuous fuel for innovative research. But equally important is that all this takes place with a cake and a coffee and a lot of laughs afterwards.

Last, but certainly not least, I would like to express my deepest gratitude to my family. My parents have made many sacrifices for me to come to France and to have the possibility to study here. So I would like to thank them now for their trust. I want to thank you Emil for believing in me. You are my main source of self-confidence. Thank you for making the effort to listen to stories about biased gene conversion and recombination, and for your unconditional help. And not last, thank you for Vlad, this amazing little kid. People say it is difficult to write a PhD and have a newborn. It is, that's for sure. But it has some good parts too. Thank you Vlad for understanding that mom was stressed sometimes, for being such a calm and smiling child. Thank you for teaching me to be more organized and giving me some perspective on things.

Notations

General abbreviations

A	Adenine
bp	base pair
C	Cytosine
CpG	A dinucleotide CG, p standing for a phosphate link.
DNA	Deoxyribonucleic acid
G	Guanine
Gb	Giga base
kb	kilo base
Mb	Mega base
Myr	Million years
N_e	effective population size
SNP	single nucleotide polymorphism
T	Thymine
TSS	Transcription Start Site

Meiosis and recombination-related abbreviations

CE	Central Element (referring to the SC)
cM	centimorgan
CO	Crossover
COI	CO Interference
COR	Crossover Rate
dHJ	double Holliday Junction
DSB	Double-Strand Break
DSBh	Double-Strand Break hotspot
DSBR	Double-Strand Break Repair model
dsDNA	double stranded DNA
F1	First generation of offspring in a crossing experiment
F2	Second generation of offspring in a crossing experiment
F/M	Female/Male ratio
HapMap1, 2, and 3	the 1st, 2nd, and 3rd respective phases of HapMap Project
HJ	Holliday Junction
HR	Homologous Recombination

HS	Heterogeneous Stock (for mouse populations)
LD	Linkage Disequilibrium
LE	Lateral Element (referring to the SC)
NAHR	Nonallelic Homologous Recombination
NCO	Non-crossover
NCOR	Non-crossover Rate
NE	Nuclear Envelope
NHEJ	Nonhomologous End Joining
PC	Pairing Centres
rDNA	ribosomal DNA
RI	Recombinant Inbred lines (in a crossing experiment)
SC	Synaptonemal Complex
SDSA	Synthesis-Dependent Strand-Annealing model
SEI	Single End Invasion
ssDNA	single stranded DNA
TF	Transverse Filaments (referring to the SC)

Mathematical symbols

C	coefficient of coincidence
C_3	Three-point coefficient of coincidence.
D'	the difference between the frequency of a two locus haplotype and the product of the component alleles, divided by the most extreme possible value, given the marginal allele frequencies, measure of LD
g	genetic distance
I	Identity matrix
m	In the counting models, m stands for the number of NCO events that separate two consecutive COs. It is a measure of the strength of interference
P	Physical length (Mb) of an interval or chromosome
p	The fraction of COs that are not subject to interference under the two-pathway model Housworth and Stahl (2003).
Q	the substitution matrix along a branch of a phylogenetic tree
R	frequency of recombinants among the offspring
r^2	correlation of alleles at different loci, measure of LD
y	The mean number of DSB events in the counting model of Foss et al. (1993).
$\#$	Number
χ^2	Chi-square distribution
Γ	Gamma distribution
λ	The rate parameter for Γ
ν	The shape parameter for Γ

σ_0	Uniform basal tensile stress in the mechanical stress model of Kleckner et al. (2004).
$\otimes$	Tensor product of matrices

BGC abbreviations

BGC	Biased Gene Conversion
BER	base excision repair
gBGC	GC Biased Gene Conversion
GC*	equilibrium or stationary GC-content
MMR	mismatch repair

Other abbreviations

AIC	Akaike Information Criterion
BIC	Bayesian Information Criterion
CEPH	Centre d'Etude du Polymorphism Humain
CI	Confidence Interval
DAF	Derived Allele Frequency
DT	Distance to Telomeres
HGMD	Human Gene Mutation Database
H-W test	Hotteling-William's t-test
L	Likelihood
LCR	Low Copy Repeat
LDT	Log Distance to Telomeres
LINE	Long interspersed nuclear element
LOD	Logarithm of Odds
MHC	Major Histocompatibility Complex
PAR	Pseudoautosomal Region
PCR	Polymerase Chain Reaction
RE	Repetitive Element
SINE	Short Interspersed Nuclear Element
TE	Transposable Element

Definitions

Information assembled from Stumpf and McVean (2003); Arnheim et al. (2007); Lynch (2007); Paigen and Petkov (2010); DB-NCBI

Allele One of the variant forms of a DNA sequence at a particular locus, or location, on a chromosome.

Backcross Crossing experiment in which individuals in the first generation are crossed back with one or both their parents to obtain the second generation of offspring.

Bouquet formation The clustering of telomeres together on the nuclear membrane early in meiosis.

Centimorgan Unit of genetic distance between markers that lie close enough to one another so that 1% of the meiotic products will exhibit a crossover between them (in a single generation)

Chiasmata A chiasma (plural chiasmata) is the cytologically visible physical connection between homologous chromatids during meiosis that corresponds to the sites of genetic crossing over.

Chromatid The product of chromosome replication in meiosis I. Chromatids are distinguished from chromosomes by the fact that the two daughter chromatids of one chromosome remain attached at their centromeres through meiosis I cell division.

Crossover (CO) Recombination product consisting of a reciprocal exchange of DNA sequences, usually between a pair of homologous chromosomes

Cytokinesis The division of the cytoplasm between two daughter cells following nuclear division.

Diploid Having two gene copies at a genetic locus; as in virtually all animals and land plants.

Double-strand break (DSB) Cleavage of both strands of a DNA molecule at a specific site.

Effective Population Size (Ne) Represents the size of an ideal population (identical individuals, random mating, no overlapping generations) accounting for realistic demographic and structure features. It determines the rate of change in the composition of a population caused by genetic drift.

Bottleneck A temporary marked reduction in population size.

equilibrium GC-content (GC*) A statistic resuming the matrix of substitutions. It is the GC-content reached by a sequence under a constant substitution pattern. $GC* = \frac{AT \to GC}{AT \to GC + GC \to AT}$

F2 intercrosses Crossing experiment in which the F2 mapping population is produced by intercrossing F1 individuals.

Four-gamete test If all four possible gametes are observed for two bi-allelic loci then this test infers that a recombination event must have occurred between them (under an infinite sites mutation model).

Gene conversion The process by which one participant in a recombination event is converted to the sequence of the partner participant; occurs during almost all recombination events, but not necessarily associated with cross-over

Genetic distance Distance between DNA markers on a chromosome measured as the amount of crossover between them. A genetic map is an ordered list of markers along the chromosome and the intermarker genetic distances.

Genetic drift The change in the frequency of a gene variant (allele) in a population due to random sampling.

Genetic interference The presence of a recombinational event in one region that affects the occurrence of recombinational events in adjacent regions. Positive interference, which is seen in eukaryotes, reduces the probability of using nearby hotspots in the same meiosis and causes a more even spacing of crossover than would occur by chance.

Genotyping The process by which DNA is analyzed to determine which genetic variant (allele) is present for a certain marker.

haploid Having a single gene copy at a genetic locus; as in all prokaryotes, germ cells, and some unicellular eukaryotes.

Haplotype A set of genetic markers that are present on a single chromosome and that show complete or nearly complete linkage disequilibrium - that is, they are inherited through generations without being changed by crossing over or other recombination mechanisms.

Hardy-Weinberg equilibrium Both allele and genotype frequencies in a randomly-mating population remain constant across generations, unless specific disturbing influences are introduced.

Holliday junction The point at which the strands of the two dsDNA molecules exchange partners as an intermediate step in crossing over.

Infinite sites mutation model A model that assumes that there are an infinite number of nucleotide sites and consequently that each new mutation occurs at a different locus.

Linkage disequilibrium (LD) The nonrandom association of alleles at two or more loci.

Mutational load Represents a reduction of the mean fitness of a population subsequent to mutations accumulation.

Non-crossover (NCO) Recombination product consisting in the swap of small DNA segment

Panmixia Random mating.

Physical distance Distance between DNA markers on a chromosome measured in the number of nucleotide base pairs. A physical map is an ordered list of markers along a chromosome and the inter-marker physical distances

Polymerase chain reaction (PCR) PCR is a technique that amplifies a specific region of DNA as defined by two primer sequences. It is a very useful technique as it generates many copies of one specific genetic material, and thus uses very small amounts of DNA as starting material. PCR is a three stage process: DNA is denaturated (made single stranded), then the primers bind or anneal to their complementary sequence, and in the end, the primers are extended by the addition of nucleotides complementary to that on the template sequences. This process is repeated multiple times. The end result is amplification of the sequence between and including the primer sequence.

Positive selection A process by which natural selection favors a single beneficial genotype over other genotypes and may drive this genotype to a high frequency in a population.

Pseudoautosomal A region on a sex chromosome that is homologous between the X and Y chromosomes. Successful meiosis in males requires a crossover in this region.

Recombinant inbred (RI) lines Crossing experiment in which inbred recombinant lines are obtained from an F1 generation resulting from a cross between parents homozygous at every locus.

Recombination Exchange of DNA sequence information within or between chromosomes.

Recombination nodules The early, visible manifestations of sites of chiasmata and crossovers. They are recognized by immunochemical staining, typically for the proteins of late recombination nodules.

Single-nucleotide polymorphism (SNPs) SNPs distinguish the chromosomes of two

individuals or strains. There are millions of SNPs in mammalian genomes, and they have become the preferred markers for genetic studies.

Synaptonemal complex A linear protein complex that forms the backbone of each chromatid during prophase I of meiosis and promotes genetic recombination. The DNA of the chromatid is attached to the complex in long loops. The name is derived from the word synapsis, which has been used to describe chromatid pairing.

Three-point coefficient of coincidence (C_3) The coefficient of coincidence calculated in a pair of adjacent intervals.

Zinc finger A protein loop in which cysteine or cysteine-histidine residues coordinate a zinc ion to form the base of the loop. Three of the amino acids in the loop cooperate to recognize three base pairs of DNA, and a tandem array of zinc fingers can show considerable DNA-binding specificity.

Contents

Préambule

Dès le passage à l'agriculture et à l'élevage des animaux, les hommes ont entamé les premières expériences génétiques, en étudiant et manipulant la transmission des caractères à la descendance. Mais ce n'est qu'au 19ème siècle qu'un support génétique a été identifié pour les caractères, support que Gregor Mendel a nommé "facteurs héréditaires". Il a découvert que l'expression d'un caractère chez un individu est régulée par une paire de facteurs, un provenant du père et le deuxième de la mère. Une des lois énoncée par Mendel affirme que les différents caractères sont hérités indépendamment les uns des autres. Des expériences ultérieures ont remis en question la disjonction indépendante des caractères. Ces patrons héréditaires inhabituels, quand certains caractères ségrègent ensemble plus souvent qu'attendu, a donné la définition de la **liaison génétique**. Thomas Morgan a associé la liaison entre les facteurs à leur appartenance à un même chromosome et a rapporté la force de cette liaison à la distance qui sépare les facteurs. Toutefois, certains facteurs montrent des niveaux différents de liaison : entre ségrégation indépendante et liaison complète. Morgan a suggéré que la liaison entre des facteurs appartenant à un même chromosome peut être brisée par la recombinaison lors de la méiose, à travers les chiasmata. Les chiasmata sont les sites visibles de l'échange de matériel génétique entre les chromosomes des deux parents, nommé crossover (CO).

En brisant la liaison génétique entre les gènes, les COs remanient le matériel génétique et génèrent, ainsi, des nouvelles combinaisons entre les différents variants des gènes. La recombinaison fait donc partie des quatre forces fondamentales qui influencent l'évolution des espèces. La **sélection** explique l'adaptation des espèces au fils des générations par la propagation des traits favorisant la survie et la reproduction. La **mutation** constitue la principale source de variation sur laquelle la sélection agit. La **recombinaison** trie la variation génétique et constitue une importante source d'innovation. La **dérive génétique** garantit la déviation des fréquences des allèles dans une population, indépendamment des autres forces évolutives.

*"**Evolution** is a population genetic process governed by **four fundamental forces**, which jointly dictate the relative abilities of genotypic variants to expand throughout a species. Darwin articulated a clear but informal description of one of those forces, **selection** (including natural and sexual selection), whose central role in the evolution of complex phenotypic traits is universally accepted, and for which an elaborate formal theory in terms of change in genotypic frequencies now exists. The remaining three evolutionary forces, however, are non-adaptive in the sense that they are not a function of the fitness properties of individuals : **mutation** (broadly including insertions, deletions, and duplications) is the fundamental source of variation on which natural selection acts ; **recombination** (including crossing-over and gene conversion) assorts variation within*

and among chromosomes ; and ***random genetic drift*** *ensures that gene frequencies will deviate a bit from generation to generation independently of other forces. Given the century of theoretical and empirical work devoted to the study of evolution, the only logical conclusion is that these four broad classes of mechanisms are, in fact, the only fundamental forces of evolution. Their relative intensity, directionality, and variation over time define the way in which evolution proceeds in a particular context."*(Lynch, 2007)

Objectifs et plan de la thèse

Cette thèse a pour objectif l'analyse, dans un contexte évolutif, des mécanismes de la recombinaison et leur impact sur les génomes. La problématique de la quantification des différences liées à la recombinaison entre espèces y est abordée, des bases moléculaires du phénomène jusqu'à une estimation plus générale d'un modèle. L'impact de la recombinaison sur le patron des substitutions nucléotidiques et la fréquence des allèles dans la population est aussi étudié.

La thèse est structurée en deux grandes parties. La première partie, composée des chapitres I et II, passe en revue les techniques et approches existantes pour l'étude de la recombinaison et la conversion génique biaisée. La seconde partie, chapitres III, IV, et V, présente des approches nouvelles ayant pour but d'améliorer notre compréhension des mécanismes évolutifs de la recombinaison et des structures génomiques.

Première partie

Dans le premier chapitre, section I.1, nous présentons **les mécanismes moléculaires méiotiques à la base de la recombinaison**. Lors de la méiose, les chromosomes homologues s'apparient sur leur longueur. Cet appariement est indépendant de la recombinaison dans certaines espèces comme *C. elegans* ou *D. melanogaster* (Gerton and Hawley, 2005; Zickler, 2006). Toutefois, pour la majorité des espèces, l'union complète des homologues nécessite la formation des cassures double-brin (revue dans Joyce and McKim (2007)). Les **cassures double-brin** ont été identifiées comme étant les précurseurs des événements de recombinaison (Szostak et al., 1983). Plusieurs modèles ont été proposés pour expliquer le passage entre les cassures double-brin et leur réparation en crossovers (COs) ou non-crossovers (NCOs) (Szostak et al., 1983; Allers and Lichten, 2001b; Constantinou et al., 2002; Wu and Hickson, 2003). La section I.2 offre une vue d'ensemble des facteurs génomiques contrôlant la production des événements de recombinaison et plus particulièrement des COs. L'émergence des techniques à haute résolution dans l'étude de la recombinaison a permis l'analyse de la distribution de ces événements le long des chromosomes dans quelques espèces modèles, comme l'homme et la levure. Nous savons maintenant que la recombinaison a lieu dans des régions restreintes (quelques kb) du génome appelées **points chauds de recombinaison** (Jeffreys et al., 2004; Myers et al., 2005). De plus, la recombinaison n'est pas répartie aléatoirement le long des chromosomes à cause de **l'interférence** tant entre les cassures double-brin (Anderson et al., 2001; de Boer et al., 2006) qu'entre les événements de recombinaison (Bishop and Zickler, 2004). Au niveau des régions chromosomiques, la recombinaison est localisée principalement à proximité

des telomères et est reduite dans les gènes et à côté du centromère (Myers et al., 2005; Mancera et al., 2008; Paigen et al., 2008). Récemment, **le gène *Prdm9***, un déterminant majeur des points chauds de recombinaison a été identifié chez l'homme et la souris (Myers et al., 2009; Baudat et al., 2009). Chez l'homme, la protéine à doigts de zinc produite par ce gène se lie à un motif dégénéré de 13 nucléotides qui est spécifique de 40% des points chauds de recombinaison chez cette espèce (Myers et al., 2008).

Tandis que d'importantes avancées ont été faites dans notre compréhension des mécanismes de la recombinaison, ces études dans quelques espèces modèles ont mis en évidence d'importantes différences dans ce processus, non seulement entre les espèces, mais aussi entre les sexes et les individus d'une même population. Ainsi, **le caryotype** (nombre et longueur des chromosomes), ainsi que l'histoire démographique et l'évolution des protéines liées à la recombinaison, semblent des facteurs importants pour expliquer les différences entre espèces (section I.2.4.1). La différence de recombinaison entre les sexes, nommée hétérochiasmie, affecte non seulement le nombre des COs mais aussi leur distribution le long des chromosomes (Shifman et al., 2006; Broman et al., 1998; Kong et al., 2002; Paigen et al., 2008; Wong et al., 2010) (table I.3). La variabilité inter-individus, quant à elle, semble intimement liée aux allèles du gène *Prdm9* porté par ceux-ci (Cheung et al., 2007; Baudat et al., 2009; Berg et al., 2010).

Les expériences réalisées depuis les travaux de Morgan montrent un rôle double de la recombinaison. Premièrement, la recombinaison a un rôle essentiel dans la progression de la méiose, en assurant **la bonne ségrégation des homologues**, entraînant ainsi sa forte régulation. Deuxièmement, la recombinaison est un processus qui évolue rapidement conduisant à de fortes différences au sein même d'une population. Un autre rôle évolutif de la recombinaison consiste à **façonner le paysage génomique** au niveau des nucléotides. Dans la section I.3, nous présentons comment la recombinaison influence la production et l'évolution des isochores (longues régions du génome caractérisées par un contenu homogène en GC) à travers **la conversion génique biaisée** (pour revue, voir Duret and Galtier (2009)). Les particularités des isochores ainsi que leur association à d'autres caractéristiques génomiques sont présentées dans ce chapitre.

Toutes ces avancées dans l'étude de la recombinaison ont été possibles grâce à des progrès technologiques majeurs. Ces percées technologiques ont facilité l'acquisition de nombreuses données à forte résolution. Dans le chapitre II, section II.1, nous décrivons les principales techniques pour mesurer la recombinaison : cartes génétiques et de déséquilibre de liaison et l'analyse par sperm-typing. Les cartes génétiques représentent l'outil le plus ancien pour l'étude de la recombinaison, depuis leur première mise en place par Sturtevant (1913). Elles se basent sur le dépistage de la transmission des marqueurs génétiques au sein des familles. Les cartes génétiques constituent, pour le moment, le seul moyen de quantifier les COs à l'échelle des génomes dans les deux sexes (Lynn et al., 2004; Cheung et al., 2007). Toutefois, elles sont dépendantes de la taille de la famille étudiée, ainsi que du nombre de méioses représentatives qui résultent souvent dans des cartes à faible résolution, particulièrement chez les eucaryotes (Arnheim et al., 2003).

L'étude du déséquilibre de liaison à l'intérieur d'une population a permis l'étude des événements historiques de recombinaison (Lewontin and Kojima, 1960). Malgré les limites de cette technique pour l'étude de l'hétérochiasmie, le nombre important d'individus

étudiés assure une forte résolution des COs à l'échelle du génome (Myers et al., 2005). Le déséquilibre de liaison sert de guide pour l'identification, localement, des potentiels points chauds de recombinaison qui peuvent ensuite être étudiés à très haute résolution dans la lignée germinale mâle, grâce au sperm-typing (Li et al., 1988).

L'acquisition des données de recombinaison ne représente que le premier pas dans l'étude de ces mécanismes. Dans la section II.2 nous décrivons quelques uns des modèles principaux pour l'étude de la distribution des COs. Ces modèles se concentrent principalement sur la modélisation des distances entre les COs. Le premier modèle, **counting model**, considère que deux COs vont être séparés par un certain nombre de NCOs (Foss et al., 1993). La distance entre deux COs suit donc une loi de Γ dont le paramètre, estimé sur la longueur génétique des chromosomes, décrit la force d'interférence. Le modèle **mechanical stress model**, quant à lui, modélise l'apparition des COs en prenant en compte des phénomènes physiques qui génèrent des tensions au niveau des chromosomes lors de la méiose (Kleckner et al., 2004).

La dernière section de ce chapitre, II.3, présente des modèles développés pour l'inférence des patrons de substitutions sous l'influence de la recombinaison à travers la conversion génique biaisée. Arndt et al. (2003) a utilisé les méthodes de maximum de vraisemblance pour inférer le patron de substitution dans une espèce en se basant soit sur un triple alignement entre des espèces proches, soit sur l'alignement entre la séquence actuelle et son équivalent ancestral. En utilisant ce valeurs de substitution dans les séquences neutres du génome humain, Duret and Arndt (2008) proposent un modèle quantifiant l'effet de la conversion génique biaisée sur ce patron. A la fin de ce chapitre II.3.3, nous présentons nos résultats de simulation de l'effet de la conversion génique biaisée sur la fréquence des allèles délétères dans la population humaine (Necşulea et al., 2011). Nous montrons que la conversion génique biaisée peut contrecarrer la sélection et engendrer le maintien des mutations délétères à de hautes fréquences dans les populations.

Deuxième partie

Comme mentionné précédemment, la recombinaison est un processus très dynamique conduisant à de multiples différences entre espèces, sexes, et individus. Dans le but de caractériser les différences inter-espèces dans la recombinaison, nous avons développé **un nouveau modèle basé sur les cartes génétiques**, détaillé dans le chapitre III. Ce modèle met en relation la longueur génétique totale des chromosomes (representant le nombre total de COs) et leur longueur physique. Des notions biologiques importantes sur le processus de la recombinaison sont prises en compte pour la construction de ce modèle : la nécessité d'un CO obligatoire par paire d'homologues pour assurer leur bonne ségrégation et la force d'interférence entre COs. L'ajustement de ce modèle aux données donne l'estimation de deux paramètres de la recombinaison : **le taux de production de COs supplémentaires par Mb** et **la force moyenne d'interférence par espèce**, définie comme la distance physique entre des COs consécutifs. Puisque le modèle implique une analyse au niveau global du caryotype, il peut être ajusté même sur des cartes génétiques de faible résolution, permettant ainsi l'étude de nombreuses espèces. Dans le chapitre III, nous montrons que ce modèle s'ajuste bien sur les 24 vertébrés et non-vertébrés analysés, même dans les cas qui ne peuvent pas être expliqués par un modèle linéaire

simple.

L'étude des distances inter-COs n'ayant été menée que chez quelques espèces. Les estimations de nos paramètres d'interférence dans ces espèces sont en accord avec les valeurs obtenues par ces études montrant le grand potentiel de notre modèle à étudier l'interférence. Les estimations obtenues pour de nouvelles espèces fournissent des données originales sur la distribution des COs. En outre, nous avons utilisé les valeurs prédites du taux de CO par Mb pour comparer les espèces entre elles et déterminer celles qui se ressemblent. Les espèces avec des paramètres similaires peuvent aussi partager un processus et des complexes protéiques de la recombinaison similaires.

Dans le but d'étudier l'hétérochiasmie, nous avons ajusté le modèle sur les cartes génétiques mâle et femelle appartenant à 6 vertébrés. Comme attendu, le sexe ayant la plus petite distance inter-CO présente également plus de COs, qui en outre, sont distribués plus uniformément. Pour 4 des 6 vértébrés, ces tendances engendrent également un taux de production des COs par Mb plus important. En revanche, pour l'opossum, les deux paramètres sont plus élevés chez la femelle que chez le mâle. Est-ce que cela résulte de la faible résolution de la carte génétique pour cette espèce, ou traduit un comportement à part chez la femelle ? Cela mérite une analyse plus approfondie. Nos résultats, ainsi que des nouvelles données (Elferink et al., 2010), remettent en question le manque d'héterochiasmie précédemment consentie chez le poulet.

L'analyse des causes de l'héterochiasmie a motivé notre deuxième étude, présentée dans le chapitre IV. **La présence et le sens de l'héterochiasmie** varient entre les espèces. De plus, nous savons que la recombinaison a un impact important sur les séquences nucléotidiques à travers la conversion génique biaisée. Jusqu'à maintenant, des études chez l'homme ont montré que la recombinaison mâle était le facteur principal dans l'évolution du contenu en GC (Webster et al., 2005; Duret and Arndt, 2008). Dans le chapitre IV, nous analysons la question de **l'impact du sexe sur la relation GC/recombinaison** chez 5 vertébrés. Nos résultats montrent que l'effet plus fort du mâle n'est pas valable pour toutes les espèces. Même chez l'homme, cet effet est principalement engendré par des régions proches des teloméres, qui contiennent principalement des points chauds de recombinaison mâle. Ces résultats montrent un impact important des forts taux de recombinaison sur la composition en nucléotides, indépendamment du sexe. La différence entre les sexes dans la localisation et l'intensité des points chauds de recombinaison est le facteur important de l'impact différentiel du sexe sur la relation GC/recombinaison selon la localisation chromosomique.

En outre, nous avons étudié l'impact du patron de substitution sur l'évolution du contenu en GC. Pour des échelles de temps faibles, la divergence homme-chimpanzé, le GC actuel des séquences est très différent du GC attendu à l'équilibre (Meunier and Duret, 2004; Duret and Arndt, 2008). Dans le chapitre IV, nous montrons que pour des échelles de temps plus longues, le contenu en GC des régions neutres, soumises à la mutation, à la conversion génique biaisée, et à la dérive génétique, est proche de l'équilibre. Nous proposons une hypothèse pour ces résultats apparemment contradictoires. Tout d'abord, les points chauds de recombinaison sont très dynamiques, comme l'indique l'absence de conservation entre des espèces proches comme l'homme et le chimpanzé (Ptak et al., 2005; Winckler et al., 2005). Ensuite, certaines régions chromosomiques comme celles proches

des telomméres conservent une haute densité en points chauds de recombinaison chez une majorité des espèces. Ces observations indiquent que, même si le contenu en GC oscille sous la pression des biais mutationnels et de la conversion génique, à long terme les deux biais s'atténuent réciproquement.

Les résultats présentés dans cette thèse amènent des éléments nouveaux pour la compréhension de l'influence réciproque entre caryotype, sexe, recombinaison, et composition en nucléotides. Cependant, comme le titre de la section d'où provient la citation précédente l'indique : "Nothing in evolution makes sense except in the light of population genetics" (L'évolution ne fait sens qu'à la lumière de la génétique des populations) (Lynch, 2007). En accord avec ce principe, le travail présenté dans cette thèse s'inscrit dans un projet plus important qui vise à intégrer les nouvelles informations sur le processus de recombinaison dans un modèle décrivant son impact évolutif dans les populations.

Introduction

When humankind first started practicing agriculture and animal breeding, it also initiated the first genetic experiments, by studying and influencing the transmission of traits to the offspring. It was not until the 1800s that the traits were found to have a discrete material support, which Gregor Mendel called "factors". It was Mendel that discovered that factors for one trait come in pairs, one from the father and one from the mother. One of the laws stated by Mendel is that different factors are passed on to the offspring separately from one another. Subsequent experiments have emphasized important deviations from this law of independent assortment. The notion of linkage arose when unusual patterns of inheritance were observed between certain factors, when certain traits were found to segregate together more often than not. Thomas Morgan associated the linkage between factors to their belonging on the same chromosome, and related the strength of this linkage to the distance separating them. However, despite a localization on the same chromosome and short physical distances, the transmission of certain factors showed incomplete linkage. It was Morgan who suggested that breaks in the linkage between factors on the same chromosome were the consequence of recombination, through chiasmata observed during meiosis. Chiasmata are the visible sites of the exchange of genetic material between the chromosomes from the two parents, also termed crossover.

By breaking the linkage between genes, crossovers mix the genetic material and thus, create new combinations of gene variants. Hence, recombination is creating variation and represents a powerful source of innovation. In chapter I, section I.1, we offer a detailed description of the molecular mechanisms leading to the advent of recombination. Section I.2 provides our latest understanding of the genomic features generating and controlling recombination, and particularly crossovers. High-resolution studies in a few organisms, such as human and yeast, have provided valuable information on the distribution of recombination events along chromosomes. However, important differences have been observed in their localization and frequency, not only between species, but also between sexes and individuals of the same species. The results obtained since the work of Morgan describe a dual role of recombination. First, recombination plays an essential role in the progress of meiosis, and thus, it is highly regulated. Second, recombination is perceived as a highly dynamic process. Another important evolutionary role of recombination consists in influencing the genomic sequences at the nucleotide level. In section I.3, we describe how recombination can generate isochore structures (long regions of relatively homogeneous GC-content) through the mechanism of biased gene conversion. The characteristics of these structures as well as their correlation to other genomic features are also provided in this section.

However, all these advancements have been possible thanks to a major technological

progress. These technological breakthroughs have facilitated the acquisition of large amount of high-quality data. In chapter II, section II.1, we present the main genetic methods that have led to the study of recombination: linkage and LD maps, as well as sperm-typing results. As the data on recombination increased considerably, a new need emerged: the necessity of models to describe them. Section II.2 describes some main modeling techniques of the distribution of crossovers. These models are mainly focusing on the distance separating two consecutive crossovers, as these events are not distributed randomly, but interfere with each other. The last section of this chapter, II.3, deals with the models for the analysis and quantification of the impact of recombination on the nucleotide changes, through the influence of the substitution pattern. Also in this section II.3.3 we present our analysis of the influence of biased gene conversion on the frequencies of alleles in a population. Notably, we focus on the modeling of the role played by biased gene conversion in the maintenance of deleterious alleles in a population.

As previously mentioned, recombination is a highly dynamic process, as multiple differences can be observed between species and sexes. However, these differences could be analyzed only in a few species, and the models describing the distribution of crossovers characterize only a subset of these species. In order to understand the evolutionary role of recombination, it is important to describe its mechanism at a much larger scale. In chapter III, we make use of the availability of low-resolution genetic maps in a wide variety of species to model the distribution of crossovers. The model we propose takes into account the constraint of one obligatory crossover per pair of homologs in order to ensure their correct segregation. This model is characterized by two parameters, which represent the rate with which supplementary crossovers are produced and the strength of interference. The estimation of these parameters in 24 vertebrates and non-vertebrates yields important information on their role in creating differences among species and sexes.

The differences in recombination between sexes (heterochiasmy) have been found to account for a differential impact of sex on the nucleotide composition. Thus, in human, the male, rather than female, recombination seems to correlate better with the GC-content of sequences. In chapter IV, we investigate this relation in four additional vertebrates. We compare the heterochiasmy differential impact with the localization along the chromosomes. This analysis allows us to understand the role played by sex on the relation between recombination and nucleotide composition, under biased gene conversion.

A summary of all our results is provided in chapter V. In view of the results presented in this thesis, we further discuss the future leads they offer to improving our understanding of the evolution of recombination and its impact on the nucleotide landscape of genomes.

Chapter I

Molecular mechanisms of recombination

This chapter provides the necessary basis to understand the molecular mechanisms of recombination and its impact on the genome. The first section summarily describes the phases of meiosis, with a detailed presentation of the recombination mechanism in the second section. The third section characterizes the process of biased gene conversion and its implications on the isochore structures.

I.1 Meiosis

Most sexually-reproducing species have diploid cells, e.g. they have two copies of each chromosome, one from each parent. When, in turn, such an individual reproduces, it transmits only half of its genetic material to the offspring, through specialized cells termed gametes. An essential step in the sexual reproduction of species is the generation of haploid gametes from diploid cells, which prevents the doubling in genetic material with each generation. The reduction in ploidy is achieved through a special type of cellular division, called meiosis.

The specialized diploid cells in ovaries and testis (germinal cells) contain two copies of each chromosome (paternal and maternal), also known as **homologs**. A preceding step to meiosis consists in the replication of DNA in germinal cells, with the duplication of chromosomes. At the end of this phase each chromosome consists of two sister chromatids linked at the level of the centromere. Two cell divisions follow, which halve the number of chromosomes in the gametes, thus resulting in four haploid cells.

I.1.1 The phases of meiosis

The first meiotic division is particularly long, representing more than 90% of the total time of meiosis. It is also known as *reductional division* since it produces two haploid cells. The passage from a diploid number of chromosomes to a haploid stage is done in four phases: prophase, metaphase, anaphase and telophase as in figure I.1. Two important events, specific to meiosis, take place during **prophase I**: the synapsis of homologous chromosomes and recombination. In turn, prophase I is divided into five phases: leptotene,

zygotene, pachytene, diplotene, diakinezis. Of the wide range of proteins acting during prophase I, some are mentioned here after and a detailed description can be found in the additional table A. At **metaphase I**, the paired chromosomes become attached to the meiotic spindle and line up. The chromosomes are condensed at their maximum and the chiasmata (the points of contact between homologs) are visible. The resolution of the chiasmata takes place during **anaphase I**, when the two replicated homologs (each still consisting of two sister chromatids) separate and are pulled to opposite poles. The chromosomes reach the poles of the meiotic spindle during the **telophase I** and the cell divides resulting in two sister cells, each inheriting two copies of either the maternal or the paternal homolog of each pair. Each daughter cell contains half the number of chromosomes, which consists of a pair of sister chromatids, closely attached at the level of the centromere.

The actual formation of gametes is taking place during the second meiotic division, also known as *equational division*. The transition between the two meiotic divisions takes place rapidly, during a short interphase period, with no DNA replication. The nuclear envelope (NE) of each daughter cell breaks down in **prophase II**, and a new meiotic spindle forms. In **metaphase II**, single condensed chromosomes, as opposed to homologous pairs of chromosomes in metaphase I, line up on the spindle. The two sister chromatids making up each chromosome are separated at the centromere during **anaphase II**. They segregate to opposites poles of the cell, thus generating two haploid nuclei, each containing a single chromatid. At **telophase II**, the nuclear envelope of each one of the four cells is formed, producing four gametes, each with a haploid set of chromosomes.

I.1.2 Pairing and synapsis of homologs during prophase I

The pairing of chromosomes starts at leptotene, as the homologs overcome spatial separation from complete dissociation to co-alignment. In order to achieve this long-range alignments, homologous chromosomes must find and recognize each other. In a few organisms, the establishment of a physical contact between homologs may occur prior to meiosis (reviewed in McKee (2004); Zickler (2006)). This phenomenon is encountered in Dipterans, as it is especially necessary for the initiation of meiotic association in *Drosophila* males which lack both recombination and a synaptonemal complex (SC) (Vazquez et al., 2002). The spatial association of homologs has also been reported in somatic cells during mitotic interphase, when chromosomes occupy distinct territories according to their length and gene-density, but this association is infrequent and seems to occur randomly (Cremer and Cremer, 2001; Mora et al., 2006). However, the premeiotic interactions are far from being an universal feature. Even when these type of interactions are present, it is difficult to assess their influence on the pairing of chromosomes during meiosis.

Prior to pairing, chromosome ends are linked to the cytoskeleton network through the inner and outer nuclear membrane complex proteins SUN/KASH (Tzur et al., 2006). Figure I.2 depicts the attachment between the microtubules in the cytoplasm and the chromosomes inside the nucleus through specific nuclear envelope (NE) proteins. These NE bridges allow cytoplasmic forces to induce chromosome movement inside the nucleus (Penkner et al., 2009). The motion of chromosomes is supposed to help the pairing of homologs by creating the opportunity of encounter, but most importantly by disrupting the nonhomologous

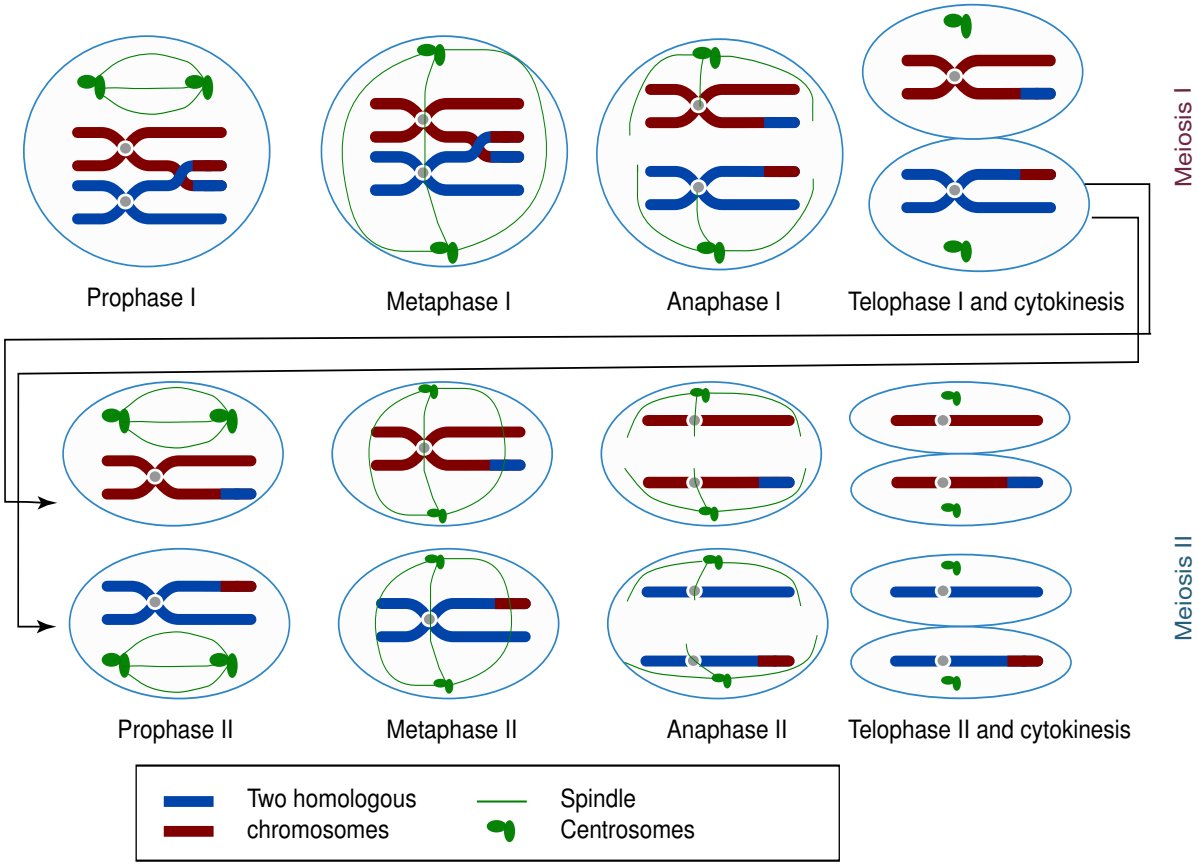

Figure I.1: *The representation of the two meiotic divisions: Meiosis I and II. Each meiotic division is further classified into four phases: prophase, metaphase, anaphase and telophase. At the end of meiosis, four gametes are produced each with a haploid set of chromosomes. Prophase I is characterized by the exchange of genetic material between homologous chromosomes, also known as crossovers (CO). During metaphase, chromosomes attach to the spindle formed between centrosomes. The genetic material, homologs for Meiosis I and sister chromatids for Meiosis II, segregate at opposite poles during anaphase. The telophase results in the reconstruction of the NE around each homologous chromosome or sister chromatid for the first and second meiotic divisions respectively. Cytokinesis results in the division of the cytoplasm in order to form two daughter cells.*

associations (reviewed in Koszul and Kleckner (2009)). At late leptotene the chromosomes migrate into a specific meiotic organization called the "bouquet" arrangement (Zickler and Kleckner, 1998; Scherthan, 2001). The "chromosomal bouquet" is a conserved feature of eukaryotes, characterized by the telomeres being anchored to the NE and the chromosomes being clustered within a delimited volume of the nucleus (Zickler, 2006). Although the role of the "bouquet" configuration in the pairing between homologs is not well defined, it has been suggested that the clustering of chromosomes in a limited area, as well as their rapid movement in and out of the "bouquet" are essential for the resolution of entanglements of chromosomes as well as the prevention of nonhomologous contacts (Zickler, 2006).

The chromosome dynamics during meiosis is indeed an essential step in their synapsis, but the question still remains as to how homologs recognize each other at very long

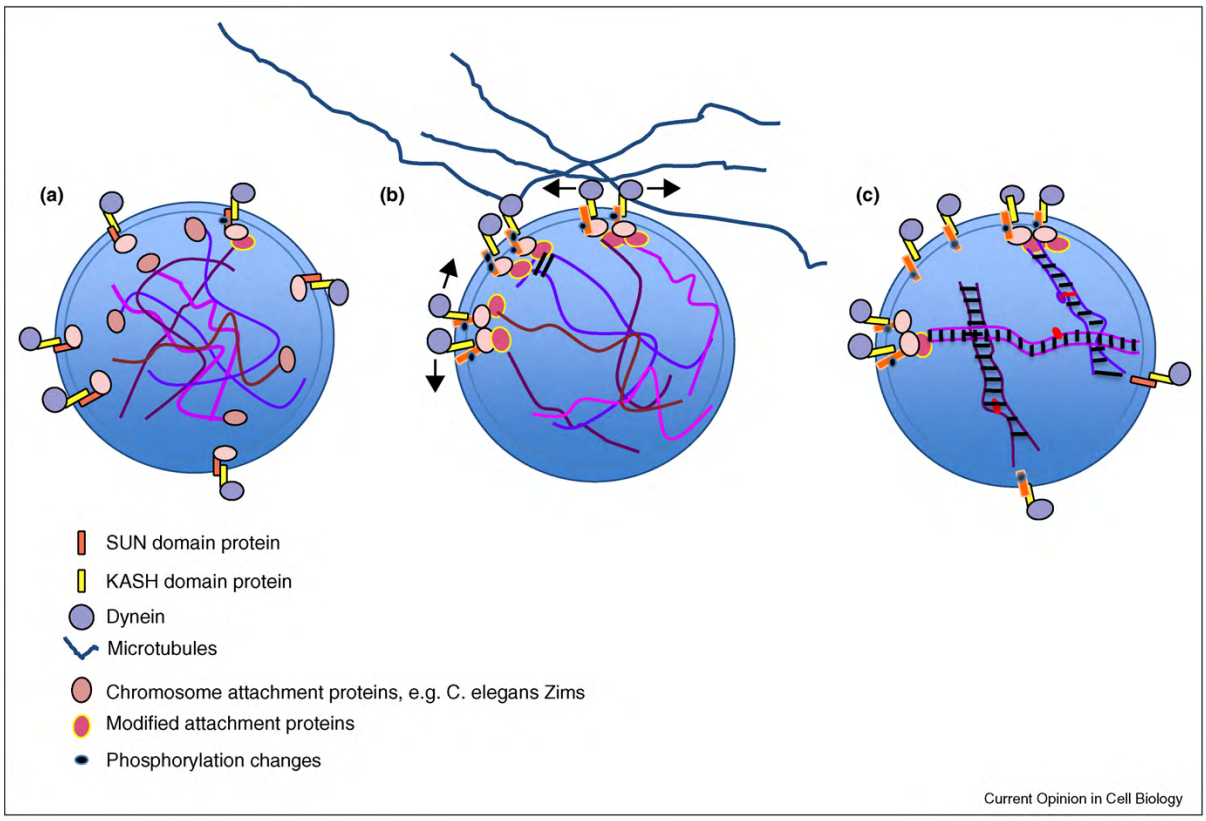

Figure I.2: *Attachment to the nuclear envelope promotes chromosome movements and homologous attachments.(a) A SUN/KASH domain complex bridges the nuclear envelope (NE) connecting with dynein on the cytoplasmic face. Chromosomes attach via telomeres or specialized pairing center sequences to the NE complex. In* C. elegans*, Sun-1 phosphorylation (black circle on SUN protein with attached chromosome) early in zygotene is required for subsequent events. (b) The SUN/KASH/chromosomal foci cluster together and mature into large patches, as additional phosphorylation of Sun-1 is observed. In patches, dynein-mediated forces stress chromosomes, leading to detachment of non-homologous attachments and synapsis and homologous ones. (c) As homologs fully synapse and execute DSB repair, SUN-1 phosphorylation status again changes, leading to dispersal of chromosomes into pachytene morphology. From Yanowitz (2010).*

distances. Recombinases (such as Rad51) are known to facilitate the homology recognition, but at a local scale, when the interacting molecules are already aligned (Rao et al., 1995; Barzel and Kupiec, 2008). Moreover, in many organisms the homologous pairing is independent of recombination (Gerton and Hawley, 2005; Zickler, 2006). Multiple recombination-independent mechanisms for homology search have been proposed. The clustering of chromosomes could be attained through specific *cis*-acting pairing centres (PC). In *C. elegans*, homologue-recognition regions (HRR) have been found to localize along each chromosome, and are essential to the local stabilizing of pairing and the initiation of SC polymerization (McKim et al., 1988; MacQueen et al., 2005). Highly transcribed ribosomal DNA (rDNA) regions in *D. melanogaster* also play a role as PC between the X and Y chromosomes (McKee, 1996). The pericentric heterochromatic regions too, could act as pairing sites between chromosomes, in *S. cerevisiae* and *D. melanogaster* (Kemp

et al., 2004; Dernburg et al., 1996). Another mechanism for long-range pairing is based on the observation that during meiosis, chromosomes pair only when transcriptionally active (Cook, 1997). DNA regions that are under active transcription form loops attached to specialized **transcription factories**. Multiple homologous loops may share the same transcription factory allowing for a transient binding between DNA sequences, and the subsequent paring of homologs (Xu and Cook, 2008). Even if considered less probable, the model of **DNA-DNA** direct contacts is based on long-range attractive interactions between double-stranded DNA (Danilowicz et al., 2009). These interactions result from the spatial modulation of charge distribution in DNA helices (Kornyshev and Wynveen, 2009), even in protein-free conditions.

I.1.3 Double strand break (DSB) dependent pairing and the Synaptonemal Complex (SC)

For the majority of species, full homologous pairing seems to be intimately linked to the initiation of recombination via double-strand breaks (DSB) (reviewed in Joyce and McKim (2007)). However, knock-out mutants for the proteins responsible for DSB formation in *Caenorhabditis elegans* and females of *Drosophila* can still build a synaptonemal complex (SC) structure and establish inter-homologs synapsis (Dernburg et al., 1998; McKim et al., 1998). SC is a well-conserved tripartite proteinaceous structure consists of two lateral elements (LE) and a central element (CE), connected together by transverse filaments (TF), with the two homologous chromatides disposed in loops around the corresponding LE (Schmekel and Daneholt, 1995) (figure I.3). The chromosome axes begin to assemble in short fragments, at leptotene, as a result of the incorporation of cohesin (*e.g* Rec8) and axial proteins, such as SCP2 and SCP3 in mammals (Eijpe et al., 2003). The bits will then fuse and form full-length LE as part of the SC (Schalk et al., 1998). Also at leptotene, DSBs are induced on the chromatin loops through the action of the evolutionary conserved endonuclease Spo11 (Keeney et al., 1997; Blat et al., 2002; Keeney and Neale, 2006). The Mre11 complex of proteins further removes Spo11 from the DNA ends and continues to degrade the DSBs from the 5' to the 3' end (Borde and Cobb, 2009). Even if DSBs may occur on chromatin loop, it has been proposed that the sequence containing the DSB and the chromosome axis will become spatially associated via DNA/protein recombination complexes (Blat et al., 2002) (figure I.4). It has been observed that the sites of DSBs form 400 nanometers (nm) local bridges between the homologous chromosome axes (Tessé et al., 2003). The exact mechanism of DSB-mediated alignment is not fully understood, nevertheless a complex of proteins has been identified as being involved in the interaxis bridges assembly (Storlazzi et al., 2010). Strand exchange proteins, such as Rad51 and Dmc1, will form nucleoprotein filaments, binding the resulting single stranded DNA (ssDNA) and catalyzing homologous strand invasion (Shinohara et al., 1992; Kagawa and Kurumizaka, 2010).

At zygotene, a small subset of the DSB bridges, the ones that have matured into axial associations, and that later will form crossovers (CO), are also developing sites for the SC (Page and Hawley, 2004). An overview of the recombination and SC processes is pictured in figure I.5. Contemporary to the initiation of the CE in SC, the 3' ssDNA

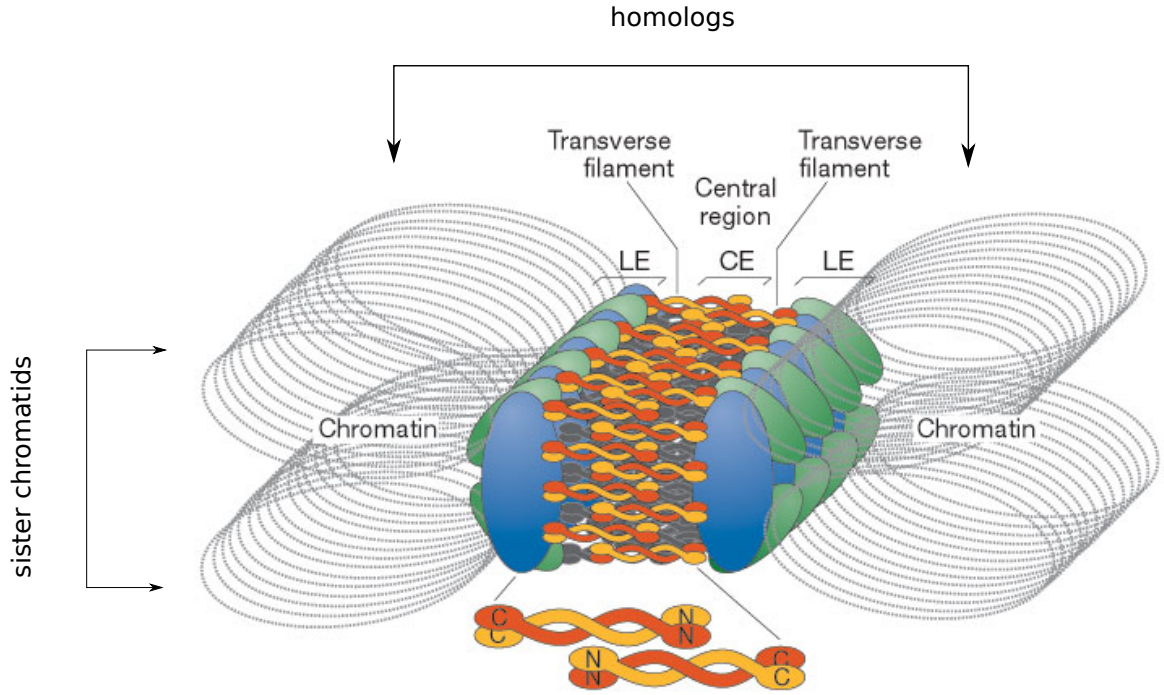

Figure I.3: *Model of the synaptonemal complex structure. The lateral element (LE) comprises cohesins (Rec8/C(2)M/SYN1, STAG3/Rec11, SMC1-β and SMC3) (blue ovals), the structural proteins SCP2 and SCP3 and the HORMA-domain proteins Hop1/HIM3/Asy1 (all other LE proteins - green ovals). The transverse filaments (TF) are formed by the proteins Zip1/SCP1/C(3)G/SYP1 (shown also at the bottom). Adapted from Castro and Lorca (2005), originally adapted from Page and Hawley (2004).*

invades the homologous double strand DNA (dsDNA), through a process called single-end invasion (SEI) (Hunter and Kleckner, 2001). Homology is recognized between the two sequences through sequential cycles of binding, sampling and release of the dsDNA (Neale and Keeney, 2006). The proteins responsible for the CE nucleation (SCP1 protein, in mammals and ZMM proteins, in yeast) polymerize along homologs leading to the full assembly of the SC at mid-pachytene (Meuwissen et al., 1997; Zickler, 2006). The stable connection between homologs via the SC is called synapsis (Zickler and Kleckner, 1998).

I.1.4 Molecular mechanisms of recombination

Spontaneous DSBs arise frequently, and without a correct repair mechanism, they would be highly deleterious leading to chromosome mis-segregation, rearrangements or apoptosis. The repair of the DNA break can proceed either by non-homologous end joining (NHEJ) or by homologous recombination (HR), using a DNA template (Haber, 2000). NHEJ is widely used in mammalian mitotic cells and consists of directly ligating the broken ends of the DNA (Weterings and van Gent, 2004). The process itself needs no or very little homology and is very prone to errors (Lieber et al., 2003). The repair of DSBs generated during meiosis exhibits low levels of NHEJ in mammals, and is mainly performed through HR (Goedecke et al., 1999; Haber, 2000). HR uses a template DNA sequence, that can be either the sister chromatid, the homologous chromosome or an ectopic sequence, in order

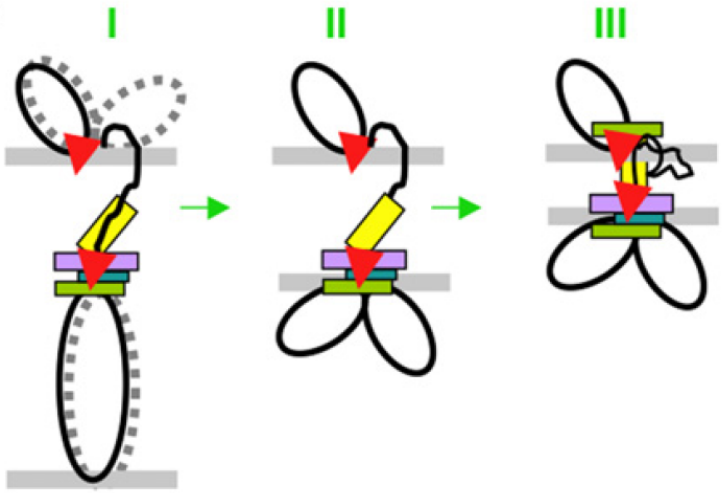

Figure I.4: *Possible architecture of the DNA/protein recombination complexes mediating homolog pairing. (I) One DSB end (lower red arrowhead) interacts with a homologous chromatin loop, thereby initiating the assembly of a protein complex containing at least four post-DSB recombination proteins (for details of the proteins see Storlazzi et al. (2010)). The other DSB end (upper red arrowhead) associates with the axis of the DSB "donor" chromosome. (II) The complex formed with the partner chromosome in (I) becomes axis-associated, thereby bringing donor and recipient chromosomes into closer proximity, with asymmetric evolution of the recipient chromosome complex. (III) The chromosome axes are separated by a distance of 400 nm. From Storlazzi et al. (2010).*

to rebuild the missing DNA. The use of the homologous chromosome as a template is preferred during meisois as it is essential for the accurate segregation of homologs at the end of meiosis I (Schwacha and Kleckner, 1997). Hereafter, HR will refer to the recombination process that takes place between homologous chromatids during meiosis. The repair through HR yields two types of final products: crossovers (CO) and non-crossovers (NCO). While a CO supposes the reciprocal exchange of large portions of genetic material between the homologous chromosomes, a NCO is a non-reciprocal, highly local event which results in the swap of only a small DNA segment.

The mechanisms leading to these two recombination products are not yet fully understood, but all the models for HR are based on the formation of a single-end invasions (SEI) intermediate. One of the first models to account for the production of both COs and NCOs, the double-strand break repair (DSBR) model (Szostak et al., 1983), is based on the resolution of a cross-stranded structure, the Holliday junction (HJ) (Holliday, 1964) (figure I.6). Following the SEI, the loop (also called D-loop) formed by the coming apart of the homologous dsDNA, is enlarged through new DNA synthesis and captures the opposing free 5' end. Ligation of the two ends as well as gap repair of the missing DNA on the sister chromatid completes the formation of a second recombination intermediate, the double HJ (dHJ). Endonucleases resolve the dHJ by introducing symmetric nicks in the strands with the same polarity, which are then ligated. If, like in figure I.6, the cuts (arrows) are made on the two sides of the dHJ, thus affecting all four strands, a CO is produced.

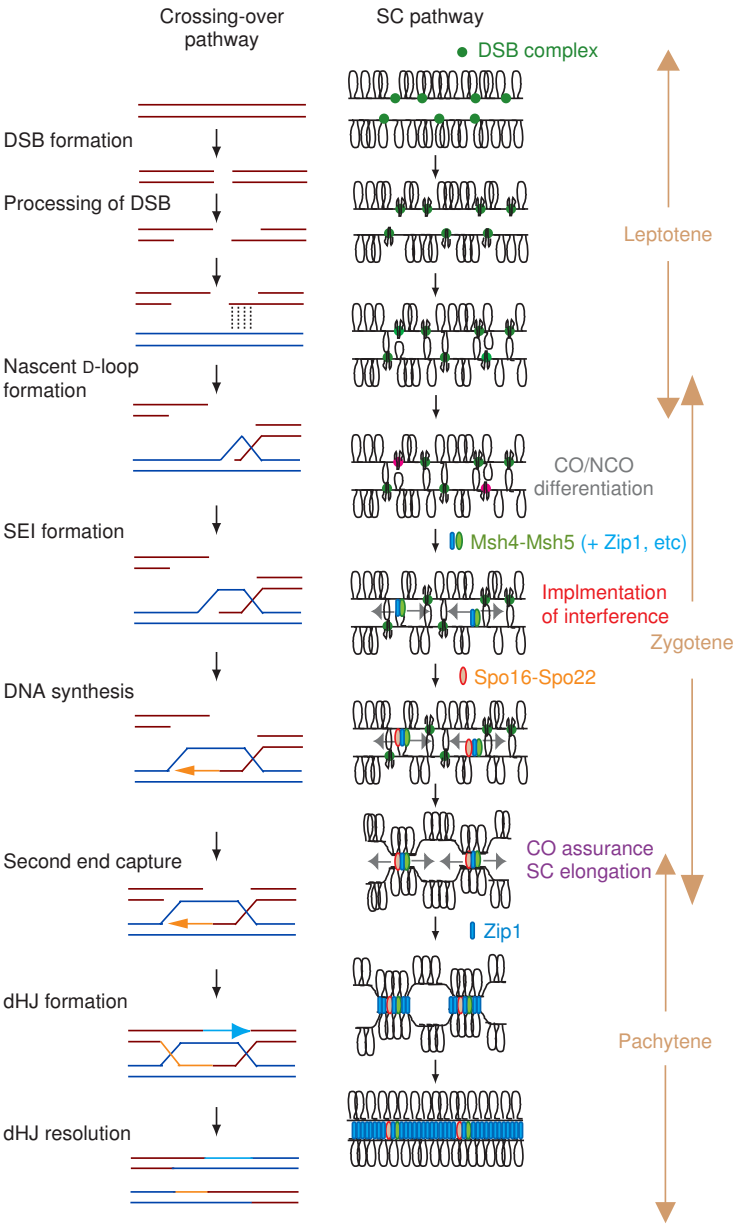

Figure I.5: *Model of the parallel between recombination and synaptonemal complex (SC) formation and timing in yeast. At the beginning of leptotene, DSBs appear on the chromatin loops along the chromosomes. At zygotene, bridges are formed between the axes of the two homologous chromosomes at the sites of DSBs by single-end invasion (SEI) and the formation of D-loops. DSBs can be solved either as crossovers (CO) or non-crossovers (NCO). The sites of future CO resolution will recruit proteins such as Zip1, which is a component of the central element (CE) constituting the SC. At mid-pachytene, the polymerization of Zip1 results in the full assembly of the SC. The resolution of the recombination intermediates, double Holliday junctions (dHJ) yields CO recombination products. From Shinohara et al. (2008).*

Two cuts on the same side of the dHJ, affecting only two homologous strands out of the four, produce a NCO. Many predictions of the DSBR model have come true, starting with the observation of the dHJ intermediates deduced from 2D gel analysis (Schwacha and Kleckner, 1995; Allers and Lichten, 2001a). Recently, two long-awaited eukaryotic resolvases of the dHJ have been identified: GEN1/Yen1 (Sharples, 2001; Ip et al., 2008; Bailly et al., 2010) and SLX4/BTBD12/MUS312/Him-18 complex (Fekairi et al., 2009; Saito et al., 2009; Svendsen et al., 2009).

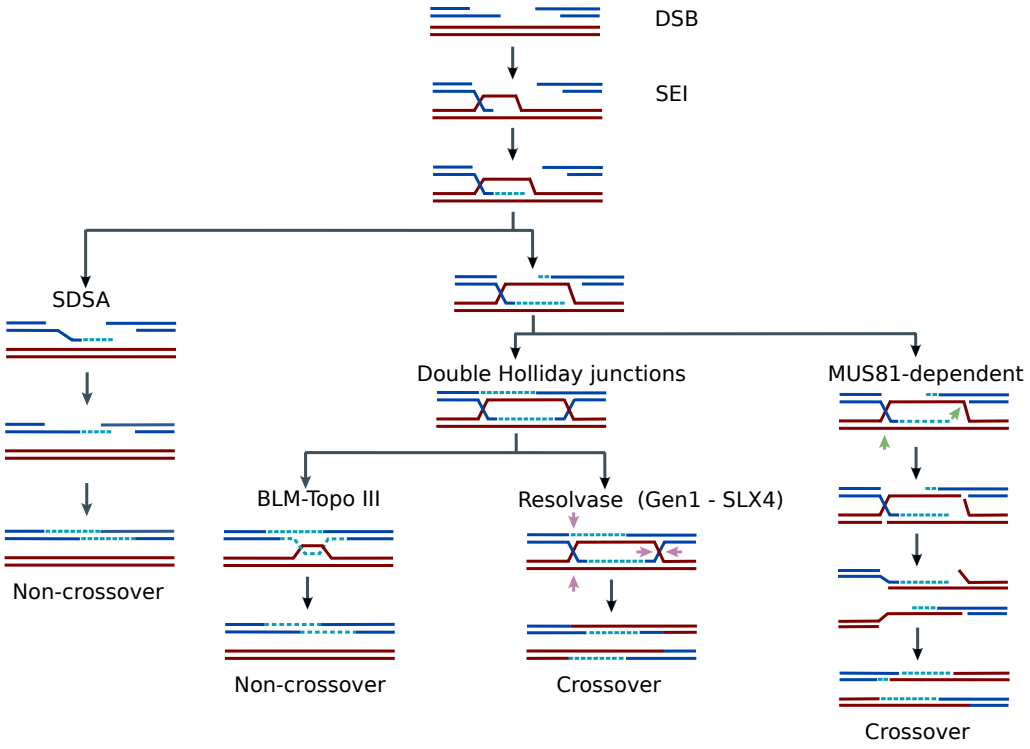

Figure I.6: *Homologous recombination. Summary of our current understanding of recombination pathways that are initiated by a DNA double-strand break (DSB) and which lead to gene conversion with or without crossover. First, the ends of the DSB are cut, producing single stranded DNA that recruits the recombination protein RAD51. The assembly of a RAD51 nucleoprotein filament leads to interactions with homologous duplex DNA and strand invasion. This process is known as single-end invasion (SEI). In some pathways for recombination (centre), SEI is followed by the capture of the second DNA end. This intermediate can proceed to form double Holliday junctions, and any remaining gaps might be filled by new DNA synthesis. The resulting Holliday junctions might then serve as the substrate for a classic Holliday-junction-resolution reaction or be dissociated by the combined actions of BLM (Bloom's syndrome protein) and topoisomerase IIIα (Topo III). The BLM-Topo-III reaction primarily leads to the formation of non-crossover products, as mutations in BLM cause an increase in crossover formation. Recombinants can also form by a MUS81-dependent pathway that does not involve Holliday-junction formation (right). Similarly, DSBs can be repaired by synthesis-dependent strand annealing (SDSA) (left) to produce non-crossovers. Adapted from Liu and West (2004).*

Despite the advantage of offering an integrated view of the process generating CO and NCO, the DSBR model does not account for all the biological observations, especially regarding the production of NCOs (reviewed in McMahill et al. (2007)). The resolution of the dHJ as a NCO is expected to generate a heteroduplex DNA to the left of the DSB in one of the chromatids and to the right in the other chromatid. However, several studies have found that in the majority of cases the heteroduplex is present only in one of the chromatids, and even in cases of two tracts of heteroduplex, they were localized on the same chromatid (Allers and Lichten, 2001a; Merker et al., 2003; Jessop et al., 2005). Additionally, knock-out mutants for proteins involved in CO production reduce drastically their number, but yield no influence on the production of NCOs (reviewed in Bishop and Zickler (2004)). The current hypothesis seems to be that the majority of physically observed HJ are processed into COs (Allers and Lichten, 2001a). These observations as well as the discovery of additional protein complexes involved in CO/NCO production have led to the description of alternative pathways as represented in figure I.6. An alternative pathway of dHJ resolution involves the helicase BLM. BLM together with RMI1 and TOPIII form a protein complex (BLM*) which catalyzes the dissolution of the dHJ and generates NCOs as the final products by preventing the exchange of flanking sequences (Wu and Hickson, 2003, 2006).

Another pathway acting on the HJ and leading to the exclusive production of COs, involves the Mus81-Eme1 protein complex (Mus81*) (Constantinou et al., 2002). In *Schizosaccharomyces pombe*, the majority, if not all COs, are dependent on this pathway (Boddy et al., 2001). It was first thought that Mus81-null mutants, in mouse, were viable, but it was recently demonstrated that they are also subject to severe meiotic defects (Holloway et al., 2008). Studies in *S. cerevisae* and *Arabidopsis thaliana* have pointed to the particularity of Mus81 in generating interference-independent COs (de los Santos et al., 2003; Berchowitz et al., 2007). Although the mechanism involving Mus81* is not yet clear, the preferred hypothesis consists in a HJ cleavage activity, which has been observed *in vitro* (Cromie et al., 2006; Taylor and McGowan, 2008). It has been suggested that Mus81* acts on the D-loops before their full maturation into dHJs, by making two cuts on the opposing strands of the homologous chromatid, transforming the four-way branched structure into two linear products (Gaillard et al., 2003; Osman et al., 2003). The linear products are further resolved by DNA synthesis and ligation, resulting in final COs.

Additional to DSBR and BLM* models, most NCOs result from synthesis-dependent strand-annealing (SDSA) without the formation of a HJ (Allers and Lichten, 2001b). Following the SEI and the extension of the invading end past the site of the DSB, the D-loop is disrupted. The displaced DNA strand will further anneal with its complementary ssDNA on the other side of the DSB. DNA synthesis and nick ligation will complete the process, resulting in a NCO (McMahill et al., 2007). Intermediates of the SDSA pathway have been detected in *S. cerevisae* meiotic cells (McMahill et al., 2007).

The current view is that multiple pathways may be used for the formation of recombination products. Moreover, these pathways are not completely independent as there is evidence of cross-talk among them. SLX4 and Mus81 interact, and it has been proposed that the SLX1-SLX4 may be part of the Mus81* pathway as well, with SLX1 making the initial nick of the HJ and Mus81 cutting the nicked HJ generated by the second end

capture (Svendsen and Harper, 2010). Also, BLM is known to interact with Mus81 in somatic cells and with MLH1, representing a possible bridge between the DSBR and Mus81 pathways (Holloway et al., 2008, 2010).

I.1.5 Postsynaptic phase

By mid-late pachytene, the mature CO products are observed cytologically at chiasmata sites (Hunter and Kleckner, 2001; Guillon et al., 2005). From late pachytene to diplotene, the SC is disassembled as its CE proteins dissociate from the chromosome arms (Tsubouchi and Roeder, 2005). Following the dissociation of the SC, chiasmata become visible. The homologous chromosomes still attached at the centromere as well as at chiasmata sites, prepare to attach the meiotic spindle upon the entry in metaphase I (Zickler and Kleckner, 1999; Zickler, 2006).

Review articles for this sub-chapter: Liu and West (2004); Zickler (2006); Ding et al. (2010); Storlazzi et al. (2010); Székvölgyi and Nicolas (2010); Yanowitz (2010)

I.2 Recombination

I.2.1 Distribution of recombination events

I.2.1.1 DSB distribution

Are the recombination events: DSBs, COs and NCOs, evenly distributed along the chromosomes? What makes a genomic region likely to host some of these products? These open questions have lately benefited from advances in microarray and cytological technologies, especially in yeast. Thus, DSBs have been found to cluster in small regions, called **DSB hotspots** (DSBh) separated by long regions with few or no DSB events (de Massy et al., 1995; Lichten and Goldman, 1995; Baudat and Nicolas, 1997; Petes, 2001). Hotspots are regions having a higher fraction of events compared to their surrounding environment. In *S. cerevisae*, DSBs may occur at many sites within regions of a few hundred base pairs (bp) (de Massy et al., 1995; Liu et al., 1995; Xu and Kleckner, 1995). Mutations at the putative DNA-binding surface of Spo11 have been shown to affect the distribution of DSBs, but the effect is weak and **no specific motif** has been found at the sequence level to explain the existence of DSBh (Liu et al., 1995; Murakami and Nicolas, 2009). Despite a lack of specificity in the binding of Spo11, some epigenetic features have been found to correlate with the distribution of DSBhs. DSBs are preferentially initiated in the **chromatin loops** rather than the chromatin bound to chromosome axes (Blat et al., 2002). But not all chromatin loops contain a DSB, the distance between DSBs exceeding the average size of the DNA loops (Gerton et al., 2000). **Local chromatin accessibility** is another important factor in the initiation of DSBs, since DSBh are preferentially located in nuclease-hypersensitive regions (Ohta et al., 1994; Wu and Lichten, 1994; Berchowitz et al., 2009). Histone modifications, especially H3 lysine 4 trimethylation, associated with active chromatin, are also marks of recombination initiation sites in *S. cerevisae* (Borde et al., 2009) and mouse (Buard et al., 2009).

In *S. cerevisae*, two chromosomal landmarks are considered cold DSB regions: the **chromosome ends** (also known as telomeres) (Su et al., 2000; Blitzblau et al., 2007; Buhler et al., 2007) and **ribosomal DNA (rDNA)** (Petes and Botstein, 1977; Blitzblau et al., 2007) in *S. cerevisae*. It was postulated that DSB initiation sites avoid highly repetitive DNA, as it could lead to nonhomologous interactions between chromosomes and loss of rDNA repeats (Barton et al., 2003). Despite the first 20 kb of the chromosome ends in *S. cerevisae* being cold, the following 30 Kb are hot, suggesting that **telomeres act as promoters for a strong recombination activity** in adjacent regions (Blitzblau et al., 2007; Buhler et al., 2007; Barton et al., 2008). At first, centromeric regions were also considered cold regions (Gerton et al., 2000; Borde et al., 2004). However, important DSB hotspots have been found in the pericentromeric region of *S. cerevisae* (Blitzblau et al., 2007; Buhler et al., 2007). Interestingly, pericentromeric sequences have also an open chromatin structure (Berchowitz et al., 2009).

At the sequence level, DSBs form preferentially in **intergenic** regions (Baudat and Nicolas, 1997; Gerton et al., 2000; Cromie et al., 2007). In *S. cerevisae*, most recombination initiation sites occur in the vicinity of transcription promoters (Baudat and Nicolas, 1997; Gerton et al., 2000). Some DSB hotspots have been found to require the presence of transcription factors, however the level of transcription doesn't seem to affect the frequency of DSB events (Gerton et al., 2000). In *S. pombe*, an association has been reported between recombination hotspots and long non-coding RNA loci, which was proposed to result from the role of these RNA loci in binding factors, such as transcription factors, and thus remodeling the chromatin (Wahls et al., 2008).

The existence of hot and cold DSB regions results in the non-random distribution of DSB events. This distribution is related to the observed phenomenon of interference between the recombination initiation sites. Positive interference (simply termed interference hereafter) supposes the existence of an inhibition zone around events, preventing the formation of additional recombination occurrences. Even if DSB interference is detected in the studies mentioned previously, it is certainly underestimated, as the DSB mapping techniques account for the combined results of thousands of independent meioses (Berchowitz and Copenhaver, 2010). Cytological evidence of DSB interference includes the observation of distances between early recombination nodules (structures associated to the SC), in plants (Anderson et al., 2001), as well as between early MSH4 foci in mouse (de Boer et al., 2006). Both early nodules and MSH4 foci are associated with Rad51/Dmc1, and are considered representative of the DSB sites (Zickler and Kleckner, 1999). Another indication of competition between DSBs has been observed by deleting DSB hotspots, which stimulated the formation of DSBs at adjacent sites (Wu and Lichten, 1995). Also, insertion of a DSB hotspot results in the reduction of DSB activity in the neighboring hotspots (Wu and Lichten, 1995; Fan et al., 1997; Robine et al., 2007). The existence of interference suggests that even if the distribution of DSBs is variable from one meiosis to another, their number is subject to little variability, as for example in yeast, it varies between 150 and 170 events per meiosis (Buhler et al., 2007; Robine et al., 2007).

I.2.1.2 CO and NCO distribution

Figure I.7 depicts the distribution of DSB and recombination rates along chromosome III

in yeast. Additional to the hotspot organization of DSBs, the recombination products, COs and NCOs, are also subject to a non random distribution. Moreover, some DSBh seem more favorable to NCOs while other host preferentially COs, suggesting higher levels of interference (Mancera et al., 2008) (figure I.8). Techniques such as genetic mapping, linkage disequilibrium and sperm-typing (see chapter II.1 for details), have permitted the extensive study of CO distribution in a wide variety of species. In humans, a majority of CO events (60%) are part of a known **CO hotspots** (COh) (Coop et al., 2008), and 60-70% of these known COh are hosted within **10% of the genome** (Myers et al., 2005). A CO hotspot is a region **1-2 kb wide** (Jeffreys et al., 2004), surrounded by CO-depleted regions, on average 50-100 Kb long (Myers et al., 2005). Despite an evolutionary conserved length (Mancera et al., 2008; Wu et al., 2010), COh display a wide variety of intensities (Arnheim et al., 2007; Wu et al., 2010). The CO frequency associated with COh, in mouse, ranges from 0.0027% to 1.1%(Wu et al., 2010), for an average CO rate (COR) of 0.5 cM/Mb per genome (Cox et al., 2009). In human too, the high resolution characterization of recombination hotspots through sperm-typing (additional table B) indicates a wide variety in COh intensity, for a genome-wide average COR of 1.1 cM/Mb (Kong et al., 2002).

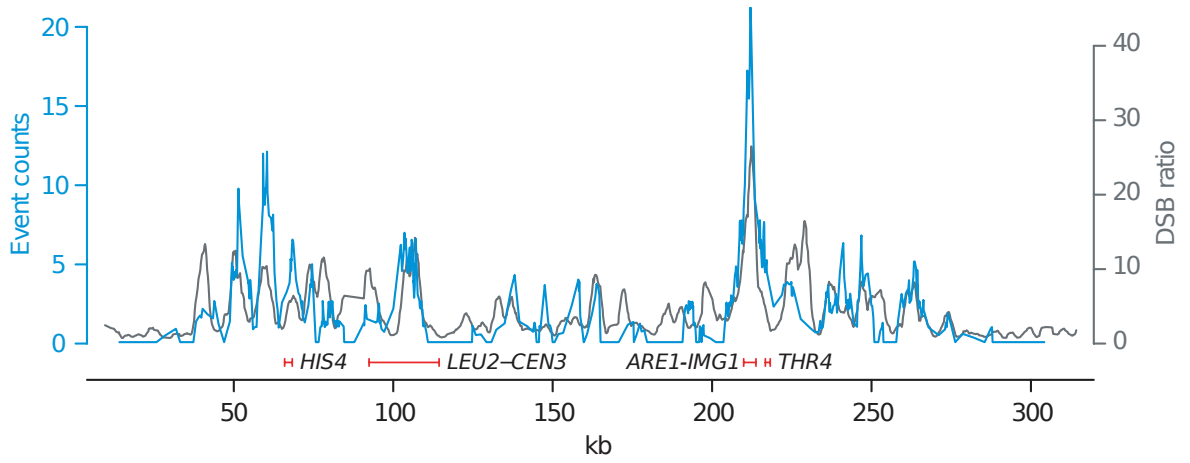

Figure I.7: *Comparison of DSB and recombination rates along chromosome III in yeast. DSB smoothed fluorescence ratios in a SK1 strain (dmc1D, grey) are compared with recombination event counts in a S288c/YJM789 hybrid strain (blue), after adjusting the latter for varying intermarker interval size. Peak locations largely agree despite distinct strain backgrounds, although some fine-scale differences exist. Previously known hotspots are indicated by red segments. From Mancera et al. (2008).*

The CO frequency in the centromeric and pericentromeric regions is very low, both in human and yeast, (Myers et al., 2005; Chen et al., 2008; Mancera et al., 2008). Experiments in yeast suggest that this reduction might be the consequence of a low number of recombination initiation events in this region (Gerton et al., 2000; Borde et al., 2004). However, as previously mentioned, the pericentromeric regions are not completely devoid of DSBh. A reason for the reduction in recombination may be that COs close to centromeres can interfere with meiotic chromosome segregation (Rockmill et al., 2006). It

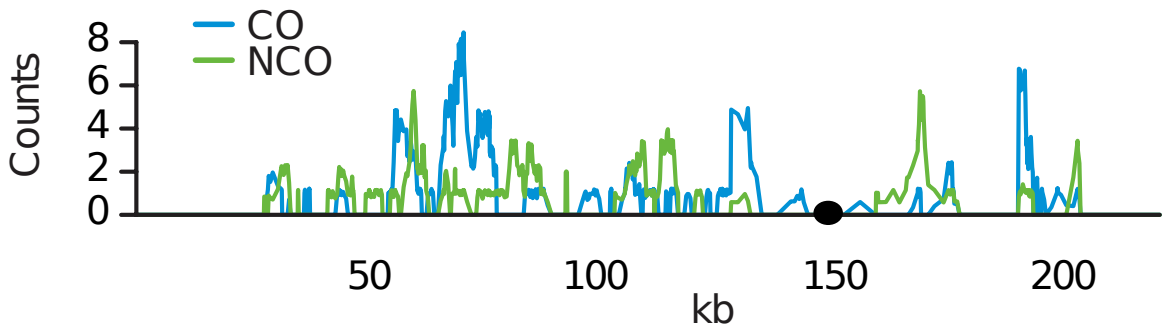

Figure I.8: *Crossover and non-crossover rates along chromosome I in yeast. Crossover (CO, blue) and non-crossover (NCO, green) counts, adjusted for varying intermarker interval size. The black circle represents the centromere. From Mancera et al. (2008).*

has been postulated that the CO pathway is inhibited, and repair might favor the sister rather than the homologous chromatin, which results in a reduction of the number of COs in this region (Blitzblau et al., 2007; Chen et al., 2008). This inhibition is not due to a more compact chromatin configuration, as the chromatin structure in the pericentromeric area has been found to be open (Berchowitz et al., 2009). Moreover, recent studies in the genome of maize have shown the existence of NCO events at centromere (Shi et al., 2010).

The distribution of CO and NCO events at telomeres is more ambiguous. It has been suggested that the recombination events are depleted at telomeres as they could generated high rates of non-homologous recombination, due to the high density of repeat elements in this region (Barton et al., 2003). However, the ambiguity may also simply result from the difficulty of studying repetititve sequences (Mancera et al., 2008). In mammals, the recombination rates are highly increased in the regions adjacent to chromosome ends (Myers et al., 2005; Paigen et al., 2008), while in *S. cerevisae* the landscape is patchy, with some chromosomes having no event long before the telomeres, while others show a strong activity near telomeres (Mancera et al., 2008).

Although DSBs in *S. cerevisae* are associated with promoter regions, only 25% of the CO hotspots overlap a promoter (Mancera et al., 2008). In humans too, CORs are low near the transcription start site (TSS), but start increasing 10 Kb away from the TSS (Coop et al., 2008). A lower COR near genes seems to be a general feature of mammals and plants (Drouaud et al., 2006; Kauppi et al., 2007; Paigen et al., 2008; Wu et al., 2010). Moreover, the use of 3.1 million human single nucleotide polymorphisms (SNPs) has revealed an asymmetry in the distribution of recombination around genes, as illustrated in figure I.9, with regions 3' of transcribed domains having more CO activity than the 5' regions (International HapMap Consortium, 2007).

A degenerated 13-mer sequence motif (CCNCCNTNNCCNC) has been identified in association with human CO hotspots (Myers et al., 2008) (figure I.10). LD studies, as well as sperm-typing analyses (Webb et al., 2008), have located the presence of the motif in **40% of the CO hotspots**. The function of this motif in humans has been associated with the binding of the zinc-finger protein PRDM9 (figure I.10) (Myers et al.,

2009). The binding sequence of **PRDM9** is an exact match of the 13-mer motif, with a degeneracy at positions 3, 6, 8, 9, and 12 and no degeneracy at the remaining 8 positions. An independent study (Baudat et al., 2009) of the recombination activity in mouse found that the gene coding for PRDM9 was located at the *double-strand break control 1* (*Dsbc1*) genetic locus which controls for the activity and distribution of recombination hotspots in this species (Grey et al., 2009). Additional to the zinc fingers, PRDM9 also contains a SET-methyltransferase domain, involved in the tri-methylation of the 4th lysine in histone 3 (H3K4me3) (Baudat et al., 2009). The H3K4me3 is associated with the initiation of recombination in both *S. cerevisae* and mouse (Borde et al., 2009; Buard et al., 2009). Despite the association between PRDM9 and hotspot activity, the exact mechanism of this interaction is not yet understood. Both in mouse and human, the sequencing of the *Prdm9* gene has revealed the existence of multiple alleles in a population, resulting in variants with variable number of Zn fingers (Parvanov et al., 2009; Berg et al., 2010). The CO activity and hotspot distribution is highly dependent on the type of *Prdm9* alleles carried by individuals. In humans, the allele associated with the 13-mer motif is termed A. It controls the motif association to recombination hotspots in different genetic backgrounds: repeat and nonrepeat DNA, male and female, as well as for generating ectopic recombination (Myers et al., 2008). While the PRDM9 variant coded by allele A is responsible for the recognition of the 13-mer motif, it has been found to trigger recombination even at hotspots depleted of the motif (Berg et al., 2010). On the other hand, other PRDM9 variants, not recognizing the motif, generate high levels of recombination at hotspots containing the motif (Berg et al., 2010). These results imply that PRDM9 might explain more than the 40% of the hotspots containing the motif (McVean and Myers, 2010). It may be that PRDM9 binds even diverged motifs, while additional flanking sequences stabilize the bond (Myers et al., 2008). Additionally, the H3K4me3 activity of PRDM9 might allow the recruitement of the recombination protein complex containing Spo11 (Baudat et al., 2010).

Finally, the comparison of *Prdm9* sequences in multiple metazoans has revealed that *Prdm9* is under an accelerated evolution (Oliver et al., 2009). A particularly high divergence rate characterizes this gene between human and chimpanzee. Given this rapid evolution, it is not puzzling that the binding sequence of the chimpanzee PRDM9 differs from the 13-mer CO hotspot motif found in humans (Myers et al., 2009; Oliver et al., 2009). If PRDM9 is indeed an attribute of CO hotspots in all species, its species-specific analysis would reveal different sequence motifs associated with hotspot activity.

I.2.2 Non-allelic homologous recombination (NAHR)

Recombination can also take place between non-allelic sequences situated at different genomic locations. This non-allelic homologous recombination (NAHR) (a.k.a ectopic recombination) occurs mainly between repeats. Low copy repeats (LCRs), resulting from the duplication of a few hundred kb long sequences, that display high sequence similarity, represent preferred NAHR sites (Bailey and Eichler, 2006). Studies of LCR have demonstrated that both allelic recombination and NAHR are similar processes (reviewed in Sasaki et al. (2010)). NAHR is initiated by DSBs and is localized in hotspots 1-2 Kb inside the LCRs. The 13-mer degenerated motif associated with allelic recombination is also indicative of NAHR hotspots (Myers et al., 2008).

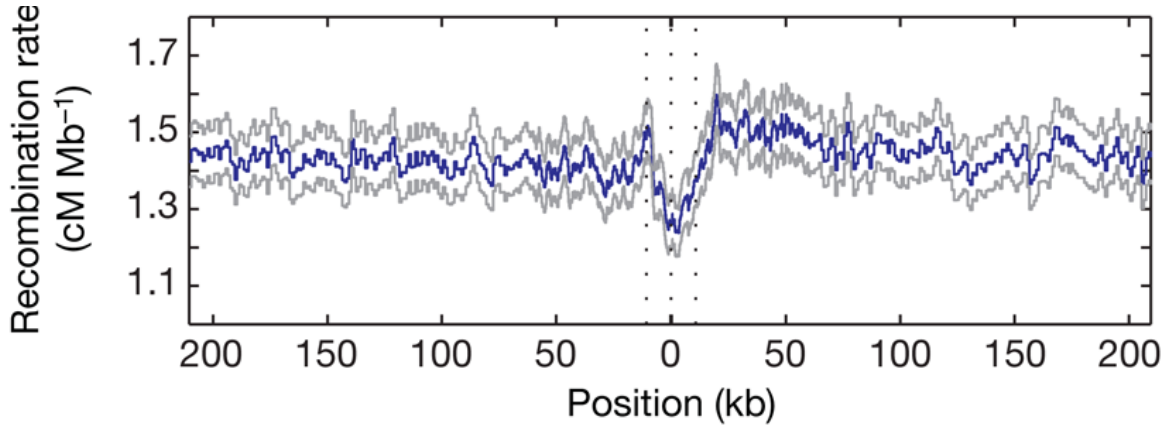

Figure I.9: *The recombination rate around genes in human. The blue line indicates the mean. Grey lines indicate the quartiles of the distribution. Values were calculated separately 5' from the transcription start site (the first dotted line) and 3' from the transcription end site (third dotted line) and were joined at the median midpoint position of the transcription unit (central dotted line). Note the sharp drop in recombination rate within the transcription unit, the local increase around the transcription start site and the broad decrease away from the 3' end of genes. Adapted from International HapMap Consortium (2007).*

Figure I.10: *The extended 39-bp human hotspot motif. This extended motif contains the estimated degeneracy of the human 13-bp hotspot motif (relative letter height proportional to estimated probability of hotspot activity, total letter height determined by degree of base specificity). In silico prediction of the binding consensus for PRDM9, aligned with the 13-mer, with more influential positions (P<0.01) shown in red. Underlined is an additional 8-bp matching sequence. The logo shows predicted degeneracy within this consensus. Below the text is the human sequence of four predicted DNA-contacting amino acids for the 13 successive human PRDM9 zinc fingers (1 oval per finger, differing colors for differing fingers, separated finger is gapped N-terminal from others), and their predicted base contacts within the motif. Adapted from Myers et al. (2009).*

Recombination between non-allelic sequences results in genomic rearrangements, such as deletions, duplications, inversions or isodicentric chromosome formation (reviewed in Sasaki et al. (2010)) (figure I.11). These rearrangements can induce genomic disorders. In human, two such disorders have been found to contain the degenerated CO hotspot motif (Berg et al., 2010). The study of the effect of *Prdm9* alleles on the frequency of rearrangements in these regions has revealed that this allelic associated recombination gene is also characteristic of NAHR hotspots (Berg et al., 2010).

I.2.3 Interference between recombination products

Keeping homologs together during the reductional division of meiosis is essential for their correct segregation (figure I.1). This role is fulfilled by COs as they first initiate the formation of the SC and keep the homologous chromosomes connected during their migration to the poles (Roeder, 1997). Thus, it has been observed that the majority of species have at least one CO per pair of homologs, also known as the obligate CO (de Villena and Sapienza, 2001). A control mechanism, known as CO homeostasis, promotes the formation of CO events at the expense of NCOs, in order to assure the obligate CO (Martini et al., 2006). Moreover, the distribution of COs is not random along the chromosomes. Instead they are subject to interference, thus ensuring a more uniform distribution and reducing the risks of non-disjunction (Bishop and Zickler, 2004).

The construction of the first genetic map in *Drosophila* allowed the first observation of **the interference phenomenon** between adjacent COs (Sturtevant, 1913). Despite the validation of CO interference (COI) in a vast majority of species (table I.1), the genetic mechanisms underlying this phenomenon are poorly understood. At first, it was thought that polymerization around the initiation sites of the SC prevented CO formation in adjacent regions (Maguires, 1988). However, the sites of SC initiation, some of which are also the sites of future COs, exhibit interference long before the assembly of the SC (Fung et al., 2004). Moreover, mutants for the spo16 gene in *S. cerevisae*, which show defects in the extension of the SC, exhibit normal distribution of interference (Shinohara et al., 2008). Cells with a defective SC in mouse are also having normal interference levels (de Boer et al., 2006).

The modern view on COI is that it is not dependent on the formation of the SC, and is probably established very close to the transition between DSBs and SEI formation (Hunter and Kleckner, 2001; Bishop and Zickler, 2004). Recent studies support the idea that interference takes place at the time when Msh4-Msh5 complexes stabilize SEI (Shinohara et al., 2008) (figure I.5). In *S. cerevisae*, mutants of *tid1* gene, coding for the Tid1 protein, involved in the regulation of strand invasion, have normal levels of COs, but the interference is greatly decreased (Shinohara et al., 2003). Dissociation of the strand invasion events, regulated by RTEL1, promotes COI by preventing adjacent DSBs to be repaired through the DSBR pathway, and generating NCOs, possibly through SDSA (Barber et al., 2008; Youds et al., 2010).

The study of COI is further complicated by the existence of two types of COs: interfering and non-interfering. The first type is generated through the DSBR pathway (Msh4-Msh5 COs), while the other uses the Mus81 pathway, as described in section I.1.4. The distribution of these two kinds of COs varies widely between species, from *S. pombe* and

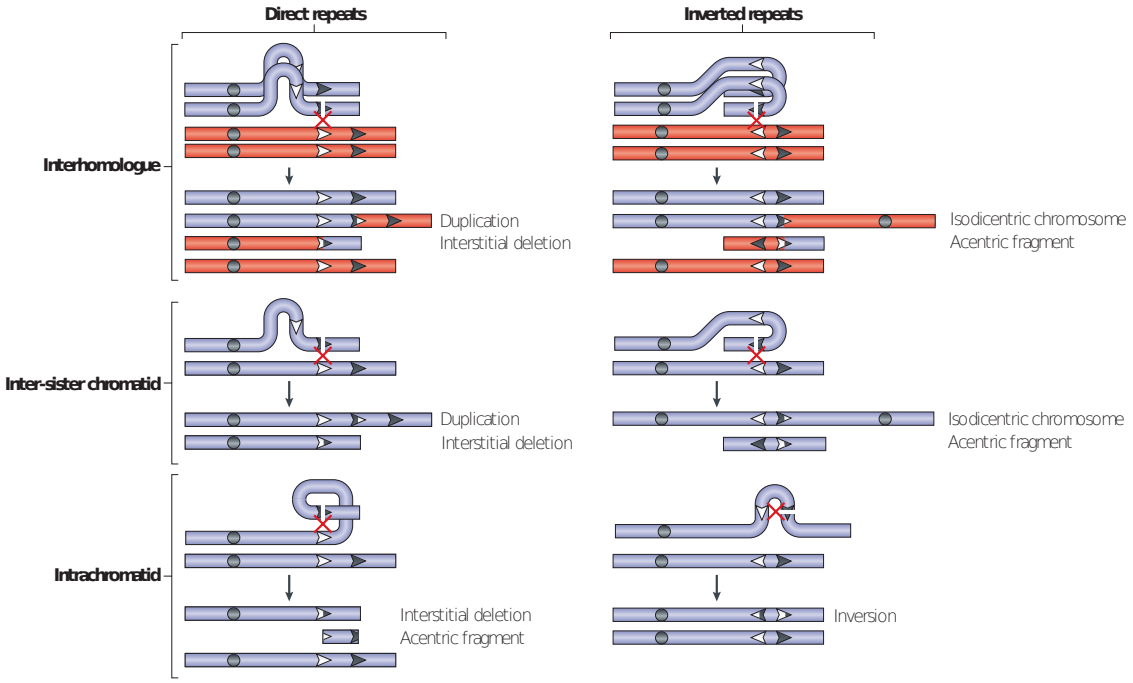

Figure I.11: *Genome rearrangement by non-allelic homologous recombination. Crossover recombination between repeated DNA sequences at non-allelic positions can generate a deletion, a duplication, an inversion or an isodicentric chromosome. Depicted here are six chromosomal outcomes of non-allelic homologous recombination (NAHR) between repeats located on the same chromosome, with two orientations of repeats relative to one another (direct or inverted) for each of three types of interactions (between homologues, between sister chromatids or in the same chromatid). Homologous chromosomes are shown in blue and red, and sister chromatids are depicted in the same colour (homologous chromosomes are not shown in the schematics depicting inter-sister chromatid or intrachromatid exchanges for simplicity). Low-copy repeats (LCRs) are shown as white and black arrowheads. From Sasaki et al. (2010).*

Aspergillus nidulans with no interference, to *C. elegans* with complete interference (table I.1). While models have been proposed and proteins have been identified for the two pathways, the mechanisms creating interference in one but not the other type of CO are still unknown. Two hypotheses have been advanced regarding the difference between the interfering and non-interfering COs (reviewed in Berchowitz and Copenhaver (2010)). The "toolbox hypothesis" postulates that Mus81-Eme1 and Msh4-Msh5 protein complexes are both recruited at all the DSB sites (Berchowitz and Copenhaver, 2010). The majority of DSBs is then repaired through the DSBR pathway, and only a subset of recombination intermediates (aberrant ones) will be later resolved by the Mus81-Eme1 protein complex. The idea of a recruitment for both protein complexes at the recombination initiation sites comes from the observation of co-localization of AtMUS81 with AtRAD51 and AtMSH4 foci at leptotene, in *A. thaliana* (Higgins et al., 2008). Moreover, mutants of the *mus81* gene in *S. cerevisae* show an accumulation of aberrant recombination intermediates (single HJ, intersister and multichromatid molecules) which prevent the correct segregation of homologs and fail to divide nuclei (Jessop and Lichten, 2008; Oh et al., 2008).

The "two-phase hypothesis", proposed by (Getz et al., 2008), supposes that COs are specialized, with one class of COs contributing to "pairing" of homologs, while the other assures their "disjunction". The "pairing" COs occur early during meiosis and are non-interfering as shown by the *ndj1* mutants, which have a delay in pairing and also a decrease in interference (Conrad et al., 1997). Additionally, the lack of non-interfering, "pairing", COs in *C. elegans* and *Drosophila* might have been responsible for alternative, CO independent pairing mechanisms between homologs (table I.1). However, if the "pairing" COs are the ones dependent on MUS81, a few inconsistencies arise: why hasn't any pairing defect been observed in *mus81* mutants? and how about the timing of MUS81, which has been reported to act late on the recombination intermediates (Jessop and Lichten, 2008; Oh et al., 2008)? It is possible that the "pairing" COs represent a new pathway of DSB repair, independent of both Mus81 and Msh4-Msh5. This possibility is supported by the existence of an average 0.85 chiasmata per cell that is not explained by either AtMSH4 or AtMUS81, in *A. thaliana* (Higgins et al., 2008). The "disjunction" COs are supposed to occur late during meiosis and be subject to interference, being dependent on MSH4.

NCO events, not associated to COs, don't seem to interfere with one another in *S. cerevisae* (Malkova et al., 2004; Mancera et al., 2008). However, the influence between COs and adjacent NCOs is more ambiguous. If interference between COs is generated by the adjacent recombination intermediates being resolved as NCOs, a negative interference is expected between the two recombination products. Biological observations of the CO-NCO distances have yielded contradictory results. While studies of discrete intervals, associated with precise loci, have found no or negative interference between COs and NCOs (Malkova et al., 2004; Getz et al., 2008), the genome wide study of recombination products, in *S. cerevisae*, found that this same distance is 13.1 Kb larger than expected by chance (Mancera et al., 2008). This might imply, that at the genome scale, NCOs inhibit the formation of COs in their vicinity. Moreover, (Berchowitz and Copenhaver, 2010) proposes that the discordance between the two types of biological results, might reflect the existence of two classes of NCOs, as for the COs, interfering and non-interfering with the single locus studies having found only the non-interfering class.

Review articles for this sub-chapter: Berchowitz and Copenhaver (2010); Martinez-Perez and Colaiácovo (2009); Yanowitz (2010); Székvölgyi and Nicolas (2010)

I.2.4 Differences in recombination

The particular distribution of recombination events and the constraint of an obligate CO suggest that recombination is subject to a strong control. Misplaced or too few recombination products can generate gametes and offspring with an abnormal number of chromosomes (aneuploidy) (Baker et al., 1976; Hunt and Hassold, 2002; Petronczki et al., 2003). Furthermore, recombination plays an important adaptive role by breaking up and reshuffling chromosome segments, thus producing novel multilocus haplotypes that serve as potential selective alternatives for adaptive evolution (Otto and Lenormand, 2002; Marais and Charlesworth, 2003). However, recombination rates display high levels of variation among individuals and species. Unraveling the mechanisms generating this variation is fundamental for the understanding of genome evolution.

Organism	N	SC?	Interf.?	Msh4-Msh5 COs?	Mus81* COs?	COs / meiosis	References
Saccharomyces cerevisiae	16	Yes	Yes	Yes	Yes	90	de los Santos et al. (2003); Argueso et al. (2004); Buhler et al. (2007); Mancera et al. (2008)
Schizosaccharomyces pombe	3	No	No	No	Yes	38	Munz (1994); Hollingsworth and Brill (2004); Cromie et al. (2006)
Neurospora crassa	7	Yes	Yes	n.d.[1]	n.d.	20	Perkins (1962); Foss et al. (1993)
Aspergillus nidulans	8	No	No	n.d.	n.d.	n.d.	Strickland (1958); Egel-Mitani et al. (1982)
Caenorhabditis elegans	6	Yes	Yes	Yes	No	6	Tsai et al. (2008)
Arabidopsis thaliana	5	Yes	Yes	Yes	Yes	10	Copenhaver et al. (2002); Higgins et al. (2004); Drouaud et al. (2007); Higgins et al. (2008)
Solanum lycopersicum	12	Yes	Yes	Likely	Likely[2]	21	Sherman and Stack (1995); Anderson et al. (2001); Lhuissier et al. (2007)
Zea mays	20	Yes	Yes	n.d.	n.d.[3]	20	Anderson et al. (2003); Li et al. (2007); Falque et al. (2009)
Drosophila melanogaster	4	Yes	Yes	No	No[4]	6	Carpenter (1975); Foss et al. (1993)
Danio rerio	25	Yes	Yes	n.d.	n.d.	25-40	Moens (2006); Kochakpour and Moens (2008)
Mus musculus	20	Yes	Yes	Yes	Yes	22-28	de Boer et al. (2006); Baudat and de Massy (2007); de Boer et al. (2007); Holloway et al. (2008)
Homo sapiens	23	Yes	Yes	Yes	Likely	50-70	Lynn et al. (2002); Tease et al. (2006); Vallente et al. (2006); Baudat and de Massy (2007); Holloway et al. (2008)

Table I.1: *CO interference comparisons across model genetic organisms. Haploid chromosome number (N) and presence or absence of the synaptonemal complex (SC) or CO interference is noted. Also shown are presence or absence of Msh4-Msh5 (interference-sensitive) and Mus81-Eme1 (interference-insensitive) mediated CO pathways. Adapted from Berchowitz and Copenhaver (2010)*

[1] n.d. - not defined

[2] The distribution of Mlh1 foci (characteristic of CO sites) suggests the existence of two types of COs: interfering and non-interfering (Lhuissier et al., 2007). The presence of Mus81-Eme1 like proteins has not been studied yet for this species.

[3] (Falque et al., 2009) applied mathematical models on the distribution of late nodules of recombination (indicative of CO products) and found that the data was explained by the existence of two types of COs: interfering and non-interfering.

[4] MEI-9 is a protein similar to MUS81 in *Drosophila*, and it has been associated with the CO production, possibly involving HJ intermediates (Yildiz et al., 2002).

I.2.4.1 Differences among species

The availability of genetic maps in a large panel of species has demonstrated that, at a large genomic scale, taxa are subject to different CO rates (COR) (table I.2). Under the condition of an obligated CO per pair of homologs, the number of chromosomes should predict the total number of COs. However, in most species, there are far more COs than the number of chromosomes. In mammals, it has been found that the number of chromosome arms is a better predictor of the total number of COs (de Villena and Sapienza, 2001; Coop and Przeworski, 2007) (figure I.12). Previous cytological observations supported this view, as, except for the acrocentric chromosomes (with the centromere near the end of the chromosome), at least one chiasmata was observed on each chromosome arm, in human (Hassold et al., 2004). It has been proposed that especially for metacentric chromosomes (centromere near the middle of the chromosome), one CO per chromosome may not be sufficient for the correct segregation of homologs (de Villena and Sapienza, 2001; Coop and Przeworski, 2007). Nevertheless, recent studies on human pedigree data have shown that proper disjunction does not require the presence of COs on each chromosome arm, and instead the obligated CO condition applies to the whole chromosome (Fledel-Alon et al., 2009).

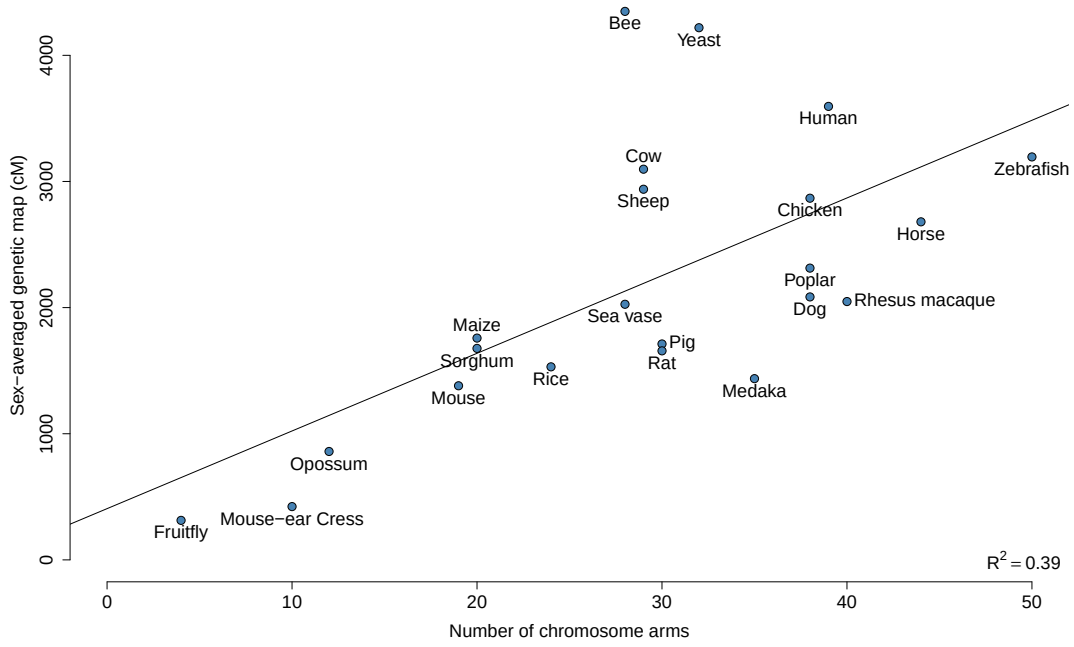

Figure I.12: *Genetic maps in vertebrates and non-vertebrates. The y-axis shows an estimate of the total sex averaged genetic-map length of autosomes. The x-axis shows the total number of autosomes arms in each species, excluding the small arms of telocentric and acrocentric chromosomes. Also shown is the line of best fit. Adapted from Coop and Przeworski (2007) with information from table I.2.*

If the karyotype can partially explain differences in the total number of chiasmata between species, additional factors are involved in the ample variation observed at a local scale. Comparison of syntenic blocks, 5 and 10 Mb long, across human, rat and

mouse found only very small positive correlation between the corresponding CORs (Jensen-Seaman et al., 2004). The intervals with different CORs between two very close mouse subspecies (< 1% sequence divergence), *Mus musculus castaneus* and *Mus musculus musculus*, cover more than 19% of the mouse genome (Dumont et al., 2010). Likewise, the analysis of the linkage disequilibrium pattern in a 14 Mb region homologous between human and chimpanzee showed no similarity between the two species in the distribution of recombination hotspots (Ptak et al., 2005). An independent study comparing the recombination rates among human and chimpanzee has confirmed that, despite the high similarity at the DNA sequence, the two species don't share the same hotspots (Winckler et al., 2005). The 13-mer degenerated motif, previously found to co-localize with CO hotspots in human, is present in higher number copies in the chimpanzee (Myers et al., 2009). However, it is inactive in all the 22 loci investigated for containing the motif in chimpanzee (Myers et al., 2009). Moreover, PRDM9, the zinc-finger protein that binds the motif, has a highly different binding sequence in chimpanzees compared with humans (Myers et al., 2009). A comparison of the *prdm9* gene between human and chimpanzee has shown a divergence level 5-fold higher than the genome-wide average (Oliver et al., 2009). This same study found a rapid evolution in the *prdm9* gene among different taxa. This rapid evolution is responsible for the fast-evolving binding sequences of the DNA motif associated with recombination hotspots (Myers et al., 2009; Oliver et al., 2009).

I.2.4.2 Differences among sexes and age classes

Linkage analyses in *D. melanogaster* have revealed for the first time sex-differences in recombination between autosomes (Morgan, 1912; Sturtevant, 1913; Morgan, 1914). The lack of DSB-mediated recombination in one sex compared to the other is termed **achiasmy**. This phenomenon has been observed for males *Drosophila* and some *Scorpionidae*, for females of some *Lepidoptera* (Miao et al., 2005; Yamamoto et al., 2006), some *Trichoptera*, and some *Crustacea*, as well as for isolated species of molluscs, water-mites, grasshoppers, and alderflies (Bell, 1982). The phylogenetic distribution of achiasmy observations indicates that this phenomenon has originated independently in at least 20 metazoan lineages (Bell, 1982; Burt et al., 1991).

More often both sexes have recombination, but one sex exhibits more recombination than the other (**heterochiasmy**). For most known heterochiasmy cases, the female has more COs than the male (table I.3). However, the majority of marsupials investigated (*Smithopsis crasicaudata*, *Macropus eugenii*, *Monodelphis domestica*) show the opposite pattern, with recombination occurring more frequently in male than female meiosis. So far, only one marsupial species has been found to deviate from this tendency, with *Bettongia penicillata* showing no significant difference in the number of chiasmata between the sexes (Hayman et al., 1990). Male-biased heterochiasmy has been reported in two more vertebrates, a mammal, the sheep (*Ovis aries*) and a bird, the flycatcher (*Ficedula albicolis*), as well as for *Arabidopsis thaliana* (Drouaud et al., 2007). The male bias in domestic sheep is not a characteristic of the *Ovis* genus, as linkage-based studies in bighorn sheep (*Ovis canadensis*), a wild relative, show a F/M ratio >1. The overall genetic map of flycatchers is longer than female, but there are also individual linkage intervals in which the female recombination is higher (Backström et al., 2008). Nevertheless, pronounced heterochiasmy

Organism	**Genetic length (cM)**	**Genome size (Mb)**	**N**	**Chromosome size (Mb)**	**CO Rate (cM/Mb)**	**References**
Apis mellifera	4348.4	∼229	16	13.57 ± 5.06	20.11 ± 1.81	Beye et al. (2006); Honeybee Genome Sequencing Consortium (2006)
Saccharomyces cerevisiae	4219	∼12	16	0.75 ± 0.35	372.56 ± 72.39	Cherry et al. (1997); Mancera et al. (2008); DB-Ensembl; DB-SGD
Homo sapiens	3595.3	∼3200	22	130.35 ± 58.47	1.33 ± 0.24	International Human Genome Sequencing Consortium (2001); Matise et al. (2007)
Danio rerio	2296.1	∼1505	25	51.08 ± 9.06	0.60 ± 0.11	Kelly et al. (2000); DB-ZFIN; DB-FishMap
Bos taurus	3097.3	∼3250	29	87.79 ± 31.84	1.26 ± 0.19	Arias et al. (2009); Zimin et al. (2009)
Ovis aries	2938.50	∼2800	26	102.14 ± 66.16	1.11 ± 0.29	Poissant et al. (2010)
Gallus gallus	2867.36 [1]	∼1050	32	35.43 ± 49.45	6 ± 5.51 [1]	Chicken Genome Sequencing Project Consortium (2004); Groenen et al. (2009)
Equus caballus	2680.1	∼2430	31	73.37 ± 34.76	1.30 ± 0.36	Swinburne et al. (2006); Horse Genome Sequencing Project Consortium (2009)
Populus trichocarpa	2313.11	∼485	19	16.20 ± 6.20	7.79 ± 2.09	Tuskan et al. (2006); DB-Ensembl; DB-NCBI
Canis familiaris	2084.8	∼2500	38	61.00 ± 21.31	0.97 ± 0.29	Lindblad-Toh et al. (2005); Wong et al. (2010)

Macaca mulatta	2048.1	∼3100	20	135.49 ± 46.33	0.78 ± 0.26	Rogers et al. (2006); Rhesus Macaque Genome Sequencing and Analysis Consortium (2007)
Ciona intestinalis	2026.4	∼116	14	7.21 ± 2.88	21.48 ± 14.45	Dehal et al. (2002); Kano et al. (2006)
Zea mays	1758.3	∼2300	10	204.63 ± 49.12	0.87 ± 0.11	Anderson et al. (2003); Schnable et al. (2009); Sen et al. (2010); DB-NCBI; DB-MaizeSequence
Sus scrofa	1711.8	∼2400	18	118.71 ± 55.05	0.85 ± 0.41	Vingborg et al. (2009); Archibald et al. (2010)
Sorghum bicolor	1676.61	∼730	10	65.92 ± 7.35	2.53 ± 0.34	Kim et al. (2005); Paterson et al. (2009)
Rattus norvegicus	1657.3	∼2700	20	127.91 ± 59.89	0.71 ± 0.19	Rat Genome Sequencing Project Consortium (2004); S. T. A. R. Consortium (2008)
Vitis vinifera	1646.81	∼498	19	15.95 ± 4.83	5.77 ± 1.85	Doligez et al. (2006); Jaillon et al. (2007); DB-Ensembl; DB-NCBI
Trypanosoma brucei	1556.94	∼ 35	11	2.37 ± 1.37	61.32 ± 14.62	MacLeod et al. (2005); Jackson et al. (2010); DB-NCBI
Plasmodium falciparum	1542.6	∼23	14	1.63 ± 0.72	66.96 ± 6.07	Su et al. (1999); Gardner et al. (2002); DB-Ensembl
Oryza sativa	1530.4	∼500	12	30.89 ± 5.94	4.11 ± 0.26	Harushima et al. (1998); Project (2005); DB-Ensembl; DB-NCBI
Oryzias latipes	1436.8	∼725	24	30.17 ± 4.16	1.98 ± 0.25	Ahsan et al. (2008)

Mus musculus	1380.42	∼3400	19	130.12 ± 32.98	0.57 ± 0.08	Gregory et al. (2002); Cox et al. (2009)
Cryptococcus neoformans	1365.6	∼ 20	14	1.38 ± 0.48	78.29 ± 29.20	Marra et al. (2004); DB-NCBI
Monodelphis domestica	859.5	∼3600	8	427.88 ± 169.15	0.25 ± 0.03	Mikkelsen et al. (2007); Samollow et al. (2007)
Arabidopsis thaliana	422.5	∼119	5	23.87 ± 4.96	4.74 ± 0.84	Initiative (2000); Singer et al. (2006)
Drosophila melanogaster	313.3	∼169	4	19.59 ± 10.49 [2]	3.04 ± 1.31 [2]	Myers et al. (2000); DB-FlyBase; DB-Ensembl
Caenorhabditis elegans	231.47	∼100	5	16.51 ± 2.81	2.87 ± 0.55	C. elegans Sequencing Consortium (1998); Rockman and Kruglyak (2009); DB-AceDB

[1] The genetic map is available for the first 28 chromosomes.

[2] Chromosomes 2R, 2L, 3R, 3L.

Table I.2: *CO number and rate comparisons across some model genetic eukaryotes. Sex-averaged total genetic length, genome size, autosome number (N), average and standard deviation of chromosome size and the crossover rate (COR) between autosomes. Information on the genome size and karyotype DB-Ensembl.*

is not an universal feature of all species as the cow (Barendse et al., 1997) and several galliform birds display no or very small differences in recombination between sexes (Reed et al., 2005; Groenen et al., 2009).

Heterochiasmy is also present at a local level, with differential CO distribution along the chromosomes. In mouse and human, male COR is higher near telomeres, while in the rest of the genome the F/M COR is > 1, especially near centromeres (Shifman et al., 2006; Broman et al., 1998; Kong et al., 2002; Paigen et al., 2008; Wong et al., 2010) (figure I.13). It has been suggested that the higher male COR near telomeres in these species is a consequence of the necessity to form the *bouquet* configuration in a shorter period of time in this sex (reviewed in Paigen and Petkov (2010)). Cytological observations of chiasma distribution suggest that in opossum (*Monodelphis domestica*) the pattern is reversed, with female COs being concentrated near telomeres as opposed to a more uniform distribution in males (Sharp and Hayman, 1988). On the other hand, in most linkage intervals COR are similar among the sexes in domestic sheep, but male COR is highly increased in both subtelomeric and pericentric regions (Poissant et al., 2010). The new human genetic map (deCode 2010) has revealed that **high COR regions correspond to intergenic sequences in female, and genic sequences in male** (Kong et al., 2010).

No universal explanation has been found to describe the particularities of heterochiasmy in different species. Nevertheless, several hypotheses have been formulated. Based on the observation that in the vast majority of achiasmate species, the sex devoid of recombination is also the heterogametic one, Haldane and, later, Huxley (reviewed in Lenormand and Dutheil (2005)) proposed that achiasmy is linked to the suppression of recombination between the sex chromosomes. However, the study of sex differences in an extensive dataset yielded no link between heterochiasmy and the type of sex chromosomes (Lenormand and Dutheil, 2005). The authors propose an explanation of heterochiasmy based on differences in selection between the two sexes. This difference in selection is mainly generated by situations in which a haploid locus, rather than diploid alleles, generates phenotypic heritable traits. For example, in eutherian mammals, the female, as opossed to male, meiosis is taking place close to fertilization and the time spend in haploid phase is very restrained. This has led to the hypothesis that the sex experiencing more selection at the haploid stage (the male in eutherian mammals) will have a reduced COR in order to minimize the recombination load (mean reduction in fitness due to recombination) induced by the shuffling of previously defined combinations of genes (Lenormand and Dutheil, 2005). The extent to which haploid selection can explain such great differences between sexes is still unclear, since in mouse at most 3.3% of the genome may be under haploid selection (Joseph and Kirkpatrick, 2004). Following the same line of thought, sex-differential expression of epistatically interacting genes during meiosis might explain the local variability in the distribution of COs between sexes (Lenormand, 2003). Factors influencing the sex-differential CORs in humans include an inversion on chromosome 17 which results in an increase in female recombination activity in hotspots (Stefansson et al., 2005) and the *RNF* gene on chromosome 4, which, depending on its polymorphic variants, influences the total CO events in male and female (Kong et al., 2008). Whatever the biological explanation of heterochiasmy, the differences in COR between sexes have been found to correlate with differences in the formation of the synaptonemal complex (SC)

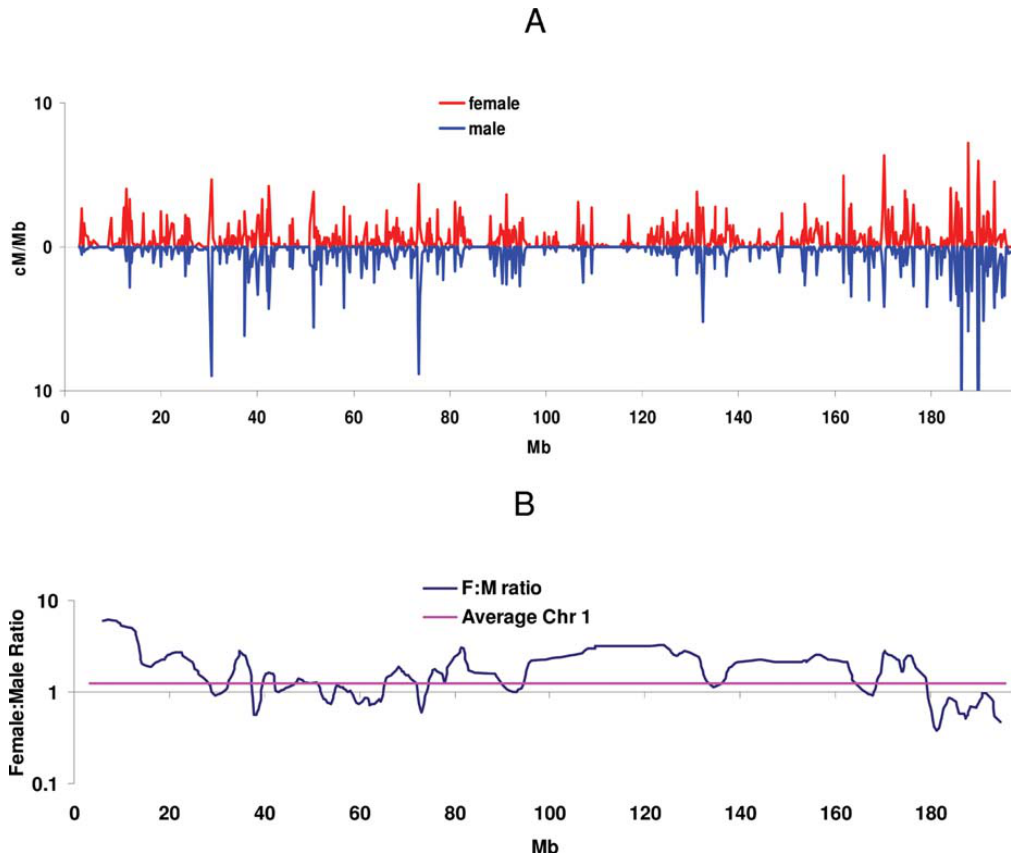

Figure I.13: *Sex specificity of mouse recombination. A. Sex-specific recombination map of chromosome 1. Red line, female recombination rates; blue line, male recombination rates. B. Female:male ratio along the chromosome. Dark blue line - female:male ratio; purple line - sex-averaged recombination rate over the entire chromosome 1. Chromosome 1 in mouse, as all other chromosomes of this species, is acrocentric. At the far left part of the graph is the centromere and at the right part, the telomere. From Paigen et al. (2008).*

(section I.1.2). In humans and mice, **the SC is longer in females than in males** (Lynn et al., 2002; Tease and Hultén, 2004), with the average μm distance between COs being the same for the two sexes. This results in a smaller Mb interference distance in females, giving rise to more COs and a higher overall COR (Petkov et al., 2007; Paigen et al., 2008).

The examination of CO hotspots in human and mouse has revealed that a majority of hotspots are active in both sexes but used with different intensities (Coop et al., 2008; Paigen et al., 2008). A recent high density genetic map in humans has revealed that 15% of hotspots are sex-specific (Kong et al., 2010). However, even if a hotspot is specific of one sex, the COR of the other sex is also higher relative to the local COR. Thus, it seems that in these two mammals, females use more frequently low and medium intensity hotspots, while males prefer to use intensly a smaller subset of hotspots (Paigen et al., 2008).

Another variable affecting recombination in humans is the age of women. The age factor has been identified following the positive correlation between maternal age and the rate of aneuploidy (Hassold and Hunt, 2001). The age effect does not seem to result from fewer recombination events, but rather from multiple errors during the two phases of

meiosis (Lamb et al., 2005; Coop and Przeworski, 2007; Hunt and Hassold, 2008). Indeed, pedigree studies in humans have demonstrated that the percentage of correct segregation of chromosome 21 in the absence of COs is frequent (Fledel-Alon et al., 2009). The causes for aneuploidy during maternal meiosis are multiple, as errors at different stages of oocyte development can result in bad disjunction of chromosomes (Hunt and Hassold, 2008). Thus, it seems that not the number but failure of the maternal check-point mechanisms of meiosis, which in turn may be age-dependent, might be responsible for outcomes such as trisomies (Hunt and Hassold, 2008; Fledel-Alon et al., 2009). In view of these results, the observation that mothers over 35 years old have a COR on average 3.1 higher than mothers under 35 years old suggests that there is selection on the number of COs in order to balance the decrease in meiotic efficiency linked to maternal age (Kong et al., 2004; Coop et al., 2008). In all these studies, the nondisjunction errors are associated only with maternal, and not paternal, age (Kong et al., 2004; Coop et al., 2008; Allen et al., 2009).

I.2.4.3 Differences among individuals of the same species

Cytogenetic methods for labeling meiotic proteins such as MLH1, have identified significant variation in the number of total exchanges per cell between human males (reviewed in Lynn et al. (2004)). The use of linkage-based analyses in human families have also revealed inter-individual differences (Broman et al., 1998; Kong et al., 2002; Cheung et al., 2007). In initial studies, the small number of individuals and/or meiosis per individual tested, have suggested that the differences in recombination number and activity concerned only females (Broman et al., 1998; Kong et al., 2002). However, this result was contradicted by recent linkage studies which showed that even if there is more variability among women, both men and women show inter-individual recombination variation (Cheung et al., 2007). These differences do not concern only the total number of events but also variations in COR along chromosomes. Furthermore, another linkage-based study in a Hutterite population, showed that individuals (either male or female) have differential use of LD-based hotspots, and this preferential usage is heritable (Coop et al., 2008).

Sperm-typing techniques allow the investigation of recombination activity in individual hotspots among different male individuals. A region containing the major histocompatibility complex (MHC) shows a 2-fold difference in COR among 5 men (Yu et al., 1996). The human NID1 hotspot of recombination exhibits individual polymorphism in the rates of both COs and NCOs (Jeffreys and Neumann, 2005). The analysis of two other recombination hotspots, MSTM1a and MSTM1b, described for the first time the existence of a hotspot, MSTM1a, which was active in only a few men and completely inactive in the rest (Neumann and Jeffreys, 2006). Multiple PRDM9 alleles have been identified in humans, accounting for the differential hotspot usage among individuals (Baudat et al., 2009). Moreover, PRDM9 alleles explain the polymorphism in recombination even in the absence of the specific 13-mer motif, suggesting a more global role of this protein in the creation of hotspots (Berg et al., 2010) (figure I.14).

Review articles for this sub-chapter: Lynn et al. (2004); Arnheim et al. (2007); Coop and Przeworski (2007); Paigen and Petkov (2010).

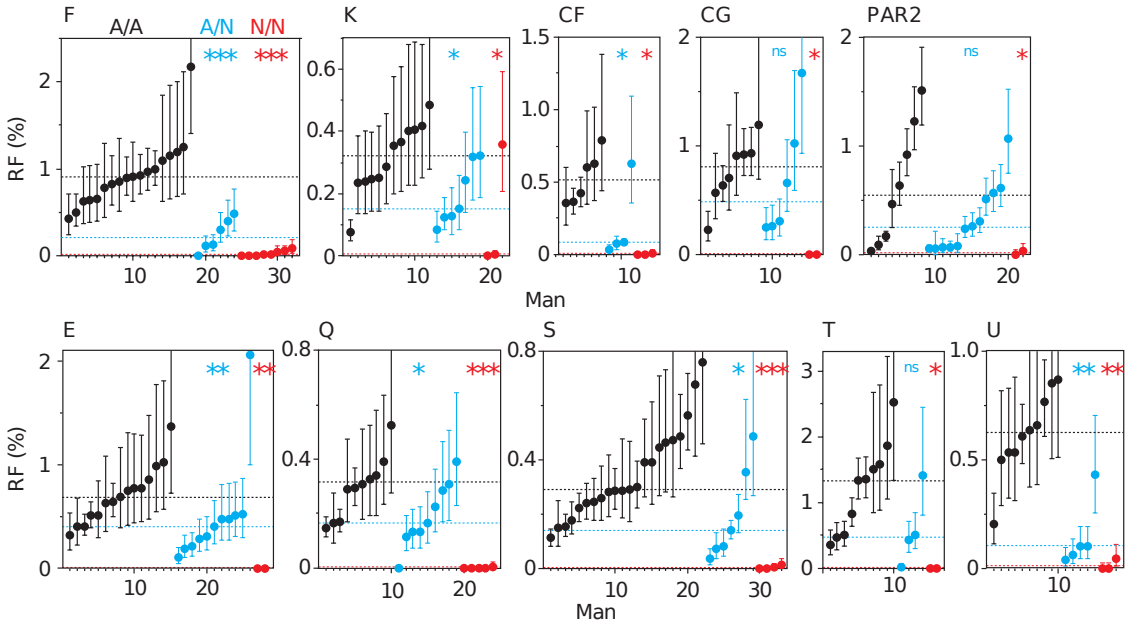

Figure I.14: *PRDM9 variants and CO hotspot activity in human sperm. The allele A of* Prdm9 *gene generates the PRDM9 variant associated with the 13-mer motif of human hotspots. The other alleles are termed N. Three type of individuals have been sequenced: individuals with two A alleles (A/A, shown in black), only one A allele (A/N, shown in blue), and two non-A alleles (N/N, shown in red). Upper part: the examination of 5 recombination hotspots, containing a central 13-mer motif. For each hotspot, the recombination frequency has been estimated in multiple individuals. The confidence intervals for each estimate of recombination frequency are shown, and median recombination frequencies within each group are indicated by dotted lines. The significance of differences between the A/A group and the A/N or N/N groups, for the Mann-Whitney test, are given at the top right (ns, not significant, $P > 0.05$; $*P < 0.05$; $**P < 0.01$; $***P < 0.001$). Lower part: corresponding analyses for five hotspots laking the 13-mer motif. Adapted from Berg et al. (2010).*

I.3 Biased gene conversion (BGC)

I.3.1 Meiotic drive

So far, 32 996 putative hotspots of recombination have been identified from LD maps (International HapMap Consortium, 2007), 72% of which overlap with COs observed in pedigree-based studies in the Hutterite population (Coop et al., 2008). Another pedigree study in humans, involving diverse populations, has identified a total of 6 938 CO hotspots (Kong et al., 2010). However, only 46 recombination hotspots have been studied in detail in humans and only in males, using sperm-typing techniques (additional table B). By typing the exact chromosome sequence at the CO junction in many individuals for each one of the 46 CO hotspots, a striking observation emerged: for some hotspots the transmission of alleles was not symmetric (Jeffreys and Neumann, 2002; Arnheim et al., 2007). This asymmetry affects both the segregation ratio of alleles as well as the position of the CO

Organism	**Data**	**♀(cM or nbr of chiasma)**	**♂ (cM or nbr of chiasma)**	**Ratio**	**Reference**
Danio rerio	LM	2582.7	942.5	2.74	Singer et al. (2002)
Felis silvestris catus	LM	5710.7	3113.7	1.83	Menotti-Raymond et al. (2009)
Takifugu rubripes	LM	1213.5	697.1	1.74	Kai et al. (2005)
Sus scrofa	LM	2336.1	1441.5	1.62	Vingborg et al. (2009)
Homo sapiens	LM	4401.2	2831.8	1.55	Matise et al. (2007)
Canis familiaris	LM	2276.3	1909.1	1.19	Wong et al. (2010)
Ovis canadensis	LM	3159.3	2824.9	1.11	Poissant et al. (2010)
Gallus gallus	LM [1]	3015.3	2778.3	1.09	Groenen et al. (2009)
Mus musculus	LM	1495.3	1375.3	1.09	Cox et al. (2009)
Taeniopygia guttata	LM	1330	1304	1.02	Backström et al. (2010)
Bos taurus	LM	3586.5	3530.6	1.02	Barendse et al. (1997)
Oryzias latipes	LM	1455.2	1453.5	1	Kimura et al. (2005)
Anas platyrhynchos	LM	1387.6	1415	0.98	Huang et al. (2006)
Ovis aries	LM	2720.4	3178.4	0.86	Poissant et al. (2010)
Ficedula albicolis	LM	1627	1982	0.82	Backström et al. (2008)
Macropus eugenii	LM	ND	ND	0.78	Zenger et al. (2002)
Smithopsis crasicaudata	CC	10.2	13.6	0.75	Burt et al. (1991)
Monodelphis domestica	LM	515.5	948.2	0.54	Samollow et al. (2007)

Table I.3: *Sex CO number and rate comparisons across some model genetic vertebrates. For each sex, the total genetic length (cM) or the total number of observed chiasma of autosomes and the recombination ratio between female and male. LM stands for linkage mapping and CC for chiasma count. The results are ordered according to the ♀:♂ ratio. Adapted from Lenormand and Dutheil (2005).*

[1] The genetic map is available for the first 28 chromosomes.

junction. A graphical illustration of such an asymmetric CO is represented in figure I.15. The junction position is not the same in all individuals, as it is displaced by a few to a few hundred bp in different individuals. This displacement is due to one allele initiating preferentially the DSB (the red allele in figure I.15), and thus being systematically replaced by the other allele. As many as 11 out of the 46 human COs described by sperm-typing are asymmetric (additional table B). As illustrated in figure I.15, the asymmetry results in the differential transmission of alleles at one or multiple polymorphic sites. For example, the segregation distortion at DNA2 hotspot is affected by two polymorphic sites, FG11G/A and FG5AT (Jeffreys et al., 2001; Jeffreys and Neumann, 2002). For the FG11 and FG5 sites respectively, 93% and 81% of the COs, carried alleles FG11G and FG5A instead of FG11A and FG5T (Jeffreys and Neumann, 2002). This results in a mean distortion of 87:13. Asymmetric profiles are also present in mouse and with an apparent higher frequency than in humans (reviewed in Wu et al. (2010)). Moreover, the variation in the transmission of alleles at one site can be inherited (Jeffreys and Neumann, 2009).

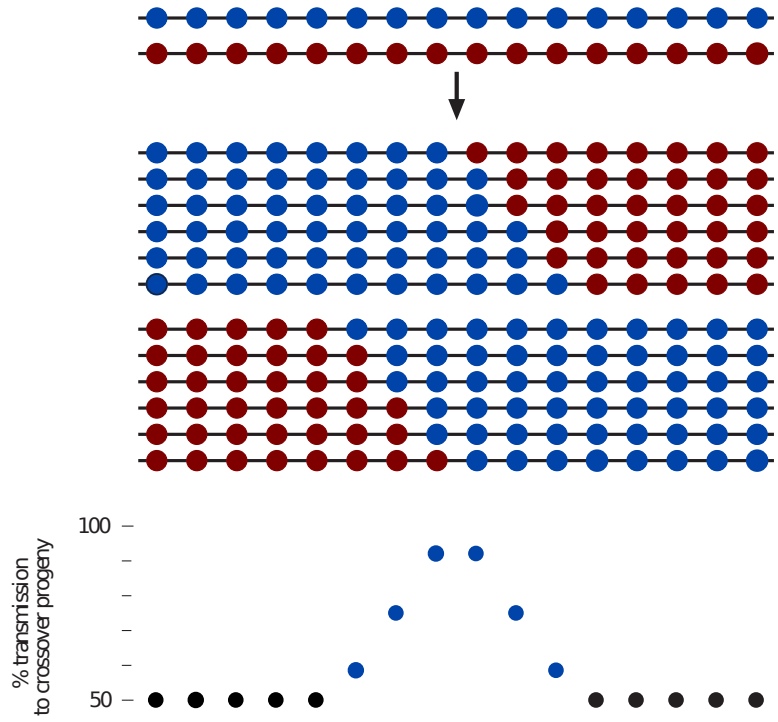

Figure I.15: *Example of reciprocal crossover asymmetry and meiotic drive in a recombination hotspot. Asymmetry arises if reciprocal exchanges between the blue and red haplotypes occur at the same rate but show exchange points mapping to different locations. If blue exchanges are displaced relative to red exchanges, then crossover progeny will show non-Mendelian over-transmission of alleles from the blue haplotype for markers closest to the center of the hotspot, as indicated in the lower part of the graph. In this part of the graph, blue circles represent blue alleles that deviate from the 50% segregation ratio, while black circles represent alleles with a 50% transmission rate. Adapted from Jeffreys et al. (2004).*

This preferential transmission of one allele over the other during recombination has

been termed meiotic drive or biased gene conversion (BGC). The choice of the haplotype for initiating recombination can generate the deviation from 50: 50 in the transmission of alleles during gene conversion (Jeffreys and Neumann, 2002). Moreover, this bias has an effect at larger scales as the main mismatch repair (MMR) mechanism during recombination involves the degradation of long patches of DNA and re-synthesis, leading to the loss of information on the strand containing the DSB (Surtees et al., 2004). This can lead to asymmetric COs (Duret and Galtier, 2009). Important implications result from the initiation bias, since in hotspots under segregation distortion, the initiating allele is going to be replaced in the population. If no other *trans*-factor controls the hotspot activity, this replacement will result in the reduction of CO activity and the eventual loss of the hotspot. This phenomenon is known as the hotspot-paradox, thus explaining the death of recombination hotspots (Boulton et al., 1997; Coop and Przeworski, 2007) and their lack of conservation between closely related species, such as human and chimpanzee (Ptak et al., 2005; Winckler et al., 2005). New insights in the birth of recombination hotspots have been provided by the recent studies of the *Prdm9* gene, which is associated with recombination activity. The variants of *Prdm9* link sequence motifs that mark recombination hotspots. The rapid evolution of this gene leads to changes in the motifs linked to hotspot, thus, generating new hotspots along the sequences and offering a solution to the hotspot-paradox (Baudat et al., 2010).

I.3.2 The molecular mechanism of GC biased gene conversion

Another bias that can result in the uneven transmission of alleles acts on the repair of single base mismatches during recombination. The repair of DSBs results in the juxtaposition of homologous chromatids, thus generating heteroduplex DNA, during both CO and NCO formation. Holliday observed that these heteroduplexes contain mismatches (Holliday, 1964). Indeed, if the pairing region between homologs contains a polymorphic site, this site will be interpreted as a mismatch. Even before the discovery of proteins involved in mismatch repair, Holliday proposed that the repair of these mismatches might explain the nonreciprocal transmission of genetic information during gene conversion (Holliday, 1964; Liu and West, 2004). Experimental studies in simian and human mitotic cells have revealed that the repair of single-base mismatches is biased towards Guanine (G) and Cytosine (C), rather than Adenine (A) and Thymine (T) (Brown and Jiricny, 1989), leading to GC biased gene conversion (gBGC) (Duret and Galtier, 2009) (figure I.16). A GC bias repair is also affecting G/T and C/A mismatches in Chinese hamster ovary cells (Bill et al., 1998). In yeast, the analysis of mitotic, as well as meiotic heteroduplex mismatch repair, is also indicative of a GC bias (Birdsell, 2002). Transfections of single mismatch containing DNA, in somatic or germ cells revealed that the bias in repair mechanisms is widespread among yeast, *Xenopus* and mammals (reviewed in Marais (2003)). Moreover, the genome-wide analysis of CO and NCO sequences in yeast showed that both recombination products are associated with an increase in the GC-content of the converted sequences (Mancera et al., 2008).

In mitotic cells, it has been suggested that the bias is associated with the base excision repair (BER) mechanism (Brown and Jiricny, 1989). BER involves the use of DNA glycosylases, which remove nucleotides from damaged DNA sites (Krokan et al., 2000).

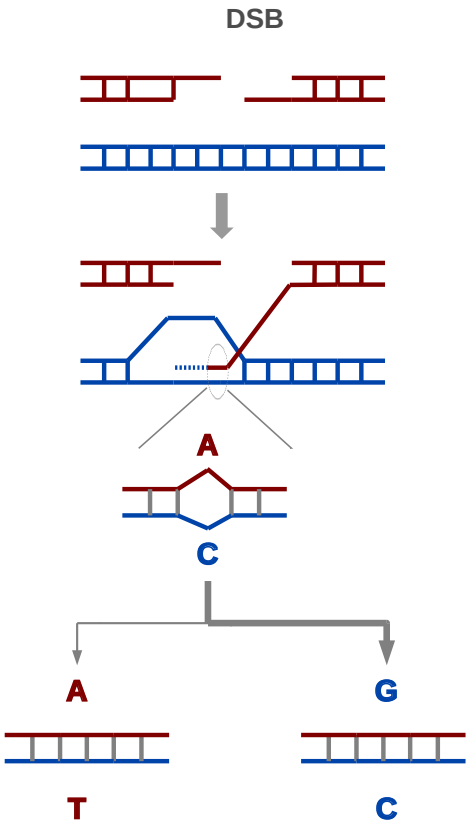

Figure I.16: *The model of gBGC. In this example, recombination is initiated through a DSB on the maternal (red) chromatid. The repair of the DSB leads to the invasion of the homologous, paternal (blue) chromatid. A heteroduplex is formed. This heteroduplex may cover a polymorphic site, an A on the maternal and a C on the paternal chromatids respectively. This polymorphism is going to be interpreted as a mismatch and repaired preferentially towards G and C rather than A and T.*

The gap is then re-filled by a DNA polymerase using the undamaged base. If during BER, the base that is removed is more often an A or T rather than C or G, this can lead to a GC bias. Since in a majority of organisms DNA glycosylases have a specificity for deaminated bases, some theories argue that the GC bias in BER has a counteracting effect for the deamination of methylated cytosines to thymines, which induce a high AT biased mutation rate (Brown and Jiricny, 1987; Birdsell, 2002; Fryxell and Zuckerkandl, 2000). However, even if BER associated enzymes are active in human and rat germ cells (Olsen et al., 2001), the extent to which BER is involved in the repair of heteroduplex associated mismatches during meiotic recombination is still unknown (Duret and Galtier, 2009).

I.3.3 Genomic evidence for gBGC

Under the model of gBGC, GC substitutions are favored, leading to an increase in the local GC-content of sequences, thus resulting in a strong correlation between the recombination rate and the GC-content (Meunier and Duret, 2004). Several genomic analyses favor

this prediction. Studies of single nucleotide polymorphism (SNP) in human populations reveal that AT→GC mutations segregate at a higher frequency than GC→AT in human noncoding regions, consistent with an explanation involving gBGC or selection (Lercher et al., 2002). Furthermore, the association between SNPs and recombination is indicative of an increase in G and C frequencies close to CO hotspots (Spencer et al., 2006). Comparison of human and chimpanzee sequences demonstrate that AT→GC substitutions tend to cluster in regions close to telomeres, characterized by high COR (Dreszer et al., 2007). The pattern of substitutions in 1 Mb noncoding windows along the human genome is also indicative of a GC bias linked to high recombination rates (Duret and Arndt, 2008). Moreover, strong correlations have been found between COR and GC-content in mammals, birds, turtle, nematode, *Drosophila*, paramecium, green alga and plants (reviewed in Duret and Galtier (2009)). Figure I.17 illustrates the strong correlations between these variables in human and chicken.

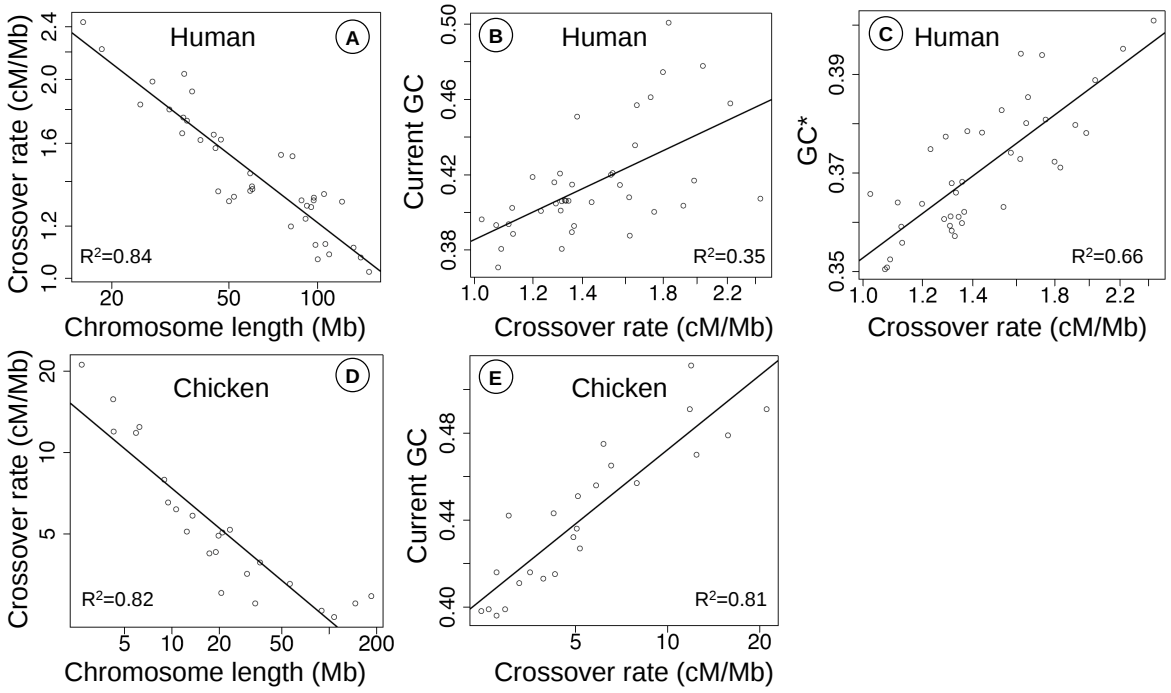

Figure I.17: *Correlations between chromosome length, crossover rate and GC-content in human and chicken autosomes. The stationary GC-content (GC*), is the GC-content that would be reached by a sequence under a constant substitution pattern. It is a statistics summarizing the matrix of substitution. Chromosome length and crossover rates are plotted in Log scale. Regression lines and Pearson's correlation coefficients (R2) are indicated. From Duret and Galtier (2009).*

The GC bias in gBGC is also affected by non-allelic homologous recombination (NAHR) (section I.2.2). In mouse and human, multigene histone families which undergo frequent non-allelic recombination have also an increased GC-content compared to single gene families which probably do not experience NAHR (Galtier, 2003). This result holds true for the *Hsp70* gene family in human and mouse (Kudla et al., 2004), the multicopies gene *HINTW* in birds (Backström et al., 2005) and the *Bex* gene family in mammals

(Zhang, 2008). An interesting example of the impact of recombination on the nucleotide composition comes from the *Fxy* gene (Galtier and Duret, 2007) (figure I.18). This gene is situated in the X-specific region in human, rat and *Mus spretus*. But in mouse, *Mus musculus*, it has been recently (less than 3 million years (Myr)) translocated such that it partially overlaps the pseudosomal region (PAR) on chromosome X. PAR is characterized by a high rate of recombination (Soriano et al., 1987), which has led to a rapid increase in the GC-content of the *Fxy* portion overlapping it. As hypothesized by the gBGC model, such an increase is the result of a high substitution rate, with all 28 amino acid substitutions in *M. musculus* being caused by AT→GC nucleotide substitutions.

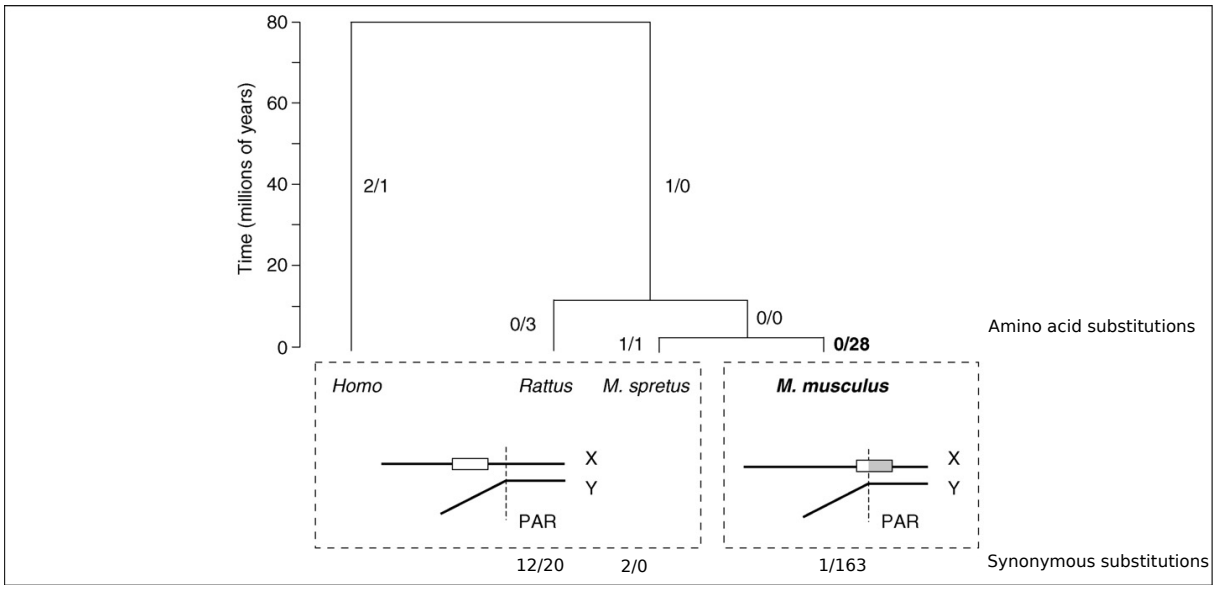

Figure I.18: *Evolutionary history of Fxy in mammals. The Fxy gene, 667 amino acids long, was translocated into M. musculus from an X-linked position to a new position, in which it overlaps the pseudoautosomal boundary (inset boxes). The time scale is given in millions of years. For each branch, the numbers of amino acid changes that have occurred in the 5' and 3' ends of the gene, respectively, are given. A strong increase in amino acid substitution rate occurred in the M. musculus lineage for the translocated fragment only. For comparison, the estimated numbers of synonymous substitutions in the Rattus, M. spretus and M. musculus branches are 12, 2 and 1 (respectively) for the 5' end of the gene, and 20, 0 and 163 (respectively) for the 3' end. From Galtier and Duret (2007).*

In agreement with a relation between recombination and nucleotide composition, differences in COR reflect differences in the GC-content. The length of chromosomes has been found to explain differences in COR among species (section I.2.4.1). Under the obligate CO condition, the COR is inversely proportional to the length of the chromosome. In chicken and zebra finch, species with a large panel of chromosome sizes, microchromosomes have high COR and are GC-rich, while longer chromosomes have lower COR corresponding to a decrease in the GC-content (figure I.17) (Chicken Genome Sequencing Project Consortium, 2004; Groenen et al., 2009; Backström et al., 2010). At the other extreme, the opossum has only 8 very long chromosomes, with a very small COR and low GC-content (Mikkelsen et al., 2007). Heterochiasmy has also an impact on the relation between COR and nucleotide composition. In humans, the females have more COs than males, however,

the male, rather than the female, COR is a better predictor of the GC content (Webster et al., 2005; Duret and Arndt, 2008). This puzzling result will be discussed in view of our results in chapter IV.

Despite such large amount of evidence in favor of gBGC, several observations constitute exceptions to this model. COR correlates negatively with the GC-content along chromosome 4 in *Arabidopsis* (Drouaud et al., 2006). A set of Y-linked, non-recombining genes have elevated values of GC (Eyre-Walker and Hurst, 2001). A strong correlation between recombination rates and GC-content is also found in yeast (Marsolier-Kergoat and Yeramian, 2009). However, in this species, the AT→GC substitution pattern is not correlated with recombination. This result has been interpreted in favor of the hypothesis that it is the high GC-content of sequences that promotes recombination, such as GC-rich regions might represent sites that favor a chromatin structure that is open to the recombination machinery (Gerton et al., 2000; Petes, 2001; Blat et al., 2002; Petes and Merker, 2002).

I.3.4 Impact of gBGC on the genomic landscape: isochores

The mechanism of gBGC has emerged, alongside mutation, selection and genetic drift, as an important evolutionary force in shaping the structure of genomes. Genomes are characterized by a strong variability in base composition. Density gradient centrifugation techniques allow the separation of DNA fragments according to their GC-content (Macaya et al., 1976; Bernardi et al., 1985). These techniques have led to the discovery of **isochores** in mammals. Isochores were defined as long (>300 kb), well-delimited (Bernardi et al., 1985; Fukagawa et al., 1995), DNA sequences, with a relatively homogeneous base composition (homogeneous GC-content). While their length, degree of homogeneity and boundaries have been redefined according to the new available genomic analyses in different species, proofs of their existence have emerged in mammals (reviewed(Costantini et al., 2009)), birds (Costantini et al., 2007b), but also reptiles, amphibians (reviewed in Costantini et al. (2009)), fishes (Costantini et al., 2007a; Melodelima and Gautier, 2008) and non-vertebrates (Cammarano et al., 2009).

According to their GC-content, families of isochores have been defined, ranging from low- to high-GC. In humans, the average GC-content in isochores ranges from 33% to 60%, leading to the definition of 5 isochore classes (L1, L2, H1, H2 and H3) (Bernardi et al., 1985). The isochore organization reflects differences in genomic features. Thus, GC-rich isochores correspond to high gene density regions (Mouchiroud et al., 1991; Zoubak et al., 1996), compact genes enriched in short introns (Duret et al., 1995; Dunham et al., 2003), high Alu and low LINE insertions (Soriano et al., 1983; Smit, 1999; International Human Genome Sequencing Consortium, 2001), early DNA replication timing (Federico et al., 1998; Watanabe et al., 2002), and high CORs (Eyre-Walker, 1993; Fullerton et al., 2001; Meunier and Duret, 2004). In view of these correlations, the molecular mechanisms responsible for the origin of isochores are important for understanding genome evolution.

While similar isochore structures seem conserved in eutherian mammals and birds (Costantini et al., 2009, 2007b), fish and amphibian genomes are more homogeneous. This has led to the interpretation that compositional heterogeneity evolved in the ancestor of amniotes (Hughes et al., 1999) and that GC-rich isochores are specific of warm-blooded

species, such as mammals and birds (Bernardi, 1993). This interpretation is based on the hypothesis that RNA and proteins benefit from a higher thermo-stability in a GC-rich context (Bernardi, 2007). However, the sequencing of ectothermic species has revealed that these genomes are also subject to heterogeneities (Hughes et al., 1999). Even among mammals, the isochore structures as highly different, with mouse and opossum having a lower, more homogeneous genomic GC-content than other mammals (Mouchiroud et al., 1988; Gregory et al., 2002; Mikkelsen et al., 2007). Moreover, no empirical evidence has been found in support of this hypothesis and it fails to explain the GC-enrichment in non-coding regions which are not under selective pressures (reviewed in Duret and Galtier (2009)). Another possible explanation was a differential bias in the mutation pattern along the genome (Wolfe et al., 1989). However, it is the bias in the fixation of mutations and not their type that is associated with genomic GC heterogeneities (Eyre-Walker and Hurst, 2001). Thus, the gBGC hypothesis has emerged as the most probable explanation for the apparition of isochores (Holmquist, 1992; Eyre-Walker, 1993; Duret and Galtier, 2009).

The rapid evolution of recombination rates along genomes has been proposed to mirror the evolution of GC heterogeneity (Duret and Galtier, 2009). Indeed, isochores are also dynamic structures as the GC-content of rich isochores has been found to decrease, leading to the hypothesis of "erosion" in some mammals (Duret et al., 2002). This "erosion" could be explained by multiple rearrangements events, such in the muridae lineage, that will affect the rates of recombination and thus, decrease the impact of gBGC (Mouchiroud et al., 1988; Rat Genome Sequencing Project Consortium, 2004; Duret and Galtier, 2009). Furthermore, fusions, such as those having occured in the opossum branch, lead to an increase in chromosome size and consequently may account for a decrease in COR and gBGC (Mikkelsen et al., 2007; Samollow et al., 2007). A recent study involving 33 species of mammals has confirmed the "erosion" of isochores in primates and muridae, however, this process seems to affect only the 20% most GC-rich genes (Romiguier et al., 2010). In other species, such as tenrec, shrew, microbat, and rabbit, even GC-rich genes show an increase in their GC-content, suggesting that the "erosion" of isochores is not an universal phenomenon.

In addition to the large-scale genomic impact of gBGC, this neutral molecular process also affects functional sequences. Thus, recombination hotspots have been named metaphorically the Achilee's heel of our genome, as they can promote the fixation of slightly deleterious AT→GC mutations (Galtier and Duret, 2007). In primates, there is evidence that gBGC is driving the fixation of deleterious mutations in proteins (Galtier et al., 2009). This same result was observed in some grass species (Glémin et al., 2006). A model in population genetics theoretically validates the role of gBGC in maintaining recessive deleterious mutations for long periods of time (Glémin, 2010). A study of the impact of gBGC on functional sequences is detailed in chapter V.

Review articles for this sub-chapter: Marais (2003); Galtier and Duret (2008); Duret and Galtier (2009).

I.4 Conclusion

Since the experiments of Gregor Mendel in the 19th century, our understanding of the molecular processes controlling the transmission of characters to the offspring has vastly progressed. It is now known that in sexually reproducing species, the choice of the genetic material that will be inherited by the descendants is made during meiosis. Each new individual possesses a diploid number of chromosomes, half of which come from the father and half from the mother. When, in turn, this new individual produces offspring of its own, it transmits only half of its chromosomes. The simple solution as to which chromosomes to transmit would be a random choice. However, the process is more complicated, as the homologs inter-exchange genetic material. This exchange is the result of recombination and ensures the correct segregation of homologous chromosomes.

Recombination is initiated in the early phases of meiosis through double-strand breaks (DSB). DSBs are positioned preferentially in hotspots. These hotspots are not distributed randomly along the chromosomes. Certain chromosome regions such as telomeres and ribosomal DNA are inhibitory for DSB formation. While no sequence specificity has yet been associated with DSB hotspots, features such as open chromatin structure, binding sites for transcription factor, intergenic or promoter regions have a certain influence on their positioning.

Multiple pathways have been proposed to explain the repair of DSBs. However, all these pathways result in one or both recombination products, which are crossovers (COs) and non-crossovers (NCOs). Apart from the mechanisms and proteins leading to their production, COs and NCOs differ in the quantity of genetic material exchanged between homologs. COs result in the transfer of long homologous sequences and they are the most easily observed and studied among the two recombination products. Similar to DSBs, COs also cluster in hotspots and are influenced by certain genomic features. In human, part of the CO hotspots are characterized by the presence of a 13-mer degenerated nucleotide motif. Recently, this motif has been found to be the binding site of the PRDM9 zinc finger protein. The structure as well as the evolution of the *Prdm9* gene represent key elements in the study of CO production and distribution.

Recombination plays a major role in the shaping of the nucleotide landscape of genomes. First, by shuffling the pre-existing combination of alleles, it facilitates the action of natural selection. Second, it impacts on the nucleotide composition of sequences, by promoting GC biased gene conversion (gBGC), thus leading to an increase in GC content. gBGC represents one of the main processes leading to the organization of genomes in isochores. Isochores are long regions of relatively homogeneous GC-content, which correlated with multiple genomic features, such as gene density, gene length, repeat element insertion, and replication timing.

Understanding the evolution of recombination is thus essential for the study of genome evolution. In particular, differences in recombination between species, sexes, and individuals, are interesting in view of their differential impact on the genomic structures. We will present in detail in the following chapter some of the methods developed for the study of recombination products and the relation between recombination and GC-content.

Chapter II

Methods for studying recombination

This chapter describes in detail the methods developed for the study of recombination and its role in biased gene conversion. Section II.1 presents the state-of-art in genetic map construction, their use and limitations. Section II.2 offers an insight into the models proposed to describe the distribution of recombination events along the chromosomes. Finally, section II.3 provides some insights into the mathematical tools used for the study of biased gene conversion. In section II.3.3 we present the model we developed for the study of the impact of biased gene conversion on the frequency of deleterious alleles in human population.

II.1 Detecting and measuring recombination

Mendel's principle of independent assortment states that allele pairs separate independently during the formation of gametes. At the beginning of the 20th century, experiments conducted on sweet peas by William Bateson, Edith Rebecca Saunders, and Reginald Punnet seemed to contradict this principle, as described in figure II.1. The two traits they were examining, flower color and the shape of pollen grains, were transmitted together. The observed deviation from the independent assortment was termed "coupling" or linkage. But it was not until the work of Morgan and his students studies on fruit fly, *Drosophila melanogaster*, that the role of genetic linkage was fully understood (Morgan, 1911). By quantifying how much traits on the same chromosome were linked (the frequency of crossover (CO) between linked traits), the order and relative distance between factors could be mapped. This observation led to the first linkage map of six sex-linked factors, in *D. melanogaster*, by one of Morgan's students, Alfred Sturtevant (Sturtevant, 1913).

Proceeding from these early works, the accepted measure of recombination rate was defined as the expected number of recombination events between two loci per generation. In honor of T. Morgan, John Burdon Sanderson Haldane defined the unit of this measure a *morgan* (Haldane, 1919). When recombination leads to the separation of two loci with an expected frequency of 1%, the loci are said to be 1 centimorgan (cM) apart. The presence of a recombination event can be detected by comparing the position of a set of polymorphic markers between loci, in different meioses. While COs act on large genomic regions, resulting in the exchange of markers flanking a recombination hotspot, non-crossovers (NCOs) have a small-scale impact, affecting only markers within the hotspot (Hellenthal

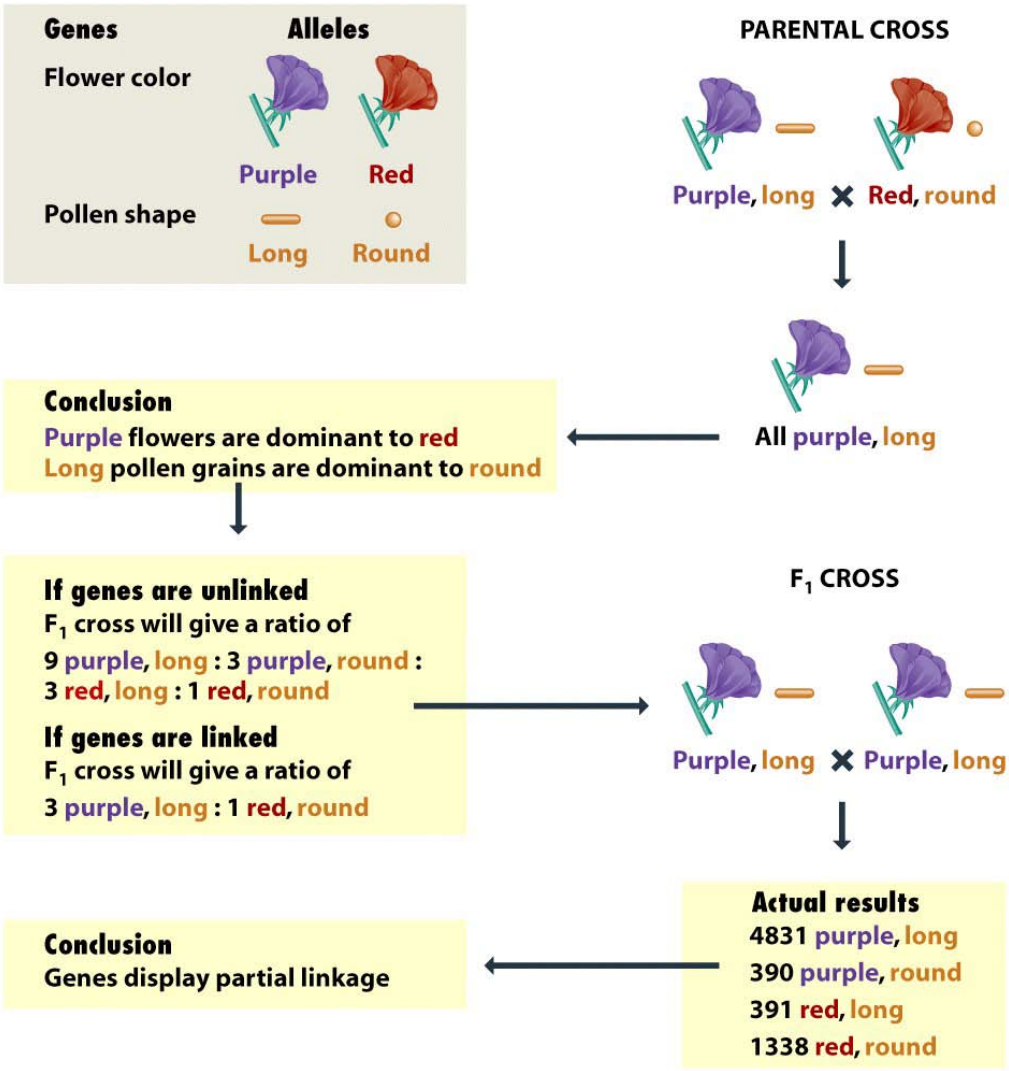

Figure II.1: *The cross shown here was carried out by Bateson, Saunders and Punnett in 1905 with sweet peas. The parental cross gives a typical dihybrid result, with all the F1 plants displaying the same phenotype, indicating that the dominant alleles are purple flowers and long pollen grains. The F1 cross gives unexpected results as the progeny show neither a 9: 3: 3: 1 ratio (expected for genes on different chromosomes) nor a 3: 1 ratio (expected if the genes are completely linked). An unusual ratio is typical of partial linkage. From Brown (2002).*

and Stephens, 2006). As markers are widely spaced along the genomes, it is difficult to detect NCO events and thus, the main focus of recombination rates estimations is on the computation of crossover rates (COR). The estimation of non-crossover rates (NCOR) will be described in Chapter II.1.5.

II.1.1 Genetic markers

A critical stage while studying the linkage patterns, resides in identifying **markers** along chromosomes. A genetic marker is a DNA locus occupying a definite position in the

genome and that is subject to inter-individual variability. Ideally, any genetic marker should be heritable in a simple Mendelian fashion, easily traceable, display high levels of polymorphism (a high number of variants generates a high proportion of heterozygotes), have only a low mutation frequency, be in Hardy-Weinberg equilibrium (selection, mutation, genetic drift, meiotic drive, etc. do not influence the allele and genotype frequencies), the alleles follow a co-dominant mode of inheritance (all genotypes, homozygous and heterozygous, can be ascertained) (Ziegler et al., 2010). The first markers for linkage studies were genes that generated different phenotypes and could thus be visually examined. However, the number, distribution, and degree of polymorphism of genes along the genomes generate low-resolution genetic maps (Brown, 2002). New types of DNA markers have since emerged as more suitable candidates for linkage mapping. They are classified in different categories according to the type of polymorphism they display, in length (*i.e.* microsatellites) or sequence (*i.e.* single nucleotide polymorphism).

Microsatellites are tandem repeats of short motif sequences (1-6 bp). For instance, $(CA)_n$ is a 2 bp motif repeated n times. Shortly after their discovery in the 1980s (Hamada and Kakunaga, 1982), microsatellites have become one of the most used genetic markers, especially due to their high variability in length and wide distribution in eukaryotic genomes (Hamada et al., 1982; International Human Genome Sequencing Consortium, 2001; Ellegren, 2004). Microsatellites can be uniquely amplified along the genome through polymerase chain reaction (PCR), which uses specific primers for the unique sequences flanking the marker. The alleles at a particular microsatellite locus are further identified by assessing their length through agarose gel electrophoresis (Feingold J., 1998).

Recently, due to novel, cost-effective genotyping platforms, **single nucleotide polymorphism** (SNP) has emerged as the favored marker for linkage studies. A SNP is a stable variation of the DNA sequence, involving a single base at a genomic position in multiple individuals, with alleles displaying relatively high frequencies in a population. For example, agtActt and agtTctt could correspond to two sequences at the same locus, in two individuals, the polymorphism being A/T. For such a variation to be considered a SNP, both alleles should reach frequencies $> 1\%$ in a population (HapMap). SNPs occur every 100 to 300 bases along the more than 3 billion bases of the human genome, and, thus, their number is estimated to be 10-30 million (HapMap). As for microsatellites, the DNA sequences are amplified by PCR, but the genotyping results either from the hybridization to an array containing anchored oligonucleotides or direct sequencing.

Several techniques have been developed in order to study the distribution of recombination rates along chromosomes. Since the first linkage map developed by Sturtevant (1913), crosses and pedigree studies have led to the building of **genetic maps** in different organisms ranging from viruses to mammals. Lack of recombination can result in tight associations of alleles at multiple loci leading to their segregation in blocks, which are known as haplotypes and are subject to strong **linkage disequilibrium (LD)**. Recombination hotspots can be inferred by analyzing the LD patterns in population genetic data. Another technique consists in the direct quantification of recombination intensities in sperm cells, through the molecular experiments of **sperm typing**.

II.1.2 Genetic maps

Genetic maps are the oldest and most widespread tool for studying the meiotic behavior of chromosomes. Their construction relies on genotyping and tracking the transmission of polymorphic markers in a large number of individuals in families. This step requires the identification of the parental origin of each allele for a certain marker in the progeny. The transmission of markers during meiosis can be followed by analyzing the genotypes of the resulting gametes. Figure II.2 depicts the genotypes of the four gametes at two biallelic loci depending on the existence and/or position of CO(s). However, the majority of experiments in eukaryotes are not focused on the direct examination of gametes, but instead infer the inheritance of parental gametes in the diploid offspring. In organisms that can be manipulated genetically and that have sufficiently short generation times to be tractable, the most common experimental procedure to genetic map building are crosses (figure II.3). Parental lines are usually chosen from highly divergent inbred populations. Individuals in the first generation (F1) can then be crossed with their parents in order to produce the F2 generation (**backcross**). If the F2 mapping population is obtained by intercrossing F1 individuals, the experiment is termed **F2 intercrosses**. Another approach consists in producing **recombinant inbred lines** (RIL). This technique starts with inbred parental lines, homozygous at every locus and continues by inbreeding each of the resulting heterozygous F1 individuals until recombinant inbred lines are obtained.

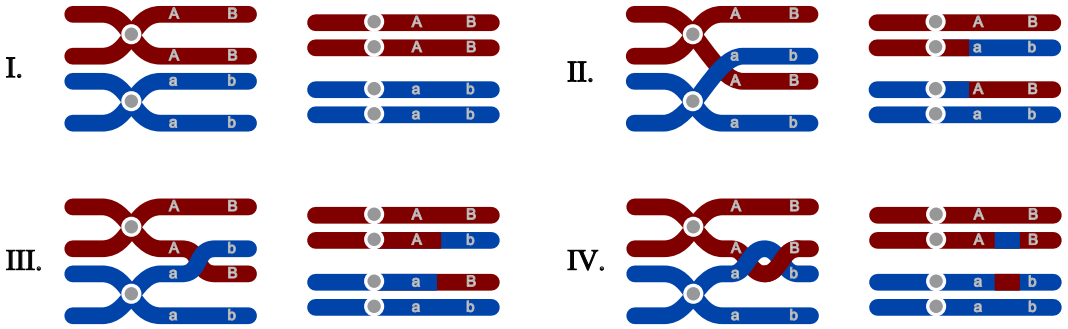

Figure II.2: *Transmission of two markers in the offspring. Two parents, mother (red chromosome) and father (blue chromosome), produce four type of gametes. Both mother and father are homozygous at two consecutive markers, AABB for the mother and aabb for the father. I. There is complete linkage between the two loci, leading to four non-recombinant gametes, identical to each one of the parents. II. A CO event takes place before the A/a locus. In the situation when no other markers are defined prior to the A/a locus, this recombination will not be detected in the offspring. III. A CO event takes place inside the A/a - B/b interval and will be observed in the form of recombinant gametes. IV. Two COs have occurred in the A/a - B/b interval. The two events cancel each other and no recombinant genotype is observed if no additional marker is present in this interval.*

In natural populations for which it is not possible to create inbred lines or in the case of crosses being constrained for ethical reasons or long generation times, **pedigree studies** can be employed for map building. Such is the case in humans, in which recombination frequencies are calculated by examining the genotypes of individuals in a family for several

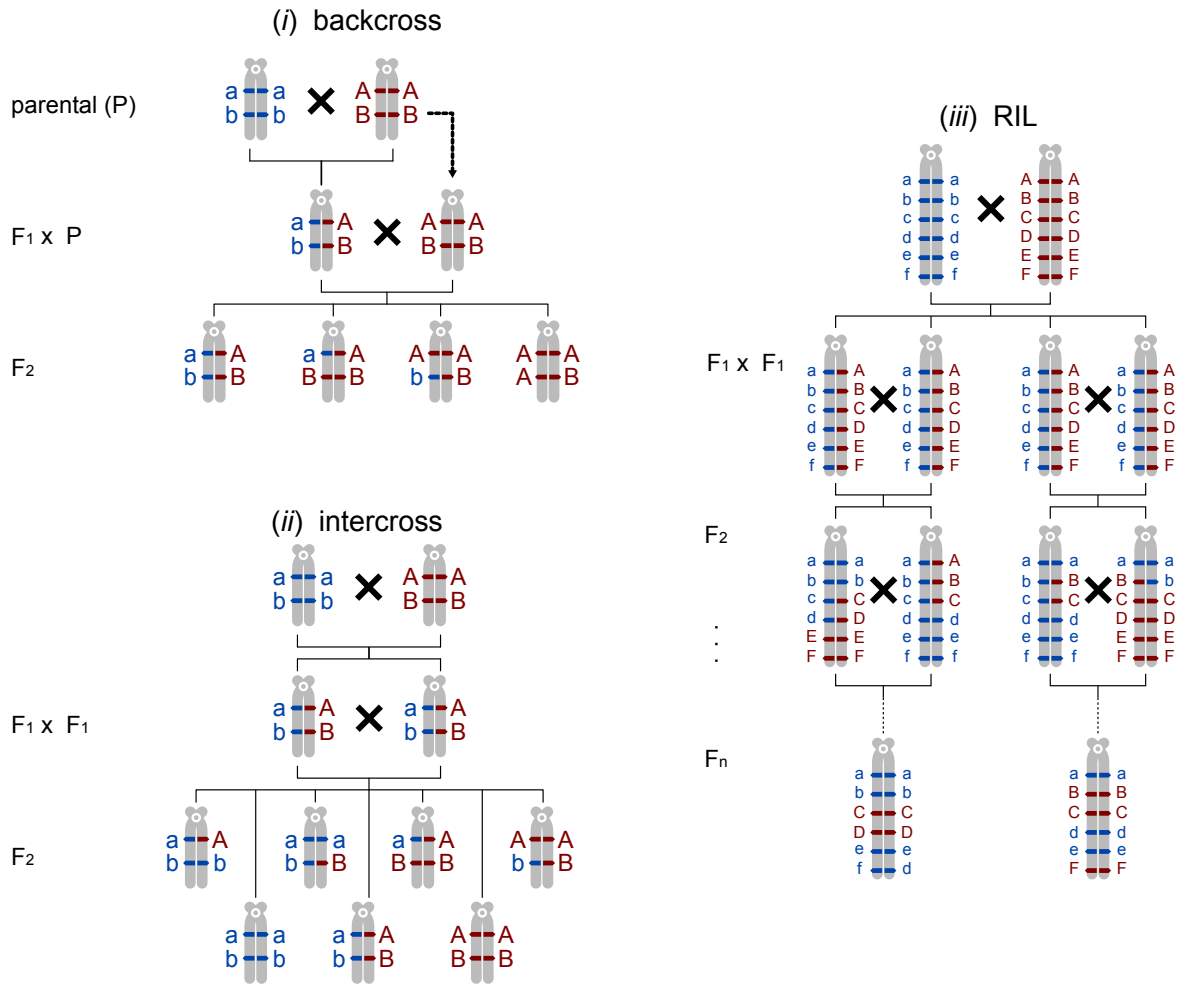

Figure II.3: *Crossing experiments. I. In a backcross experiment, each individual in the F1 generation is crossed with one of its parents, generating the F2 generation. II. An intercross results from the mating of F1 individuals among themselves. III. Recombinant inbred lines result from repeated sibling mating of the individuals in F1 resulting in genomes that are mosaics of the original parental genomes, homozygous at every locus.*

successive generations. An example of a two-generation family is given in figure II.4.

II.1.2.1 Ordering markers

Once genetic markers have been defined along the genome, the following step consists in assessing their physical linkage and order by analyzing their segregation in pedigrees. This can be achieved by analyzing the fraction of recombinants (R) observed in the offspring. Given two markers segregating at two distinct loci, if they are subject to **complete linkage**, the parental combination of alleles will be transmitted as a whole to the offspring. As in figure II.2 I., the alleles of the two markers (A/a and B/b) are always transmitted according to their parental association (AB and ab). Independent segregation can result from the two markers being separated, either because of localization on different chromosomes or because of systematic COs between the two markers (figure

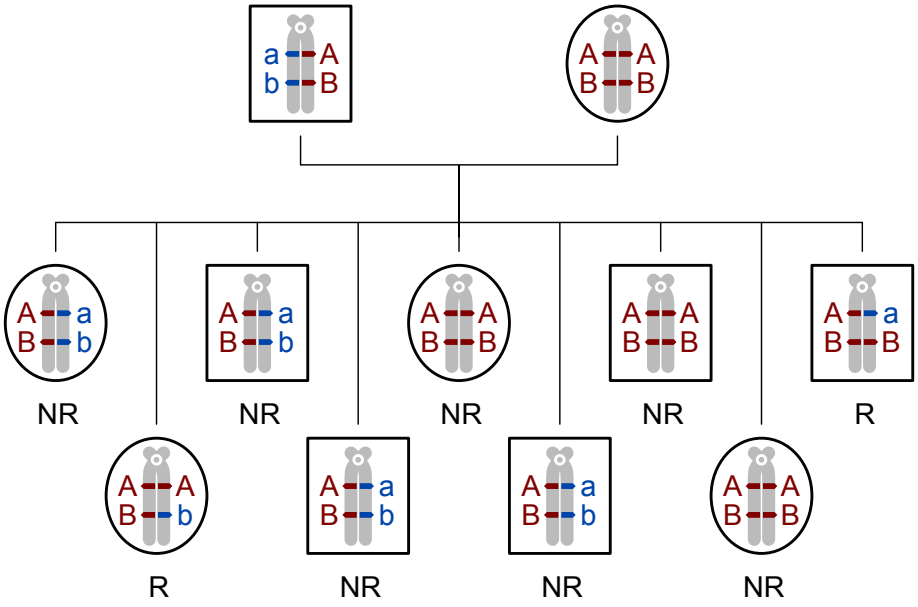

Figure II.4: *The pedigree analysis at two biallelic loci of a family with 9 children. The mother (circle) is homozygous and the father (rectangle) has a known heterozygous genotype. The haplotypes of the 9 children are inferred from the known haplotypes of the parents. For these two loci, 7 out of the 9 children have Non Recombinant (NR) haplotypes, while the other 2 are Recombinants (R). Furthermore, in this example, given the parental haplotypes, the recombination event can be attributed to the paternal line. Adapted from Backström (2009).*

II.2 II.). In this case, the parental/recombinant chromosomes are inherited in a 50/50 ratio. Depending on the frequency with which COs disrupt the association between these markers, deviations are expected to be observed from the 50% ratio. The example in figure II.4 represents a cross between two parents with two distinctive loci, A/a and B/b (Backström, 2009). In this example, the mother is homozygous (AB - AB) at the two loci, while the father is heterozygous (AB - ab). Among the 9 offspring, 2 carry recombinant chromosomes (R). Thus, the frequency of recombinants in this dataset is $R = \frac{2}{9} = 0.22$, implying a genetic distance (g) between the two markers of 22 cM and a deviation from the 50% independent segregation.

However, in real linkage experiments, family sizes are small and it is often unknwow how the alleles are joined in the heterozygous parent (AB - ab or Ab - aB). These limitations complicate the detection of recombinant and non-recombinant chromosomes in the offspring (Backström, 2009). In order to test if two markers are linked, a maximum likelihood log (base 10) score (a.k.a LOD for *logarithm of odds*) is calculated. It represents the ratio between the likelihood of the two alternative hypotheses, linkage and independent segregation, as a function of the observed R (Equation II.1):

$$LOD(R) = log_{10}\left(\frac{L(R)}{L(0.5)}\right) = log_{10}\left(R^{n_R}(1-R)^{n_T-n_R}2^{n_T}\right) \quad \text{(II.1)}$$

where L is the likelihood and is defined as the probability of the observed genotypes

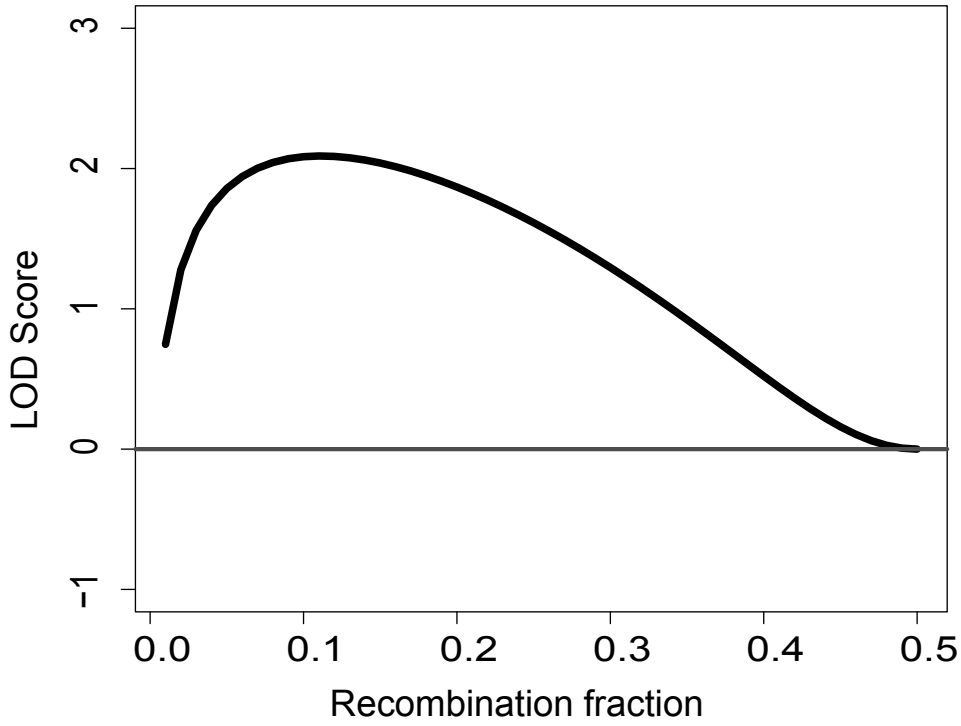

Figure II.5: *The LOD score curve as function of the recombination rate (R) for a total number of 9 children. The maximum of the curve is reached for an R close to 0.1.*

given R, n_R is the number of offspring with a recombinant genome and n_T the total number of offspring. By convention, a LOD score ≥ 3 is indicative of a linkage between the two markers with R recombination frequency. The threshold of 3 represents a 1000 to 1 odds that the linkage observed did not occur by chance. A LOD≤ -2 is evidence, with at least 100 to 1 odds, against linkage. In between these threshold values, the pedigree information is not conclusive to distinguish among the two hypotheses. In agreement with this last case, in the example from figure II.4, the LOD score is 0.64.

The next step in building a genetic map consists in ordering the markers. For example, for 3 markers A, B, and C, there are three possible orders ABC, ACB, BAC (reverse combinations are equivalent). If the recombinant frequencies between pairs of markers are $R_{AB} = 0.15$, $R_{AC} = 0.10$, and $R_{BC} = 0.07$, ACB is the "best" order. The lack of additivity between recombination frequencies will be discussed hereafter. There are two major difficulties raised by this approach: 1) the definition of "best" order is complex; 2) its generalization for a large number of markers is impossible.

In order to determine the "best" order, all possible marker combinations should be compared by calculating their individual likelihood (the distance between all markers given the data) (for example using algorithms such as CRI-MAP (Matise et al., 1995) or MAP-O-MAT (Kong and Matise, 2005)). As the number of possible orders for n markers is $\frac{n!}{2}$, the above procedure becomes tedious, even for computer algorithms. In order to overcome the problem of multiple testing, the first step consists in assigning markers to linkage groups, ideally equal to the haploid chromosomes number. Two markers are assigned to the same linkage group if their LOD score is higher than a threshold value.

The order inside a linkage group is established by starting with a small number of markers and gradually adding new markers one by one. As each new marker is added the maximum likelihood of the orders is computed progressively. This procedure is repeated a number of times by seeding with different groups of starting markers. For species with an assembled genomic sequence, the order of the markers is also given by the physical map.

II.1.2.2 Calculating genetic distances

The recombination fraction quantifies the linkage between markers. This measure is appropriate for small genetic intervals. However, as two loci are further apart, the probability of two CO events very close to each other (double CO events) increases. The existence of such events means that R values of adjacent intervals are not additive. A classical modelisation of the relation between recombination frequencies in adjacent segments is given by equation II.2:

$$R_{x+y} = R_x + R_y - 2CR_xR_y \tag{II.2}$$

where C is the coefficient of coincidence defined as the ratio between the observed and expected number of double COs. The interference strength between COs is 1-C. Defined between 0 and 1, a C of 1 stands for no interference, while $C = 0$ when the two intervals show complete interference (no CO separates them). Figure II.2 IV. illustrates the outcome of a double CO between two consecutive markers. If the two COs occur between the same two markers their effect will cancel each other out and no recombinant chromatids will be observed, leading to an underestimation of the recombination fraction if not accounted for.

Another measure of the linkage between markers is the genetic distance (g), measured in cM, which is the expected number of COs per meiosis. By definition the genetic distance is additive. In the case of small intervals (less than 10 cM) the recombination frequency and the genetic distance are equivalent (Kosambi, 1944). The functions linking R and g integrate the probability of multiple CO events per interval. Thus, the expected proportion of single chromosomes having k COs in an interval is noted p_k. In order to account for these events, the recombination frequency is defined as $R = \sum_{k=odd} p_k$, since even number of COs in the same interval cancel each other (figure II.2.IV). One method considers that p_k follows a Poisson distribution $p_k = \frac{e^{-g}g^k}{k!}$ and that CO events are independent of one another (Haldane, 1919). The resulting relation (equation II.3) is called Haldane's map function:

$$R = \frac{1}{2}\left(1 - e^{-2g}\right) \Leftrightarrow g = -\frac{1}{2}ln(1-2R) \tag{II.3}$$

The map function from equation II.3 considers $C = 1$. However, the availability of large sets of genetic markers has proven that this hypothesis is incorrect, as interference is a widespread phenomenon (Chapter I.2.3). Kosambi (1944) set $C = 2R$, as when $R = 0.5$, $C = 1$ and there is no interference as in Haldane's map function. Integrating interference levels, for $C = 2R$, a new mapping function has been established (Kosambi, 1944) (equation II.4):

$$R = \frac{1}{2}\frac{e^{4g}-1}{e^{4g}+1} \Leftrightarrow g = \frac{1}{4}ln\frac{1+2R}{1-2R} \quad \text{(II.4)}$$

II.1.2.3 Sex-averaged and sex-specific genetic maps

Like physical maps, genetic maps provide information on the order and distance between genetic loci along chromosomes. Given this common trait, the two maps are often compared in order to confirm the position of markers in the genome. However, of the two, only genetic maps quantify the number and distribution of CO events during meiosis. In figure II.6, the two types of map, physical and genetic are shown in parallel. The combined information of these two sources enables the estimation of recombination rates per physical distance, expressed in cM/Mb.

The most common type of genetic maps is sex-averaged, the genetic distance of an interval corresponding to the expected number of COs in both sexes. But maps specific to male or female meiosis can also be built. One approach consists in directly measuring the recombination intensity in the male and female germline. Such experiments, as the sperm-typing technique described in Chapter II.1.4, provide a high resolution description of local recombination events. On the other hand, pedigree linkage studies generate genome-wide sex-specific genetic maps as in figure II.6.

In human, by comparing the haplotype of mother, father and children, the position of parental COs can be inferred. A recent example of such an approach is the study of Hutterite related individuals (Coop et al., 2008). The procedure consists in identifying "informative" markers along the chromosomes. A marker is informative for the mother if it is heterozygous for the mother and homozygous for the father. Markers informative for the father show the reverse pattern. The next step consists in detecting the type of alleles inherited by the children. An example of the identification of CO events between 10 consecutive markers in a family with 4 children is shown in figure II.7. In order to identify a CO event in the paternal line, the type of alleles inherited from the father are compared in pairs of siblings. If the father is AT and the mother AA, and both genotypes of two of their children are AT or AA at this locus, they have both inherited the *same paternal allele*. Instead, if one of the siblings is AT and the other AA, they have *different paternal alleles*. If in a pair of siblings, there is a switch between two paternal consecutive markers in *same paternal alleles* and *different paternal alleles* states, then a recombination event has taken place in the paternal line (Chowdhury et al., 2009). A detailed explanation of CO inference in figure II.7 is given in table II.1. Based on a similar methodology, sex-specific maps can be computed in three-generation pedigrees, by analyzing informative markers that are shared between grandparent and grandchild, thus identifying CO events occurring in the parents (Cheung et al., 2007).

In the case of large pedigrees, accounting for multiple meioses, and with numerous markers distributed along chromosomes, the resulting genetic maps are a valuable tool for identifying recent CO events with a fairly good resolution. Moreover, linkage mapping is the only technique that allows the genome-wide analysis of female and male recombinations separately. Given major heterochiasmy levels in some species (table I.3), sex-specific genetic maps seem the most pertinent tool for studying meiotic recombination (Lynn et al., 2004; Cheung et al., 2007). However, the resolution of a genetic map depends on the number

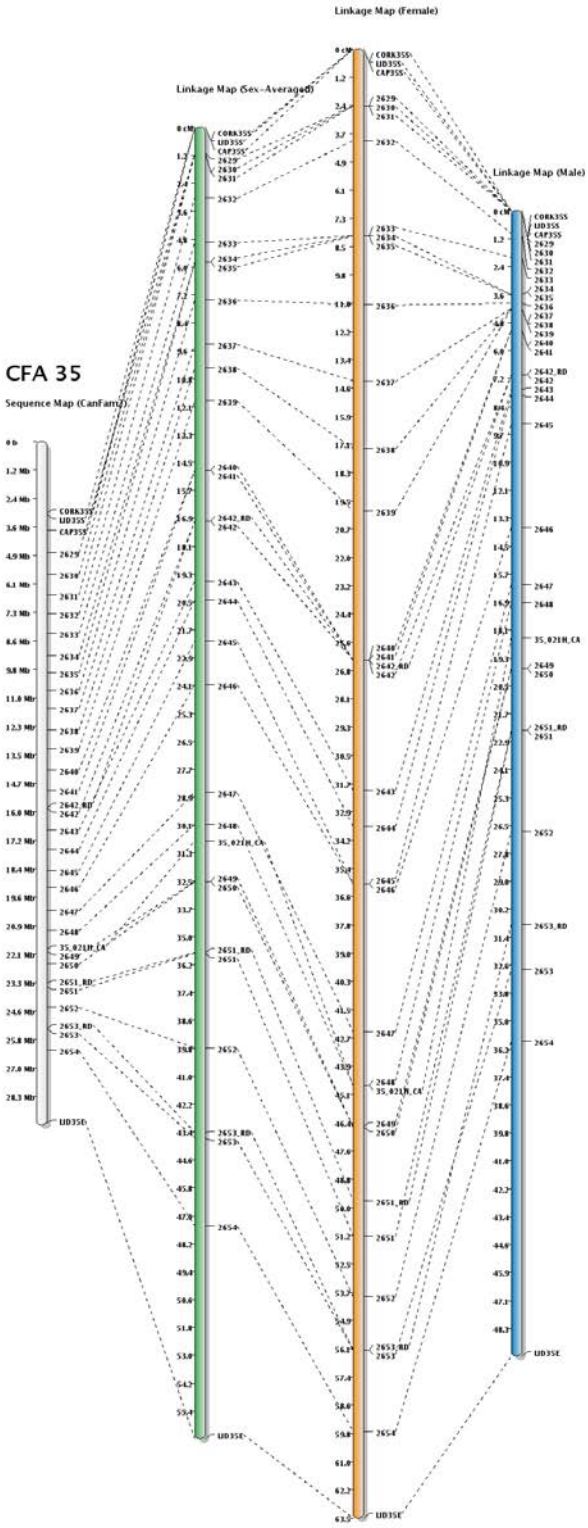

Figure II.6: *The mapping of markers along the physical map and sex-averaged, female, and male genetic maps of the dog chromosome 35. From* `http://www.vgl.ucdavis.edu/dogmap/`.

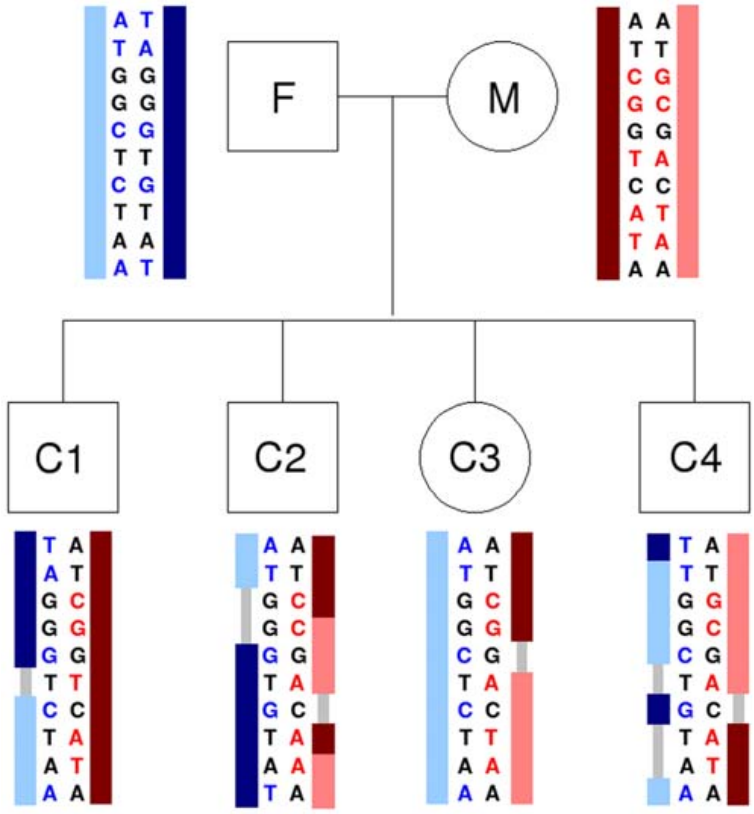

Figure II.7: *Identification of recombination events. Genotypes at 10 consecutive SNPs for two parents and their four children in this pedigree are provided. Informative makers are marked in red and blue for the mother and father, respectively. Recombination events are shown by color switches (for example from dark to light red), or, if regions are large, they are shown in gray. From Chowdhury et al. (2009).*

of COs that can be inferred, and in eukaryotes, the size of the pedigree and the small number of meioses are a limiting factor for linkage mapping, leading to low resolution maps (Arnheim et al., 2003).

II.1.3 Linkage Disequilibrium

II.1.3.1 Quantifying LD and recombination

The resolution problem of genetic maps, raised by a limited number of individuals in a family, can be solved by inferring historical CO rates from samples of population data. Thus, the detection of recombination events is based on the study of disruptions in the associations of alleles at two loci, in unrelated individuals. The deviation from random segregation of alleles at two loci has been termed **linkage disequilibrium** (LD) (Lewontin and Kojima, 1960), and recombination results in the reduction of LD.

There are two statistics widely used for the quantification of LD: D' and r^2. Both $|D'|$ and r^2 take values between 0 and 1, 0 for no LD and 1 for complete LD. Given two bi-allelic

Markers	Siblings, vs ref. C1 C2	C3	C4	CO events
M1	≠	≠	=	in C4 in the paternal line between M1 and M2
M2	≠	≠	≠	
M2	≠	≠	≠	in C2 in the paternal line somewhere btw. M2 and M5
M5	=	≠	≠	
M5	=	≠	≠	in C1 and C4 in the paternal line somewhere btw. M5 and M7
M7	≠	=	≠	
M7	≠	=	≠	in C4 in the paternal line somewhere btw. M7 and M10
M10	≠	=	=	
M3	=	=	≠	in C2 in the maternal line somewhere btw. M3 and M4
M4	≠	=	≠	
M4	≠	=	≠	in C3 in the maternal line somewhere btw. M4 and M6
M6	≠	≠	≠	
M6	≠	≠	≠	in C2 and C4 in the maternal line somewhere btw. M6 and M8
M8	=	≠	=	
M8	=	≠	=	in C2 in the maternal line between M8 and M9
M9	≠	≠	=	

Table II.1: *The inference of recombination events in the paternal and maternal line from the genotypes in figure II.7. The reference child is C1. For each pair of siblings between C1 and the other three children, the informative markers for the father are colored blue and for the mother in red. A symbol of equality (=) stands for the two children in a pair sharing the same parental allele, and ≠ for different alleles. If a switch between the two symbols is observed between two consecutive informative markers, a CO events has taken place. If such a switch has taken place between C1 and only one other child, the CO has occurred in the sibling. If instead, the majority of sibling pairs show switches between two markers, the CO is assigned to C1.*

loci, A/a and B/b respectively, f_{AB} is the frequency of the haplotype with A at the first locus and B at the second, and f_A, f_a, f_B, and f_b are the frequencies of alleles A, a, B, and b respectively. In the case of linkage equilibrium, the frequency of each gametic type is equal to the product of the respective allele frequencies (*i. e.* $f_{AB} = f_A f_B$) (Lewontin and Kojima, 1960). The deviation from equilibrium is noted D and is defined in equation II.5. In the case $D = 0$, there is linkage equilibrium, meaning a random association of alleles. However, this measure is sensitive to allele frequencies, leading to the definition of the D' statistics (equation II.6), where D_{max} is the maximum value of the deviation of the actual gametic frequencies from LD, defined as $D_{max} = min(f_A f_b, f_a f_B)$ if $D > 0$ or $D_{max} = min(f_A f_B, f_a f_b)$ if $D < 0$ (Lewontin, 1964).

$$D = f_{AB} - f_A f_B \tag{II.5}$$

$$D' = \frac{D}{D_{max}} \tag{II.6}$$

The other measure of LD between loci is r^2 (Hill and Robertson, 1968), defined in equation II.7. It represents the correlation of alleles in a pair of loci. The r^2 statistics, by contrast with D', is robust to biased allele frequency and small sample sizes, making it the measure of choice for LD mapping (Backström, 2009).

$$r^2 = \frac{D^2}{f_A f_B f_a f_b} \tag{II.7}$$

LD is expected to decrease from one generation to the other (Slatkin, 2008). Both recombination and recurrent mutation can result in reduced levels of LD between loci. The relation between the decrease in LD and the recombination frequency is formulated in equation II.8, where t is time in generations (reviewed in (Slatkin, 2008)). One method used to identify the impact of recombination on the patterns of LD is the **four-gamete test** (Hudson and Kaplan, 1985). In order to distinguish the effect of recombination and mutation, the test supposes that mutations are not recurrent, meaning that once a mutation has affected a site, no new mutations will occur at this site. This hypothesis, also known as the infinite sites model (Kimura, 1969), is valid in populations such as humans, with large genomes and small mutation rates. The four-gamete test considers that if in a population, all four possible haplotypes are present for two biallelic loci, then no recombination event has taken place and $|D'| = 1$. Otherwise, recombination must have broken down the LD between the two loci and $D' = 0$. Relative $|D'|$ values, different from 0 and 1, are difficult to translate into recombination rates as there is no linear relation between the two variables. Moreover, by affecting only the values 0 and 1 to the statistics $|D'|$, the four-gamete test misses many recombination events (Myers and Griffiths, 2003). Other more sophisticated and powerful methods have been developed in order to infer the minimum recombination events that explain the evolution of an observed set of haplotypes in a population sample (reviewed in (Stumpf and McVean, 2003)). In figure II.8.b the LD pattern in a region of the human MHC allows the identification of two potential CO hotspots (vertical arrows) between two LD blocks.

$$D(t+1) = (1-R)D(t) \tag{II.8}$$

II.1.3.2 HapMap Project

Another very important role for linkage mapping is the detection of alleles and combination of alleles responsible for complex diseases. The analysis of a large amount of SNPs in case-control samples leads to the identification of variants associated with disease-risk (Clark et al., 2010). For this purpose, **the HapMap project** aims at finding the subset of SNPs that describes at best the human genetic variation (International HapMap Consortium, 2005). The initial aim of HapMap was to achieve the resolution of **one SNP every 5 kb**. The first phase (HapMap1) resulted in the genotyping of more than one million SNPs in 269 individuals from 4 different geographic regions: 30 parents-child trios from Yoruba in Ibadan, Nigeria; 30 parents-child trios from Utah USA, with a northern and western european ancestry; 45 unrelated Han Chinese individuals from Beijing; and 44 unrelated Japanese individuals from Tokyo. The SNPs were evenly distributed along non-repetitive portions of autosomes and the X chromosome, with on average 1 SNP every

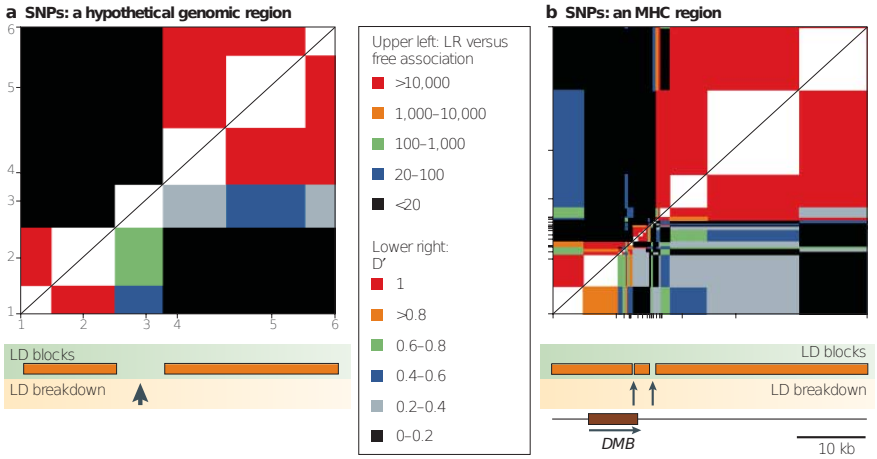

Figure II.8: *Graphical presentation of linkage-disequilibrium data. a. Linkage disequilibrium (LD) in a hypothetical genomic region. In this example, there are 6 SNPs, numbered 1-6 (labeled below and to the left of the plot). LD is calculated for each pair of SNPs from genotypes of a panel of individuals. Each pairwise comparison is plotted as a rectangle centred on each SNP, and extending half way to adjacent markers. The $|D'|$ values are plotted by colour code (see key) below the diagonal. The statistical significance of each $|D'|$ value can be deduced from the LIKELIHOOD RATIO (LR) versus free association, which is shown above the diagonal. In this example, $|D'| = 1$ between markers 1 and 2, indicating complete LD. This value is supported by a high (>10,000) LR. By contrast, the $|D'|$ value between markers 3 and 5 is 0.4-0.6, but this value is accompanied by poor odds (LR versus free association <20). A region of LD breakdown is apparent between markers 2 and 4 (arrow). b. LD in a portion of the human major histocompatibility complex (MHC). Patterns of LD near the DMB gene are shown. Only SNPs with a minor allele frequency ≥ 0.15 have been included. LD blocks are identified visually and are shown as orange boxes. There are two regions of LD breakdown (vertical arrows). From Kauppi et al. (2004).*

5 kb (International HapMap Consortium, 2005). HapMap1 offered a detailed description of the LD patterns along human genomes, organized in LD blocks 7 to 16 kb long, interrupted by hotspots of recombination hotspots (Kauppi et al., 2007).

Phase II of HapMap Project (HapMap2) resulted in the final genotyping of 3.1 million SNPs, distributed on average **1 SNP per kb** (International HapMap Consortium, 2007). Additional to its role in the identification of disease-associated markers, HapMap2 data has provided a high-resolution LD map that led to the identification of more than 25 000 hotspots of recombination along the human genome (Myers et al., 2006). Wavelet analysis of the genomic features linked to the presence of these hotspots, latter led to the identification of the 13-mer degenerate motif responsible for recruiting more than 40% of the previously identified recombination hotspots (Myers et al., 2008).

Recently, the Phase III HapMap (HapMap3) data has been published (International HapMap 3 Consortium et al., 2010). While HapMap1 was focused on the genotyping of SNPs with minor allele frequencies $> 5\%$ in order to avoid genotyping errors, **HapMap3 contains both common and rare alleles**. In order to identify these rare alleles, 1.6

million SNPs have been genotyped in a total of 1 184 individuals from 11 populations world-wide. This new type of data will further improve our understanding of the evolution of recombination rates at local level.

II.1.3.3 Potential biases in estimating recombination with LD

During the establishment of inbred laboratory strains, only a small number of meioses subject to CO events are generated, leading to the detection of relatively few COs (Kauppi et al., 2007). The study of LD patterns in such inbred lines in mouse and rat, has generated longer LD blocks than those found in humans (Guryev et al., 2006). However, linkage data between inbred strains (Kauppi et al., 2007) or in wild population mice (Laurie et al., 2007), have revealed linkage decays comparable with those observed in humans, and have resulted in the identification of additional recombination hotspots in this species (Kauppi et al., 2007). Other species, for which LD patterns have been analyzed include: cow (Gautier et al., 2007), sheep (McRae et al., 2002), pig (Nsengimana et al., 2004), dog (Sutter et al., 2004; Lindblad-Toh et al., 2005), zebra finch (Stapley et al., 2010)... However, none of these species has yet reached the resolution of linkage studies achieved in humans.

While, LD and genetic maps are very well correlated at 1 Mb level, LD analysis offers a far increased resolution, at the kb level, for the study of recombination rates (Myers et al., 2005). Furthermore, sperm-typing studies together with LD data have led to the characterization of recombination hotspots. However, CO rate (COR) estimated from LD data can not distinguish heterochiasmatic behavior as they are sex-averaged (Myers et al., 2005). While genetic maps describe recent recombination events taking place in the individuals of a family, LD patterns reflect long evolutionary history (Arnheim et al., 2003). At this time-scale, LD patterns are influenced by multiple factors, such as the structure and demographic history of a population, as well as natural selection and genetic drift (reviewed in (Slatkin, 2008)). Levels of LD are expected to increase both with the level of subdivision in a population and with the existence of bottlenecks. Mixture of individuals from sub-populations with different allele frequencies create LD. An extreme example is when one sub-population is fixed for alleles A and B and another for a and b. Any mixture between these two sub-populations will generate only the AB and ab haplotypes, which are interpreted as perfect linkage. Moreover, bottlenecks result in a LD increase because of a rapid loss of haplotyes. An example of the effect of bottleneck on linkage patterns is illustrated by two species: domestic dog (*Canis familiaris*) and Norway spruce (*Picea abies*) (Backström, 2009). While the dog has been through at least two bottlenecks in the past 10 000 years and presents extensive LD (Sutter et al., 2004), the LD is decaying rapidly between close loci in Norway spruce in agreement with a large population size conserved for a long period of time (Heuertz et al., 2006). In cases of very strong selective pressures, beneficial alleles can become fixed in a population, along with physically linked loci and thus account for LD patterns (McVean, 2007). To some extent, genetic drift can also generate LD between loci as it can lead to the loss or fixation of some haplotypes by chance alone (Slatkin, 2008).

II.1.4 Sperm-typing

Both genetic and LD maps contain information on the distribution of CO events along the genomes. However, the **sperm typing** technique monitors the transmission of recombinant sequences in the sperm of one individual (Li et al., 1988). Thus, it results in the characterization of CO hotspots at a very local level, both in position and intensity. Figure II.9 illustrates in detail the two sperm typing techniques. The single sperm typing method consists in genotyping each allele in each sperm cell, and thus, finds the position of COs between markers, by comparing the allelic state of each marker to the alleles in the diploid male sperm donor (reviewed in (Arnheim et al., 2003)). The sperm pool typing method identifies only recombinant sequences by using specific primers that amplify only one CO type from a pool of sequences (Jeffreys et al., 1998). In order to design the primers, the sperm-typing technology needs first to identify in advance the position of putative COs. Such positions are inferred from LD maps. Coupled with LD studies and using SNPs as genetic markers, the sperm pool typing method has led to the detailed characterization of 46 recombination hotspots (additional table B). LD patterns, as in figure II.8, can guide the identification of potential recombination hotspots, that can be thoroughly analyzed through sperm typing.

Sperm typing offers the best resolution for identifying recombination as it consists in the study of numerous meioses in individuals. However, the present lack of an automatic process linked to this method, makes such studies technically difficult and prevents their adaptation to the genome scale (Coop et al., 2008). Another drawback is that it can only be applied in males.

II.1.5 Gene conversion rates

NCOs have usually been ignored when studying the distribution of recombination events given that the resolution needed for their detection was lacking. The sperm-typing technique is one of the few genetic methods with sufficient resolution to identify these events. The sperm-typing analysis of some human recombination hotspots has revealed that NCOs occur 4 to 15 times as frequently as COs, with conversion tracts ranging from 50 to 2 000 bp (Jeffreys and May, 2004). In addition to their evolutionary role, gene conversion events also affect the estimation of CO rates from linkage (Mancera et al., 2008) and LD studies (Pritchard and Przeworski, 2001). Recently, thanks to the large quantity of genetic markers and the new genotyping methods developed, more attention has been given to the characterization of NCOs.

NCOs affect the LD pattern by causing a decrease in LD at small scales. In humans, over short genomic intervals there is less LD than expected based on the patterns of COs alone (Pritchard and Przeworski, 2001). However, NCOs are highly localized affecting only markers within a recombination hotspot, and thus, are more difficult to detect individually (Hellenthal and Stephens, 2006). The rates of recombination inferred from population data represent the combined effects of COs and NCOs (International HapMap Consortium, 2005; Myers et al., 2005). Although statistical models have been developed that perform the joint estimation of CO and NCO rates from population data (Gay et al., 2007; Yin et al., 2009), such estimations are still difficult to implement at a genome-wide level.

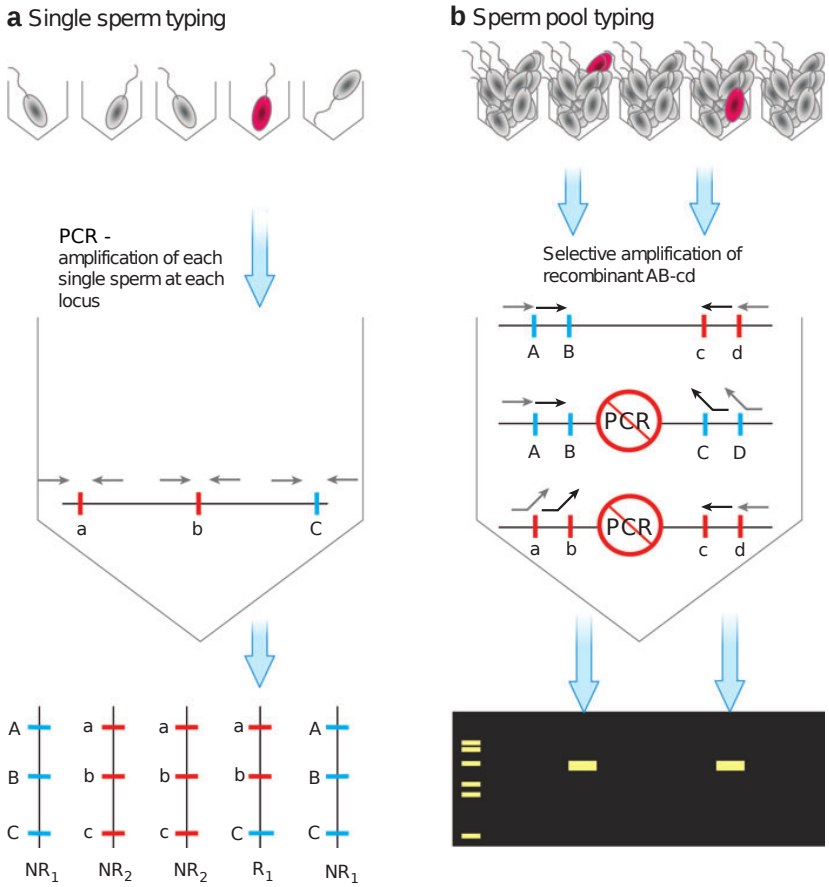

Figure II.9: *Two different approaches for sperm typing. (a) Individual sperm cells are isolated in different wells. All sperm, including one that is recombinant (red), are lysed and each one is amplified at each locus through techniques such as PCR. Primers flanking the SNP region of the single target molecule produce amplicons that can be genotyped to determine the allelic state of that molecule. If the diploid male sperm donor is ABC on one chromosome and abc on the other (ABC/abc), then typed sperm could be either ABC or abc (nonrecombinants, NR) or Abc, aBC, ABc, abC (recombinants, R). In the example, a single-sperm (haploid) molecule typed as abC (R1) must have been produced by a crossover between the B/b and C/c loci. The reciprocal crossover product can also be identified (ABc, not shown), as will crossovers in the A/a-B/b interval (Abc and aBC, also not shown). (b) Typing of total sperm DNA pools is achieved by having a selective PCR that only amplifies one of the crossover types. Usually required are two informative SNPs flanking each side of the interval and sperm from a donor that is heterozygous at all four SNP sites (AB-CD/ab-cd). Crossovers within the interval will recombine these two groups of flanking polymorphisms, thereby distinguishing the two recombinants genetically, not only from one another (AB-cd and ab-CD), but also from both nonrecombinants (AB-CD and ab-cd). For example, to amplify the AB-cd recombinant, the first round of PCR employs a primer that perfectly matches A (and is mismatched with SNPa and a second primer that matches d (gray arrow). The second round (applied to a small amount of first round product) uses primers (black arrows) that match B and c. Using these allele-specific primers, only product from wells containing the recombinant (AB-cd) will render a visible PCR product on a gel. From Arnheim et al. (2007).*

The most conclusive studies of NCO events so far results from sperm-typing studies. Human recombination hotspots DNA3 (Jeffreys and May, 2004) and NID1 (Jeffreys and Neumann, 2005) CO hotspots have been found to host NCO events as well. The presence and frequency of such events differ between hotspots (NCOs are only a quarter of COs in NID1 and in DNA3, gene conversion occurs 4 to 15 times more often than COs). In the majority of cases, NCO tracts were shorter and the conversion activity was centered on the hotspot center (Jeffreys and May, 2004).

As previously mentioned, genetic maps lack the resolution to detect all CO events and even more NCOs. However, in yeast, a high-resolution map was obtained by genotyping $\sim$ 52 000 markers in the four viable spores of 51 meioses (Mancera et al., 2008). Due to this high density of markers, a majority of recombination hotspots contain multiple markers within. A recombination event was detected when there was a switch in the genotype of two nearby loci. If the switch was observed in different spores, a CO was inferred, if it took place in the same spore, a NCO must have occurred. This study has resulted in the estimation of an average of 90.5 COs and 46.2 NCOs per meiosis, the number of NCOs being probably underestimated.

Review articles: (Feingold J., 1998; Stumpf and McVean, 2003; Arnheim et al., 2007; Clark et al., 2010), dissertations: (Auton, 2007; Backström, 2009; Wahlberg, 2009), and books: (Samollow, 2010; Ziegler et al., 2010) for this sub-chapter.

II.2 Modeling the distribution of recombination events

The comparison of physical and genetic lengths between markers has revealed that their respective lengths are not linearly related. This discrepancy is generated by a non-uniform distribution of recombination events along chromosomes. As detailed in chapter I.2.1, different regions of the chromosomes (telomeric, interstitial or centromeric) present different affinities for the recombination machinery. Recombination takes place mainly in restricted regions, called hotspots. Furthermore, COs undergo interference, with a CO discouraging the presence of other COs in its vicinity. We describe in this chapter some of the models developed to answer the questions on the number and distribution of COs.

II.2.1 Counting model

Map functions such as Haldane's and Kosambi's, in equations II.3 and II.4, respectively, describe the relationship between the frequency of recombinants and the genetic length of chromosomes. However, following the advancements in our understanding of the molecular process of meiosis, mathematical models have emerged that address the double-strand break (DSB) process directly.

Counting models consider that DSBs are distributed randomly along chromosomes and their number follows a Poisson distribution of parameter y (the mean number of events). The impact of interference is measured by considering that any two consecutive COs should be separated on average by m NCOs (Foss et al., 1993) (figure II.10). In this class of models, interference depends on the genetic rather than physical length. While the

strength of interference is indeed controlled by the density of CO events (genetic map), it can take a wide range of values at the physical scale, in different organisms (Foss et al., 1993; Berchowitz and Copenhaver, 2010). In the model from Foss et al. (1993) the genetic sizes of all intervals are constant.

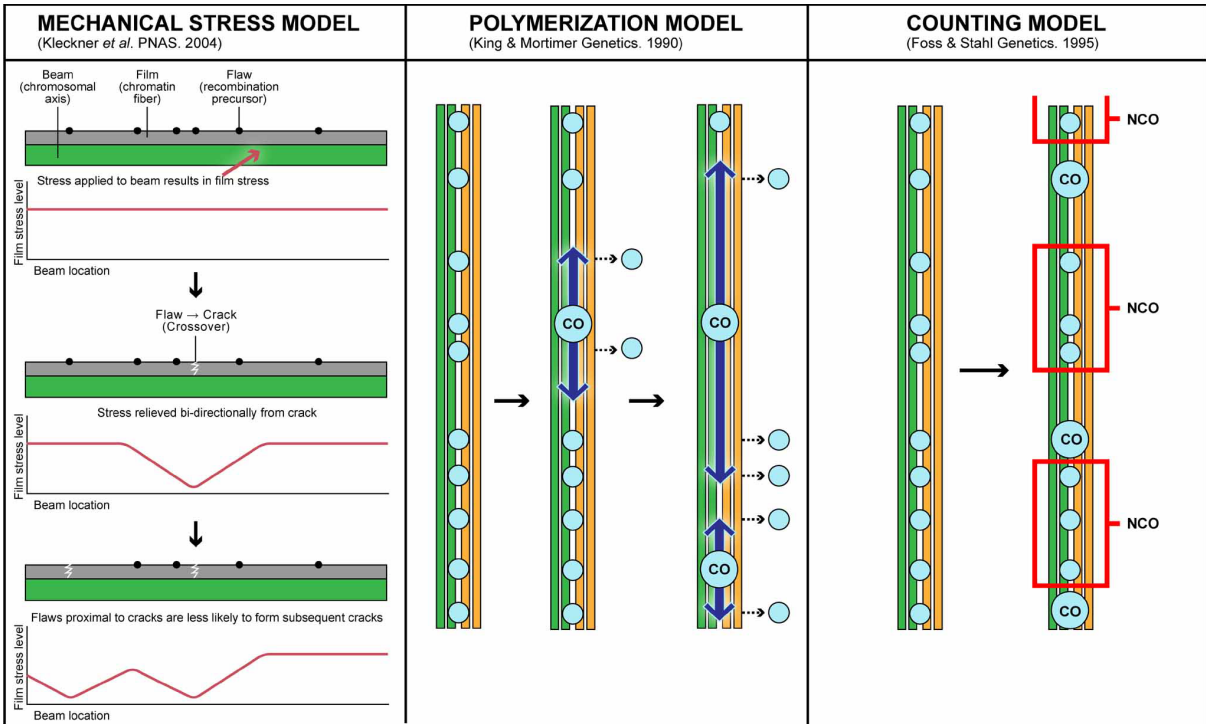

Figure II.10: *Interference Models. The left panel depicts the beam-film demonstration of the* ***mechanical stress model*** *proposed by Kleckner et al. (2004). The beam (chromosomal axis; green), film (chromatin fiber; gray), flaws (CO precursors; black dots). Diagrams depicting the stress level are shown under each beam in which the x axis represents beam position and stress level on the y. The center panel depicts the* ***polymerization model*** *proposed by King and Mortimer (1990). Chromatids are shown in green (parent 1) and yellow (parent 2). Small light blue circles represent recombination precursors and CO designates are shown as larger circles marked with 'CO'. The interference polymer is shown as a large arrow emanating from CO sites, and CO precursors removed by the polymer are shown to the right accompanied with a dashed arrow. The right panel depicts the* ***counting model*** *proposed by Foss et al. (1993). Chromatids are shown in green (parent 1) and yellow (parent 2). Small light blue circles represent recombination precursors and CO designates are shown as larger circles marked with 'CO'. In this diagram, $m = 3$ and intervening NCOs between COs are outlined in a red box. From Berchowitz and Copenhaver (2010).*

The counting model can also generate map functions. The general definition of the recombination frequency is the probability that half the number of chromatids contain at least 1 CO, $R = \frac{1}{2}P(\#_{CO} \geq 1)$. The number k, of DSB events in an interval, has the Poisson probability $P(\#_{DSB} = k, y) = \frac{y^k e^{-y}}{k!}$ and $P(\#_{DSB} > m) = 1 - \sum_{k=0}^{k=m} P(\#_{DSB} = k, y)$.

The probability of 1 CO given the number of DSBs and m is:

$$P(\#_{CO}=1,\#_{DSB}=k,m)=\begin{cases}\frac{k}{m+1}, & \text{if } k\leq m\\ 1, & \text{otherwise}\end{cases} \tag{II.9}$$

Thus, the formula for recombination frequency given the parameter m is defined as:

$$\begin{aligned}R&=\frac{1}{2}\sum_{k=0}^{m}P(\#_{DSB}=k,y)+P(\#_{DSB}>m)\\&=1-e^{-y}\sum_{k=0}^{m}\frac{y^k}{k!}\left(1-\frac{k}{m+1}\right)\end{aligned} \tag{II.10}$$

In an interval of map length g Morgans, the mean number of DSB events, y per tetrad is $y=2(m+1)g$ (Foss and Stahl, 1995). By replacing the expression in equation II.10 a relation is established between R and g, leading to a new map function.

In addition to providing a mapping function, the counting model can also quantify interference along chromosomes. It is also called the chi-square model, as the inter CO distance follows a χ^2 distribution with $2(m+1)$ degrees of freedom (Broman and Weber, 2000). The χ^2 distribution is a member of the Γ distribution family. The Γ distribution has two parameters: shape (ν) and rate (λ). In the case of the counting model, the shape parameter is the number of Poisson events (DSBs) needed to ensure a CO, $m+1$. The rate parameter is twice this same number, to account for tetrads $2(m+1)$. The density and cumulative distribution functions of inter CO distances are:

$$\begin{aligned}f(x|\lambda=2(m+1),\nu=(m+1))&=\frac{\lambda^\nu}{\Gamma(\nu)}x^{\nu-1}e^{-\lambda x}\\F(x|\lambda=2(m+1),\nu=(m+1))&=\sum_{k=m+1}^{\infty}e^{-\lambda x}\frac{(\lambda x)^k}{k!}\end{aligned} \tag{II.11}$$

When applied to data from *Drosophila* and *Neurospora*, with parameter m taking values 4 and 2 respectively, the model adjusts well to the data (Foss et al., 1993). However, in budding yeast many intervals separating successive COs were extremly short leading to incorrect estimations of m (Foss and Stahl, 1995). In many organisms, two types of COs have been identified: interfering and non-interfering (table I.1). The counting model was thus adapted to account for the two types of COs. This has resulted in the **two-pathway model** (Housworth and Stahl, 2003).

The two-pathway model considers that a fraction p, of COs are not subject to interference ($m=0$). The inter CO distances for the interfering type is given by the Γ distribution as in equation II.11, with parameter $\lambda=2(1-p)(m+1)$ and $\nu=m+1$. Given the series of inter CO distances g_0, g_1, ..., g_n along a chromosome (where g_0 and g_n are the distances between the start of the chromosome to the first CO, and from the last CO to the end of the chromosome, respectively), the algorithm considers all 2^n possibilities to assign the n COs into the two types. The distributions of g_0 and g_n are calculated separately under the assumption of stationarity (the start and end of the chromosome do not influence the positions of the first and respectively last COs). The inter CO distances are further

divided in two sets: y_0, y_1, ..., y_j for non-interfering, and z_0, z_1, ..., z_k for interfering COs. The relation between g_i, y_i, and z_i is $\sum_{i=0}^{n} g_i = \sum_{i=0}^{j} y_i = \sum_{i=0}^{k} z_i = G$, where G is the total genetic length of the chromosome. The probability of the inter CO distances for the two types of COs is calculated separately and their sum over all 2^n possible divisions gives the probability of the observed g_i sequence under the two-pathway model:

$$P(g_0, g_1, ..., g_n|p, m) = \sum_{(y_0, y_1, ..., y_j)(z_0, z_1, ..., z_k)} P(y_0, y_1, ..., y_j|p, 0)P(z_0, z_1, ..., z_k|1-p, m) \quad \text{(II.12)}$$

The product of the above probabilities for a collection of meiotic products generates the likelihood of the model parameters, p and m, given the data. By maximizing the likelihood function, parameters have been estimated in *S. cerevisiae* (Stahl et al., 2004), *A. thaliana* (Copenhaver et al., 2002; Lam et al., 2005), maize (Falque et al., 2009), and humans (Housworth and Stahl, 2003; Fledel-Alon et al., 2009).

Another improvement to the counting model consists in considering that m, the number of NCOs between successive COs, is not constant (Lange et al., 1997). For the **Poisson-skip model**, the random number of NCOs is chosen according to a Poisson distribution, s_n. The number of skipped events is random at each run. The χ^2 model is a special case of the Poisson-skip model, with $s_n = 1$. The inter CO distribution for the Poisson-skip model has a cumulative distribution of: $\sum_{m=0}^{\infty} s_m \sum_{k=m+1}^{\infty} e^{-\lambda x} \frac{(\lambda x)^k}{k!}$, which for $s_m = 1$ gives the relation in equation II.11.

Given the small number of parameters to be estimated, counting models have been widely used to assess the strength of interference at the chromosome level. A recent model based on the χ^2 model of (Foss et al., 1993) was proposed that integrates the condition of one obligated CO per chromosome (Falque et al., 2007). The **forced initial CO (FIC)** model starts by choosing the position of the first obligatory CO from an uniform distribution. Additional COs are generated towards each end of the chromosome according to the counting model. Data from mouse have been fitted by the FIC model, and proved to yield better estimates for the number of COs per chromosome than the standard counting model.

Counting models have the advantage of relying on easy to implement mathematical functions with few parameters. However, from a biological perspective, the model predicts an overall reduction in the number of COs and NCOs following a reduction in the number of DSBs (Berchowitz and Copenhaver, 2010). This is not the case for real data, as CO homeostasis ensures that CO rates are kept at high levels despite a decrease of DSB frequencies (Martini et al., 2006).

II.2.2 Mechanical Stress Model

Another model for the distribution of COs is the **mechanical stress** model (Kleckner et al., 2004). The chromosome is represented in this model as a beam covered by a film (the chromatin fiber) (figure II.10). A uniform basal tensile stress σ_0 is present all along

the film. This stress will increase monotonically, as meiotic chromosomes compress and expand, until a crack will appear in the film. The crack corresponds to a CO. As the chromosome is under stress, the obligatory CO is always ensured. For a film of length G, a parameter of the model is N, the number of DSBs (also termed "flaws" in the model). Once a CO has taken place, it will result in the relief of the stress on either side of the crack, over a distance l. A CO at position g will result in a new distribution of stress $\sigma(g)$. The next step consists in placing a new CO at a locus having reached a critical stress value, usually outside the stress relief domain of previous COs. At each step the stress function is recalculated until the overall stress value σ_0 reaches a limit $(\sigma_0)_{max}$. Special conditions are applied for the stress distribution at the ends of the chromosomes.

While the break induced by DSBs can account for a reduction in the stress along the chromosomes, the role played by COs in the stress-relief process is more ambiguous (Berchowitz and Copenhaver, 2010). Moreover, the mechanical view of CO distribution can not explain the existence of non-interfering COs. Although based on a physical phenomenon, the mechanical forces generating the COs are not easy to quantify. Simulations of the model in *Drosophila* and *Chorthippus brunneus* have led to the estimation of the number and inter CO distances (Kleckner et al., 2004). Recently, Falque et al. (2009) developed a method to measure the goodness of fit of the mechanical stress model on the data, and thus find the best estimators for the parameters. This method calculates a projected likelihood score (PLS), which is defined as:

$$PLS = \begin{cases} P(k) \text{ if } k \in [0,1] \\ P(k) \sum_{i=1}^{k-1} \rho_k(g_i) \text{ if } k > 1 \end{cases} \tag{II.13}$$

where k is the number of COs predicted by bivalent, $P(k)$ the probability of k COs/bivalent, ρ_k the probability density of inter-CO distances for a bivalent with k COs, and g_i the genetic distance between CO i and $i+1$. Both the counting and the mechanical stress models have been implemented in a software package, named CODA (CrossOver Distribution Analyzer), that estimates the parameters of the two models (Gauthier et al., 2011).

II.2.3 Polymerization model

The **polymerization** model has been proposed as a mechanistic explanation of interference propagation along the chromosomes, through the synaptonemal complex (King and Mortimer, 1990). The model starts by placing randomly early nodules (EN), thought to correspond to DSB sites, along a chromosome arm. The number of EN has a Poisson distribution of mean 2 times the observed number of COs. As in figure II.10, at each time step, each EN site has a probability of becoming a CO and initiating polymerization. The bidirectional polymerization at one such CO sites has two consequences: it prevents the binding of other EN and it ejects the previously bound ENs from the SC. The polymerization initiated by a CO has a certain probability of being terminated, or ends once it has reached the polymerization from another CO, a telomere or centromere. Simulations of the polymerization algorithm have been performed in budding yeast and *Drosophila* (King

and Mortimer, 1990).

This model has the advantage of offering a biological representation of the interference model along the SC in μm, at which level the interference distances are more similar between species, than at the Mb scale (Berchowitz and Copenhaver, 2010). However, no such polymer has been observed during meiosis and the role of SC in promoting interference has been refuted by experimental data (Fung et al., 2004; de Boer et al., 2006; Shinohara et al., 2008).

II.2.4 Recombination and karyotype

The previous models are all focused on the local distribution of CO events according to the genetic length. Another type of model, that is especially treated in this thesis, tries to model the wide variability in COR between species but also among sexes and chromosomes. At the species level, COR are greatly influenced by the number and lengths of chromosomes (Chapter I.2.4.1). At a local level, COR can be defined as the ratio between genetic (G measured in cM) and physical (P measured in Mb) lengths. However, at the scale of the chromosome, this relation is inappropriate as a strong constraint of at least one CO per chromosome (de Villena and Sapienza, 2001) ensures that even very small chromosomes have a basal G of 50 cM.

A two-parameter model has been developed to describe the relation between G and P at the level of chromosomes (Li and Freudenberg, 2009):

$$G = G_0 + kP \tag{II.14}$$

where G_0 is the minimum genetic length and k is the COR. The parameters of the model are estimated by applying a linear regression on the data. The authors compare the goodness-of-fit of the linear model with (equation II.14) and without an intercept (*e.g.* G0=0), through the Akaike and Bayesian information criterions, AIC and BIC respectively:

$$\begin{aligned} AIC &= 2p - log(L) = 2p + nlog\left(\frac{RSS}{n}\right) \\ BIC &= log(n)p - 2log(L) = log(n)p + nlog\left(\frac{RSS}{n}\right) \end{aligned} \tag{II.15}$$

where p is the number of parameters in the model, L the maximum likelihood estimated from the data and n, the number of samples used to calculate the likelihood. These criteria compare the accuracy of a given model at the same time, penalizing models with many parameters. L is calculated with respect to the variance of the model errors, RSS. RSS is defined as $\sum_{i=1}^{n} \hat{\epsilon}_i^2$, with $\hat{\epsilon}_i$ representing the residual between the model and data point i. $\hat{\epsilon}_i$ equals $(G_i - G_0 - kP_i)$ or $(G_i - kP_i)$ for the models with and without interference. Depending on the sign of the criterion difference between the two models, the better model can be defined. For example, $AIC_{2-parameter} - AIC_{1-parameter} < 0$ is indicative of a better two-parameter model. The fitting of the two parameter model in seven species, human, mouse, rat, chicken, honey-bee, yeast, and worm, yields better estimates in the majority

of cases emphasizing the important role played by the obligatory CO in defining the COR (Li and Freudenberg, 2009).

Figure II.11 provides a representation of a the two-parameter regression on the genetic and physical lengths of chromosomes and chromosomes arms for female, sex-averaged and male human data. These regressions indicate that in humans, heterochiasmy is mainly due to higher COR in female than in males, while the intercept is very similar among sexes. While in the case of human and yeast G_0 values are close to 50 cM, in the other species these values are much smaller than 50 cM and even negative (in the case of honey-bee and worm) despite a valid condition of an obligate CO. Thus, the G_0 parameter is difficult to interpret from a biological point of view.

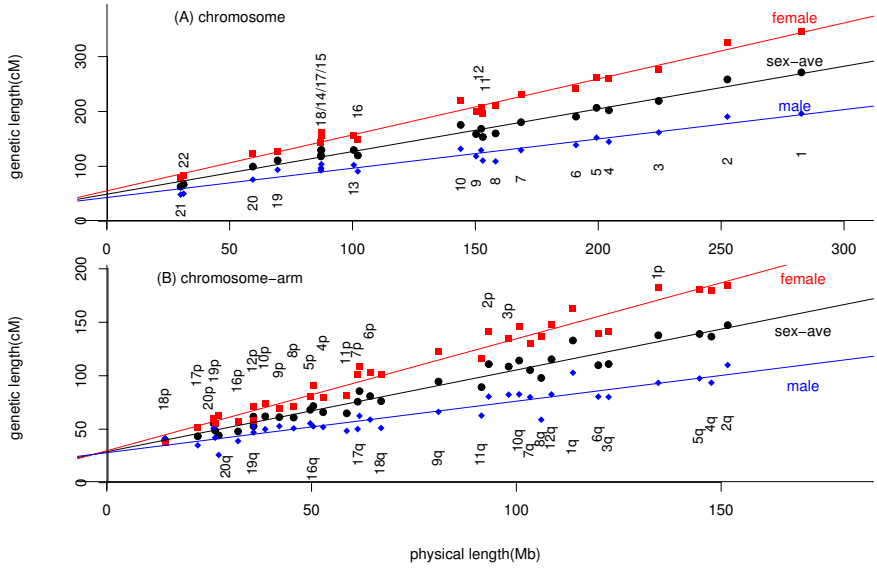

Figure II.11: *Two-parameter regression of human genetic length over physical length. (A) Analysis at the chromosome scale. Female (red), male (blue), and sex-averaged (black) genetic length of each chromosome (in cM) is plotted against its physical length (in Mb). The least-square regression lines are:* $G = 54.2 + 1.02P$ *(female),* $G = 42.0 + 0.52P$ *(male),* $G = 48.1 + 0.78P$ *(sex-average). (B) Analysis of metacentric chromosome at the chromosome-arm scale. The best fit regression lines are:* $G = 29.0 + 1.05P$ *(female),* $G = 27.1 + 0.48P$ *(male),* $G = 28.0 + 0.77P$ *(sex-average). From Li and Freudenberg (2009).*

There is still an unresolved debate if the condition of one obligatory CO applies to the whole chromosome or to each chromosome arm independently. In figure II.11, the data from the two chromosomal lengths are compared. The two-parameter model has been further applied to distinguish between these two hypotheses of one obligatory CO per chromosome ($G_0 = 50$ cM) or per arm ($G_0 = 100$ cM) (Li et al., 2010). The AIC criterion (equation II.15) finds that the model of 1 CO per chromosome is consistently better in human.

Review articles for this sub-chapter: (Drouaud et al., 2007; Mézard et al., 2007; Berchowitz and Copenhaver, 2010).

II.3 The impact of biased gene conversion on the nucleotide composition

II.3.1 Equilibrium GC-content

When studying the relation between COR and the GC-content of sequences, it is important to consider that these two variables do not evolve at the same time scales. Hotspots of recombination vary in position and intensity even among individuals of the same species (chapter I.2.4.3), while the genomic sequence is highly conserved between closely related species (Chimpanzee Sequencing and Analysis Consortium, 2005). One approach to this problem consists in estimating the current substitution pattern of DNA sequences, and deducing the equilibrium GC-content reached by sequences evolving with that substitution pattern. Thus, the GC content reached by a sequence, also termed **equilibrium GC-content (GC*)**, is indicative of the current processes affecting the evolution of the genome, such as COR (Meunier and Duret, 2004). Under a simple model of nucleotide substitution, with no neighbor-dependence, GC* is calculated as the percentage of AT→GC substitutions among all AT→GC and GC→AT substitutions (Sueoka, 1962):

$$GC* = \frac{r_{AT\to GC}}{r_{AT\to GC} + r_{GC\to AT}} \tag{II.16}$$

A simple method to compute the substitution frequencies is through **parsimony** (Meunier and Duret, 2004). The principle of parsimony consists in minimizing the number of changes that can occur on the branches of a given phylogenetic tree. Moreover, in parsimony all the substitutions have the same probability of occurrence. Figure II.12 illustrates a three species tree. If at a site, the bases C, A, and C are observed in species 1, 2, and 3 respectively, the most parsimonious scenario would be that the ancestral sequences 4 and 0 contain also a C. A substitution $C \to A$ is inferred on the branch from 4 to 2. However, as divergence levels increase, the power to detect substitutions by parsimony decreases.

II.3.1.1 Maximum-likelihood GC* estimation on a tree with N branches

A more reliable methodology for substitution estimations is the maximum-likelihood approach (Arndt et al., 2003; Duret and Arndt, 2008). This method allows the estimation of a different substitution matrix along each branch of a phylogenetic tree. Assuming the phylogenetic tree of three species, illustrated in figure II.12, the evolutionary dynamics can be summed up along each branch through a substitution matrix, Q, for which $r_{\alpha\to\beta}$ stands for the substitution rate from α to β on that branch:

$$Q = \begin{pmatrix} r_{AA} & r_{C\to A} & r_{G\to A} & r_{T\to A} \\ r_{A\to C} & r_{CC} & r_{G\to C} & r_{T\to C} \\ r_{A\to G} & r_{C\to G} & r_{GG} & r_{T\to G} \\ r_{A\to T} & r_{C\to T} & r_{G\to T} & r_{TT} \end{pmatrix} \tag{II.17}$$

A **general** substitution matrix considers all 12 possible rates of substitution between nucleotides. By definition, the diagonal elements are constrained by the requirement that

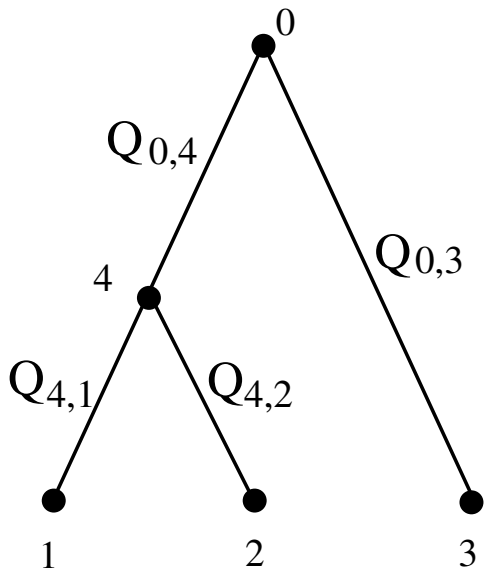

Figure II.12: *A phylogenetic tree of three species. The tree has 3 leaves corresponding to the 3 species 1-3, one internal node 4 and a root 0. $Q_{i,j}$ stands for the substitution matrix along the branch that joins i to j.*

Symbol	**Definition**
α, β	Nucleotides at a site in the sequence
$r_{\alpha\to\beta}$	Substitution rate from α to β
Q	Substitution matrix
t	Time
$p_\beta(t)$	The probability of nucleotide β at a certain position, at time t
$P(t)$	Probability matrix of occurence for all possible substitutions during a period of time t
N	The number of current day sequences in the phylogenetic tree
M	The number of internal nodes in the tree
S	The gap-free length of the alignment between sequences
$\overrightarrow{\alpha}^i$	The sequence of length S corresponding to the leaf i in the tree
$p^{(0)}(\overrightarrow{\alpha}^0)$	The probability to have $\overrightarrow{\alpha}^0$ as the ancestral sequence
$P_{r^{(i,j)}(\overrightarrow{\alpha}^j\|\overrightarrow{\alpha}^i)}$	Substitution probabilities of sequences along the branch (i,j)
L	Likelihood to observe the current day sequences on the leaf nodes under a given phylogeny

Table II.2: *Parameters of the maximum-likelihood approach for the inference of substitution patterns along branches in a tree.*

all the elements of a column should sum to 0. Moreover, a simplifying hypothesis about the matrix II.17 consists in ignoring multiple substitutions at a given site. This could be true under the infinite site model in the case of small branch lengths. Along any branch, the substitution rates are constant, and the model is termed **homogeneous**.

Given the double-stranded nature of the DNA molecule, nucleotides are paired between the two complementary strands, A - T and C - G. For neutrally evolving sites, under

no mutation or repair strand bias, the rates of substitution are not strand-dependent. The strand **complement symmetry** property of the matrix II.17 reduces the number of parameters from 12 to 6 (*i.e.* $r_{C\to A} = r_{G\to T}$).

The probabilities of occurrence of all possible substitutions during a period of time t are then derived. For a small increment time Δt, at a certain position in a sequence, the change in frequency of a nucleotide β is dependent on the probability of β at time t, minus the substitutions from β, plus the substitutions towards β in the interval Δt:

$$p_\beta(t+\Delta t) = p_\beta(t) - p_\beta(t) r_{\beta\to\beta}\Delta t + \sum_{\alpha\neq\beta} p_\alpha(t) r_{\alpha\to\beta}\Delta t$$

The relation between the evolution of the probability P and the substitution matrices Q is:

$$\begin{aligned} P(t+\Delta t) &= P(t) + QP(t)\Delta t \\ \frac{\partial}{\partial t}P_\beta(t) &= \sum_\alpha Q_{\alpha\beta}P_\alpha(t) \\ P(t) &= e^{Qt} \end{aligned} \tag{II.18}$$

For the model described above, evolution is assumed to proceed independently at each nucleotide site. The model from Arndt and Hwa (2005) considers the existence of a second type of substitutions which are site-dependent. Some mutation processes, such as the deamination of cytosines to thymine are highly influenced by the genomic context. In mammals, the substitution rate of C and G nucleotides is increased 10-fold for CG dinucleotides (termed *CpG* dinucleotides, where p stands for a phosphate bond) than for the rest of the dinucleotides (Hess et al., 1994). Hypermutations of CpG dinucleotides leads to their depletion and an excess of TpG and CpA. Any dinucleotide dependence can be modeled with a substitution matrix, by comparing the type of dinucleotides on the ancestral sequence, $\alpha\beta$, with the corresponding pair on the present sequence, $\alpha'\beta'$:

$$Q^{CpG}_{\alpha'\beta'\alpha\beta} = \begin{cases} r_{CpG\to CpA/TpG} & \text{if}(\alpha\beta = CG \text{ and } \alpha'\beta' = CA) \text{ or } (\alpha\beta = CG \text{ and } \alpha'\beta' = TG) \\ -2r_{CpG\to CpA/TpG} & \text{if}(\alpha\beta = CG \text{ and } \alpha'\beta' = CG) \\ 0 & \text{otherwise} \end{cases} \tag{II.19}$$

In practice, the model addresses the CpG dinucleotides hypermutability by examining the substitutions in sets of three consecutive nucleotides. The $4^3 \times 4^3$ substitution matrix is defined in this case as:

$$Q = Q\otimes I\otimes I + I\otimes Q\otimes I + I\otimes I\otimes Q + Q^{CpG}\otimes I + I\otimes Q^{CpG} \tag{II.20}$$

where I is the identity matrix and the operator $\otimes$ represents the tensor product of matrices defined as:

$$A\otimes B = \begin{bmatrix} a_{11}B & \cdots & a_{1n}B \\ \vdots & \ddots & \vdots \\ a_{m1}B & \cdots & a_{mn}B \end{bmatrix} \tag{II.21}$$

As in equation II.18, the time evolution of the probability matrix for any consecutive tri-nucleotides $\beta_1\beta_2\beta_3$ is:

$$\frac{\partial}{\partial t}P_{\beta_1\beta_2\beta_3}(t) = \sum_{\alpha_1\alpha_2\alpha_3} Q_{\beta_1\beta_2\beta_3\alpha_1\alpha_2\alpha_3}P_{\alpha_1\alpha_2\alpha_3}(t) \tag{II.22}$$

II.3.1.2 Estimating parameters

In order to estimate the equilibrium GC-content towards which the sequences are evolving, one needs to infer the nucleotide substitution rates, knowing the genomic sequences of present day species. This amounts to estimating the parameters of the evolutionary models described above. A simple, widely-used model, consists in affecting the same Q matrix to all branches in a tree. This is a simplifying model, and in reality the evolutionary forces - such as mutation and selection - need not be the same on each branch of the phylogenetic tree. Nevertheless, a trade-off has to be found between the number of parameters to estimate and the biological reality of the model.

The data consists of gap-free alignment between N current day sequences ($\overrightarrow{\alpha}^i$), of equal length S, and a phylogenetic tree with M internal nodes. The likelihood of the observed sequences knowing the phylogeny is the sum over all the internal nodes representing the ancestral unknown sequences:

$$L = \sum_{\overrightarrow{\alpha}^0,\ \overrightarrow{\alpha}^{N+1},\ \cdots,\ \overrightarrow{\alpha}^{N+M}} p^{(0)}(\overrightarrow{\alpha}^0)\prod_{i,j} P_{r^{i,j}}(\overrightarrow{\alpha}^j|\overrightarrow{\alpha}^i) \tag{II.23}$$

For a three species tree, equation II.23 becomes:

$$L = \sum_{\overrightarrow{\alpha}^0,\ \overrightarrow{\alpha}^4} p^{(0)}(\overrightarrow{\alpha}^0)P_{r^{0,4}}(\overrightarrow{\alpha}^0|\overrightarrow{\alpha}^4)P_{r^{0,3}}(\overrightarrow{\alpha}^0|\overrightarrow{\alpha}^3)P_{r^{4,1}}(\overrightarrow{\alpha}^4|\overrightarrow{\alpha}^1)P_{r^{4,2}}(\overrightarrow{\alpha}^4|\overrightarrow{\alpha}^2) \tag{II.24}$$

The estimation of the substitution frequencies proceeds from the maximization of the likelihood functions obtained by varying all the substitution parameters along each branch of the tree. Under the assumption of independence between sites in a sequence, the likelihood of the alignment is the product of the likelihoods at each single site. The likelihood function at a site is defined as the sum of all possible internal unknown states starting from the root (Squartini, 2010). In the case of independent sites, the likelihood of a tree can be easily computed by algorithms such as the *pruning* algorithm (Felsenstein, 1981). This algorithm consists in storing and using the conditional probabilities at intermediary nodes to compute the likelihoods of superior levels towards the root. After the pruning, equation II.24 becomes:

$$L = \sum_{\overrightarrow{\alpha}^0} p^{(0)}(\overrightarrow{\alpha}^0)P_{r^{0,3}}(\overrightarrow{\alpha}^0|\overrightarrow{\alpha}^3)P_{r^{0,4}}(\overrightarrow{\alpha}^0|\overrightarrow{\alpha}^4)\sum_{\overrightarrow{\alpha}^4} P_{r^{4,1}}(\overrightarrow{\alpha}^4|\overrightarrow{\alpha}^1)P_{r^{4,2}}(\overrightarrow{\alpha}^4|\overrightarrow{\alpha}^2) \tag{II.25}$$

where $\sum_{\overrightarrow{\alpha}^4} P_{r^{4,1}}(\overrightarrow{\alpha}^4|\overrightarrow{\alpha}^1)P_{r^{4,2}}(\overrightarrow{\alpha}^4|\overrightarrow{\alpha}^2)$ is the conditional probability at internal node 4.
If the substitution matrices are different along each branch, the number of parameters to be estimated for each site is $6(N+M)$ and 3 additional free parameters for the frequencies of nucleotides (A, C, G, T) in the ancestral sequence. The maximization of the likelihood functions for each parameter yields estimates of the substitution frequencies.

In the case of neighbor dependent substitutions, the likelihood of a tree is no longer the product of the likelihoods at all sites. *L*s are computed using a Monte-Carlo Maximum-Likelihood approach (Arndt and Hwa, 2005). The sequences at all the internal nodes are set as being the consensus of the descendant sequences or one random descendant sequence if no consensus can be built. For each branch the substitution frequencies between ancestral and present-day sequences are then estimated using a maximum-likelihood approach (Arndt et al., 2003). Given these substitution frequencies, each sequence at the internal nodes is re-estimated. A nucleotide in a triplet of sites is replaced by the one with the maximum likelihood, given the substitution matrix and the neighboring nucleotides. The algorithm is repetead multiple times until the convergence of tri-nucleotide distribution in the sequence at the root node.

The implementation of this model has led to the estimation of the GC* in human (Arndt et al., 2003; Webster et al., 2005; Duret and Arndt, 2008). These estimations were also performed along a three species tree, human, chimpanzee, macaque, or from alignments between the ancestral (consensus) and the current day sequences of transposable elements (TEs) (Arndt et al., 2003; Webster et al., 2005, 2006).

II.3.1.3 Maximum-likelihood GC* estimation on transposable elements

When GC* is estimated from the alignment of transposable elements (TEs) to their consensus sequences (Arndt et al., 2003), the model is simplified as the ancestral sequence of each TE subfamily is supposed to be known. Each TE subfamily is considered to evolve under a substitution matrix Q. For each position along the alignment of a TE to its corresponding ancestral sequence, the number of occurrences of $\alpha_i \rightarrow \beta_i$ knowing the adjacent bases, α_{i-1} and α_{i+1}, are counted. The total number of observed substitutions in such an alignment is $C(\beta|\alpha_L, \alpha, \alpha_R)$. The set of all counts for a given TE subfamily is noted $\{C\}$. The set of substitutions that best explains these counts is obtained by maximizing the following likelihood:

$$L(\{C\}|Q) = \prod_{\beta,\alpha_L,\alpha,\alpha_R} P(*\beta*|\alpha_L, \alpha, \alpha_R; Q)^{C(\beta|\alpha_L,\alpha,\alpha_R)} \quad \text{(II.26)}$$

where $P(*\beta*|\alpha_L, \alpha, \alpha_R; Q)$ is the probability that the ancestral sequence $\alpha_L, \alpha, \alpha_R$ will result, under the substitution model Q, in the $\alpha \rightarrow \beta$ substitution.

Review articles and dissertations for this sub-chapter: (Arndt et al., 2003; Arndt and Hwa, 2005; Duret and Arndt, 2008; Squartini, 2010).

II.3.2 Theoretical gBGC model

Meunier and Duret (2004) have shown that COR is a better predictor of the GC* than of the current GC-content. This result is the consequence of the evolution at different time scales of recombination and nucleotide composition. Recombination affects the substitution pattern through GC biased gene conversion (gBGC) (chapter I.3). However, recombination hotspots are short lived (chapter I.2.4). At such short time scales the bias in substitution rates doesn't generate extensive changes in the GC-content of genomic sequences in a population. Moreover, the extent of this impact was questioned by the low density of

Symbol	Definition
S	Nucleotide of type C or G
W	Nucleotide of type A or T
f	Fraction of the genomic regions involved in a hotspot
$\mu_{S\to W}$	Mutation rate from S to W
$\mu_{W\to S}$	Mutation rate from W to S
N	Effective population size
$r_{S\to W}$	Substitution rate from S to W
$r_{W\to S}$	Substitution rate from W to S
s	Strength of gBGC
$P(s)$	Probability of a mutation under gBGC to get fixed in a population
$P(0)$	Probability of a mutation without gBGC to get fixed by genetic drift
i	Recombination rate
k	Constant depending on hotspot length and bias in gBGC repair

Table II.3: *Parameters of model for the gBGC impact on the substitution frequencies (Duret and Arndt, 2008).*

recombination hotspots in the human genome (only 3% of the genome) (Myers et al., 2005; Spencer et al., 2006). Duret and Arndt (2008) proposed a theoretical model, with realistic parameters, for the role of gBGC on the substitution pattern and subsequently on the GC-content of sequences.

The model considers that gBGC is active only in the f% of the genomic region corresponding to recombination hotspots. The mutation process is considered identical within the hotspots and outside of them. The rates of substitution in neutrally evolving genomic regions are:

$$\begin{aligned} r_{W\to S} &= (1-f)2N\mu_{W\to S}P(0) + f2N\mu_{W\to S}P(s) \\ r_{S\to W} &= (1-f)2N\mu_{S\to W}P(0) + f2N\mu_{S\to W}P(-s) \end{aligned} \tag{II.27}$$

where the parameters are explained in table II.3 and $P(0) = \frac{1}{2N}$. The parameter s is proportional to the intensity of recombination hotspots, $s = k \times i$, with different k values estimated from (Spencer et al., 2006). The probability of fixation for a mutation subject to s is:

$$P(s) = \frac{1 - e^{-2s}}{1 - e^{-4Ns}} \tag{II.28}$$

Two applications of this model at a 1 Mb scale, termed M1 and M2, have been explored. The M1 approach considers that the density of hotspots, f, in a 1 Mb window varies between 0.05% and 10.7%. For the M2 approach, i varies from 0.66 to 157 cM/Mb. The correlation of the GC* predicted by the model II.27, under M1 and M2, to the COR is very close with the GC* inferred on real data using the maximum-likelihood method described in chapter II.3.1. No significant difference has been found between the predictions of the two models (Duret and Arndt, 2008). The model of (Duret and Arndt, 2008) explains the relation between COR and GC* observed in humans. It further quantifies the impact of gBGC at the nucleotide level, concluding that indeed the current strength of gBGC is not

sufficient to maintain the isochore structure observed in this species. This is consistent with the phenomenon of isochore erosion (Duret et al., 2002).

II.3.3 Our model on the effect of gBGC on the frequency of deleterious mutations in human populations

Studying the impact of recombination on the level of polymorphism in different genomes is important in view of the dual role of this mechanism. First, recombination is thought to affect local effective population size along the genome, and consequently polymorphism levels (Otto and Barton, 1997). The effective population size (N_e) is a measure of the total population size that overcomes difficulties raised by complex mating systems, spatial structure, and overlapping generations by considering an ideal population representative of the real one. In regions of low recombination, the linkage between genes is high, and selective sweep of advantageous mutations (Smith and Haigh, 1974) or elimination of deleterious alleles (the so-called background selection) (Charlesworth et al., 1993) are expected to reduce diversity at linked loci. Positive correlations between recombination rate and genomic diversity have been reported in multiple organisms (Stephan and Langley, 1989; Nachman et al., 1998; Charlesworth, 2009). Second, loci subject to recombination and gBGC should experiment a reduction in polymorphism, due to the bias favoring the fixation of G and C alleles.

In 1982, Lamb and Helmi have proposed a first model of the joint effects of selection, mutation and biased gene conversion in an infinite population (Lamb and Helmi, 1982). A model of this same mechanism, in a finite population has been developed by Nagylaki (Nagylaki, 1983a,b) for a subset of cases (the dominance aspects were not treated). The conclusion of these studies was that BGC is acting as directional selection. Based on these results, Duret and Arndt (2008) have modeled the impact of gBGC and mutation on neutrally evolving sequences, in human. Given realistic parameters for the intensity and density of recombination hotspots, the model predicts very well the observed relation between recombination and substitution pattern. Thus, it represents a quantitative proof of the important role played by gBGC in shaping the isochore landscape of the human genome. Additionally, gBGC has also been found to affect sequences under selection. A first line of evidence in this sense came from the study of the substitution pattern in regulatory elements, which were found to have an important GC bias (Galtier and Duret, 2007). More recently, this type of bias was also observed in human protein-coding exons (Berglund et al., 2009; Galtier et al., 2009). It follows that even in functional sequences, gBGC can sometimes counteract the effect of selection and lead to the accumulation of deleterious mutations. This has led to the recombination hotspots being named 'Achilles' heel' of our genome (Galtier and Duret, 2007).

In collaboration with Laurent Duret, Anamaria Necsulea, David Cooper, and Peter Stenson, we have started to work on the quantification of the part played by gBGC in the maintenance of deleterious alleles in the human population (Necşulea et al., 2011). Our approach was based on simulations of the impact of gBGC on the frequencies of alleles in a finite size population. We considered that sites are independent, and subject to mutation, selection, and gBGC. For simplification considerations, the initial population is considered

homozygous, subject to random mating, and following a Fisher-Wright probabilistic model with multinomial sampling, ensuring a constant population size over time. Simulations were run for over 20 000 generations, for a population of effective size: 10^4 individuals (Yu et al., 2004), with a mutation rate of 10^{-8} mutations per base-pair per individual per generation (Nachman and Crowell, 2000).

The alleles that can segregate at each locus belong to one of two classes: S(trong) (G and C) or W(eak) (A and T). The action of selection was modeled through the fitness coefficients ω_{SS}, ω_{SW} and ω_{WW} for the individuals **SS**, **SW** and **WW**, respectively. These coefficients lie within the interval $[0, 1]$ ($\omega = 1$ for neutrally evolving loci, and $\omega < 1$ for loci under negative selection). The mean fitness value is $\overline{\omega}$:

$$\overline{\omega} = z_{SS}\omega_{SS} + z_{SW}\omega_{SW} + z_{WW}\omega_{WW}$$

where z denotes the zygotic frequencies.

For individuals that were heterozygous at a given locus (**SW**), we termed u the probability of conversion $S \longrightarrow W$ and v the probability of conversion $W \longrightarrow S$. The gene conversion bias at this site is measured through $\delta = v - u$ and has positive values when GC-biased gene conversion occurs. We termed the frequency of the S allele p and hence the frequency of allele W is $1 - p$. The model describes the transition from one generation, n, to the next, $n + 1$, admitting panmixia, with the following equations:

adults n **:** $f_{SS} = p_n^2$; $f_{WS} = 2p_n(1 - p_n)$; $f_{WW} = (1 - p_n)^2$

gametes $n + 1$ **:** $g_S = p_n + (p_n - p_n^2)\delta$; $g_W = 1 - g_S$

zygotes $n + 1$ **:** $z_{SS} = g_S^2$; $z_{SW} = 2g_S g_W$; $z_{WW} = g_W^2$

adults $n + 1$ **:** ${f^*}_{SS} = \frac{\omega_{SS}}{\overline{\omega}} z_{SS}$; ${f^*}_{WS} = \frac{\omega_{SW}}{\overline{\omega}} z_{SW}$; ${f^*}_{WW} = \frac{\omega_{WW}}{\overline{\omega}} z_{WW}$

alleles $n + 1$ **:** $p_{n+1} = {f^*}_{SS} + \frac{1}{2}{f^*}_{SW}$

where f represents the frequency of individuals at generation n, g the frequency of gametes at generation $n + 1$, f^* the frequency of individuals at generation $n + 1$ and p_{n+1} the frequency of the S allele at generation $n + 1$.

Here we have only analyzed mutations that were both deleterious and recessive. We termed s the selection coefficient, so that the fitness of individuals homozygous for the mutant allele is $\omega = 1 - s$. Thus, for the simulations of the fate of a newly-arisen $W \to S$ mutation in a WW population, we have $\omega_{SS} = \omega$, $\omega_{SW} = 1 - (1 - \omega)h$ and $\omega_{WW} = 1$. For the simulations of the fate of a newly-arisen $S \to W$ mutation in an SS population, the fitness coefficients are $\omega_{SS} = 1$, $\omega_{SW} = 1 - (1 - \omega)h$ and $\omega_{WW} = \omega$.

Combinations with the following parameters values have been simulated :

- N_e, the effective population size, 10^4.
- μ, the mutation rate per bp, per individual, per generation, is 10^{-8}.

- δ, the coefficient of biased gene conversion. Three values were tested: 0 (no BGC, only selection), 0.00013 (a mild hotspot of recombination - $40\frac{cM}{Mb}$) and 0.0013 (a strong hotspot of recombination - $400\frac{cM}{Mb}$). The population scaled gBGC coefficient ($N_e\delta$) in the human genome was estimated by Spencer et al. (2006) by analyzing the DAF spectra of noncoding SNPs. In genomic regions of high recombination, their estimate was $N_e\delta = 0.325$. Given that, in the human genome, the average COR of recombination hotspots is 40 cM/Mb Myers et al. (2006), it is expected that the gBGC coefficient should be about 16 times higher in these hotspots. We therefore considered two values of the population-scale gBGC coefficient: $N_e\delta = 1.3$ and $N_e\delta = 13$.

- ω, the fitness coefficient of the homozygous derived allele, can take the following values : 1 (neutral), $1 - 10^{-4}$ (recessive and slightly deleterious), $1 - 10^{-3}$ (recessive and mildly deleterious), $1 - 10^{-2}$ (recessive and fairly deleterious) and $1 - 10^{-1}$ (recessive and highly deleterious).

- h, the fitness coefficient of the heterozygous allele takes the values 0 and 0.3.

The results of these simulations, for different parameter values, are given in figure II.13. These results show that due to gBGC, $AT \rightarrow GC$ substitutions segregate at higher frequency than $GC \rightarrow AT$ one. This effect is detected for nearly neutral ($|N_e s| = 1$) and mildly deleterious ($|N_e s| = 10$) mutations. In the case of intense recombination hotspots, the impact of gBGC could be detected even for highly deleterious mutations ($|N_e s| = 100$). The same trends are observed when the heterozygous is also negatively selected ($h = 0.3$), the main difference being that the recessive alleles are very quickly eliminated from the population and only a few could attain non-null frequencies.

The conclusions of our simulations are in agreement with observations from real data. In the article by Necşulea et al. (2011), the relation between the substitution pattern and recombination rate is studied for three types of genomic regions: intergenic, synonymous, and non-synonymous. SNPs are retrieved from HapMap Project phase III (International HapMap Consortium, 2007). The non-synonymous sites are further classified according to their association to diseases as predicted by PolyPhen (Sunyaev et al., 2001) and human gene mutation (HGMD) database (Stenson et al., 2009). Figure II.14 represents the DAF of $AT \rightarrow GC$ and $GC \rightarrow AT$ mutations in regions of high recombination. For all classes of mutations considered, the $AT \rightarrow GC$ mutations segregate at higher frequencies than the $GC \rightarrow AT$ mutations. This result holds true even for non-synonymous sites associated with human disease. This analysis confirms the impact of gBGC on functional sites, and quantifies for the first time that gBGC is responsible for the maintenance of deleterious mutations at high frequencies in human population.

These results have been recently confirmed through a theoretical study of the effect of mutation, selection, drift, and gBGC on the fate of deleterious mutations (Glémin, 2010). Furthermore, the population genetics model developed by Glémin (2010) characterizes the effect of gBGC on the mutational load. The mutational load results from an accumulation of mutations leading to a reduction in the mean fitness of a population. BGC was previously reported to increase the mutational load (Bengtsson, 1990), up to an estimated values of 10% in human (Glémin, 2010).

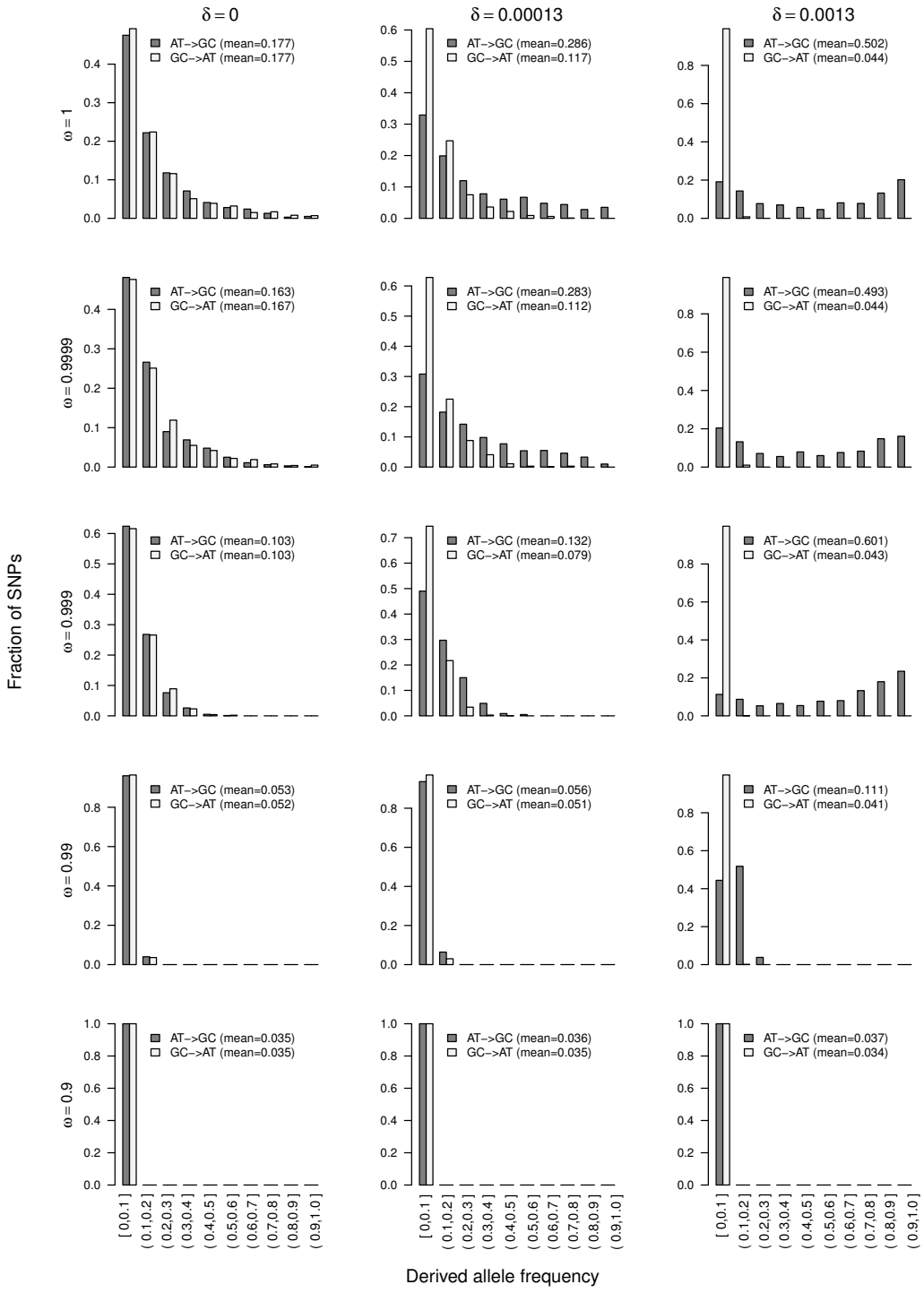

Figure II.13: *Derived allele frequencies spectrum obtained through simulations with different parameter sets. Represented in light gray are the distributions of derived allele frequencies for GC → AT alleles, and in dark gray, those of AT → GC alleles. The population-scaled selection coefficient ($N_e s$) and the population-scaled biased gene conversion parameter ($N_e \delta$) are indicated for each graph. For this graph the deleterious allele is completely recessive ($h = 0$). From Necşulea et al. (2011).*

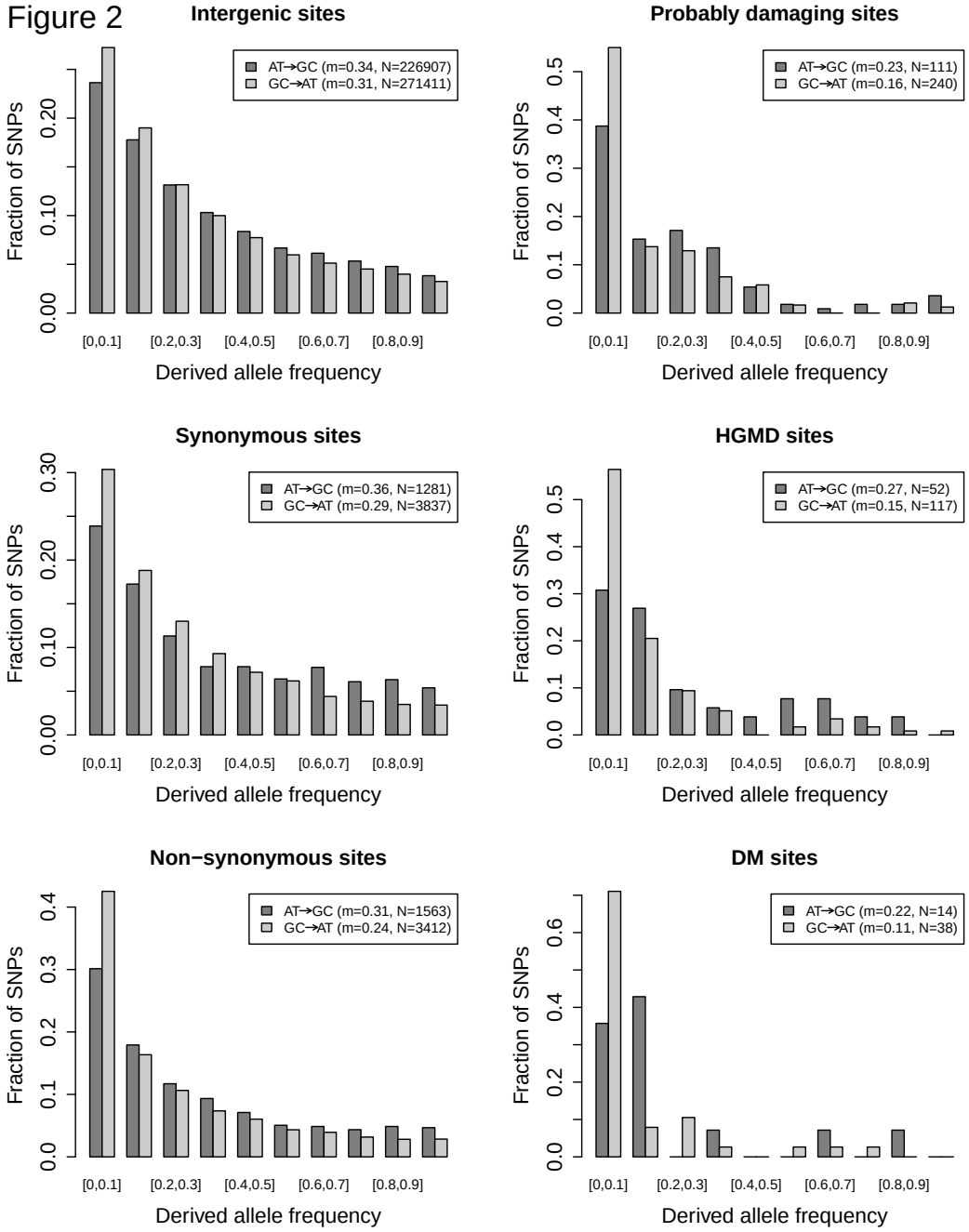

Figure II.14: *Derived allele frequency spectra for the HapMap YRI sample, for different genomic regions and classes of nonsynonymous SNPs. The data presented here relate only to the high recombination class. Dark gray: AT → GC mutations, light gray: GC → AT mutations. The left column represents the DAF of HapMap SNPs classified according to their genomic class. The right column contains the DAFs of non-synonymous SNPs related to disease-association. "Probably damaging sites" are predicted by PolyPhen. "HGMD sites" are all the sites described in HGMD database. "DM sites" is a subset of the "HGMD sites" representing only those mutations regarded as being a direct cause of disease. From Necşulea et al. (2011).*

II.4 Conclusion

This chapter presents some of the major models and approaches that have been used to identify, quantify, and characterize recombination rates and the evolution of nucleotide composition. The main experimental methods for the localization of recombination events are described in the first section of this chapter. These include linkage and linkage disequilibrium (LD) mapping as well as sperm-typing techniques. Each one of these methods presents advantages and disadvantages. **Linkage mapping** offer low-resolution but sex-specific, genomes wide information about CO distribution. **LD studies** characterize CO production with a high-resolution, but this information is sex- and time-averaged. The higher resolution of recombination hotspots, indicative of both CO and NCO events, is attained through **sperm-typing techniques**, however, this approach is very regional and indicative only of male meiosis.

The data generated by such methods are then used to model the distribution of CO events along chromosomes in order to better understand the variations we observe in terms of localization and intensity. Four models are presented in detail. The first three models study the distribution of CO events and strength of interference. **The counting model** (Foss et al., 1993) considers that the inter-CO distance follows a χ^2 distribution with a parameter representing the average number of NCOs separating two consecutive COs. **The mechanical stress model** (Kleckner et al., 2004) represents COs as cracks in the chromatin fiber generated by the tension of chromosome condensation. **The polymerization model** (King and Mortimer, 1990) considers that COs initiate the polymerization along the synaptonemal complex. As polymerization extends from a CO it will prevent the formation of additional COs. While these three models of interference are based on the local distribution of CO events along chromosomes, the fourth model described in this section focuses on the relation between the total number of COs on a chromosome and its physical length (Li and Freudenberg, 2009). A linear relation is established between these two variables.

The last section of this chapter is focused on the relation between COR and the nucleotide composition of sequences. It is important, when analyzing this relation, to understand that these two variables are not acting at the same time scale. A maximum-likelihood model is presented that calculates the GC* of sequences under a constant substitution pattern. The second model quantifies the impact of COR on GC*. In order to understand the molecular mechanisms of biological processes such as recombination and gBGC, it is essential to put forward hypotheses. Models have proven an useful tool for validating hypotheses, especially given the availability of ever growing experimental data. We also present here our simulations of the impact of gBGC on the maintenance of deleterious mutations in the human population. Our results agree with the real data showing that even at non-synonymous sites, AT→GC disease-associated mutations segregate at a higher frequency than the GC→AT ones.

Chapter III

Karyotype and recombination pattern

In this chapter we detail the mathematical model we built in order to describe the influence of the karyotype on the rate and distribution of crossovers. Section III.1 contains a brief introduction of the models already existent in the literature, which were largely detailed in chapter II.2, as well as the motivation for a new model. Section III.2 specifies our model between the genetic and physical length of chromosomes. The data and mathematical tools used for testing the fitting of the model are also presented in this section. The last section III.3, presents the results we obtained by applying the model to vertebrate and invertebrate data.

III.1 Introduction

As we have thoroughly emphasized in chapter I, recombination is an essential process for the segregation of homologous chromosomes and for the evolution of genomes. Our current understanding of the molecular mechanisms of recombination results from experiments in a few model organisms. These experiments consist in the observation of recombination frequency and distribution, through two main types of methods : cytological and genetic. Cytological and immunofluorescent observations consist in directly examining and measuring the number and distribution of chiasma foci along chromosomes. Genetic studies, as described in chapter II.1, involve crossing, pedigree, linkage disequilibrium (LD), and sperm-typing experiments. However, the ideal data for the study of recombination would imply a genome-wide, sex-specific, high-resolution analysis, that provide information on both crossovers (CO) and non-crossovers (NCO). Unfortunately, few are the species for which the data that could be produced meet all or even some of these criteria. Recently, the first high-resolution (a few tens of bp) map of meiotic recombination (both COs and NCOs) covering the whole genome has been produced for *Saccharomyces cerevisiae* (Mancera et al., 2008). Although at lower resolutions (from a few ten kb to 1 Mb), the construction of human sex-specific CO maps has shed light on the important features of recombination in vertebrates (Coop et al., 2008; Kong et al., 2010). Such maps (∼200 kb resolution) were also constructed for chromosome 1 in *Mus musculus* (Paigen et al., 2008) and chromosome 4 in *Arabidopsis thaliana* (Drouaud et al., 2007). A deeper analysis of several vertebrate

recombination hotspots has been achieved through fine-scale sperm-typing experiments, covering a few kb, in human and mouse (Jeffreys et al., 2001; Yauk et al., 2003; Jeffreys et al., 2005; Kauppi et al., 2007; Webb et al., 2008; Wu et al., 2010).

Many conserved features of recombination, such as hotspot length, CO distribution, protein complexes involved, have been described in these model species (chapter I). However, experiments even in this limited number of species have uncovered **multiple important evolutionary differences in the process of recombination**. At the molecular level, the major known determinant of recombination hotspots, the *Prdm9* gene, has known a rapid evolution in different taxa (Oliver et al., 2009). At a more global level, differences in karyotype underline important inter-species variations in CO rate (COR) (chapter I). The karyotype of a species is characterized both by the number of chromosomes and their respective length. The number of chromosomes is intimately linked to the number of COs in a species through the requirement of at least one CO per chromosome (or chromosome arm) to ensure the proper segregation of homologs during meiosis (de Villena and Sapienza, 2001) (figure I.12). However, this relation is further complicated by the existence of multiple COs along some chromosomes. It is expected that the longer a chromosome is, the higher the number of COs that it can host. Nevertheless, the interference phenomenon demonstrates the existence of strong constraints on the distribution of recombination events (Bishop and Zickler, 2004). The molecular mechanisms responsible for interference ensure that consecutive COs do not form too close to each other. Figure III.1 illustrates the relation between the COR and the physical length of chromosomes in 13 vertebrates. As the physical size of chromosomes increases, the COR decreases until it reaches a "plateau" with a constant COR value (figure III.1). This plateau is generated by the strength of interference which limits the number of possible COs. There is a strong variability of this phenomenon not only among species, but also among chromosomes of the same species (Lian et al., 2008). Moreover, two types of COs exist, interfering and non-interfering (reviewed in Mézard (2006)). These differences in interference indicate that the relation between the number of chromosomes, the chromosomes sizes, and the number of COs is not straightforward.

In chapter II.2 we give an overview of the different models of the impact of interference-related mechanisms on the distribution of COs along the length of chromosomes. When measuring recombination rates, three distinct aspects of chromosome length are relevant: the genetic, physical, and synaptonemal complex (SC) lengths. The genetic length of a chromosome, expressed in cM, represents the average number of COs. The physical length, expressed in Mb, is a measure of the number of nucleotides making up the chromosome. The mitotic and SC lengths, both expressed in μm, by comparison with the Mb length, indicate the degree of chromosome condensation as cytologically observed at the mitotic or meiotic stage respectively.

The major models quantifying the strength of interference (counting and mechanical stress models) are adjusted on the genetic length of chromosomes (chapter II.2) and thus provide information on the expected number of NCOs separating consecutive COs. Cytological observations of the distribution of MLH1 foci (characteristic of CO sites) in mouse indicated that the distance between consecutive foci was approximately 70% of the SC length (Froenicke et al., 2002). However, different chromosomes have different levels of

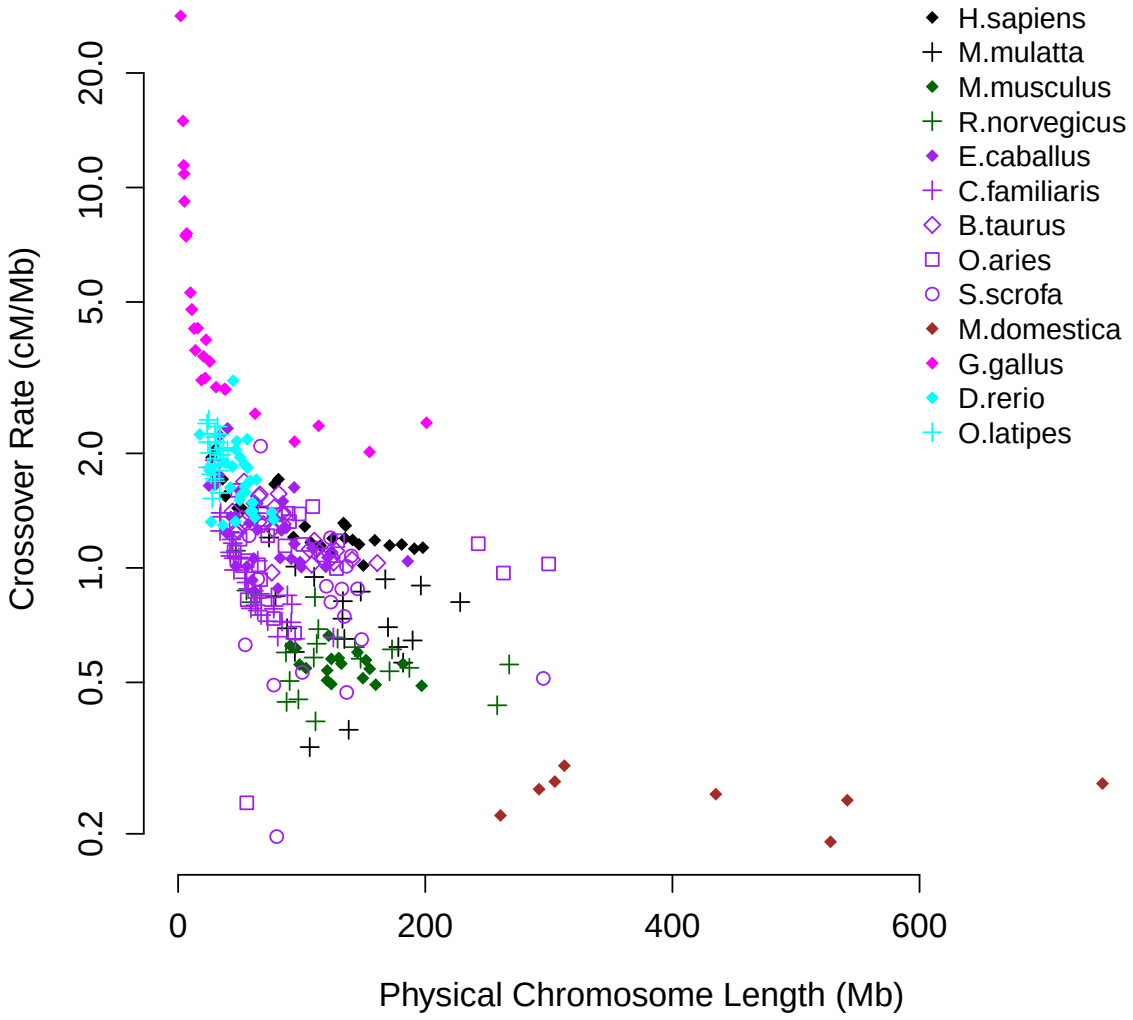

Figure III.1: *Representation of the relation between the CO rate (cM/Mb) and the physical length of chromosomes (Mb) for 13 vertebrates. The colors of the points are associated with the phylogenetic groups described hereafter in table III.1. The shape of the points differentiates between species inside each phylogenetic group. The y-axis is represented on a log-scale.*

condensation (Codina-Pascual et al., 2006; Froenicke et al., 2002) and thus, the inter-CO distance expressed in absolute SC length (μm) is variable from one chromosome to another (Froenicke et al., 2002).

While both the genetic and SC length are important determinants of CO interference, these studies were performed in a few model organisms with a fair resolution of the observed CO maps. However, low-resolution genetic maps, informative at the whole chromosome level, are available for a wide range of species. An inter-species comparison of these low-resolution maps could provide a better insight in the conserved and divergent features

of recombination. Li and Freudenberg (2009) have analyzed the relationship between the genetic and the physical length of chromosomes, with a linear regression model (figure II.11). The slope of the regression line corresponds to the estimated COR and the intercept represents the inferred recombination rate for infinitely small chromosomes. Despite a good fit on the data, the intercept of the linear model is generally smaller than the 50 cM expected under the obligatory CO condition. It is thus difficult to interpret this parameter biologically. Moreover, as shown in figure III.2, the model does not fit well the data for the chicken, which is particular in that it has a number of very small chromosomes.

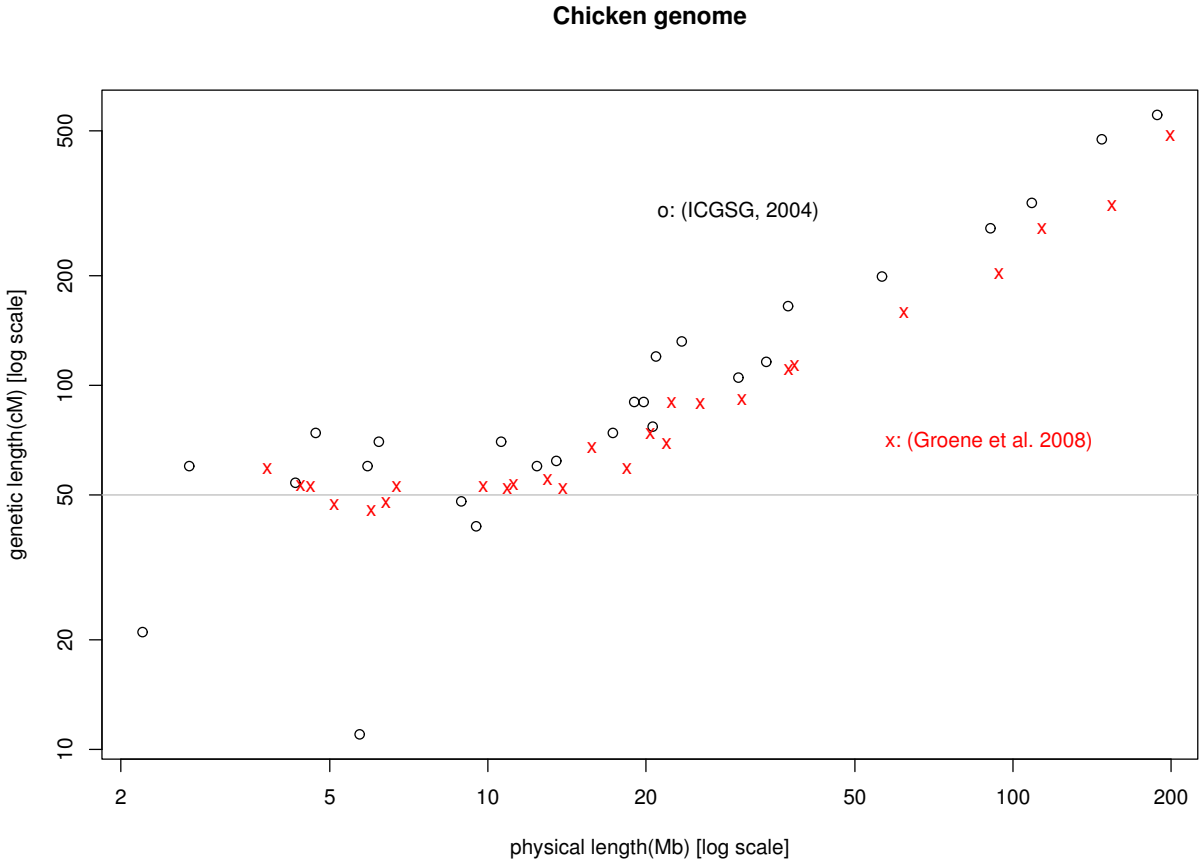

Figure III.2: *The genetic length (in cM) vs. physical length (in Mb) plot for chicken genome (*Gallus gallus*) in log-log scale. Data from two genetic maps are shown: International Chicken Genome Sequencing Consortium (2004) (black circles) and Groenen et al. (2009) (red "x"). From Li and Freudenberg (2009).*

The focus of our work, as for Li and Freudenberg (2009), is **the modelisation of the link between the genetic and the physical length of chromosomes**, at the global level of a species. The motivation behind this model is to capture the relation between CO production and karyotype in a wide variety of species. As opposed to the previous models from the literature, our model is non-linear. It is defined by two parameters: the mean rate of additional COs accumulation and the average interference strength. We apply our model to 13 vertebrate and 14 invertebrate species. For 10 out of the 13 vertebrate species, for which sex-specific genetic maps where available, we applied the model separately for the male and female recombination. The estimation of these parameters yields new insights in

the evolution of recombination differences in CO number and distribution between species, but also on differences between sexes.

III.2 Methods and data

III.2.1 Modeling the influence of karyotype on the recombination pattern

For a given species, let P be the total physical length (Mb) of a chromosome. We define $f(P)$ a function that links P to G, the total genetic length (cM) of the same chromosome. Considering the molecular mechanism of recombination, $f(P)$ is subject to different constraints. First, an obligatory CO per chromosome results in $f(P) \geq G_0 = 50$, for all P. Second, given an average length per species between two consecutive COs (hereafter termed interference length), P_I, all chromosomes of this species with $P \leq P_I$ can have one and only one CO, $f(P) = 50$ cM. This is equivalent with the first derivative of $f(P)$, $f'(P)$, becoming 0 as P tends to 0: $f'(P) \to 0$. Third, we make the hypothesis that for long chromosomes the relation between the genetic and physical lengths is linear and the slope of the linear relationship (r) equals the average rate with which additional COs are produced per chromosome: $f'(\infty) \to r$.

Under these assumptions, the non-linear model is:

$$\begin{aligned} \mathbf{G = f(P) = 50 + r \ln(1 + e^{P - P_I})} \\ f'(P) = r\frac{e^{P-P_I}}{1+e^{P-P_I}} \end{aligned} \tag{III.1}$$

Parameter r represents the increase in the COR, expressed in cM/Mb, subsequent to the obligatory CO per chromosome. Parameter P_I is the average estimate for a species of the physical interference length, accounting for both interfering and non-interfering COs as well as intra- and inter-chromosome variation.

The values in 0 and ∞ of the first derivative in equation III.1 are:

$$\begin{aligned} f'(0) = r\frac{1}{1+e^{P_I}} \\ f'(\infty) = r \end{aligned} \tag{III.2}$$

The function and its derivative in equation III.1 are drawn in figure III.3 for different parameter values. The parameter P_I represents the "plateau" before the linear part of the model, and r the slope of this linear domain.

III.2.2 Fitting models

III.2.2.1 Linear and non-linear least squares

Given a set of n data points, (P_1,G_1), (P_2,G_2), ..., (P_n,G_n), corresponding to the total physical and genetic lengths of n chromosomes, and a model $f(P, \theta)$, where θ represents the set of parameters, we want to find the best fitting curve. For this purpose, a score

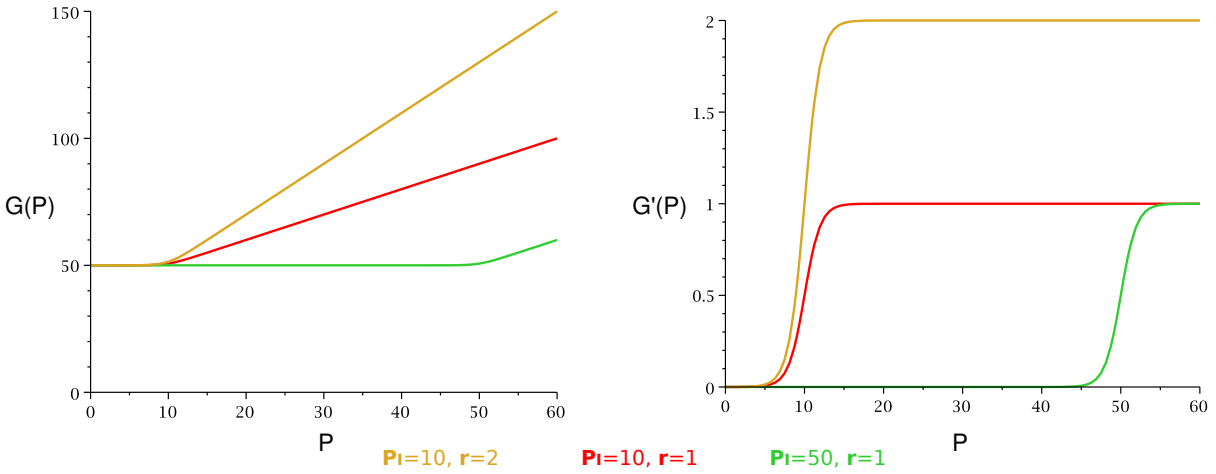

Figure III.3: *The model in equation III.1 and its derivative, for parameter values:* $\mathbf{P_I = 10}$**,** $\mathbf{r = 2}$ *in orange,* $\mathbf{P_I = 10}$**,** $\mathbf{r = 1}$ *in red and* $\mathbf{P_I = 50}$**,** $\mathbf{r = 1}$ *in green.*

function is needed to compute the goodness of fit of $f(P, \theta)$ to the dataset. One of the most commonly used score functions is the least-squares. For each pair (P_k,G_k), the residuals of a fit are $R_k = G_k - f(P_k, \theta)$. The least-squares method consists in minimizing the residual sum of squares (S):

$$S = \sum_{k=1}^{n} {R_k}^2 \tag{III.3}$$

Finding the minimum values of S is equivalent to finding the zero-values of its first derivative with respect to each parameter. In the case where $f(P, \theta)$ depends linearly on θ (a.k.a linear least-squares), the minimization is achievable in one step and yields an analytical solution. In the case of non-linear least squares problems, the minimization procedure is implemented iteratively, resulting in an heuristic solution. This further implies that starting values need to be provided to start the iterations.

An important step when fitting a non-linear least squares is the choice of starting parameter values. In order to avoid convergence towards local minima, for each of the two parameters we tested starting values ranging from 0.5 to 200. For the majority of starting values, the algorithm converged toward the $\hat{\theta}$ values. For simplicity and time considerations, we decided to use the following starting values for each species in our dataset: for r the slope of the linear model of Li and Freudenberg (2009) that we previously fitted on the data; and for P_I the minimum value of physical chromosome length. The linear model of Li and Freudenberg (2009) is described in equation II.14. All the mathematical analyses were performed in R.

III.2.2.2 Confidence interval

The outcome of model fitting yields estimates of the parameter values. As for any statistical inference, it is important to measure the uncertainty of these estimates by defining confidence intervals (CI). However, for **the non-linear estimations CIs can be only approximated, and are usually asymmetric** (Bates and Watts, 1988). In

order to find the approximate CIs for the parameters, we use a likelihood *profile* method (Bates and Watts, 1988). A *profile* consists in systematically fixing one parameter in the model at a specific value while varying the remaining parameters, identifying the best fit, and comparing it to the original model fit. For a given parameter (θ), the extremes of a CI are estimated by attributing θ a series of values (θ_1, ..., θ_m) above and below its estimated value ($\hat{\theta}$). For each one of these values, a statistic τ is calculated, which represents the signed root square of the ratio between the change in the residual sum of squares and the residual standard error (s^2):

$$\tau(\theta_i) = sign(\theta_i - \hat{\theta})\sqrt{\frac{S(\theta_i) - S(\hat{\theta})}{s^2}} \quad \text{(III.4)}$$

where $s^2 = \sum_{k=1}^{n} \frac{{R_k}^2}{n-p}$, with p the number of parameters of the model.

All $\tau(\theta_i)$ are then interpolated and the endpoints of the CI are found by comparing the τ and t-distributions. For all the species, except the methaterian *Monodelphis domestica* the CI is calculated with a p-value of 0.95. For *M. domestica*, the lack of data in the horizontal part of the model (50 cM) and the limited number of values in the linear part (only 8 chromosomes) complicate the adjustment of parameters. In order to infer the CI for this type of species, we release the constraints on the parameters and consider the CI for a corresponding p-value of 0.75. The same p-value was used for *Sus scrofa*, *Arabidopsis thaliana*, *Sorghum bicolor*, and *Zea mays*.

III.2.2.3 Comparing and grouping species

The purpose of estimating parameters r and P_I is to compare them between species in order to find resemblances, but also to better understand differences in the recombination mechanism. Given that for the majority of species, chromosomes lie mainly in the linear part of the model, the scarcity of data points on the 50 cM plateau results in big CIs for the parameter P_I. These CIs are frequently overlapping between species, thus, rendering their comparison difficult. When comparing species, we focused on the parameter r, indicative of an average per species rate of CO production additional to the obligatory CO. When the CIs of r values between two species overlap, the two species are considered similar in their average COR.

III.2.3 Data

For each species, we fit the model on the total physical and genetic length of all autosomes. Sexual chromosomes were excluded from this analysis, given their particular selective constraints, recombination activity, and data availability as opposed to autosomes.

III.2.3.1 Sex-averaged maps

We have acquired the sex-averaged genetic or linkage disequilibrium maps of 13 vertebrates and 14 non-vertebrates. The species are distributed according to different phylogenetic groups as detailed in table III.1. They have been chosen according to the availability of the

information on CO number, karyotype, and sequence assembly. The above classification does not account for the same level of variety inside each class. The group of *Primata* has a maximum divergence time of approximately 25 Million years (Myr) (Rhesus Macaque Genome Sequencing and Analysis Consortium, 2007), while that of *Teleostei* contains species having diverged ∼323 Myr ago (Kasahara et al., 2007). Nevertheless, this division is only qualitative and even though conclusions might be valid for one group, all analyses are performed individually for each species.

Phylogeny	**Latin name**	**Common name**	**References**
Primata	*Homo sapiens*	Human	(Matise et al., 2007)
Primata	*Macaca mulatta*	Rhesus Monkey	(Rogers et al., 2006)
Rodentia	*Mus musculus*	Mouse	(Cox et al., 2009)
Rodentia	*Rattus norvegicus*	Rat	(Jensen-Seaman et al., 2004)
Laurasiatheria	*Equus caballus*	Horse	(Swinburne et al., 2006)
Laurasiatheria	*Canis familiaris*	Dog	(Wong et al., 2010; DB-DogMap)
Laurasiatheria	*Bos taurus*	Cow	(Arias et al., 2009)
Laurasiatheria	*Ovis aries*	Sheep	(Poissant et al., 2010)
Laurasiatheria	*Sus scrofa*	Pig	(Vingborg et al., 2009)
Metatheria	*Monodelphis domestica*	Opossum	(Samollow et al., 2007)
Aves	*Gallus gallus*	Chicken	(Groenen et al., 2009)
Teleostei	*Danio rerio*	Zebrafish	MGH map (DB-ZFIN)
Teleostei	*Oryzias latipes*	Medaka	(Ahsan et al., 2008; DB-MedakaMap)
Insecta	*Apis mellifera*	Bee	(Beye et al., 2006)
Insecta	*Drosophila melanogaster*	Fruitfly	(DB-FlyBase)
Metazoa	*Ciona intestinalis*	Sea Vase	(Kano et al., 2006)
Metazoa	*Caenorhabditis elegans*	Round Worm	(DB-AceDB)
Fungi	*Saccharomyces cerevisiae*	Baker's Yeast	(DB-SGD)
Fungi	*Cryptococcus neoformans*	N.A.	(Marra et al., 2004)
Protista	*Trypanosoma brucei*	N.A.	(MacLeod et al., 2005)
Protista	*Plasmodium falciparum*	Malaria Parasite	(Su et al., 1999)
Plantae	*Populus trichocarpa*	Western Balsam Poplar	(DB-NCBI)
Plantae	*Vitis vinifera*	Grape Vine	(Doligez et al., 2006)
Plantae	*Arabidopsis thaliana*	Mouse-ear Cress	(Singer et al., 2006)

Plantae	*Oryza sativa*	Rice	(Harushima et al., 1998; DB-NCBI)
Plantae	*Zea mays*	Maize	(DB-NCBI)
Plantae	*Sorghum bicolor*	Sorghum	(Kim et al., 2005)

Table III.1: *Information on the sex-averaged genetic maps of the 27 vertebrates and non-vertebrates used in this study. Additional information about these species has previously been provided in table I.2.*

III.2.3.2 Sex-specific vertebrate genetic maps

In order to compare parameters between sexes, we fit the model separately for male and female genetic maps. The sex-specific genetic maps are available for the following 10 species: *Homo sapiens*, *Mus musculus*, *Canis familiaris*, *Ovis aries*, *Bos taurus*, *Sus scrofa*, *Monodelphis domestica*, *Gallus gallus*, *Danio rerio*, *Oryzias latipes*. The sex-specific maps for *Bos taurus* are based on Barendse et al. (1997). The female map of *Danio rerio* is the Heat Shock (HS) map from DB-ZFIN. The sex-specific maps for *Oryzias latipes* are based on Kimura et al. (2005). For the rest of the species, the source of the sex-specific genetic maps is the same as the above sex-averaged maps.

All the species will be referred in the rest of this manuscript, according to the initial of their genus name followed by the species name (*i.e Homo sapiens* is referred to as *H. sapiens*).

III.3 Inter-species differences in CO number and distribution

III.3.1 Estimates of the sex-averaged CO interference length and rate of COs

III.3.1.1 Vertebrate parameter values

We have fitted both the linear model of Li and Freudenberg (2009) (equation II.14) and our non-linear model (equation III.1) to the set of 13 vertebrates described in section III.2.3. Figure III.4 shows the data as well as the adjustment of the two models. For the majority of species, the two models result in similar interpolations of the data points (figure III.4). However, for species such as *C.familiaris* the interest of our non-linear model, which accounts for the plateau at 50 cM, is straightforward. For each vertebrate, **our non-linear model infers the length of the plateau at 50 cM, representing the interference length**. Given the strong heterogeneity of the genetic maps (the number, as well as the physical and the genetic lengths of chromosomes), the interference length is highly variable (figure III.4). Despite a general good fit of the models, the data of the mammal *S.scrofa* in figure III.4 seem to be poorly explained by both models. This might very well be an artifact of the quality of the genetic map. The genetic map of *S.scrofa*

has only 462 markers which are not distributed along the entire lengths of chromosomes (Vingborg et al., 2009).

In table III.2 we estimate the parameters for the linear and non-linear models. For the majority of species, the genetic maps are characterized by data points with a linear trend (figure III.4). This results in **estimates of the slope of the linear model and the *parameter* r of our non-linear model being very similar** (*i.e.* their corresponding CIs overlap) (table III.2). In these cases, the two parameters are both estimators of the average COR per species. However, the genetic map of *C. familiaris* is characterized by many chromosomes having a genetic length of approximately 50 cM (figure III.4). In this case, the linear model cannot predict their behavior. On the other hand, our non-linear model describes well the plateau at 50 cM (figure III.4). This results in differences among the estimates of the slope (0.341 cM/Mb) and r (0.482 cM/Mb) parameters, for the linear and non-linear models respectively (table III.2). A plateau at 50 cM for *G. gallus* has been previously observed by Li and Freudenberg (2009) (figure III.2). However, since the plateau region is not very wide for this species, the difference between the slope and the r estimates is not very strong (table III.2). Our model estimates an average COR of 0.448 cM/Mb for the *M. musculus* species, which is considerably stronger than the 0.398 cM/Mb estimated by the linear model (table III.2). The difference between the two estimates is mainly influenced by one data value, corresponding to the chromosome with the smallest physical length. This chromosome is also the only one in *M.musculus* to represent the 50 cM plateau (figure III.4). The same trend is also observed for *R. norvegicus*. A particularly marked difference characterizes the fish *O. latipes*, with the linear model predicting an average COR of 1.91 cM/Mb, smaller than our prediction of 2.57 cM/Mb (table III.2). For these last two species, the discrepancy between the linear and non-linear models comes mainly from the chromosomes with a genetic length smaller than 50 cM (figure III.4). All these species (*M. musculus*, *R.norvegicus*, *C. familiaris*, *G. gallus*, and *O. latipes*) have estimates of the *parameter* r similar but with higher values than the slope estimates of the linear model. One species shows the opposite trend. For *E. caballus*, r (0.886 cM/Mb) is slightly smaller than the slope (0.894 cM/Mb) (table III.2). By eliminating chromosomes 1 and 14, that show the highest dispersion (figure III.4), the two estimates become equal. **The non-linear model seems thus more robust to potential "outliers".**

Except for *H. sapiens* and *D. rerio*, the intercept of the linear model is smaller than 50 cM or even null (table III.2). However, the values of the total genetic length per chromosome indicate that all these vertebrates are subject to the obligatory CO condition (figure III.4). It is thus difficult to interpret the biological significance of this parameter.

One major improvement of the non-linear model over the previous models is that it yields estimates of the *interference parameter*, P_I, which measures the length of the 50 cM plateau (figure III.4). Considering that many species have mainly chromosomes with genetic lengths superior to 50 cM, **the estimates of this parameter are error-prone**. In some cases, like *H. sapiens*, *M. mulatta*, *S. scrofa*, and *D. rerio*, P_I cannot be considered statistically different from zero, as its corresponding CI contains the null value (table III.2). In order to improve the estimates of P_I for *H. sapiens* and *D. rerio*, the arms of metacentric chromosomes are considered separately. This approach is based on the observation in humans, that the obligatory CO constraint

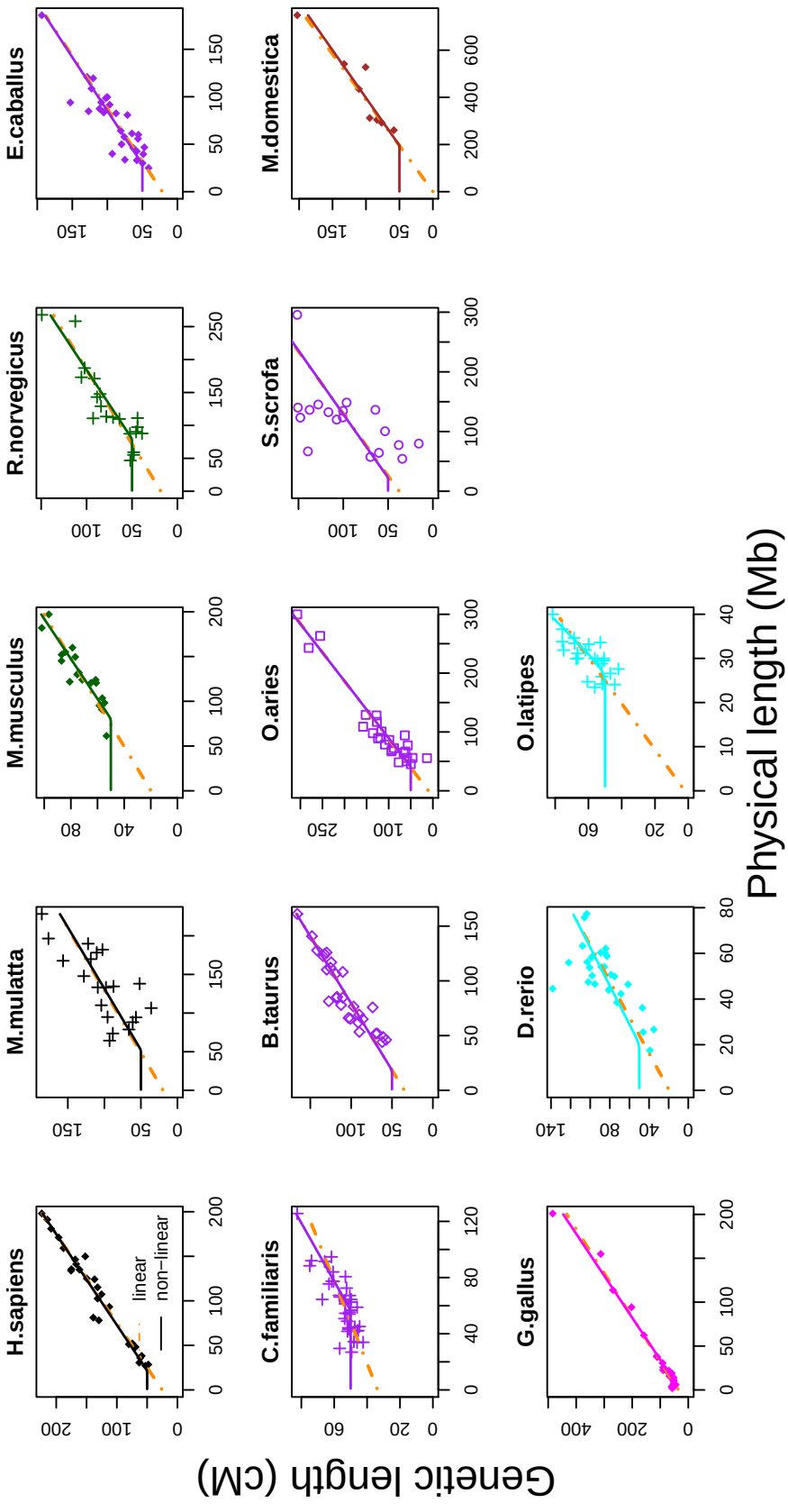

Figure III.4: *Representation of the physical and genetic lengths of chromosomes for all the 13 vertebrates investigated. These lengths relate to the entire chromosomes, except for the metacentric chromosomes of the primate* H. sapiens *and the fish* D.rerio *for which they involve the chromosome arms. Colors of the data points indicate the classification detailed in table III.1. The shape of points for each species is used to distinguish among species inside each phylogenetic group. The full and dashed lines represent the fitting of the non-linear and linear models respectively.*

Species	Linear model		Non-linear model	
	Intercept	Slope	P_I	r
H. sapiens - entire chromosomes	**43.4**±6.14*	**0.916**±0.043*	7.16 ∈ [-7.49 ; 19.4]	**0.916** ∈ [0.827 ; 1.01]
H. sapiens - metacentric arms	**28.5**±4.93*	**0.981**±0.0414*	**21.9** ∈ [12.6 ; 30.3]	**0.981** ∈ [0.896 ; 1.07]
M. mulatta	17±19.1	**0.63**±0.134*	52.3 ∈ [-19.9 ; 110]	**0.63** ∈ [0.349 ; 1.23]
M. musculus	**23.9**±6.19*	**0.398**±0.0462*	**80.4** ∈ [63.4 ; 92.2]	**0.448** ∈ [0.347 ; 0.55]
R. norvegicus	**18.7**±7.13*	**0.439**±0.0507*	**79.6** ∈ [54.2 ; 96.4]	**0.479** ∈ [0.358 ; 0.596]
E. caballus	**23.8**±6.98*	**0.894**±0.0863*	**28.8** ∈ [15.28 ; 43.6]	**0.886** ∈ [0.712 ; 1.13]
C. familiaris	**34.1**±2.60*	**0.341**±0.0402*	**56.6** ∈ [47.4 ; 62.7]	**0.482** ∈ [0.346 ; 0.616]
B. taurus	**34.9**±6.13*	**0.819**±0.0658*	**18.4** ∈ [3.62 ; 29.3]	**0.819** ∈ [0.684 ; 0.954]
O. aries	7.7±8.15	**1.03**±0.0673*	**41** ∈ [27.7 ; 53.1]	**1.03** ∈ [0.892 ; 1.17]
S. scrofa	39.4±20.8	**0.469**±0.16*	22.6 ∈ [-49.0 ; 55.8]	**0.469** ∈ [0.297 ; 0.66]
M. domestica	1.50±16.8	**0.248**±0.0367*	**196** ∈ [133 ; 242]	**0.248** ∈ [0.201 ; 0.294]
G. gallus	**34.1**±3.37*	**2.04**±0.056*	**10.5** ∈ [6.69 ; 14.7]	**2.09** ∈ [1.97 ; 2.21]
D. rerio - entire chromosomes	**59.7**±19.6*	0.592±0.356	-16.3 ∈ [-59.1 ; 26.4]	**0.592** ∈ [0.237 ; 0.948]
D. rerio - metacentric arms	23.1±13.4	**1.23**±0.258*	**20.5** ∈ [12.2 ; 28.8]	**1.19** ∈ [0.892 ; 1.48]
O. latipes	2.16±11.4	**1.91**±0.375*	**27.1** ∈ [23.2 ; 29.3]	**2.57** ∈ [1.48 ; 3.85]

Table III.2: *Values of the parameters for the linear and non-linear models. In the case of the linear model, the standard deviation of each parameter is given and the * stands for parameters which are statistically different from 0. For the non-linear model, the confidence interval of each parameter is shown, through the minimum and maximum values. For* H. sapiens *and* D. rerio*, two sets of data values were used: the "entire chromosomes" data, like for the rest of the species, are made up of entire lengths of chromosomes; the "metacentric arms" data contain also the entire lengths of chromosomes, except for the metacentric, which arms are considered independently.*

is regulated per chromosome arm for large metacentric chromosomes (de Villena and Sapienza, 2001; Fledel-Alon et al., 2009) (chapter I.2.4.1). This procedure is applied to metacentric chromosomes 1, 2, 16, 19, and 20 in *H. sapiens*, and metacentric chromosomes 6 and 7 in *D. rerio*. Centromere physical positions are retrieved from (DB-Ensembl) for *H. sapiens*. These centromere positions were subsequently reported on the genetic maps of the chromosomes and the genetic length per chromosome arm is inferred (Matise et al., 2007). For *D. rerio*, the centromere-linked markers Z6767 and Z20932 for chromosome 6 (Mohideen et al., 2000) and Z3412 for chromosome 11 (Phillips et al., 2006) give the genetic and physical position of centromeres on the MGH genetic map (DB-ZFIN). For both species the predicted values for P_I become different from zero. For *M. mulatta* and *S. scrofa*, the low resolution of the genetic maps does not allow the inference of physical and genetic lengths per chromosome arm.

III.3.1.2 Invertebrate parameter values

We have further tested our model on the data from 14 non-vertebrates, spanning different phylogenetic groups (table III.1). Figure III.5 illustrates the fitting of the linear and non-linear model on these genetic maps. Similar to vertebrates (figure III.4), the non-linear model adjusts very well on the data. However, neither model explains the data from the fruitfly (*D. melanogaster*) and the nematode (*C. elegans*) (figure III.5). Only the females of *D. melanogaster* undergo recombination (reviewed in Zickler (2006)). The recombination landscape of *C. elegans* is singular, as all its five chromosomes have one and only one CO, suggesting that complete interference is present (Hammarlund et al., 2005). The particularities of these species together with their reduced number of chromosomes lead to a poor fitting of the two models studied here (figure III.5).

Except for these two species and *Z. mays*, **the r *parameter* is very high in non-vertebrates compared to vertebrates**, ranging from 2.04 cM/Mb in *V. vinifera* to 286 cM/Mb in *S. cerevisiae* (table III.3). The chromosomes of these species have small physical lengths, from a few ten Mb for *A. mellifera*, plants, and *C. elegans* to an order of hundreds or a few kb for the remaining species (table I.2). Combined with the obligatory CO condition per chromosome, these characteristics result in high COR values. *Z. mays* is the only invertebrate among the 14 in this study that has very long chromosomes, with an average physical length greater than *H. sapiens* (table I.2). It follows that the *r parameter* is much smaller for this species compared with the rest of non-vertebrates (table III.3).

Six species out of the fourteen have P_I values with corresponding CIs containing zero: *D. melanogaster*, *C. elegans*, *S. cerevisiae*, *C. neoformans*, *V. vinifera*, and *Z. mays* (table III.3). The results for *D. melanogaster* and *C. elegans* have been discussed above. However, null estimates of P_I don't necessarily imply that these species lack interference. For both *C. neoformans* and *V. vinifera*, data values of the genetic maps are diffuse, due mainly to a low number of markers, 301 for 14 chromosomes (Marra et al., 2004) and 515 for 19 chromosomes (Doligez et al., 2006), respectively. A higher number of markers, assuring a better coverage of chromosome lengths, could lead to better adjustments of the models. Nevertheless, the density of markers is not an argument for *S. cerevisiae* and *Z. mays* for which COs are known to undergo interference (Mancera et al., 2008; Falque et al., 2009). The failure of the non-linear model to estimate correctly the inter-CO distance is

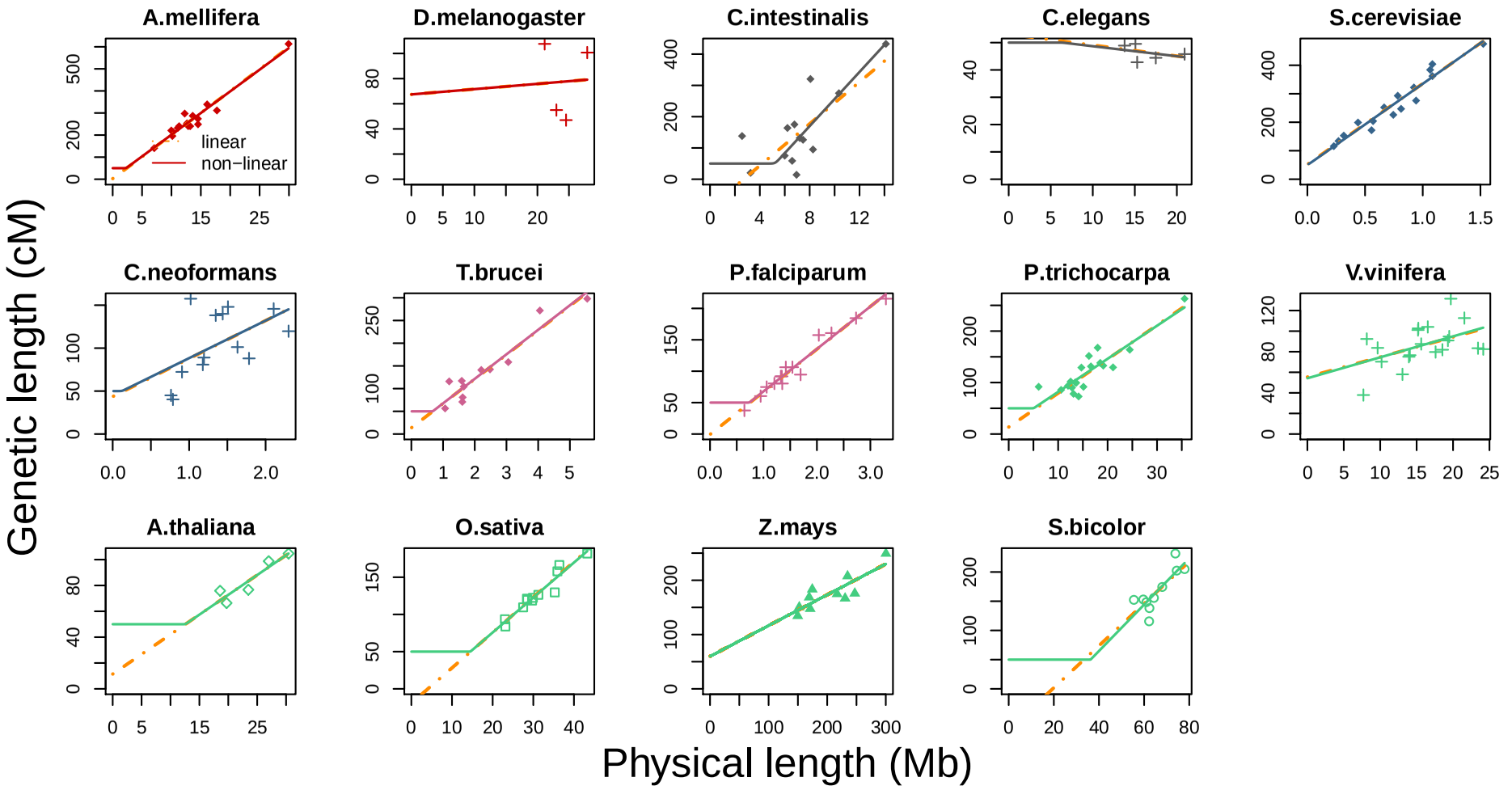

Figure III.5: *Representation of the relationship between the per chromosome physical and sex-averaged genetic lengths for the 14 non-vertebrates investigated. The colors indicate the classification detailed in table III.1. The shape of points for each species is used to distinguish among species inside each phylogenetic group. The full and dashed lines represent the fitting of the non-linear and linear model, respectively.*

Species	Linear model		Non-linear model	
	Intercept	Slope	P_I	r
A. mellifera	5.09±18.7	**19.6**±1.29*	**2.29** ∈ [0.287 ; 3.83]	**19.6** ∈ [16.9 ; 22.4]
D. melanogaster	67.4±186	0.42±7.65	-41.4 ∈ [-1263 ; 1153]	0.42 ∈ [-7.23 ; 8.07]
C. intestinalis	-72.2±65.4	**31.6**±8.46*	**5.23** ∈ [2.01 ; 6.87]	**43.3** ∈ [21.3 ; 66.7]
C. elegans	**52.3**±9.24*	-0.367±0.553	6.4 ∈ [-9.32 ; 22.1]	-0.367 ∈ [-0.92 ; 0.187]
S. cerevisiae	**49.3**±16.3*	**286**±19.8*	0.00231 ∈ [-0.140 ; 0.110]	**286** ∈ [243 ; 328]
C. neoformans	44.7±30.6	43.7±21	0.122 ∈ [-0.524 ; 0.767]	**43.7** ∈ [22.7 ; 64.7]
T. brucei	13.6±15.5	**54**±5.73*	**0.675** ∈ [0.0322 ; 1.11]	**54** ∈ [41 ; 67]
P. falciparum	-1.65±6.77	**68.5**±3.81*	**0.736** ∈ [0.574 ; 0.885]	**67.6** ∈ [58.8 ; 77.1]
P. trichocarpa	17.6±13.5	**6.43**±0.782*	**5.04** ∈ [0.785 ; 9.26]	**6.43** ∈ [4.78 ; 8.77]
V. vinifera	**54.2**±14.8*	**2.04**±0.888*	-2.05 ∈ [-10.2 ; 6.06]	**2.04** ∈ [1.15 ; 2.93]
A. thaliana	11.4±17.4	**3.07**±0.718*	**12.6** ∈ [5.25 ; 16]	**3.07** ∈ [1.89 ; 4.24]
O. sativa	-18.2±13.5	**4.7**±0.428*	**14.5** ∈ [10.1 ; 17.5]	**4.7** ∈ [3.74 ; 5.65]
Z. mays	59.8±26.5	**0.567**±0.126*	-17.3 ∈ [-119 ; 36.8]	**0.567** ∈ [0.390 ; 0.743]
S. bicolor	-94±63.9	**3.97**±0.964*	**36.3** ∈ [27.9 ; 41.7]	**3.97** ∈ [3.11 ; 4.83]

Table III.3: *Values of the parameters for the linear and non-linear models for 14 non-vertebrates. In the case of the linear model, the standard deviation of each parameter is given and the * stands for parameters which are found to be statistically different from 0. For the non-linear model, the confidence interval of each parameter is shown, through the minimum and maximum values. Bold text stands for parameters different from 0 or confidence intervals that exclude 0.*

intrinsic to its own definition. In order to ensure the obligatory CO condition, we force a minimum length of 50 cM per chromosome (equation III.1). However, the distribution of data points in *S. cerevisiae* and *Z. mays* is so that the smallest genetic lengths are 116 and 134.4 cM respectively. The misrepresentation of points close to the 50 cM plateau leads to estimations of the P_I *parameter* that are not significantly different from zero.

III.3.1.3 Examining the interference parameter

Previous studies in a few species with a high resolution of CO events resulted in the estimation of the inter-CO distance (table III.4). For *H. sapiens*, the inter-CO distance was estimated for male and female separately, using the two-pathway model (described in chapter II.2.1) (Fledel-Alon et al., 2009). The parameter of the gamma model (the number of NCOs between consecutive COs) is 6.96 for females and 9.17 for males, with a proportion of non-interfering COs of 0.06 and 0.08 respectively (table III.4). The P_I estimate of our non-linear model is 7.16 Mb (table III.2). It is difficult to compare directly our estimates of the interference strength (*parameter* P_I in Mb) with these values as they are expressed in different units of measure.

Species	Inter-CO distance	% non-interfering COs	References
H. sapiens **- ♀**	6.96 NCOs	6	(Fledel-Alon et al., 2009)
H. sapiens **- ♂**	9.17 NCOs	8	(Fledel-Alon et al., 2009)
M. musculus **- sex-average**	70% of SC	10	(Froenicke et al., 2002; de Boer et al., 2006)
M. musculus **- ♀**	102 Mb	10	(Petkov et al., 2007; de Boer et al., 2006)
M. musculus **- ♂**	122 Mb	10	(Petkov et al., 2007; de Boer et al., 2006)
C. familiaris **- ♂**	60% of SC (first 7 chromosomes)	NA	(Basheva et al., 2008)
A. mellifera	2.70 NCOs	NA	(Solignac et al., 2007)
S. cerevisiae	0.072 Mb	30	(Mézard et al., 2007; Mancera et al., 2008)
A. thaliana **- sex-average**	44.1 cM	15	(Drouaud et al., 2006; Mézard et al., 2007)
A. thaliana **- ♀**	~33 cM	NA	(Drouaud et al., 2007)
A. thaliana **- ♂**	~44.8 cM	NA	(Drouaud et al., 2007)
Z. mays	~6 NCOs	~15	(Falque et al., 2009)
S. bicolor	<50 cM	NA	(Bowers et al., 2003)

Table III.4: *The experimental values of the interference distance in different species. This distance is expressed in different units: cM, % of SC, Mb, number of NCOs predicted by the two-pathway model (chapter II.2.1). When available, the proportion of non-interfering COs is also provided. NA stands for Not Available.*

While *M. mulatta* has a similar karyotype to *H. sapiens*, involving only 4 major inter-

chromosomal rearrangements (fusions and fissions) (Rhesus Macaque Genome Sequencing and Analysis Consortium, 2007), the predicted r value (0.63 cM/Mb) is smaller (although not significantly different) than the one for *H. sapiens* (0.9 cM/Mb) (table III.2). This result is consistent with the cytological observation of a smaller number of MLH1 foci in this species, for the male sex (Hassold et al., 2009). Estimates of the P_I parameter suggest that this reduction in the number of COs, could be explained by an increased interference length in *M. mulatta* (52.3 Mb) compared to *H. sapiens* (7.16 Mb) (table III.2).

For *M. musculus* the fitting of the non-linear model, yields an estimate of the interference parameter of 80.4 Mb (table III.2). At the μm scale, the interference length has been predicted to represent 70 % of the total *M. musculus* SC length (Froenicke et al., 2002) (table III.4). Relating this value to the average physical chromosome length, 129.6 Mb (table I.2), results in a mean interference length of 90.72 Mb. Moreover, 10% of the COs in *M. musculus* are not subject to interference (de Boer et al., 2006) (table III.4). Thus, our predicted value of P_I is in accordance with the experimental estimates.

The first 7 chromosomes of male *C. familiaris* are characterized by a mean inter-COs distance of 60% of their SC length (Basheva et al., 2008) (table III.4). Given an average physical length of these chromosomes of 93.83 Mb (table I.2), the equivalent of this inter-CO distance would be 56.3 Mb, which is very close to the estimated value of 56.6 Mb for the interference parameter P_I (table III.2).

The high resolution data of *S. cerevisiae* has allowed the detailed investigation of both CO and NCO distributions (Mancera et al., 2008). Thus, the mean distance between consecutive COs is 71.8 kb (table III.4). Our non-linear model evaluates the *parameter* P_I at 2.3 kb (table III.3). However, as discussed above, our model can't estimate correctly this parameter. This is also the case for *Z. mays*.

We have access to the experimental estimates of inter-CO distances for three other species: *A. mellifera*, *S. bicolor*, and *A. thaliana* (table III.4). In *A. mellifera*, the fitting of a gamma distribution (counting model (Foss et al., 1993)) on the distances between double COs yielded an interference length of 2.7 NCOs (table III.4). We infer a P_I value of 2.29 Mb (table III.3). For *S. bicolor*, a depletion of double COs has been observed for intervals less than 50 cM (table III.4), which represents the same magnitude as the interference value of 36.3 Mb estimated by our non-linear model (table III.3). The interference strength has been estimated at a high resolution on the chromosome 4 of *A. thaliana*, both for the sex-averaged and the sex-specific genetic maps (Drouaud et al., 2006, 2007). On this particular chromosome, the predicted values of the interference strength is approximately 44 cM (table III.4). Our non-linear model results in an estimate of inter-CO distance of 12.6 Mb (table III.3). It is difficult to compare these values directly, as one is an estimate for one chromosome, expressed in cM, while the other is an average value per all chromosomes, expressed in Mb.

The estimations of the interference *parameter* P_I for both vertebrates and non-vertebrates are directly proportional to the physical length of chromosomes (tables III.2, III.3, and I.2). We have modeled this relation in figure III.6. Indeed, we observe a very strong correlation ($R^2 = 0.89$) between these variables (figure III.6). This result implies that **the interference phenomenon is intimately linked to the physical length of chromosomes**. Furthermore, the equation of the regression in figure III.6 is:

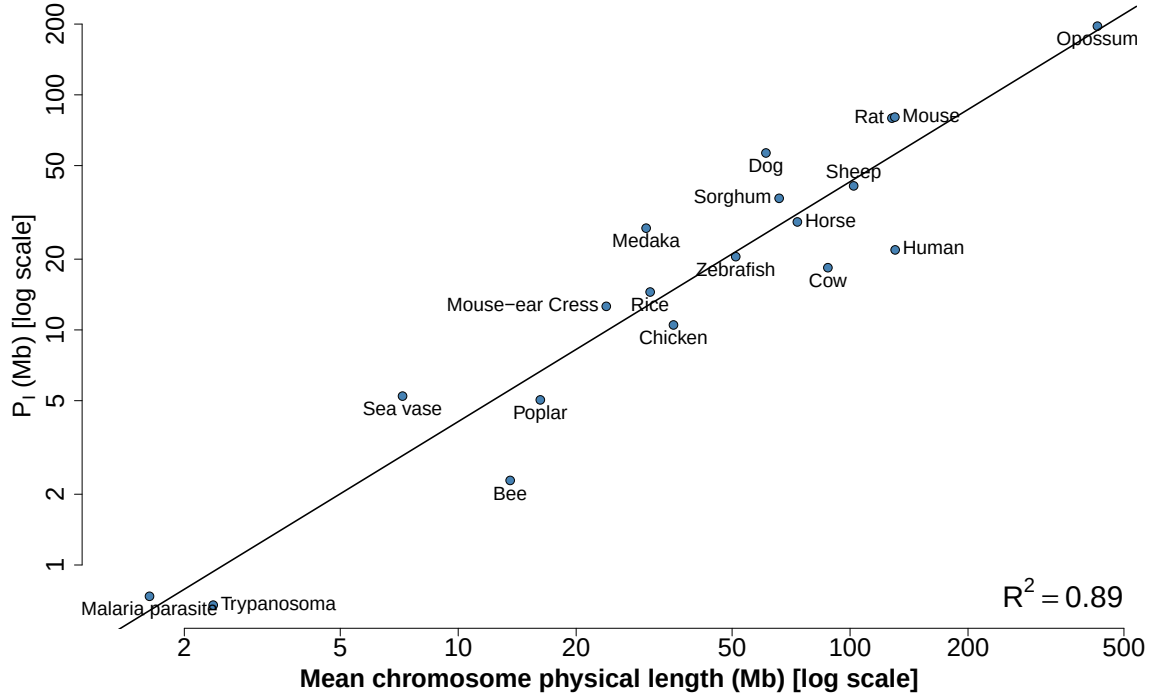

Figure III.6: *Relation between the non-null values of the interference* parameter P_I *of our non-linear model (tables III.2 and III.3) and the mean physical length of chromosomes (table I.2). Both axes are in log scale. The* R^2 *coefficient of this regression is also provided.*

$$P_I = 0.80877 + 0.45303 \times \overline{P} \tag{III.5}$$

where $\overline{P}$ represents the mean physical length of all the chromosomes of a species (Mb). **Equation III.5 represents a new tool for inferring the interference strength even for species for which genetic maps are not available yet**.

III.3.1.4 Resemblance among species

In order to understand the mutual evolution of recombination and karyotype, it is important to identify species with similar recombination patterns. For this purpose, **we analyze the variability of the parameter** r**, which represents the rate of additional CO production**. In figure III.7, we observe that this variability is greatly explained ($R^2 = 0.91$) by the total genome size (Mb) of species. A similar negative trend between the total CO rate (defined as the ratio between genetic and physical lengths of chromosomes) and the genome size was previously reported (Lynch, 2006). This trend is mainly explained by the lack of correlation between the number of chromosomes and the size of the genome (Lynch, 2006). Coupled with a limited number of COs per chromosome and the obligatory CO, it results that species with a smaller genome size will have a higher CO rate. In agreement with this explanation, the relation in figure III.7 shows that the rate of CO additional to the obligatory CO (r) is also negatively correlated with the total size of the genomes. Thus, **mechanisms acting on the physical scales of genomes are strong**

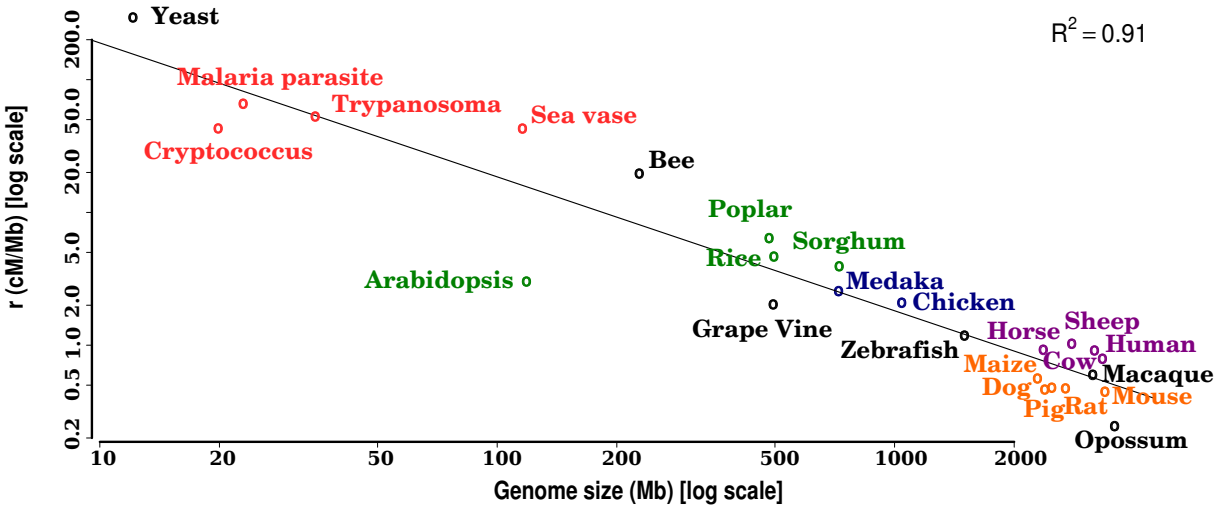

Figure III.7: *Relation between the non-null values of the additional CO rate* parameter *r of our non-linear model (tables III.2 and III.3) and the total genome size (table I.2). Both axes are in log scale. The* R^2 *coefficient of this regression is also provided. Species are colored according to similar r values. Black color is used for species that are not part of any group. The similarities inside a group result from an overlap between the CIs of* parameter *r between all species.*

determinants of recombination rates.

Moreover as for the interference parameter (P_I), the relation in figure III.7 is defined by the equation III.6:

$$ln(r) = 3.2854 - 1.0087 \times ln(\sum P) \quad \text{(III.6)}$$

We further infer the way species cluster according to the estimate of the parameter r. The overlap between CIs in tables III.2 and III.3 indicates a similarity between species. Species with similar r values are grouped in figure III.7 by the same color. Species colored in black are not part of any group. For simplicity reasons, we emphasize only large groups of similarity and ignore those species that overlap only partially with a group. For example, the CIs of *parameter r* of *A. mellifera* and *C. intestinalis* overlap. However, *A. mellifera* is not similar, according to this criteria, to the other species in the group of *C. intestinalis* (table III.3 and figure III.7). Another example is *V. vinifera*, for which the r value overlaps with *G. gallus* and *O. latipes*, and *A. thaliana*, but these three species are part of two distinct groups.

Some species were excluded from the comparison. This is the case of *M. mulatta* which has a large CI overlapping all the other eutherian species (table III.2). We also omit *D. rerio* as it overlaps multiple groups at a time. Two other species, *D. melanogaster* and *C. elegans* are excluded as the models don't fit the data.

Five major groups emerge (figure III.7). A first group is formed by species with small genomes *T. brucei*, *P. falciparum*, *C. neoformans*, and *C. intestinalis* but very high CO rates ($r > 43$ cM/Mb) (figure III.7). A second cluster emerges, which represents the plants: *O. sativa*, and *S. bicolor*, with which overlap *A. thaliana* and *P. trichocarpa* (figure III.7). However, *A. thaliana* is also similar to *V. vinifera*, *G. gallus*, and *O. latipes*. The group of

G. gallus and *O. latipes* is characterized by high values of r (> 1.5 cM/Mb) compared to the rest of the vertebrates (table III.2). At the other extreme, the group of *Z. mays*, *M. musculus*, *R. norvegicus*, *C. familiaris*, and *S. scrofa* (figure III.7) is distinguished by small values of r (< 0.5 cM/Mb) (table III.2). In between, the cluster of *H. sapiens*, *E. caballus*, *B. taurus*, and *O. aries* (figure III.7) yields values of r close to unity (table III.2).

Phylogeny seems a powerful tool to explain the clustering of species according to similarities of the r parameter. Where phylogeny fails to explain such clusters, another important factor, the size of the genome, succeeds (figure III.7). For example, *Z. mays*, a plant, has a genome size comparable to vertebrates and a recombination rate closer to the group of *M. musculus*. However, there are groups that neither of these factors can predict. *H. sapiens* has a very similar genome size to *M. musculus* and is closer phylogenetically to this species than with the rest of its group. Nevertheless *H. sapiens* has a r value (~ 1) two times higher than *M. musculus* (table III.2). The exploration of additional factors, such as for example the rate of rearrangement, could yield new insights into the process of recombination.

III.3.2 Heterochiasmy in vertebrates

III.3.2.1 Parameter values

Heterochiasmy, the difference in recombination rates and distribution between sexes, is a widely spread phenomenon, as detailed in chapter I.2.4.2. **Our model**, by quantifying both the rate of additional COs and interference strength, **allows the comparison and quantification of sex-differences**.

We have fitted the linear and non-linear models on the sex-specific genetic maps of 10 vertebrates (figure III.8). We observe important sex-specific differences in the strength of interference and the rate of additional COs (figure III.8). However, for two species out of the ten, the model doesn't adjust well on the data: *S. scrofa* and *O. latipes* (figure III.8). The sex-specific genetic maps of these species have a low density of markers ($\sim$200), unequally distributed along the chromosomes (Kimura et al., 2005; Vingborg et al., 2009), which could lead to low-quality data. However, there could also be additional recombination mechanisms specific to each of these species that are not integrated in our model. We do not further discuss the results in these species. Another species for which we can't compare heterochiasmy is *D. rerio*. While an updated high density female (Heat Shock) genetic map exists (DB-ZFIN), the male genetic map is very poor in quality (Singer et al., 2002).

The predicted values of the corresponding parameters, together with their precision, are reported in table III.5. For *H. sapiens*, the interference parameter P_I of the non-linear model estimated on the total length of chromosomes is null for the two sexes. By replacing metacentric chromosomes with their corresponding lengths per chromosome arm, P_I values are different from zero. In *H. sapiens*, our estimates of the female and male interference parameter (19.2 and 25.6 Mb, respectively) (table III.5), show the same tendency as it was previously reported for this species (6.96 and 9.17 NCOs, respectively) (table III.4).

There is a good correspondence between predicted values of P_I (table III.5) and experimentally inferred values of the interference strength (table III.4). However, while the trend (female, male) and order of magnitude is the same between our estimates and

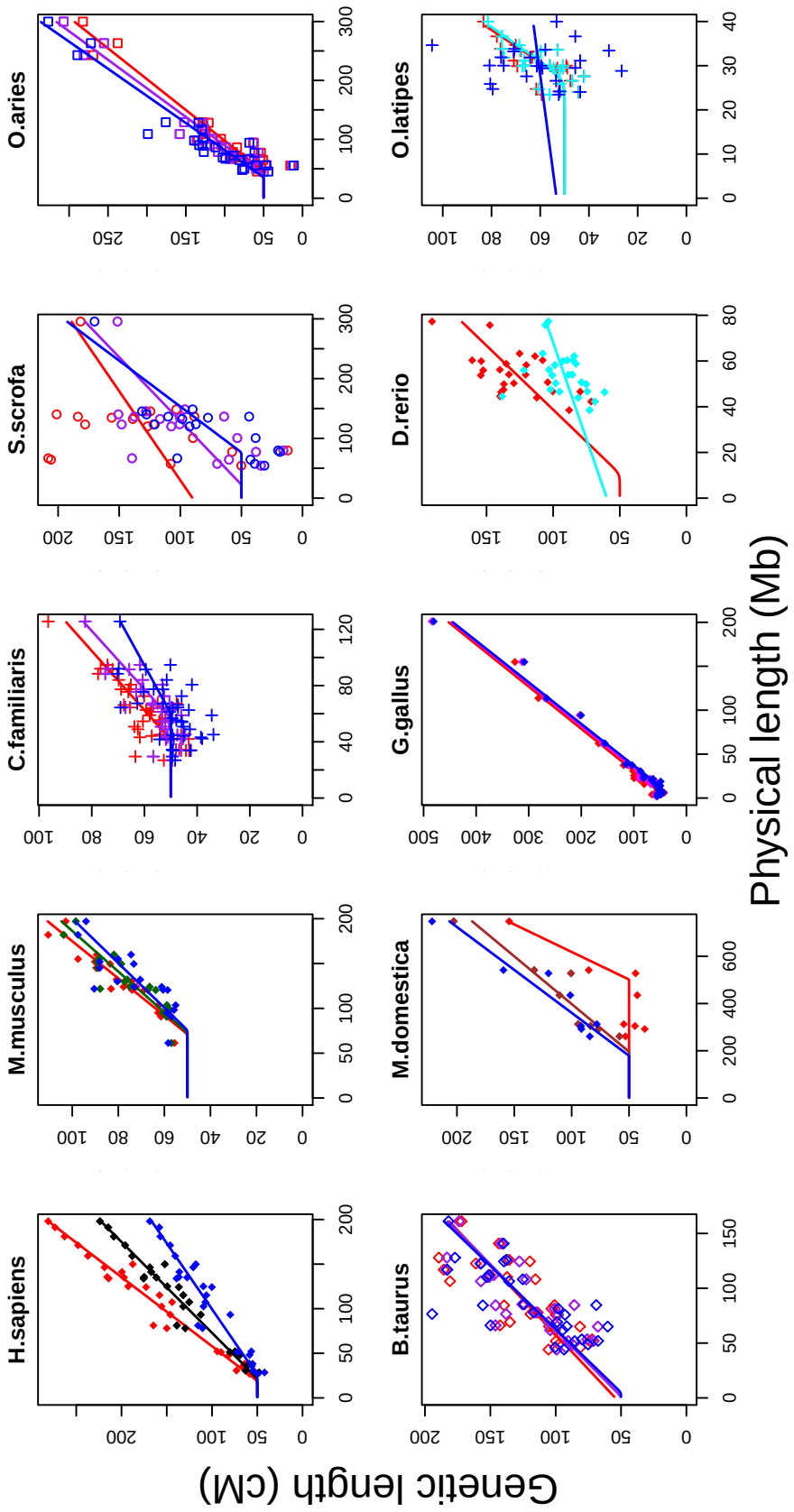

Figure III.8: *Representation of the relationship between the physical and genetic lengths for both sex-specific and sex-averaged maps of 10 vertebrates investigated. These lengths relate to the entire chromosomes, except for the metacentric chromosomes of* H. sapiens*, for which they involve the chromosome arms. Colors of the sex-averaged data points are indicative of the classification detailed in section III.2.3. Female and male data points are represented by the colors red and blue respectively. The shape of points for each species is used to distinguish among species inside each phylogenetic group. The lines represent the fitting of the non-linear model.*

Species	Linear model		Non-linear model	
	Intercept	Slope	P_I	r
H. sapiens - ♀	**25.2**±5.88*	**1.29**±0.0495*	**19.2** ∈ [10.5 ; 26.8]	**1.29** ∈ [1.19 ; 1.39]
H. sapiens - ♂	**32.6**±4.97*	**0.678**±0.0418*	**25.6** ∈ [11.9 ; 37.7]	**0.677** ∈ [0.591 ; 0.766]
M. musculus - ♀	**20.9**±4.96*	**0.446**±0.0372*	**71.2** ∈ [54.7 ; 81.7]	**0.482** ∈ [0.389 ; 0.568]
M. musculus - ♂	**26.8**±8.44*	**0.352**±0.0632*	**75.7** ∈ [32.9 ; 94.7]	**0.402** ∈ [0.237 ; 0.555]
C. familiaris - ♀	**33.8**±3.10*	**0.428**±0.0481	**41.4** ∈ [33.2 ; 52.3]	**0.475** ∈ [0.366 ; 0.651]
C. familiaris - ♂	**34.8**±3.73*	**0.252**±0.0579*	**60.2** ∈ [43.1 ; 83.4]	**0.294** ∈ [0.117 ; 0.706]
S. scrofa - ♀	**90**±32*	0.335±0.246	-119±-298	0.335±0.0893
S. scrofa - ♂	8.43±16.9	**0.604**±0.13*	**76.5** ∈ [15.6 ; 102]	**0.651** ∈ [0.327 ; 0.944]
O. aries - ♀	7.27±7.45	**0.953**±0.0616*	**45.2** ∈ [32.1 ; 58]	**0.956** ∈ [0.826 ; 1.1]
O. aries - ♂	11.3±10	**1.09**±0.0827*	**35.6** ∈ [19.1 ; 48.9]	**1.09** ∈ [0.916 ; 1.26]
B. taurus - ♀	**54.4**±11.9*	**0.79**±0.128*	-5.53 ∈ [-53.6 ; 19.4]	**0.79** ∈ [0.528 ; 1.05]
B. taurus - ♂	**45.7**±15.1*	**0.866**±0.162*	4.99 ∈ [-48.8 ; 30.1]	**0.866** ∈ [0.534 ; 1.2]
M. domestica - ♀	-17.3±24.6	**0.191**±0.054*	**502** ∈ [466 ; 528]	**0.429** ∈ [0.339 ; 0.519]
M. domestica - ♂	0.837±17.3	**0.275**±0.038*	**179** ∈ [125 ; 220]	**0.275** ∈ [0.232 ; 0.318]
G. gallus - ♀	**38.9**±3.19*	**2.05**±0.0531*	**6.77** ∈ [3.17 ; 10.4]	**2.08** ∈ [1.97 ; 2.20]
G. gallus - ♂	**30.9**±3.51*	**2.03**±0.0584*	**13.4** ∈ [9.2 ; 17.8]	**2.11** ∈ [1.99 ; 2.24]
D. rerio - ♀	30.9±25.6	**1.78**±0.464*	10.7 ∈ [-40.7 ; 26.4]	**1.78** ∈ [0.824 ; 2.74]
O. latipes - ♀	1.86±12.6	**1.95**±0.412*	**27.1** ∈ [22.8 ; 29.4]	**2.73** ∈ [1.5 ; 4.16]
O. latipes - ♂	53.3±27.2	0.241±0.895	0.241±-0.653	-13.6 ± -177

Table III.5: *Values of the parameters for the linear and non-linear models. In the case of the linear model, the standard deviation of each parameter is given and the stands for parameters which are found to be statistically different from 0. For the non-linear model, the confidence interval of each parameter is shown, through the minimum and maximum values. Bold text stands for parameters different from 0 or confidence intervals that exclude 0. For* H. sapiens *the parameter values in this table are estimated on the entire lengths of chromosomes, except for the metacentric chromosomes, whose arms are considered independently.*

the experimental approach, some differences are apparent. It should be emphasized that P_I values are expressed in Mb, and correspond to average values for all chromosomes. The values in table III.4 are expressed either in number of NCOs (*H. sapiens*), Mb (*M. musculus*), or % of SC (*C. familiaris* male). Moreover, these values are estimated only on chromosome 1 for *M. musculus*, and on the first 7 chromosomes for *C. familiaris*.

III.3.2.2 Comparing male and female

The vertebrates that we analyze here show different heterochiasmy trends (table III.6). The females of *H. sapiens*, *M. musculus*, and *C. familiaris* have more COs than the males (table I.3). The reverse heterochiasmy is seen in *O. aries* and *M. domestica* table I.3. Groenen et al. (2009) observed no heterochiasmy in *G. gallus* when all chromosomes (autosomes and sex chromosomes) as well as linkage groups were considered. However, when we consider the total genetic lengths of only autosomes in table I.3, the ratio female to male is 1.09 and is equal to that observed in *M. musculus*. *B. taurus* has comparable linkage maps in males and females (Barendse et al., 1997).

The general rule when comparing P_I values between sexes is that the sex with the smallest P_I has more COs. This rule is logical, as a smaller interference length leads to a more equal distribution of COs along the chromosome and subsequently an increase in their number. The prediction of the interference parameter P_I is highly variable leading to large CIs (table III.6). Except for the metatherian *M. domestica*, the corresponding CIs of P_I overlap between the sexes. Though not significant, the difference observed for this parameter between males and females is a predictor of the heterochiasmy relations previously reported (table III.6). The only species for which the CIs of P_I don't overlap is *M. domestica* (table III.5). The average interference length in female, P_I♀, is more than 2.5 fold greater than the corresponding male value (P_I♂). This observation is in agreement with the genetic and cytological observations that chiasma are mainly localized close to telomeres for female *M. domestica*, as opposed to a more uniform distribution along chromosomes in male (Sharp and Hayman, 1988; Samollow et al., 2007). Moreover, at the chromosome level, all chromosomes have an expected male CO number that exceeds the one obligatory CO per chromosome (figure III.8). For female, the P_I♀ is so strong, that only the last, longest chromosomes have an excess of COs (figure III.8).

When comparing the r coefficients between female and male, different trends emerge (table III.6). As for the P_I parameter, the CIs of the r coefficient overlap between sexes for the majority of species (table III.5). Two species, however, show distinct differences among female and male r values, *H. sapiens* and *M. domestica*. The general trend for *H. sapiens*, *M. musculus*, *C. familiaris*, and *O. aries* is that **the sex with the smallest P_I, and subsequently, the highest CO number, has a higher r as well** (table III.6). This result is consistent with higher r values indicating a higher density of COs per Mb (Kong et al., 2002; Maddox and Cockett, 2007; Matise et al., 2007; Kong et al., 2010). However, this trend is not true for *M. domestica* or *G. gallus* (table III.6).

Similar to *O. aries*, the total genetic length of female linkage map is smaller than the male's for *M. domestica* (Samollow et al., 2007). Thus, when estimating sex-specific CORs as the ratio between genetic and physical length of the linkage map of *M. domestica*, the male COR was found to be higher than the female COR (Samollow et al., 2007).

Species	r♀ r♂	P_I♀ P_I♂	Nbr CO♀ Nbr CO♂
H. sapiens	$>*$	$<$	$>$
M. musculus	$>$	$<$	$>$
C. familiaris	$>$	$<$	$>$
O. aries	$<$	$>$	$<$
B. taurus	$<$	**0**	$=$
M. domestica	$>*$	$>*$	$<$
G. gallus	$<$	$<$	$\geq$

Table III.6: *Table of the relations between male and female estimates of the parameters r and P_I, as well as the total number of COs for all the autosomes (no sex-chromosomes). A star (*) stands for female and male CIs not overlapping. The other inequalities between sex-specific estimates reflect tendencies as their corresponding CIs overlap. The differences in the total number of COs between sexes are obtained from litterature as represented in table I.3. Red: the female values are greater than the male values. Blue: the opposite pattern is observed. 0: parameter values not significantly different from 0. =: no difference. The $\geq$ relation between total number of COs in female and male* G. gallus *is based on the values in table I.3, for which the ratio F/M is the same as for* M. musculus*. However, the main bibliographic resource for* G. gallus *reports the lack of heterochiasmy for this species when all chromosomes, autosomes and sex chromosomes, are considered, hence the equality sign (Groenen et al., 2009).*

Intriguingly, when comparing the estimated r values, we find that $r♀ > r♂$ (table III.6). It should be noted, however, that the estimation of the r parameter in female is performed only on the last two out of the eight chromosomes, that have more than 1 CO (figure III.8). Moreover, important information is lacking for this species on the COR in the vicinity of telomeres, and these regions are known to host more COs in female than male (Sharp and Hayman, 1988). Our observations on the P_I values in this species show that the majority of chromosomes in female have only one obligatory CO as opposed to males (figure III.8), due to a much longer interference distance in this sex (table III.5). Our results suggest that once additional COs start to accumulate in female, they do so at a rate much higher than in males (r values represent the rate with which additional COs accumulate).

G. gallus was previously reported to lack heterochiasmy (Groenen et al., 2009). However, we find that the P_I estimate is ∼2-fold larger in male than female, while the r values are very close (table III.5). A smaller P_I *parameter* in females implies that this sex has more CO events than the male. Indeed, when we compared the sex-specific genetic maps from Groenen et al. (2009) only for autosomes (no sex-chromosomes), the female to male ratio we obtained was similar to that of *M. musculus* (table I.3). **Our results suggest that *G. gallus* is also subject to heterochiasmy**, with the female having more COs than the male (table III.6). A recent linkage study in two populations of *G. gallus*, with an increased number of markers, has revealed a new perspective on the heterochiasmy in this species (Elferink et al., 2010). In these two populations, the number of COs was higher for males than for females. When we fit our model on this new data, we obtain the following parameter values: $r♀ = 1.62 \in [1.53; 1.7]$, $r♂ = 1.98 \in [1.89; 2.07]$,

$P_I♀ = 8.33 \in [4.67; 11.9]$, and $P_I♂ = 10.3 \in [7.1; 13.6]$. The difference between sexes in their P_I value shows the same trend $P_I♀ < P_I♂$. This trend is opposed to the previous hypothesis that smaller interference length is indicative of higher CO number. However, the difference between the sex-specific interference lengths is much smaller on this new data. Moreover, with the new dataset we estimate that the accumulation of additional COs is also significantly slower in female than male ($r♀ < r♂$). **Our work on *G. gallus* rises new questions on the existence and the sens of heterochiasmy in this species.**

III.4 Conclusion

In this chapter, we have developed a **non-linear model linking the number of COs to the karyotype structure of species, including the obligatory CO per chromosome condition**. The two parameters of the model can be interpreted biologically: the average rate of supplementary CO per Mb (following the obligatory CO) (r) and the average physical inter-CO distance (or interference parameter) (P_I). The interference mechanism describes the zone of inhibition imposed by a CO for the formation of subsequent COs. When interference is relatively strong, small chromosomes will experiment only one CO and thus, form a plateau at 50 cM. As larger chromosomes are analyzed, their relation to the number of chromosomes becomes linear. This phenomenon is well observed in species such as *Canis familiaris* and *Gallus gallus*. Due to its formula, our model characterizes simultaneously both the plateau and the linear behavior of chromosomes.

We have fitted the model on data from 13 vertebrates and 14 non-vertebrates. Based on these 27 datasets we quantified both the rate of CO production and the inter-CO distance. Chromosome lengths are relatively high in vertebrates, from a few ten to a few hundred Mb leading to long range interference strength and small *r parameter* estimates. For the 14 non-vertebrates, the chromosome lengths are small, a few Mbs, resulting in small inter-CO distances but very high CORs compared to vertebrates. Previously, two methods were used to compute the COR: the ratio between genetic and physical length or the slope of a linear model fitting the data (Li and Freudenberg, 2009). The parameter r is similar to the slope of the linear model for the majority of species. However, when the physical chromosome lengths span both the plateau and the linear behavior of genetic maps (*C. familiaris*, *G. gallus*, female *M. domestica*), our model is a better predictor. We have also compared the inferred P_I value to experimental estimates of inter-CO distances (table III.4). The heterogeneity of these experiments, the resolution of the data, the number of chromosomes involved, as well as the unit of measure complicate such comparisons. Nevertheless, there are strong similarities in magnitude between our average prediction of interference per species, and the experimental analysis of inter-CO length based on double COs. Our estimates of interference strength, even for species previously lacking such information, are indicative of the distribution of CO events along chromosomes. These estimates could represent the starting point when defining the resolution of the linkage analysis that will allow, for example, the observation of double COs. Moreover, we show that **there is a strong relation between the average physical length of chromosomes and the P_I *parameter***. This relation could represent a good estimate of the interference strength even for species for which genetic maps are not available.

The use of the r parameter, as an indicator of the average COR, has also allowed us to compare species and group those with similar r values. We find a strong negative correlation between r and the total size of the genomes (figure III.7). This relation, in agreement with previous observations, indicates that the regulation of CO rate is highly influenced at the Mb scale. As for the interference parameter, the relation in equation III.6 could be used as a tool to estimate parameter r even for species without an available genetic map. Phylogeny and genome size emerge as important factors for the similarity in CO rates among species (figure III.7). Based on the r parameter we find five clusters of species with similar values (figure III.7). However, neither phylogeny, nor genome size can explain the two clusters of vertebrates: the one containing *H. sapiens* and the other containing *M. musculus*. Given the high rate of rearrangements characterizing *M. musculus*, *R. norvegicus*, and *C. familiaris*, as compared to the rest of eutherians (Wienberg, 2004; Murphy et al., 2005; Kemkemer et al., 2009), it could indicate a higher genome instability, affecting recombination hotspots and leading to a subsequent decrease in the COR. Such clusters of species, with similar CO rates, indicates similar karyotype structure and dynamics. The in-depth study of the karyotype evolution of species inside each group could further lead to the identification of recombination mechanisms shared by species inside each cluster. Moreover, the analysis of extremely divergent species would help understand the evolution of differences in the recombination mechanism.

In order to better understand the mechanisms of heterochiasmy, we have compared the parameters values estimated in female and male in 10 vertebrates (tables III.5 and III.6). For *H. sapiens*, *M. musculus*, and *C. familiaris*, the female has more COs than the male (Kong et al., 2002; Cox et al., 2009; Wong et al., 2010). This is mainly a cause of sex-differential interference strength, as we estimate that $P_I♀ < P_I♂$. Our quantification of P_I in the two sexes, in agreement with previous studies, suggests that the shorter interference length in female, leads to a more uniform distribution of COs, thus, generating more such events in this sex. As expected, we estimate that $r♀ > r♂$. The reverse pattern is seen in *O. aries*, for which the male sex has more COs than the female (Maddox and Cockett, 2007): COs are more equally distributed in male, leading to an increase in the CO number and per Mb rate in this sex (table III.6). Like *O. aries*, *M. domestica* has more COs in male than in female (Samollow et al., 2007), resulting from a smaller P_I in this sex. However, the per Mb COR shows an unexpected trend: $r♀ > r♂$. A possible explanation for this result would be that despite the majority of chromosomes in female *M. domestica* exhibiting only one CO, the rate at which additional COs are produced, once the interference barrier is overcome, is stronger in female. The present genetic map of *M. domestica* is built with only 150 markers, with a misrepresentation of regions close to telomeres (Samollow et al., 2007). These regions are also known to host COs, especially for female (Sharp and Hayman, 1988) and are expected to increase the genetic length of chromosomes for this sex. Additional data would help understand better the heterochiasmy in this species. The last vertebrate for which we analyzed the inter-sexes differences in recombination is *G. gallus*. Our results, as well as recent genetic maps, raise new questions about the heterochiasmy in this species.

Chapter IV

Sex-specific impact of recombination on the nucleotide composition

What is the relation between sex, recombination and nucleotide composition? In order to understand how the GC biased gene conversion is influenced by heterochiasmy, I analyze the interactions between these variables in different vertebrates.

IV.1 Introduction

In taxa ranging from yeast to mammals, the within-genome variation in the intensity of recombination events appears strongly correlated with regional differences in the GC-content of sequences (Gerton et al., 2000; Fullerton et al., 2001; Birdsell, 2002; Marais et al., 2001; Kong et al., 2002; Meunier and Duret, 2004; Webster et al., 2005; Duret and Arndt, 2008; Berglund et al., 2009; Backström et al., 2010). However, the two variables (crossover rate (COR) and GC-content) evolve at different time-scales. For example, the recombination hotspots are not conserved between closely related species such as human and chimpanzee (Ptak et al., 2005; Winckler et al., 2005) (chapter I), while the nucleotide sequence divergence between the two genomes is only of 1.1 % (Chimpanzee Sequencing and Analysis Consortium, 2005). This has led to the introduction of a new variable that characterizes the nucleotide composition, the equilibrium GC-content (GC*) (Meunier and Duret, 2004). GC* is the GC-content a sequence would reach if it evolved for an infinite period of time under a constant substitution pattern, and it correlates strongly with COR in human (Meunier and Duret, 2004; Webster et al., 2005; Duret and Arndt, 2008). Since, in human, the correlation between GC/COR is weaker than the one between GC*/COR (Duret and Arndt, 2008), it has been suggested that **recombination, through the GC biased gene conversion mechanism (gBGC), promotes the increase in GC-content by favoring the fixation of AT→GC substitutions** (Eyre-Walker and Hurst, 2001) (figure IV.1 1.). While this is the preferred causality model for the human genome, in yeast, no evidence of a correlation between COR and the substitution pattern has been detected (Marsolier-Kergoat and Yeramian, 2009). This result in yeast is in agreement with the interpretation that GC-rich sequences might act as hotspots of recombination by facilitating a chromatin structure with a high affinity to the recombination machinery (Gerton et al., 2000; Petes, 2001; Blat et al., 2002; Petes and Merker, 2002; Marsolier-

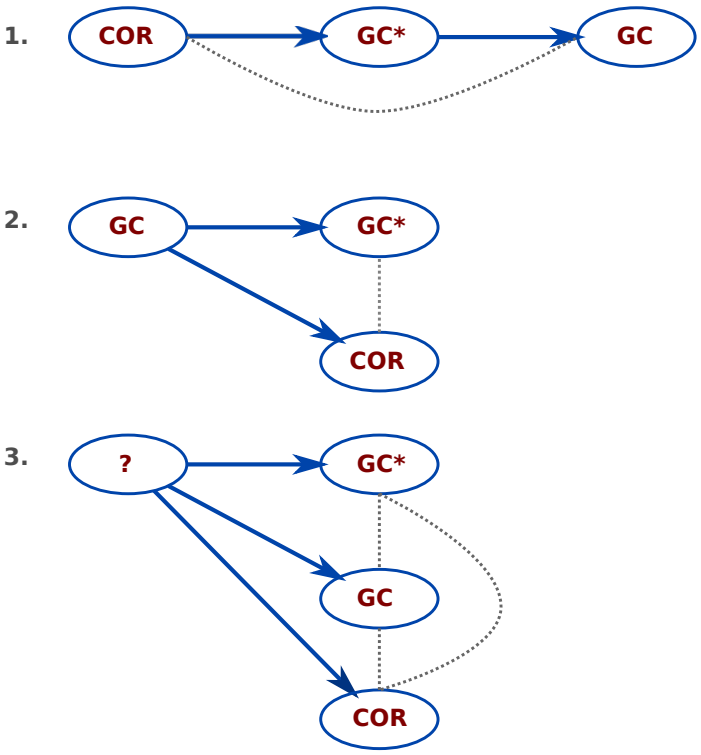

Figure IV.1: *Schematic representation of the relation between the three variables: current GC-content (GC), equilibrium GC-content (GC*), and crossover rate (COR). The interrogation mark (?) stands for an yet unidentified forth variable. Blue solid arrows represent main causality relations implied by the different hypotheses, while dashed gray lines stand for secondary weaker relations between variables. 1. COR influences the substitution pattern, and thus GC*, through gBGC, which in turn leads to the modification of GC. 2. GC-content impacts on the GC* by controlling for the DNA melting temperature, and thus the mutation rate, and in parallel, it controls the chromatin structure, and thus the accessibility of the recombination machinery and subsequently the COR. 3. A forth unknown parameter is the main driver of all three variables.*

Kergoat and Yeramian, 2009). Moreover, a high GC-content is indicative of a high DNA melting temperature (a predominant double-strand state of the DNA), which in turn is rate-limiting for the cytosine deamination - the main source of $C \rightarrow T$ mutations - thus directly influencing the GC* (Fryxell and Zuckerkandl, 2000) (figure IV.1 2.). However, these results should be interpreted with caution, as the resolution of the analysis is different between these studies (1 Mb windows in human and a few kb in yeast). The fact that the substitution patterns are not related to the recombination rate in yeast might also be caused by a reduction in the efficiency of gBGC in this species generated by a low frequency of sexual reproduction (Tsai et al., 2010). These causality models are not mutually exclusive, and there might also be still unidentified factors affecting both recombination and GC-content (figure IV.1 3.).

Whatever the causality between these two genomic variables, it seems to be sex-specific. As discussed in chapter I.2.4.2, there are considerable differences in COR among sexes, both at the global (total number of COs) and local level (distribution along chromosomes)

(heterochiasmy). At first, it was expected that since the female sex has more COs than the male in the majority of eutherian mammals, it would correlate better with the GC-content. This result was observed on a dataset of 33 loci in human (Meunier and Duret, 2004). However, the analysis of the substitution patterns in *Alu* elements along all autosomes showed that the **GC* content of a region is more strongly correlated with male rather than female local recombination rate** (Webster et al., 2005). These findings were later confirmed by inferring the substitution pattern in 1Mb windows of non-coding sequences in humans (Duret and Arndt, 2008). Moreover, also in humans, AT→GC biased substitution hotspots have been found to correlate strongly with male instead of female recombination patterns (Dreszer et al., 2007).

Two non-exclusive hypotheses have been proposed to explain why male, rather than female, recombination could have a higher impact on GC content. One explanation is based on a sex-differential strength of gBGC (Duret and Arndt, 2008). For example, a potential factor could be that **the repair mechanisms is more biased in one sex than the other**. This explanation is based on the role of gBGC in counteracting the effects induced by the hypermutability of methylated cytosines (Brown and Jiricny, 1989). Since in male mammals (contrary to females) recombination takes place on a strongly methylated genome (Lees-Murdock and Walsh, 2008; Sasaki and Matsui, 2008), the bias in gBGC could be greater in order to "correct" the more frequent C→T mutations. The second hypothesis supposes that despite a higher COR in females than in males, the **ratio between the number of DSBs resolved to COs versus NCOs is different between the sexes** (Duret and Arndt, 2008). However, known recombination rates reflect only those DSBs that result in successful COs. The conversion tract of a NCO is smaller than for a CO (Mancera et al., 2008), which makes NCOs difficult to detect with the methods usually employed, and thus its role in gBGC is still not known. If NCOs also generate gBGC and if the CO/NCO ratio is far more variable in females, COR could be a weaker estimator of the total recombination rate and subsequently of the gBGC impact in females (Duret and Arndt, 2008).

In this chapter, we address the question of **the differential impact of sex-specific COR on the nucleotide composition landscape of several vertebrate genomes**: three eutherians (human, mouse and dog), a metatherian (opossum), and a bird (chicken). These five vertebrates are subject to different heterochiasmy aspects. For the three eutherian mammals, the female sex has more COs than the male, while the opposite pattern is present in opossum (table I.3). While the F/M ratio of the total number of COs on autosomes in chicken is comparable to the one observed in mouse, this species has been previously considered as lacking heterochiasmy (Groenen et al., 2009). We further investigate this relation from the perspective of chromosomal localization.

The only vertebrate for which the impact of heterochiasmy on the GC-content of sequences was previously analyzed was the human (Webster et al., 2005; Duret and Arndt, 2008). In the study of Duret and Arndt (2008), the authors compute the GC*, in non-coding regions, on the triple alignment between human, chimpanzee, and macaque. At a 1 Mb resolution, they found a strong COR/GC* correlation, which is stronger for male than female ($R^2 = 0.27$ and $R^2 = 0.15$ respectively). However, similar studies in other species are difficult to accomplish, due to the absence of sequenced genomes for closely

related species. This difficulty has been overcome by inferring the substitution pattern in transponsable elements (TE) (Arndt et al., 2003; Webster et al., 2005). TEs represent large percentages of the genomes of different vertebrates (table IV.1). Their number evolves through insertion bursts, and a classification in families and subfamilies has been made in order to distinguish the events that likely arose during different insertion bursts (reviewed in Pace and Feschotte (2007); Cordaux and Batzer (2009)). In recent years, algorithms were developed for the identification of TEs (Smit et al., 2010) as well as for the reconstruction of their corresponding ancestral sequences (Jurka, 2000). Moreover, TEs are considered functionally neutral (Cordaux et al., 2006). These characteristics make them suitable candidates for the study of neutral substitution patterns in vertebrate genomes. For this purpose, a maximum-likelihood model has been developed, which infers the substitution rates between TEs and their consensus sequences, at the same time accounting for the hypermutability of cytosines at CpG sites (Arndt et al., 2003) (section II.3.1). Making use of this technique, we have computed the GC* from the substitution pattern inferred from TEs (LINE and SINE).

	Human	**Mouse**	**Dog**	**Opossum**	**Chicken**
Euchromatic genome size (Mb)	2,880	2,550	2,330	3,475	1,050
Karyotype					
Haploid number	23	20	39	9	33
Autosomal size range (Mb)	47-247	61-197	27-125	258-748	5-201
X(Z) chromosome size (Mb)	155	167	127	76	75
Segmental duplications					
Autosomal (%)	5.2	5.3	2.5	1.7	10.4
Intrachromosomal duplications (%)	46	84	ND	76	ND
Median length between duplications (Mb)	2.2	1.6	0.33	0.18	0.03
X (Z) chromosome (%)	4.1	13	1.7	3.3	NA
Interspersed repeats (%)					
Total	45.5	40.9	35.5	52.2	9.4
LINE/non-LTR retrotransposon	20.0	19.6	18.2	29.2	6.5
SINE	12.6	7.2	10.2	10.4	NA
Endogenous retrovirus	8.1	9.8	3.7	10.6	1.3
DNA transposon	2.8	0.8	1.9	1.7	0.8
GC-content (%)					
Autosomal	40.9	41.8	41.1	37.7	41.5
X (Z) chromosome	39.5	39.2	40.2	40.9	44.8
CpG content (%)					
Autosomal	2.0	1.7	2.2	0.9	2.1
X (Z) chromosome	1.7	1.2	1.9	1.4	NA
Recombination rate (cM/Mb)					
Autosomal	1-2	0.5-1	1.3-3.4	$\approx$0.2-0.3	2.5-21
X (Z) chromosome	0.8	0.3	0.88	$\geq$0.44	3.1
Synaptonemal complex length (μm)					
Autosomal female	461.5-674.7	185-330	NA	NA	1.71-25.62
Autosomal male	263.6,290.6	120-165	194-307	NA	NA

NA: Not Applicable

Table IV.1: *Comparative analysis of genome landscape in five amniotes. Adapted from Mikkelsen et al. (2007). Additional information from: Pigozzi (2001); Lynn et al. (2002); Tease and Hultén (2004); Basheva et al. (2008)*

Species	initialNbrRE	nonOverlapRE	nonOverlapNoExonsRE
Human	2,300,977	1,876,441	1,849,171
Mouse	1,679,995	1,306,524	1,292,658
Dog	1,985,367	1,595,603	1,594,867
Gallus	79,405	66,499	66,373
Opossum	3,230,405	2,647,349	2,645,638

Table IV.2: *This table contains the number of repeats analyzed for each species after the application of different filters. The first column contains all the repeats of type LINE or SINE, with minimum length of 250 bp for the LINEs and 100 bp for the SINEs. The second column lists the number of REs that do not overlap other REs. The third column specifies the final number of REs after eliminating the REs overlapping exons.*

IV.2 Materials and Methods

The main variables: crossover rates (COR), GC* and current GC content (GC), are calculated in windows along each of the genomes of five species: human, mouse, dog, opossum and chicken. The substitution pattern and the GC* are computed for transposable elements (Arndt et al., 2003).

The Transposable Elements

The whole-genome alignments of transposable elements (TE) to their consensus sequences are provided by the RepeatMasker program (Smit et al., 2010) for the species human, mouse, chicken and opossum. Since the data were not available for dog, we launched Repeat Masker on the whole dog genome assembly (CanFam 2.0, May 2006). Following the protocol of Arndt et al. (2003), we considered only the TEs of type LINE and SINE, with alignment lengths longer than 250 and 100 base-pairs respectively. Moreover, TEs that overlapped in these alignment files have been discarded in order to avoid ambiguities in the correct inference of substitution rates. All TEs overlapping exons were eliminated, because selection on exon sequences could have confounding effects on our gBGC analysis. Exon positions for all genes have been retrieved from version 57 of the Ensembl database (DB-Ensembl).Information regarding the number of TEs used for each species, before and after the application of different filters, is available in table IV.2.

Windows

We focused the analysis on autosomes only. In the case of human, mouse, dog, and chicken, each chromosome was divided into non-overlapping windows of one Megabase (Mb). All TEs within a window were considered. Windows containing no transposable elements were discarded. The TEs overlapping the limits of consecutive windows were reassigned to the window for which the TE had the greatest overlap.

For the opossum data, due to the low number of genetic markers, a window was defined by the positions of two consecutive genetic markers on the genetic map from Samollow et al. (2007). The assignment of TEs to each window was made according to the same principle described above.

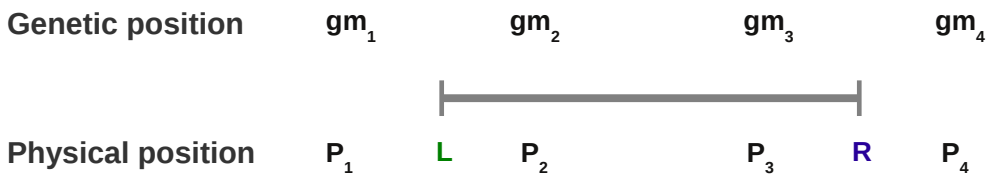

Figure IV.2: *The definition of the genetic markers that characterize a window for human, mouse, dog, and chicken. The window is defined by the physical limits left (**L**) and right (**R**). The four genetic markers associated to this window (named from 1 to 4) have the physical positions P_1, P_2, P_3, and P_4 respectively. Their genetic positions are gm_1, gm_2, gm_3, and gm_4. The recombination rate, RR, is calculated as:* $RR = \frac{\frac{gm_2-gm_1}{P_2-P_1}(P_2-L)+\frac{gm_3-gm_2}{P_3-P_2}(P_3-P_2)+\frac{gm_4-gm_3}{P_4-P_3}(R-P_3)}{R-L}$.

The distribution of window lengths is reported as the distance between the first and last TE assigned to each window.

Recombination

The recombination data were calculated from the sex-specific and sex-averaged genetic maps: human (Matise et al., 2007), mouse (Cox et al., 2009), dog (DB-DogMap; Wong et al., 2010), opossum (Samollow et al., 2007), and chicken (Groenen et al., 2009).

In the case of human, mouse, dog, and chicken, each window was characterized by four genetic markers: the two closest genetic markers left and right from each one of the window limits. The recombination rates, expressed in centimorgans per megabase (cM/Mb), for each window were computed as the average of the recombination rate between each pair of the above four consecutive genetic markers weighted by their overlap to the window. Only windows defined by at least three genetic markers were analyzed. The detailed calculation of the recombination rate is represented in figure IV.2.

All the markers with inverted positions on the genetic and physical maps were discarded. In the end, we were left with 117 useful markers.

Equilibrium GC content (GC) and current GC content*

The maximum-likelihood method of Arndt et al. (2003) is used to compute substitution rates for individual nucleotides accounting for CpG hypermutability. The substitutions are inferred between the ancestral and the current sequences of TEs. The consensus sequence of TEs is supposed to be a good approximation of the ancestral sequence. We aligned the transposable element sequences with their respective consensus sequences inside each window. Based on these alignments, the method infers seven substitution rates, supposing strand symmetry (*e.g* A→G = T→C): four transversions, two transitions, and the CpG transition rate. The equilibrium GC content (GC*), the GC-content towards which a sequence will evolve under a constant substitution pattern, is thus calculated on all non-CpG substitutions in each window according to the model of Sueoka (1962) as being the percentage of AT→GC substitutions among all AT→GC and GC→AT substitutions. In order to obtain a high precision in the estimation of the substitution frequency, we

eliminated from the analysis any windows containing concatenated alignments that had less than 100 Kb of uninterrupted, unambiguous nucleotide sequences (no indels and no N) (Duret and Arndt, 2008). Since the number of TEs in the chicken genome are much less numerous (table IV.2), for this species, we eliminated any window containing alignments with less than 20 Kb of informative nucleotide sequences (the value of the first quartile).

For each window, the current GC content was computed based on the genomic sequences after eliminating the exons. The positions of the exons were retrieved from version 57 of the Ensembl database (DB-Ensembl).

The centromere positions

The centromere positions for human, mouse, dog, and chicken were retrieved from (Karolchik et al., 2004). The cytological determination of centromere positions in opossum was done according to Duke et al. (2007).

Statistics

We quantified the strength of the correlation between any two variables using Pearson's ρ correlation coefficient. In the case of two correlations sharing one common variable, in order to test which is stronger, we apply a Hotteling-William's t-test (H-W test), with the null hypothesis $r_{XY} = r_{XZ}$.

$$t = |r_{XY} - r_{XZ}| \sqrt{\frac{(N-1)(1+r_{YZ})}{2\frac{N-1}{N-3}|R| + \frac{(r_{XY}+r_{XZ})^2}{4}(1-r_{YZ})^3}}$$

where $|R| = 1 - r_{XY}^2 - r_{XZ}^2 - r_{YZ}^2 + 2r_{XY}r_{XZ}r_{YZ}$ (the determinant of the 3x3 correlation matrix) and N is the number of observations in each variable X and Y. This ratio follows a Student's t distribution with *N-3* degrees of freedom.

In the main text, when comparing the strength of two correlations, reported in brackets, are the p-values of the H-W test.

IV.3 Recombination, nucleotide composition, sex and chromosome localization

IV.3.1 Sex-specific impact in vertebrates

Our analyses of all five vertebrates result in overall strong correlations between the equilibrium GC content (GC*) and the sex-specific CO rates (COR) (table IV.3). For the first time, we find that the GC* is more strongly correlated with male rather than female local COR in all three eutherian mammals (human p-value$< 10^{-6}$, mouse p-value$< 10^{-6}$, and dog (p-value$< 10^{-6}$) (table IV.3). This result is consistent with previous observation in human (Webster et al., 2005; Duret and Arndt, 2008). However, we detect no sex-specific impact in chicken (p-value=0.1434) and for the opossum, the female COR is a better predictor of GC* than the male COR (p-value=0.028) (figure IV.3 and table IV.3). **The differential impact of sex according to the organisms is a first observation that the male recombination is not a driving factor in all species.**

ρ_{COR,GC^*}	Human		Mouse		Dog		Opossum		Chicken	
	♀	♂	♀	♂	♀	♂	♀	♂	♀	♂
all data	$0.29^{***} < 0.40^{***}$		$0.25^{***} < 0.36^{***}$		$-0.03 < 0.22^{***}$		$0.47^{*} > 0.31^{*}$		$0.54^{***} \approx 0.58^{***}$	
no centromeres	$0.31^{***} < 0.41^{***}$		$0.24^{***} < 0.36^{***}$		$-0.001 < 0.25^{***}$		N.D.		$0.52^{***} \approx 0.57^{***}$	
no telomeres	$0.39^{***} > 0.26^{***}$		$0.25^{***} < 0.37^{***}$		$0.002 < 0.15^{**}$		N.D.		$0.56^{***} \approx 0.60^{***}$	
interstitial regions	$0.41^{***} > 0.28^{***}$		$0.25^{***} < 0.37^{***}$		$0.03 < 0.18^{**}$		N.D.		$0.50^{***} \approx 0.57^{***}$	

Table IV.3: *Pearson's ρ correlation coefficient between recombination rate and GC* for four data sets: 1. all windows along the chromosomes, 2. no centromere windows, 3. no subtelomeric windows, and 4. only the interstitial windows (5 Mb away from telomeres and centromeres).*

*** p-values of the correlation test $\leq 10^{-16}$
** p-values of the correlation test $\leq 10^{-10}$
* p-values of the correlation test ≤ 0.05
N.D. no available data
$>$, $<$ and $\approx$ - statistical difference between female and male ρ values (Hotteling-William's t-test) are reported as inequalities.

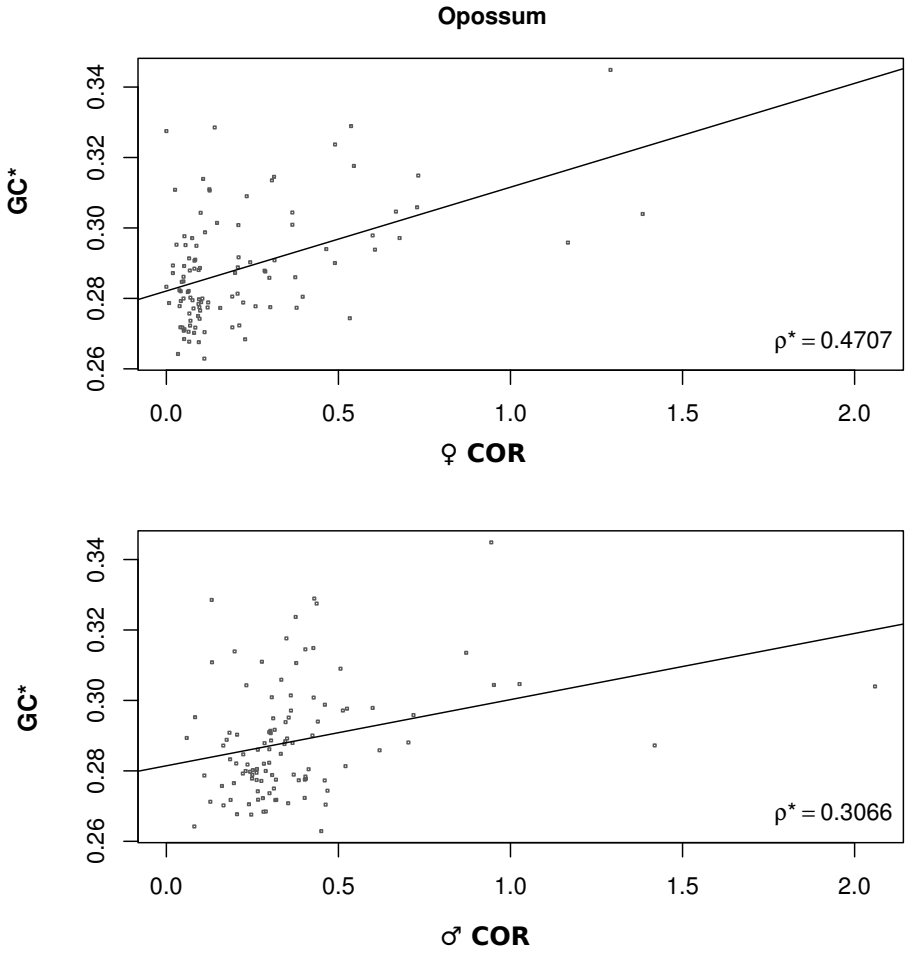

Figure IV.3: *The correlations between recombination rates (in male and female) and GC* in opossum. The value of Pearson's ρ correlation coefficient is reported for each graph. One asterisk near these values stand for p-values ≤ 0.05 of the correlation test.*

IV.3.2 Chromosome localization

Males and females differ not only in the total number of COs but also in their localization along chromosomes (chapter I.2.4.2). Our results in chapter III emphasize the role played by the mechanisms responsible for interference in the distribution of CO events. We have found that the sex with the smallest inter-CO distance has a more uniform distribution of recombination. On the other hand, the sex with the largest interference strength will usually have fewer COs. The placement of these COs is not random along the chromosomes. As described in chapter I.2.1.1, telomeres promote recombination activity in their vicinity while centromeres act as suppressors. Moreover, the intensity of CO production in these regions varies between sexes (chapter I.2.4.2).

Thus, **the telomeric, subtelomeric, and centromeric regions play an important role in the study of the sex-specific differences in recombination**. This observation has motivated us to classify windows according to their position on the chromosome: 5 Mb close to telomeres, centromeres or interstitial. This chromosomal representation reveals that, especially in human, the male COR/GC* correlation is driven mainly by windows situated less than 5 Mb away from telomeres (figure IV.4). In table IV.3, the "no telomeres" data shows a stronger COR/GC* correlation in females than in males. We interpret this high regional impact in the male dataset as the result of highly localized male CO hotspots close to telomeres as opposed to a more uniform distribution in females (Kong et al., 2002). In human, the elimination of windows situated within 5 Mb of telomeres (hereafter called subtelomeric) reduces drastically the genome-wide variability in male recombination rate and causes the female, rather than male, COR to correlate better with GC* (p-value$< 10^{-6}$) (table IV.3). This shift in the sex impact is characterized by a difference between sexes of the same magnitude as the one favoring the male sex, on all windows. This result attests that even **at the level of one genome both sexes can drive the COR/GC* correlation with equal strengths according to the chromosomal localization**.

In spite of a similar sex specific distribution of recombination hotspots in the other two eutherians, mouse and dog (Cox et al., 2009; Wong et al., 2010), these species are not subject to a reversal in the sex influence outside subtelomeric regions (p-value$< 10^{-6}$) (table IV.3). According to the initial hypothesis, the male sex could be the genuine driving factor of the COR/GC* correlation in these species, independent of chromosomal regions. However, these genomes, as opposed to the human genome, are **highly diverged relative to the common ancestor of eutherian mammals, presenting multiple chromosomal rearrangements** (O'Brien et al., 1999; Nash et al., 2001; Wienberg, 2004; Murphy et al., 2005; Kemkemer et al., 2009). This recent shuffling of the genomic sequence can result in younger chromosome ends as well as the distribution of telomeric- and subtelomeric- specific structures along the interiors of chromosomes in these species (Meyne et al., 1990), thus influencing the disposition of hotspots of recombination and the nucleotide substitution patterns on a wider chromosomal range. This hypothesis is consistent with the observation that in mouse a substantial proportion of hotspots is shared between the two sexes outside the subtelomeric regions (Paigen et al., 2008).

As discussed in chapter III, it is uncertain whether the chicken is or is not subject to heterochiasmy. Our results indicate the lack of sex-specific impact on the correlation

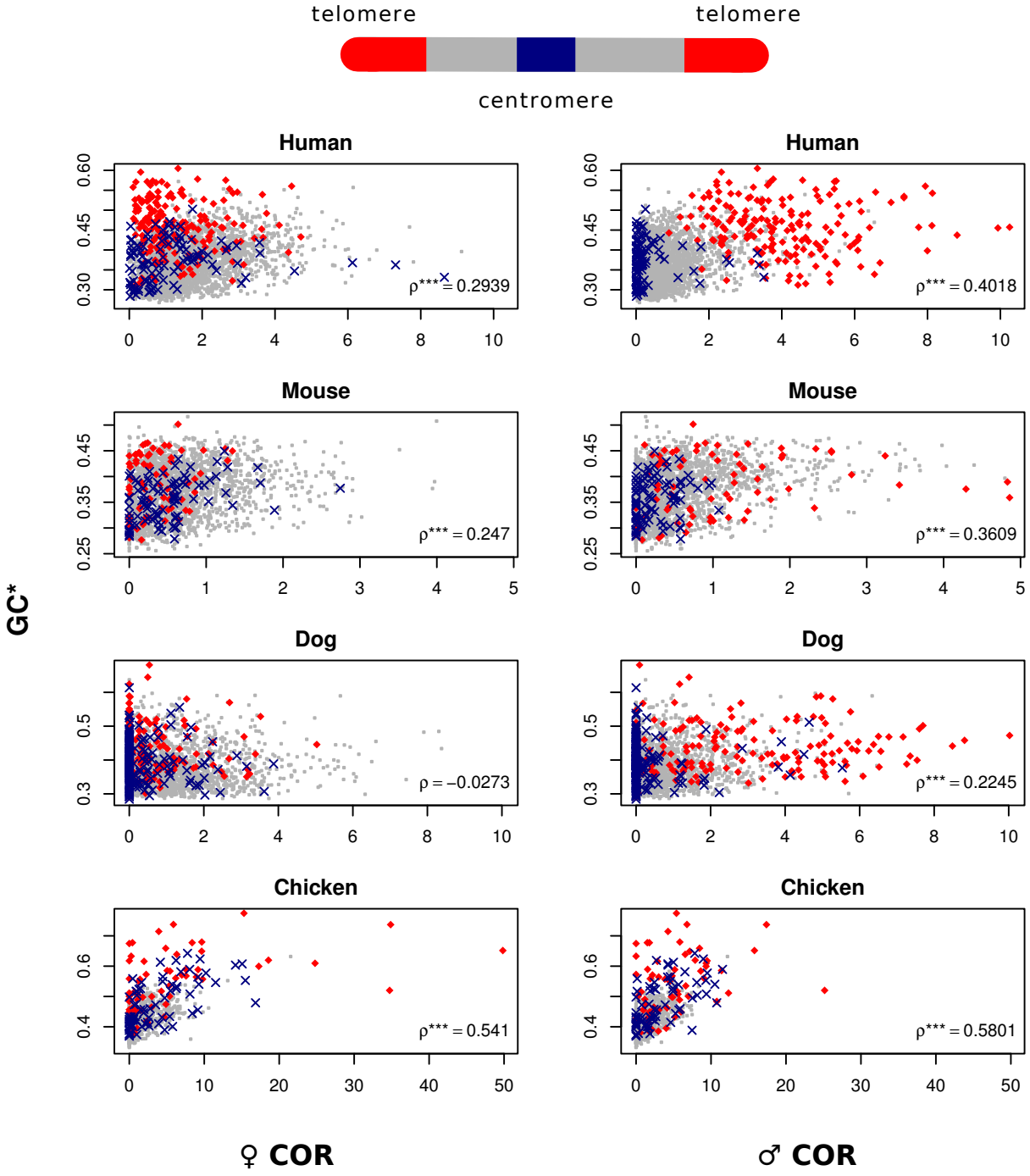

Figure IV.4: *The correlations between recombination rates (in male and female) and the equilibrium GC-content in human, mouse, dog, and chicken. Each point represents the values of the variables in an approximately 1 Mb window. The blue x-shaped like points represent windows that are within 5 Mb (left and right) of the centromere. The red diamonds are windows that are within 5 Mb of the telomeres. The gray points represent the rest of the windows which are not localized near centromeres or telomeres. The value of Pearson's ρ correlation coefficient is reported for each graph. Three, two and one asterisks near these values stand for p-values of the correlation test* $\leq 10^{-16}$, $\leq 10^{-10}$ *and* ≤ 0.05 *respectively.*

between COR and base composition in this species, before or after the elimination of windows in the vicinity of telomeres (table IV.3). It is not yet clear whether these results favor the absence of heterochiasmy or they follow from the lack of power to detect sex-specific differences. This lack of power can result from a limited number of data points due to a low density of TEs in chicken (International Chicken Genome Sequencing Consortium, 2004).

The opossum genome is distinguished from those of most other vertebrates by exhibiting a much higher COR in males compared to females, and thus provides a new model organism for the study of sex-specific impact on the recombination and genomic landscape (Samollow et al., 2007). Interestingly, we find that in opossum the female, rather than the male, COR correlates more strongly to the nucleotide substitution pattern (figure IV.3 and table IV.3). We expect this sex-specific impact to become even more pronounced as more linkage data, especially in subtelomeric regions, become available. Consistent with this prediction, recent addition of new linkage map markers in subtelomeric regions, beyond the previous map ends, has greater impact on increasing the length of female linkage map than that of the male map (unpublished data, discussed by Samollow (2010)). This result is supported by cytologic studies of meiotic cells of the opossum (Sharp and Hayman, 1988) which revealed that chiasmata are concentrated near the ends of chromosomes in female metaphase I nuclei, while those of males are much more evenly distributed. To the extent that this physical pattern reflects the actual distribution of chromosomal exchange events, we expect the female recombination rate to be greater than male recombination rate in subtelomeric regions. Furthermore, subtelomeric regions in opossum are also rich in GC (Mikkelsen et al., 2007).

In additional figures C.1 and C.2, we represent the correlation of the distance to telomeres (DT) with GC*, female COR, and male COR. In agreement with previous studies, we observe that DT is a strong predictor of COR (reviewed in Backström et al. (2010)). Table IV.4 presents a summary of these correlations as well as the comparison of the DT/COR strength between female and male. We show that the DT/COR correlation is stronger in males than in females for the three eutherian mammals: human, mouse, and dog (table IV.4). This result is mainly a consequence of a strong localization of male recombination hotspots in telomeric and subtelomeric regions for these species (Kong et al., 2002; Shifman et al., 2006; Cheung et al., 2007). In chicken, the strength of DT/COR correlation is comparable among sexes ($r = -0.156$ and $r = -0.165$ for female and male respectively) (H-T p-value=0.78). While in opossum, markers in telomeric and subtelomeric regions are scarce, given the preliminary data, we find that the DT is more strongly negatively correlated with female ($r = -0.483$) than with male ($r = -0.211$) COR (H-T p-value=2.5×10^{-4}) (figure C.2 and table IV.4).

IV.3.3 Quantifying the impact on GC*

We identified, so far, three variables that have an influence on the GC*: **COR, sex, and DT**. The starting premises of this work are that potential sex-linked biases in the gBGC mechanism or a sex-specific additional information brought by NCOs affect the COR/GC* correlations. Comparisons of linear regression models incorporating these variables, separately or together, shed light on these premises (table IV.5). First, for all

	DT		
	GC*	**♀COR**	**♂COR**
Human	-0.41***	-0.32***	-0.49***
H-W test p-value		4.6×10^{-16}	
Mouse	-0.28***	-0.06*	-0.28***
H-W test p-value		1.1×10^{-20}	
Dog	-0.07*	-0.21***	-0.33***
H-W test p-value		3.7×10^{-5}	
Opossum	-0.17	-0.48*	-0.21*
H-W test p-value		2.5×10^{-4}	
Chicken	-0.29**	-0.16*	-0.16*
H-W test p-value		0.78	

Table IV.4: *Pearson's ρ correlation coefficients of distance to telomeres (DT) and equilibrium GC (GC*), female, and male crossover rate (♀COR and ♂COR respectively), in human, mouse, dog, opossum, and chicken. The Hotteling-William's t-test p-value (H-W test p-value) that compares the strengths of correlation DT/COR between male and female are also reported.*

*** p-values of the correlation test $\leq 10^{-16}$
** p-values of the correlation test $\leq 10^{-10}$
* p-values of the correlation test ≤ 0.05

Species	**GC*=**			
	f(♀COR)	**f(♂COR)**	**f(COR_{Avg})**	**f(COR_{Avg},LDT)**
Human	0.09	0.16	0.18	0.28
Mouse	0.06	0.13	0.14	0.16
Dog	0.00	0.05	0.02	0.04
Opossum	0.22	0.09	0.18	0.19
Chicken	0.29	0.34	0.38	0.47

Table IV.5: *The R^2 for linear regression models explaining the evolution of GC* as function of different factors: female COR, male COR, and sex-averaged COR. LDT stands for the log distance to telomeres. All R^2 are significantly different from 0, except the $GC^* = f($♀COR model in dog.*

species, except the dog, the COR/GC* correlation is at least as strong for the sex-averaged as it is for the strongest of the sex-specific genetic maps. This result implies that the sex-averaged COR is an equally good predictor of the nucleotide composition (explaining between 14 and 38% of the variability) (table IV.5). The particularity of the dog genome, for which the COR in female is not correlated to the GC*, results in a decrease of the sex-averaged COR/GC* relation (see section IV.3.4).

Second, we increase the proportion of variability by adding the variable log DT (LDT): $GC^* = f(COR_{Avg}, LDT)$ (table IV.5). This increase is significant except for the opossum. Since for the opossum we lack data values close to telomeres, the LDT parameter doesn't

	Number of windows	GC* - ♀COR	GC - ♀COR	GC* - ♂COR	GC - ♂COR	GC* - GC	♀COR - ♂COR
Human	2653	0.2939	0.2486	0.4018	0.3567	0.9619	0.2761
H-W p-value		8.798×10^{-19}		4.129×10^{-20}			
Mouse	2240	0.2470	0.1796	0.3609	0.3333	0.8885	0.3354
H-W p-value		3.529×10^{-12}		2.983×10^{-3}			
Dog	2137	-0.0284	-0.0536	0.2320	0.1973	0.9553	0.0632
H-W p-value		9.514×10^{-5}		3.341×10^{-8}			
Opossum	107	0.4721	0.3766	0.3079	0.2992	0.933	0.6441
H-W p-value		0.0026		0.8014			
Chicken	646	0.5410	0.5292	0.5801	0.5963	0.8807	0.6337
H-W p-value		0.463		0.289			

Table IV.6: *Pearson's ρ correlation coefficients between male and female recombination, GC* and current GC in human, mouse, dog, opossum, and chicken. The Hotteling-William's t-test p-value (H-W p-value) that compares the strengths of correlation are reported.*

significantly improve the model in this species (p-value=0.26). We conclude that **the sex-averaged COR and LDT are good predictors of the GC***, in all species. In addition to the significant interaction between these two predictors, LDT might also account for the variation in GC* independently of COR. CO hotspots are not the only recombination products that have a preferential subtelomeric localization. The existence of DSB hotspots in these regions might generate NCOs (Blitzblau et al., 2007; Buhler et al., 2007; Barton et al., 2008) (chapter I.2.1.1), which could also impact on the nucleotide composition of these regions, thus generating a supplementary effect of LDT on the GC*.

IV.3.4 The particular case of the dog

The sex-specific GC*/COR correlations in dog follow the same trend as for human and mouse (stronger in male than female) (table IV.3). However, the difference between sexes is particularly strong for this species, as the COR/GC* and COR/current GC correlations are null in female (table IV.6). The dog is also the species with the poorest correlation between male and female COR (table IV.6). Such null or negative correlations, between COR and GC-related features, in female dog, have been previously reported (Wong et al., 2010). The degenerated motif typical of human recombination hotspots is associated with high male CORs, but not female (Wong et al., 2010). Together with the poor correlation between sex-specific CORs (table IV.6), these results suggest that the relation between genomic features and local CORs is mediated differently in the two sexes (Wong et al., 2010).

IV.3.5 Cause-effect implications

Each one of the variables, recombination, nucleotide composition, and substitution pattern, is influenced by and influences the others (figure IV.1). In figure IV.1, we represent three

possible models explaining the relation between variables. While all three variables seem inter-connected, in order to understand the molecular mechanism animating the evolution of genomes, the question of which is the strongest factor arises. It has been proposed that a stronger correlation of COR to GC*, rather than the current GC indicates that the causality is more likely recombination→GC rather than GC→recombination (Meunier and Duret, 2004). In agreement with this hypothesis, we find that the COR/GC* correlation is systematically higher than the COR/current GC (table IV.6). The lack of statistical difference between these correlations in chicken might be due to the GC-content of this species being close to equilibrium (Webster et al., 2006), as its karyotype is similar to the ancestral karyotype of amniotes (Bourque et al., 2005). As we have discussed in chapter III, the karyotype is a major determinant of recombination in different species. Thus, a stable karyotype could indicate a conserved recombination landscape, which in turn ensures a constant substitution pattern, under gBGC.

These results are in agreement with the hypothesis of recombination being the driving factor in the evolution of base composition by affecting the substitution pattern through the process of gBGC. Nevertheless, these results do not exclude the alternative explanations of the GC-content affecting both the GC* and recombination, or the influence of other, yet unidentified factors, acting on all the above mentioned variables.

IV.3.6 Reviewing the hypothesis of sex-specific impact

Two hypotheses have been proposed so far to explain the stronger male impact on nucleotide composition observed in humans (Duret and Arndt, 2008). One explanation is that in males the bias induced by gBGC is stronger than in females. This might be due to the fact that male meiosis takes place on a strongly methylated genome (unlike the female meiosis) (Lees-Murdock and Walsh, 2008; Sasaki and Matsui, 2008) which would be more prone to mutations through cytosine deamination. A stronger bias towards the fixation of $AT \rightarrow GC$ mutations would counter-balance this higher cytosine mutation rate, and would thus be advantageous. According to our results, the sex with a dominant effect on the nucleotide composition varies depending on the species (table IV.3). In opossum, it is the female COR that is driving GC*, while in chicken we detect no sex-specific effect. Thus, consistent with this first hypothesis, in opossum, the female, contrary to male, recombination should be taking place on a methylated genome in this species and no difference of methylation should affect the chicken genome. Moreover, in human, we find that the COR/GC* correlation in interstitial regions of the chromosomes is stronger in female than male (table IV.3. **Future investigations of the epigenetic processes during early gametogenesis** are expected to shed light on the pertinence of this hypothesis.

The second hypothesis is that the crossover rate in male is a better estimator of the total recombination rate than in female (Duret and Arndt, 2008). Indeed, the total recombination rate includes all the DSB products, CO and NCO events, which can all affect the nucleotide composition through gBGC. If the ratio between these events were more variable in females, it would generate a weaker correlation between the analyzed CO rate and the total DSB rate and thus account for a less marked impact of COs on the GC* in this sex. Data on the number and distribution of NCOs will allow further testing of this hypothesis. The strong contribution of the DT variable to the model linking sex-averaged

COR to GC* is indirect evidence of **an additional role played by NCOs** (table IV.5). We hypothesize that DT is indicative not only of the CO, but also NCO distribution and thus brings a supplementary explanation to the variation of GC*.

In addition to the above-mentioned hypotheses and in view of our results, we propose **an alternative explanation** based neither on the strength of gBGC, nor on the additional impact of NCO events, but **on the difference in the distribution and usage of COs between sexes** (chapitre I.2.4.2). **At a local level, high CORs, independent of the sex inducing them, will experiment more gBGC events and thus, generate a stronger influence on the regional nucleotide composition**. The observed sex-linked difference in the COR/GC* correlation is mainly linked to the strategies for the distribution of CO events (heterochiasmy). In male eutherian mammals and in female opossum, the crossovers are mainly localized in the telomeric and subtelomeric regions, while the opposite sex presents a more uniform distribution of these events (Sharp and Hayman, 1988; Matise et al., 2007; Cox et al., 2009; Wong et al., 2010). Moreover, in eutherians, the usage of hotspots is also sex-dependent, with the fewer male hotspots exhibiting an intense activity, whereas the many COs in female correspond to low and medium recombination hotspots (Petkov et al., 2007; Coop et al., 2008). Thus, the subtelomeric, intensely used male COR hotspots account for a greater GC* in eutherian mammals, than the evenly distributed, moderately female COR hotspots. In agreement with this hypothesis, we detect no sex-specific impact in chicken, for which no notable differences between male and female COR distribution and number have been observed (Groenen et al., 2009) (figure IV.4). While the molecular mechanism responsible for the sex-specific number and distribution of recombination hotspots is still unclear, it is intimately linked with a difference in Mb interference between the sexes (reviewed in Paigen and Petkov 2010). The sex with stronger interference will generally have less COs (chapter III.3.2), and since chromosome ends are rich in recombination hotspots (chapter I.2.1.1), these COs are usually situated close to telomeres. It follows that one sex will have intense telomeric CO hotspots, while the other will have a more even distribution of recombination events and intensities along the chromosomes. Moreover, the physical interference distance is intimately linked to the compaction of chromosomes during meiosis (de Boer et al., 2006). Although in eutherians, the interference distance (when measured in microns) is the same between the sexes, a different compaction level of the chromatids determines the COs to be further away at the Mb scale in males (de Boer et al., 2006).

IV.4 Discussing the methodology

IV.4.1 Using TEs

The use of TEs has allowed the above analyses to be performed in other vertebrates than human, in the absence of multiple-species whole-genome alignments. The results obtained from organisms with different heterochiasmy patterns have allowed us to propose a new hypothesis for the role of sex in the COR/GC* correlation. We have thus formulated our hypothesis that heterochiasmy itself, and no other sex-factor is the main factor impacting on the GC*. However, the insertion of TEs is not random, as in human, Alus are preferentially

$\rho_{COR,GC*}$	♀	♂	**H-W p-value**
decode2002 AllData	0.387	0.515	4.6×10^{-11}
decode2002 NoTelo	0.492	0.420	6.5×10^{-5}

Table IV.7: *Pearson's ρ correlation coefficient between human 2002 decode genetic maps (Kong et al., 2002) and GC* inferred from human-chimpanzee-macaque triple alignment (Duret and Arndt, 2008) in: all windows (AllData) along the chromosomes and no subtelomeric windows (5 Mb away from telomeres) (NoTelo). Also, the p-value of the Hotelling-William test (H-W p-value) for the comparison of correlation strength between male and female.*

fixed in GC-rich regions, while LINEs prefer GC-poor sequences Soriano et al. (1983); Smit (1999); International Human Genome Sequencing Consortium (2001). This insertion bias could in principle account for the observed substitution bias.

Moreover, TEs have been generated by bursts of insertion at different times in evolution. Thus, the TEs present in a genome have different ages and the substitution pattern they generate are indicative of multiple substitution processes taking place over longer periods of time. Meanwhile, the COR rates inferred from genetic maps correspond to the current recombination process, which is dynamic, with the perpetual birth and death of recombination hotspots Ptak et al. (2005); Winckler et al. (2005). A better description of the COR/GC* relation is expected if using recently diverged sequences.

However, the conclusions we obtain in the human genome, by using TEs, hold true on the triple human-chimpanzee-macaque non-coding sequences from Duret and Arndt (2008) (table IV.7). We have also filtered the TE subfamilies according to their divergence. We retained only those subfamilies with a mean divergence $\leq 20\%$ and standard deviation $\leq 5\%$. Furthermore, all copies with $> 20\%$ divergence were eliminated. In order to have enough data points, the windows containing concatenated alignments that had more than 20 kb, instead of 100 kb, of uninterrupted, unambiguous nucleotide sequences were analyzed. It follows that by reducing this constraint, the number of data points is similar between this analysis (table IV.8) and the previous unfiltered data (table IV.6). This filter could not be applied on the data in chicken because of the drastic reduction in the number of windows left. The conclusions on the difference between sex-specific impact and chromosome localization remain unchanged after applying this divergence filter (table IV.8).

A puzzling effect of the use of TEs when inferring the GC* is that, contrary to earlier studies Webster et al. (2005); Duret and Arndt (2008), our estimations of the GC*/current GC correlation coefficients are much higher than those previously reported (table IV.6). Since these very high correlations could be indicative of a bias in the method and/or the data, we tested our methodology only on Alu subfamilies like in Webster et al. (2005), but using 1 Mb windows (Table IV.9). There is a decrease in the strength of the GC*/current GC correlation with the decrease in Alu subfamily divergence (AluJ: 0.746, AluS: 0.725, AluY: 0.483). A decrease in this correlation is also observed in all species after applying the divergence filter (*e. g.* 0.8708 instead of 0.9619 in human) thus we believe that there

	Number of windows	**GC* - ♀COR**	**GC - ♀COR**	**GC* - ♂COR**	**GC - ♂COR**	**GC* - GC**	**♀COR - ♂COR**
Human	2678	0.308	0.245	0.4657	0.3613	0.8708	0.2777
H-W p-value		1.911×10^{-11}		1.315×10^{-32}			
Mouse	1852	0.213	0.1967	0.2722	0.3005	0.6648	0.3417
H-W p-value		0.3706		0.1179			
Dog	2153	-0.0638	-0.059	0.3072	0.1974	0.7868	0.0607
H-W p-value		0.736		3.643×10^{-16}			
Opossum	107	0.447	0.3773	0.272	0.299	0.8069	0.6441
H-W p-value		0.2031821		0.643			
Chicken	646	0.5410	0.5292	0.5801	0.5963	0.8807	0.6337
H-W p-value		0.463		0.289			

Table IV.8: *Pearson's ρ correlation coefficients between male and female recombination, GC* and current GC in human, mouse, dog and chicken on the REs filtered for family divergence. Only families with a mean divergence $\leq 20\%$ and standard deviation $\leq 5\%$, and copies with $\leq 20\%$ divergence have been analyzed. Only alignments containing > 20 kb of repeat sequence were retained for further analysis. The Results are reported for 1Mb windows correlations. The Hotteling-William's t-test p-value (H-W p-value) that compares the strengths of correlation are reported.*

Source	**Nb. of windows**	**Mean length**	Pearson's correlation coefficient (ρ)					**Divergence (Myr)**
			*GC GC**	*GC** ♀COR	*GC* ♀COR	*GC** ♂COR	*GC* ♂COR	
AluJ Webster et al.	3819	595±467 kb	0.68	0.218	0.189	0.409	0.277	73
AluJ This study	1207	1Mb	0.746	0.086	0.083	0.445	0.326	
AluS Webster et al.	3843	592±468 kb	0.632	0.235	0.25	0.376	0.297	43
AluS This study	2528	1Mb	0.725	0.214	0.221	0.379	0.348	
AluY Webster et al.	3799	598±467 kb	0.503	0.243	0.186	0.434	0.388	28
AluY This study	554	1Mb	0.483	0.061	0.016	0.492	0.347	

Table IV.9: *The correlations between the GC*, sex-specific recombination rate and current GC on the three Alu families are compared between our approach and the one used in Webster et al. (2005). Our results are based on the 1 Mb windows but, as proposed by Webster et al., retaining only alignments containing > 20 kb of repeat sequence.*

is no bias in the method (table IV.8).

When computing the substitution rates on highly diverged neutrally evolving sequences, we measure an average substitution pattern which has been under the impact of mutation and fixation biases as well as selective sweeps and background selection. The substitution

process is heterogeneous both in time and at the genomic level. The recombination hotspots are local structures (covering ≈2 Kb of genomic sequence Myers et al. 2005) with a short lifespan (Ptak et al., 2005; Winckler et al., 2005). When a hotspot covers a neutrally evolving region it can induce biases in the substitution pattern, such as the GC bias induced by gBGC. This bias would lead to an increase in the local GC content. However, once the recombination hotspot has died away, the recombination-induced bias would also disappear and the local GC-content is expected to decrease under an AT-biased mutation rate (Sueoka, 1988; Hershberg and Petrov, 2010). The average substitution rates inferred at long time scales, in 1 Mb windows, generate a GC* that will fluctuate around the local genomic GC-content, explaining the increase in GC*/current GC correlation with the increase in divergence. Despite performing a correlation between time-averaged substitution patterns and present recombinational landscape, all our analyses confirm previous results. While recombination hotspots are short-lived, chromosomal regions, such as telomeres, consistently experience recombination events, being thus informative of the time-averaged recombination rate.

IV.4.2 Window length

Because of the opossum low-density genetic map, we define windows in this species between two adjacent genetic markers (table IV.2). This procedure generates windows with a mean average of 27 Mb and a standard deviation of 25 Mb. The length and variability of window sizes can have a confounding effect on the interpretation of the sex-specific impact on the correlation between recombination and GC*. Previous studies in human (Duret and Arndt, 2008) and yeast (Marsolier-Kergoat and Yeramian, 2009) detect an increase in the recombination/GC* correlation coefficients with the size of the windows. Notably, at the scale of a few Kb, the locations of crossover hotspots are known to vary strongly among individuals of the same species (Neumann and Jeffreys, 2006; Jeffreys and Neumann, 2009), while it has been proposed that, at the Mb scale, the recombination regions are more stable in time (Myers et al., 2005). It is thus difficult to compare the results in opossum, with the 1 Mb-resolution observations in the other four vertebrates.

In order to bypass this difficulty, we tested the effect of different window sizes (between 0.5 Mb and 20 Mb) on the strength of COR/GC* for female and male, in both human and mouse. In both species, the stronger male effect on recombination rate/GC* correlation is detectable for small window sizes (human $\leq$ 10 Mb and mouse $\leq$ 15 Mb), and as the size of the windows increases, the sex-specific difference diminishes and disappears (Table IV.10). In contrast, the stronger female COR correlation with GC* persists in opossum, even for windows with a mean size $>$ 20 Mb. Moreover, when dividing the opossum dataset into windows smaller and larger than 20 Mb (additional figures D.1 and D.2), the stronger female effect is conserved at different scales.

IV.5 Conclusion

In this chapter, we present a study of the relation between nucleotide composition, COR and sex. The starting point is the observation that, in human, the COR/GC* correlation is

	Human			Mouse		
Window Size (Mb)	**Nr data**	$\rho_{COR♂,GC*} - \rho_{COR♀,GC*}$	**p-val**	**Nr data**	$\rho_{COR♂,GC*} - \rho_{COR♀,GC*}$	**H-W p-val.**
0.5	5355	0.1046	0	4786	0.0819	0
1	2688	0.1137	0	2393	0.1104	0
1.5	1794	0.1152	0	1593	0.1356	0
2	1350	0.1211	0	1197	0.1398	0
2.5	1087	0.1368	0	954	0.1346	0
5	541	0.1773	10^{-4}	473	0.1247	0.0037
7	383	0.0873	0.0637	335	0.1258	0.0093
10	267	0.0424	0.382	230	0.1068	0.0516
12	221	10^{-4}	0.9988	191	0.1292	0.0356
15	175	-0.0134	0.7786	149	0.0936	0.1374
20	128	-0.0422	0.3231	110	0.0594	0.3953

Table IV.10: *The value of $\rho_{COR♂,GC*} - \rho_{COR♀,GC*}$ and the p-value of the Hotteling-William's t-test (H-W p-value) for the significance of this quantity, for different window sizes in human and mouse.*

stronger in male than in female (Webster et al., 2005; Duret and Arndt, 2008). Our analysis of this correlation in different vertebrates has revealed that **the main factor explaining GC* is the intensity in COR**. In turn, heterochiasmy generates sex-differential COR distribution and intensities along the chromosomes. Thus, **independent of the sex, a region with a high COR will have a significant bias in the substitution pattern, through mechanisms such as gBGC, and in time result in an increase in GC-content.**

The analysis of the base composition characteristics based on TEs has proven a valuable tool in the study of patterns of genome evolution. The availability of less diverged triple alignments of non-coding sequences between relatively closely related species will provide a complementary method for inferring the substitution pattern in other species as well. Finally, in light of the reversals in sex effects seen in opossum recombinational and base composition characteristics, increasing resolution of the opossum genetic map promises to provide a naturally occurring, comparative model for investigating processes that differentiate the location and resolution of meiotic DSB events between the sexes. The sequencing of other genomes from species with the same heterochiasmy pattern, such as two other metatherian mammals, *Sminthopsis crassicaudata* and *Macropus eugenii*, or the sheep, *Ovis aries*, will represent a valuable tool to study into more depth the heterochiasmy impact on the nucleotide landscape.

Chapter V

Conclusions and Perspectives

*"**Evolution** is a population genetic process governed by **four fundamental forces**, which jointly dictate the relative abilities of genotypic variants to expand throughout a species. Darwin articulated a clear but informal description of one of those forces, **selection** (including natural and sexual selection), whose central role in the evolution of complex phenotypic traits is universally accepted, and for which an elaborate formal theory in terms of change in genotypic frequencies now exists. The remaining three evolutionary forces, however, are non-adaptive in the sense that they are not a function of the fitness properties of individuals: **mutation** (broadly including insertions, deletions, and duplications) is the fundamental source of variation on which natural selection acts; **recombination** (including crossing-over and gene conversion) assorts variation within and among chromosomes; and **random genetic drift** ensures that gene frequencies will deviate a bit from generation to generation independently of other forces. Given the century of theoretical and empirical work devoted to the study of evolution, the only logical conclusion is that these four broad classes of mechanisms are, in fact, the only fundamental forces of evolution. Their relative intensity, directionality, and variation over time define the way in which evolution proceeds in a particular context."*(Lynch, 2007)

The aim of this thesis is the study of one of these evolutionary forces: **recombination**. For this purpose, we used two main approaches. First, we analyzed **the patterns of recombination rate variation between species**. Second, we examined **the differential impact of inter-sexes variation in recombination on the nucleotide composition of genomes**. Both these approaches aim to enhance our understanding of the molecular processes driving the evolution of recombination and of genomic sequences.

To characterize the differences in recombination among species, **we have developed a new model based on genetic maps**, as described in chapter III. This model addresses the relation between the total genetic length of chromosome (indicative of the total number of crossovers (COs)) and their physical length. The model incorporates important biological knowledge of the recombination process: the necessity for at least one CO per pair of homologs in order to ensure their correct segregation and the notion of interference between COs. It further allows the estimation of the rate with which additional COs are produced per Mb and the average strength of interference, defined as the physical distance between consecutive COs. As the model implies an analysis at the global level of the karyotype, it can be applied even on low-resolution data and, hence, **results in the exploration**

of multiple species. In chapter III, we showed that our model adjusts well on all 27 vertebrate and non-vertebrate genetic maps we analyzed, especially on those that cannot be predicted by a simple linear model.

Studies of the inter-COs distance have been previously performed in only a handful of species, and for these species, our estimates of the interference strength are in agreement with these previously estimated values. Moreover, the estimates of this parameter in the remaining species yields useful, novel information for understanding the distribution of CO events. **We show that the interference distance in Mb is highly correlated to the average physical length of chromosomes**. This linear relation represents **an important tool for characterizing inter-CO distances even for species lacking recombination data**. We further compared species based on similarities in the estimated rate of production of additional COs per Mb. **Species with comparable parameter values might also share similar karyotype structures and dynamics affecting the recombination pattern.**

Furthermore, fitting the model on sex-specific genetic maps of 10 vertebrates has proven a valuable tool for the study of heterochiasmy. As expected, **the sex with the smallest inter-CO distance has the higher number of COs, which are more uniformly distributed**. In the majority of cases this also leads to an increased rate of additional CO production. Exceptions were observed for 2 species. In the metatherian *Monodelphis domestica* both parameters (rate of additional CO production and interference strength) are higher in female than male. Whether this is due to a low-resolution of the genetic map in this species, or it is indicative of a particular mechanism in female, this result is worth further investigations. The other species with a peculiar sex-associated trend in the values of parameters is the bird *Gallus gallus*. Our results, as well as recent studies, question the lack of heterochiasmy, previously accepted for this species.

Our results on heterochiasmy have led to the second analysis presented in this thesis (chapter IV). Different patterns of heterochiasmy have been observed in different species. Additionally, we know that recombination has an impact on the nucleotide composition of sequences, through biased gene conversion. So far, studies in human have found the male recombination to be the leading factor in the evolution of GC-content (Webster et al., 2005; Duret and Arndt, 2008). In chapter IV, we have asked the question of the sex-specific impact on the relation between recombination and nucleotide composition, by analyzing this relation in a set of 5 vertebrates. Our results show that the stronger impact of male recombination, observed for humans, is not true for all species. Moreover, even in human, it is mainly driven by regions close to telomeres, which are enriched in male recombination hotspots. These results suggest that **strong CORs are correlated with the nucleotide composition, independent of the sex generating them**. The difference between sexes in the localization and intensity of recombination hotspots is responsible for the differential impact of sex on the GC/COR relation in different regions of the chromosomes. As we have previously demonstrated (chapter III), the sex-differential distribution of CO events is related to the strength of interference.

Moreover, we studied the impact of substitution patterns on the evolution of the GC-content. At a smaller time-scale, the divergence human-chimpanzee, the GC-content has been found very dissimilar to its equilibrium. In chapter IV, we show that **at large**

time-scales, the GC-content of neutrally evolving sequences, subject to mutation, genetic drift, and biased gene conversion **is approaching its equilibrium**. We offer a hypothesis for these apparently contradictory observations. First, recombination hotspots are highly dynamic, indicated by their lack of conservation between close species such as human and chimpanzee Ptak et al. (2005); Winckler et al. (2005). Second, chromosomal regions such as those close to telomeres maintain a very high density in recombination hotspots in a majority of species. These results imply that while the GC-content is fluctuating due to mutation and gene conversion biases, in the long run the two biases might attenuate each other.

The results presented so far give new perspectives on the reciprocal influence of karyotype, sex, recombination, and nucleotide composition. However, as the title of the section containing the opening citation of this chapter states: "Nothing in evolution makes sense except in the light of population genetics" (Lynch, 2007). In agreement with this point of view, the work presented here is part of a bigger project aiming to integrate newly available information on the recombination process, in a model describing its evolutionary impact in a population. In chapter II.3.3 **we show that gBGC plays a major role in the maintenance at relatively high frequencies of deleterious alleles in the human population**. The results of our simulations agree with the real data showing that even at non-synonymous sites, AT→GC disease-associated mutations segregate at a higher frequency than the GC→AT ones.

The work described in this thesis has revealed important factors of the differences in recombination between species and sexes. The sex, karyotype, interference strength, and chromosomal localization of hotspots are only some of the determinants of recombination. Maybe the most important conclusion of this thesis is that the study of recombination would need to integrate data from multiple species. Understanding the differences in the mechanism is equally important to the study of its conserved features. Many questions still remain unanswered. What generates a differential condensation of chromosomes between male and female? How can we explain the differences and similarities in recombination between species accounting for populations dynamics? Model of the gBGC impact have been built mainly for human, what do these models predict for other organisms? How long do recombination hotspots live? How does this dynamics affect the evolution of sequences under gBGC? How about the organization of genomes in isochores and their related genomic features? While huge progress has been made in the last years for the study of recombination and more is to be expected, we hope that the work presented in this thesis can form a starting point for the study of these yet unanswered questions.

Bibliography

The symbol ↪ has been used to denote pages in this work where the publication has been referenced. Whenever available, Digital Object Identifiers (DOIs) have been made clickable and point to online versions of the publications, via `http://dx.doi.org/`

Aboussekhra A., Chanet R., Adjiri A., and Fabre F. Semidominant suppressors of srs2 helicase mutations of saccharomyces cerevisiae map in the rad51 gene, whose sequence predicts a protein with similarities to procaryotic reca proteins. *Mol Cell Biol*, 12(7):3224–3234, Jul 1992. ↪p.171

Ahsan B., Kobayashi D., Yamada T., Kasahara M., Sasaki S., Saito T. L., Nagayasu Y., Doi K., Nakatani Y., Qu W., Jindo T., Shimada A., Naruse K., Toyoda A., Kuroki Y., Fujiyama A., Sasaki T., Shimizu A., Asakawa S., Shimizu N., ichi Hashimoto S., Yang J., Lee Y., Matsushima K., Sugano S., Sakaizumi M., Narita T., Ohishi K., Haga S., Ohta F., Nomoto H., Nogata K., Morishita T., Endo T., Shin-I T., Takeda H., Kohara Y., and Morishita S. Utgb/medaka: genomic resource database for medaka biology. *Nucleic Acids Res*, 36(Database issue):D747–D752, Jan 2008. doi:10.1093/nar/gkm765. ↪p.32, 90

Allen E. G., Freeman S. B., Druschel C., Hobbs C. A., O'Leary L. A., Romitti P. A., Royle M. H., Torfs C. P., and Sherman S. L. Maternal age and risk for trisomy 21 assessed by the origin of chromosome nondisjunction: a report from the atlanta and national down syndrome projects. *Hum Genet*, 125(1):41–52, Feb 2009. doi:10.1007/s00439-008-0603-8. ↪p.36

Allers T. and Lichten M. Intermediates of yeast meiotic recombination contain heteroduplex dna. *Mol Cell*, 8(1):225–231, Jul 2001a. ↪p.17, 18

Allers T. and Lichten M. Differential timing and control of noncrossover and crossover recombination during meiosis. *Cell*, 106(1):47–57, Jul 2001b. ↪p.2, 18

Anderson L. K., Hooker K. D., and Stack S. M. The distribution of early recombination nodules on zygotene bivalents from plants. *Genetics*, 159(3):1259–1269, Nov 2001. ↪p.2, 20, 28

Anderson L. K., Doyle G. G., Brigham B., Carter J., Hooker K. D., Lai A., Rice M., and Stack S. M. High-resolution crossover maps for each bivalent of zea mays using recombination nodules. *Genetics*, 165(2):849–865, Oct 2003. ↪p.28, 32

Archibald A. L., Bolund L., Churcher C., Fredholm M., Groenen M. A. M., Harlizius B., Lee K.-T., Milan D., Rogers J., Rothschild M. F., Uenishi H., Wang J., Schook L. B., and Consortium S. G. S. Pig genome sequence–analysis and publication strategy. *BMC Genomics*, 11:438, 2010. ↪p.32

Argueso J. L., Wanat J., Gemici Z., and Alani E. Competing crossover pathways act during meiosis in saccharomyces cerevisiae. *Genetics*, 168(4):1805–1816, Dec 2004. doi:10.1534/genetics.104.032912. ↪p.28

Arias J. A., Keehan M., Fisher P., Coppieters W., and Spelman R. A high density linkage map of the bovine genome. *BMC Genet*, 10:18, 2009. doi:10.1186/1471-2156-10-18. ↪p.31, 90

Arndt P. F. and Hwa T. Identification and measurement of neighbor-dependent nucleotide substitution processes. *Bioinformatics*, 21(10):2322–2328, May 2005. doi:10.1093/bioinformatics/bti376. ↪p.73, 75

Arndt P. F., Petrov D. A., and Hwa T. Distinct changes of genomic biases in nucleotide substitution at the time of mammalian radiation. *Mol Biol Evol*, 20(11):1887–1896, Nov 2003. doi:10.1093/molbev/msg204. ↪p.4, 71, 75, 112, 113, 114

Arnheim N., Calabrese P., and Nordborg M. Hot and cold spots of recombination in the human genome: the reason we should find them and how this can be achieved. *Am J Hum Genet*, 73 (1):5–16, Jul 2003. doi:10.1086/376419. ↪p.3, 57, 61, 62

Arnheim N., Calabrese P., and Tiemann-Boege I. Mammalian meiotic recombination hot spots. *Annu Rev Genet*, 41:369–399, 2007. doi:10.1146/annurev.genet.41.110306.130301. ↪p.xiii, 21, 36, 37, 63, 64

Auton A. *The Estimation of Recombination Rates from Population Genetic Data*. PhD thesis, Hertford College, University of Oxford, 2007. ↪p.64

Backström N. *Gene Mapping in Ficedula Flycatchers*. PhD thesis, Uppsala Universitet, 2009. ↪p.52, 59, 61, 64

Backström N., Ceplitis H., Berlin S., and Ellegren H. Gene conversion drives the evolution of hintw, an ampliconic gene on the female-specific avian w chromosome. *Mol Biol Evol*, 22(10): 1992–1999, Oct 2005. doi:10.1093/molbev/msi198. ↪p.42

Backström N., Karaiskou N., Leder E. H., Gustafsson L., Primmer C. R., Qvarnström A., and Ellegren H. A gene-based genetic linkage map of the collared flycatcher (ficedula albicollis) reveals extensive synteny and gene-order conservation during 100 million years of avian evolution. *Genetics*, 179(3):1479–1495, Jul 2008. doi:10.1534/genetics.108.088195. ↪p.30, 38

Backström N., Forstmeier W., Schielzeth H., Mellenius H., Nam K., Bolund E., Webster M. T., Ost T., Schneider M., Kempenaers B., and Ellegren H. The recombination landscape of the zebra finch Taeniopygia guttata genome. *Genome Res*, 20(4):485–495, Apr 2010. doi:10.1101/gr.101410.109. ↪p.38, 43, 109, 119

Bailey J. A. and Eichler E. E. Primate segmental duplications: crucibles of evolution, diversity and disease. *Nat Rev Genet*, 7(7):552–564, Jul 2006. doi:10.1038/nrg1895. ↪p.23

Bailis J. M. and Roeder G. S. Synaptonemal complex morphogenesis and sister-chromatid cohesion require mek1-dependent phosphorylation of a meiotic chromosomal protein. *Genes Dev*, 12(22):3551–3563, Nov 1998. ↪p.170

Bailly A. P., Freeman A., Hall J., Déclais A.-C., Alpi A., Lilley D. M. J., Ahmed S., and Gartner A. The caenorhabditis elegans homolog of gen1/yen1 resolvases links dna damage signaling to dna double-strand break repair. *PLoS Genet*, 6(7):e1001025, 2010. doi:10.1371/journal.pgen.1001025. ↪p.17

Baker B. S., Carpenter A. T., Esposito M. S., Esposito R. E., and Sandler L. The genetic control of meiosis. *Annu Rev Genet*, 10:53–134, 1976. doi:10.1146/annurev.ge.10.120176.000413. ↪p.27

Barber L. J., Youds J. L., Ward J. D., McIlwraith M. J., O'Neil N. J., Petalcorin M. I. R., Martin J. S., Collis S. J., Cantor S. B., Auclair M., Tissenbaum H., West S. C., Rose A. M., and Boulton S. J. Rtel1 maintains genomic stability by suppressing homologous recombination. *Cell*, 135(2):261–271, Oct 2008. doi:10.1016/j.cell.2008.08.016. ↪p.25, 172

Barendse W., Vaiman D., Kemp S. J., Sugimoto Y., Armitage S. M., Williams J. L., Sun H. S., Eggen A., Agaba M., Aleyasin S. A., Band M., Bishop M. D., Buitkamp J., Byrne K., Collins F., Cooper L., Coppettiers W., Denys B., Drinkwater R. D., Easterday K., Elduque C., Ennis S., Erhardt G., Li L., and Lil L. A medium-density genetic linkage map of the bovine genome. *Mamm Genome*, 8(1):21–28, Jan 1997. ↪p.34, 38, 91, 105

Barton A. B., Su Y., Lamb J., Barber D., and Kaback D. B. A function for subtelomeric dna in saccharomyces cerevisiae. *Genetics*, 165(2):929–934, Oct 2003. ↪p.20, 22

Barton A. B., Pekosz M. R., Kurvathi R. S., and Kaback D. B. Meiotic recombination at the ends of chromosomes in saccharomyces cerevisiae. *Genetics*, 179(3):1221–1235, Jul 2008. doi:10.1534/genetics.107.083493. ↪p.20, 121

Barzel A. and Kupiec M. Finding a match: how do homologous sequences get together for recombination? *Nat Rev Genet*, 9(1):27–37, Jan 2008. doi:10.1038/nrg2224. ↪p.12

Basheva E. A., Bidau C. J., and Borodin P. M. General pattern of meiotic recombination in male dogs estimated by mlh1 and rad51 immunolocalization. *Chromosome Res*, 16(5):709–719, 2008. doi:10.1007/s10577-008-1221-y. ↪p.98, 99, 112

Bates D. and Watts D. *Nonlinear regression analysis and its applications*, volume 2. Wiley Online Library, 1988. ↪p.88, 89

Baudat F. and de Massy B. Regulating double-stranded dna break repair towards crossover or non-crossover during mammalian meiosis. *Chromosome Res*, 15(5):565–577, 2007. doi:10.1007/s10577-007-1140-3. ↪p.28

Baudat F. and Nicolas A. Clustering of meiotic double-strand breaks on yeast chromosome iii. *Proc Natl Acad Sci U S A*, 94(10):5213–5218, May 1997. ↪p.19, 20

Baudat F., Buard J., Grey C., Fledel-Alon A., Ober C., Przeworski M., Coop G., and de Massy B. Prdm9 is a major determinant of meiotic recombination hotspots in humans and mice. *Science*, Dec 2009. doi:10.1126/science.1183439. ↪p.3, 23, 36, 173

Baudat F., Buard J., Grey C., and de Massy B. Identification d'une protéine-clé pour le contrôle des sites de recombinaison méiotique= Prdm9, a key control of mammalian recombination hotspots. *MS. Médecine sciences*, 26(5):468–470, 2010. ISSN 0767-0974. ↪p.23, 40

Bell G. *The masterpiece of nature: the evolution and genetics of sexuality*. CUP Archive, 1982. ISBN 0856647535. ↪p.30

Bengtsson B. O. The effect of biased conversion on the mutation load. *Genet Res*, 55(3):183–187, Jun 1990. ↪p.79

Berchowitz L. E. and Copenhaver G. P. Genetic interference: don't stand so close to me. *Curr Genomics*, 11(2):91–102, Apr 2010. doi:10.2174/138920210790886835. ↪p.20, 26, 27, 28, 65, 67, 68, 69, 70

Berchowitz L. E., Francis K. E., Bey A. L., and Copenhaver G. P. The role of atmus81 in interference-insensitive crossovers in a. thaliana. *PLoS Genet*, 3(8):e132, Aug 2007. doi:10.1371/journal.pgen.0030132. ↪p.18

Berchowitz L. E., Hanlon S. E., Lieb J. D., and Copenhaver G. P. A positive but complex association between meiotic double-strand break hotspots and open chromatin in saccharomyces cerevisiae. *Genome Res*, 19(12):2245–2257, Dec 2009. doi:10.1101/gr.096297.109. ↪p.19, 20, 22

Berg I., Neumann R., Lam K., Sarbajna S., Odenthal-Hesse L., May C., and Jeffreys A. PRDM9 variation strongly influences recombination hot-spot activity and meiotic instability in humans. *Nature Genetics*, 42(10):859–863, 2010. ISSN 1061-4036. ↪p.3, 23, 25, 36, 37, 175

Berglund J., Pollard K. S., and Webster M. T. Hotspots of biased nucleotide substitutions in human genes. *PLoS Biol*, 7(1):e26, Jan 2009. doi:10.1371/journal.pbio.1000026. ↪p.77, 109

Bernardi G. The vertebrate genome: isochores and evolution. *Mol Biol Evol*, 10(1):186–204, Jan 1993. ↪p.45

Bernardi G. The neoselectionist theory of genome evolution. *Proc Natl Acad Sci U S A*, 104(20): 8385–8390, May 2007. doi:10.1073/pnas.0701652104. ↪p.45

Bernardi G., Olofsson B., Filipski J., Zerial M., Salinas J., Cuny G., Meunier-Rotival M., and Rodier F. The mosaic genome of warm-blooded vertebrates. *Science*, 228(4702):953–958, May 1985. ↪p.44

Beye M., Gattermeier I., Hasselmann M., Gempe T., Schioett M., Baines J. F., Schlipalius D., Mougel F., Emore C., Rueppell O., Sirviö A., Guzmán-Novoa E., Hunt G., Solignac M., and Page R. E. Exceptionally high levels of recombination across the honey bee genome. *Genome Res*, 16(11):1339–1344, Nov 2006. doi:10.1101/gr.5680406. ↪p.31, 90

Bill C. A., Duran W. A., Miselis N. R., and Nickoloff J. A. 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(4): 1935–1943, Aug 1998. ↪p.40

Birdsell J. A. Integrating genomics, bioinformatics, and classical genetics to study the effects of recombination on genome evolution. *Mol Biol Evol*, 19(7):1181–1197, Jul 2002. ↪p.40, 41, 109

Bishop D. K. and Zickler D. Early decision; meiotic crossover interference prior to stable strand exchange and synapsis. *Cell*, 117(1):9–15, Apr 2004. ↪p.2, 18, 25, 84

Bishop D. K., Park D., Xu L., and Kleckner N. Dmc1: a meiosis-specific yeast homolog of e. coli reca required for recombination, synaptonemal complex formation, and cell cycle progression. *Cell*, 69(3):439–456, May 1992. ↪p.171

Blat Y., Protacio R. U., Hunter N., and Kleckner N. Physical and functional interactions among basic chromosome organizational features govern early steps of meiotic chiasma formation. *Cell*, 111(6):791–802, Dec 2002. ↪p.13, 19, 44, 109

Blitzblau H. G., Bell G. W., Rodriguez J., Bell S. P., and Hochwagen A. Mapping of meiotic single-stranded dna reveals double-stranded-break hotspots near centromeres and telomeres. *Curr Biol*, 17(23):2003–2012, Dec 2007. doi:10.1016/j.cub.2007.10.066. ↪p.20, 22, 121

Boddy M. N., Gaillard P. H., McDonald W. H., Shanahan P., Yates J. R., and Russell P. Mus81-eme1 are essential components of a holliday junction resolvase. *Cell*, 107(4):537–548, Nov 2001. ↪p.18

Borde V. and Cobb J. Double functions for the mre11 complex during dna double-strand break repair and replication. *Int J Biochem Cell Biol*, 41(6):1249–1253, Jun 2009. doi:10.1016/j.biocel.2008.12.013. ↪p.13, 171

Borde V., Lin W., Novikov E., Petrini J. H., Lichten M., and Nicolas A. Association of mre11p with double-strand break sites during yeast meiosis. *Mol Cell*, 13(3):389–401, Feb 2004. ↪p.20, 21

Borde V., Robine N., Lin W., Bonfils S., Géli V., and Nicolas A. Histone h3 lysine 4 trimethylation marks meiotic recombination initiation sites. *EMBO J*, 28(2):99–111, Jan 2009. doi:10.1038/emboj.2008.257. ↪p.19, 23

Boulton A., Myers R. S., and Redfield R. J. The hotspot conversion paradox and the evolution of meiotic recombination. *Proc Natl Acad Sci U S A*, 94(15):8058–8063, Jul 1997. ↪p.40

Bourque G., Zdobnov E. M., Bork P., Pevzner P. A., and Tesler G. Comparative architectures of mammalian and chicken genomes reveal highly variable rates of genomic rearrangements across different lineages. *Genome Res*, 15(1):98–110, Jan 2005. doi:10.1101/gr.3002305. ↪p.122

Bowers J. E., Abbey C., Anderson S., Chang C., Draye X., Hoppe A. H., Jessup R., Lemke C., Lennington J., Li Z., Lin Y.-R., Liu S.-C., Luo L., Marler B. S., Ming R., Mitchell S. E., Qiang D., Reischmann K., Schulze S. R., Skinner D. N., Wang Y.-W., Kresovich S., Schertz K. F., and Paterson A. H. A high-density genetic recombination map of sequence-tagged sites for sorghum, as a framework for comparative structural and evolutionary genomics of tropical grains and grasses. *Genetics*, 165(1):367–386, Sep 2003. ↪p.98

Broman K. W. and Weber J. L. Characterization of human crossover interference. *Am J Hum Genet*, 66(6):1911–1926, Jun 2000. doi:10.1086/302923. ↪p.66

Broman K. W., Murray J. C., Sheffield V. C., White R. L., and Weber J. L. Comprehensive human genetic maps: individual and sex-specific variation in recombination. *Am J Hum Genet*, 63(3):861–869, Sep 1998. doi:10.1086/302011. ↪p.3, 34, 36

Brown T. A. *Genomes (2nd edition)*. Wiley-Liss, 2002. ↪p.48, 49

Brown T. C. and Jiricny J. A specific mismatch repair event protects mammalian cells from loss of 5-methylcytosine. *Cell*, 50(6):945–950, Sep 1987. ↪p.41

Brown T. C. and Jiricny J. Repair of base-base mismatches in simian and human cells. *Genome*, 31(2):578–583, 1989. ↪p.40, 111

Buard J., Barthès P., Grey C., and de Massy B. Distinct histone modifications define initiation and repair of meiotic recombination in the mouse. *EMBO J*, 28(17):2616–2624, Sep 2009. doi:10.1038/emboj.2009.207. ↪p.19, 23

Buhler C., Borde V., and Lichten M. Mapping meiotic single-strand dna reveals a new landscape of dna double-strand breaks in saccharomyces cerevisiae. *PLoS Biol*, 5(12):e324, Dec 2007. doi:10.1371/journal.pbio.0050324. ↪p.20, 28, 121

Burt A., Bell G., and Harvey P. Sex differences in recombination. *Journal of Evolutionary Biology*, 4(2):259–277, 1991. ISSN 1420-9101. ↪p.30, 38

C. elegans Sequencing Consortium. Genome sequence of the nematode c. elegans: a platform for investigating biology. *Science*, 282(5396):2012–2018, Dec 1998. ↪p.33

Cammarano R., Costantini M., and Bernardi G. The isochore patterns of invertebrate genomes. *BMC Genomics*, 10:538, 2009. doi:10.1186/1471-2164-10-538. ↪p.44

Carpenter A. T. Electron microscopy of meiosis in drosophila melanogaster females: Ii. the recombination nodule–a recombination-associated structure at pachytene? *Proc Natl Acad Sci U S A*, 72(8):3186–3189, Aug 1975. ↪p.28

Castro A. and Lorca T. Exploring meiotic division in cargèse. meeting on meiotic divisions and checkpoints. *EMBO Rep*, 6(9):821–825, Sep 2005. doi:10.1038/sj.embor.7400504. ↪p.14

Charlesworth B. Fundamental concepts in genetics: effective population size and patterns of molecular evolution and variation. *Nat Rev Genet*, 10(3):195–205, Mar 2009. doi:10.1038/nrg2526. ↪p.77

Charlesworth B., Morgan M. T., and Charlesworth D. The effect of deleterious mutations on neutral molecular variation. *Genetics*, 134(4):1289–1303, Aug 1993. ↪p.77

Chen S. Y., Tsubouchi T., Rockmill B., Sandler J. S., Richards D. R., Vader G., Hochwagen A., Roeder G. S., and Fung J. C. Global analysis of the meiotic crossover landscape. *Dev Cell*, 15 (3):401–415, Sep 2008. doi:10.1016/j.devcel.2008.07.006. ↪p.21, 22

Cherry J. M., Ball C., Weng S., Juvik G., Schmidt R., Adler C., Dunn B., Dwight S., Riles L., Mortimer R. K., and Botstein D. Genetic and physical maps of saccharomyces cerevisiae. *Nature*, 387(6632 Suppl):67–73, May 1997. ↪p.31

Cheung V. G., Burdick J. T., Hirschmann D., and Morley M. Polymorphic variation in human meiotic recombination. *Am J Hum Genet*, 80(3):526–530, Mar 2007. doi:10.1086/512131. ↪p.3, 36, 55, 119

Chicken Genome Sequencing Project Consortium. A physical map of the chicken genome. *Nature*, 432(7018):761–764, Dec 2004. doi:10.1038/nature03030. ↪p.31, 43

Chimpanzee Sequencing and Analysis Consortium. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*, 437(7055):69–87, Sep 2005. ↪p.71, 109

Chowdhury R., Bois P. R. J., Feingold E., Sherman S. L., and Cheung V. G. Genetic analysis of variation in human meiotic recombination. *PLoS Genet*, 5(9):e1000648, Sep 2009. doi:10.1371/journal.pgen.1000648. ↪p.55, 57

Chua P. R. and Roeder G. S. Zip2, a meiosis-specific protein required for the initiation of chromosome synapsis. *Cell*, 93(3):349–359, May 1998. ↪p.169

Clark A. G., Wang X., and Matise T. Contrasting methods of quantifying fine structure of human recombination. *Annu Rev Genomics Hum Genet*, 11:45–64, Sep 2010. doi:10.1146/annurev-genom-082908-150031. ↪p.59, 64

Codina-Pascual M., Campillo M., Kraus J., Speicher M. R., Egozcue J., Navarro J., and Benet J. Crossover frequency and synaptonemal complex length: their variability and effects on human male meiosis. *Mol Hum Reprod*, 12(2):123–133, Feb 2006. doi:10.1093/molehr/gal007. ↪p.85

Conrad M. N., Dominguez A. M., and Dresser M. E. Ndj1p, a meiotic telomere protein required for normal chromosome synapsis and segregation in yeast. *Science*, 276(5316):1252–1255, May 1997. ↪p.27, 168

Constantinou A., Chen X.-B., McGowan C. H., and West S. C. Holliday junction resolution in human cells: two junction endonucleases with distinct substrate specificities. *EMBO J*, 21(20): 5577–5585, Oct 2002. ↪p.2, 18, 173

Cook P. R. The transcriptional basis of chromosome pairing. *J Cell Sci*, 110 (Pt 9):1033–1040, May 1997. ↪p.13

Coop G. and Przeworski M. An evolutionary view of human recombination. *Nat Rev Genet*, 8 (1):23–34, Jan 2007. doi:10.1038/nrg1947. ↪p.29, 36, 40

Coop G., Wen X., Ober C., Pritchard J. K., and Przeworski M. High-resolution mapping of crossovers reveals extensive variation in fine-scale recombination patterns among humans. *Science*, 319(5868):1395–1398, Mar 2008. doi:10.1126/science.1151851. ↪p.21, 22, 35, 36, 37, 55, 62, 83, 123

Copenhaver G. P., Housworth E. A., and Stahl F. W. Crossover interference in arabidopsis. *Genetics*, 160(4):1631–1639, Apr 2002. ↪p.28, 67

Cordaux R. and Batzer M. A. The impact of retrotransposons on human genome evolution. *Nat Rev Genet*, 10(10):691–703, Oct 2009. doi:10.1038/nrg2640. ↪p.112

Cordaux R., Lee J., Dinoso L., and Batzer M. A. Recently integrated alu retrotransposons are essentially neutral residents of the human genome. *Gene*, 373:138–144, May 2006. doi:10.1016/j.gene.2006.01.020. ↪p.112

Costantini M., Auletta F., and Bernardi G. Isochore patterns and gene distributions in fish genomes. *Genomics*, 90(3):364–371, Sep 2007a. doi:10.1016/j.ygeno.2007.05.006. ↪p.44

Costantini M., Filippo M. D., Auletta F., and Bernardi G. Isochore pattern and gene distribution in the chicken genome. *Gene*, 400(1-2):9–15, Oct 2007b. doi:10.1016/j.gene.2007.05.025. ↪p.44

Costantini M., Cammarano R., and Bernardi G. The evolution of isochore patterns in vertebrate genomes. *BMC Genomics*, 10:146, 2009. doi:10.1186/1471-2164-10-146. ↪p.44

Cox A., Ackert-Bicknell C. L., Dumont B. L., Ding Y., Bell J. T., Brockmann G. A., Wergedal J. E., Bult C., Paigen B., Flint J., Tsaih S.-W., Churchill G. A., and Broman K. W. A new standard genetic map for the laboratory mouse. *Genetics*, 182(4):1335–1344, Aug 2009. doi:10.1534/genetics.109.105486. ↪p.21, 33, 38, 90, 108, 114, 117, 123

Cremer T. and Cremer C. Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nat Rev Genet*, 2(4):292–301, Apr 2001. doi:10.1038/35066075. ↪p.10

Cromie G. A., Hyppa R. W., Taylor A. F., Zakharyevich K., Hunter N., and Smith G. R. Single holliday junctions are intermediates of meiotic recombination. *Cell*, 127(6):1167–1178, Dec 2006. doi:10.1016/j.cell.2006.09.050. ↪p.18, 28

Cromie G. A., Hyppa R. W., Cam H. P., Farah J. A., Grewal S. I. S., and Smith G. R. A discrete class of intergenic dna dictates meiotic dna break hotspots in fission yeast. *PLoS Genet*, 3(8): e141, Aug 2007. doi:10.1371/journal.pgen.0030141. ↪p.20

Danilowicz C., Lee C. H., Kim K., Hatch K., Coljee V. W., Kleckner N., and Prentiss M. Single molecule detection of direct, homologous, dna/dna pairing. *Proc Natl Acad Sci U S A*, 106 (47):19824–19829, Nov 2009. doi:10.1073/pnas.0911214106. ↪p.13

DB-AceDB, 2010. URL `ftp://ftp.sanger.ac.uk/pub/acedb/celegans/`. ↪p.33, 90

DB-DogMap. Canine genetic linkage map, 2008. URL `http://www.vgl.ucdavis.edu/dogmap/`. ↪p.90, 114

DB-Ensembl. Ensembl project, 2010. URL `http://www.ensembl.org`. ↪p.31, 32, 33, 95, 113, 115

DB-FishMap. FishMap2.0 : A community resource for Zebrafish Genomics (Zv8 Update), 2010. URL `http://fishmap2.igib.res.in/`. ↪p.31

DB-FlyBase, 2010. URL `http://flybase.org/maps/chromosomes/maps.html`. ↪p.33, 90

DB-MaizeSequence. The Maize Genome Sequencing Project, 2010. URL `http://beta.maizesequence.org/Zea_mays/Info/Index`. ↪p.32

DB-MedakaMap. Utgb medaka genome browser, 2010. URL `http://medaka.utgenome.org/`. ↪p.90

DB-NCBI. National center for biotechnology information, 2010. URL `http://www.ncbi.nlm.nih.gov/`. ↪p.xiii, 31, 32, 33, 90, 91

DB-SGD. Saccharomyces genome database, 2010. URL `http://www.yeastgenome.org/pgMaps/pgI.shtml`. ↪p.31, 90

DB-ZFIN, 2010. URL `http://zfin.org/cgi-bin/mapper_select.cgi`. ↪p.31, 90, 91, 95, 102

de Boer E., Stam P., Dietrich A. J. J., Pastink A., and Heyting C. Two levels of interference in mouse meiotic recombination. *Proc Natl Acad Sci U S A*, 103(25):9607–9612, Jun 2006. doi:10.1073/pnas.0600418103. ↪p.2, 20, 25, 28, 69, 98, 99, 123

de Boer E., Dietrich A. J. J., Höög C., Stam P., and Heyting C. Meiotic interference among mlh1 foci requires neither an intact axial element structure nor full synapsis. *J Cell Sci*, 120(Pt 5): 731–736, Mar 2007. doi:10.1242/jcs.003186. ↪p.28

de los Santos T., Hunter N., Lee C., Larkin B., Loidl J., and Hollingsworth N. M. The mus81/mms4 endonuclease acts independently of double-holliday junction resolution to promote a distinct subset of crossovers during meiosis in budding yeast. *Genetics*, 164(1):81–94, May 2003. ↪p.18, 28

de Massy B., Rocco V., and Nicolas A. The nucleotide mapping of dna double-strand breaks at the cys3 initiation site of meiotic recombination in saccharomyces cerevisiae. *EMBO J*, 14(18): 4589–4598, Sep 1995. ↪p.19

de Villena F. P.-M. and Sapienza C. Recombination is proportional to the number of chromosome arms in mammals. *Mamm Genome*, 12(4):318–322, Apr 2001. doi:10.1007/s003350020005. ↪p.25, 29, 69, 84, 95

Dehal P., Satou Y., Campbell R. K., and et al. J. C. The draft genome of ciona intestinalis: insights into chordate and vertebrate origins. *Science*, 298(5601):2157–2167, Dec 2002. doi:10.1126/science.1080049. ↪p.32

Dernburg A. F., Sedat J. W., and Hawley R. S. Direct evidence of a role for heterochromatin in meiotic chromosome segregation. *Cell*, 86(1):135–146, Jul 1996. ↪p.13

Dernburg A. F., McDonald K., Moulder G., Barstead R., Dresser M., and Villeneuve A. M. Meiotic recombination in c. elegans initiates by a conserved mechanism and is dispensable for homologous chromosome synapsis. *Cell*, 94(3):387–398, Aug 1998. ↪p.13

Ding D.-Q., Haraguchi T., and Hiraoka Y. From meiosis to postmeiotic events: alignment and recognition of homologous chromosomes in meiosis. *FEBS J*, 277(3):565–570, Feb 2010. doi:10.1111/j.1742-4658.2009.07501.x. ↪p.19

Doligez A., Adam-Blondon A. F., Cipriani G., Gaspero G. D., Laucou V., Merdinoglu D., Meredith C. P., Riaz S., Roux C., and This P. An integrated ssr map of grapevine based on five mapping populations. *Theor Appl Genet*, 113(3):369–382, Aug 2006. doi:10.1007/s00122-006-0295-1. ↪p.32, 90, 95

Dreszer T. R., Wall G. D., Haussler D., and Pollard K. S. Biased clustered substitutions in the human genome: the footprints of male-driven biased gene conversion. *Genome Res*, 17(10): 1420–1430, Oct 2007. doi:10.1101/gr.6395807. ↪p.42, 111

Drouaud J., Camilleri C., Bourguignon P.-Y., Canaguier A., Bérard A., Vezon D., Giancola S., Brunel D., Colot V., Prum B., Quesneville H., and Mézard C. Variation in crossing-over rates across chromosome 4 of arabidopsis thaliana reveals the presence of meiotic recombination "hot spots". *Genome Res*, 16(1):106–114, Jan 2006. doi:10.1101/gr.4319006. ↪p.22, 44, 98, 99

Drouaud J., Mercier R., Chelysheva L., Bérard A., Falque M., Martin O., Zanni V., Brunel D., and Mézard C. Sex-specific crossover distributions and variations in interference level along arabidopsis thaliana chromosome 4. *PLoS Genet*, 3(6):e106, Jun 2007. doi:10.1371/journal.pgen.0030106. ↪p.28, 30, 70, 83, 98, 99

Duke S. E., Samollow P. B., Mauceli E., Lindblad-Toh K., and Breen M. Integrated cytogenetic bac map of the genome of the gray, short-tailed opossum, monodelphis domestica. *Chromosome Res*, 15(3):361–370, 2007. doi:10.1007/s10577-007-1131-4. ↪p.115

Dumont B. L., White M. A., Steffy B., Wiltshire T., and Payseur B. A. Extensive recombination rate variation in the house mouse species complex inferred from genetic linkage maps. *Genome Res*, Oct 2010. doi:10.1101/gr.111252.110. ↪p.30

Dunham I., Beare D. M., and Collins J. E. The characteristics of human genes: analysis of human chromosome 22. *Comp Funct Genomics*, 4(6):635–646, 2003. doi:10.1002/cfg.335. ↪p.44

Duret L. and Arndt P. F. The impact of recombination on nucleotide substitutions in the human

genome. *PLoS Genet*, 4(5):e1000071, May 2008. doi:10.1371/journal.pgen.1000071. ↪p.4, 5, 42, 44, 71, 75, 76, 77, 109, 111, 115, 122, 124, 126, 127, 130

Duret L. and Galtier N. Biased gene conversion and the evolution of mammalian genomic landscapes. *Annu Rev Genomics Hum Genet*, 10:285–311, 2009. doi:10.1146/annurev-genom-082908-150001. ↪p.3, 40, 41, 42, 45

Duret L., Mouchiroud D., and Gautier C. Statistical analysis of vertebrate sequences reveals that long genes are scarce in gc-rich isochores. *J Mol Evol*, 40(3):308–317, Mar 1995. ↪p.44

Duret L., Semon M., Piganeau G., Mouchiroud D., and Galtier N. Vanishing gc-rich isochores in mammalian genomes. *Genetics*, 162(4):1837–1847, Dec 2002. ↪p.45, 77

Egel-Mitani M., Olson L. W., and Egel R. Meiosis in aspergillus nidulans: another example for lacking synaptonemal complexes in the absence of crossover interference. *Hereditas*, 97(2): 179–187, 1982. ↪p.28

Eijpe M., Offenberg H., Jessberger R., Revenkova E., and Heyting C. Meiotic cohesin rec8 marks the axial elements of rat synaptonemal complexes before cohesins smc1beta and smc3. *J Cell Biol*, 160(5):657–670, Mar 2003. doi:10.1083/jcb.200212080. ↪p.13

Elferink M. G., van As P., Veenendaal T., Crooijmans R. P. M. A., and Groenen M. A. M. Regional differences in recombination hotspots between two chicken populations. *BMC Genet*, 11:11, 2010. doi:10.1186/1471-2156-11-11. ↪p.5, 106

Ellegren H. Microsatellites: simple sequences with complex evolution. *Nat Rev Genet*, 5(6): 435–445, Jun 2004. doi:10.1038/nrg1348. ↪p.49

Eyre-Walker A. Recombination and mammalian genome evolution. *Proc Biol Sci*, 252(1335): 237–243, Jun 1993. doi:10.1098/rspb.1993.0071. ↪p.44, 45

Eyre-Walker A. and Hurst L. D. The evolution of isochores. *Nat Rev Genet*, 2(7):549–555, Jul 2001. doi:10.1038/35080577. ↪p.44, 45, 109

Falque M., Mercier R., Mézard C., de Vienne D., and Martin O. C. Patterns of recombination and mlh1 foci density along mouse chromosomes: modeling effects of interference and obligate chiasma. *Genetics*, 176(3):1453–1467, Jul 2007. doi:10.1534/genetics.106.070235. ↪p.67

Falque M., Anderson L. K., Stack S. M., Gauthier F., and Martin O. C. Two types of meiotic crossovers coexist in maize. *Plant Cell*, 21(12):3915–3925, Dec 2009. doi:10.1105/tpc.109.071514. ↪p.28, 67, 68, 95, 98

Fan Q. Q., Xu F., White M. A., and Petes T. D. Competition between adjacent meiotic recombination hotspots in the yeast saccharomyces cerevisiae. *Genetics*, 145(3):661–670, Mar 1997. ↪p.20

Federico C., Saccone S., and Bernardi G. The gene-richest bands of human chromosomes replicate at the onset of the s-phase. *Cytogenet Cell Genet*, 80(1-4):83–88, 1998. ↪p.44

Feingold J. S. M. Fellous M. *Principes de génétique humaine*. Hermann, Éditeurs des sciences et des arts, 1998. ↪p.49, 64

Fekairi S., Scaglione S., Chahwan C., Taylor E. R., Tissier A., Coulon S., Dong M.-Q., Ruse C., Yates J. R., Russell P., Fuchs R. P., McGowan C. H., and Gaillard P.-H. L. Human slx4 is a holliday junction resolvase subunit that binds multiple dna repair/recombination endonucleases. *Cell*, 138(1):78–89, Jul 2009. doi:10.1016/j.cell.2009.06.029. ↪p.17, 173

Felsenstein J. Evolutionary trees from dna sequences: a maximum likelihood approach. *J Mol Evol*, 17(6):368–376, 1981. ↪p.74

Fledel-Alon A., Wilson D. J., Broman K., Wen X., Ober C., Coop G., and Przeworski M.

Broad-scale recombination patterns underlying proper disjunction in humans. *PLoS Genet*, 5 (9):e1000658, Sep 2009. doi:10.1371/journal.pgen.1000658. ↪p.29, 36, 67, 95, 98

Foss E., Lande R., Stahl F. W., and Steinberg C. M. Chiasma interference as a function of genetic distance. *Genetics*, 133(3):681–691, Mar 1993. ↪p.x, 4, 28, 64, 65, 66, 67, 82, 99

Foss E. J. and Stahl F. W. A test of a counting model for chiasma interference. *Genetics*, 139(3): 1201–1209, Mar 1995. ↪p.66

Fridkin A., Penkner A., Jantsch V., and Gruenbaum Y. Sun-domain and kash-domain proteins during development, meiosis and disease. *Cell Mol Life Sci*, 66(9):1518–1533, May 2009. doi:10.1007/s00018-008-8713-y. ↪p.168

Froenicke L., Anderson L. K., Wienberg J., and Ashley T. Male mouse recombination maps for each autosome identified by chromosome painting. *Am J Hum Genet*, 71(6):1353–1368, Dec 2002. doi:10.1086/344714. ↪p.84, 85, 98, 99

Fryxell K. J. and Zuckerkandl E. Cytosine deamination plays a primary role in the evolution of mammalian isochores. *Mol Biol Evol*, 17(9):1371–1383, Sep 2000. ↪p.41, 110

Fukagawa T., Sugaya K., Matsumoto K., Okumura K., Ando A., Inoko H., and Ikemura T. A boundary of long-range g + clocus: pseudoautosomal boundary-like sequence exists near the boundary. *Genomics*, 25(1):184–191, Jan 1995. ↪p.44

Fullerton S. M., Carvalho A. B., and Clark A. G. Local rates of recombination are positively correlated with gc content in the human genome. *Mol Biol Evol*, 18(6):1139–1142, Jun 2001. ↪p.44, 109

Fung J. C., Rockmill B., Odell M., and Roeder G. S. Imposition of crossover interference through the nonrandom distribution of synapsis initiation complexes. *Cell*, 116(6):795–802, Mar 2004. ↪p.25, 69

Gaillard P.-H. L., Noguchi E., Shanahan P., and Russell P. The endogenous mus81-eme1 complex resolves holliday junctions by a nick and counternick mechanism. *Mol Cell*, 12(3):747–759, Sep 2003. ↪p.18

Galtier N. Gene conversion drives gc content evolution in mammalian histones. *Trends Genet*, 19(2):65–68, Feb 2003. ↪p.42

Galtier N. and Duret L. Adaptation or biased gene conversion? extending the null hypothesis of molecular evolution. *Trends Genet*, 23(6):273–277, Jun 2007. doi:10.1016/j.tig.2007.03.011. ↪p.43, 45, 77

Galtier N. and Duret L. Biased gene conversion and its impact on human genome evolution. *Wiley Online Library*, 2008. ↪p.45

Galtier N., Duret L., Glémin S., and Ranwez V. Gc-biased gene conversion promotes the fixation of deleterious amino acid changes in primates. *Trends Genet*, 25(1):1–5, Jan 2009. doi:10.1016/j.tig.2008.10.011. ↪p.45, 77

Gardner M. J., Hall N., Fung E., White O., Berriman M., Hyman R. W., Carlton J. M., Pain A., Nelson K. E., Bowman S., Paulsen I. T., James K., Eisen J. A., Rutherford K., Salzberg S. L., Craig A., Kyes S., Chan M.-S., Nene V., Shallom S. J., Suh B., Peterson J., Angiuoli S., Pertea M., Allen J., Selengut J., Haft D., Mather M. W., Vaidya A. B., Martin D. M. A., Fairlamb A. H., Fraunholz M. J., Roos D. S., Ralph S. A., McFadden G. I., Cummings L. M., Subramanian G. M., Mungall C., Venter J. C., Carucci D. J., Hoffman S. L., Newbold C., Davis R. W., Fraser C. M., and Barrell B. Genome sequence of the human malaria parasite plasmodium falciparum. *Nature*, 419(6906):498–511, Oct 2002. doi:10.1038/nature01097. ↪p.32

Gauthier F., Martin O. C., and Falque M. Coda (crossover distribution analyzer): quantitative characterization of crossover position patterns along chromosomes. *BMC Bioinformatics*, 12: 27, 2011. doi:10.1186/1471-2105-12-27. ↪p.68

Gautier M., Faraut T., Moazami-Goudarzi K., Navratil V., Foglio M., Grohs C., Boland A., Garnier J. G., Boichard D., Lathrop G. M., Gut I. G., and Eggen A. Genetic and haplotypic structure in 14 european and african cattle breeds. *Genetics*, 177(2):1059–1070, Oct 2007. doi:10.1534/genetics.107.075804. ↪p.61

Gay J., Myers S., and McVean G. Estimating meiotic gene conversion rates from population genetic data. *Genetics*, 177(2):881–894, Oct 2007. doi:10.1534/genetics.107.078907. ↪p.62

Gerton J. L. and Hawley R. S. Homologous chromosome interactions in meiosis: diversity amidst conservation. *Nat Rev Genet*, 6(6):477–487, Jun 2005. doi:10.1038/nrg1614. ↪p.2, 12

Gerton J. L., DeRisi J., Shroff R., Lichten M., Brown P. O., and Petes T. D. Inaugural article: global mapping of meiotic recombination hotspots and coldspots in the yeast saccharomyces cerevisiae. *Proc Natl Acad Sci U S A*, 97(21):11383–11390, Oct 2000. doi:10.1073/pnas.97.21.11383. ↪p.19, 20, 21, 44, 109

Getz T. J., Banse S. A., Young L. S., Banse A. V., Swanson J., Wang G. M., Browne B. L., Foss H. M., and Stahl F. W. Reduced mismatch repair of heteroduplexes reveals "non"-interfering crossing over in wild-type saccharomyces cerevisiae. *Genetics*, 178(3):1251–1269, Mar 2008. doi:10.1534/genetics.106.067603. ↪p.27

Glémin S. Surprising fitness consequences of gc-biased gene conversion: I. mutation load and inbreeding depression. *Genetics*, 185(3):939–959, Jul 2010. doi:10.1534/genetics.110.116368. ↪p.45, 79

Glémin S., Bazin E., and Charlesworth D. Impact of mating systems on patterns of sequence polymorphism in flowering plants. *Proc Biol Sci*, 273(1604):3011–3019, Dec 2006. doi:10.1098/rspb.2006.3657. ↪p.45

Goedecke W., Eijpe M., Offenberg H. H., van Aalderen M., and Heyting C. Mre11 and ku70 interact in somatic cells, but are differentially expressed in early meiosis. *Nat Genet*, 23(2): 194–198, Oct 1999. doi:10.1038/13821. ↪p.14

Gregory S. G., Sekhon M., Schein J., Zhao S., Osoegawa K., Scott C. E., Evans R. S., Burridge P. W., Cox T. V., Fox C. A., Hutton R. D., Mullenger I. R., Phillips K. J., Smith J., Stalker J., Threadgold G. J., Birney E., Wylie K., Chinwalla A., Wallis J., Hillier L., Carter J., Gaige T., Jaeger S., Kremitzki C., Layman D., Maas J., McGrane R., Mead K., Walker R., Jones S., Smith M., Asano J., Bosdet I., Chan S., Chittaranjan S., Chiu R., Fjell C., Fuhrmann D., Girn N., Gray C., Guin R., Hsiao L., Krzywinski M., Kutsche R., Lee S. S., Mathewson C., McLeavy C., Messervier S., Ness S., Pandoh P., Prabhu A.-L., Saeedi P., Smailus D., Spence L., Stott J., Taylor S., Terpstra W., Tsai M., Vardy J., Wye N., Yang G., Shatsman S., Ayodeji B., Geer K., Tsegaye G., Shvartsbeyn A., Gebregeorgis E., Krol M., Russell D., Overton L., Malek J. A., Holmes M., Heaney M., Shetty J., Feldblyum T., Nierman W. C., Catanese J. J., Hubbard T., Waterston R. H., Rogers J., de Jong P. J., Fraser C. M., Marra M., McPherson J. D., and Bentley D. R. A physical map of the mouse genome. *Nature*, 418 (6899):743–750, Aug 2002. doi:10.1038/nature00957. ↪p.33, 45

Grey C., Baudat F., and De Massy B. Genome-wide control of the distribution of meiotic recombination. *PLoS Biol*, 7(2):e1000035, 2009. ↪p.23

Groenen M. A. M., Wahlberg P., Foglio M., Cheng H. H., Megens H.-J., Crooijmans R. P. M. A., Besnier F., Lathrop M., Muir W. M., Wong G. K.-S., Gut I., and Andersson L. A high-density snp-based linkage map of the chicken genome reveals sequence features correlated

with recombination rate. *Genome Res*, 19(3):510–519, Mar 2009. doi:10.1101/gr.086538.108. ↪p.31, 34, 38, 43, 86, 90, 105, 106, 111, 114, 123

Guillon H., Baudat F., Grey C., Liskay R. M., and de Massy B. Crossover and noncrossover pathways in mouse meiosis. *Mol Cell*, 20(4):563–573, Nov 2005. doi:10.1016/j.molcel.2005.09.021. ↪p.19

Guryev V., Smits B. M. G., van de Belt J., Verheul M., Hubner N., and Cuppen E. Haplotype block structure is conserved across mammals. *PLoS Genet*, 2(7):e121, Jul 2006. doi:10.1371/journal.pgen.0020121. ↪p.61

Haber J. E. Partners and pathwaysrepairing a double-strand break. *Trends Genet*, 16(6):259–264, Jun 2000. ↪p.14

Haldane J. The combination of linkage values and the calculation of distances between the loci of linked factors. *J. Genet*, 8(29):309, 1919. ↪p.47, 54

Hamada H. and Kakunaga T. Potential z-dna forming sequences are highly dispersed in the human genome. *Nature*, 298(5872):396–398, Jul 1982. ↪p.49

Hamada H., Petrino M. G., and Kakunaga T. A novel repeated element with z-dna-forming potential is widely found in evolutionarily diverse eukaryotic genomes. *Proc Natl Acad Sci U S A*, 79(21):6465–6469, Nov 1982. ↪p.49

Hammarlund M., Davis M. W., Nguyen H., Dayton D., and Jorgensen E. M. Heterozygous insertions alter crossover distribution but allow crossover interference in caenorhabditis elegans. *Genetics*, 171(3):1047–1056, Nov 2005. doi:10.1534/genetics.105.044834. ↪p.95

HapMap. International hapmap project, 2010. URL `http://snp.cshl.org/abouthapmap.html`. ↪p.49

Harushima Y., Yano M., Shomura A., Sato M., Shimano T., Kuboki Y., Yamamoto T., Lin S. Y., Antonio B. A., Parco A., Kajiya H., Huang N., Yamamoto K., Nagamura Y., Kurata N., Khush G. S., and Sasaki T. A high-density rice genetic linkage map with 2275 markers using a single f2 population. *Genetics*, 148(1):479–494, Jan 1998. ↪p.32, 91

Hassold T. and Hunt P. To err (meiotically) is human: the genesis of human aneuploidy. *Nat Rev Genet*, 2(4):280–291, Apr 2001. doi:10.1038/35066065. ↪p.35

Hassold T., Judis L., Chan E. R., Schwartz S., Seftel A., and Lynn A. Cytological studies of meiotic recombination in human males. *Cytogenet Genome Res*, 107(3-4):249–255, 2004. doi:10.1159/000080602. ↪p.29

Hassold T., Hansen T., Hunt P., and VandeVoort C. Cytological studies of recombination in rhesus males. *Cytogenet Genome Res*, 124(2):132–138, 2009. doi:10.1159/000207519. ↪p.99

Hayman D., Smith M., and Rodger J. A comparative study of chiasmata in male and femaleBettongia penicillata (Marsupialia). *Genetica*, 83(1):45–49, 1990. ISSN 0016-6707. ↪p.30

Hellenthal G. and Stephens M. Insights into recombination from population genetic variation. *Curr Opin Genet Dev*, 16(6):565–572, Dec 2006. doi:10.1016/j.gde.2006.10.001. ↪p.47, 62

Hershberg R. and Petrov D. A. Evidence that mutation is universally biased towards at in bacteria. *PLoS Genet*, 6(9), Sep 2010. doi:10.1371/journal.pgen.1001115. URL `http://dx.doi.org/10.1371/journal.pgen.1001115`. ↪p.126

Hess S. T., Blake J. D., and Blake R. D. Wide variations in neighbor-dependent substitution rates. *J Mol Biol*, 236(4):1022–1033, Mar 1994. ↪p.73

Heuertz M., Paoli E. D., Källman T., Larsson H., Jurman I., Morgante M., Lascoux M., and Gyllenstrand N. Multilocus patterns of nucleotide diversity, linkage disequilibrium and

demographic history of norway spruce [picea abies (l.) karst]. *Genetics*, 174(4):2095–2105, Dec 2006. doi:10.1534/genetics.106.065102. ↪p.61

Higgins J. D., Armstrong S. J., Franklin F. C. H., and Jones G. H. The arabidopsis muts homolog atmsh4 functions at an early step in recombination: evidence for two classes of recombination in arabidopsis. *Genes Dev*, 18(20):2557–2570, Oct 2004. doi:10.1101/gad.317504. ↪p.28

Higgins J. D., Buckling E. F., Franklin F. C. H., and Jones G. H. Expression and functional analysis of atmus81 in arabidopsis meiosis reveals a role in the second pathway of crossing-over. *Plant J*, 54(1):152–162, Apr 2008. doi:10.1111/j.1365-313X.2008.03403.x. ↪p.26, 27, 28

Hill W. and Robertson A. Linkage disequilibrium in finite populations. *TAG Theoretical and Applied Genetics*, 38(6):226–231, 1968. ISSN 0040-5752. ↪p.59

Holliday R. A mechanism for gene conversion in fungi. *Genetics Research*, 5(02):282–304, 1964. ↪p.15, 40

Hollingsworth N. M. and Brill S. J. The mus81 solution to resolution: generating meiotic crossovers without holliday junctions. *Genes Dev*, 18(2):117–125, Jan 2004. doi:10.1101/gad.1165904. ↪p.28

Hollingsworth N. M. and Byers B. Hop1: a yeast meiotic pairing gene. *Genetics*, 121(3):445–462, Mar 1989. ↪p.169

Holloway J. K., Booth J., Edelmann W., McGowan C. H., and Cohen P. E. Mus81 generates a subset of mlh1-mlh3-independent crossovers in mammalian meiosis. *PLoS Genet*, 4(9): e1000186, 2008. doi:10.1371/journal.pgen.1000186. ↪p.18, 19, 28

Holloway J. K., Morelli M. A., Borst P. L., and Cohen P. E. Mammalian blm helicase is critical for integrating multiple pathways of meiotic recombination. *J Cell Biol*, 188(6):779–789, Mar 2010. doi:10.1083/jcb.200909048. ↪p.19

Holloway K., Lawson V. E., and Jeffreys A. J. Allelic recombination and de novo deletions in sperm in the human beta-globin gene region. *Hum Mol Genet*, 15(7):1099–1111, Apr 2006. doi:10.1093/hmg/ddl025. ↪p.174

Holmquist G. P. Chromosome bands, their chromatin flavors, and their functional features. *Am J Hum Genet*, 51(1):17–37, Jul 1992. ↪p.45

Honeybee Genome Sequencing Consortium. Insights into social insects from the genome of the honeybee Apis mellifera. *Nature*, 443:931–949, 2006. ↪p.31

Horse Genome Sequencing Project Consortium. Genome sequence, comparative analysis, and population genetics of the domestic horse. *Science*, 326(5954):865–867, Nov 2009. doi:10.1126/science.1178158. ↪p.31

Housworth E. A. and Stahl F. W. Crossover interference in humans. *Am J Hum Genet*, 73(1): 188–197, Jul 2003. doi:10.1086/376610. ↪p.x, 66, 67

Huang Y., Zhao Y., Haley C. S., Hu S., Hao J., Wu C., and Li N. A genetic and cytogenetic map for the duck (anas platyrhynchos). *Genetics*, 173(1):287–296, May 2006. doi:10.1534/genetics.105.053256. ↪p.38

Hudson R. R. and Kaplan N. L. Statistical properties of the number of recombination events in the history of a sample of dna sequences. *Genetics*, 111(1):147–164, Sep 1985. ↪p.59

Hughes S., Zelus D., and Mouchiroud D. Warm-blooded isochore structure in nile crocodile and turtle. *Mol Biol Evol*, 16(11):1521–1527, Nov 1999. ↪p.44, 45

Hunt P. A. and Hassold T. J. Sex matters in meiosis. *Science*, 296(5576):2181–2183, Jun 2002. doi:10.1126/science.1071907. ↪p.27

Hunt P. A. and Hassold T. J. Human female meiosis: what makes a good egg go bad? *Trends Genet*, 24(2):86–93, Feb 2008. doi:10.1016/j.tig.2007.11.010. ↪p.36

Hunter N. and Kleckner N. The single-end invasion: an asymmetric intermediate at the double-strand break to double-holliday junction transition of meiotic recombination. *Cell*, 106(1): 59–70, Jul 2001. ↪p.14, 19, 25

Initiative A. G. Analysis of the genome sequence of the flowering plant arabidopsis thaliana. *Nature*, 408(6814):796–815, Dec 2000. ↪p.33

International Chicken Genome Sequencing Consortium. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature*, 432(7018): 695–716, Dec 2004. ↪p.86, 119

International HapMap 3 Consortium, Altshuler D. M., Gibbs R. A., Peltonen L., Altshuler D. M., Gibbs R. A., Peltonen L., Dermitzakis E., Schaffner S. F., Yu F., Peltonen L., Dermitzakis E., Bonnen P. E., Altshuler D. M., Gibbs R. A., de Bakker P. I. W., Deloukas P., Gabriel S. B., Gwilliam R., Hunt S., Inouye M., Jia X., Palotie A., Parkin M., Whittaker P., Yu F., Chang K., Hawes A., Lewis L. R., Ren Y., Wheeler D., Gibbs R. A., Muzny D. M., Barnes C., Darvishi K., Hurles M., Korn J. M., Kristiansson K., Lee C., McCarrol S. A., Nemesh J., Dermitzakis E., Keinan A., Montgomery S. B., Pollack S., Price A. L., Soranzo N., Bonnen P. E., Gibbs R. A., Gonzaga-Jauregui C., Keinan A., Price A. L., Yu F., Anttila V., Brodeur W., Daly M. J., Leslie S., McVean G., Moutsianas L., Nguyen H., Schaffner S. F., Zhang Q., Ghori M. J. R., McGinnis R., McLaren W., Pollack S., Price A. L., Schaffner S. F., Takeuchi F., Grossman S. R., Shlyakhter I., Hostetter E. B., Sabeti P. C., Adebamowo C. A., Foster M. W., Gordon D. R., Licinio J., Manca M. C., Marshall P. A., Matsuda I., Ngare D., Wang V. O., Reddy D., Rotimi C. N., Royal C. D., Sharp R. R., Zeng C., Brooks L. D., and McEwen J. E. Integrating common and rare genetic variation in diverse human populations. *Nature*, 467(7311):52–58, Sep 2010. doi:10.1038/nature09298. ↪p.60

International HapMap Consortium. A haplotype map of the human genome. *Nature*, 437(7063): 1299–1320, Oct 2005. ↪p.59, 60, 62

International HapMap Consortium. A second generation human haplotype map of over 3.1 million snps. *Nature*, 449(7164):851–861, Oct 2007. doi:10.1038/nature06258. ↪p.22, 24, 37, 60, 79

International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, Feb 2001. ↪p.31, 44, 49, 124

Ip S. C. Y., Rass U., Blanco M. G., Flynn H. R., Skehel J. M., and West S. C. Identification of holliday junction resolvases from humans and yeast. *Nature*, 456(7220):357–361, Nov 2008. doi:10.1038/nature07470. ↪p.17, 172

Ira G., Malkova A., Liberi G., Foiani M., and Haber J. E. Srs2 and sgs1-top3 suppress crossovers during double-strand break repair in yeast. *Cell*, 115(4):401–411, Nov 2003. ↪p.173

Jackson A. P., Sanders M., Berry A., McQuillan J., Aslett M. A., Quail M. A., Chukualim B., Capewell P., MacLeod A., Melville S. E., Gibson W., Barry J. D., Berriman M., and Hertz-Fowler C. The genome sequence of trypanosoma brucei gambiense, causative agent of chronic human african trypanosomiasis. *PLoS Negl Trop Dis*, 4(4):e658, 2010. doi:10.1371/journal.pntd.0000658. ↪p.32

Jaillon O., Aury J., Noel B., Policriti A., Clepet C., Casagrande A., Choisne N., Aubourg S., Vitulo N., Jubin C., et al. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature*, 449(7161):463–467, 2007. ISSN 0028-0836. ↪p.32

Jeffreys A. J. and May C. A. Intense and highly localized gene conversion activity in human meiotic crossover hot spots. *Nat Genet*, 36(2):151–156, Feb 2004. doi:10.1038/ng1287. ↪p.62, 64

Jeffreys A. J. and Neumann R. Reciprocal crossover asymmetry and meiotic drive in a human recombination hot spot. *Nat Genet*, 31(3):267–271, Jul 2002. doi:10.1038/ng910. ↪p.37, 39, 40, 174

Jeffreys A. J. and Neumann R. Factors influencing recombination frequency and distribution in a human meiotic crossover hotspot. *Hum Mol Genet*, 14(15):2277–2287, Aug 2005. doi:10.1093/hmg/ddi232. ↪p.36, 64, 174

Jeffreys A. J. and Neumann R. The rise and fall of a human recombination hot spot. *Nat Genet*, 41(5):625–629, May 2009. doi:10.1038/ng.346. ↪p.39, 126, 175

Jeffreys A. J., Murray J., and Neumann R. High-resolution mapping of crossovers in human sperm defines a minisatellite-associated recombination hotspot. *Mol Cell*, 2(2):267–273, Aug 1998. ↪p.62

Jeffreys A. J., Kauppi L., and Neumann R. Intensely punctate meiotic recombination in the class ii region of the major histocompatibility complex. *Nat Genet*, 29(2):217–222, Oct 2001. doi:10.1038/ng1001-217. ↪p.39, 84, 174

Jeffreys A. J., Holloway J. K., Kauppi L., May C. A., Neumann R., Slingsby M. T., and Webb A. J. Meiotic recombination hot spots and human dna diversity. *Philos Trans R Soc Lond B Biol Sci*, 359(1441):141–152, Jan 2004. doi:10.1098/rstb.2003.1372. ↪p.2, 21, 39

Jeffreys A. J., Neumann R., Panayi M., Myers S., and Donnelly P. Human recombination hot spots hidden in regions of strong marker association. *Nat Genet*, 37(6):601–606, Jun 2005. doi:10.1038/ng1565. ↪p.84, 174

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. Comparative recombination rates in the rat, mouse, and human genomes. *Genome Res*, 14(4):528–538, Apr 2004. doi:10.1101/gr.1970304. ↪p.30, 90

Jessop L. and Lichten M. Mus81/mms4 endonuclease and sgs1 helicase collaborate to ensure proper recombination intermediate metabolism during meiosis. *Mol Cell*, 31(3):313–323, Aug 2008. doi:10.1016/j.molcel.2008.05.021. ↪p.26, 27

Jessop L., Allers T., and Lichten M. Infrequent co-conversion of markers flanking a meiotic recombination initiation site in saccharomyces cerevisiae. *Genetics*, 169(3):1353–1367, Mar 2005. doi:10.1534/genetics.104.036509. ↪p.18

Joseph I., Jia D., and Lustig A. J. Ndj1p-dependent epigenetic resetting of telomere size in yeast meiosis. *Curr Biol*, 15(3):231–237, Feb 2005. doi:10.1016/j.cub.2005.01.039. ↪p.168

Joseph S. and Kirkpatrick M. Haploid selection in animals. *Trends in Ecology & Evolution*, 19 (11):592–597, 2004. ↪p.34

Joyce E. F. and McKim K. S. When specialized sites are important for synapsis and the distribution of crossovers. *Bioessays*, 29(3):217–226, Mar 2007. doi:10.1002/bies.20531. ↪p.2, 13

Jurka J. Repbase update: a database and an electronic journal of repetitive elements. *Trends Genet*, 16(9):418–420, Sep 2000. ↪p.112

Kagawa W. and Kurumizaka H. From meiosis to postmeiotic events: uncovering the molecular roles of the meiosis-specific recombinase dmc1. *FEBS J*, 277(3):590–598, Feb 2010. doi:10.1111/j.1742-4658.2009.07503.x. ↪p.13, 171

Kai W., Kikuchi K., Fujita M., Suetake H., Fujiwara A., Yoshiura Y., Ototake M., Venkatesh B., Miyaki K., and Suzuki Y. A genetic linkage map for the tiger pufferfish, takifugu rubripes. *Genetics*, 171(1):227–238, Sep 2005. doi:10.1534/genetics.105.042051. ↪p.38

Kano S., Satoh N., and Sordino P. Primary genetic linkage maps of the ascidian, ciona intestinalis. *Zoolog Sci*, 23(1):31–39, Jan 2006. ↪p.32, 90

Karolchik D., Hinrichs A., Furey T., Roskin K., Sugnet C., Haussler D., and Kent W. The ucsc table browser data retrieval tool, 2004. URL `http://genome.ucsc.edu/`. ↪p.115

Kasahara M., Naruse K., Sasaki S., Nakatani Y., Qu W., Ahsan B., Yamada T., Nagayasu Y., Doi K., Kasai Y., Jindo T., Kobayashi D., Shimada A., Toyoda A., Kuroki Y., Fujiyama A., Sasaki T., Shimizu A., Asakawa S., Shimizu N., Hashimoto S.-I., Yang J., Lee Y., Matsushima K., Sugano S., Sakaizumi M., Narita T., Ohishi K., Haga S., Ohta F., Nomoto H., Nogata K., Morishita T., Endo T., Shin-I T., Takeda H., Morishita S., and Kohara Y. The medaka draft genome and insights into vertebrate genome evolution. *Nature*, 447(7145):714–719, Jun 2007. doi:10.1038/nature05846. ↪p.90

Kauppi L., Jeffreys A. J., and Keeney S. Where the crossovers are: recombination distributions in mammals. *Nat Rev Genet*, 5(6):413–424, Jun 2004. doi:10.1038/nrg1346. ↪p.60

Kauppi L., Stumpf M. P. H., and Jeffreys A. J. Localized breakdown in linkage disequilibrium does not always predict sperm crossover hot spots in the human mhc class ii region. *Genomics*, 86(1):13–24, Jul 2005. doi:10.1016/j.ygeno.2005.03.011. ↪p.174

Kauppi L., Jasin M., and Keeney S. Meiotic crossover hotspots contained in haplotype block boundaries of the mouse genome. *Proc Natl Acad Sci U S A*, 104(33):13396–13401, Aug 2007. doi:10.1073/pnas.0701965104. ↪p.22, 60, 61, 84

Keeney S. and Neale M. J. Initiation of meiotic recombination by formation of dna double-strand breaks: mechanism and regulation. *Biochem Soc Trans*, 34(Pt 4):523–525, Aug 2006. doi:10.1042/BST0340523. ↪p.13

Keeney S., Giroux C. N., and Kleckner N. Meiosis-specific dna double-strand breaks are catalyzed by spo11, a member of a widely conserved protein family. *Cell*, 88(3):375–384, Feb 1997. ↪p.13, 171

Kelly P. D., Chu F., Woods I. G., Ngo-Hazelett P., Cardozo T., Huang H., Kimm F., Liao L., Yan Y. L., Zhou Y., Johnson S. L., Abagyan R., Schier A. F., Postlethwait J. H., and Talbot W. S. Genetic linkage mapping of zebrafish genes and ests. *Genome Res*, 10(4):558–567, Apr 2000. ↪p.31

Kemkemer C., Kohn M., Cooper D. N., Froenicke L., Högel J., Hameister H., and Kehrer-Sawatzki H. Gene synteny comparisons between different vertebrates provide new insights into breakage and fusion events during mammalian karyotype evolution. *BMC Evol Biol*, 9:84, 2009. doi:10.1186/1471-2148-9-84. ↪p.108, 117

Kemp B., Boumil R. M., Stewart M. N., and Dawson D. S. A role for centromere pairing in meiotic chromosome segregation. *Genes Dev*, 18(16):1946–1951, Aug 2004. doi:10.1101/gad.1227304. ↪p.12

Kim J.-S., Klein P. E., Klein R. R., Price H. J., Mullet J. E., and Stelly D. M. Chromosome identification and nomenclature of sorghum bicolor. *Genetics*, 169(2):1169–1173, Feb 2005. doi:10.1534/genetics.104.035980. ↪p.32, 91

Kimura M. The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics*, 61(4):893–903, Apr 1969. ↪p.59

Kimura T., Yoshida K., Shimada A., Jindo T., Sakaizumi M., Mitani H., Naruse K., Takeda H., Inoko H., Tamiya G., and Shinya M. Genetic linkage map of medaka with polymerase chain

reaction length polymorphisms. *Gene*, 363:24–31, Dec 2005. doi:10.1016/j.gene.2005.07.043. ↪p.38, 91, 102

King J. S. and Mortimer R. K. A polymerization model of chiasma interference and corresponding computer simulation. *Genetics*, 126(4):1127–1138, Dec 1990. ↪p.65, 68, 82

Kironmai K. M., Muniyappa K., Friedman D. B., Hollingsworth N. M., and Byers B. Dna-binding activities of hop1 protein, a synaptonemal complex component from saccharomyces cerevisiae. *Mol Cell Biol*, 18(3):1424–1435, Mar 1998. ↪p.169

Kleckner N., Zickler D., Jones G. H., Dekker J., Padmore R., Henle J., and Hutchinson J. A mechanical basis for chromosome function. *Proc Natl Acad Sci U S A*, 101(34):12592–12597, Aug 2004. doi:10.1073/pnas.0402724101. ↪p.xi, 4, 65, 67, 68, 82

Kochakpour N. and Moens P. B. Sex-specific crossover patterns in zebrafish (danio rerio). *Heredity*, 100(5):489–495, May 2008. doi:10.1038/sj.hdy.6801091. ↪p.28

Kolas N. K. and Cohen P. E. Novel and diverse functions of the dna mismatch repair family in mammalian meiosis and recombination. *Cytogenet Genome Res*, 107(3-4):216–231, 2004. doi:10.1159/000080600. ↪p.172

Kong A., Gudbjartsson D. F., Sainz J., Jonsdottir G. M., Gudjonsson S. A., Richardsson B., Sigurdardottir S., Barnard J., Hallbeck B., Masson G., Shlien A., Palsson S. T., Frigge M. L., Thorgeirsson T. E., Gulcher J. R., and Stefansson K. A high-resolution recombination map of the human genome. *Nat Genet*, 31(3):241–247, Jul 2002. doi:10.1038/ng917. ↪p.3, 21, 34, 36, 105, 108, 109, 117, 119, 124

Kong A., Barnard J., Gudbjartsson D. F., Thorleifsson G., Jonsdottir G., Sigurdardottir S., Richardsson B., Jonsdottir J., Thorgeirsson T., Frigge M. L., Lamb N. E., Sherman S., Gulcher J. R., and Stefansson K. Recombination rate and reproductive success in humans. *Nat Genet*, 36(11):1203–1206, Nov 2004. doi:10.1038/ng1445. ↪p.36

Kong A., Thorleifsson G., Stefansson H., Masson G., Helgason A., Gudbjartsson D. F., Jonsdottir G. M., Gudjonsson S. A., Sverrisson S., Thorlacius T., Jonasdottir A., Hardarson G. A., Palsson S. T., Frigge M. L., Gulcher J. R., Thorsteinsdottir U., and Stefansson K. Sequence variants in the rnf212 gene associate with genome-wide recombination rate. *Science*, 319(5868): 1398–1401, Mar 2008. doi:10.1126/science.1152422. ↪p.34

Kong A., Thorleifsson G., Gudbjartsson D. F., Masson G., Sigurdsson A., Jonasdottir A., Walters G. B., Jonasdottir A., Gylfason A., Kristinsson K. T., Gudjonsson S. A., Frigge M. L., Helgason A., Thorsteinsdottir U., and Stefansson K. Fine-scale recombination rate differences between sexes, populations and individuals. *Nature*, 467(7319):1099–1103, Oct 2010. doi:10.1038/nature09525. ↪p.34, 35, 37, 83, 105

Kong X. and Matise T. C. Map-o-mat: internet-based linkage mapping. *Bioinformatics*, 21(4): 557–559, Feb 2005. doi:10.1093/bioinformatics/bti024. ↪p.53

Kornyshev A. A. and Wynveen A. The homology recognition well as an innate property of dna structure. *Proc Natl Acad Sci U S A*, 106(12):4683–4688, Mar 2009. doi:10.1073/pnas.0811208106. ↪p.13

Kosambi D. The estimation of map distances from recombination values. *Ann. Eugen*, 12(1944): 172–175, 1944. ↪p.54

Koszul R. and Kleckner N. Dynamic chromosome movements during meiosis: a way to eliminate unwanted connections? *Trends Cell Biol*, 19(12):716–724, Dec 2009. doi:10.1016/j.tcb.2009.09.007. ↪p.11

Krokan H. E., Nilsen H., Skorpen F., Otterlei M., and Slupphaug G. Base excision repair of dna in mammalian cells. *FEBS Lett*, 476(1-2):73–77, Jun 2000. ↪p.40

Kudla G., Helwak A., and Lipinski L. Gene conversion and gc-content evolution in mammalian hsp70. *Mol Biol Evol*, 21(7):1438–1444, Jul 2004. doi:10.1093/molbev/msh146. ↪p.42

Lam S. Y., Horn S. R., Radford S. J., Housworth E. A., Stahl F. W., and Copenhaver G. P. Crossover interference on nucleolus organizing region-bearing chromosomes in arabidopsis. *Genetics*, 170(2):807–812, Jun 2005. doi:10.1534/genetics.104.040055. ↪p.67

Lamb B. and Helmi S. The extent to which gene conversion can change allele frequencies in populations. *Genetical Research*, 39:199–217, 1982. ↪p.77

Lamb N. E., Yu K., Shaffer J., Feingold E., and Sherman S. L. Association between maternal age and meiotic recombination for trisomy 21. *Am J Hum Genet*, 76(1):91–99, Jan 2005. doi:10.1086/427266. ↪p.36

Lammers J. H., Offenberg H. H., van Aalderen M., Vink A. C., Dietrich A. J., and Heyting C. The gene encoding a major component of the lateral elements of synaptonemal complexes of the rat is related to x-linked lymphocyte-regulated genes. *Mol Cell Biol*, 14(2):1137–1146, Feb 1994. ↪p.169

Lange K., Zhao H., and Speed T. The Poisson-skip model of crossing-over. *The Annals of Applied Probability*, 7(2):299–313, 1997. ISSN 1050-5164. ↪p.67

Laurie C. C., Nickerson D. A., Anderson A. D., Weir B. S., Livingston R. J., Dean M. D., Smith K. L., Schadt E. E., and Nachman M. W. Linkage disequilibrium in wild mice. *PLoS Genet*, 3 (8):e144, Aug 2007. doi:10.1371/journal.pgen.0030144. ↪p.61

Leem S. H. and Ogawa H. The mre4 gene encodes a novel protein kinase homologue required for meiotic recombination in saccharomyces cerevisiae. *Nucleic Acids Res*, 20(3):449–457, Feb 1992. ↪p.170

Lees-Murdock D. J. and Walsh C. P. Dna methylation reprogramming in the germ line. *Epigenetics*, 3(1):5–13, 2008. ↪p.111, 122

Lenormand T. The evolution of sex dimorphism in recombination. *Genetics*, 163(2):811–822, Feb 2003. ↪p.34

Lenormand T. and Dutheil J. Recombination difference between sexes: a role for haploid selection. *PLoS Biol*, 3(3):e63, Mar 2005. doi:10.1371/journal.pbio.0030063. ↪p.34, 38

Lercher M. J., Smith N. G. C., Eyre-Walker A., and Hurst L. D. The evolution of isochores: evidence from snp frequency distributions. *Genetics*, 162(4):1805–1810, Dec 2002. ↪p.42

Leu J. Y., Chua P. R., and Roeder G. S. The meiosis-specific hop2 protein of s. cerevisiae ensures synapsis between homologous chromosomes. *Cell*, 94(3):375–386, Aug 1998. ↪p.170

Lewontin R. The interaction of selection and linkage. I. General considerations; heterotic models. *Genetics*, 49(1):49, 1964. ↪p.58

Lewontin R. and Kojima K. The evolutionary dynamics of complex polymorphisms. *Evolution*, 14(4):458–472, 1960. ISSN 0014-3820. ↪p.3, 57, 58

Lhuissier F. G. P., Offenberg H. H., Wittich P. E., Vischer N. O. E., and Heyting C. The mismatch repair protein mlh1 marks a subset of strongly interfering crossovers in tomato. *Plant Cell*, 19(3):862–876, Mar 2007. doi:10.1105/tpc.106.049106. ↪p.28

Li H. H., Gyllensten U. B., Cui X. F., Saiki R. K., Erlich H. A., and Arnheim N. Amplification and analysis of dna sequences in single human sperm and diploid cells. *Nature*, 335(6189): 414–417, Sep 1988. doi:10.1038/335414a0. ↪p.4, 62

Li J., Harper L. C., Golubovskaya I., Wang C. R., Weber D., Meeley R. B., McElver J., Bowen B., Cande W. Z., and Schnable P. S. Functional analysis of maize rad51 in meiosis and double-

strand break repair. *Genetics*, 176(3):1469–1482, Jul 2007. doi:10.1534/genetics.106.062604. ↪p.28

Li W. and Freudenberg J. Two-parameter characterization of chromosome-scale recombination rate. *Genome Res*, Oct 2009. doi:10.1101/gr.092676.109. ↪p.69, 70, 82, 86, 88, 91, 92, 107

Li W., He C., and Freudenberg J. A mathematical framework for examining whether a minimum number of chiasmata is required per metacentric chromosome or chromosome arm in human. *Genomics*, Dec 2010. doi:10.1016/j.ygeno.2010.11.007. ↪p.70

Lian J., Yin Y., Oliver-Bonet M., Liehr T., Ko E., Turek P., Sun F., and Martin R. H. Variation in crossover interference levels on individual chromosomes from human males. *Hum Mol Genet*, 17(17):2583–2594, Sep 2008. doi:10.1093/hmg/ddn158. ↪p.84

Lichten M. and Goldman A. S. Meiotic recombination hotspots. *Annu Rev Genet*, 29:423–444, 1995. doi:10.1146/annurev.ge.29.120195.002231. ↪p.19

Lieber M. R., Ma Y., Pannicke U., and Schwarz K. Mechanism and regulation of human non-homologous dna end-joining. *Nat Rev Mol Cell Biol*, 4(9):712–720, Sep 2003. doi:10.1038/nrm1202. ↪p.14

Lin F.-M., Lai Y.-J., Shen H.-J., Cheng Y.-H., and Wang T.-F. Yeast axial-element protein, red1, binds sumo chains to promote meiotic interhomologue recombination and chromosome synapsis. *EMBO J*, 29(3):586–596, Feb 2010. doi:10.1038/emboj.2009.362. ↪p.169

Lindblad-Toh K., Wade C. M., Mikkelsen T. S., and et al. E. K. K. Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature*, 438(7069):803–819, Dec 2005. doi:10.1038/nature04338. ↪p.31, 61

Liu J., Wu T. C., and Lichten M. The location and structure of double-strand dna breaks induced during yeast meiosis: evidence for a covalently linked dna-protein intermediate. *EMBO J*, 14 (18):4599–4608, Sep 1995. ↪p.19

Liu J. G., Yuan L., Brundell E., Björkroth B., Daneholt B., and Höög C. Localization of the n-terminus of scp1 to the central element of the synaptonemal complex and evidence for direct interactions between the n-termini of scp1 molecules organized head-to-head. *Exp Cell Res*, 226(1):11–19, Jul 1996. doi:10.1006/excr.1996.0197. ↪p.168

Liu Y. and West S. C. Happy hollidays: 40th anniversary of the holliday junction. *Nat Rev Mol Cell Biol*, 5(11):937–944, Nov 2004. doi:10.1038/nrm1502. ↪p.17, 19, 40

Lynch M. The origins of eukaryotic gene structure. *Mol Biol Evol*, 23(2):450–468, Feb 2006. doi:10.1093/molbev/msj050. URL `http://dx.doi.org/10.1093/molbev/msj050`. ↪p.100

Lynch M. *The origins of genome architecture*. {WH Freeman & Company}, 2007. ↪p.xiii, 2, 6, 129, 131

Lynn A., Koehler K. E., Judis L., Chan E. R., Cherry J. P., Schwartz S., Seftel A., Hunt P. A., and Hassold T. J. Covariation of synaptonemal complex length and mammalian meiotic exchange rates. *Science*, 296(5576):2222–2225, Jun 2002. doi:10.1126/science.1071220. ↪p.28, 35, 112

Lynn A., Ashley T., and Hassold T. Variation in human meiotic recombination. *Annu Rev Genomics Hum Genet*, 5:317–349, 2004. doi:10.1146/annurev.genom.4.070802.110217. ↪p.3, 36, 55

Macaya G., Thiery J. P., and Bernardi G. An approach to the organization of eukaryotic genomes at a macromolecular level. *J Mol Biol*, 108(1):237–254, Nov 1976. ↪p.44

MacLeod A., Tweedie A., McLellan S., Taylor S., Hall N., Berriman M., El-Sayed N. M., Hope M., Turner C. M. R., and Tait A. The genetic map and comparative analysis with the physical map

of trypanosoma brucei. *Nucleic Acids Res*, 33(21):6688–6693, 2005. doi:10.1093/nar/gki980. ↪p.32, 90

MacQueen A. J., Phillips C. M., Bhalla N., Weiser P., Villeneuve A. M., and Dernburg A. F. Chromosome sites play dual roles to establish homologous synapsis during meiosis in c. elegans. *Cell*, 123(6):1037–1050, Dec 2005. doi:10.1016/j.cell.2005.09.034. ↪p.12

Maddox J. and Cockett N. An update on sheep and goat linkage maps and other genomic resources. *Small Ruminant Research*, 70(1):4–20, 2007. ISSN 0921-4488. ↪p.105, 108

Maguires M. P. Crossover site determination and interference. *J Theor Biol*, 134(4):565–570, Oct 1988. ↪p.25

Malkova A., Swanson J., German M., McCusker J. H., Housworth E. A., Stahl F. W., and Haber J. E. Gene conversion and crossing over along the 405-kb left arm of saccharomyces cerevisiae chromosome vii. *Genetics*, 168(1):49–63, Sep 2004. doi:10.1534/genetics.104.027961. ↪p.27

Mancera E., Bourgon R., Brozzi A., Huber W., and Steinmetz L. M. High-resolution mapping of meiotic crossovers and non-crossovers in yeast. *Nature*, 454(7203):479–485, Jul 2008. doi:10.1038/nature07135. ↪p.3, 21, 22, 27, 28, 31, 40, 62, 64, 83, 95, 98, 99, 111

Marais G. Biased gene conversion: implications for genome and sex evolution. *Trends Genet*, 19 (6):330–338, Jun 2003. ↪p.40, 45

Marais G. and Charlesworth B. Genome evolution: recombination speeds up adaptive evolution. *Curr Biol*, 13(2):R68–R70, Jan 2003. ↪p.27

Marais G., Mouchiroud D., and Duret L. Does recombination improve selection on codon usage? lessons from nematode and fly complete genomes. *Proc Natl Acad Sci U S A*, 98(10):5688–5692, May 2001. doi:10.1073/pnas.091427698. ↪p.109

Marra R. E., Huang J. C., Fung E., Nielsen K., Heitman J., Vilgalys R., and Mitchell T. G. A genetic linkage map of cryptococcus neoformans variety neoformans serotype d (filobasidiella neoformans). *Genetics*, 167(2):619–631, Jun 2004. doi:10.1534/genetics.103.023408. ↪p.33, 90, 95

Marsolier-Kergoat M.-C. and Yeramian E. Gc content and recombination: reassessing the causal effects for the saccharomyces cerevisiae genome. *Genetics*, 183(1):31–38, Sep 2009. doi:10.1534/genetics.109.105049. ↪p.44, 109, 126

Martinez-Perez E. and Colaiácovo M. P. Distribution of meiotic recombination events: talking to your neighbors. *Curr Opin Genet Dev*, 19(2):105–112, Apr 2009. doi:10.1016/j.gde.2009.02.005. ↪p.27

Martini E., Diaz R. L., Hunter N., and Keeney S. Crossover homeostasis in yeast meiosis. *Cell*, 126(2):285–295, Jul 2006. doi:10.1016/j.cell.2006.05.044. ↪p.25, 67

Masson J. Y. and West S. C. The rad51 and dmc1 recombinases: a non-identical twin relationship. *Trends Biochem Sci*, 26(2):131–136, Feb 2001. ↪p.171

Matise T. C., Schroeder M. D., Chiarulli D. M., and Weeks D. E. Parallel computation of genetic likelihoods using cri-map, pvm, and a network of distributed workstations. *Hum Hered*, 45(2): 103–116, 1995. ↪p.53

Matise T. C., Chen F., Chen W., Vega F. M. D. L., Hansen M., He C., Hyland F. C. L., Kennedy G. C., Kong X., Murray S. S., Ziegle J. S., Stewart W. C. L., and Buyske S. A second-generation combined linkage physical map of the human genome. *Genome Res*, 17(12): 1783–1786, Dec 2007. doi:10.1101/gr.7156307. ↪p.31, 38, 90, 95, 105, 114, 123

May C. A., Shone A. C., Kalaydjieva L., Sajantila A., and Jeffreys A. J. Crossover clustering and

rapid decay of linkage disequilibrium in the xp/yp pseudoautosomal gene shox. *Nat Genet*, 31 (3):272–275, Jul 2002. doi:10.1038/ng918. ↪p.174

McKee B. D. The license to pair: identification of meiotic pairing sites in drosophila. *Chromosoma*, 105(3):135–141, Sep 1996. ↪p.12

McKee B. D. Homologous pairing and chromosome dynamics in meiosis and mitosis. *Biochim Biophys Acta*, 1677(1-3):165–180, Mar 2004. doi:10.1016/j.bbaexp.2003.11.017. ↪p.10

McKim K. S., Howell A. M., and Rose A. M. The effects of translocations on recombination frequency in caenorhabditis elegans. *Genetics*, 120(4):987–1001, Dec 1988. ↪p.12

McKim K. S., Green-Marroquin B. L., Sekelsky J. J., Chin G., Steinberg C., Khodosh R., and Hawley R. S. Meiotic synapsis in the absence of recombination. *Science*, 279(5352):876–878, Feb 1998. ↪p.13

McMahill M. S., Sham C. W., and Bishop D. K. Synthesis-dependent strand annealing in meiosis. *PLoS Biol*, 5(11):e299, Nov 2007. doi:10.1371/journal.pbio.0050299. ↪p.18

McRae A. F., McEwan J. C., Dodds K. G., Wilson T., Crawford A. M., and Slate J. Linkage disequilibrium in domestic sheep. *Genetics*, 160(3):1113–1122, Mar 2002. ↪p.61

McVean G. The structure of linkage disequilibrium around a selective sweep. *Genetics*, 175(3): 1395–1406, Mar 2007. doi:10.1534/genetics.106.062828. ↪p.61

McVean G. and Myers S. PRDM9 marks the spot. *Nature genetics*, 42(10):821, 2010. ISSN 1546-1718. ↪p.23

Melodelima C. and Gautier C. The gc-heterogeneity of teleost fishes. *BMC Genomics*, 9:632, 2008. doi:10.1186/1471-2164-9-632. ↪p.44

Menotti-Raymond M., David V. A., Schäffer A. A., Tomlin J. F., Eizirik E., Phillip C., Wells D., Pontius J. U., Hannah S. S., and O'Brien S. J. An autosomal genetic linkage map of the domestic cat, felis silvestris catus. *Genomics*, 93(4):305–313, Apr 2009. doi:10.1016/j.ygeno.2008.11.004. ↪p.38

Merker J. D., Dominska M., and Petes T. D. Patterns of heteroduplex formation associated with the initiation of meiotic recombination in the yeast saccharomyces cerevisiae. *Genetics*, 165(1): 47–63, Sep 2003. ↪p.18

Meunier J. and Duret L. Recombination drives the evolution of gc-content in the human genome. *Mol Biol Evol*, 21(6):984–990, Jun 2004. doi:10.1093/molbev/msh070. ↪p.5, 41, 44, 71, 75, 109, 111, 122

Meuwissen R. L., Offenberg H. H., Dietrich A. J., Riesewijk A., van Iersel M., and Heyting C. A coiled-coil related protein specific for synapsed regions of meiotic prophase chromosomes. *EMBO J*, 11(13):5091–5100, Dec 1992. ↪p.168

Meuwissen R. L., Meerts I., Hoovers J. M., Leschot N. J., and Heyting C. Human synaptonemal complex protein 1 (scp1): isolation and characterization of the cdna and chromosomal localization of the gene. *Genomics*, 39(3):377–384, Feb 1997. doi:10.1006/geno.1996.4373. ↪p.14, 168

Meyne J., Baker R. J., Hobart H. H., Hsu T. C., Ryder O. A., Ward O. G., Wiley J. E., Wurster-Hill D. H., Yates T. L., and Moyzis R. K. Distribution of non-telomeric sites of the (ttaggg)n telomeric sequence in vertebrate chromosomes. *Chromosoma*, 99(1):3–10, Apr 1990. ↪p.117

Mézard C. Meiotic recombination hotspots in plants. *Biochem Soc Trans*, 34(Pt 4):531–534, Aug 2006. doi:10.1042/BST0340531. ↪p.84

Mézard C., Vignard J., Drouaud J., and Mercier R. The road to crossovers: plants have their say. *Trends Genet*, 23(2):91–99, Feb 2007. doi:10.1016/j.tig.2006.12.007. ↪p.70, 98

Miao X.-X., Xub S.-J., Li M.-H., Li M.-W., Huang J.-H., Dai F.-Y., Marino S. W., Mills D. R., Zeng P., Mita K., Jia S.-H., Zhang Y., Liu W.-B., Xiang H., Guo Q.-H., Xu A.-Y., Kong X.-Y., Lin H.-X., Shi Y.-Z., Lu G., Zhang X., Huang W., Yasukochi Y., Sugasaki T., Shimada T., Nagaraju J., Xiang Z.-H., Wang S.-Y., Goldsmith M. R., Lu C., Zhao G.-P., and Huang Y.-P. Simple sequence repeat-based consensus linkage map of bombyx mori. *Proc Natl Acad Sci U S A*, 102(45):16303–16308, Nov 2005. doi:10.1073/pnas.0507794102. ↪p.30

Mikkelsen T. S., Wakefield M. J., Aken B., Amemiya C. T., Chang J. L., Duke S., Garber M., Gentles A. J., Goodstadt L., Heger A., Jurka J., Kamal M., Mauceli E., Searle S. M. J., Sharpe T., Baker M. L., Batzer M. A., Benos P. V., Belov K., Clamp M., Cook A., Cuff J., Das R., Davidow L., Deakin J. E., Fazzari M. J., Glass J. L., Grabherr M., Greally J. M., Gu W., Hore T. A., Huttley G. A., Kleber M., Jirtle R. L., Koina E., Lee J. T., Mahony S., Marra M. A., Miller R. D., Nicholls R. D., Oda M., Papenfuss A. T., Parra Z. E., Pollock D. D., Ray D. A., Schein J. E., Speed T. P., Thompson K., VandeBerg J. L., Wade C. M., Walker J. A., Waters P. D., Webber C., Weidman J. R., Xie X., Zody M. C., Platform B. I. G. S., Team B. I. W. G. A., Graves J. A. M., Ponting C. P., Breen M., Samollow P. B., Lander E. S., and Lindblad-Toh K. Genome of the marsupial Monodelphis domestica reveals innovation in non-coding sequences. *Nature*, 447(7141):167–177, May 2007. doi:10.1038/nature05805. ↪p.33, 43, 45, 112, 119

Moens P. B. Zebrafish: chiasmata and interference. *Genome*, 49(3):205–208, Mar 2006. doi:10.1139/g06-021. ↪p.28

Mohideen M. A., Moore J. L., and Cheng K. C. Centromere-linked microsatellite markers for linkage groups 3, 4, 6, 7, 13, and 20 of zebrafish (danio rerio). *Genomics*, 67(1):102–106, Jul 2000. doi:10.1006/geno.2000.6233. ↪p.95

Mora L., Sánchez I., Garcia M., and Ponsà M. Chromosome territory positioning of conserved homologous chromosomes in different primate species. *Chromosoma*, 115(5):367–375, Oct 2006. doi:10.1007/s00412-006-0064-6. ↪p.10

Morgan T. An attempt to analyze the constitution of the chromosomes on the basis of sex-limited inheritance in Drosophila. *Journal of Experimental Zoology*, 11(4):365–413, 1911. ISSN 1097-010X. ↪p.47

Morgan T. Complete linkage in the second chromosome of the male of Drosophila. *Science*, 36: 719–720, 1912. ↪p.30

Morgan T. No crossing over in the male of drosophila of genes in the second and third pairs of chromosomes. *Biological Bulletin*, 26(4):195–204, 1914. ↪p.30

Mouchiroud D., Gautier C., and Bernardi G. The compositional distribution of coding sequences and dna molecules in humans and murids. *J Mol Evol*, 27(4):311–320, 1988. ↪p.45

Mouchiroud D., D'Onofrio G., Aïssani B., Macaya G., Gautier C., and Bernardi G. The distribution of genes in the human genome. *Gene*, 100:181–187, Apr 1991. ↪p.44

Munz P. An analysis of interference in the fission yeast schizosaccharomyces pombe. *Genetics*, 137(3):701–707, Jul 1994. ↪p.28

Murakami H. and Nicolas A. Locally, meiotic double-strand breaks targeted by gal4bd-spo11 occur at discrete sites with a sequence preference. *Mol Cell Biol*, 29(13):3500–3516, Jul 2009. doi:10.1128/MCB.00088-09. ↪p.19

Murphy W. J., Larkin D. M., van der Wind A. E., Bourque G., Tesler G., Auvil L., Beever J. E., Chowdhary B. P., Galibert F., Gatzke L., Hitte C., Meyers S. N., Milan D., Ostrander E. A.,

Pape G., Parker H. G., Raudsepp T., Rogatcheva M. B., Schook L. B., Skow L. C., Welge M., Womack J. E., O'brien S. J., Pevzner P. A., and Lewin H. A. Dynamics of mammalian chromosome evolution inferred from multispecies comparative maps. *Science*, 309(5734): 613–617, Jul 2005. doi:10.1126/science.1111387. ↪p.108, 117

Myers E. W., Sutton G. G., Delcher A. L., Dew I. M., Fasulo D. P., Flanigan M. J., Kravitz S. A., Mobarry C. M., Reinert K. H., Remington K. A., Anson E. L., Bolanos R. A., Chou H. H., Jordan C. M., Halpern A. L., Lonardi S., Beasley E. M., Brandon R. C., Chen L., Dunn P. J., Lai Z., Liang Y., Nusskern D. R., Zhan M., Zhang Q., Zheng X., Rubin G. M., Adams M. D., and Venter J. C. A whole-genome assembly of drosophila. *Science*, 287(5461): 2196–2204, Mar 2000. ↪p.33

Myers S., Bottolo L., Freeman C., McVean G., and Donnelly P. A fine-scale map of recombination rates and hotspots across the human genome. *Science*, 310(5746):321–324, Oct 2005. doi:10.1126/science.1117196. ↪p.2, 3, 4, 21, 22, 61, 62, 76, 126

Myers S., Spencer C. C. A., Auton A., Bottolo L., Freeman C., Donnelly P., and McVean G. The distribution and causes of meiotic recombination in the human genome. *Biochem Soc Trans*, 34(Pt 4):526–530, Aug 2006. doi:10.1042/BST0340526. ↪p.60, 79

Myers S., Freeman C., Auton A., Donnelly P., and McVean G. A common sequence motif associated with recombination hot spots and genome instability in humans. *Nat Genet*, 40(9): 1124–1129, Sep 2008. doi:10.1038/ng.213. ↪p.3, 22, 23, 60

Myers S., Bowden R., Tumian A., Bontrop R. E., Freeman C., Macfie T. S., McVean G., and Donnelly P. Drive against hotspot motifs in primates implicates the prdm9 gene in meiotic recombination. *Science*, Dec 2009. doi:10.1126/science.1182363. ↪p.3, 22, 23, 24, 30, 173

Myers S. R. and Griffiths R. C. Bounds on the minimum number of recombination events in a sample history. *Genetics*, 163(1):375–394, Jan 2003. ↪p.59

Nachman M. W. and Crowell S. L. Estimate of the mutation rate per nucleotide in humans. *Genetics*, 156(1):297–304, Sep 2000. ↪p.78

Nachman M. W., Bauer V. L., Crowell S. L., and Aquadro C. F. Dna variability and recombination rates at x-linked loci in humans. *Genetics*, 150(3):1133–1141, Nov 1998. ↪p.77

Nagylaki T. Evolution of a large population under gene conversion. *Proc Natl Acad Sci U S A*, 80(19):5941–5945, Oct 1983a. ↪p.77

Nagylaki T. Evolution of a finite population under gene conversion. *Proc Natl Acad Sci U S A*, 80(20):6278–6281, Oct 1983b. ↪p.77

Nash W. G., Menninger J. C., Wienberg J., Padilla-Nash H. M., and O'Brien S. J. The pattern of phylogenomic evolution of the canidae. *Cytogenet Cell Genet*, 95(3-4):210–224, 2001. ↪p.117

Neale M. J. and Keeney S. Clarifying the mechanics of dna strand exchange in meiotic recombination. *Nature*, 442(7099):153–158, Jul 2006. doi:10.1038/nature04885. ↪p.14

Necşulea A., Popa A., Cooper D. N., Stenson P. D., Mouchiroud D., Gautier C., and Duret L. Meiotic recombination favors the spreading of deleterious mutations in human populations. *Hum Mutat*, 32(2):198–206, Feb 2011. doi:10.1002/humu.21407. ↪p.4, 77, 79, 80, 81

Neumann R. and Jeffreys A. J. Polymorphism in the activity of human crossover hotspots independent of local dna sequence variation. *Hum Mol Genet*, 15(9):1401–1411, May 2006. doi:10.1093/hmg/ddl063. ↪p.36, 126, 174

Newnham L., Jordan P., Rockmill B., Roeder G. S., and Hoffmann E. The synaptonemal complex protein, zip1, promotes the segregation of nonexchange chromosomes at meiosis i. *Proc Natl Acad Sci U S A*, 107(2):781–785, Jan 2010. doi:10.1073/pnas.0913435107. ↪p.169

Nsengimana J., Baret P., Haley C. S., and Visscher P. M. Linkage disequilibrium in the domesticated pig. *Genetics*, 166(3):1395–1404, Mar 2004. ↪p.61

O'Brien S. J., Menotti-Raymond M., Murphy W. J., Nash W. G., Wienberg J., Stanyon R., Copeland N. G., Jenkins N. A., Womack J. E., and Graves J. A. M. The promise of comparative genomics in mammals. *Science*, 286(5439):458–62, 479–81, Oct 1999. ↪p.117

Offenberg H. H., Schalk J. A., Meuwissen R. L., van Aalderen M., Kester H. A., Dietrich A. J., and Heyting C. Scp2: a major protein component of the axial elements of synaptonemal complexes of the rat. *Nucleic Acids Res*, 26(11):2572–2579, Jun 1998. ↪p.169

Oh S. D., Lao J. P., Taylor A. F., Smith G. R., and Hunter N. Recq helicase, sgs1, and xpf family endonuclease, mus81-mms4, resolve aberrant joint molecules during meiotic recombination. *Mol Cell*, 31(3):324–336, Aug 2008. doi:10.1016/j.molcel.2008.07.006. ↪p.26, 27

Ohta K., Shibata T., and Nicolas A. Changes in chromatin structure at recombination initiation sites during yeast meiosis. *EMBO J*, 13(23):5754–5763, Dec 1994. ↪p.19

Oliver P. L., Goodstadt L., Bayes J. J., Birtle Z., Roach K. C., Phadnis N., Beatson S. A., Lunter G., Malik H. S., and Ponting C. P. Accelerated evolution of the prdm9 speciation gene across diverse metazoan taxa. *PLoS Genet*, 5(12):e1000753, Dec 2009. doi:10.1371/journal.pgen.1000753. ↪p.23, 30, 84

Oliver-Bonet M., Campillo M., Turek P. J., Ko E., and Martin R. H. Analysis of replication protein a (rpa) in human spermatogenesis. *Mol Hum Reprod*, 13(12):837–844, Dec 2007. doi:10.1093/molehr/gam076. ↪p.171

Olsen A. K., Bjørtuft H., Wiger R., Holme J., Seeberg E., Bjørås M., and Brunborg G. Highly efficient base excision repair (ber) in human and rat male germ cells. *Nucleic Acids Res*, 29(8): 1781–1790, Apr 2001. ↪p.41

Osman F., Dixon J., Doe C. L., and Whitby M. C. Generating crossovers by resolution of nicked holliday junctions: a role for mus81-eme1 in meiosis. *Mol Cell*, 12(3):761–774, Sep 2003. ↪p.18

Otto S. P. and Barton N. H. The evolution of recombination: removing the limits to natural selection. *Genetics*, 147(2):879–906, Oct 1997. ↪p.77

Otto S. P. and Lenormand T. Resolving the paradox of sex and recombination. *Nat Rev Genet*, 3(4):252–261, Apr 2002. doi:10.1038/nrg761. ↪p.27

Pace J. K. and Feschotte C. The evolutionary history of human dna transposons: evidence for intense activity in the primate lineage. *Genome Res*, 17(4):422–432, Apr 2007. doi:10.1101/gr.5826307. URL `http://dx.doi.org/10.1101/gr.5826307`. ↪p.112

Page S. L. and Hawley R. S. The genetics and molecular biology of the synaptonemal complex. *Annu Rev Cell Dev Biol*, 20:525–558, 2004. doi:10.1146/annurev.cellbio.19.111301.155141. ↪p.13, 14

Paigen K. and Petkov P. Mammalian recombination hot spots: properties, control and evolution. *Nat Rev Genet*, 11(3):221–233, Mar 2010. doi:10.1038/nrg2712. ↪p.xiii, 34, 36, 123

Paigen K., Szatkiewicz J. P., Sawyer K., Leahy N., Parvanov E. D., Ng S. H. S., Graber J. H., Broman K. W., and Petkov P. M. The recombinational anatomy of a mouse chromosome. *PLoS Genet*, 4(7):e1000119, Jul 2008. doi:10.1371/journal.pgen.1000119. ↪p.3, 22, 34, 35, 83, 117

Parisi S., McKay M. J., Molnar M., Thompson M. A., van der Spek P. J., van Drunen-Schoenmaker E., Kanaar R., Lehmann E., Hoeijmakers J. H., and Kohli J. Rec8p, a meiotic recombination

and sister chromatid cohesion phosphoprotein of the rad21p family conserved from fission yeast to humans. *Mol Cell Biol*, 19(5):3515–3528, May 1999. ↪p.170

Parvanov E. D., Petkov P. M., and Paigen K. Prdm9 controls activation of mammalian recombination hotspots. *Science*, Dec 2009. doi:10.1126/science.1181495. ↪p.23

Paterson A. H., Bowers J. E., Bruggmann R., Dubchak I., Grimwood J., Gundlach H., Haberer G., Hellsten U., Mitros T., Poliakov A., Schmutz J., Spannagl M., Tang H., Wang X., Wicker T., Bharti A. K., Chapman J., Feltus F. A., Gowik U., Grigoriev I. V., Lyons E., Maher C. A., Martis M., Narechania A., Otillar R. P., Penning B. W., Salamov A. A., Wang Y., Zhang L., Carpita N. C., Freeling M., Gingle A. R., Hash C. T., Keller B., Klein P., Kresovich S., McCann M. C., Ming R., Peterson D. G., ur Rahman M., Ware D., Westhoff P., Mayer K. F. X., Messing J., and Rokhsar D. S. The sorghum bicolor genome and the diversification of grasses. *Nature*, 457(7229):551–556, Jan 2009. doi:10.1038/nature07723. ↪p.32

Penkner A. M., Fridkin A., Gloggnitzer J., Baudrimont A., Machacek T., Woglar A., Csaszar E., Pasierbek P., Ammerer G., Gruenbaum Y., and Jantsch V. Meiotic chromosome homology search involves modifications of the nuclear envelope protein matefin/sun-1. *Cell*, 139(5): 920–933, Nov 2009. doi:10.1016/j.cell.2009.10.045. ↪p.10

Perkins D. D. Crossing-over and interference in a multiply marked chromosome arm of neurospora. *Genetics*, 47:1253–1274, Sep 1962. ↪p.28

Petes T. D. Meiotic recombination hot spots and cold spots. *Nat Rev Genet*, 2(5):360–369, May 2001. doi:10.1038/35072078. ↪p.19, 44, 109

Petes T. D. and Botstein D. Simple mendelian inheritance of the reiterated ribosomal dna of yeast. *Proc Natl Acad Sci U S A*, 74(11):5091–5095, Nov 1977. ↪p.20

Petes T. D. and Merker J. D. Context dependence of meiotic recombination hotspots in yeast: the relationship between recombination activity of a reporter construct and base composition. *Genetics*, 162(4):2049–2052, Dec 2002. ↪p.44, 109

Petkov P. M., Broman K. W., Szatkiewicz J. P., and Paigen K. Crossover interference underlies sex differences in recombination rates. *Trends Genet*, 23(11):539–542, Nov 2007. doi:10.1016/j.tig.2007.08.015. ↪p.35, 98, 123

Petronczki M., Siomos M. F., and Nasmyth K. Un ménage à quatre: the molecular biology of chromosome segregation in meiosis. *Cell*, 112(4):423–440, Feb 2003. ↪p.27

Pezza R. J., Voloshin O. N., Vanevski F., and Camerini-Otero R. D. Hop2/mnd1 acts on two critical steps in dmc1-promoted homologous pairing. *Genes Dev*, 21(14):1758–1766, Jul 2007. doi:10.1101/gad.1562907. ↪p.170

Phillips R. B., Amores A., Morasch M. R., Wilson C., and Postlethwait J. H. Assignment of zebrafish genetic linkage groups to chromosomes. *Cytogenet Genome Res*, 114(2):155–162, 2006. doi:10.1159/000093332. ↪p.95

Pigozzi M. I. Distribution of mlh1 foci on the synaptonemal complexes of chicken oocytes. *Cytogenet Cell Genet*, 95(3-4):129–133, 2001. ↪p.112

Poissant J., Hogg J. T., Davis C. S., Miller J. M., Maddox J. F., and Coltman D. W. Genetic linkage map of a wild genome: genomic structure, recombination and sexual dimorphism in bighorn sheep. *BMC Genomics*, 11:524, 2010. doi:10.1186/1471-2164-11-524. ↪p.31, 34, 38, 90

Pritchard J. K. and Przeworski M. Linkage disequilibrium in humans: models and data. *Am J Hum Genet*, 69(1):1–14, Jul 2001. doi:10.1086/321275. ↪p.62

Project I. R. G. S. The map-based sequence of the rice genome. *Nature*, 436(7052):793–800, Aug 2005. ↪p.32

Ptak S. E., Hinds D. A., Koehler K., Nickel B., Patil N., Ballinger D. G., Przeworski M., Frazer K. A., and Pääbo S. Fine-scale recombination patterns differ between chimpanzees and humans. *Nat Genet*, 37(4):429–434, Apr 2005. doi:10.1038/ng1529. ↪p.5, 30, 40, 109, 124, 126, 131

Rao B. J., Chiu S. K., Bazemore L. R., Reddy G., and Radding C. M. How specific is the first recognition step of homologous recombination? *Trends Biochem Sci*, 20(3):109–113, Mar 1995. ↪p.12

Rat Genome Sequencing Project Consortium. Genome sequence of the brown norway rat yields insights into mammalian evolution. *Nature*, 428(6982):493–521, Apr 2004. ↪p.32, 45

Reed K. M., Chaves L. D., Hall M. K., Knutson T. P., and Harry D. E. A comparative genetic map of the turkey genome. *Cytogenet Genome Res*, 111(2):118–127, 2005. doi:10.1159/000086380. ↪p.34

Revenkova E., Eijpe M., Heyting C., Hodges C. A., Hunt P. A., Liebe B., Scherthan H., and Jessberger R. Cohesin smc1 beta is required for meiotic chromosome dynamics, sister chromatid cohesion and dna recombination. *Nat Cell Biol*, 6(6):555–562, Jun 2004. doi:10.1038/ncb1135. ↪p.170

Rhesus Macaque Genome Sequencing and Analysis Consortium. Evolutionary and biomedical insights from the rhesus macaque genome. *Science*, 316(5822):222–234, Apr 2007. doi:10.1126/science.1139247. ↪p.32, 90, 99

Robine N., Uematsu N., Amiot F., Gidrol X., Barillot E., Nicolas A., and Borde V. Genome-wide redistribution of meiotic double-strand breaks in saccharomyces cerevisiae. *Mol Cell Biol*, 27 (5):1868–1880, Mar 2007. doi:10.1128/MCB.02063-06. ↪p.20

Rockman M. V. and Kruglyak L. Recombinational landscape and population genomics of caenorhabditis elegans. *PLoS Genet*, 5(3):e1000419, Mar 2009. doi:10.1371/journal.pgen.1000419. ↪p.33

Rockmill B. and Roeder G. S. Red1: a yeast gene required for the segregation of chromosomes during the reductional division of meiosis. *Proc Natl Acad Sci U S A*, 85(16):6057–6061, Aug 1988. ↪p.169, 170

Rockmill B., Voelkel-Meiman K., and Roeder G. S. Centromere-proximal crossovers are associated with precocious separation of sister chromatids during meiosis in saccharomyces cerevisiae. *Genetics*, 174(4):1745–1754, Dec 2006. doi:10.1534/genetics.106.058933. ↪p.21

Roeder G. S. Meiotic chromosomes: it takes two to tango. *Genes Dev*, 11(20):2600–2621, Oct 1997. ↪p.25

Rogers J., Garcia R., Shelledy W., Kaplan J., Arya A., Johnson Z., Bergstrom M., Novakowski L., Nair P., Vinson A., Newman D., Heckman G., and Cameron J. An initial genetic linkage map of the rhesus macaque (macaca mulatta) genome using human microsatellite loci. *Genomics*, 87(1):30–38, Jan 2006. doi:10.1016/j.ygeno.2005.10.004. ↪p.32, 90

Romiguier J., Ranwez V., Douzery E. J. P., and Galtier N. Contrasting gc-content dynamics across 33 mammalian genomes: relationship with life-history traits and chromosome sizes. *Genome Res*, 20(8):1001–1009, Aug 2010. doi:10.1101/gr.104372.109. ↪p.45

S. T. A. R. Consortium. Snp and haplotype mapping for genetic analysis in the rat. *Nat Genet*, 40(5):560–566, May 2008. doi:10.1038/ng.124. ↪p.32

Saito T. T., Youds J. L., Boulton S. J., and Colaiácovo M. P. Caenorhabditis elegans him-18/slx-4 interacts with slx-1 and xpf-1 and maintains genomic integrity in the

germline by processing recombination intermediates. *PLoS Genet*, 5(11):e1000735, Nov 2009. doi:10.1371/journal.pgen.1000735. ↪p.17, 173

Samollow P. B. *Marsupial Genetics and Genomics*, chapter Marsupial linkage maps., pages 75–99. Springer, Dordrecht, Heidelberg, London, New York, 2010. ↪p.64, 119

Samollow P. B., Gouin N., Miethke P., Mahaney S. M., Kenney M., VandeBerg J. L., Graves J. A. M., and Kammerer C. M. A microsatellite-based, physically anchored linkage map for the gray, short-tailed opossum (Monodelphis domestica). *Chromosome Res*, 15(3):269–281, 2007. doi:10.1007/s10577-007-1123-4. ↪p.33, 38, 45, 90, 105, 108, 113, 114, 119

Sasaki H. and Matsui Y. Epigenetic events in mammalian germ-cell development: reprogramming and beyond. *Nat Rev Genet*, 9(2):129–140, Feb 2008. doi:10.1038/nrg2295. ↪p.111, 122

Sasaki M., Lange J., and Keeney S. Genome destabilization by homologous recombination in the germ line. *Nat Rev Mol Cell Biol*, 11(3):182–195, Mar 2010. doi:10.1038/nrm2849. ↪p.23, 25, 26

Schalk J. A., Dietrich A. J., Vink A. C., Offenberg H. H., van Aalderen M., and Heyting C. Localization of scp2 and scp3 protein molecules within synaptonemal complexes of the rat. *Chromosoma*, 107(8):540–548, Dec 1998. ↪p.13

Schalk J. A., Offenberg H. H., Peters E., Groot N. P., Hoovers J. M., and Heyting C. Isolation and characterization of the human scp2 cdna and chromosomal localization of the gene. *Mamm Genome*, 10(6):642–644, Jun 1999. ↪p.169

Scherthan H. A bouquet makes ends meet. *Nat Rev Mol Cell Biol*, 2(8):621–627, Aug 2001. doi:10.1038/35085086. ↪p.11

Schmekel K. and Daneholt B. The central region of the synaptonemal complex revealed in three dimensions. *Trends Cell Biol*, 5(6):239–242, Jun 1995. ↪p.13

Schnable P. S., Ware D., Fulton R. S., and et al. J. C. S. The b73 maize genome: complexity, diversity, and dynamics. *Science*, 326(5956):1112–1115, Nov 2009. doi:10.1126/science.1178534. ↪p.32

Schwacha A. and Kleckner N. Identification of double holliday junctions as intermediates in meiotic recombination. *Cell*, 83(5):783–791, Dec 1995. ↪p.17

Schwacha A. and Kleckner N. Interhomolog bias during meiotic recombination: meiotic functions promote a highly differentiated interhomolog-only pathway. *Cell*, 90(6):1123–1135, Sep 1997. ↪p.15

Sen T. Z., Harper L. C., Schaeffer M. L., Andorf C. M., Seigfried T. E., Campbell D. A., and Lawrence C. J. Choosing a genome browser for a model organism database: surveying the maize community. *Database (Oxford)*, 2010:baq007, 2010. doi:10.1093/database/baq007. ↪p.32

Sharp P. J. and Hayman D. L. An examination of the role of chiasma frequency in the genetic system of marsupials. *Heredity*, 60 (Pt 1):77–85, Feb 1988. ↪p.34, 105, 106, 108, 119, 123

Sharples G. J. The x philes: structure-specific endonucleases that resolve holliday junctions. *Mol Microbiol*, 39(4):823–834, Feb 2001. ↪p.17

Sherman J. D. and Stack S. M. Two-dimensional spreads of synaptonemal complexes from solanaceous plants. vi. high-resolution recombination nodule map for tomato (lycopersicon esculentum). *Genetics*, 141(2):683–708, Oct 1995. ↪p.28

Shi J., Wolf S. E., Burke J. M., Presting G. G., Ross-Ibarra J., and Dawe R. K. Widespread gene conversion in centromere cores. *PLoS Biol*, 8(3):e1000327, Mar 2010. doi:10.1371/journal.pbio.1000327. ↪p.22

Shifman S., Bell J. T., Copley R. R., Taylor M. S., Williams R. W., Mott R., and Flint J. A high-resolution single nucleotide polymorphism genetic map of the mouse genome. *PLoS Biol*, 4(12):e395, Nov 2006. doi:10.1371/journal.pbio.0040395. ↪p.3, 34, 119

Shinohara A., Ogawa H., and Ogawa T. Rad51 protein involved in repair and recombination in s. cerevisiae is a reca-like protein. *Cell*, 69(3):457–470, May 1992. ↪p.13, 171

Shinohara M., Sakai K., Shinohara A., and Bishop D. K. Crossover interference in saccharomyces cerevisiae requires a tid1/rdh54- and dmc1-dependent pathway. *Genetics*, 163(4):1273–1286, Apr 2003. ↪p.25, 172

Shinohara M., Oh S. D., Hunter N., and Shinohara A. Crossover assurance and crossover interference are distinctly regulated by the zmm proteins during yeast meiosis. *Nat Genet*, 40 (3):299–309, Mar 2008. doi:10.1038/ng.83. ↪p.16, 25, 69, 170

Singer A., Perlman H., Yan Y., Walker C., Corley-Smith G., Brandhorst B., and Postlethwait J. Sex-specific recombination rates in zebrafish (danio rerio). *Genetics*, 160(2):649–657, Feb 2002. ↪p.38, 102

Singer T., Fan Y., Chang H.-S., Zhu T., Hazen S. P., and Briggs S. P. A high-resolution map of arabidopsis recombinant inbred lines by whole-genome exon array hybridization. *PLoS Genet*, 2(9):e144, Sep 2006. doi:10.1371/journal.pgen.0020144. ↪p.33, 90

Slatkin M. Linkage disequilibrium–understanding the evolutionary past and mapping the medical future. *Nat Rev Genet*, 9(6):477–485, Jun 2008. doi:10.1038/nrg2361. ↪p.59, 61

Smit, AFA, Hubley R., and Green P. Repeatmasker open-3.0, 2010. URL `http://www.repeatmasker.org/PreMaskedGenomes.html`. ↪p.112, 113

Smit A. F. Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Curr Opin Genet Dev*, 9(6):657–663, Dec 1999. ↪p.44, 124

Smith J. M. and Haigh J. The hitch-hiking effect of a favourable gene. *Genet Res*, 23(1):23–35, Feb 1974. ↪p.77

Snowden T., Acharya S., Butz C., Berardini M., and Fishel R. hmsh4-hmsh5 recognizes holliday junctions and forms a meiosis-specific sliding clamp that embraces homologous chromosomes. *Mol Cell*, 15(3):437–451, Aug 2004. doi:10.1016/j.molcel.2004.06.040. ↪p.172

Snowden T., Shim K.-S., Schmutte C., Acharya S., and Fishel R. hmsh4-hmsh5 adenosine nucleotide processing and interactions with homologous recombination machinery. *J Biol Chem*, 283(1):145–154, Jan 2008. doi:10.1074/jbc.M704060200. ↪p.172

Solignac M., Mougel F., Vautrin D., Monnerot M., and Cornuet J.-M. A third-generation microsatellite-based linkage map of the honey bee, apis mellifera, and its comparison with the sequence-based physical map. *Genome Biol*, 8(4):R66, 2007. doi:10.1186/gb-2007-8-4-r66. ↪p.98

Soriano P., Meunier-Rotival M., and Bernardi G. The distribution of interspersed repeats is nonuniform and conserved in the mouse and human genomes. *Proc Natl Acad Sci U S A*, 80 (7):1816–1820, Apr 1983. ↪p.44, 124

Soriano P., Keitges E. A., Schorderet D. F., Harbers K., Gartler S. M., and Jaenisch R. High rate of recombination and double crossovers in the mouse pseudoautosomal region during male meiosis. *Proc Natl Acad Sci U S A*, 84(20):7218–7220, Oct 1987. ↪p.43

Spencer C. C. A., Deloukas P., Hunt S., Mullikin J., Myers S., Silverman B., Donnelly P., Bentley D., and McVean G. The influence of recombination on human genetic diversity. *PLoS Genet*, 2(9):e148, Sep 2006. doi:10.1371/journal.pgen.0020148. ↪p.42, 76, 79

Squartini F. *Stationarity and Reversibility in the Nucleotide Evolutionary Process*. PhD thesis, Freien Universität Berlin, 2010. ↪p.74, 75

Stahl F. W., Foss H. M., Young L. S., Borts R. H., Abdullah M. F. F., and Copenhaver G. P. Does crossover interference count in saccharomyces cerevisiae? *Genetics*, 168(1):35–48, Sep 2004. doi:10.1534/genetics.104.027789. ↪p.67

Stapley J., Birkhead T. R., Burke T., and Slate J. Pronounced inter- and intrachromosomal variation in linkage disequilibrium across the zebra finch genome. *Genome Res*, 20(4):496–502, Apr 2010. doi:10.1101/gr.102095.109. ↪p.61

Stefansson H., Helgason A., Thorleifsson G., Steinthorsdottir V., Masson G., Barnard J., Baker A., Jonasdottir A., Ingason A., Gudnadottir V. G., Desnica N., Hicks A., Gylfason A., Gudbjartsson D. F., Jonsdottir G. M., Sainz J., Agnarsson K., Birgisdottir B., Ghosh S., Olafsdottir A., Cazier J.-B., Kristjansson K., Frigge M. L., Thorgeirsson T. E., Gulcher J. R., Kong A., and Stefansson K. A common inversion under selection in europeans. *Nat Genet*, 37 (2):129–137, Feb 2005. doi:10.1038/ng1508. ↪p.34

Stenson P. D., Mort M., Ball E. V., Howells K., Phillips A. D., Thomas N. S., and Cooper D. N. The human gene mutation database: 2008 update. *Genome Med*, 1(1):13, 2009. doi:10.1186/gm13. ↪p.79

Stephan W. and Langley C. H. Molecular genetic variation in the centromeric region of the x chromosome in three drosophila ananassae populations. i. contrasts between the vermilion and forked loci. *Genetics*, 121(1):89–99, Jan 1989. ↪p.77

Storlazzi A., Gargano S., Ruprich-Robert G., Falque M., David M., Kleckner N., and Zickler D. Recombination proteins mediate meiotic spatial chromosome organization and pairing. *Cell*, 141(1):94–106, Apr 2010. doi:10.1016/j.cell.2010.02.041. ↪p.13, 15, 19

Strickland W. N. An analysis of interference in aspergillus nidulans. *Proc R Soc Lond B Biol Sci*, 149(934):82–101, Jul 1958. ↪p.28

Stumpf M. P. H. and McVean G. A. T. Estimating recombination rates from population-genetic data. *Nat Rev Genet*, 4(12):959–968, Dec 2003. doi:10.1038/nrg1227. ↪p.xiii, 59, 64

Sturtevant A. The linear arrangement of six sex-linked factors in Drosophila, as shown by their mode of association. *Journal of Experimental Zoology*, 14(1):43–59, 1913. ISSN 1097-010X. ↪p.3, 25, 30, 47, 49

Su X., Ferdig M. T., Huang Y., Huynh C. Q., Liu A., You J., Wootton J. C., and Wellems T. E. A genetic map and recombination parameters of the human malaria parasite plasmodium falciparum. *Science*, 286(5443):1351–1353, Nov 1999. ↪p.32, 90

Su Y., Barton A. B., and Kaback D. B. Decreased meiotic reciprocal recombination in subtelomeric regions in saccharomyces cerevisiae. *Chromosoma*, 109(7):467–475, Nov 2000. ↪p.20

Sueoka N. On the genetic basis of variation and heterogeneity of DNA base composition. *Proc Natl Acad Sci U S A*, 48:582–592, Apr 1962. ↪p.71, 114

Sueoka N. Directional mutation pressure and neutral molecular evolution. *Proc Natl Acad Sci U S A*, 85(8):2653–2657, Apr 1988. ↪p.126

Sunyaev S., Ramensky V., Koch I., Lathe W., Kondrashov A. S., and Bork P. Prediction of deleterious human alleles. *Hum Mol Genet*, 10(6):591–597, Mar 2001. ↪p.79

Surtees J. A., Argueso J. L., and Alani E. Mismatch repair proteins: key regulators of genetic recombination. *Cytogenet Genome Res*, 107(3-4):146–159, 2004. doi:10.1159/000080593. ↪p.40

Sutter N. B., Eberle M. A., Parker H. G., Pullar B. J., Kirkness E. F., Kruglyak L., and Ostrander E. A. Extensive and breed-specific linkage disequilibrium in canis familiaris. *Genome Res*, 14 (12):2388–2396, Dec 2004. doi:10.1101/gr.3147604. ↪p.61

Svendsen J. M. and Harper J. W. Gen1/yen1 and the slx4 complex: Solutions to the problem of holliday junction resolution. *Genes Dev*, 24(6):521–536, Mar 2010. doi:10.1101/gad.1903510. ↪p.19

Svendsen J. M., Smogorzewska A., Sowa M. E., O'Connell B. C., Gygi S. P., Elledge S. J., and Harper J. W. Mammalian btbd12/slx4 assembles a holliday junction resolvase and is required for dna repair. *Cell*, 138(1):63–77, Jul 2009. doi:10.1016/j.cell.2009.06.030. ↪p.17, 173

Swinburne J. E., Boursnell M., Hill G., Pettitt L., Allen T., Chowdhary B., Hasegawa T., Kurosawa M., Leeb T., Mashima S., Mickelson J. R., Raudsepp T., Tozaki T., and Binns M. Single linkage group per chromosome genetic linkage map for the horse, based on two three-generation, full-sibling, crossbred horse reference families. *Genomics*, 87(1):1–29, Jan 2006. doi:10.1016/j.ygeno.2005.09.001. ↪p.31, 90

Sym M., Engebrecht J. A., and Roeder G. S. Zip1 is a synaptonemal complex protein required for meiotic chromosome synapsis. *Cell*, 72(3):365–378, Feb 1993. ↪p.169

Székvölgyi L. and Nicolas A. From meiosis to postmeiotic events: homologous recombination is obligatory but flexible. *FEBS J*, 277(3):571–589, Feb 2010. doi:10.1111/j.1742-4658.2009.07502.x. ↪p.19, 27

Szostak J. W., Orr-Weaver T. L., Rothstein R. J., and Stahl F. W. The double-strand-break repair model for recombination. *Cell*, 33(1):25–35, May 1983. ↪p.2, 15

Taylor E. R. and McGowan C. H. Cleavage mechanism of human mus81-eme1 acting on holliday-junction structures. *Proc Natl Acad Sci U S A*, 105(10):3757–3762, Mar 2008. doi:10.1073/pnas.0710291105. ↪p.18

Tease C. and Hultén M. A. Inter-sex variation in synaptonemal complex lengths largely determine the different recombination rates in male and female germ cells. *Cytogenet Genome Res*, 107 (3-4):208–215, 2004. doi:10.1159/000080599. ↪p.35, 112

Tease C., Hartshorne G., and Hultén M. Altered patterns of meiotic recombination in human fetal oocytes with asynapsis and/or synaptonemal complex fragmentation at pachytene. *Reprod Biomed Online*, 13(1):88–95, Jul 2006. ↪p.28

Tessé S., Storlazzi A., Kleckner N., Gargano S., and Zickler D. Localization and roles of ski8p protein in sordaria meiosis and delineation of three mechanistically distinct steps of meiotic homolog juxtaposition. *Proc Natl Acad Sci U S A*, 100(22):12865–12870, Oct 2003. doi:10.1073/pnas.2034282100. ↪p.13

Tiemann-Boege I., Calabrese P., Cochran D. M., Sokol R., and Arnheim N. High-resolution recombination patterns in a region of human chromosome 21 measured by sperm typing. *PLoS Genet*, 2(5):e70, May 2006. doi:10.1371/journal.pgen.0020070. ↪p.174, 175

Tsai C. J., Mets D. G., Albrecht M. R., Nix P., Chan A., and Meyer B. J. Meiotic crossover number and distribution are regulated by a dosage compensation protein that resembles a condensin subunit. *Genes Dev*, 22(2):194–211, Jan 2008. doi:10.1101/gad.1618508. ↪p.28

Tsai I. J., Burt A., and Koufopanou V. Conservation of recombination hotspots in yeast. *Proc Natl Acad Sci U S A*, 107(17):7847–7852, Apr 2010. doi:10.1073/pnas.0908774107. ↪p.110

Tsubouchi T. and Roeder G. S. A synaptonemal complex protein promotes homology-independent centromere coupling. *Science*, 308(5723):870–873, May 2005. doi:10.1126/science.1108283. ↪p.19, 169

Tuskan G., Difazio S., Jansson S., Bohlmann J., Grigoriev I., Hellsten U., Putnam N., Ralph S., Rombauts S., Salamov A., et al. The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). *Science*, 313(5793):1596, 2006. ↪p.31

Tzur Y. B., Wilson K. L., and Gruenbaum Y. Sun-domain proteins: 'velcro' that links the nucleoskeleton to the cytoskeleton. *Nat Rev Mol Cell Biol*, 7(10):782–788, Oct 2006. doi:10.1038/nrm2003. ↪p.10, 168

Vallente R. U., Cheng E. Y., and Hassold T. J. The synaptonemal complex and meiotic recombination in humans: new approaches to old questions. *Chromosoma*, 115(3):241–249, Jun 2006. doi:10.1007/s00412-006-0058-4. ↪p.28

Vazquez J., Belmont A. S., and Sedat J. W. The dynamics of homologous chromosome pairing during male drosophila meiosis. *Curr Biol*, 12(17):1473–1483, Sep 2002. ↪p.10

Vingborg R. K. K., Gregersen V. R., Zhan B., Panitz F., Høj A., Sørensen K. K., Madsen L. B., Larsen K., Hornshøj H., Wang X., and Bendixen C. A robust linkage map of the porcine autosomes based on gene-associated snps. *BMC Genomics*, 10:134, 2009. doi:10.1186/1471-2164-10-134. ↪p.32, 38, 90, 92, 102

Wahlberg P. *Chicken Genomics - Linkage and QTL mapping*. PhD thesis, Uppsala Universitet, 2009. ↪p.64

Wahls W. P., Siegel E. R., and Davidson M. K. Meiotic recombination hotspots of fission yeast are directed to loci that express non-coding rna. *PLoS One*, 3(8):e2887, 2008. doi:10.1371/journal.pone.0002887. ↪p.20

Watanabe Y., Fujiyama A., Ichiba Y., Hattori M., Yada T., Sakaki Y., and Ikemura T. Chromosome-wide assessment of replication timing for human chromosomes 11q and 21q: disease-related genes in timing-switch regions. *Hum Mol Genet*, 11(1):13–21, Jan 2002. ↪p.44

Webb A. J., Berg I. L., and Jeffreys A. Sperm cross-over activity in regions of the human genome showing extreme breakdown of marker association. *Proc Natl Acad Sci U S A*, 105 (30):10471–10476, Jul 2008. doi:10.1073/pnas.0804933105. ↪p.22, 84, 175

Webster M. T., Smith N. G. C., Hultin-Rosenberg L., Arndt P. F., and Ellegren H. Male-driven biased gene conversion governs the evolution of base composition in human alu repeats. *Mol Biol Evol*, 22(6):1468–1474, Jun 2005. doi:10.1093/molbev/msi136. ↪p.5, 44, 75, 109, 111, 112, 115, 124, 125, 127, 130

Webster M. T., Axelsson E., and Ellegren H. Strong regional biases in nucleotide substitution in the chicken genome. *Mol Biol Evol*, 23(6):1203–1216, Jun 2006. doi:10.1093/molbev/msk008. ↪p.75, 122

Weterings E. and van Gent D. C. The mechanism of non-homologous end-joining: a synopsis of synapsis. *DNA Repair (Amst)*, 3(11):1425–1435, Nov 2004. doi:10.1016/j.dnarep.2004.06.003. ↪p.14

Wienberg J. The evolution of eutherian chromosomes. *Curr Opin Genet Dev*, 14(6):657–666, Dec 2004. doi:10.1016/j.gde.2004.10.001. ↪p.108, 117

Winckler W., Myers S. R., Richter D. J., Onofrio R. C., McDonald G. J., Bontrop R. E., McVean G. A. T., Gabriel S. B., Reich D., Donnelly P., and Altshuler D. Comparison of fine-scale recombination rates in humans and chimpanzees. *Science*, 308(5718):107–111, Apr 2005. doi:10.1126/science.1105322. ↪p.5, 30, 40, 109, 124, 126, 131

Wolfe K. H., Sharp P. M., and Li W. H. Mutation rates differ among regions of the mammalian genome. *Nature*, 337(6204):283–285, Jan 1989. doi:10.1038/337283a0. ↪p.45

Wong A. K., Ruhe A. L., Dumont B. L., Robertson K. R., Guerrero G., Shull S. M., Ziegle J. S., Millon L. V., Broman K. W., Payseur B. A., and Neff M. W. A comprehensive linkage map of the dog genome. *Genetics*, 184(2):595–605, Feb 2010. doi:10.1534/genetics.109.106831. ↪p.3, 31, 34, 38, 90, 108, 114, 117, 121, 123

Wu L. and Hickson I. D. The bloom's syndrome helicase suppresses crossing over during homologous recombination. *Nature*, 426(6968):870–874, Dec 2003. doi:10.1038/nature02253. ↪p.2, 18, 173

Wu L. and Hickson I. D. Dna helicases required for homologous recombination and repair of damaged replication forks. *Annu Rev Genet*, 40:279–306, 2006. doi:10.1146/annurev.genet.40.110405.090636. ↪p.18, 173

Wu T. C. and Lichten M. Meiosis-induced double-strand break sites determined by yeast chromatin structure. *Science*, 263(5146):515–518, Jan 1994. ↪p.19

Wu T. C. and Lichten M. Factors that affect the location and frequency of meiosis-induced double-strand breaks in saccharomyces cerevisiae. *Genetics*, 140(1):55–66, May 1995. ↪p.20

Wu Z. K., Getun I. V., and Bois P. R. J. Anatomy of mouse recombination hot spots. *Nucleic Acids Res*, 38(7):2346–2354, Apr 2010. doi:10.1093/nar/gkp1251. ↪p.21, 22, 39, 84

Xu L. and Kleckner N. Sequence non-specific double-strand breaks and interhomolog interactions prior to double-strand break formation at a meiotic recombination hot spot in yeast. *EMBO J*, 14(20):5115–5128, Oct 1995. ↪p.19

Xu M. and Cook P. R. The role of specialized transcription factories in chromosome pairing. *Biochim Biophys Acta*, 1783(11):2155–2160, Nov 2008. doi:10.1016/j.bbamcr.2008.07.013. ↪p.13

Yamada T., ichi Mizuno K., Hirota K., Kon N., Wahls W. P., Hartsuiker E., Murofushi H., Shibata T., and Ohta K. Roles of histone acetylation and chromatin remodeling factor in a meiotic recombination hotspot. *EMBO J*, 23(8):1792–1803, Apr 2004. doi:10.1038/sj.emboj.7600138. ↪p.

Yamamoto K., Narukawa J., Kadono-Okuda K., Nohata J., Sasanuma M., Suetsugu Y., Banno Y., Fujii H., Goldsmith M. R., and Mita K. Construction of a single nucleotide polymorphism linkage map for the silkworm, Bombyx mori, based on bacterial artificial chromosome end sequences. *Genetics*, 173(1):151–161, May 2006. doi:10.1534/genetics.105.053801. ↪p.30

Yanowitz J. Meiosis: making a break for it. *Curr Opin Cell Biol*, Sep 2010. doi:10.1016/j.ceb.2010.08.016. ↪p.12, 19, 27

Yauk C. L., Bois P. R. J., and Jeffreys A. J. High-resolution sperm typing of meiotic recombination in the mouse mhc ebeta gene. *EMBO J*, 22(6):1389–1397, Mar 2003. doi:10.1093/emboj/cdg136. ↪p.84

Yildiz O., Majumder S., Kramer B., and Sekelsky J. J. Drosophila mus312 interacts with the nucleotide excision repair endonuclease mei-9 to generate meiotic crossovers. *Mol Cell*, 10(6): 1503–1509, Dec 2002. ↪p.28

Yin J., Jordan M. I., and Song Y. S. Joint estimation of gene conversion rates and mean conversion tract lengths from population snp data. *Bioinformatics*, 25(12):i231–i239, Jun 2009. doi:10.1093/bioinformatics/btp229. ↪p.62

Youds J. L., Mets D. G., McIlwraith M. J., Martin J. S., Ward J. D., ONeil N. J., Rose A. M., West S. C., Meyer B. J., and Boulton S. J. Rtel-1 enforces meiotic crossover interference and homeostasis. *Science*, 327(5970):1254–1258, Mar 2010. doi:10.1126/science.1183112. ↪p.25, 172

Yu J., Lazzeroni L., Qin J., Huang M. M., Navidi W., Erlich H., and Arnheim N. Individual variation in recombination among human males. *Am J Hum Genet*, 59(6):1186–1192, Dec 1996. ↪p.36

Yu N., Jensen-Seaman M. I., Chemnick L., Ryder O., and Li W.-H. Nucleotide diversity in gorillas. *Genetics*, 166(3):1375–1383, Mar 2004. ↪p.78

Yuan L., Liu J. G., Zhao J., Brundell E., Daneholt B., and Höög C. The murine scp3 gene is required for synaptonemal complex assembly, chromosome synapsis, and male fertility. *Mol Cell*, 5(1):73–83, Jan 2000. ↪p.169

Zenger K. R., McKenzie L. M., and Cooper D. W. The first comprehensive genetic linkage map of a marsupial: the tammar wallaby (Macropus eugenii). *Genetics*, 162(1):321–330, Sep 2002. ↪p.38

Zhang L. Adaptive evolution and frequent gene conversion in the brain expressed x-linked gene family in mammals. *Biochem Genet*, 46(5-6):293–311, Jun 2008. doi:10.1007/s10528-008-9148-8. ↪p.43

Zickler D. From early homologue recognition to synaptonemal complex formation. *Chromosoma*, 115(3):158–174, Jun 2006. doi:10.1007/s00412-006-0048-6. ↪p.2, 10, 11, 12, 14, 19, 95

Zickler D. and Kleckner N. The leptotene-zygotene transition of meiosis. *Annu Rev Genet*, 32: 619–697, 1998. doi:10.1146/annurev.genet.32.1.619. ↪p.11, 14

Zickler D. and Kleckner N. Meiotic chromosomes: integrating structure and function. *Annu Rev Genet*, 33:603–754, 1999. doi:10.1146/annurev.genet.33.1.603. ↪p.19, 20, 169

Ziegler A., König I., and Pahlke F. *A Statistical Approach to Genetic Epidemiology: Concepts and Applications, with an E-learning Platform*. Wiley-VCh, 2010. ISBN 3527323899. ↪p.49, 64

Zimin A. V., Delcher A. L., Florea L., Kelley D. R., Schatz M. C., Puiu D., Hanrahan F., Pertea G., Tassell C. P. V., Sonstegard T. S., Marçais G., Roberts M., Subramanian P., Yorke J. A., and Salzberg S. L. A whole-genome assembly of the domestic cow, bos taurus. *Genome Biol*, 10(4):R42, 2009. doi:10.1186/gb-2009-10-4-r42. ↪p.31

Zoubak S., Clay O., and Bernardi G. The gene distribution of the human genome. *Gene*, 174(1): 95–102, Sep 1996. ↪p.44

Additional Material

Appendix A

Proteins involved in meiosis.

Appendix B

Human recombination hotspots analyzed by sperm-typing

Appendix C

Correlations between distance to telomeres, GC*, and sex-specific COR

Appendix D

Opossum correlation windows smaller and larger than 20 Mb

A Proteins involved in meiosis

Protein name	Category	Organism	Role	References
Ndj1	Chromosome alignment	budding yeast	Ndj1 is a meiosis-specific protein involved in the processing of telomere repeats, and required for the localization of telomeres to the NE and the "bouquet" configuration.	Conrad et al. (1997); Joseph et al. (2005)
SUN - KASH	Chromosome alignment	budding yeast, *C. elegans*	Proteins from these two domain proteins form complexes that span the nuclear envelope, linking on one side the telomeres with the cytoskeleton on the other side. They are thus participating in the movement of chromosomes, essential for the recognition and alignment of homologues during meiosis.	Tzur et al. (2006); Fridkin et al. (2009)
SCP1/ Syn1	SC Component - central region	rat, mouse, human / hamster	These proteins are the major component of the SC transverse filaments, with their N-terminal end located within the CE and the C-terminal close to or within the LE. Their recruitment starts at zygotene and the SCP1/Syn1 are fully assembled to the SC at pachytene.	Meuwissen et al. (1992); Liu et al. (1996); Meuwissen et al. (1997)

Protein name	Category	Organism	Role	References
Zip1	SC Component - central region	yeast	Component of the central region of the SC, Zip1 is localized specifically at centromeres early in prophase, before the homologs become aligned. This early attachment mediates centromere pairing and segregation of the partners at meiosis I. After the initiation of recombination, Zip1 holds homologs together.	Sym et al. (1993); Tsubouchi and Roeder (2005); Newnham et al. (2010)
SCP2	SC Component - axial core	rat, human	This axis-associated protein seems to bind AT-rich DNA and be involved in the structural organization of meiotic prophase chromosomes.	Offenberg et al. (1998); Schalk et al. (1999)
SCP3 / Cor1 / Sycp3	SC Component - axial core	rat, mouse / hamster / human	It has been proposed that SCP3 is essential for the formation of the lateral elements of the SC in spermatocytes.	Lammers et al. (1994); Yuan et al. (2000)
Hop1	SC Component - axial core	yeast	Hop1 binds strongly to non-specific duplex DNA, adjacent to sites of DSB formation, probably via its Zinc-finger domain.	Hollingsworth and Byers (1989); Kironmai et al. (1998); Zickler and Kleckner (1999)
Red1	SC Component - axial core	yeast	Red1 recruits Hop1 proteins to the chromosomes. The Hop1/Red1 complexes, together with Zip1 are responsible for the initiation and elongation of SCs.	Rockmill and Roeder (1988); Lin et al. (2010)
Zip2	SC Component	yeast	Zip2 is necessary for the polymerization of Zip1 along the chromosomes and the formation of SC. It is also involved in the synapsis initiation.	Chua and Roeder (1998); Tsubouchi and Roeder (2005)

Protein name	Category	Organism	Role	References
Hop2	SC Component	yeast	Hop2 is part of protein complexes that assure the specific association between homologous, by preventing non-homologous interactions. They are supposed to facilitate the search for homology through the capture of potential partner chromosomes. These complexes also act on the Dmc1-ssDNA filaments, by stabilizing them.	Leu et al. (1998); Pezza et al. (2007)
Spo16 - Spo22	SC Component	yeast	These two proteins functions together to facilitate Zip1 polymerization from sites of recombination.	Shinohara et al. (2008)
Mek1 (aka Mre4)	Protein kinase	yeast	Mek1 functions in combination with Red1 and Hop1 to ensure interhomologue (IH) recombination by preventing the use of a sister chromatid as the template in DNA repair	Rockmill and Roeder (1988); Leem and Ogawa (1992); Bailis and Roeder (1998)
Rec8	Cohesin	yeast, human	Rec8 is a phosphoprotein necessary for meiotic sister chromatin cohesion and the formation of chiasmata.	Parisi et al. (1999)
Smc1Beta	Cohesin	human	Smc1Beta is part of the ATPase subunits complex that is responsible, as its name suggests, in the structural maintainance of chromosomes during mitosis as well as meiosis. Being part of the cohesin complex, the role of Smc1Beta is in the maintaining cohesion between sister chromatids but not only. It is also important in the assembly of the axial elements of the SC and the formation of loops along the AEs.	Revenkova et al. (2004)

Protein name	Category	Organism	Role	References
Spo11	DSB formation / repair	all organisms undergoing meiotic recombination	Spo11 is the endonuclease that makes meiotic DSBs.	Keeney et al. (1997)
Mre11 - Rad50 - Nbs1 / Xrs2	DSB formation / repair	all organisms undergoing meiotic recombination	This protein complex removes Spo11 and is involved in the 5' to 3' resection of the DSB.	Borde and Cobb (2009)
RPA	DSB formation / repair	eukaryotes	This replication protein A binds the 3' ssDNA in DSBs and in collaboration with other proteins loads Rad51 and Dmc1 to the ssDNA.	Oliver-Bonet et al. (2007)
Rad51	DSB formation / repair	most eukaryotes	Rad51 is a recombinase essential for both mitotic and meiotic recombination. Like Dmc1, it plays an essential role in the repair of DSBs by DNA strand transfer. After binding to the single strand DNA, Rad51 plays a role in searching for DNA that shares sequence homology.	Aboussekhra et al. (1992); Shinohara et al. (1992)
Dmc1	DSB formation / repair	most eukaryotes	Dmc1 is a meiosis specific recombinase, that localizes to meiotic nodules, together with RAD51, and forms nucleoprotein filaments that catalyze strand invasion and double Holliday junction formation. Dmc1 is thought to direct Rad51 towards inter-homologue recombination, rather than inter-sister exchange.	Bishop et al. (1992); Masson and West (2001); Kagawa and Kurumizaka (2010)

Protein name	Category	Organism	Role	References
GEN1 / GEN-1 / YEN1	DSB formation /repair	human / *C. elegans* / yeast respectively	Recently identified resolvase of the double Holliday Junctions. They are enzymes producing symmetric nicks in chromatids of the same polarity. The resulting nicked chromatids are later ligated to form COs or NCOs.	Ip et al. (2008)
RTEL1	DSB formation /repair	*C. elegans*, human	RTEL1 is an anti-recombinase that disassembles D loop-recombination intermediates, promoting SDSA pathway to generate NCO. It is thus involved in the interference between COs.	Barber et al. (2008); Youds et al. (2010)
Tid1	DSB formation / repair	yeast	Tid1 is a protein that facilitates strand invasion. tid1 mutants have normal levels of COs but the interference us decreased.	Shinohara et al. (2003)
MSH4 & MSH5	Mismatch repair	yeast, *C. elegans*, mouse, human	These two proteins are specific to the repair of DNA mismatches during meiotic recombination. They are required to stabilize the single strand invasion recombination intermediates and can bind to the Holliday jonction.	Snowden et al. (2004, 2008)
MLH1 & MLH3	Mismatch repair	yeast, mouse, human	MLH1 and MLH3 mismatch repair proteins function as an heterodimer which in conjunction with the MSH4-MSH5 complex are involved in the formation and stabilization of crossovers during meiosis. This pathway generates interference-dependent COs.	Kolas and Cohen (2004)

Protein name	Category	Organism	Role	References
Mus81*	CO / NO Production	most eukaryotes	Mus81* is an endonuclease, containing Mus81 and Eme1 (or in *S. cerevisae* Mms4 proteins, which is involved in the generation of interference-independent COs. In S. pombe, Mus81 generates the majority of COs.	Constantinou et al. (2002)
BLM*	CO / NO Production	yeast, human	BLM* is a complex of proteins, containing the Bloom's syndrom helicase, BLM in human (Sgs1 in *S. cerevisae*), which alongside RMI1 and TOPIII, participate in the dissolution of double Holliday Junctions and the formation of NCOs exclusively.	Wu and Hickson (2003); Ira et al. (2003); Wu and Hickson (2006)
SLX1 - SLX4 / MUS312 / Him-18 / BTBD12	CO / NO Production	most eukaryotes	These proteins act as resolvase of the dHJ, promoting the production of COs. In mutants for these proteins, the number of COs is reduced and the recombination process is delayed.	Fekairi et al. (2009); Saito et al. (2009); Svendsen et al. (2009)
PRDM9	CO / NO Production	human, mouse	PR domain containing 9 (PRDM9) is a zinc-finger protein that binds the 13-mer specific CO hotspot motif and is considered to be a key determinant of CO formation. It is also a H3K4 histone methyltransferase which has been found to associate with DSB initiation sites in *S. cerevisae*.	Myers et al. (2009); Baudat et al. (2009)

B Human recombination hotspots analyzed by sperm-typing.

Name	Intensity	Width (kb)	Location	BGC	References
DPA1	27	1.6	MHC class II (intergenic)	No	Kauppi et al. (2005)
DNA1	0.4	1.9	MHC class II (intergenic)	No	Jeffreys et al. (2001)
DNA2	8	1.3	MHC class II (intergenic)	87:13	Jeffreys et al. (2001); Jeffreys and Neumann (2002)
DNA3	100	1.2	MHC class II (intergenic)	No	Jeffreys et al. (2001)
DMB1	5	1.8	MHC class II (intron)	No	Jeffreys et al. (2001)
DMB2	45	1.2	MHC class II (intergenic)	No	Jeffreys et al. (2001)
TAP2	8	1.0	MHC class II (intron)	No	Jeffreys et al. (2001)
MSTM1a	15	1.2	Chr 1 (intergenic)	72:28	Jeffreys et al. (2005); Neumann and Jeffreys (2006)
MSTM1b	16	1.6	Chr 1 (intergenic)	No	Jeffreys et al. (2005); Neumann and Jeffreys (2006)
NID1	70	1.5	Chr 1 (intron)	74:26	Jeffreys and Neumann (2005); Jeffreys et al. (2005)
NID2a	10	1.4	Chr 1 (intron)	No	Jeffreys et al. (2005)
NID2b	4	1.1	Chr 1 (intron)	No	Jeffreys et al. (2005)
NID3	70	2	Chr 1 (intergenic)	No	Jeffreys et al. (2005)
MS32	40	1.5	Chr 1 (intergenic)	No	Jeffreys et al. (2005)
MSTM2	0.9	1.3	Chr 1 (intergenic)	No	Jeffreys et al. (2005)
SHOX	300	2	Chr X/Y PAR region	No	May et al. (2002)
β-globin	200	1.2	Chr 11 (intron)	No	Holloway et al. (2006)
PCP-1a	21	2.2	Chr 21 (intron)	No	Tiemann-Boege et al. (2006)

Name	Intensity	Width (kb)	Location	BGC	References
PCP-1b	25	0.5	Chr 21 (intron)	No	Tiemann-Boege et al. (2006)
PCP-2	77.5	1.2	Chr 21 (intron)	No	Tiemann-Boege et al. (2006)
A		1.3	Chr 21	No	Webb et al. (2008)
B		1.3	Chr 21	83:17	Webb et al. (2008)
C1		1.2	Chr 6	No	Webb et al. (2008)
C2		1.9	Chr 6	No	Webb et al. (2008)
D		1.3	Chr 6	56:44	Webb et al. (2008)
E		1.4	Chr 8	56:44	Webb et al. (2008)
F		1.4	Chr 12	No	Webb et al. (2008)
G1		1.3	Chr 8	No	Webb et al. (2008)
G2		1.5	Chr 8	No	Webb et al. (2008)
H		1.2	Chr 3	No	Webb et al. (2008)
J1		1.3	Chr 5	83:17	Webb et al. (2008)
J2		1.5	Chr 5	No	Webb et al. (2008)
K		1.4	Chr 8	No	Webb et al. (2008)
L		1.9	Chr 18	No	Webb et al. (2008)
M		1.5	Chr 2	No	Webb et al. (2008)
N		1.6	Chr 18 (intron)	No	Webb et al. (2008)
P		1.4	Chr 13	56:44	Webb et al. (2008)
Q		1.3	Chr 1	56:44	Webb et al. (2008)
R		1.9	Chr 20	No	Webb et al. (2008)
S1	0.15%	1.5	Chr 3	62:38	Jeffreys and Neumann (2009)
S2	0.1%	1.0	Chr 3	74:26	Jeffreys and Neumann (2009)
T	$\tilde{7}60$		Chr 3 (intergenic)		Berg et al. (2010)
U	$\tilde{3}50$		Chr 9 (intergenic)		Berg et al. (2010)
CF	$\tilde{8}75$		Chr 8 (intron)		Berg et al. (2010)
CG	$\tilde{6}50$		Chr 16 (intergenic)		Berg et al. (2010)
PAR2	$\tilde{2}30$		Chr X/Y PAR region		Berg et al. (2010)

C Correlations between distance to telomeres, GC*, and sex-specific COR

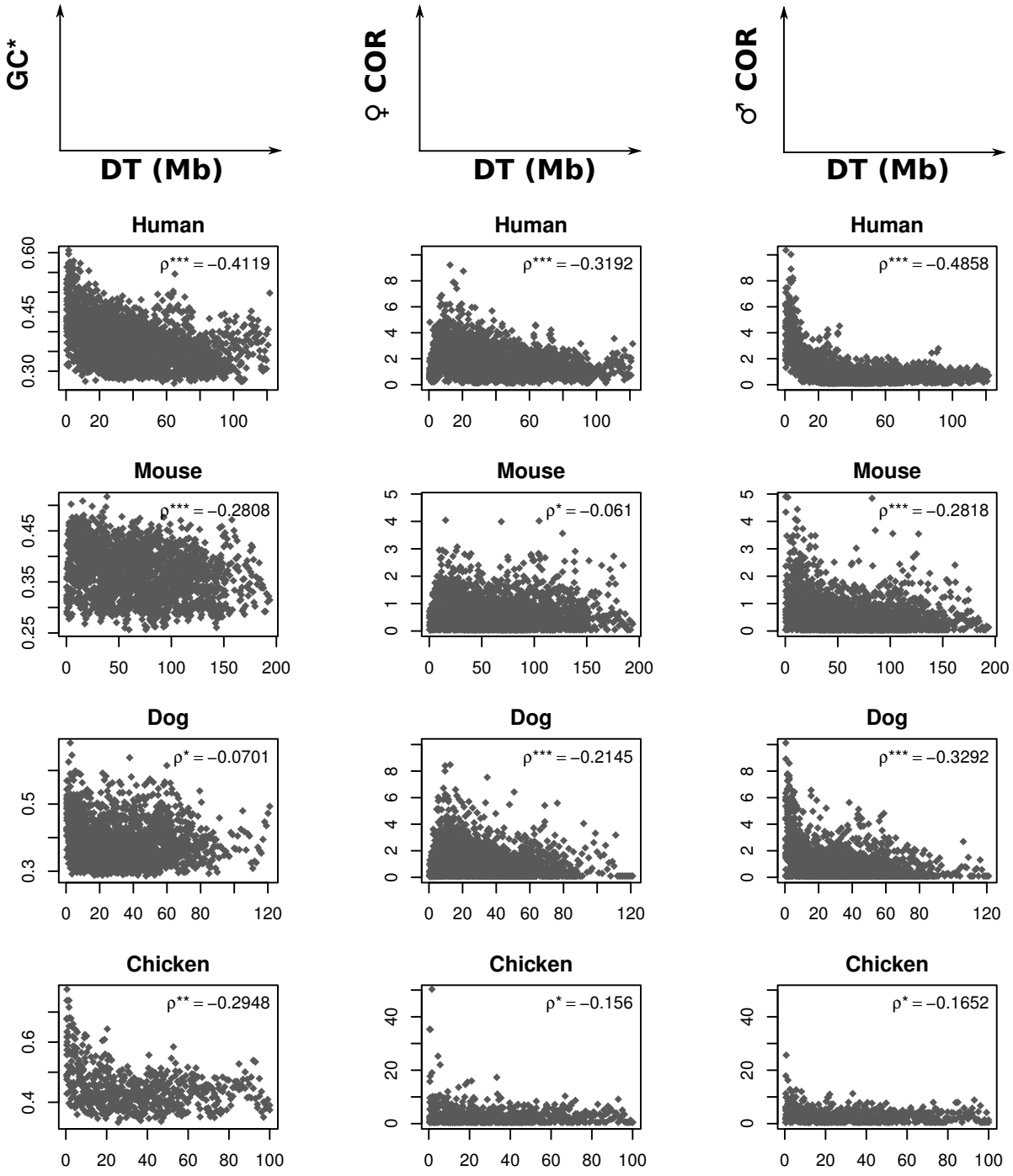

Figure C.1: *The correlations between GC*, female (♀) and male (♂) crossover rates (COR) and the distance to telomeres (DT) in human, mouse, dog and chicken. Pearson's ρ correlation coefficients are given. Three, two and one stars near these values stand for p-values of the correlation test inferior to* 10^{-16}, 10^{-10} *and* 0.05 *respectively.*

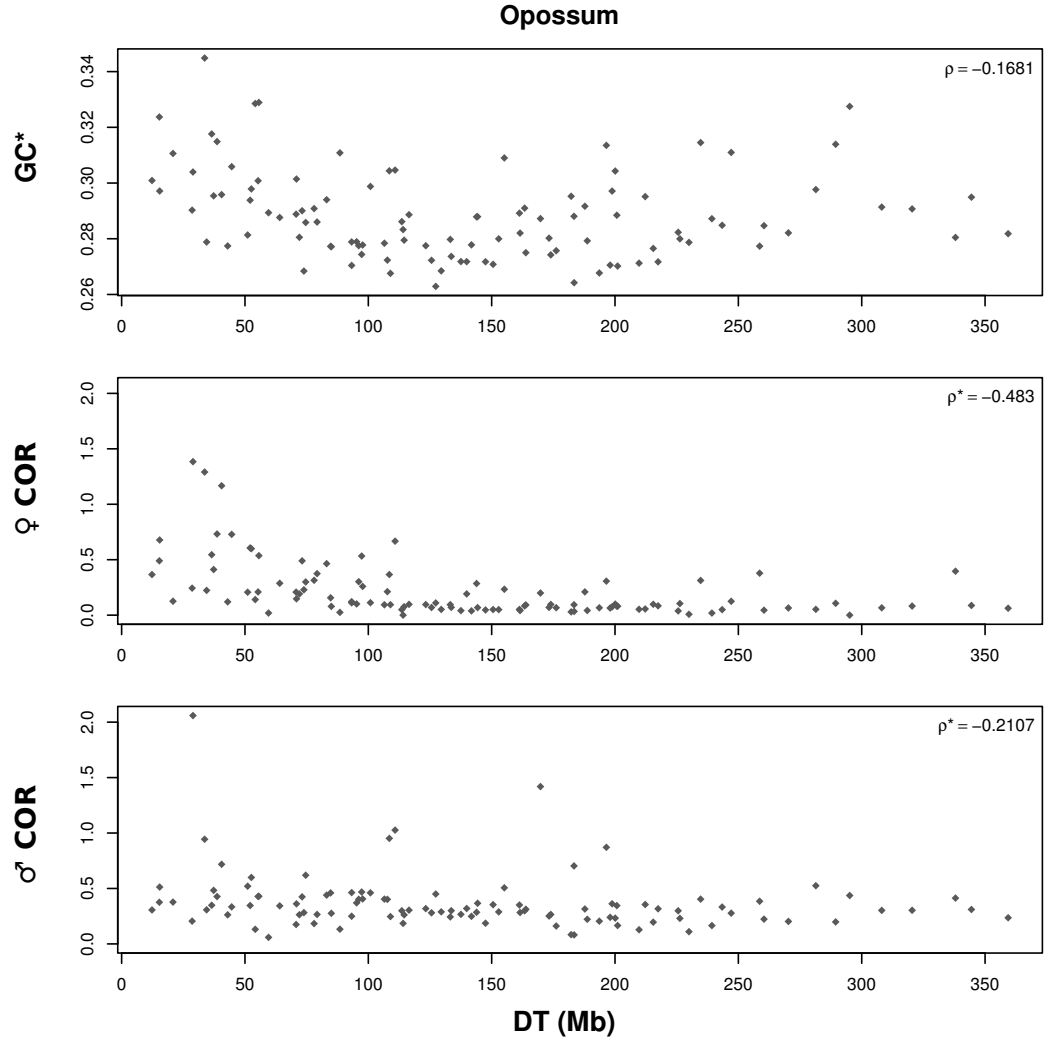

Figure C.2: *The correlations between GC*, female (♀) and male (♂) crossover rates (COR) and the distance to telomeres (DT) in opossum. Pearson's ρ correlation coefficients are given. Stars near these values stand for p-values of the correlation test inferior to* 0.05.

D Opossum correlation windows smaller and larger than 20 Mb

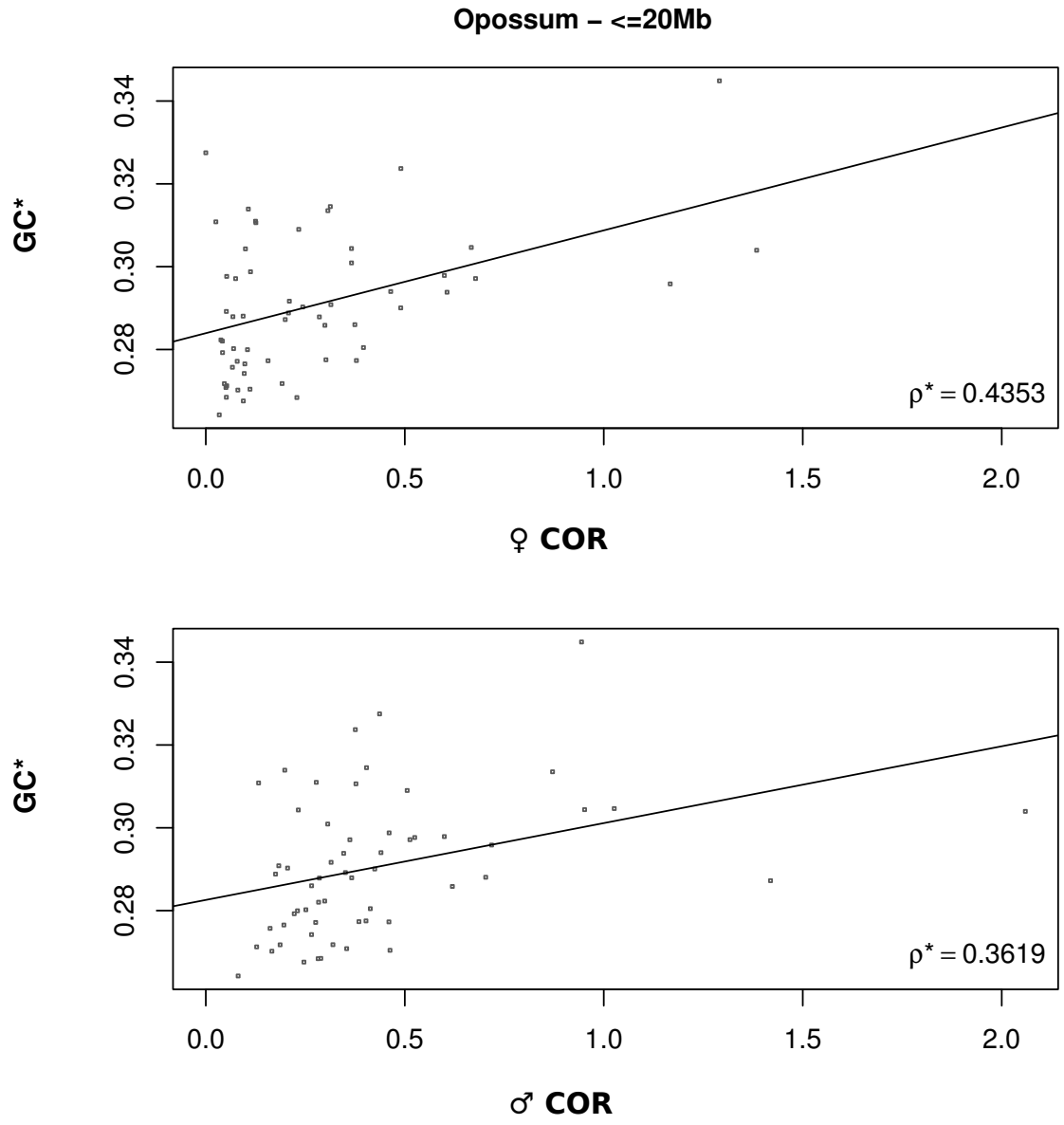

Figure D.1: *The correlations between GC*, female and male COR for window sizes inferior to 20Mb in opossum.*

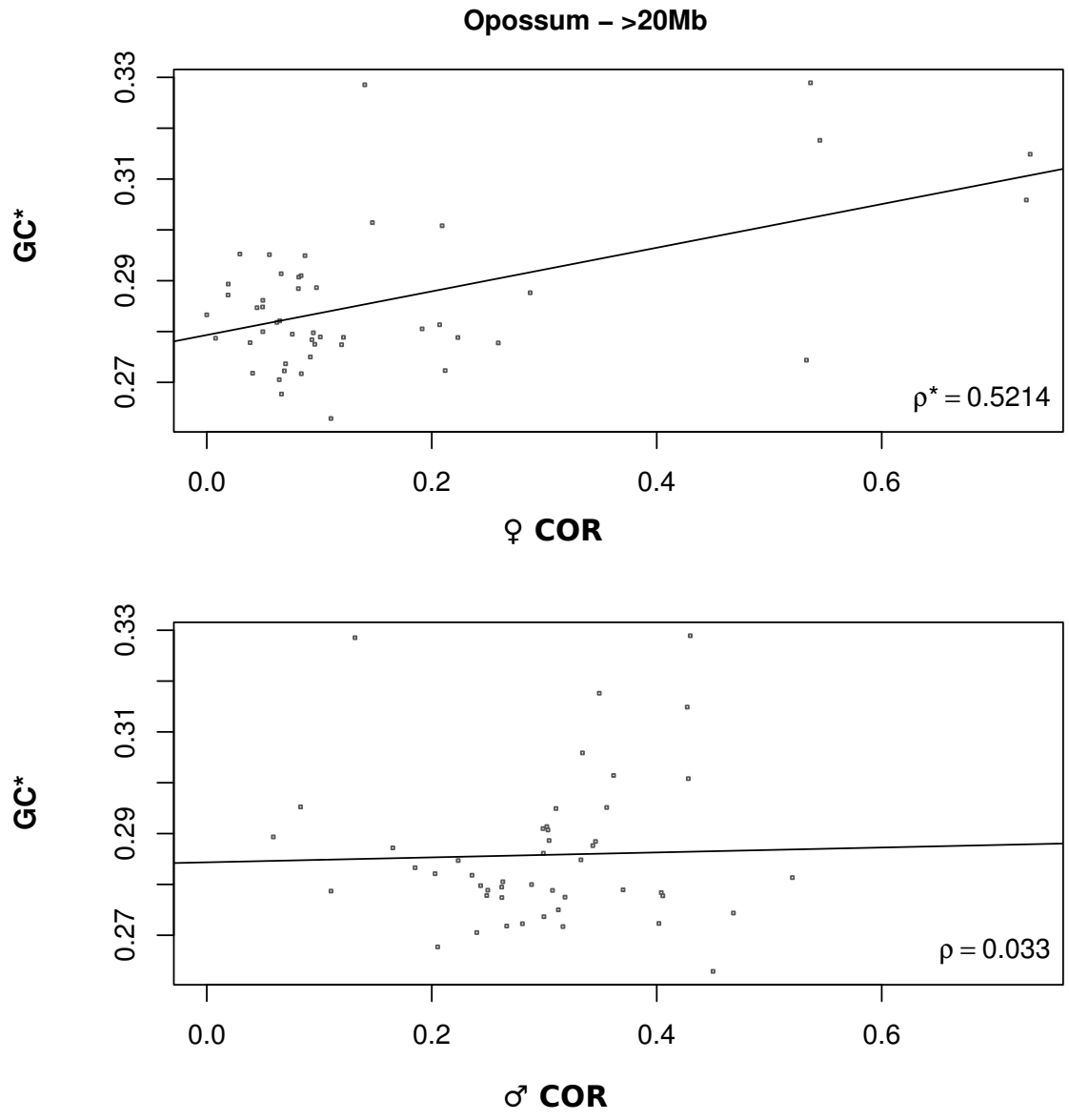

Figure D.2: *The correlations between GC*, female and male COR for window sizes superior to 20Mb in opossum.*