INFERENCE OF MOLECULAR MECHANISMS FROM SEQUENCE PATTERNS IN HUMAN DNA VARIATION

Thesis for the degree of Philosophiae Doctor (PhD)

Sigve Nakken

Centre for Molecular Biology and Neuroscience (CMBN)

Institute of Medical Microbiology, Rikshospitalet

Oslo University Hospital

Faculty of Medicine

University of Oslo, Norway

Series of dissertations submitted to the
Faculty of Medicine, University of Oslo
No. 1047

ISBN 978-82-8072-554-7

Cover: Inger Sandved Anfinsen.
Printed in Norway: AIT Oslo AS.

Produced in co-operation with Unipub.
The thesis is produced by Unipub merely in connection with the
thesis defence. Kindly direct all inquiries regarding the thesis to the copyright
holder or the unit which grants the doctorate.

ACKNOWLEDGEMENTS

The work presented in this thesis was primarily carried out in the Bioinformatics group at the Centre for Molecular Biology and Neuroscience (CMBN) and the Institute of Medical Microbiology at Rikshospitalet in Oslo, Norway. The project has received financial support from two primary sources, CMBN and the Norwegian Research Council.

Initially, I want to express gratitude to *Professor George Karypis* at the University of Minnesota in USA who introduced me to the field of computational genomics in 2003. His passion for research and science inspired me to pursue an academic career. I would further like to thank most sincerely my two supervisors, *Professor Torbjørn Rognes* and *Professor Eivind Hovig*. Most importantly, I have appreciated your encouragement and faith in me as a PhD student. Thanks to Torbjørn, for bringing me into CMBN and for several important contributions to my work. Eivind's enthusiasm, eagerness, and broad competence have most of all been a great inspiration to me, but also vital to the progress of my work. Our common interest in genetic variation and human mutation research will hopefully lead to many exciting projects in the years to come.

Next I would thank all co-authors, in particular Einar Andreas Rødland for valuable statistical guidance. Thanks to my closest colleague, Dr. Gard Thomassen, for much fun during the last years, both inside and outside of the office. At times during my PhD education, I have worked as a software developer in the companies of Sencel and PubGene. I wish to thank all colleagues in these companies for an exciting time. It has further been a true pleasure to contribute to the ongoing research at Spermatech AS, which has provided me with hands-on experience with germ cell biology. I have very much appreciated collaborative research projects with other groups. Thanks to Jarle Breivik (Institute for Basic Medical Sciences, UiO), Trond Paul Leren (Institute of Medical Genetics, Rikhospitalet), Robert Lyle and Dag Erik Undlien (Institute of Medical Genetics, Ullevål Universitetssykehus), Esten Nakken (my oldest brother, Institute for Experimental Medical Research, Ullevål Universitetssykehus), Ole Petter Ottersen and Erlend Nagelhus (Institute for Basic Medical Sciences, UiO), and Torstein Tengs (National Veterinary Institute).

Finally, I would thank my friends and closest family. A particular thanks goes to my dear girlfriend Siw for her laughter, smiles, and unconditional support during the work with this thesis.

Oslo, November 2010

Sigve Nakken

TABLE OF CONTENTS

LIST OF PAPERS

I. Large-scale inference of the point mutational spectrum in human segmental duplications

Nakken S, Rødland EA, Rognes T, Hovig E

BMC Genomics. (2009) **10**:43.

II. The disruptive positions in human G-quadruplex motifs are less polymorphic and more conserved than their neutral counterparts

Nakken S, Rognes T, Hovig E

Nucleic Acids Res (2009) **37**(17):5749-56

III. Impact of DNA physical properties on local sequence bias of human mutation

Nakken S, Rødland EA, Hovig E

Manuscript.

IV. Unstable DNA repair genes shaped by their own sequence modifying phenotypes

Falster D, Nakken S, Bergem-Ohr M, Rødland EA, Breivik J

J Mol Evol (2010) **70**(3):266-74

LIST OF ABBREVIATIONS

5mC	5-methylcytosine
BER	base excision repair
BGC	biased gene conversion
CNV	copy number variant
CPD	cyclobutane pyrimidine dimer
dbSNP	database of single nucleotide polymorphism at NCBI
DIM	duplication-inferred mutation
DSB	double strand break
EST	expressed sequence tag
GWAS	genome-wide association study
HGP	the human genome project
HGSDB	human genome segmental duplication database (TCAG, Canada)
Indel	insertion-deletion
LINE	long interspersed repetitive element
MMR	mismatch repair
MSV	multisite variant
NCBI	national center for biotechnology information (USA**)**
PP	pyrimidine-pyrimidone (6-4) dimer
PSV	paralogous sequence variant
RMF	relative mutation fraction
ROS	reactive oxygen species
SAM	S-adenosylmethionine
SINE	short interspersed repetitive element
SNP	single nucleotide polymorphism
STR	short tandem repeat
TCR	transcription-coupled repair
TSS	transcription start site
UTR	untranslated region
VNTR	variable number tandem repeat

1 INTRODUCTION

> *Mutation fascinates because of its three faces: the variability it generates that conditions all evolutionary change, the disease it generates that consumes a substantial proportion of our resources, and the means it offers for dissecting all facets of biological phenomena.*
>
> *John W. Drake (1991)*

Deoxyribonucleic acid (DNA) is a biological macromolecule that is vital to all known living organisms. All living cells on earth store their hereditary information in the form of double-stranded DNA molecules. DNA molecules are long polymer chains that are formed from a restricted set of monomers. Each monomer in a single DNA strand consists of two parts, a sugar (deoxyribose) with a phosphate group attached to it, and a unique base. The four unique bases found in DNA include adenine (A), guanine (G), cytosine (C) and thymine (T). Though principally encoding epigenetic information, the biological importance of 5-methylcytosine (5mC) is so great that many scientists currently think of it the as "the fifth base". The most important property of DNA lies within this five-letter alphabet of chemical bases, which have the remarkable capacity of encoding all the essential instructions that are needed for the cell to grow, function, and reproduce.

Following developments in DNA sequencing technology in the early 1970's, determining the sequence of bases in a DNA molecule has now become a standard biochemical technique [Shendure and Ji 2008]. As a result of great collaborative sequencing efforts, initiated both within the private industry and through government funding, a draft of the human genome sequence was published in 2001 [Lander et al. 2001; Venter et al. 2001]. The draft sequence contained approximately 3.2 billion bases in total. The quality of the sequence is still being improved on a continuous basis. The establishment of a reference genome was initially of great importance for the identification and annotation of genes and other functional elements in our DNA. Additional sequencing of other mammalian genomes, such as the chimpanzee and mouse genomes, opened up a new era for comparative genomics research [Mouse Genome Sequencing Consortium 2002; Boffelli et al. 2004; Chimpanzee Sequencing and Analysis Consortium 2005]. At the same time, huge efforts were initiated towards an understanding of human genome variation, that is, the interindividual DNA variations that are present at significant frequencies in the human population. Genome-wide polymorphism data quickly revealed substantial genetic

differences among humans [Sachidanandam et al. 2001; Hinds et al. 2005; International HapMap Consortium 2005].

The current ability to perform high-throughput discovery and genotyping of genomic DNA variants provide valuable data that holds great potential for an increased understanding of human genetics. The greatest expectations have probably been for our understanding of the genetic constituents of human disease. The large number of accessible human polymorphisms has the potential of revealing novel associations between variation in the primary DNA sequence and the susceptibility of clinical phenotypes. In contrast to the low frequent, gene-specific variants that underlie Mendelian disorders, it has been suggested that complex disease phenotypes could be explained by multiple DNA variants that act in concert with environmental influences [Bodmer and Bonilla 2008; Manolio et al. 2008]. It is also believed that many of these variants exert their effect on the phenotype in a regulatory manner, rather than through alteration of the coding DNA sequence and the encoded protein [Pastinen and Hudson 2004; Knight 2005]. Several genome-wide association studies have within the last very few years identified many risk variants associated with complex diseases [Wellcome Trust Case Control Consortium 2007; Frazer et al. 2009]. Other disciplines within human genetics have also experienced great benefits from the accumulation of DNA variation data. Specifically, variation data provide indications to the impact of natural selection in the genome and may thus highlight functionally important genes that have been targets for adaptive evolution [Akey et al. 2002; Biswas and Akey 2006]. Estimation of evolutionary pressure by means of the allele frequency spectrum has further been shown to be powerful in studying the biological function of specific genomic elements such as gene promoters and microRNAs [Chen and Rajewsky 2006; Sethupathy et al. 2008].

The discovery of DNA variants in the human genome also brings forth new opportunities for large-scale studies of human mutation. The different types of DNA sequence variants are many and complex, ranging from single nucleotide polymorphisms (SNPs) to structural variations at the kilobase level [Feuk et al. 2006]. The most common variants are by far SNPs, originally defined as polymorphic sites in which the minor allele is present with a frequency of at least 1% in a given population. SNPs are caused by the most frequent mutation events, DNA point mutations. The focus of the work presented in this thesis is the relationship between genetic variation and the molecular mechanisms that underlie their genomic distribution and DNA sequence specificity. Insight into this matter bears importance not only for our understanding of the mechanisms that cause genetic

disease and the progression of cancer, but also for the processes that fuels DNA evolution.

Which mechanisms determine the patterns of genetic variation? Or more specifically, what generates human SNPs? Figure 1 illustrates how the densities of genetic variants vary along human chromosome 1. Generally, genetic differences between human individuals are considered to have originated through germline mutations in nuclear DNA of specific individuals. Downstream of the initiating mutation event, the fate of the mutant allele (i.e. its allele frequency in the gene pool) will be determined by the interaction of two evolutionary forces: random genetic drift and natural selection [Serre and Hudson 2006; Hurst 2009]. Considering the initiating DNA mutation events only, it is clear that these could be triggered by factors of both endogenous and exogenous origins. Studies in model organisms have demonstrated that cellular exposure to elements such as ionizing radiation and ultraviolet light, alkylating agents, anticancer drugs, and other mutagenic chemicals may induce mutations and chromosome abnormalities in DNA [Auerbach and Kilbey 1971; Miller 1985; Murnane 1996; Witt and Bishop 1996; Nohmi and Masumura 2005]. However, none of these exogenous factors seem to underlie the spectrum of mutations causing human genetic disorders [Otake et al. 1990; Czeizel et al. 1991; Rudiger 1991; Byrne et al. 1998]. Considerable evidence points on the other hand towards a dominant role for endogenous mechanisms in the generation of human mutations [Cooper and Krawczak 1993]. Here, a complex collection of factors comes into play. First of all, it has been recognized that normal cellular processes introduce DNA lesions, and thus exhibit a mutagenic potential. One primary example involves the error-prone incorporation of nucleotides by DNA polymerases, a process that is central in both replication and repair [Kunkel 2004]. Another example involves the necessity for double-strand breaks (DSBs) during meiotic recombination [Duret and Galtier 2009]. An additional source of endogenous damage is the spontaneous base alterations that follow from continuous chemical reactions of DNA with oxygen and water (e.g. methylation and deamination). Such modifications can introduce base mispairs in the double helix and lead to mutations in subsequent rounds of DNA replication. Both the DNA backbone and the different bases are furthermore attacked by reactive oxygen species (ROS), which are produced frequently during normal cellular metabolism [Lindahl 1993]. Secondly, although cells are equipped with an advanced repertoire of DNA repair pathways that counteracts the majority of lesions, it is apparent that the effectiveness and specificity of repair enzymes may vary in different environments, presenting an additional endogenous threat to genetic integrity [Rajski et al. 2000; Jackson and Bartek 2009]. Last but not least, the physical and topological characteristics of a

genomic sequence context along with its higher-order chromatin structure and epigenetic modifications are potentially important modulators of genetic variation [Liu et al. 2007; Cummings et al. 2008; Lieberman-Aiden et al. 2009; Sasaki et al. 2009].

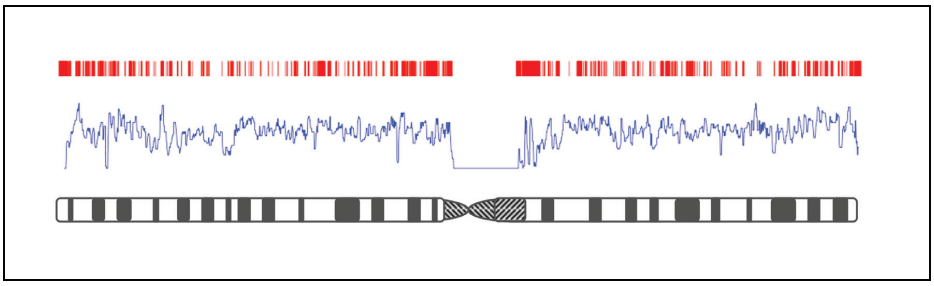


Figure 1: Presence of copy number variants (in red; from Database of Genomic Variants (http://projects.tcag.ca/variation/)) and density of biallelic single nucleotide polymorphisms (in blue; validated set from dbSNP (build 130, April 2009)) along human chromosome 1. Note that chromosome bands are for illustration purposes only.

These properties constitute important dimensions in the interactions between DNA, damaging agents and repair enzymes. In summary, it is clear that the patterns of genetic variation are determined by a complex interplay between a range of molecular factors and basic evolutionary forces. However, although mechanisms underlying many mutagenic processes have been elucidated in great detail, the relative impact of the different molecular factors to the observed spectrum of genomic variation is less clear, and warrants further study.

It has been hypothesized that the efficacy of many DNA interactions in the cell are governed by local sequence specificities. Indeed, it has been confirmed in many instances that such dependencies exist. Distinct sequence preferences have been observed with respect to DNA adduct formation, polymerase base misincorporations, and successful repair by base excision and mismatch repair enzymes [Kunkel and Alexander 1986; Eftedal et al. 1993; Cloutier et al. 2001; Donigan and Sweasy 2009]. Based on these findings, one may suspect that predominant sequence dependencies have left historic fingerprints in the current DNA variation patterns, visible as local sequence hotspots (and coldspots) of mutation. Moreover, in addition to local dependencies, it is evident that the endogenous factors that modify or interact with DNA display large-scale regional biases along chromosomes. This has been illustrated in genome-wide studies of oxidative damage patterns, DNA methylation profiles, recombination frequencies and DNA replication timing [Myers et al. 2005; Ohno et

al. 2006; Karnani et al. 2007; Lister et al. 2009a]. Hence, the mechanisms underlying DNA variation patterns appear to have both local and global dimensions in the human genome. This conjecture forms the basis of the topic addressed in this thesis. We employ computational genomics techniques to quantify the distribution and sequence patterns in genome-wide spectra of DNA variation, and attempt to associate these patterns to underlying mutational mechanisms. The primary goal is to increase our understanding of the relative contributions by different mutational mechanisms at work in the human genome. A secondary goal is to demonstrate the usefulness of adopting a genetic variation and mutation perspective in the analysis of genome function and evolution.

The rest of this introductory chapter is organized as follows. Initially, a brief historical perspective is provided. This perspective highlights scientific discoveries that have been crucial for our current comprehension of genetic variation and how the genetic material is able to undergo change through mutation. Next, we illustrate and describe the various types of DNA sequence variants that have been discovered in the human genome. Further, the public resources and databases that are available for studying human genetic variation are presented. At the end of the chapter, we define the scope of the work that is part of this thesis.

1.1 HISTORICAL PERSPECTIVE[1]

More than a century ago, long before it was recognized that DNA is the hereditary material, scientists discovered basic mechanisms of heredity that has been vital to the present understanding of genetic variation and human mutation. At the time in which Mendel's proposed hereditary laws were rediscovered in independent studies of flowering plants [Correns 1900; Tschermak 1900], evidence had also been brought forward as to how the hereditary material could be governed by thread-like structures within the cell-nucleus [Flemming 1882; van Beneden 1883]. The discovery of chromosomes and an understanding of their division process during mitosis and meiosis led Sutton and De Vries to formulate a chromosomal theory of Mendelian heredity [De Vries 1903; Sutton 1903], a theory that

[1] The brief historical perspective presented here is to a large extent based on two different sources. Further details can be found in the comprehensive overview given by Whitehouse [Whitehouse 1973], and in the introductory chapter of Human Gene Mutation [Cooper and Krawczak 1993].

developed further to include the physical event of chromosomal crossovers at chiasmata [Janssens 1909].

The notion of the gene as a hereditary unit and the recognition that genes apparently could contain multiple alleles (a well-known example being the human blood-groups [Bernstein 1925]) launched the concept of mutation. Originally introduced by De Vries after his extensive studies of *Oenothera lamarckiana* (Evening primrose) [De Vries 1901], mutations were considered to be the mechanistic explanations for the sudden appearances of new forms that arose in each generation. The rate at which mutations occurred in the genetic material, and the factors that influenced this rate, was subsequently pioneered by the work of Muller in the late 1920's, using *Drosophila melanogaste*r (Fruit fly) as the model organism. Muller devised novel genetic techniques that crossed male flies with female flies containing a recessive lethal gene in one of their X-chromosomes, enabling the detection of lethal mutations in the X-chromosome of male *Drosophila* trough the presence of grandsonless-lineages [Muller 1927; Muller 1928a]. An incorporation of a dominant gene in the female chromosome carrying the lethal gene made it further possible to measure the rate of non-lethal mutations. It became apparent that the number of lethal mutations greatly outnumbered the non-lethal ones. Muller demonstrated that a rise of temperature, as well as radiation by X-rays, produced significant increases in the mutation rate [Muller 1928b]. Although Muller's results pointed to a linear relationship between X-ray dose and the number of induced mutations, it also became evident that different genes mutated with different frequencies. Furthermore, it was shown that X-rays could cause structural changes and rearrangements within chromosomes, revealed through linkage studies following irradiation [Muller and Altenburg 1930]. Some fifteen years after Muller's outstanding experiments with X-rays, studies provided the first evidence for a chemical induction of mutations. Auerbach and Robson demonstrated that $\beta-\beta'-\text{dichloro}-\text{diethyl}-\text{sulphide}$ (mustard-gas) was highly mutagenic [Auerbach and Robson 1947]. Subsequently, a wide range of chemicals was found to cause mutation, including urethane, nitrogen mustard, and hydrogen peroxide. Importantly however, both X-rays and chemicals produced a wide range of mutation types, and no evidence was brought forward at the time for specific types of mutations induced by specific mutagens.

Through the study of genetic transformation in *Diplococcus pneumoniae*, it became evident that the hereditary material must be carried by DNA alone [Avery et al. 1944]. Nearly a decade later, Watson and Crick's discovery of the double helical structure of DNA provided an elegant model for the gene and how genetic information is replicated [Watson

and Crick 1953]. The chemical model of DNA paved the way for an understanding of mutations in terms of DNA structure and replication. Studies in bacteriophage T4 demonstrated that DNA replication could represent an endogenous origin of mutations, when base analogues (i.e. 5-bromouracil) were incorporated erroneously opposite the regular DNA bases [Benzer and Freese 1958]. Importantly, the mutagenic effect of 5-bromouracil was highly non-random, causing mutation at sites with characteristic frequencies. Other mutagens, such as proflavin, 2-aminopurine, and hydroxylamine, displayed different, yet also distinct patterns [Brenner et al. 1958; Freese et al. 1961]. These studies demonstrated for the first time that the actions of mutagens are specific for certain sites in DNA. Notably, the mutations that occurred spontaneously without any exposure to mutagens also appeared to be clustered in hotspots [Freese 1959; Benzer 1961].

Elucidation of the key mechanisms and pathways underlying DNA repair, initially in studies of *E. coli* and subsequently in eukaryotic cells, introduced a new dimension in mutation research (a comprehensive overview is presented by Friedberg et. al [Friedberg et al. 2006]). It was realized that the mutagenic potentials of different DNA lesions is guided to a significant extent by the efficiency of repair. In addition to repair, it became apparent that many lesions were replication-dependent [Drake and Baltz 1976]. The importance of DNA polymerase fidelity was recognized, that is, how accurately the polymerase selects bases during DNA synthesis [Kunkel and Loeb 1981]. Different models of polymerase-induced errors were proposed, suggesting that a physical misalignment of template and primer could introduce mispairs during replication of repetitive or palindromic DNA sequences [Streisinger et al. 1966; Ripley 1982; Kunkel and Soni 1988].

1.2 DNA SEQUENCE VARIANTS

DNA sequence variants may be gross, i.e. at the level of the chromosome, or very small, at the single base pair level. The complex spectrum of observed DNA variants in the human genome illustrates a great mechanistic variety in which mutations can alter the genetic material. Although the focus here lies on single point mutations and human SNPs, it seems relevant to briefly characterize the full spectrum of common variants. The nomenclature used to describe sequence variants is not completely standard, and different terms are usually adopted depending on phenotype or the relative population frequency of the variant (i.e. *mutation* versus *polymorphism*). Here, the main purpose is to illustrate the spectrum based on the sequence alterations they cause in DNA. We have therefore decided to merely

use the term *variant* for the different types of variations depicted in Figure 2. In Figure 2, chromosomal copies that belong to two different individuals (red and blue face) are shown for each type of variant.

Figure 2: Examples of sequence variants in the human genome. Variant sequence is shown in green colour, fixed duplicated sequence in red.

Single nucleotide variants (2A) denote single base, allelic differences between members of a species (or between paired chromosomes in an individual). They represent the most common type of allelic variation in the human genome. The variant alleles of a particular site are collectively referred to as a SNP if the minor allele is present at a significant frequency (>= 1%) in the population. The single point mutations that lead to

single nucleotide variants can be subclassified into transition mutations and transversion mutations (Figure 3). Transitions involve a change of one purine for another, or a pyrimidine for another. Transversions involve the interchange of a purine for a pyrimidine, or the interchange of a pyrimidine for a purine. In humans and most other species, transitions dominate the point mutational spectrum.

The potential molecular effects of a particular sequence variant depend on where in the genome it occurs. If a variant occurs in the coding sequence of a gene it can: 1) change the codon for an amino acid to another (non-synonymous or missense change; causing alteration of protein function or stability [Wang and Moult 2001; Ng and Henikoff 2006]), 2) change the codon for an amino acid to a stop codon (nonsense change; leading to premature termination of protein synthesis [Chang and Kan 1979; Cooper 1993]), or 3) change the codon for an amino acid to a different codon for the same amino acid (synonymous or silent change; causing no alteration of the encoded protein but may affect splicing or mRNA stability [Cartegni et al. 2002; Wang et al. 2005]).

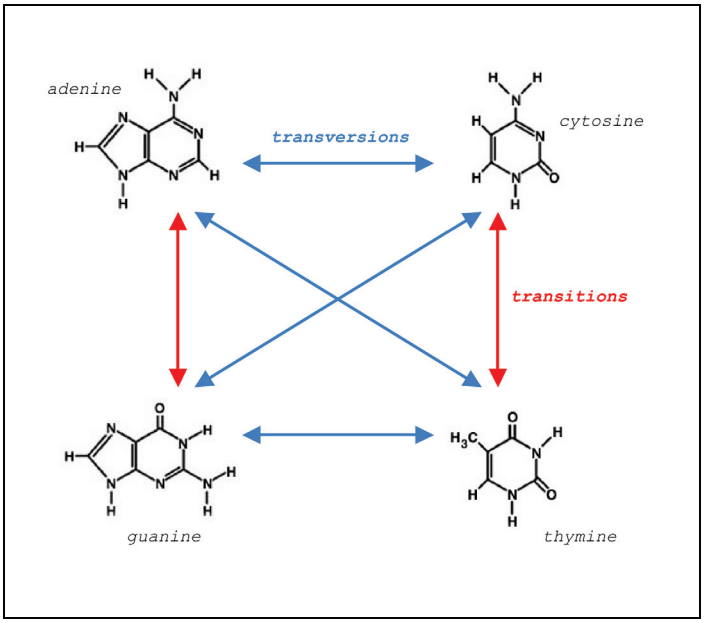


Figure 3: Base changes in DNA can be classified as transitions and transversions.

Gene sequence variants that occur outside the protein-coding part may also have significant molecular effects. For example, these variants can disrupt the promoters or regulatory elements of genes and thereby alter gene expression levels [Hoogendoorn et al. 2003].

Intronic variants at the exon-intron junction may cause defective splicing [Busslinger et al. 1981; ElSharawy et al. 2009] and variants in *cis*-acting elements in untranslated regions (UTRs) may alter properties such as polyadenylation, mRNA stability, subcellular location and translation efficiency [Conne et al. 2000; Jansen 2001; Chen et al. 2006; Wang et al. 2006]. The function of microRNAs, which regulate a substantial proportion of protein-coding genes, may also be impaired by specific sequence variants [Abelson et al. 2005]. Importantly however, most common single nucleotide variants observed in the human genome are generally considered to cause no particular effects on the cellular phenotype, since the majority occurs in the large proportion of the genome that contains apparently non-functional DNA [Levy et al. 2007].

Insertion-deletion variants (2B) or *indels* are sequence differences in which a relatively small piece of DNA has been inserted or deleted. The majority of human indels have been found to be microscopic, with lengths <= 10bp [Mills 2006; Levy et al. 2007]). These variants may have a variety of phenotypic consequences depending on where in the genome it occurs, and many of these are similar to the effects observed for single nucleotide variants. Although highly repressed in protein-coding DNA [Clark et al. 2007], indels may cause a shift in the translational reading frame (frameshift mutations). Frameshift variants may have significant effects on protein function, and some human variants have been associated with disease phenotypes [Yeo et al. 1998; Ogura et al. 2001]. Furthermore, studies in non-coding DNA have demonstrated that short indels are frequently part of a tandemly repeated DNA sequences [Messer and Arndt 2007]. These may undergo mutation via physical slippage during DNA replication or unequal crossing-over at recombination [Levinson and Gutman 1987; Kondrashov and Rogozin 2004; Kvikstad et al. 2007]. In fact, deletions and insertions of tandem repeats lead to a distinct class of variation known as *variable number tandem repeats (VNTRs),* in which the length of the tandem repeat varies between individuals [Jeffreys et al. 1988]. VNTRs form the basis of DNA fingerprinting, which is a technique used to uniquely identify individuals based on their DNA sequence profile [Jeffreys et al. 1985a; Jeffreys et al. 1985b]. There are two principal families of VNTRs, minisatellites and microsatellites, which essentially differ with respect to the length of the repeat. The instability of trinucleotide repeats in germline DNA is further responsible for a broad class of genetic disorders [Pearson et al. 2005]. The most well-known and severe example is probably Huntington's disease, characterized by an expansion of CAG (polyGlu) repeats in the coding regions of the Huntingtin gene [Walker 2007]. A different class of common repeat structures also produces genetic variation. Transposons and transposon-like

repetitive elements, such as SINEs (short interspersed repetitive elements) and LINEs (long interspersed repetitive elements), collectively occupy more than 40% of the human genome [Lander et al. 2001]. They represent mobile genetic elements that serve as an ongoing source of genetic variation [Dewannieux et al. 2003; Bennett et al. 2004; Mills et al. 2007].

Copy number variants (CNVs, 2C) generally encompass sequence variations in which individual chromosomes carry a relative gain or loss of a large (>1kb) DNA segment [Feuk et al. 2006; Freeman et al. 2006]. These variants have been shown to be present at significant frequencies in healthy human individuals [Redon et al. 2006; Kidd et al. 2008], but their associations to human diseases indicate that they also may convey distinct clinical phenotypes [Sebat et al. 2007; McCarroll et al. 2008]. Diseases related to copy number variations may arise via gene function defects following deletions of a coding gene sequence, or through abnormal expression levels following duplications in dosage-sensitive genes [Lupski and Stankiewicz 2005]. Although CNVs are far less frequent than SNPs in the human genome, the full amount of variant sequence generated by CNVs is considerably larger than for the total set of SNPs (approximately 30% of the human genome sequence is affected by CNVs as opposed to a mere 1% for SNPs [Zhang et al. 2009]). It should be mentioned that CNV discovery and analysis is a relatively new discipline compared to SNP research, and that the phenotypic influences by CNVs are only beginning to be explored.

Single base differences between duplicated segments of the human genome lead to another type of sequence variation that has been coined *paralogous sequence variants* [Bailey et al. 2002]. In Figure 2D, an example of an intrachromosomal segmental duplication is shown in red, and the paralogous variation are indicated as single base differences. This kind of variation forms when either of the two duplications is mutated. In contrast to ordinary allelic polymorphisms that constitute variation *between* genomes, these variants primarily represent differences *within* identical parts of the genome. Many of these variants could be fixed (i.e. no variation between individuals, and hence the identical sequence and variants in both individuals depicted in Figure 2D).

Other and less common variants that have not been exemplified in Figure 2 include orientational variants (e.g. inversions) and positional variants (e.g. translocations). An *inversion variant* is an example of an orientational variation, being a variant in which the order of the base pairs is reversed in a defined section of a chromosome [Sharp et al. 2006]. A well-characterized inversion variant in humans involves a section of chromosome 17 in which a ~900kb segment is in the reverse order in approximately 20% of individuals with Northern European ancestry [Stefansson et al. 2005]. A *translocation variant* occurs when

parts of nonhomologous chromosomes are rearranged, a phenomenon that is frequently observed in human cancers [Rabbitts 1994]. Translocations may produce fusion genes, which are hybrid genes made up of otherwise separated genes [Eichler 2001].

1.3 RESOURCES FOR THE STUDY OF HUMAN DNA VARIATION

There have been significant improvements in DNA sequencing and genotyping technology during the last decade. The so-called high-throughput technologies produce vast amounts of sequence and genetic variation data. Fortunately, the establishments of centralized repositories of genetic variation have accompanied the data increase. The free availability of genetic variation data trough online databases has proven to be a success factor in many respects. SNP databases have not only enabled rapid progress in mapping complex disease traits, but also served as a great resource for evolutionary and population genetics research.

In this section, available resources for the analysis of genetic variants in the human genome will be presented. Emphasis will be put on the two primary resources that are dedicated towards SNPs, the dbSNP database and the HapMap database.

dbSNP

The database of single nucleotide polymorphisms at NCBI, USA (dbSNP) is the principal database of SNP information [Sherry et al. 1999; Sherry et al. 2001]. It is publicly available via web and file transfer protocol (http://www.ncbi.nlm.nih.gov/SNP). The collection of polymorphisms in dbSNP includes SNPs, small-scale indel polymorphisms as well as variations in repeats (i.e. short tandem repeats (STRs)). The database generally holds variations in any species, and from all parts of the genome. Since the release of dbSNP in late 1998, the size of the database has increased rapidly owing to continuous submissions of variation data from academic research laboratories and private research companies. Outputs from large-scale SNP discovery and validation initiatives such as The SNP Consortium [Weissman et al. 2001], the Perlegen genotyping initiative [Hinds et al. 2005], the HapMap project [International HapMap Consortium 2003] (see below), and most recently the 1000 Genomes project [Kaiser 2008], are all merged into nonredundant SNP clusters in dbSNP. The nonredundant SNP clusters (identified through a unique rsID) are established by way of a dbSNP clustering procedure of all submitted variants, ensuring that the same SNP discovered from independent laboratories are annotated with the same rsID. Each entry (i.e. rsID) in the dbSNP database includes the sequence context of the polymorphism, its

occurrence frequency (by population or individual, if available), and the experimental methods and protocols used to assay the variation. The variation data are regularly updated in synchrony with genome rebuilds, ensuring the highest quality of SNP locus mapping and scrutiny.

A key concern with the content in dbSNP is the quality of the millions of variant entries in the database. It is evident that the database holds a great number of sequence variants that have been predicted rather than discovered experimentally. Historically, computational and probabilistic techniques have been a dominant force in the detection of variants that have been submitted to dbSNP. These techniques were largely based on advanced screenings of sequence alignments between highly identical expressed sequence tags (ESTs) or shotgun reads [Buetow et al. 1999; Altshuler et al. 2000; Irizarry et al. 2000; Marth et al. 2001]. Although such *insilico* SNP discovery proved powerful in many respects, their specificity was not optimal. Sequencing errors in the raw sequence tags or shotgun reads probably introduced a significant number of false positive SNPs [Platzer et al. 2006]. The duplicated nature of the human genome was also not particularly well handled by these SNP discovery schemes, introducing an additional element of noise in the set of reported SNPs [Nakken et al. 2009a; Musumeci et al. 2010]. As a response to the need for more extensive quality control of the database, dbSNP has developed a validation scheme for variant entries. Currently, a SNP is considered to be valid if it satisfies either of the following criteria:

i. The SNP contains at least two independent submissions, of which one is assayed by a non-computational technique (‘byCluster’)
ii. The SNP contains allele frequencies and genotype frequencies provided by a submitter (‘byFrequency’)
iii. Both alleles of the SNP have been seen at least twice in an experimental assay (‘by2Hit2Allele’)
iv. The SNP has been genotyped by the HapMap project or another large-scale project (‘byHapMap’ or ‘byOtherPop’)

In addition to criteria above, dbSNP reports whether a polymorphism maps uniquely or not in the human genome. It has been shown that human SNPs that have been validated in this manner have a significantly higher probability of being truly polymorphic than the non-validated ones [Edvardsen et al. 2006]. In 2004, a large-scale validation study of more than

220,000 dbSNP entries was reported [Nelson et al. 2004]. It was demonstrated that approximately 80% of the SNPs that have been annotated as validated were common (minor allele frequency >= 0.1) in Caucasians. This number dropped to ~50% when validation status were not taken into account. Still, there is a great gap between the total set of unique SNP clusters and the set of SNP clusters that has been validated. As of April 2010, there are approximately 14.7 million nonredundant SNPs that have a positive validation status, whereas the total number of nonredundant SNPs in the database amounts to 23.7 million (illustrated in Figure 4).

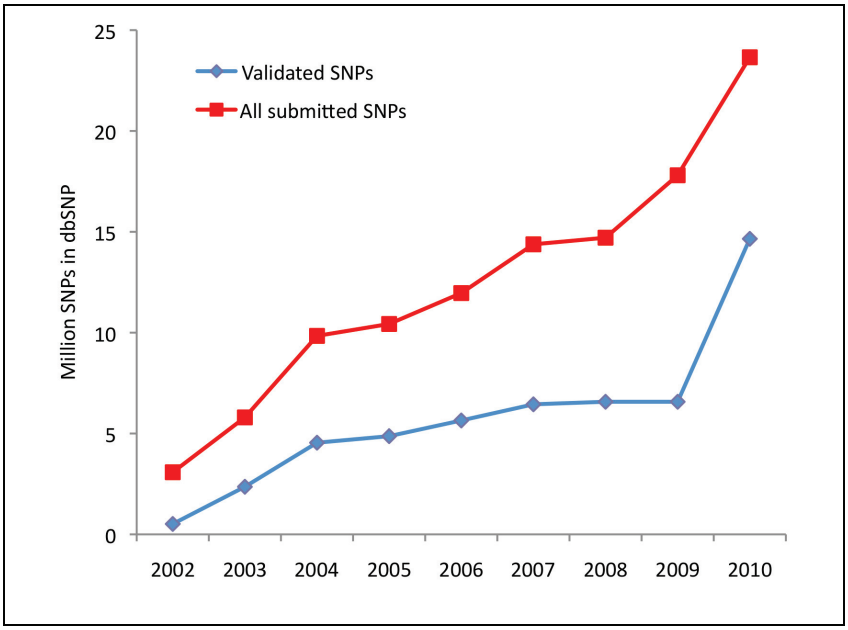


Figure 4: Increase in total number of nonredundant human SNP entries in dbSNP during the last eight years. Note the recent peak, likely resulting from increased discovery enabled by high-throughput sequencing.

By means of sequence comparisons with the genomes of chimpanzee and orangutan (or macaque), it is currently possible to infer the ancestral and derived alleles for most human SNPs. In this manner, one can establish the direction of the underlying nucleotide substitution. One of the two SNP alleles along with its neighbouring sequence are likely to map uniquely to orthologous regions in chimpanzee and orangutan, and the direction of substitution in the human lineage can further be inferred through a maximum parsimony

approach [Jiang and Zhao 2006]. Orthologous alleles for human SNPs are currently available via the UCSC Genome Browser data resources [Thomas et al. 2007].

The HapMap project

The International HapMap Project was designed to create a public, genome-wide database of patterns of common human sequence variation to guide genetic studies of human health and disease [International HapMap Consortium 2003; Manolio et al. 2008]. As a result of genotyping more than 3 million SNPs in four different populations, the HapMap consortium has successfully determined allele frequencies and association patterns between common SNPs [International HapMap Consortium 2005; International HapMap Consortium 2007]. A key output from the project was the estimated patterns of linkage disequilibrium (LD) in the human genome, that is the non-random association of alleles at two or more sites on the same chromosome. Variation in LD patterns exists because recombination occurs non-randomly along human chromosomes, concentrated in particular hotspots. The estimation of LD patterns, or more specifically the haplotype information across human populations, has made it possible to conduct cost-efficient genome-wide association studies (GWAS) of common human diseases [Manolio et al. 2008].

Apart from its major impact on GWAS, the HapMap project has been a significant contributor towards the validation of a large proportion of human SNPs. A total of 269 individuals from four different populations have been genotyped at approximately 3 million SNPs. The HapMap thus provides a great resource for reliable SNP data. Considering SNPs with significant allele frequencies within the HapMap populations is currently one of the most robust ways for establishing a genome-wide set of polymorphic sites with zero false positive SNPs.

Other resources

The study of human DNA variation is clearly not limited to human SNPs. One key resource lies within the sequence variants that underlie Mendelian disorders, which have been logged over several years and deposited into the Human Gene Mutation Database (HGMD) [Cooper and Ball 1998]. With respect to large-scale, structural variants that have been discovered in healthy individuals, most of these variants have been recorded in the Database of Genomic Variants [Iafrate et al. 2004; Zhang et al. 2006]. Finally yet importantly, one may study DNA variation patterns by means of genomic comparisons with closely related species. Analyses of sequence differences between the human and chimpanzee genome have

provided novel insights into fundamental aspects of human mutation and natural selection [Kehrer-Sawatzki and Cooper 2007].

1.4 SCOPE AND OVERVIEW OF THESIS

The aim of this thesis is to explore the role and impact of different molecular mechanisms in the generation of human DNA variation patterns. Taking into account the breadth of the studies included in the thesis, we consider this scope to encompass: 1) characterization of DNA variation patterns in the human genome, 2) tests of hypotheses for mechanisms that could underlie the distribution of DNA variation, and 3) exploration of variation patterns as an aid to understand genome function and evolution. A cornerstone in each of the undertaken studies is the use of computational and statistical genomics techniques to assess potential sequence biases of DNA variants in the human genome. In fact, all studies are purely computational, utilizing publicly annotated genome sequences and polymorphisms in dbSNP as the primary data material. With respect to DNA variation types, the focus lies on single point mutations and human polymorphisms (SNPs).

The rest of the thesis is organized as follows. In Chapter 2, an overview of known molecular factors that could influence DNA variation patterns is presented. In Chapter 3, we provide a summary of the present investigation, describing the main findings in the four studies that have been undertaken. In Chapter 4, we discuss the findings and methodology in our studies in relation to previous work, along with future prospects.

2 FACTORS AFFECTING DNA VARIATION PATTERNS

This chapter discusses the present knowledge of molecular factors that influence the genomic distribution and context-dependency of DNA variation. A temptation has been to organize the chapter in a manner that follows the main points of control in the generation of mutations and standing DNA variation. The first section is about the mutational input, which encompasses both exogenous and endogenous sources of DNA damage. A special emphasis will be placed on the different mechanisms of endogenous DNA damage and the various models that have been proposed for DNA replication errors. The second section concerns the mutational output, involving the different DNA repair pathways and to what extent repair enzymes exhibit sequence biases with respect to efficient repair of DNA lesions. The third section is about the relationships between genetic variation and physical properties of DNA, including chromatin and nucleosome dynamics. We then discuss the impact of natural selection in the fourth section. Notably, the presentations of DNA damage and repair factors in the first two sections have not been restricted to analyses and findings in germline cells. In fact, a substantial proportion of the current knowledge with respect to DNA adduct formation, replication fidelity, and repair specificity have been established from analyses of somatic cells in lower order organisms, such as *E.coli* and yeast. In order to retain the germline perspective, we will in the fifth section of the chapter provide a brief discussion on distinct properties of germline cells. Finally, we present some non-biological factors that could modulate the observed variation patterns.

It is worth noting that the current understanding of factors that influence variation patterns have been obtained from a diverse set of approaches. Genetic and biochemical approaches have lately been complemented by genomics-based approaches. These approaches are based upon the increased availability of annotated genome sequences and mutation/polymorphism data. Through the use of advanced computational and statistical techniques, high-throughput genomics have successfully discovered many novel relationships between genomic properties and genetic variation.

2.1 MUTATIONAL INPUT

The mutational input or DNA damage is often divided into two major classes; environmental (exogenous) and endogenous. This classification is not always ideal as some types of damage may have both endogenous and exogenous origins (e.g. alkylating agents

and reactive oxygen species [Friedberg et al. 2006]). It is also clear that endogenous cellular processes modulate the impact of many exogenous mutagens. The endogenous class of damage can nevertheless be subclassified into two principal categories: 1) spontaneous chemical modifications of DNA through reactions with oxygen and water, and 2) DNA replication errors, including base misincorporations by the polymerase and copying passed damaged templates. With respect to the exogenous class of DNA damage, a number of different mutagens have been identified [Friedberg et al. 2006]. Some of these external mutagens preferentially form adducts in specific sequence contexts [Gordon and Haseltine 1982; Richardson and Richardson 1990]. The overrepresentation of these contexts among somatic cancer mutations suggests a distinct role for exogenous mutagens in the generation of particular human cancers [Pfeifer et al. 2002; Pfeifer and Besaratinia 2009; Pleasance et al. 2010]. As opposed to the case for somatic mutations, there is limited data suggestive of a significant impact of exogenous damage in the generation of germline mutations. Studies on large numbers of children born to atomic bomb survivors (i.e. persons that have been highly exposed to ionizing radiation) failed to detect significant genetic consequences [Otake et al. 1990; Rudiger 1991]. There are thus obvious incongruencies between the potency of an external agent to cause DNA damage in a cell, and its role in the induction of genetic effects in the offspring. In light of these observations, we have decided to focus on the endogenous mechanisms underlying mutational input. However, it is relevant to first describe some of the most important exogenous sources of DNA damage.

2.1.1 Exogenous sources

Ionizing radiation may lead to the formation of ionized and excited molecules in human cells, and the variety of induced DNA lesions is huge [Ward 1988; Cadet et al. 2004]. Humans are exposed to cosmic radiation and naturally occurring radionuclides, which are the primary two sources of external background exposure [Friedberg et al. 2006]. Internally, exposure arises from decay of naturally deposited radionuclides within tissues (e.g. potassium-40). Ionizing radiation may damage DNA both directly and indirectly. DNA may absorb the radiation energy directly, causing ionization of bases or sugars [Ward 1988]. Indirect effects occur when DNA reacts with species generated by radiation in water or other surrounding molecules. Radiolysis of water generates several reactive species, of which the hydroxyl radical (·OH) is particularly important with respect to DNA damage [Cadet et al. 1999]. The hydroxyl radical typically attacks the C5=C6 double bond in pyrimidines, but nearly 20% of the radicals react with the sugars in the DNA backbone,

potentially causing strand breakage in a sequence-dependent manner [Ward 1985; Breen and Murphy 1995] (see *Oxidative damage* below). A localized attack of each DNA strand by two or more ·OH radicals may induce double-strand breaks, which have lethal effects if unrepaired [Ward 1990]. Although ionizing radiation result in a wide array of lesions, the relative importance of individual lesions with respect to genetic end points such as survival, mutagenesis or chromosome aberrations has been difficult to establish.

UV radiation exposure has been used extensively for investigating the biological consequences of damage in mammalian cells. It is relevant because organisms have been exposed to solar UV light since the beginning of the evolution of life. The UV spectrum is dived into three segments: UV-A (320 to 400 nm), UV-B (295 to 320 nm) and UV-C (100 to 295 nm). While solar radiation consists mainly of UV-A and UV-B, most laboratory studies expose cells to UV-C light. The same lesions are apparently produced at the longest wavelengths, but more efficiently induced from UV-C [Kuluncsics et al. 1999]. A frequent photoproduct produced by UV radiation is a covalent linkage between adjacent pyrimidines, a structure referred to as a cyclobutane pyrimidine dimer (CPD) [Mitchell et al. 1991; Cadet et al. 2005]. The formation of CPDs in DNA induces a bend and local distortion of the helical structure [Park et al. 2002], which could be important with respect to recognition by repair enzymes. Furthermore, steady-state formation of CPD is also influenced by the nature of the flanking nucleotides. Thymine dimers preferentially formed when flanked by adenines [Gordon and Haseltine 1982]. Cytosine methylation has also been shown to increase the frequency of CPD formation by nearly two-fold [Rochette et al. 2009], and C:G→T:A transitions has been suggested to be the most frequent point mutation caused by CPD formation in human cells [Keohavong et al. 1991]. A different type of photoproduct called pyrimidine-pyrimidone (6-4) adduct (PP) links the ring structures of two adjacent pyrimidines in a different manner than CPDs do [Franklin et al. 1985]. PPs do however occur at a frequency that is some three-fold lower than that of CPD. In addition to the formation CPDs and PPs, high doses of UV radiation may cause strand breakage [Rosenstein and Ducore 1983].

A number of chemical agents may introduce lesions in DNA. Research in this area dates back to 1940's, when the chemical substance of mustard gas was found to be highly mutagenic. A range of other agents has later been shown to react with DNA and form alkyl DNA adducts [Singer 1975; Singer and Kusmierek 1982]. Agents that are commonly used to induce alkylation damage in wetlab studies include methylnitrosourea (MNU), *N*-methyl-*N'*-nitro-*N*-nitrosoguanidine (MNNG) and methyl methanesulfonate (MMS). Methyl

chloride (MeCl) represents an example of a naturally occurring alkylation agent in the environment, emitted primarily from biomass burning [Crutzen and Andreae 1990]. Generally, an alkylation agent may react with a number of different sites within all four bases. The ring nitrogens of the bases appear however to be the most nucleophilic sites [Singer and Kusmierek 1982]. As with CPDs, formation of alkylation products is highly nonrandom with respect to the DNA sequence [Richardson and Richardson 1990]. The sequence-specific reaction of alkylating agents with DNA may thus play a role in the generation of mutational hot spots. There are also alkylation sources in relation to normal cellular metabolism, most notably the methyl group donor S-adenosylmethionine (SAM, see more in *Endogenous sources*) [Sakata et al. 1993; Chiang et al. 1996; Lu 2000].

A particular class of alkylating agents can react with two different nucleophilic centers in DNA, causing interstrand DNA cross-links or intrastrand adducts [Singer 1975; Millard et al. 1991; Friedberg et al. 2006]. This type of damage is important in the sense that it may prevent strand separation, and thus potentially block the processes of DNA replication and transcription. It is for this reason that these agents are attractive in cancer chemotherapy. Examples of bifunctional alkylation agents are nitrous acid (also formed intracellularly from nitrites), cisplatin and photoactivated psoralens.

Other exogenous damaging agents that have been relatively well-studied are agents with substantial carcinogenic effects, including aflatoxins (originates from natural metabolism in fungi) [Smela et al. 2001], tobacco-specific nitrosamines [Hecht 1999], and benzo[*a*]pyrene [Cosman et al. 1992]. Interestingly, the latter mutagen can induce either transitions or transversions in DNA, and the nature of the induced mutation is largely determined by the sequence context of the adduct [Kozack et al. 2000].

As mentioned previously, the quantitative impact of exogenous mutagens to damage and mutation in germline DNA is questionable. Environmental mutagens are nevertheless of clear relevance with respect to human mutagenesis, considering the significant associations that have been observed between the sequence specificity of particular mutagens and mutation spectra in somatic cancer genomes [Pfeifer and Besaratinia 2009; Pleasance et al. 2010].

2.1.2 Endogenous sources

In this section, we discuss various endogenous sources that are potentially damaging to germline DNA. Focus will lay on single base damage, including common base modifications (e.g. oxidative or hydrolytic), DNA replication errors, and damage associated with meiotic recombination. Spontaneous loss of bases that occur through enzymatic cleavage of the N-glycosyl bond in DNA (i.e. *depurination* and *depyrimidination*) may represent another important endogenous source of DNA damage in the germline. We have however omitted a treatment of this topic, as we found limited literature on the potential relationship between depurination/depyrimidination events and human mutation spectra. Figure 5 illustrates some common base modifications that are discussed throughout the section, including also some lesions induced by previously discussed exogenous sources.

Deamination

In the cell, temperature and pH-dependent reactions between nucleotides and oxygen and water constitute a substantial part of the endogenous damage to DNA [Lindahl 1993; De Bont and van Larebeke 2004]. A common spontaneous lesion introduced by these reactions is a deaminated base, being a base that has lost its exocyclic amino group. The loss of the amino group alters the chemical properties of the bases, including their pairing properties in the double helix, potentially giving rise to mutations during semiconservative replication of DNA.

One of the most frequent deamination events is the hydrolytic deamination of cytosine to uracil. It has been estimated that between 100 and 500 cytosines/cell/day are deaminated to uracil [Barnes and Lindahl 2004]. Importantly, the rate of cytosine deamination in duplex DNA is less than 1% of that in single-stranded DNA [Lindahl and Nyberg 1974]. This implies that the reaction is more prone to occur at instances in which DNA is transiently denatured, such as replication forks and transcription bubbles. The reaction may be further enhanced by a number of other factors, both steric and chemical. The formation of CPDs or the presence of a mismatched base opposite cytosine can promote deamination, as can reactions with nitrous acid and sodium bisulphite [Hayatsu et al. 1970; Frederico et al. 1993; Lemaire and Ruzsicska 1993; Caulfield et al. 1998]. If a U:G mispair is not repaired before the next round of DNA replication, it may cause C:G→T:A transition mutations, since polymerases efficiently incorporate A opposite U in the template [Krokan et al. 2002]. Subsequent removal of U and insertion of T would fixate the transition. The

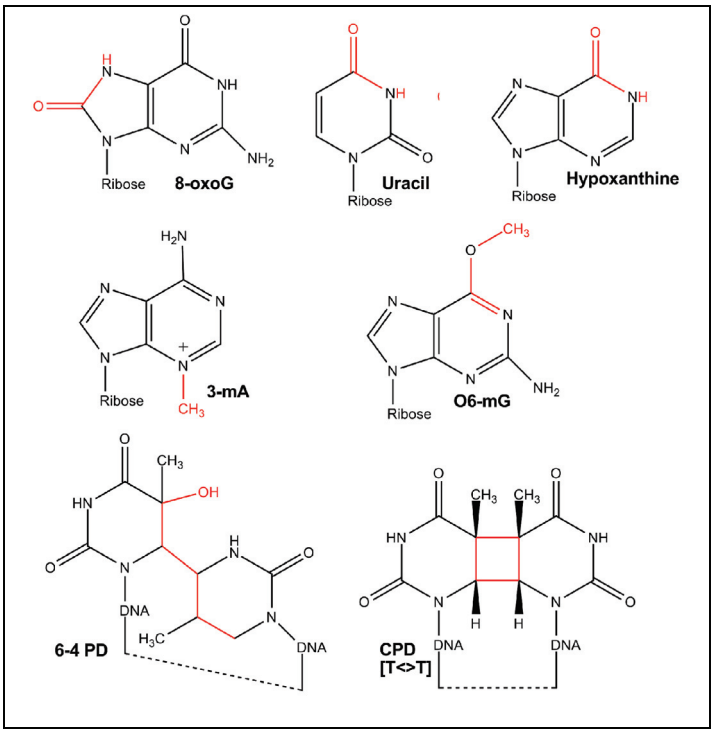


Figure 5: Examples of common base lesions in genomic DNA. 8-oxoG, 7,8-dihydro-8-oxoguanine; 3-mA, 3-methyladenine; O6-mG, O6-methylguanine; 6-4 PD, 6-4 pyrimidine dimer (here, thymine–thymine dimer [T(6-4)T]); CPD, cyclobutane pyrimidine dimer (here, thymine– thymine dimer [T o4 T]). Illustration is adapted from [Dalhus et al. 2009].

biological importance of cytosine deamination has been demonstrated in human cells that are deficient in the removal of uracil from DNA, which experience a significant increase in the rate of nucleotide transitions [Radany et al. 2000].

Under given physiological conditions, adenine and guanine can also deaminate, to hypoxanthine and xanthine, respectively. These events are however much less frequent than deaminations of cytosine (conversion of adenine to hypoxanthine occurs at 2% of the rate of the conversion of cytosine to uracil [Lindahl 1993]). Both xanthine and hypoxanthine are however both potentially mutagenic. Hypoxanthine can base pair with cytosine, and it may thus cause transition mutations following DNA replication. Xanthine is unable to pair with both cytosine and thymine, and can hence stall replication.

5-methylcytosine and mutagenesis at the CpG dinucleotide

The enzymatic conversion of cytosine to 5-methylcytosine (5mC) by DNA methyltransferases is an epigenetic modification that is essential in normal embryonic development of mammals [Li et al. 1992; Okano et al. 1999]. Generally, cytosine methylation functions to maintain a repressed chromatin state and downregulate gene expression [Bird and Wolffe 1999; Suzuki and Bird 2008]. In eukaryotes, cytosine methylation occurs predominantly within the CpG dinucleotide context, yet non-CpG methylation appears to play a significant role in human stem cells [Ramsahoye et al. 2000; Lister et al. 2009b]. Not all CpG sites are methylated, and methylation patterns are tissue specific [Meissner et al. 2008; Rakyan et al. 2008]. Methylation is however also considered to be a frequent source of damage and mutation [Gonzalgo and Jones 1997]. Owing most likely to the relatively high rate of hydrolytic deamination of 5mC to thymine (approximately two-fold higher than deamination of cytosine to uracil [Shen et al. 1994]) and the supposedly low-fidelity repair of T:G mispairs [Brown and Jiricny 1987; Walsh and Xu 2006], C:G→T:A transitions are hotspot mutations in the CpG context [Lutsenko and Bhagwat 1999].

Nearest-neighbour sequence analyses of both human SNPs and germline disease mutations have highlighted the CpG context as a hotspot of human mutation. In their pioneering studies of germline DNA disease mutations, Cooper and colleagues estimated the mutability of CpG (i.e. to TpG or CpA) to be approximately five to seven times the base mutation rate [Cooper and Youssoufian 1988; Cooper and Krawczak 1990; Cooper and Krawczak 1993; Krawczak et al. 1998]. Of approximately 7000 different lesions causing human genetic disease, 23% were found to be transitions at the CpG dinucleotide [Krawczak et al. 1998]. Studies of human SNPs have obtained similar results, showing that CpGs are six- to seven-fold more abundant at polymorphic sites than expected by chance [Zhao and Boerwinkle 2002; Tomso and Bell 2003; Zhao and Zhang 2006]. Recent genomics-based studies have further shown that the rate of C:G→T:A in CpGs is strongly negatively correlated with the regional GC content, although the influence appears very local (i.e. 2kb) [Elango et al. 2008]. This finding is consistent with the idea that the rate of deamination in a given genomic region is related to its propensity of being in a transiently single-stranded form, a property which in turn is dictated by DNA melting and therefore also base composition [Fryxell and Moon 2005; Zhao and Jiang 2007]. Additional support of the methylation-deamination model came through the observation that C/T and A/G SNPs were underrepresented in CpG-rich clusters known as CpG islands [Jiang and Zhao

2006], which are essentially unmethylated domains that coincide with transcribed regions [Bird 1986; Cross and Bird 1995]. Suppression of CpG sequence variation in CpG islands have also been confirmed with respect to sequence variation inferred from duplicated DNA [Nakken et al. 2009a]. In summary, a number of parameters affect DNA variation induced by the methylation-deamination process at CpGs: global methylation patterns, the presence of CpG islands, regional variability in base composition, and potentially a less-than-perfect repair mechanism of T:G mispairs. Recent technological advances in mapping DNA methylomes at base resolution are likely to further increase our understanding of cytosine methylation patterns and their relationship to mutation and DNA variation [Lister and Ecker 2009].

The observation that CpG appears hypermutable irrespective of methylation-mediated deamination has brought forward alternative mechanisms with respect to CpG mutagenesis [Pfeifer 2006]. One pathway may involve methyltransferases, which have been shown to promote deaminations at CpG targets in bacteria [Wyszynski et al. 1994]. A more direct pathway may occur via activation-induced cytidine deaminase (AID), which is required for somatic hypermutation [Muramatsu et al. 2000]. It is also noteworthy that CpG enhances adduct formation by several external mutagens [Denissenko et al. 1997; Tommasi et al. 1997].

Endogenous alkylating agents

In human tissues, DNA alkylation adducts are frequently formed from reactive endogenous agents [Singer 1975]. The most important of these agents is the reactive methyl group donor S-adenosylmethionine (SAM). In addition to its important role in enzymatic methylation (i.e. 5mC), SAM may introduce mutagenic adducts through non-enzymatic methylation of DNA. It has been estimated that an intracellular concentration of 4 x 10^{-5} M SAM produces approximately 4000 7-methylguanine (7mG), 300 3-methyladenine (3mA) and 10-30 O^6-methylguanine residues/cell/day in mammals [Rydberg and Lindahl 1982]. While 7-methylguanine is relatively harmless, 3-methyladenine is a cytotoxic DNA lesion that blocks replication, and O^6-methylguanine can cause both G:C→A:T and T:A→C:G transitions upon DNA replication [De Bont and van Larebeke 2004]. There appears to be a range of different factors that could determine the formation of alkylation adducts: the local sequence context, the chemical nature of the alkylating agent, and chromatin structure [Cloutier et al. 2001]. Similarly, there is data that point to a significant sequence context-

dependency with respect to base excision repair of methylpurines in DNA [Ye et al. 1998] (see more in *Mutational Output – DNA repair*).

Oxidative damage

The majority of DNA damage is most likely caused by reactive oxygen species (ROS) in the intracellular environment. ROS may arise either directly through normal metabolic reactions or indirectly through exposure to external agents such as ionizing radiation or UV light. The main oxidative damaging agents are superoxide radical ($\cdot O_2^-$), hydrogen peroxide (H_2O_2) and the highly reactive hydroxyl radical ($\cdot OH$). The damage produced by ROS includes more than 80 different products, primarily various forms of oxidized bases and single-strand breaks [Cadet et al. 1999; Cadet et al. 2003; De Bont and van Larebeke 2004].

Among the different bases in DNA, guanine is most susceptible to oxidation by ROS. Its simplest oxidized form 7-8-dihydro-8-oxoguanine (8-oxoG) appears to be of most biological importance [Cooke et al. 2003]. The accumulation of 8-oxoG in DNA is a result of the incorporation of 8-oxo-dGTP generated in nucleotide pools as well as of the direct oxidation of guanine in DNA. It has been estimated that there are several thousands 8-oxoG residues per nuclear genome of normal human tissues [Gedik and Collins 2005]. 8-oxoG is known as a potent premutagenic lesion, because it can pair with adenine as well as cytosine during DNA replication, and thus potentially cause a C:G→A:T transversion mutation [Shibutani et al. 1991; Cheng et al. 1992]. A recent study performed fluorescence in situ detection of 8-oxoG in human metaphase chromosomes (prepared from peripheral lymphocytes), and discovered significant variation in 8-oxoG intensity across the genome [Ohno et al. 2006]. This variation was furthermore conserved between four human individuals. Interestingly, this study also found a correlation between 8-oxoG intensity and human SNP density, although this association could be confounded by other factors such as local GC content, CpG density, and recombination rate. Furthermore, no outstanding relationship was discovered between the frequency of transversion SNPs and 8-oxoG intensity [Ohno et al. 2006]. With respect to the impact of local sequence context on the mutagenicity of 8-oxoG, studies in yeast have found that the 5' neighbouring base could have a significant effect on proper lesion bypass by translesion polymerases [Yung et al. 2007; Yung et al. 2008].

The hydroxyl radical ($\cdot OH$) may also cause strand breaks through abstraction of a deoxyribose hydrogen atom. It has been revealed that the accessible surface areas of the hydrogen atoms of the DNA backbone modulate the propensity for these breaks

[Balasubramanian et al. 1998]. Importantly, the cleavage patterns produced by the hydroxyl radical in a DNA sequence will thus reflect the local DNA topography and structure. A recent study used hydroxyl radical footprinting data as a basis to develop a solid sequence-based predictor of this structural information of DNA [Greenbaum et al. 2007]. A subsequent investigation demonstrated that non-CpG mutation spectra in mammals correlated significantly with the predicted patterns of hydroxyl radical cleavage, suggesting that local DNA topography (or merely the intensity of oxidative damage) has played an important part in shaping recent mammalian sequence evolution [Stoltzfus 2008].

DNA polymerases and replication fidelity

Molecular mechanisms involved in DNA synthesis are essential in maintaining the genetic integrity of germline DNA. Highly accurate synthesis is beneficial in the sense that it will avoid mutations that can initiate human disease. On the other hand, it is clear that less accurate DNA synthesis are beneficial for evolution of the species, and more specifically for the development of a normal immune system [Kunkel 2004; Maizels 2005]. Enzymes within the DNA polymerase family largely preserve the balance between low and high synthesis accuracy. This family contains at least 14 different members in the human genome and whose primary function is to catalyze DNA synthesis in the course of replication and in the process of base excision repair [Loeb and Monnat 2008]. The accuracy by which a replicative polymerase synthesizes DNA (i.e. its *fidelity*) appears exceptionally high in eukaryotes, with a base substitution error rate in the range of 10^{-8} to 10^{-9}. High replication fidelity is achieved by means of a multi-step error-correcting process involving primarily *i)* geometric selection of the incoming nucleotide (i.e. the polymerase active site utilizes the "shape fit" more than base-base hydrogen-bonding to discriminate between correct and incorrect pairings), and *ii)* a 3'→5' exonuclease proofreading activity that removes misincorporated nucleotides at the primer terminus (can increase fidelity up to 100-fold) [Friedberg et al. 2006]. If a mismatch evades these two steps, an additional error-correcting mechanism arises in the form of mismatch repair. Recently, it has been discovered that specialized polymerases with lower fidelity also exist in the human genome. These polymerases can insert and extend nucleotides opposite sites of DNA template damage (*translesion synthesis*), and are error-prone when copying undamaged DNA, lacking an inherent proofreading activity [Prakash et al. 2005].

Despite the presence of robust error-correcting machineries among the canonical eukaryotic polymerases, they produce base substitution errors at detectable rates *in vitro*,

demonstrating that mispairs can form with a stability that is sufficient for catalysis [Kunkel 1985b; Kunkel 1985a; Kunkel and Alexander 1986]. Such mispairs may for instance arise through wobble base pairing or via chemically modified bases [Johnson and Beese 2004]. An example of the latter is 8-oxoG, which can form a Hoogsteen pair with adenine [Hsu et al. 2004]. Chemical and structural constraints further yield mispair formations with significantly different frequencies [Kunkel and Alexander 1986]. The rate of a particular mispair also appears to vary between different polymerases, suggesting that differential impact of functionally distinct polymerases may influence mutation frequency.

In bacteria, high frequencies of nucleotide substitutions at repetitive and inverted DNA sequences have caught much attention. It is generally thought that these mutations do not occur via direct miscoding, but rather mediated through a transient misalignment of template and primer strands during replication. The initial model suggested that a physical slippage or misalignment frequently occur at runs of the same nucleotide (homonucleotide tracts), and that these events lead to short base deletions and insertions [Streisinger et al. 1966] (see Figure 6, *left pathway*). An extension of this model, coined *dislocation mutagenesis*, proposed that base substitution errors could occur when a transient misalignment was followed by a correct nucleotide incorporation and realignment of the primer-template [Kunkel and Soni 1988] (Figure 6, *right pathway*). The model was suggested as the underlying mutational mechanism of two hot spots for errors by Pol ß [Kunkel 1985a]. Dislocation mutagenesis is rate-limited in part by the identity of the template bases, and recent investigations have shown that guanine templates are less prone to slippage than pyrimidine templates [Chi and Lam 2007; Chi and Lam 2009]. An additional model involves hotspot mutagenesis mediated by imperfect inverted repeats, so-called *quasipalindromes* [Ripley 1982]. Here, it is thought that a transient misalignment occurs through the formation of a hairpin structure between two imperfect inverted repeats, allowing one repeat to template mutations into the other [Rosche et al. 1997; van Noort et al. 2003; Lovett 2004]. This mechanism converts imperfect inverted repeats into more perfect ones (see Figure 7).

What is known about the relationship between DNA replication mechanisms and the observed DNA variation patterns in humans? With respect to human indel polymorphisms, a significant fraction has been found to be repeat expansions of mono- or dinucleotides that contain adenine or thymine, a finding that points to physical slippage as the underlying mechanism [Mills 2006]. There is also evidence for an overrepresentation of motifs associated with DNA polymerase pausing in the vicinity of disease-causing indels [Ball et

al. 2005]. For single nucleotide variants, it appears as if many of the bacterial substitution hotspots that have been explained by means of replication errors are not as prominent in

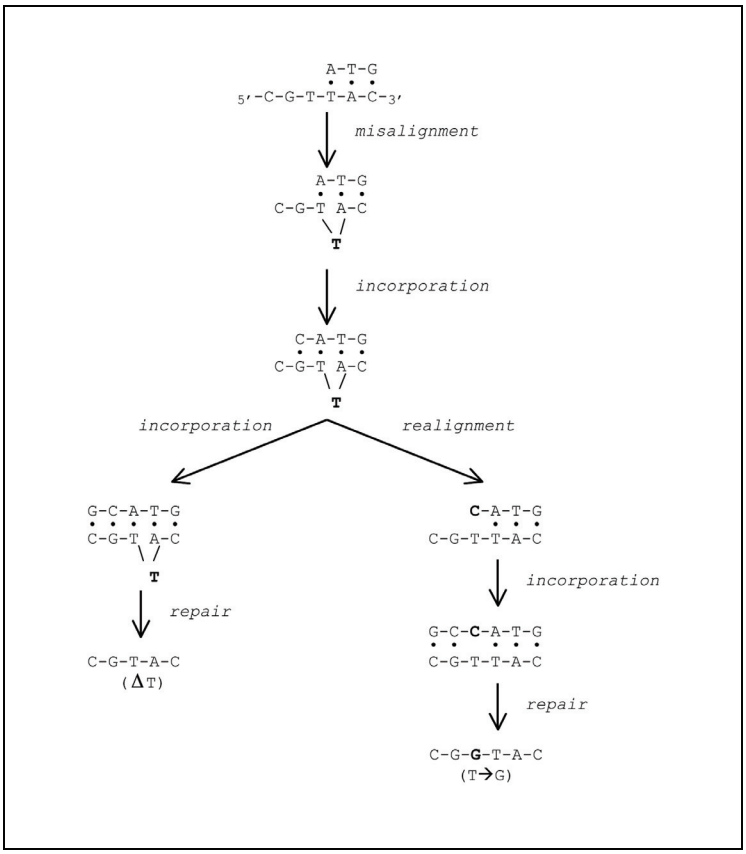


Figure 6: Pathways for base mutations following misalignment between template and primer during DNA replication (adapted from [Cooper and Krawczak 1993])

higher order organisms. The impact of quasipalindrome mutagenesis, estimated as the observed/expected ratio of closely spaced perfect inverted repeats, was found to be negligible in coding exons of human, mouse, and the *Arabidopsis thaliana* plant [Ladoukakis and Eyre-Walker 2007]. However, analyzing perfect repeats only within exonic sequence is not optimal, considering that these DNA sequences encode amino acids that are subject to natural selection. A significant association between replication and mutation has on the other hand been observed for disease mutations. Specifically, Cooper and Krawczak found a clear correlation (Spearman's rho=0.563, p=0.06) between the

relative frequencies of 139 non-synonymous mutations causing genetic disease and the estimated frequencies of *in vitro* misincorporations by vertebrate DNA polymerases [Cooper and Krawczak 1990].

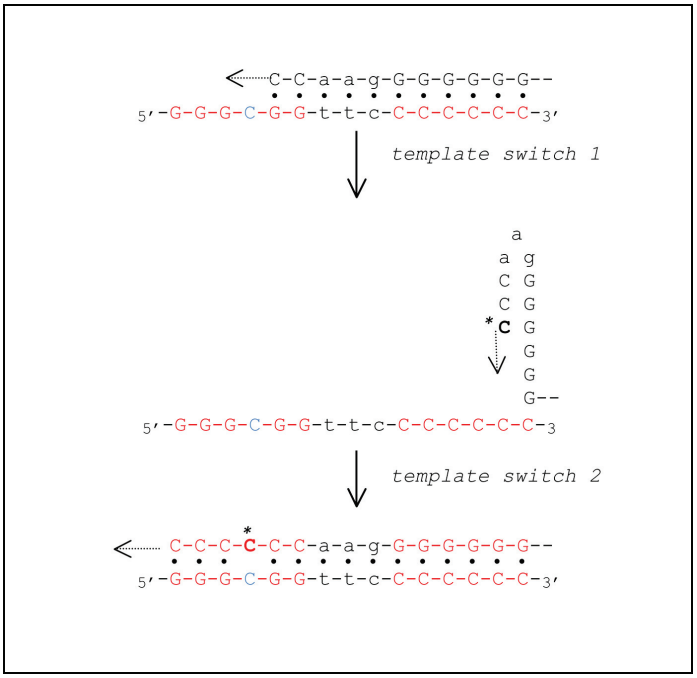


Figure 7: Quasipalindrome-associated mutagenesis. DNA replication through a pair of imperfect inverted repeats (repeats in template strand in red colour, mismatched base in blue colour). Transient formation of an intrastrand hairpin allows copying of one strand of the hairpin (putative mutation in bold, asterisk). Template switching back to the normal mode could lead to a G>C mutation and two perfect inverted repeats (adapted from [Lovett 2004])

When interpreting this association, one must keep in mind that the purified polymerases used to estimate misincorporations *in vitro* lack the 3'→5' exonuclease activity required for proofreading *in vivo*. Although the discovered relationship appears notable, the underlying dataset was very limited at the time that this was reported (139 disease mutations). To our knowledge, such a relationship has not been sufficiently explored with respect to the distribution and frequencies of human SNPs.

A recent study explored the relationship between human replication timing and diversity patterns in the human genome. The temporal order of DNA replication during an S phase displays marked regional variability [Karnani et al. 2007]. A remarkable trend was

found in which later-replicating regions exhibit an increased rate of mutations (both SNPs and fixed lineage substitutions) for all substitution types. The authors suggest that the result is most likely a consequence of an accumulation of damage-susceptible single-stranded DNA in later-replicating regions [Stamatoyannopoulos et al. 2009]. Another group studied the relationship between mammalian substitution rates and replication timing, and they obtained essentially similar results [Chen et al. 2010].

The importance of DNA replication mechanisms in the induction of point mutations (and standing genetic variation) is further highlighted when comparing the maternal and paternal contributions to genetic novelty [Ellegren 2007]. While the autosomes on average evolve with an equal influence of male- and female-originating mutations, the X chromosome experiences the male and female environment to different extents during evolution (one-third and two-thirds, respectively). Using this information in conjunction with estimated divergence rates (i.e. human-chimp sequence differences) on the autosomes and the X chromosome, one can estimate the normalized ratio of X to autosomal mutation rate. The size of this ratio, commonly known as *male mutation bias (α)*, range from 2 to 8 [Makova and Li 2002]. The explanation for male mutation bias has been attributed to an increased impact of mutagenic DNA replication events in males compared to females, essentially owing to the relative excess of cell divisions in the male versus female germ line. Importantly, it has been argued that most base substitutions in male take place in the A-pale class of spermatogonia, which are self-renewing (SrAp cells) and divide continuously throughout a man's life [Qin et al. 2007]. In contrast, oogonia cease replication during fetal life after ~30 cell generations [Drost and Lee 1995]. When comparing the average number of cell generations in sperm of 20-year old men with the cell generation history of a female gamete, this ratio is remarkably consistent with the male mutation bias. Although male mutation bias in general points to the importance of replication-dependent events in the generation of mutations, not all substitution types are affected to the same extent. CpG sites (those outside CpG islands) display a male mutation bias that is significantly reduced ($\alpha=2$) [Taylor et al. 2006]. This can be explained by the fact that transitions induced by the methylation-deamination process at CpG sites are basically replication-independent. Some authors have however postulated that the underlying cause of low male bias at CpG could be more complex, for instance that female and male germ lines may exhibit quantitative differences with respect to CpG methylation, enzymatic deamination pathways, and mismatch repair [Arnheim and Calabrese 2009] (see more below in *Germline perspectives*).

Meiotic recombination

Sexually reproducing organisms make use of meiotic recombination to maintain genetic diversity in its offspring. Crossovers at meiosis lead to haploid gametes that contain new combinations of grandpaternal and grandmaternal genetic material, and the process hence works to increase the haplotype diversity on which selection can act. The consequences of meiosis are however not limited to crossovers and reciprocal exchange of DNA sequences. In particular, recombination events involve highly localized gene conversion events that introduce a small segment of DNA from one background into the other.

Recently, statistical analyses of genome-wide genotype data have facilitated the identification of more than 25,000 recombination hotspots in the human genome, which are narrow (1-2 kb) and GC-rich regions where the majority of meiotic crossovers occur [Myers et al. 2005]. Recombination hotspots are furthermore positively associated with SNP density [Hellmann et al. 2005; Spencer et al. 2006]. One potential mechanism underlying this association comes through the action of natural selection. Both selective sweeps and background selection reduce diversity at linked neutral sites. The strength of such variation-reducing selection will thus depend on how quickly linkage breaks down, which in turn is mediated by the recombination rate. Apart from the selective explanation, there has been much debate with respect to how molecular mechanisms of recombination and sequence properties of recombination hotspots could affect SNP density. Here, we briefly discuss some of the main theories that have been put forward to explain the association between recombination and patterns of genetic diversity.

It is widely accepted that recombination events are initiated via double-strand breaks (DSBs) in one of two nonsister homologous chromatids. The DSB ends are then degraded to generate long 3' single-stranded tails of several hundred base pairs. One single-stranded DNA will further invade the homologous sequence on the other chromosome, establishing a heteroduplex DNA intermediate. The heteroduplex may be repaired by different pathways, and produce either crossovers or noncrossovers [Duret and Galtier 2009]. It has been suggested that this process in itself is mutagenic, representing a neutral explanation for the association between diversity levels and recombination rates [Hellmann et al. 2003]. The mutagenic aspect of recombination amounts both to the fact that stretches of single-stranded DNA tails are more prone to cellular damage than double-stranded DNA, and that DNA synthesis required for repair of DSBs can produce copying errors. Evidence for the latter has been observed experimentally in studies of mitotic double-strand break repair in *Saccharomyces cerevisiae* [Strathern et al. 1995]. Alternatively, the association could be

due to sequence factors that shape both mutation and recombination rates. For instance, recombination hotspots are significantly correlated with the hypermutable CpG dinucleotide [Kong et al. 2002]. More recently it was also shown that methylation-associated SNPs in the HapMap project correlated with broad recombination rates [Sigurdsson et al. 2009]. This study did not highlight the mutagenic aspect of methylation, however, but rather suggested that it could function as an epigenetic marker for recombination.

It is evident that the fate of existing mutations and SNPs could be driven by molecular biases involved in recombination. First, allelic differences with respect to the rate of DSB formation will lead to meiotic drive against the more recombinogenic allele (i.e. the recombination initiation bias model [Jeffreys and Neumann 2002]). Second, it is clear that the heteroduplexes formed during DSB repair may introduce DNA mismatches at bi-allelic loci in heterozygous individuals, and that the repair of these mismatches may be weakly biased in favour of GC alleles. This process is known as *biased gene conversion* (BGC), and provides a mechanism by which alleles at polymorphic loci can become over-transmitted from one generation to the next [Duret and Galtier 2009]. Evidence for BGC comes from the observation that A:T→G:C SNPs segregate at higher allele frequencies than G:C→A:T SNPs at the level of recombination hotspots, an observation one would not expect merely from a mutagenic effect of recombination [Spencer 2006]. It is thought that the GC bias arises from a bias in the activity of DNA glycosylases or mismatch repair enzymes (described in *Mutational Output – DNA Repair*).

Asymmetry in the DNA precursor pool

It is well recognized that a balanced supply of undamaged deoxyribonucleoside triphosphate (dNTP) precursors is critical with respect to both high fidelity DNA replication and successful DNA repair. Mutagenic mechanisms resulting from asymmetry in the precursor pool are for the most part replication-associated. A dNTP excess or deficiency may cause more misinsertions, and a dNTP in excess may also inhibit proofreading by driving DNA chain extension past a mismatch site before an editing nuclease can correct the error (next-nucleotide effect). On the other hand, the dNTP pool is tightly regulated during the cell cycle, and nearly all organisms possess natural pool asymmetries that is presumed to be optimal with respect to a balance between velocity and accuracy of DNA replication [Mathews 2006]. The most striking general feature is the underrepresentation of dGTP, which usually comprises 5-10% of the total dNTP pool [Mathews and Ji 1992]. Although the aspect of dNTP pool imbalances clearly represents a considerable factor in the induction

of replication errors, there is, to our knowledge, limited data regarding the role of DNA precursor metabolism in human germline cells and its implications for mutagenesis.

2.2 MUTATIONAL OUTPUT – DNA REPAIR

It is obvious from the previous discussions that the underlying processes of base damage and DNA synthesis errors will influence human mutation at different levels; the local and global sequence environment, the timing, and sometimes what it mutates to. However, damage is clearly not mutation. Eukaryotic cells have evolved a versatile DNA damage response in order to maintain genetic integrity and stability. With respect to base damage and mispairs occurring in DNA, the repair mechanisms involve four primary pathways in humans: (1) direct reversal, (2) mismatch repair (MMR), (3) nucleotide excision repair (NER), and (4) base excision repair (BER). DNA repair is highly regulated with substantial functional overlap between the different pathways, and recent advances in structural biology have provided detailed insights into the function of important repair enzymes. A detailed outline of how repair enzymes operate in the context of various DNA lesions is considered to go well beyond the scope of this thesis. Rather, in light of the topic that is being addressed, we will draw attention to studies that have analyzed sequence dependencies in the context of repair efficiency. Focus will be on how successful repair may be modulated by the local DNA sequence environment. Such effects could cause sequence biases in relation to the types and relative fractions of unrepaired damage events, which in turn could produce biases in human mutation and genetic variation patterns.

Base excision repair

The base excision repair pathway represents the primary mechanism for repair of chemically modified DNA bases that introduce minor helix distortions. Examples of base modifications or DNA damage in this respect include oxidative lesions (e.g. 8-oxoguanine), hydrolytic deaminations (e.g. deamination of cytosine to uracil), and alkylation adducts (e.g. 3-methyladenine and 7-methylguanine formed endogenously by SAM). The BER pathway is in most instances initiated by specialized enzymes known as DNA glycosylases, which functions to remove the damaged base, leaving an abasic site that subsequently can be filled with the correct base by DNA polymerase ß. DNA glycosylases differ primarily with respect to the kinds of damage that they target.

8-oxoguanine, representing a common oxidative lesion in genomic DNA, is removed and repaired by different mechanisms in human cells (reviewed in [David et al. 2007]). The 8-oxoG in 8-oxoG:C basepairs is removed by 8-oxoG DNA glycosylase (OGG1). To our knowledge, a mechanism in which the local sequence context of 8-oxoG modulates OGG1 specificity has not been reported. MUTYH (bacterial MutY homolog) excises the adenine in the 8-oxoG:A mispairs that may form during replication. As for OGG1, we are not aware of any context effects for the MUTYH enzyme.

N-methylpurine-DNA glycosylase (MPG) is required for removal of methylated purines in DNA, including 3mA and 7mG (O^6-methylguanine is repaired by direct damage reversal by O^6-methylguanine-DNA methyltransferase (MGMT)). These adducts yielded position-dependent repair patterns in normal fibroblast cells, in the sense that 3mA and 7mG persisted in DNA for varying periods time depending on sequence context [Ye et al. 1998]. For 3mA however, it appeared as if the position-dependency was caused by a bias in a post-DNA glycosylase stage of the repair process.

The activity of uracil DNA glycosylase (human UNG protein), which recognizes and removes uracil from DNA, has further been found to exhibit distinct sequence specificity. Uracil removal by UNG may vary as much as 10-fold depending on the identity of the bases that surround the substrate [Eftedal et al. 1993]. The consensus sequences for good and poor repair (AUA/TAT and CUC/GAG, respectively) could indicate that it is the inherent flexibility and bendability of DNA that govern part of the specificity in the UNG-DNA interaction [Seibert et al. 2002]. TDG (thymine DNA glycosylase) is in the same structural superfamily as UNG, and this enzyme initiates the repair of both T:G and U:G mismatches to G:C base pairs in DNA. With respect to T:G mispairs, the amount of incision by TDG was 3-12 fold greater in the CpG context than in the ApG, GpG and TpG contexts [Sibghat-Ullah et al. 1996]. In mammals, there is an additional glycosylase that can remove T or U mispaired with G, with a high activity in methylated CpG sites. Methyl-CpG-binding protein 4 (MBD4), containing a methyl-CpG binding domain, can thus act upon thymine in the frequently occurring G:T mismatches arising from deamination of 5mC at CpG sites [Hendrich et al. 1999]. The methyl-CpG binding domain has however been shown to be dispensable for MBD4 substrate specificity. It appears as if the substrate spectrum of MBD4, in which 5mCpG/TpG and CpG/TpG contexts are preferred, is retained fully by the mere catalytic domain of MBD4 [Petronzelli et al. 2000].

As briefly touched upon in earlier discussions, there appears to be a general supposition that T:G mispairs are repaired less efficiently than U:G mispairs, and that this

weakness could contribute to the high frequency of C:G→T:A transitions generated from deamination of 5mC. Indeed, high fidelity removal of uracil do seem as a necessity in mammalian cells, considering that uracil, unlike thymine, is not a regular DNA base. Some data also support that T:G mispairs are not exclusively repaired in favour of guanine. Brown and Jiricny measured the efficiency by which heterogeneous and homogeneous mispairs were corrected in monkey kidney cells [Brown and Jiricny 1988]. Although the T:G mismatch was the most efficiently repaired heterogeneous mispair, such mispairs were repaired in favour of thymine in 8% of the cases (90% in favour of guanine). Thus, in 8% of the cases, repair allows the C→T transition to be fixed. On the other hand, it is noteworthy that both MBD4 and TDG have a preferred activity at CpG sites. Others have also suggested that factors other than error-prone repair make significant contributions towards the hypermutability of 5mCpG. In particular, some mutagenic adducts preferentially form in 5mCpG and render the context mutable irrespective of the methylation-deamination process [Pfeifer 2006].

Mismatch repair

The DNA mismatch repair system (MMR) represents a critically important system for mutation avoidance in both eukaryotic and prokaryotic cells. The primary role for the highly conserved MMR enzymes is to correct errors that occur during DNA replication. MMR seeks to eliminate two principal types of errors; *single mismatched basepairs* that result from incorporation of an incorrect deoxynucleotide, and small *insertion-deletion loops (IDLs)* that forms from occasional physical slippage between primer and template strands during synthesis [Friedberg et al. 2006]. The activity of MMR is however not limited to DNA replication. MMR targets mismatches and IDLs that occur in heteroduplex DNA intermediates during homologous recombination. It is also evident that some mispairs induced from oxidative damage may provoke correction by the MMR system. Thus, MMR appears critical in ensuring the fidelity of both replication and recombination and is also important in the general DNA damage response.

In humans, the MMR system contains at least seven different genes, the majority of them being homologues of the MutS and MutL family of genes discovered in *E.coli*. The MutSα heterodimer complex (consisting of MutS homolog 2 (MSH2) and MutS homolog 6 (MSH6)) is responsible for locating and recognizing base pair mismatches or single-base IDLs in duplex DNA. MutSβ (MSH2 and MSH3) recognizes IDLs only. The MutL

heterodimers are recruited by the MutS complexes and consist of MLH1 dimerized with either PMS2 (MutLα), PMS1 (MutLβ) or MLH3 (MutLγ) [Shah et al. 2010].

A fundamental problem in MMR is one of specificity, that is, the timely recognition of mismatches and IDLs among the vast excess of correctly paired DNA. A number of structural, thermodynamic and chemical features characterize base mispairs in DNA, and there have been sustained efforts to determine the contribution of these in mismatch recognition [Rajski et al. 2000]. There is a general agreement that ATP-binding and hydrolysis (ATPase) activities in the MutS heterodimers play an essential role in this respect. A range of different DNA lesions activate the MSH ATPase, which in turn induces a conformational change in the MutS-DNA complex that is required for subsequent interaction of MutS with MutL [Marra and Schär 1999]. A recent report demonstrated that the ATPase activation of MutSα exhibits characteristic nearest neighbour sequence effects. Mismatches embedded in contexts containing symmetric 3'-purines enhanced, whereas symmetric 3' pyrimidines reduced, MSH2-MSH6 ATPase activation [Mazurek et al. 2009]. This effect was most prominent for G/T, G/G, and G/A mismatches. Other studies, analyzing both crystallographic and biochemical data, have emphasized the local DNA bendability as a mechanism of MutS specificity [Wang et al. 2003; Tessmer et al. 2008]. MutS-DNA complexes at specific mismatches were found to exhibit conformations in which DNA was both bent and unbent, whereas at nonspecific sites, DNA was found in bent conformations only [Wang et al. 2003]. This observation suggested that the bent conformation represents an intermediate in the formation of an unbent state that is competent for the ATP activation that leads to repair.

MMR proficiency is particularly important for maintaining the stability of microsatellite sequences, which make up an estimated 3% of the human genome [Subramanian et al. 2003]. Microsatellites are short, repetitive elements with one to six bases per repeat unit, and are nonrandomly distributed throughout the genome [Shah et al. 2010]. Microsatellites are mutation-prone due to an increased likelihood of strand slippage events during the replication of repetitive elements. Mechanistically, microsatellite alleles are primarily considered to undergo either expansion or contraction. The mutational specificity may be attributed to both intrinsic DNA features and repair proficiencies of the MMR proteins. Intrinsic DNA features refer to the repeated sequence itself, and include motif size (mononucleotide, dinucleotide, trinucleotide etc.), length (number of units) and sequence composition [Eckert and Hile 2009]. With respect to MMR proficiency, experimental studies in yeast and mice have indicated that defects in specific MMR genes

yield mutational biases [Yao et al. 1999; Otsuka et al. 2003; Hegan et al. 2006]. Mutation of MSH2 is related to the Lynch syndrome in humans, which is characterized by early development of tumours with microsatellite instability [Lynch et al. 2006]. The instability has been found to be dominated by an overrepresentation of contracted mononucleotide microsatellites [Sammalkorpi et al. 2007].

Transcription-coupled repair

Some types of DNA damage lead to bulky distortions in the double helix. Examples include UV-induced cyclopyrimidine dimers and pyrimidine-pyrimidone (6-4) adducts. These lesions are repaired through a distinct repair mechanism, nucleotide excision repair (NER). In eukaryotes, NER involves at least nine different proteins, and deficiencies in most of these proteins lead to severe disease phenotypes [Friedberg et al. 2006]. There is furthermore a significant genomic heterogenity in terms of NER efficiency in transcriptionally active and transcriptionally silent regions. This is largely due to transcription-coupled repair (TCR), a subpathway of NER that repairs the transcribed strand and transcriptionally active regions faster than transcriptionally silent DNA [Fousteri and Mullenders 2008]. There is also data indicating a link between MMR enzymes and TCR activity, although there is conflicting evidence for the existence of such an association [Kobayashi et al. 2005]. Here, we briefly describe the observed patterns of nucleotide substitutions in human transcripts, which largely have been attributed to germline TCR activity.

The mutational signatures related to transcription are most strikingly strand asymmetries with respect to the template/non-template strand. Such strand asymmetries are evident through the different rates at which reverse complementary substitution occur on the nontemplate strand. An analysis of a 1.5Mb region of orthologous sequence between human and chimpanzee found that both purine transitions (A→G and G→A) occur at a higher rate than their complementary pyrimidine transitions on the non-template strand [Green et al. 2003]. Similar biases were found among human SNPs in introns and at fourfold degenerate sites [Qu et al. 2006]. Another comparative genomics approach showed that even transversions could exhibit weak strand asymmetries, and that some of these signatures were context-dependent [Hwang and Green 2004]. The presumably TCR-induced substitutional asymmetries in transcribed regions of the human genome are further thought to underlie an observed compositional strand asymmetry along human transcripts. The

correlation of germ cell gene expression with compositional asymmetry has supported this hypothesis [Majewski 2003].

2.3 ROLE OF DNA PHYSICS AND CHROMATIN

The physical properties of DNA, essentially in terms of local thermodynamics and flexibility, represent important modulatory factors in terms of damage incidence and DNA repair efficiency. As touched upon in an earlier discussion, DNA melting appears to be a key rate-limiting step for 5-methylcytosine deamination [Zhao and Jiang 2007]. Experimental data suggests also that physical replication slippage is less common in thermodynamically stable templates [Chi and Lam 2009]. The flexibility or bendability of different DNA sequence contexts has further been suggested to play a role in uracil removal by UNG and mismatch recognition by the MSH2-MSH6 heterodimer [Seibert et al. 2002; Wang et al. 2003]. DNA thermodynamics is generally considered to be an important factor in proposed models for DNA mismatch repair. One model specifically postulates that the level of kinetic destabilization induced by a DNA mispair in a given sequence context could modulate the efficiency by which the mispair is recognized and repaired [Rajski et al. 2000]. In summary, several lines of evidence suggest that DNA physics bears an overall importance in the generation of human mutations and DNA variation patterns.

The compaction of the genetic material into a nucleoprotein complex constitutes a widely recognized, yet frequently not deeply considered, dimension of all DNA transactions in human cells [Widlak et al. 2006]. Chromatin is constructed from nucleosomes, which contains ~147 nucleotides of DNA wrapped around an octamer of core histones. The degree of compaction is further dependent on the physical separation between adjacent nucleosomes; linker regions are of variable lengths, but they typically range from 10 to 80 basepairs [Luger et al. 1997]. Chains of nucleosomes are further looped and folded into various higher order structures. Generally, it is thought that DNA in high-ordered or compact chromatin is less accessible to damaging agents than decondensed or free chromatin. Late studies on exogenous damaging factors support this idea. Benzopyrene-induced damage was observed significantly reduced in highly condensed mature spermatocyte DNA, and the intensity of UV-induced pyrimidine-pyrimidone photoproducts was found to be six-fold higher in linker DNA compared to nucleosomal DNA [Balhorn et al. 1984; Mitchell-Olds et al. 1995]. Despite a greater lesion rate in linker DNA, sequences in these regions presumably offer greater accessibility for repair proteins [Boulikas 1992]. A

study in yeast demonstrated that cyclo-pyrimidine dimers in linker sequence are more readily repaired by photolyase activity than nucleosomal sites [Suter and Thoma 2002]. Importantly, the highly regulated patterns of gene expression in human cells, as well as epigenetic phenomena (e.g. genomic imprinting), rely on specific chromatin structures. Thus, chromatin organization needs to be sufficiently plastic to allow programmed changes in transcription patterns and cellular development. Specific chemical modifications of histones help partition the genome into distinct domains such as euchromatin ("accessible") and heterochromatin ("inaccessible") [Schones and Zhao 2008].

Recent advances in mapping chromatin accessibility and nucleosome positioning at genome-wide levels have spawned studies on the general relationships between substitution rates and chromatin status. Experiments with human DNA have typically been performed on lymphocytes (e.g. CD4+ T cells), and the observations are then taken to be representative of chromatin status in germline cells. Nucleosome occupancy has primarily been mapped using the micrococcal nuclease (MNase), which preferentially cleaves linker DNA over nucleosomal DNA [Schones et al. 2008]. Digestion patterns by deoxyribonuclease (DNAse I) have been used to track the condensation/decondensation status of large chromatin regions; the hypersensitive DNase I sites are frequently used to identify genetic regulatory elements [Boyle et al. 2008]. With respect to DNA variation patterns, there has been a consistent finding that linker DNA and open regions of the genome display a lower substituion rate than DNA in nucleosomes and closed chromatin. This rate heterogenity has been observed in computational analyses of several species, including yeast, fish and human [Ying et al.; Prendergast et al. 2007; Warnecke et al. 2008; Washietl et al. 2008; Sasaki et al. 2009]. Although most studies have suggested that the difference arises from a higher background mutation rate in closed chromatin (i.e. reduced repair efficiency), a study in yeast argued that selective constraints in linker DNA sequences could be the underlying cause [Warnecke et al. 2008]. To our knowledge, the current understanding of the molecular mechanisms that contribute to chromatin-related differences in human DNA variation patterns is still limited.

2.4 SELECTION ON MUTATIONAL OUTPUT

The previous sections of this chapter have discussed the factors that influence the mutation rate in DNA. In summary, the mutational load is primarily determined by *i)* the relative impact of different sources of DNA damaging agents and replication errors, and *ii)* the

actions and efficiencies of different DNA repair pathways. Moreover, the rate of mutation varies considerably in genome, and this variation occurs at different scales and dimensions (e.g. local context variation owing to CpG, intrachromosomal variation owing to TCR, and interchromosomal variation owing to male mutation bias). However, the patterns of observed genetic variation across the genome do not only reflect the underlying mutation rate, but also evolutionary forces in the form of natural selection and genetic drift [Hurst 2009].

One school of thought suggests that most intra-specific mutations are selectively neutral, i.e. that they do not affect the fitness of the organism. The allele frequencies may change owing to chance alone (genetic drift). A second school of thought believes that most genetic variants will affect the organism's fitness and hence be subject to Darwinian selection [Nielsen 2005]. Importantly, it has been challenging to elucidate the relative contributions of natural selection and genetic drift to the current genetic variation patterns [Biswas and Akey 2006]. Moreover, selection can operate in different modes. *Positive selection* is generally defined as the spread to fixation of an allele that increases the fitness of individuals. Signatures of positive selection include a skew in the allele frequency distribution (for instance an excess of high frequency derived alleles), reduced levels of genetic variation, and elevated levels of linkage disequilibrium [Biswas and Akey 2006]. Genetic hitchhiking occurs when the advantageous allele carries along linked alleles in a selective sweep. The consequence of this phenomenon is reduced nucleotide variation at the linked neutral sites [Hurst 2009]. *Purifying* or *negative selection* eliminates deleterious alleles. An excess of low frequency derived alleles is a common signature of purifying selection. This mode of selection also forms the basis of phylogenetic footprinting, a technique that exploits measures of sequence conservation across species as evidence for purifying selection and a supposed biological function for a given motif. It can be especially challenging to distinguish between purifying selection and low mutation rates, as both processes result in low levels of nucleotide variation. *Balancing selection* is selection that favours diversity.

Dense maps of human SNPs have created new opportunities for large-scale detection of recent selection in the human genome. The development of advanced statistical tests for selection has been critical in this respect. The tests differ primarily with respect to what type of selection footprints they aim to detect (a brief overview of statistical tests for selection and problems associated with them can be found elsewhere [Biswas and Akey 2006; Nielsen et al. 2007]). With respect to specific findings in the human genome, a key

discovery involves the genic variation that causes amino acid change. These non-synonymous variants are undergoing widespread purifying selection, evident through low heterozygosity compared to synonymous variants, and an excess of low frequency derived alleles [Cargill et al. 1999; Hughes et al. 2003; International HapMap Consortium 2007]. A similar excess of low frequency derived alleles have been found for conserved microRNA binding sites and conserved noncoding elements in the human genome, indicative of weak negative selection and a functional importance of these genomic elements [Chen and Rajewsky 2006; Drake et al. 2006]. Other studies have used the human SNP spectrum in discovering selective pressures acting on transcription factor binding sites and exonic splicing enhancers [Fairbrother et al. 2004; Sethupathy et al. 2008]. Furthermore, large-scale approaches have relied on the use of LD patterns to identify haplotypes that segregate at moderate or high frequencies (i.e. incomplete selective sweeps) [Voight et al. 2006; Sabeti et al. 2007]. Such signatures point to genes or other functional elements in the genome that have been subject to recent positive selection. A famous example in this respect is the lactase locus (LCT) in the European population, which involves a high-frequency haplotype that contains an allele associated with lactase persistence [Bersaglieri et al. 2004]. Other genes that have shown evidence of selective sweeps include those related to the nervous system, pigmentation, reproduction, immunity and olfactory receptors [Nielsen et al. 2007].

In addition to the selection that can act on the organism's phenotype as a whole (i.e. its reproductive fitness in a given environment), there is mutation data that could be explained by a selective advantage at the level of premeiotic testis cells [Arnheim and Calabrese 2009]. Direct sequencing of the fibroblast growth factor receptor 2 gene (FGFR2) in human sperm DNA has revealed a highly elevated frequency of a disease-causing (Apert syndrome) C→G transversion, which proposedly could be caused by a dominant gain-of-function effect in the encoded protein [Yu et al. 2000; Goriely et al. 2003]. This finding thus illustrates an intriguing phenomenon of male germline mutagenesis; a mutation that is harmful to the organism may be advantageous in the cellular context of the testis.

As highlighted in the discussion above, use of human SNPs for the purpose of studying molecular mechanisms of mutagenesis requires careful data filtering. In order to establish a neutral spectrum of nucleotide substitutions (reflecting the mutational processes and minimal impact of natural selection), it is common practice to exclude genomic regions that are likely to be functionally constrained. Regions that ought to be excluded encompass

most importantly protein-coding exons, but also constrained elements such as splice sites, conserved non-coding elements, and gene promoters.

2.5 GERMLINE PERSPECTIVES

The events that ultimately generate human patterns of DNA variation can potentially occur at any stage in the development of germ cells in both sexes. Biological characteristics of the germ cell cycles, including the fundamental gender differences, may thus both *i)* contribute to unique patterns of germline mutations that contrasts to those found in somatic tissues, and *ii)* lead to differences in the relative contributions by paternal and maternal germlines in the induction of damage and mutation in the offspring.

As discussed above, DNA methylation is associated with hotspots of C→T transitions due to the rapid deamination of 5-methylcytosine to thymine. Recent studies have demonstrated that genome-wide methylation patterns in the germline (i.e. human sperm and testicular cells in mice) differ markedly from those observed in somatic cells [Eckhardt et al. 2006; Oakes et al. 2007]. Furthermore, methylation patterns are erased and reacquired differentially in developing male and female gametes [Schaefer et al. 2007; Trasler 2009]. In males, *de novo* methylation occurs prenatally in the entire cohort of prospermatogonia, and spermatogonia must maintain these methylation patterns for the large number of mitotic divisions that precede entry into meiosis [Carrell and Hammoud 2010]. In females, *de novo* methylation occurs shortly before ovulation in growing oocytes, after the crossing-over stage of meiosis I. The relatively greater number of cell divisions in the male germline thus increases the burden with respect to maintenance of methylation patterns, and may potentially increase the load of mutations. The origin of methylation-associated mutations may further be influenced by the profound sexual dimorphism in the nature of sequences that undergo de novo methylation in germ cells and the mechanisms by which methylation is regulated [Schaefer et al. 2007].

In addition to DNA methylation, it is evident that male germ cells undergo unique and extensive chromatin remodelling after specification. Chemical modifications to nucleosomal histones, which act alone or in concert to influence gene repression or activation, ensure tight control of gene expression through spermatogenesis [Carrell and Hammoud 2010]. Another characteristic feature of spermiogenesis is the widespread changes in chromatin structure that enables high-level compaction of the genome in the mature sperm head. This involves the exchange of most canonical histones with protamines,

basic proteins with high DNA-binding affinity [Kimmins and Sassone-Corsi 2005]. The extent to which unique chromatin remodelling processes in sperm could impact the paternal DNA integrity is to our knowledge poorly understood.

For practical reasons, the mature female germ cell has been subject to less experimental research than the male. Some mutagens are however specific to the female germline, such as hycanthane and bleomycin [Lewis 1999]. Oxidative stress, which induce oxidized base adducts such as 8-oxoG, is thought to underlie much of the DNA damage in human spermatozoa [Aitken and De Iuliis 2010]. With regards to DNA repair, it has been shown that a large number of genes in common pathways display enhanced or specialized expression during gametogenesis [Jaroudi and SenGupta 2007]. Generally, nuclear extracts from male germ cells in humans and mice display high BER activity [Intano et al. 2001; Olsen et al. 2001]. Exceptions do nevertheless exist. Importantly, the excision of 8-oxoG by OGG1 was significantly limited in human testicular cells compared to rat male germ cells [Olsen et al. 2003]. Nucleotide excision repair, which targets bulky DNA adducts, exhibited varying levels of activity across different spermatogonic cell types in rat. The level of NER was however consistently lower compared to somatic cells [Xu et al. 2005].

2.6 NON-BIOLOGICAL FACTORS

DNA sequence and polymorphism data deposited into public databases are unfortunately not error-free. In addition to the general uncertainty of computationally inferred sequence variants (as discussed in the introductory chapter), variant detection by means of DNA sequencing could also be erroneous. Thus, while the variation patterns observed in human DNA sequences for the most part reflect underlying molecular mechanisms, data noise introduced by non-biological factors will reduce reliable inference.

DNA sequencing errors may originate from multiple sources. Amplification of a target DNA sequence with the polymerase chain reaction (PCR) process may cascade artificial sequence changes in the PCR products. PCR most often involves the thermostable *Taq* polymerase, which lacks an exonuclease activity and occasionally misincorporates deoxyribonucleotides into the growing DNA chain [Chen et al. 1991]. Sequencing errors could further be introduced by sequence-dependent polymerase errors during dye-terminator sequencing. Automated software translation of fluorescence-based sequencing traces into DNA sequence (and error probabilities) adds the final layer to the traditional sequencing process [Ewing and Green 1998; Ewing et al. 1998]. The sequence assembly process, that is

the complex task of arranging individual sequenced fragments in its correct order to produce a larger and complete sequence, can furthermore be erroneous due to the presence of sequence repeats [Tammi et al. 2003].

A study that investigated genetic variation at the conserved intronic splice sites in human genes noted particular biases in terms of reported polymorphisms in dbSNP [Platzer et al. 2006]. By manually examining sequence traces for the polymorphisms, they identified sequencing errors through low signal-to-noise ratios (i.e. wrong base calling). An astonishing 93% error rate was found among 181 SNPs at the 3'-acceptor guanine, and the authors concluded that this was exclusively due to a known suppression of "G after A" incorporation in the genetically engineered DNA polymerases in dye-terminator sequencing reactions. They also found other SNP contexts with false-positive rates > 10%, specifically A(G/H)N, C(A/Y)C, and G(A/C)C (H = A, C or T, Y = C or T, N stands for any base). As demonstrated by the authors, however, most of the false-positive SNPs were filtered away when a stricter set of criteria for inclusion of SNPs from dbSNP was used. Currently, the second generation, high-throughput sequencing technologies are starting to populate databases with novel sequence variants [Shendure and Ji 2008]. As far as we know, the error distributions produced by the new platforms has not yet been characterized, calling for further research into this matter.

Another non-biological bias arises from the SNP discovery process, which purpose is to make an initial assessment of DNA sites that harbour variation among individuals [Clark et al. 2005b]. Since SNP discovery samples are known to be variable both in size and in composition, this will influence SNP properties in subsequent genotyping. The properties that are affected include the levels of variability, the distribution of allele frequencies, and levels of linkage disequilibrium [Nielsen et al. 2007]. It is thought that an ascertainment bias in the SNP discovery process could erode power of tests between SNPs and complex disorders.

3 PRESENT INVESTIGATION

3.1 AIMS OF THE STUDY

There were two primary aims of the studies undertaken in this thesis. First, we wanted to quantify the genomic distribution and context-dependency of DNA variants. Here, variants are not confined to polymorphic sites *per se* (i.e. SNPs). Variants also encompass those between duplicated areas of the human genome (i.e. **paper I**), and also sequence elements that display regional variation, such as mononucleotide repeats (i.e. **paper IV***).* Secondly, we aimed to explore in which manners and to what extent different molecular mechanisms of mutagenesis could underlie the observed sequence patterns of DNA variation. Listed below are the specific aims related to each of the different papers that are part of the thesis:

- Characterize the mutational spectra in duplicated DNA and compare its distribution and context patterns to SNPs that map to unique genomic locations

- Characterize the spectrum of human SNPs within DNA sequence motifs that can form four-stranded G-quadruplex structures; use SNP data to explore the potential for G4 selection

- Explore the quantitative relationship between physical aspects of DNA (i.e. factor such as bendability and thermodynamic stability) and the local sequence patterns at SNP sites; judge the impact of local DNA physics on mutation bias in the human genome

- Quantify the distribution of mononucleotide repeats human genome. Fit an evolutionary model to this distribution that may explain a particular repeat bias in sequences of human DNA repair genes.

3.2 SUMMARY OF PAPERS

Recent segmental duplications are large (>1kb) genomic sequences of high sequence identity (>=90%), making up approximately 5% of the human genome sequence [Bailey et al. 2002]. In **paper I**, we examined the mutational spectrum in these sequence segments of the human genome [Nakken et al. 2009a]. Our idea was to utilize the fact that duplications were once identical, and that mutation inference in sequence alignments of the present-day duplications could establish an independent spectrum of human point mutations (coined duplication-inferred mutations (DIMs)). Further, our goal was to perform a quantitative comparison with ordinary human SNPs that could shed light on specific mutational mechanisms at work in segmental duplications. Overall, we made three principal findings. First, we examined regions of segmental duplications with weak selectional pressure, and discovered that the distribution of point mutation types differed significantly from a presumably comparable set of validated SNPs. We found that the ratio of transitions to transversions was relatively lower in the spectrum drawn from segmental duplications, and that the G/C transversions had the highest relative increase among DIMs. The underlying reasons for these observations are not obvious, although both biological and methodological factors could be involved; 1) the relatively high GC content in duplications, which are rate-limiting for spontaneous transitions at methylated CpG dinucleotides, 2) the observation that the full mutation spectrum within duplications have originated within a longer evolutionary time-window than SNPs (and that this information was not incorporated in the estimate of transition bias), and 3) the knowledge that duplications frequently undergo homology-driven evolution, biased gene conversion in particular. Our second main result concerned the local sequence signature at DIMs and SNPs, i.e. how similar the two collections of mutations were with respect to their context-dependence. Except for minor deviances at the CpG dinucleotide, we found that DNA five-mer contexts at SNPs and DIMs were represented equally among SNPs and DIMs, implying a potentially similar impact of context-dependent modes of mutation in duplicated DNA and ordinary, unique DNA. Finally, we found that for approximately 80,000 SNPs submitted to dbSNP, both genomic positions and the recorded alleles overlapped perfectly with point mutations inferred from duplications. We suggest that these overlaps mirror false positives in dbSNP, and that they are more likely paralogous sequence variants (PSVs) or perhaps multisite variants (MSVs) that have mistakenly been taken as true allelic variants and submitted to dbSNP.

In **paper II**, we analyzed the spectrum of SNPs in a specific set of sequence motifs in the human genome. Guanine-rich sequence motifs in the human genome have considerable potential for adopting higher-order structures known as G-quadruplexes or G4 DNA [Simonsson 2001] (see Figure 8). G-quadruplex formation has been shown to occur both in vitro and in vivo [Sundquist and Klug 1989; Duquette et al. 2004; Paeschke et al. 2005]. Although several computational studies have reported on the huge abundance of G4 motifs in mammalian genomes and their enrichment in gene regulatory regions, their exact biological role in the context of genome regulation and stability are not well understood [Maizels 2006]. To shed additional light on G4 enrichment in the human genome, we performed a mapping of all human polymorphisms within G4 and a subsequent analysis of their distribution within the motifs [Nakken et al. 2009b]. Our main result was related to a specific bias of polymorphisms within G4 motifs; the sites that disrupt the potential for G4 formation were found to be significantly less polymorphic than the neutral complement of G4 sites. We controlled that the reason underlying this difference was not merely a difference in polymorphisms at CpG sites, in that sense suggesting different rate of polymorphisms also for non-CpG sites. We validated the result using a comparative genomics approach, confirming that disruptive sites were significantly more conserved than neutral sites. The specific pattern of polymorphisms appeared not only in the enriched G4 motifs in gene regulatory regions, but generally within all genomic G4 motifs. Furthermore, we demonstrated that the core feature of G4 motifs, i.e. 5'-GGG-3', represent the most underrepresented sequence context for single nucleotide polymorphisms in gene regulatory regions. Overall, findings obtained in this study suggests that guanine-rich sequences may impose specific constraints on point mutagenesis, and that this could have played a significant role with respect to the enrichment end evolution of G4 in the human genome.

In **paper III**, we aimed to add more information to the underlying causes of human mutation bias, that is, the non-random incidence of SNPs in human DNA sequences. We focused on the contribution by local physical properties of DNA in the mediation of mutation bias. Although physical properties of DNA such as bendability and thermodynamic stability have been implicated in several aspects of DNA damage and repair, their quantitative impact on the human mutational spectrum has not been properly determined. In order to pursue this issue, we initially established a large collection of SNPs that mapped to regions of the human genome that presumably evolve with no or weak functional constraints. Importantly, we utilized the chimpanzee and orangutan genome to reliably infer the direction of the nucleotide substitutions that underlie SNPs. Based on this

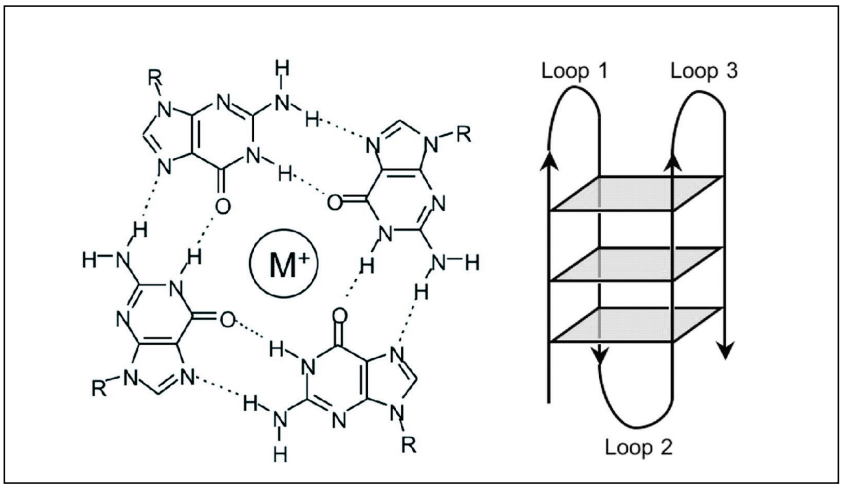


Figure 8: Illustration shows a unimolecular G-quadruplex structure, comprising a stack of G-tetrads interconnected by loop regions (shown right). G-tetrads are planar arrays of four guanines connected by Hoogsteen hydrogen bonds (shown left). A monovalent cation occupies the central position of the tetrad. Adapted from [Huppert and Balasubramanian 2005]

SNP-inferred mutation spectrum, we demonstrated that the context-dependency (i.e. at the level of DNA trimers and DNA pentamers) of nucleotide substitutions displayed some notable associations with DNA physical properties. The data thus supported the idea of a significant modulatory effect by physical DNA properties with respect to the local incidence of human mutations. Most prominently, we found that the sequence-dependent level of thermodynamic destabilization imposed by a DNA mispair displayed significant negative correlations to the incidence of nucleotide transitions. We also found considerable positive associations between the sequence-specific intensity of attack by the hydroxyl radical and the local incidence of mutation. The quantitative relationships between local DNA physical properties and the incidence of mutation displayed great variation among the different substitution types, suggesting that the physical DNA environment affects substitution-specific mechanisms of damage and repair differently.

In **paper IV**, we investigated the distribution of mononucleotide repeats in the human genome [Falster et al. 2010]. Mononucleotide repeats represent a special instance of microsatellites that are associated with genomic instability and tumour development. The stability of mononucleotide repeats is generally maintained by the mismatch repair system (MMR). Paradoxically, the DNA sequences of genes that maintain genomic stability (i.e. seven MMR genes) are themselves unstable, harbouring an increased density of

microsatellites in coding regions. Is there a mechanism that could explain this observation? In this paper, we put forward an evolutionary mechanism that seeks to explain the overrepresentation of mononucleotide repeats in MMR genes. On an evolutionary timescale, the balance between contraction and expansion of genomic mononucleotide repeats will be influenced by the different MMR phenotypes in the population. Experimental evidence suggests that the mutant MMR will be biased towards contraction, whereas the wild-type will maintain the lengths of repeats and shifting the equilibrium towards expansion. Combining this information on MMR mutation bias with the dynamics of meiotic recombination, we suggest that an allele will be more exposed to its own phenotype than to the phenotype of its alternative (i.e. mutant) alleles (see *Figure 1*, **paper IV**). Accordingly, an allele whose phenotype promotes a particular composition of nucleotides (e.g. wild-type MMR genes that promote length stability) should in general contain more of such sequence elements than other sequences in the genome. To test this hypothesis, we estimated mononucleotide repeat density within the average haplotype block (250 kb) around all human genes, as defined by the RefSeq database. We observed a 31% excess of repeats in the MMR block regions compared to all other genes. We checked that this difference between MMR and other genes could not be attributed to confounding factors such as GC content, codon bias and gene expression levels.

4 DISCUSSION

4.1 DATA QUALITY AND METHODOLOGICAL ISSUES

The establishment of reliable genetic variation or mutation data in the human genome has been an important element in each of the undertaken studies. We have further relied on available genome sequence data and annotation tracks, whose qualities are known to be variable. Additionally, we have used and developed different approaches for assessing the sequence context-dependency of point mutations. Here, we intend to discuss limitations associated with data quality and computational methods in the different papers, knowing that these factors will impact reliable inference of underlying molecular mechanisms.

Mutation sources in the recent human genome

Single nucleotide polymorphisms in the human genome, that is the collection of interindividual nucleotide variants in human populations, have been the main data source in our studies. SNPs represent the results of germline DNA mutation events in recent evolutionary time, considering that the majority most likely have occurred in the human lineage following divergence from the chimpanzees (some could be ancestral polymorphisms shared with chimpanzee and are therefore likely older in nature [Hodgkinson et al. 2009]). True SNPs are generally required to display a minor allele frequency (MAF) of at least 1% in a given population; the variants that are more rare are usually classified as rare mutations. However, considering that many reported variants in dbSNP are not associated with estimated population allele frequencies, there is thus uncertainty whether these variants are true SNPs. More extensive human genotyping studies, along with ongoing large-scale resequencing projects such as the 1000Genomes project, will hopefully contribute in further estimation of SNP allele frequencies.

In contrast to most rare mutations underlying monogenic genetic disorders, most human SNPs lie outside protein-coding regions and are presumably neutral in terms of phenotypic effects. The considerable level of DNA variation (i.e. in the form of SNPs) that is observed in apparently healthy individuals supports this notion [Levy et al. 2007]. An increasing number of SNPs have nevertheless been linked with complex diseases, although these variants rarely have shown any direct causative effects. It is generally assumed that such variants have more subtle regulatory effects than monogenic disease mutations [Cirulli and Goldstein 2010]. It is also believed that epistatic interactions between SNPs, and

interactions between SNPs and the environment represent important modulators of human health and complex disease risk [Chakravarti and Little 2003; Manolio et al. 2008; Shao et al. 2008; Frazer et al. 2009]

We have used the largest and main repository of human biallelic SNPs, dbSNP, as a starting point in **paper I, II** and **III**. As mentioned earlier, the quality of the data content in the dbSNP database is known to be variable. Due to this matter, we used careful filtering procedures to reduce the level of false positive SNPs in our data. These procedures are for the most part in accordance with several other large-scale studies of human SNPs [Fairbrother et al. 2004; Chen and Rajewsky 2006; Madsen et al. 2007; Zhao and Jiang 2007]. In particular, we trusted only the set of SNPs that were validated according to a set of criteria established by dbSNP (see Introduction). This step essentially excludes SNPs that lack multiple, independent confirmative submissions, and those that have been derived solely by computational means. Although this filtering step may exclude up to 40% of all reported variants (estimate from dbSNP build131 (May 2010)), it has been highly recommended in early studies on the subject. These studies have demonstrated that non-dbSNP-validated entries are associated with significantly more false positives than the dbSNP-validated ones [Nelson et al. 2004; Edvardsen et al. 2006]. One should however keep in mind that false-positive estimation among putative SNPs (i.e. by means of resequencing) is not straightforward. Such estimates could for instance be biased by the number of samples and the populations that are represented among the samples [International HapMap Consortium 2005]. Noting that dbSNP content has increased substantially since the studies on variant validity were undertaken, and also that many of the new entries have been discovered by means of new sequencing technologies, we find it worthwhile to re-examine the issue of false positives in the dbSNP database.

Although an exclusive analysis of validated SNPs is the most common approach for large-scale studies of SNP patterns, there are possibilities for further minimization of potential false positives. In **paper I** and **II**, we conducted SNP analyses in which available allele frequencies from the HapMap project were used to track and exclude monomorphic SNPs in our dataset. In other words, we only considered dbSNP variants that had been genotyped in the HapMap project, and that showed significant allele frequencies (MAF > 1%). The intention of using HapMap-verified SNPs was thus to ensure that all variants were true SNPs (and true underlying mutational events), with zero false positives. However, collecting true variants by means of HapMap verification data may have introduced other, unknown biases among the resulting variants. SNPs that were selected for genotyping in the

HapMap project (especially in phase I) had to meet several criteria. In phase I, selection was primarily based on the goal of genotyping at least one SNP every 5kb across the genome, and also a prioritization of non-synonymous SNPs [International HapMap Consortium 2005]. The spacing between genotyped SNPs increased however drastically with the completion of phase II, although unsuccessful design of genotyping assays also excluded many putative polymorphic sites [International HapMap Consortium 2007]. General ascertainment bias among the genotyped variants is also known to influence the SNP properties that are generated by HapMap [Clark et al. 2005a]. Notwithstanding these potential biases associated with HapMap genotyped SNPs, we are not aware to what extent these biases could affect the distribution of SNP substitution types and surrounding sequence contexts. It is the latter properties that have been subject to large-scale analyses in **paper I** and **II**. We suggest that a region-specific comparison between a set of dbSNP-validated SNPs and HapMap-verified SNPs with respect to these properties could resolve some of this uncertainty.

The relationship between an observed biallelic SNP and the mutation event that caused it has been of special interest in our studies. Considering only the two alleles associated with a SNP, it is clear that these signal limited information with respect to the mutation event. Two SNP alleles indicate whether the underlying nucleotide substitution was a transition or a transversion. However, an A/G polymorphism could be have been generated through either an A→G transition or a G→A transition. Also, we do not know on which strand the original mutation took place. Thus, since an A/G polymorphism is equivalent to a C/T polymorphism on the opposite strand, there are essentially four possible events that may underlie such a biallelic locus: A→G (or T→C), or G→A (or C→T). Determining the strand orientation of the original mutation is not feasible, so reverse complementary nucleotide substitutions (or polymorphisms) are generally treated together. The directionality of the nucleotide substitution that underlies a SNP can on the other hand be inferred by means of comparative genomics and the assumption of parsimony (discussed briefly in the Introduction). Mapping SNP loci to orthologous sites in chimpanzee for inference of ancestral and derived alleles has recently become a powerful approach in terms of large-scale SNP analysis. We utilized this approach in **paper III**, where we also used the chimpanzee and macaque genomes to infer the directionality of fixed substitutions in the human lineage. Although ancestral misidentification of SNP alleles could occur when using the parsimony approach, use of two outgroup species and the exclusion of CpG sites, as was

done in **paper III**, most likely minimized the impact of this phenomenon in our data [Hernandez et al. 2007].

In **paper I**, we investigated the mutational spectrum in recent segmental duplications. By means of a comparative analysis with SNPs in unique genomic regions, we aimed to shed light on specific molecular mechanisms of mutagenesis in these particular areas of the human genome. Segmental duplications are relatively large genomic regions (>1kb, defined as >5kb in HGSDB) of high DNA sequence identity (>90%). Our idea was to device a pipeline that traced mutational events through the single base differences in alignments of currently annotated duplications, assuming that the collection of duplication copies of a given DNA sequence all originated from one ancestral sequence. In other words, we inferred the mutational spectrum in duplications (duplication-inferred mutations (DIMs)) by recording sequence differences in sequence alignments of duplication copies. In contrast to ordinary human SNPs, which are primarily generated by mutation events in the human lineage, the complete collection of inferred mutation events in recent segmental duplications span a greater evolutionary time history. If minimal gene conversion and a neutral molecular clock for evolution are assumed, duplications with 90% identity correspond to duplication events that occurred approximately 35–40 million years ago, roughly correlating with the divergence of the New and Old World monkeys [Bailey and Eichler 2006]. Although the majority of detected duplications in the human genome appear to be relatively recent in nature (e.g. 77% of alignments have sequence identity > 95% [She et al. 2006]), our crude estimates of substitution types among DIMs and SNPs did not control for differences in the timing of DIM and SNP events. One could thus argue that the transition-transversion ratio in DIMs and SNPs are not directly comparable, since the probability of observing a transition or a transversion at a site in a DNA sequence exhibits a time-dependency [Kimura 1980]. An alternative approach would be to compare SNPs in duplications with SNPs outside duplications. However, the observation that many SNPs in duplicated areas are likely to be false positives (coinciding with DIMs and representing PSVs or MSVs rather than SNPs [Nakken et al. 2009a]) makes it challenging to establish a reliable set of true SNPs in duplications. Another idea would be to compare duplications between the chimpanzee and human genomes, infer the fixed substitution spectrum in the human lineage, and compare this spectrum to a comparable set outside duplications. Regardless of approach, however, it is clear that an analysis of nucleotide substitutions in duplications could be confounded by sequencing errors and an impact of gene conversion. Our small analysis of three-mer sequence contexts that are prone to DNA sequencing errors

indicated that these were slightly more frequent in DIMs than in HapMap-verified SNPs [Nakken et al. 2009a].

Genome annotation issues

Large-scale genomic analyses rely on genome annotation tracks, which denote the genomic locations and functions of specific DNA sequences. The most central tracks utilized in our studies include segmental duplications, G-quadruplex motifs, human protein-coding genes, and high-copy repeats.

The set of segmental duplications subject to analysis in **paper I** was taken from a database that employed a computational technique to detect duplications. This technique used a standard DNA sequence search algorithm (BLAST) to detect duplicated sequences within the human reference genome [Altschul et al. 1990; Cheung et al. 2003]. Such a procedure will thus depend very much on the quality of the reference sequence. The duplicated and repetitive regions of the genome are particularly challenging to assemble, and large duplications were in fact the most dominant feature in the remaining gaps of the initial human genome release [International Human Genome Sequencing Consortium 2004]. It is also evident that segmental duplications are responsible for much copy number variation. Henceforth, some of the duplications detected in the early reference genome could have been copy number variants rather than fixed sequence. Following the next-generation sequencing technologies, and more sequenced human genomes, it could be possible to device more reliable algorithms to locate the set of fixed segmental duplications in the human population.

In **paper II**, we mapped the locations in the human genome that matched the G-quadruplex (G4) motif. We used the *quadparser* algorithm, which searches the genome for a consensus sequence that supposedly satisfy the physical requirements for G-quadruplex formation [Huppert and Balasubramanian 2005]. The low specificity in the quadparser consensus sequence results in a large number of G-quadruplex hits in the human genome. Only a limited portion of all mapped G4 motifs are thought to actually form structures *in vivo* [Maizels 2006]. In order to study the relationship between point mutagenesis and actual physical G-quadruplex structures there is therefore a need to develop more sensitive G4 predictors. One could imagine that more specific algorithms for instance would incorporate a probabilistic profile of loop sequence preferences. It would also be of interest to identify other genomic features (i.e. other than the G4 sequence motif) that potentially enhance G-

quadruplex formation. More experimental analyses of G-quadruplex structures would be critical in this respect.

All four papers have used the RefSeq database as the source for human genes. RefSeq represents the most commonly used sequence database for human transcripts and proteins [Pruitt et al. 2005]. In one study, we analyzed particular regions of RefSeq genes (e.g. G-quadruplexes in first introns, 5', and 3' regulatory regions in **paper II**). For the most part, however, functionally constrained sequences within RefSeq genes have been excluded, which enabled analyses of mutation patterns with minimal impact of natural selection. It is also clear that some sequences outside protein-coding genes are functionally constrained, such as highly conserved non-coding elements [Drake et al. 2006]. In retrospect, the selectively neutral regions of segmental duplications could have been more carefully filtered, for instance following the procedure outlined in **paper III**.

Apart from the analysis of mononucleotide repeats in **paper IV**, we have excluded analyses of DNA variation patterns that occur within high-copy repeats. We have considered high-copy repeats to be those that are identified by RepeatMasker and Tandem Repeats Finder (with period of 12 or less), a definition we have adopted from the UCSC Genome Browser resource [Karolchik et al. 2003]. There are several reasons for having excluding these repeats in our analyses. First, repeats are more likely to be associated with assembly errors [Li et al. 2008]. Secondly, algorithms that generate multiple sequence alignments of orthologous species are primarily driven by alignments of non-repetitive sequences, making the interspersed repeat alignments potentially ambiguous and unreliable [Blanchette et al. 2004]. In **paper III**, we noted that the number of aligned bases in MultiZ sequence alignments of G4 was less frequent in high-copy repeat DNA compared to unique DNA. Thirdly, SNPs that map to unique genomic locations are generally more reliable than those that map to repetitive DNA. Finally, repeats may be biased in terms of DNA sequence composition, and therefore needs a separate treatment during analysis of DNA k-mers (see below). Nevertheless, considering that high-copy repeats constitute nearly half of the human genome sequence, we acknowledge that they should not be overlooked and that greater efforts should be put forward to understand their characteristic mutation patterns.

Assessment of mutation context-dependency

The discussions in chapter 2 revealed that numerous factors could influence the local sequence dependency of human DNA variants. The observed distribution of selectively neutral SNPs has been shaped by the relative impact of many DNA damage sources and

repair enzymes in human germline cells. Many of these factors presumably interact in a sequence-specific manner with DNA, whereas others could be relatively random in terms of local DNA interaction. In an attempt to understand how different molecular mechanisms of mutagenesis contribute to the current DNA variation patterns, a necessary starting point in our analyses has been the construction of quantitative estimates for mutation pressures in different sequence contexts. In other words, we wished to estimate the context-dependency in the final output of mammalian germline mutagenesis (e.g. sequence context of SNPs), and use these estimates to infer, either analytically or in a quantitative manner, potentially underlying molecular mechanisms.

Many other studies have looked into local context-dependencies in mammalian mutation spectra. Early studies found significant variation in neighbour-dependent substitution rates when looking at human gene/pseudogene alignments [Blake et al. 1992; Hess et al. 1994]. Others have developed mathematical models of neighbour-dependent nucleotide substitutions, and used these for an understanding of DNA evolution [Arndt et al. 2003; Arndt and Hwa 2005]. A pioneering study further estimated the most important neighbour-dependencies among human mutations that cause genetic disease [Krawczak et al. 1998]. Our methodology aimed to be somewhat intuitive to the biologist. We essentially estimated the frequencies of sequence contexts of length *k* (DNA *k*-mers) at substitutions (i.e. by mere counting), and compared these frequencies with those seen in the genomic background (i.e. non-substitution sites). In this manner, we estimated the distribution of under- and overrepresented *k*-mers at substitution sites, also referred to as *k*-mer abundance. This approach is similar to earlier studies on human SNPs [Zhao and Boerwinkle 2002; Tomso and Bell 2003; Zhao and Zhang 2006]. Some methodological variations in terms of *k*-mer analysis were however used in the papers, and these are discussed briefly in the following.

Neither sets of mutations that were compared in **paper I** (i.e. SNPs and DIMs) carried directionality of the underlying nucleotide substitutions. Reverse complementary contexts were further combined. An abundant three-mer context among mutations, say for instance C[C/A]T, could thus be attributed to the frequency of either C[C→A]T or C[A→C]T mutation events (or the reverse complementary mutation contexts). The details with respect to the exact nucleotide substitution underlying an abundant context are henceforth limited. In **paper I**, we further based our *k*-mer analysis on the estimation of expected values. The expected proportions of different variant contexts (e.g. C[C/G]A) were estimated under the assumption that a variant (i.e. the underlying substitution type) occurred

independent of surrounding sequence context. A drawback of this approach is that biases in particular contexts, which could either be the targeted biological mutation biases or potentially sequencing error biases, will influence the expectations in many other contexts.

In **paper II**, we looked at the over- and underrepresentation of DNA three-mers at polymorphic sites, adopting the approach by Tomso et al [Tomso and Bell 2003]. This approach essentially estimates the relative frequencies of three-mers at polymorphic sites (each site contributes two three-mers, one for each allele: C[C/A]T gives CCT and CAT), and compares these frequencies to the genomic background. This approach suffers from the same limitations as the method used in paper I. Particularly, an underrepresentation of polymorphisms in the GGG (reverse complementary CCC) three-mer context, as was reported in **paper II**, does not necessarily imply that the mutation of guanine (cytosine) in this context is suppressed. The frequency of GGG at polymorphic sites results from the frequencies of G[A/G]G, G[C/G]G, and G[T/G]G variant contexts. Thus, it is clear that the frequencies of T→G, C→G and A→G in this particular context could have an influence on the result. Two possible extensions of the analysis could resolve some of these uncertainties. First, we could consider only polymorphic sites in which guanine/cytosine is the allele present in the reference sequence, assuming that the reference sequence most often represent the ancestral allele. Secondly, we could annotate the SNP data with directionality from orthology mapping and perform a similar analysis as was done in **paper III**.

The goal of a *k*-mer analysis at sites of human mutations is essentially to identify DNA sequences that exert strong effects during either DNA damage or repair. Ideally, the size of *k* in such an analysis should not be limited, enabling the identification of both short and longer DNA motifs that could regulate mutagenesis. In practice however, the statistical confidence of *k*-mer abundance estimates will depend on the number of substitutions that are observed in the different *k*-mers. In other words, the total number of substitutions that are analyzed limits the size of *k*. In our analyses, we have primarily looked at *k*-mers of odd length centred at the substitution site. The reason for this choice is that such k-mers, with an equal number of bases on each side of the substitution site, is easier to analyze when the two strands are analyzed in combination (that is, combining reverse complementary *k*-mers). We have also experienced that various properties of short DNA contexts, such as thermodynamics and bendability (**paper III**), are frequently estimated experimentally in terms of DNA three-mers (*k*=3).

Other issues

In **paper IV**, we put forward a novel mechanism that could explain the high density of mononucleotide repeats in mismatch repair genes (MMR). Our model suggested that alleles (here represented by the wild-type MMR sequences) whose phenotypes promote a particular nucleotide composition (i.e. expansion or maintenance of repeats) should contain more of such elements than other sequences in the genome. The model did carry important limitations that are noteworthy when interpreting the results. The model's main underlying idea concerned how meiotic recombination of wild-type and mutant MMR alleles may generate a shift towards wild-type expansion in the MMR regions themselves. We thus made the assumption that mutant or MMR-deficient alleles exist in the human population at appreciable frequencies, yet these frequencies are generally unknown. Furthermore, although we argued for a general contraction bias among MMR mutants, the availability of quantitative data with respect to characteristic repair biases of different MMR gene mutants (i.e. more elaborate phenotypic characteristics) could have improved the analysis further. Studies in mice have observed significantly different mutational specificities for different MMR gene defects [Eckert and Hile 2009]. It remains to be determined whether these patterns can be extrapolated to humans. Finally, we widely assumed that data from animal studies and cell lines were representative for MMR activities in the human germline, which is not necessarily a valid assumption.

4.2 PATTERNS OF HUMAN DNA VARIATION

An overall aim of the thesis has been to explore context-dependent patterns of human DNA variation in relation to molecular mechanisms of mutagenesis. Here, we will discuss the findings in the different studies, with special attention to observed patterns, potentially underlying mechanisms, and possible extensions of the undertaken analyses.

In **paper I**, we inferred the mutational spectrum in human segmental duplications. A related study previously demonstrated, by means of primate genomic analyses, that such duplicated regions exhibit a higher rate of nucleotide substitutions than neighbouring, unique DNA [She et al. 2006]. Our study focused on the spectrum rather than the rate, however. Through a computational pipeline that tracked base mismatches in sequence alignments of duplication copies, we created a set of duplication-inferred mutations (DIMs). Such a strategy is in some respect similar to the early analyses of gene-pseudogene alignments [Blake et al. 1992; Hess et al. 1994]. DIMs are thought to have occurred in

recent evolution of human segmental duplications. In order to shed light on characteristic mechanisms of mutation in these regions, we compared the mutational spectrum in duplications with a SNP spectrum in non-duplicated areas of the genome. As discussed above, differences in the evolutionary time-windows associated with two spectra could have complicated the comparison. We found a significant difference in terms of the transition/transversion ratio between presumably selectively neutral sets of SNPs and DIMs. It is hard to say, however, whether this difference in distribution of substitution types stems from a differential impact of molecular mechanisms of mutagenesis in duplicated and unique DNA. Two features of duplications could induce characteristic patterns of mutation in these regions; the high GC content (as demonstrated in our work and in other studies), and the ability of duplications to undergo homology-driven evolution or gene conversion. In order to test how the characteristics of DIMs relate to GC content or gene conversion, we need to extend our analysis. One approach could be to analyze duplications (and the non-duplicated control DNA) in sequence windows of specific lengths, and estimate how the fractions of substitutions vary with GC content. Using the inferred spectrum to test for signatures of GC-biased gene conversion is more challenging. If the substitution directions of DIMs were known we could potentially have explored the distribution of "weak-to-strong" AT→GC mutations, as an overrepresentation of these are usually taken to be a signature of gene conversion [Dreszer et al. 2007].

K-mer analysis demonstrated general similarities between DIMs and SNPs, which can be interpreted as supporting evidence for similar impact of context-dependent damage and repair in the two regions. We further found that the incidence of DIM transitions at CpGs was largely suppressed in CpG islands in segmental duplications. Since CpG dinucleotides in such islands are largely unmethylated, this observation can thus be taken as support of the methylation-deamination model as a dominant mode of mutation also in segmental duplications.

Paper I also demonstrated a large overlap with the positions of inferred DIMs and the submitted SNPs in dbSNP. Although there is some evidence for an increased density of true SNPs in duplications [Fredman et al. 2004], it appears likely that many of the SNPs that overlap with DIMs in terms of both position and alleles constitute PSVs or MSVs rather than true biallelic SNPs. The development of some kind of statistical test with respect to the likelihood of SNP and DIM overlap would have strengthened the findings. We nevertheless consider our pipeline for DIM inference as a valuable resource for filtering unreliable SNPs in segmental duplications.

In **paper II** we explored context-dependent modes of SNPs in guanine-rich sequences capable of forming four-stranded DNA G-quadruplex (G4) structures. We discovered a genome-wide trend in which guanine sites that could disrupt the G4 consensus (if mutated) were less polymorphic and more conserved than the complement of sites. These data thus provided a new perspective on the highly abundant G4 motifs in the human genome, suggesting that the damage or repair of guanine (or cytosine) could exhibit context-dependencies that preserve the G4 sequences in DNA.

The role of G-quadruplex formation in terms of genome regulation and stability is not well understood. The G4 motifs are heavily enriched in gene regulatory regions of many mammals, and some experimental evidence exists for G-quadruplex formation in gene promoters [De Armond et al. 2005; Zhao et al. 2007; Du et al. 2008]. The polymorphism data we reported for G4 motifs in gene regulatory regions did however not exhibit significant differences to those in intergenic regions. One potential reason for general similarities in G4 polymorphism is the lack of knowledge with respect to motifs that actually form G-quadruplex structures *in vivo*. The comparison of G4 across genomic regions was likely to encompass guanine-rich sequences in general rather than only the sequences that actually form physical G-quadruplexes. Our preliminary analysis of guanine-rich motifs did nevertheless not support the proposed selection for G4 in gene regulatory regions. If such selection were present, we would most likely expect to see a relatively lower level of disruptive SNPs (compared to neutral) in regulatory regions. This would be expected since the sequence compositions of both neutral and disruptive G4 sites are likely the same across different genomic categories. Although the hypermutable CpG dinucleotide contributed to the global pattern (i.e. difference in SNP density between disruptive and neutral sites), we controlled that the low ratio of disruptive G4 SNPs was also valid when considering non-CpG sites only. Mechanisms other than the methylation-deamination model thus appeared to contribute.

In order to shed light on specific sequence biases associated with guanine polymorphisms, we extended our analysis to guanine-centred three-mers in general. Here, we found that the GGG/CCC triplet was underrepresented at polymorphic sites, most prominently in gene regulatory regions. We do however recognize limitations of the approach used here for quantifying polymorphism abundance at three-mers (discussed above), so caution should be made with respect to potential quantitative relationships to context-dependent molecular mechanisms. To enable proper testing of the molecular mechanisms that may contribute to guanine polymorphism patterns, we would need to

establish quantitative estimates for context-dependencies in both common guanine damage sources (e.g. 8-oxoG and replication errors at guanines) as well as their associated repair enzymes (e.g. OGG1).

In **paper III**, our aim was to add new information to the possible sources underlying local sequence biases of human DNA point mutations. These sequence biases refer to the local context-dependencies observed for human SNPs, in other words the extent to which SNP incidence is dependent upon its immediate neighbouring sequence. Such biases may for instance arise due to sequence-specific DNA mutagens or context-dependent DNA repair enzymes. Previous studies have repeatedly confirmed the well-known sequence bias for C→T (G→A) transitions, which is strongly biased towards CpG due to rapid deamination of 5-methylcytosine to thymine in this context [Tomso and Bell 2003]. Our study investigated mutation biases in non-CpG contexts only, a matter which has received less attention. Although not directly related to the study of non-CpG mutation bias, a very recent study observed that non-CpG mutation rates are contingent on local CpG content, suggesting for instance that DNA methylation may promote chromatin states that are more susceptible of mutation [Walser and Furano 2010].

The physical properties of DNA represent an important dimension of DNA interactions in human cells, and they are central in several aspects of germline mutagenesis. Varying levels of thermal motion within the double helix determine the extent of fluctuational DNA basepair openings, making normally buried groups accessible for interaction with proteins, chemicals and potential mutagens. The impact of replication slippage may further be related to the local DNA stability. The role of helix stability in mediation of mutation bias was originally investigated by Krawczak et al. [Krawczak et al. 1998]. In that study, thermodynamic measures of DNA trimer contexts were correlated with their relative mutability inferred from non-synonymous germline disease mutations. They found a general, albeit weak association in which the stability of the local DNA sequence environment increased the likelihood of mutation. In **paper III**, we observed a significant positive association between local stability and the sequence bias of C>G(G>C) transversions, but not for any of the other types of substitutions. Our genome-wide dataset of SNPs was on the other hand characteristically different in nature to the set of mutations analyzed by Krawczak et al. We did furthermore observe significant associations between the sequence bias of nucleotide transitions and the level of helix instability induced by the underlying DNA mismatches. Importantly, the thermodynamic stability of the different types of DNA mismatches is significantly context-dependent [Allawi and SantaLucia 1997;

Allawi and SantaLucia 1998a; Allawi and SantaLucia 1998b; Allawi and SantaLucia 1998c; Peyret et al. 1999]. We found that the likelihood of transitions increased when the mispair induced relatively low levels of helix instability. One possible explanation for this observation could come from the mechanism of mismatch repair, in the sense that DNA mismatches producing high levels of instability are most efficiently recognized and repaired. Alternatively, it could be that the frequency of mismatch extension by the DNA polymerase is dependent upon local mispair stability, and that it is at this level the mutation bias will be introduced. Generally, it has been a longstanding hypothesis that both structural and thermodynamic properties of DNA mismatches will influence their frequency of occurrence as well as their propensity of becoming repaired [Hunter et al. 1986; Hunter et al. 1987]. Few, if any, have investigated these issues from the perspective of observed human mutation bias. Our results, based on genome-wide statistics of SNP patterns, provide support for a significant contribution by thermodynamics of non-Watson-Crick basepairs in the generation of DNA mutation bias in the human genome.

Paper IV is conceptually different from the three other papers in several respects. First of all, we here looked at genetic variation in terms of mononucleotide repeat distribution rather than point mutations and SNPs. Secondly, the paper put forward a more complex model with respect to the underlying causes of the observed patterns of variation. Mononucleotide repeats represent the most common form of microsatellites, and the processes that govern mutational variation at microsatellites have been subject to much research. Microsatellites have been proposed to occur by slipped strand mispairing within repetitive sequences, a mechanism that can take place both during crossing over at recombination and during DNA synthesis [Eckert and Hile 2009]. Both genomic and experimental studies have shown that motif size, length and composition affect mutagenesis [Webster et al. 2002]. Indeed, we observed that [A/T) mononucleotide repeats were far more abundant than [G/C] repeats (data not shown in **paper IV**). There is however some uncertainties whether the sequence bias of microsatellite mutagenesis manifest itself at the level of replication slippage, at the level of mismatch repair, or both [Ellegren 2002].

Our work did not focus on mononucleotide repeat mutability *per se*, but rather on the distinct overrepresentation of these sequences within MMR regions. The idea that recombination of wild-type and mutant MMR alleles induce a particular bias in the very own MMR genes represents a novel explanation for this observation. The study is limited in part by not knowing the exact repair bias of what we have coined MMR deficiency and MMR proficiency in the human germline, and also the unknown population frequencies of

MMR-deficient alleles. Testing the extent to which other sequence-modifying proteins exert similar effects on their own sequence will potentially elucidate whether the mechanism represents an important and common factor in DNA evolution.

4.3 FUTURE PROSPECTS

The biological basis of human DNA variation patterns represents an intriguing field of research that continues to fascinate the scientific community. Recent large-scale computational approaches have revealed new complexities regarding the factors that could influence germline mutation processes [Hodgkinson et al. 2009]. The patterns observed by means of large-scale analyses of genomes and variation data would nevertheless need to be complemented by experimental approaches that investigate germ cell biology. Importantly, longstanding assumptions about the roles of different biological factors in human germline are only beginning to receive experimental support.

In order to quantify the relative contributions of endogenous and exogenous sources to the observed local patterns in DNA variation, it is essential to obtain more accurate estimates of context-dependencies in DNA damage sources and repair enzymes in human germ cells. This could further facilitate more sophisticated mathematical modelling of the relative impact of damage and repair factors in the generation of human mutations. Better models of mutation bias would also need to examine the contribution that comes through DNA packaging and chromatin dynamics.

The most prominent hotspot of human mutation is the CpG dinucleotide. It appears to be a consensus that the underlying molecular mechanism is a spontaneous chemical reaction, i.e. the deamination of 5-methylcytosine to thymine that is followed by a less-than-perfect repair of T:G mispairs. Other sources of DNA damage do however also exhibit preference for CpG [Pfeifer 2006]. A future challenge would be to estimate the amount of transitions at CpG that in fact is caused by the spontaneous deamination events and not some other mutagenic event. One idea could be to investigate the local sequence biases (i.e. beyond the CpG dinucleotide) in the mutational spectra of chemically induced mutagens that preferentially form at CpG, and compare these potential biases with those observed at human CpG SNPs.

The developments of second-generation DNA sequencing technologies have the potential of revolutionizing studies of DNA variation at the whole-genome level [Shendure and Ji 2008]. In contrast to Sanger sequencing, cyclic-array technologies have the ability to

process millions of sequence reads in parallel, and may thus collect genetic variants at unprecedented resolution. As pointed out in the perspectives given by Gilad et al. [Gilad et al. 2009], it is now possible to identify every mutation that arises during experimental evolution, and to compare the entire genomes of ancestral and evolved strains. Thus, performing mutation accumulation experiments with next-generation sequencing could provide further insight into the patterns of spontaneous genomic mutations. A recent study that sequenced the full genomes in a family of four provided a direct estimate of human intergenerational mutation rate, that is, how much the genome changes from one human generation to the next [Roach et al. 2010]. The latter study represents a good example with respect to potential impacts of the new technology. Importantly however, relatively little is known about the sources of variation across the different next-generation sequencing platforms. A better understanding of the sources of error and bias in the next-generation sequencing data is thus essential for future applications.

Some notable large-scale sequencing efforts have now been established, largely facilitated by the new DNA sequencing technologies. The 1000 Genomes Project is an international collaboration that aims to sequence the full genomes of 1000 individuals, and in that respect creating a complete and more detailed catalogue of human genetic variation [Siva 2008]. The consortium aims to discover >95 % of the variants (e.g. SNPs, CNVs, indels) with minor allele frequencies as low as 1% across the genome and 0.1-0.5% in gene regions. All genetic variation data will further be made available to the research community. The 1000 Genomes Project will most importantly allow for novel discoveries between rare genetic variants and complex disease phenotypes. With respect to the topic addressed in this thesis, it seems clear that the increased resolution of SNPs will be a very important resource for discovering the most significant mutational signatures in recent evolution of the human genome. It is also likely that the current notion of a single human reference genome needs to be revised.

The International Cancer Genome Consortium (ICGC) is leading an additional large-scale sequencing project that will be important for future research [Stratton et al. 2009]. By means of full-genome sequencing of tumours associated with 50 different cancer types, the project aims to identify the great majority of somatic mutations that initiate cancer development (i.e. the driver mutations). This may potentially generate some unique sequence signatures of somatic cancer mutations that may aid the identification of endogenous and exogenous mutagenic mechanisms that initiate cancer. Preliminary reports

of two different tumour genomes appear promising in this respect [Stephens et al. 2009; Pleasance et al. 2010].

REFERENCES

Abelson JF, Kwan KY, O'Roak BJ, Baek DY, Stillman AA, Morgan TM, Mathews CA, Pauls DL, Rasin M-R, Gunel M, Davis NR, Ercan-Sencicek AG, Guez DH, Spertus JA, Leckman JF et al. (2005) Sequence variants in SLITRK1 are associated with Tourette's syndrome. *Science* **310**: 317-320.

Aitken RJ, De Iuliis GN. (2010) On the possible origins of DNA damage in human spermatozoa. *Mol Hum Reprod* **16**: 3-13.

Akey JM, Zhang G, Zhang K, Jin L, Shriver MD. (2002) Interrogating a high-density SNP map for signatures of natural selection. *Genome Res* **12**: 1805-1814.

Allawi HT, SantaLucia J. (1997) Thermodynamics and NMR of internal G.T mismatches in DNA. *Biochemistry* **36**: 10581-10594.

Allawi HT, SantaLucia J. (1998a) Nearest neighbor thermodynamic parameters for internal G.A mismatches in DNA. *Biochemistry* **37**: 2170-2179.

Allawi HT, SantaLucia J. (1998b) Nearest-neighbor thermodynamics of internal A.C mismatches in DNA: sequence dependence and pH effects. *Biochemistry* **37**: 9435-9444.

Allawi HT, SantaLucia J. (1998c) Thermodynamics of internal C.T mismatches in DNA. *Nucleic Acids Res* **26**: 2694-2701.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. (1990) Basic local alignment search tool. *J Mol Biol* **215**: 403-410.

Altshuler D, Pollara VJ, Cowles CR, Van Etten WJ, Baldwin J, Linton L, Lander ES. (2000) An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature* **407**: 513-516.

Arndt PF, Burge CB, Hwa T. (2003) DNA sequence evolution with neighbor-dependent mutation. *J Comput Biol* **10**: 313-322.

Arndt PF, Hwa T. (2005) Identification and measurement of neighbor-dependent nucleotide substitution processes. *Bioinformatics* **21**: 2322-2328.

Arnheim N, Calabrese P. (2009) Understanding what determines the frequency and pattern of human germline mutations. *Nat Rev Genet* **10**: 478-488.

Auerbach C, Kilbey BJ. (1971) Mutation in eukaryotes. *Annu Rev Genet* **5**: 163-218.

Auerbach C, Robson JM. (1947) The production of mutations by chemical substances. *Proc R Soc Edinb Biol* **62**: 271-283.

Avery OT, Macleod CM, McCarty M. (1944) Studies on the Chemical Nature of the Substance Inducing Transformation of Pneumococcal Types : Induction of Transformation by a Desoxyribonucleic Acid Fraction Isolated from Pneumococcus Type Iii. *J Exp Med* **79**: 137-158.

Bailey JA, Eichler EE. (2006) Primate segmental duplications: crucibles of evolution, diversity and disease. *Nat Rev Genet* **7**: 552-564.

Bailey JA, Gu Z, Clark RA, Reinert K, Samonte RV, Adams MD, Myers EW, Li PW, Eichler EE. (2002) Recent segmental duplications in the human genome. *Science* **297**: 1003-1007.

Balasubramanian B, Pogozelski WK, Tullius TD. (1998) DNA strand breaking by the hydroxyl radical is governed by the accessible surface areas of the hydrogen atoms of the DNA backbone. *Proc Natl Acad Sci USA* **95**: 9738-9743.

Balhorn R, Weston S, Thomas C, Wyrobek AJ. (1984) DNA packaging in mouse spermatids. Synthesis of protamine variants and four transition proteins. *Exp Cell Res* **150**: 298-308.

Ball EV, Stenson PD, Abeysinghe SS, Cooper DN, Chuzhanova NA. (2005) Microdeletions and microinsertions causing human genetic disease: common mechanisms of mutagenesis and the role of local DNA sequence complexity. *Hum Mutat* **26**: 205-213.

Barnes DE, Lindahl T. (2004) Repair and genetic consequences of endogenous DNA base damage in mammalian cells. *Annu Rev Genet* **38**: 445-476.

Bennett EA, Coleman LE, Tsui C, Pittard WS, Devine SE. (2004) Natural genetic variation caused by transposable elements in humans. *Genetics* **168**: 933-951.

Benzer S. (1961) On the Topography of the Genetic Fine Structure. *Proc Natl Acad Sci U S A* **47**: 403-415.

Benzer S, Freese E. (1958) Induction of Specific Mutations with 5-Bromouracil. *Proc Natl Acad Sci U S A* **44**: 112-119.

Bernstein F. (1925) Zusammenfassende Betrachtungen über die erblichen Blutstrukturen des Menschen. *Z. VererbLehre* **37**: 237-270.

Bersaglieri T, Sabeti PC, Patterson N, Vanderploeg T, Schaffner SF, Drake JA, Rhodes M, Reich DE, Hirschhorn JN. (2004) Genetic signatures of strong recent positive selection at the lactase gene. *Am J Hum Genet* **74**: 1111-1120.

Bird AP. (1986) CpG-rich islands and the function of DNA methylation. *Nature* **321**: 209-213.

Bird AP, Wolffe AP. (1999) Methylation-induced repression--belts, braces, and chromatin. *Cell* **99**: 451-454.

Biswas S, Akey JM. (2006) Genomic insights into positive selection. *Trends Genet* **22**: 437-446.

Blake RD, Hess ST, Nicholson-Tuell J. (1992) The influence of nearest neighbors on the rate and pattern of spontaneous point mutations. *J Mol Evol* **34**: 189-200.

Blanchette M, Kent WJ, Riemer C, Elnitski L, Smit AF, Roskin KM, Baertsch R, Rosenbloom K, Clawson H, Green ED, Haussler D, Miller W. (2004) Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res* **14**: 708-715.

Bodmer W, Bonilla C. (2008) Common and rare variants in multifactorial susceptibility to common diseases. *Nat Genet* **40**: 695-701.

Boffelli D, Nobrega MA, Rubin EM. (2004) Comparative genomics at the vertebrate extremes. *Nat Rev Genet* **5**: 456-465.

Boulikas T. (1992) Evolutionary consequences of nonrandom damage and repair of chromatin domains. *J Mol Evol* **35**: 156-180.

Boyle AP, Davis S, Shulha HP, Meltzer P, Margulies EH, Weng Z, Furey TS, Crawford GE. (2008) High-resolution mapping and characterization of open chromatin across the genome. *Cell* **132**: 311-322.

Breen AP, Murphy JA. (1995) Reactions of oxyl radicals with DNA. *Free Radic Biol Med* **18**: 1033-1077.

Brenner S, Benzer S, Barnett L. (1958) Distribution of proflavin-induced mutations in the genetic fine structure. *Nature* **182**: 983-985.

Brown TC, Jiricny J. (1987) A specific mismatch repair event protects mammalian cells from loss of 5-methylcytosine. *Cell* **50**: 945-950.

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

Buetow KH, Edmonson MN, Cassidy AB. (1999) Reliable identification of large numbers of candidate SNPs from public EST data. *Nat Genet* **21**: 323-325.

Busslinger M, Moschonas N, Flavell RA. (1981) Beta + thalassemia: aberrant splicing results from a single point mutation in an intron. *Cell* **27**: 289-298.

Byrne J, Rasmussen SA, Steinhorn SC, Connelly RR, Myers MH, Lynch CF, Flannery J, Austin DF, Holmes FF, Holmes GE, Strong LC, Mulvihill JJ. (1998) Genetic disease in offspring of long-term survivors of childhood and adolescent cancer. *Am J Hum Genet* **62**: 45-52.

Cadet J, Bellon S, Douki T, Frelon S, Gasparutto D, Muller E, Pouget J-P, Ravanat J-L, Romieu A, Sauvaigo S. (2004) Radiation-induced DNA damage: formation, measurement, and biochemical features. *J Environ Pathol Toxicol Oncol* **23**: 33-43.

Cadet J, Delatour T, Douki T, Gasparutto D, Pouget JP, Ravanat JL, Sauvaigo S. (1999) Hydroxyl radicals and DNA base damage. *Mutat Res* **424**: 9-21.

Cadet J, Douki T, Gasparutto D, Ravanat J-L. (2003) Oxidative damage to DNA: formation, measurement and biochemical features. *Mutat Res* **531**: 5-23.

Cadet J, Sage E, Douki T. (2005) Ultraviolet radiation-mediated damage to cellular DNA. *Mutat Res* **571**: 3-17.

Cargill M, Altshuler D, Ireland J, Sklar P, Ardlie K, Patil N, Shaw N, Lane CR, Lim EP, Kalyanaraman N, Nemesh J, Ziaugra L, Friedland L, Rolfe A, Warrington J et al. (1999) Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat Genet* **22**: 231-238.

Carrell DT, Hammoud SS. (2010) The human sperm epigenome and its potential role in embryonic development. *Mol Hum Reprod* **16**: 37-47.

Cartegni L, Chew SL, Krainer AR. (2002) Listening to silence and understanding nonsense: exonic mutations that affect splicing. *Nat Rev Genet* **3**: 285-298.

Caulfield JL, Wishnok JS, Tannenbaum SR. (1998) Nitric oxide-induced deamination of cytosine and guanine in deoxynucleosides and oligonucleotides. *J Biol Chem* **273**: 12689-12695.

Chakravarti A, Little P. (2003) Nature, nurture and human disease. *Nature* **421**: 412-414.

Chang JC, Kan YW. (1979) beta 0 thalassemia, a nonsense mutation in man. *Proc Natl Acad Sci U S A* **76**: 2886-2889.

Chen CL, Rappailles A, Duquenne L, Huvet M, Guilbaud G, Farinelli L, Audit B, d'Aubenton-Carafa Y, Arneodo A, Hyrien O, Thermes C. (2010) Impact of replication timing on non-CpG and CpG substitution rates in mammalian genomes. *Genome Res* **20**: 447-457.

Chen J, Sahota A, Stambrook PJ, Tischfield JA. (1991) Polymerase chain reaction amplification and sequence analysis of human mutant adenine phosphoribosyltransferase genes: the nature and frequency of errors caused by Taq DNA polymerase. *Mutat Res* **249**: 169-176.

Chen J-M, Férec C, Cooper DN. (2006) A systematic analysis of disease-associated variants in the 3' regulatory regions of human protein-coding genes I: general principles and overview. *Hum Genet* **120**: 1-21.

Chen K, Rajewsky N. (2006) Natural selection on human microRNA binding sites inferred from SNP data. *Nat Genet* **38**: 1452-1456.

Cheng KC, Cahill DS, Kasai H, Nishimura S, Loeb LA. (1992) 8-Hydroxyguanine, an abundant form of oxidative DNA damage, causes G----T and A----C substitutions. *J Biol Chem* **267**: 166-172.

Cheung J, Khaja R, Lau K, Tsui L-C, Scherer SW. (2003) Genome-wide detection of segmental duplications and potential assembly errors in the human genome sequence. *Genome Biol* **4**: R25.

Chi LM, Lam SL. (2007) NMR investigation of primer-template models: structural effect of sequence downstream of a thymine template on mutagenesis in DNA replication. *Biochemistry* **46**: 9292-9300.

Chi LM, Lam SL. (2009) NMR investigation of DNA primer-template models: guanine templates are less prone to strand slippage upon misincorporation. *Biochemistry* **48**: 11478-11486.

Chiang PK, Gordon RK, Tal J, Zeng GC, Doctor BP, Pardhasaradhi K, McCann PP. (1996) S-Adenosylmethionine and methylation. *FASEB J* **10**: 471-480.

Chimpanzee Sequencing and Analysis Consortium. (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437**: 69-87.

Cirulli ET, Goldstein DB. (2010) Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nat Rev Genet* **11**: 415-425.

Clark AG, Hubisz MJ, Bustamante CD, Williamson SH. (2005a) Ascertainment bias in studies of human genome-wide polymorphism. *Genome Res* **15**: 1496-1502.

Clark AG, Hubisz MJ, Bustamante CD, Williamson SH, Nielsen R. (2005b) Ascertainment bias in studies of human genome-wide polymorphism. *Genome Res* **15**: 1496-1502.

Clark TG, Andrew T, Cooper GM, Margulies EH, Mullikin JC, Balding DJ. (2007) Functional constraint and small insertions and deletions in the ENCODE regions of the human genome. *Genome Biol* **8**: R180.

Cloutier JF, Castonguay A, O'Connor TR, Drouin R. (2001) Alkylating agent and chromatin structure determine sequence context-dependent formation of alkylpurines. *J Mol Biol* **306**: 169-188.

Conne B, Stutz A, Vassalli JD. (2000) The 3' untranslated region of messenger RNA: A molecular 'hotspot' for pathology? *Nat Med* **6**: 637-641.

Cooke MS, Evans MD, Dizdaroglu M, Lunec J. (2003) Oxidative DNA damage: mechanisms, mutation, and disease. *FASEB J* **17**: 1195-1214.

Cooper DN. (1993) Human gene mutations affecting RNA processing and translation. *Ann Med* **25**: 11-17.

Cooper DN, Ball EV. (1998) The human gene mutation database. *Nucleic Acids Res* **26**: 285-287.

Cooper DN, Krawczak M. (1990) The mutational spectrum of single base-pair substitutions causing human genetic disease: patterns and predictions. *Hum Genet* **85**: 55-74.

Cooper DN, Krawczak M. (1993) *Human Gene Mutation*. BIOS Scientific Publishers Limited.

Cooper DN, Youssoufian H. (1988) The CpG dinucleotide and human genetic disease. *Hum Genet* **78**: 151-155.

Correns C. (1900) G. Mendels Regel über das Verhalten der Nachkommenschaft der Rassenbartarde. *Ber. dt. bot. Ges.* **18**: 158-168.

Cosman M, de los Santos C, Fiala R, Hingerty BE, Singh SB, Ibanez V, Margulis LA, Live D, Geacintov NE, Broyde S, et al. (1992) Solution conformation of the major adduct between the carcinogen (+)-anti-benzo[a]pyrene diol epoxide and DNA. *Proc Natl Acad Sci U S A* **89**: 1914-1918.

Cross SH, Bird AP. (1995) CpG islands and genes. *Curr Opin Genet Dev* **5**: 309-314.

Crutzen PJ, Andreae MO. (1990) Biomass Burning in the Tropics: Impact on Atmospheric Chemistry and Biogeochemical Cycles. *Science* **250**: 1669-1678.

Cummings WJ, Bednarski DW, Maizels N. (2008) Genetic variation stimulated by epigenetic modification. *PLoS One* **3**: e4075.

Czeizel AE, Elek C, Susanszky E. (1991) The evaluation of the germinal mutagenic impact of Chernobyl radiological contamination in Hungary. *Mutagenesis* **6**: 285-288.

Dalhus B, Laerdahl JK, Backe PH, Bjoras M. (2009) DNA base repair--recognition and initiation of catalysis. *FEMS Microbiol Rev* **33**: 1044-1078.

David SS, O'Shea VL, Kundu S. (2007) Base-excision repair of oxidative DNA damage. *Nature* **447**: 941-950.

De Armond R, Wood S, Sun D, Hurley LH, Ebbinghaus SW. (2005) Evidence for the presence of a guanine quadruplex forming region within a polypurine tract of the hypoxia inducible factor 1alpha promoter. *Biochemistry* **44**: 16341-16350.

De Bont R, van Larebeke N. (2004) Endogenous DNA damage in humans: a review of quantitative data. *Mutagenesis* **19**: 169-185.

De Vries H. (1901) Die Mutationstheorie. Versuche und Beobachtungen ueber die Entstehung von Arten im Pflanzenreich.

De Vries H. (1903) Befruchtung und Bastardierung. *Leipzig*.

Denissenko MF, Chen JX, Tang MS, Pfeifer GP. (1997) Cytosine methylation determines hot spots of DNA damage in the human P53 gene. *Proc Natl Acad Sci U S A* **94**: 3893-3898.

Dewannieux M, Esnault C, Heidmann T. (2003) LINE-mediated retrotransposition of marked Alu sequences. *Nat Genet* **35**: 41-48.

Donigan KA, Sweasy JB. (2009) Sequence context-specific mutagenesis and base excision repair. *Mol Carcinog* **48**: 362-368.

Drake JA, Bird C, Nemesh J, Thomas DJ, Newton-Cheh C, Reymond A, Excoffier L, Attar H, Antonarakis SE, Dermitzakis ET, Hirschhorn JN. (2006) Conserved noncoding sequences are selectively constrained and not mutation cold spots. *Nat Genet* **38**: 223-227.

Drake JW, Baltz RH. (1976) The biochemistry of mutagenesis. *Annu Rev Biochem* **45**: 11-37.

Dreszer TR, Wall GD, Haussler D, Pollard KS. (2007) Biased clustered substitutions in the human genome: the footprints of male-driven biased gene conversion. *Genome Res* **17**: 1420-1430.

Drost JB, Lee WR. (1995) Biological basis of germline mutation: comparisons of spontaneous germline mutation rates among drosophila, mouse, and human. *Environ Mol Mutagen* **25 Suppl 26**: 48-64.

Du Z, Zhao Y, Li N. (2008) Genome-wide analysis reveals regulatory role of G4 DNA in gene transcription. *Genome Res* **18**: 233-241.

Duquette ML, Handa P, Vincent JA, Taylor AF, Maizels N. (2004) Intracellular transcription of G-rich DNAs induces formation of G-loops, novel structures containing G4 DNA. *Genes Dev* **18**: 1618-1629.

Duret L, Galtier N. (2009) Biased gene conversion and the evolution of mammalian genomic landscapes. *Annu Rev Genomics Hum Genet* **10**: 285-311.

Eckert KA, Hile SE. (2009) Every microsatellite is different: Intrinsic DNA features dictate mutagenesis of common microsatellites present in the human genome. *Mol Carcinog* **48**: 379-388.

Eckhardt F, Lewin J, Cortese R, Rakyan VK, Attwood J, Burger M, Burton J, Cox TV, Davies R, Down TA, Haefliger C, Horton R, Howe K, Jackson DK, Kunde J et al. (2006) DNA methylation profiling of human chromosomes 6, 20 and 22. *Nat Genet* **38**: 1378-1385.

Edvardsen H, Irene Grenaker Alnaes G, Tsalenko A, Mulcahy T, Yuryev A, Lindersson M, Lien S, Omholt S, Syvanen AC, Borresen-Dale AL, Kristensen VN. (2006) Experimental validation of data mined single nucleotide polymorphisms from several databases and consecutive dbSNP builds. *Pharmacogenet Genomics* **16**: 207-217.

Eftedal I, Guddal PH, Slupphaug G, Volden G, Krokan HE. (1993) Consensus sequences for good and poor removal of uracil from double stranded DNA by uracil-DNA glycosylase. *Nucleic Acids Res* **21**: 2095-2101.

Eichler EE. (2001) Recent duplication, domain accretion and the dynamic mutation of the human genome. *Trends Genet* **17**: 661-669.

Elango N, Kim S-H, Vigoda E, Yi SV. (2008) Mutations of different molecular origins exhibit contrasting patterns of regional substitution rate variation. *PLoS Comput Biol* **4**: e1000015.

Ellegren H. (2002) Mismatch repair and mutational bias in microsatellite DNA. *Trends Genet* **18**: 552.

Ellegren H. (2007) Characteristics, causes and evolutionary consequences of male-biased mutation. *Proc Biol Sci* **274**: 1-10.

ElSharawy A, Hundrieser B, Brosch M, Wittig M, Huse K, Platzer M, Becker A, Simon M, Rosenstiel P, Schreiber S, Krawczak M, Hampe J. (2009) Systematic evaluation of the effect of common SNPs on pre-mRNA splicing. *Hum Mutat* **30**: 625-632.

Ewing B, Green P. (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res* **8**: 186-194.

Ewing B, Hillier L, Wendl MC, Green P. (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res* **8**: 175-185.

Fairbrother WG, Holste D, Burge CB, Sharp PA. (2004) Single nucleotide polymorphism-based validation of exonic splicing enhancers. *PLoS Biol* **2**: E268.

Falster DS, Nakken S, Bergem-Ohr M, Rødland EA, Breivik J. (2010) Unstable DNA Repair Genes Shaped by Their Own Sequence Modifying Phenotypes. *J Mol Evol* **70**: 266-274.

Feuk L, Carson AR, Scherer SW. (2006) Structural variation in the human genome. *Nat Rev Genet* **7**: 85-97.

Flemming W. (1882) Zellsubstanz, Kern- und Zelltheilung. *Leipzig*.

Fousteri M, Mullenders LH. (2008) Transcription-coupled nucleotide excision repair in mammalian cells: molecular mechanisms and biological effects. *Cell Res* **18**: 73-84.

Franklin WA, Doetsch PW, Haseltine WA. (1985) Structural determination of the ultraviolet light-induced thymine-cytosine pyrimidine-pyrimidone (6-4) photoproduct. *Nucleic Acids Res* **13**: 5317-5325.

Frazer KA, Murray SS, Schork NJ, Topol EJ. (2009) Human genetic variation and its contribution to complex traits. *Nat Rev Genet* **10**: 241-251.

Frederico LA, Kunkel TA, Shaw BR. (1993) Cytosine deamination in mismatched base pairs. *Biochemistry* **32**: 6523-6530.

Fredman D, White SJ, Potter S, Eichler EE, Den Dunnen JT, Brookes AJ. (2004) Complex SNP-related sequence variation in segmental genome duplications. *Nat Genet* **36**: 861-866.

Freeman JL, Perry GH, Feuk L, Redon R, McCarroll SA, Altshuler DM, Aburatani H, Jones KW, Tyler-Smith C, Hurles ME, Carter NP, Scherer SW, Lee C. (2006) Copy number variation: new insights in genome diversity. *Genome Res* **16**: 949-961.

Freese E. (1959) The Difference between Spontaneous and Base-Analogue Induced Mutations of Phage T4. *Proc Natl Acad Sci U S A* **45**: 622-633.

Freese E, Bautz E, Freese EB. (1961) The chemical and mutagenic specificity of hydroxylamine. *Proc Natl Acad Sci U S A* **47**: 845-855.

Friedberg EC, Walker GC, Siede W, Wood RD, Schultz RA, Ellenberger T. (2006) *DNA Repair and Mutagenesis*. ASM Press.

Fryxell KJ, Moon W-J. (2005) CpG mutation rates in the human genome are highly dependent on local GC content. *Mol Biol Evol* **22**: 650-658.

Gedik CM, Collins A. (2005) Establishing the background level of base oxidation in human lymphocyte DNA: results of an interlaboratory validation study. *FASEB J* **19**: 82-84.

Gilad Y, Pritchard JK, Thornton K. (2009) Characterizing natural variation using next-generation sequencing technologies. *Trends Genet* **25**: 463-471.

Gonzalgo ML, Jones PA. (1997) Mutagenic and epigenetic effects of DNA methylation. *Mutat Res* **386**: 107-118.

Gordon LK, Haseltine WA. (1982) Quantitation of cyclobutane pyrimidine dimer formation in double- and single-stranded DNA fragments of defined sequence. *Radiat Res* **89**: 99-112.

Goriely A, McVean GA, Rojmyr M, Ingemarsson B, Wilkie AO. (2003) Evidence for selective advantage of pathogenic FGFR2 mutations in the male germ line. *Science* **301**: 643-646.

Green P, Ewing B, Miller W, Thomas PJ, Program NCS, Green ED. (2003) Transcription-associated mutational asymmetry in mammalian evolution. *Nat Genet* **33**: 514-517.

Greenbaum JA, Pang B, Tullius TD. (2007) Construction of a genome-scale structural map at single-nucleotide resolution. *Genome Res* **17**: 947-953.

Hayatsu H, Wataya Y, Kai K, Iida S. (1970) Reaction of sodium bisulfite with uracil, cytosine, and their derivatives. *Biochemistry* **9**: 2858-2865.

Hecht SS. (1999) DNA adduct formation from tobacco-specific N-nitrosamines. *Mutat Res* **424**: 127-142.

Hegan DC, Narayanan L, Jirik FR, Edelmann W, Liskay RM, Glazer PM. (2006) Differing patterns of genetic instability in mice deficient in the mismatch repair genes Pms2, Mlh1, Msh2, Msh3 and Msh6. *Carcinogenesis* **27**: 2402-2408.

Hellmann I, Ebersberger I, Ptak SE, Paabo S, Przeworski M. (2003) A neutral explanation for the correlation of diversity with recombination rates in humans. *Am J Hum Genet* **72**: 1527-1535.

Hellmann I, Prüfer K, Ji H, Zody MC, Pääbo S, Ptak SE. (2005) Why do human diversity levels vary at a megabase scale? *Genome Res* **15**: 1222-1231.

Hendrich B, Hardeland U, Ng HH, Jiricny J, Bird A. (1999) The thymine glycosylase MBD4 can bind to the product of deamination at methylated CpG sites. *Nature* **401**: 301-304.

Hernandez RD, Williamson SH, Bustamante CD. (2007) Context dependence, ancestral misidentification, and spurious signatures of natural selection. *Mol Biol Evol* **24**: 1792-1800.

Hess ST, Blake JD, Blake RD. (1994) Wide variations in neighbor-dependent substitution rates. *J Mol Biol* **236**: 1022-1033.

Hinds DA, Stuve LL, Nilsen GB, Halperin E, Eskin E, Ballinger DG, Frazer KA, Cox DR. (2005) Whole-genome patterns of common DNA variation in three human populations. *Science* **307**: 1072-1079.

Hodgkinson A, Ladoukakis ED, Eyre-Walker A. (2009) Cryptic Variation in the Human Mutation Rate. *PLoS Biol* **7**: e27.

Hoogendoorn B, Coleman SL, Guy CA, Smith K, Bowen T, Buckland PR, O'Donovan MC. (2003) Functional analysis of human promoter polymorphisms. *Hum Mol Genet* **12**: 2249-2254.

Hsu GW, Ober M, Carell T, Beese LS. (2004) Error-prone replication of oxidatively damaged DNA by a high-fidelity DNA polymerase. *Nature* **431**: 217-221.

Hughes AL, Packer B, Welch R, Bergen AW, Chanock SJ, Yeager M. (2003) Widespread purifying selection at polymorphic sites in human protein-coding loci. *Proc Natl Acad Sci USA* **100**: 15754-15757.

Hunter WN, Brown T, Anand NN, Kennard O. (1986) Structure of an adenine-cytosine base pair in DNA and its implications for mismatch repair. *Nature* **320**: 552-555.

Hunter WN, Brown T, Kneale G, Anand NN, Rabinovich D, Kennard O. (1987) The structure of guanosine-thymidine mismatches in B-DNA at 2.5-A resolution. *J Biol Chem* **262**: 9962-9970.

Huppert JL, Balasubramanian S. (2005) Prevalence of quadruplexes in the human genome. *Nucleic Acids Res* **33**: 2908-2916.

Hurst LD. (2009) Fundamental concepts in genetics: genetics and the understanding of selection. *Nat Rev Genet* **10**: 83-93.

Hwang DG, Green P. (2004) Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. *Proc Natl Acad Sci USA* **101**: 13994-14001.

Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, Scherer SW, Lee C. (2004) Detection of large-scale variation in the human genome. *Nat Genet* **36**: 949-951.

Intano GW, McMahan CA, Walter RB, McCarrey JR, Walter CA. (2001) Mixed spermatogenic germ cell nuclear extracts exhibit high base excision repair activity. *Nucleic Acids Res* **29**: 1366-1372.

International HapMap Consortium. (2003) The International HapMap Project. *Nature* **426**: 789-796.

International HapMap Consortium. (2005) A haplotype map of the human genome. *Nature* **437**: 1299-1320.

International HapMap Consortium. (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**: 851-861.

International Human Genome Sequencing Consortium. (2004) Finishing the euchromatic sequence of the human genome. *Nature* **431**: 931-945.

Irizarry K, Kustanovich V, Li C, Brown N, Nelson S, Wong W, Lee CJ. (2000) Genome-wide analysis of single-nucleotide polymorphisms in human expressed sequences. *Nat Genet* **26**: 233-236.

Jackson SP, Bartek J. (2009) The DNA-damage response in human biology and disease. *Nature* **461**: 1071-1078.

Jansen RP. (2001) mRNA localization: message on the move. *Nat Rev Mol Cell Biol* **2**: 247-256.

Janssens FA. (1909) Spermatogénèse dans les Batraciens. V. La théorie de la chiasmatypie. Nouvelles interprétation des cinèses de maturation. *Cellule* **25**: 387-411.

Jaroudi S, SenGupta S. (2007) DNA repair in mammalian embryos. *Mutat Res* **635**: 53-77.

Jeffreys AJ, Neumann R. (2002) Reciprocal crossover asymmetry and meiotic drive in a human recombination hot spot. *Nat Genet* **31**: 267-271.

Jeffreys AJ, Royle NJ, Wilson V, Wong Z. (1988) Spontaneous mutation rates to new length alleles at tandem-repetitive hypervariable loci in human DNA. *Nature* **332**: 278-281.

Jeffreys AJ, Wilson V, Thein SL. (1985a) Hypervariable 'minisatellite' regions in human DNA. *Nature* **314**: 67-73.

Jeffreys AJ, Wilson V, Thein SL. (1985b) Individual-specific 'fingerprints' of human DNA. *Nature* **316**: 76-79.

Jiang C, Zhao Z. (2006) Mutational spectrum in the recent human genome inferred by single nucleotide polymorphisms. *Genomics* **88**: 527-534.

Johnson SJ, Beese LS. (2004) Structures of mismatch replication errors observed in a DNA polymerase. *Cell* **116**: 803-816.

Kaiser J. (2008) DNA sequencing. A plan to capture human diversity in 1000 genomes. *Science* **319**: 395.

Karnani N, Taylor C, Malhotra A, Dutta A. (2007) Pan-S replication patterns and chromosomal domains defined by genome-tiling arrays of ENCODE genomic areas. *Genome Res* **17**: 865-876.

Karolchik D, Baertsch R, Diekhans M, Furey TS, Hinrichs A, Lu YT, Roskin KM, Schwartz M, Sugnet CW, Thomas DJ, Weber RJ, Haussler D, Kent WJ. (2003) The UCSC Genome Browser Database. *Nucleic Acids Res* **31**: 51-54.

Kehrer-Sawatzki H, Cooper DN. (2007) Understanding the recent evolution of the human genome: insights from human-chimpanzee genome comparisons. *Hum Mutat* **28**: 99-130.

Keohavong P, Liu VF, Thilly WG. (1991) Analysis of point mutations induced by ultraviolet light in human cells. *Mutat Res* **249**: 147-159.

Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, Graves T, Hansen N, Teague B, Alkan C, Antonacci F, Haugen E, Zerr T, Yamada NA, Tsang P, Newman TL et al. (2008) Mapping and sequencing of structural variation from eight human genomes. *Nature* **453**: 56-64.

Kimmins S, Sassone-Corsi P. (2005) Chromatin remodelling and epigenetic features of germ cells. *Nature* **434**: 583-589.

Kimura M. (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol* **16**: 111-120.

Knight JC. (2005) Regulatory polymorphisms underlying complex disease traits. *J Mol Med* **83**: 97-109.

Kobayashi K, Karran P, Oda S, Yanaga K. (2005) Involvement of mismatch repair in transcription-coupled nucleotide excision repair. *Hum Cell* **18**: 103-115.

Kondrashov AS, Rogozin IB. (2004) Context of deletions and insertions in human coding sequences. *Hum Mutat* **23**: 177-185.

Kong A, Gudbjartsson DF, Sainz J, Jonsdottir GM, Gudjonsson SA, Richardsson B, Sigurdardottir S, Barnard J, Hallbeck B, Masson G, Shlien A, Palsson ST, Frigge ML, Thorgeirsson TE, Gulcher JR et al. (2002) A high-resolution recombination map of the human genome. *Nat Genet* **31**: 241-247.

Kozack R, Seo KY, Jelinsky SA, Loechler EL. (2000) Toward an understanding of the role of DNA adduct conformation in defining mutagenic mechanism based on studies of the major adduct (formed at N(2)-dG) of the potent environmental carcinogen, benzo[a]pyrene. *Mutat Res* **450**: 41-59.

Krawczak M, Ball EV, Cooper DN. (1998) Neighboring-nucleotide effects on the rates of germ-line single-base-pair substitution in human genes. *Am J Hum Genet* **63**: 474-488.

Krokan HE, Drablos F, Slupphaug G. (2002) Uracil in DNA--occurrence, consequences and repair. *Oncogene* **21**: 8935-8948.

Kuluncsics Z, Perdiz D, Brulay E, Muel B, Sage E. (1999) Wavelength dependence of ultraviolet-induced DNA damage distribution: involvement of direct or indirect mechanisms and possible artefacts. *J Photochem Photobiol B* **49**: 71-80.

Kunkel TA. (1985a) The mutational specificity of DNA polymerase-beta during in vitro DNA synthesis. Production of frameshift, base substitution, and deletion mutations. *J Biol Chem* **260**: 5787-5796.

Kunkel TA. (1985b) The mutational specificity of DNA polymerases-alpha and -gamma during in vitro DNA synthesis. *J Biol Chem* **260**: 12866-12874.

Kunkel TA. (2004) DNA replication fidelity. *J Biol Chem* **279**: 16895-16898.

Kunkel TA, Alexander PS. (1986) The base substitution fidelity of eucaryotic DNA polymerases. Mispairing frequencies, site preferences, insertion preferences, and base substitution by dislocation. *J Biol Chem* **261**: 160-166.

Kunkel TA, Loeb LA. (1981) Fidelity of mammalian DNA polymerases. *Science* **213**: 765-767.

Kunkel TA, Soni A. (1988) Mutagenesis by transient misalignment. *J Biol Chem* **263**: 14784-14789.

Kvikstad EM, Tyekucheva S, Chiaromonte F, Makova KD. (2007) A macaque's-eye view of human insertions and deletions: differences in mechanisms. *PLoS Comput Biol* **3**: 1772-1782.

Ladoukakis ED, Eyre-Walker A. (2007) Searching for sequence directed mutagenesis in eukaryotes. *J Mol Evol* **64**: 1-3.

Lander ES, Linton LM, Birren BW, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J et al. (2001) Initial sequencing and analysis of the human genome. *Nature* **409**: 860-921.

Lemaire DG, Ruzsicska BP. (1993) Kinetic analysis of the deamination reactions of cyclobutane dimers of thymidylyl-3',5'-2'-deoxycytidine and 2'-deoxycytidylyl-3',5'-thymidine. *Biochemistry* **32**: 2525-2533.

Levinson G, Gutman GA. (1987) Slipped-strand mispairing: a major mechanism for DNA sequence evolution. *Mol Biol Evol* **4**: 203-221.

Levy S, Feuk L, Halpern AL, Walenz BP, Axelrod N, Huang J, Kirkness EF, Denisov G, Lin Y, Pang AWC, Stockwell TB, Tsiamouri A, Bafna V, Bansal V, Kravitz SA et al. (2007) The diploid genome sequence of an individual human. *PLoS Biol* **5**: e254.

Lewis SE. (1999) Life cycle of the mammalian germ cell: implication for spontaneous mutation frequencies. *Teratology* **59**: 205-209.

Li E, Bestor TH, Jaenisch R. (1992) Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. *Cell* **69**: 915-926.

Li H, Ruan J, Durbin R. (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res* **18**: 1851-1858.

Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, Sandstrom R, Bernstein B, Bender MA, Groudine M, Gnirke A et al. (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**: 289-293.

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

Lindahl T, Nyberg B. (1974) Heat-induced deamination of cytosine residues in deoxyribonucleic acid. *Biochemistry* **13**: 3405-3410.

Lister R, Ecker JR. (2009) Finding the fifth base: genome-wide sequencing of cytosine methylation. *Genome Res* **19**: 959-966.

Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, Nery JR, Lee L, Ye Z, Ngo Q-M, Edsall L, Antosiewicz-Bourget J, Stewart R, Ruotti V, Millar AH et al. (2009a) Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* **462**: 315-322.

Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, Nery JR, Lee L, Ye Z, Ngo QM, Edsall L, Antosiewicz-Bourget J, Stewart R, Ruotti V, Millar AH et al. (2009b) Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* **462**: 315-322.

Liu F, Tøstesen E, Sundet JK, Jenssen T-K, Bock C, Jerstad GI, Thilly WG, Hovig E. (2007) The human genomic melting map. *PLoS Comput Biol* **3**: e93.

Loeb LA, Monnat RJ. (2008) DNA polymerases and human disease. *Nat Rev Genet* **9**: 594-604.

Lovett ST. (2004) Encoded errors: mutations and rearrangements mediated by misalignment at repetitive DNA sequences. *Mol Microbiol* **52**: 1243-1253.

Lu SC. (2000) S-Adenosylmethionine. *Int J Biochem Cell Biol* **32**: 391-395.

Luger K, Mader AW, Richmond RK, Sargent DF, Richmond TJ. (1997) Crystal structure of the nucleosome core particle at 2.8 A resolution. *Nature* **389**: 251-260.

Lupski JR, Stankiewicz P. (2005) Genomic disorders: molecular mechanisms for rearrangements and conveyed phenotypes. *PLoS Genet* **1**: e49.

Lutsenko E, Bhagwat AS. (1999) Principal causes of hot spots for cytosine to thymine mutations at sites of cytosine methylation in growing cells. A model, its experimental support and implications. *Mutat Res* **437**: 11-20.

Lynch HT, Boland CR, Gong G, Shaw TG, Lynch PM, Fodde R, Lynch JF, de la Chapelle A. (2006) Phenotypic and genotypic heterogeneity in the Lynch syndrome: diagnostic, surveillance and management implications. *Eur J Hum Genet* **14**: 390-402.

Madsen BE, Villesen P, Wiuf C. (2007) A periodic pattern of SNPs in the human genome. *Genome Res* **17**: 1414-1419.

Maizels N. (2005) Immunoglobulin gene diversification. *Annu Rev Genet* **39**: 23-46.

Maizels N. (2006) Dynamic roles for G4 DNA in the biology of eukaryotic cells. *Nat Struct Mol Biol* **13**: 1055-1059.

Majewski J. (2003) Dependence of mutational asymmetry on gene-expression levels in the human genome. *Am J Hum Genet* **73**: 688-692.

Makova KD, Li WH. (2002) Strong male-driven evolution of DNA sequences in humans and apes. *Nature* **416**: 624-626.

Manolio TA, Brooks LD, Collins FS. (2008) A HapMap harvest of insights into the genetics of common disease. *J Clin Invest* **118**: 1590-1605.

Marra G, Schär P. (1999) Recognition of DNA alterations by the mismatch repair system. *Biochem J* **338 (Pt 1)**: 1-13.

Marth G, Yeh R, Minton M, Donaldson R, Li Q, Duan S, Davenport R, Miller RD, Kwok PY. (2001) Single-nucleotide polymorphisms in the public domain: how useful are they? *Nat Genet* **27**: 371-372.

Mathews CK. (2006) DNA precursor metabolism and genomic stability. *FASEB J* **20**: 1300-1314.

Mathews CK, Ji J. (1992) DNA precursor asymmetries, replication fidelity, and variable genome evolution. *Bioessays* **14**: 295-301.

Mazurek A, Johnson CN, Germann MW, Fishel R. (2009) Sequence context effect for hMSH2-hMSH6 mismatch-dependent activation. *Proc Natl Acad Sci USA* **106**: 4177-4182.

McCarroll SA, Huett A, Kuballa P, Chilewski S, Landry A, Goyette P, Zody MC, Hall J, Brant S, Cho J, Duerr R, Silverberg M, Taylor K, Rioux J, Altshuler DA et al. (2008) Deletion polymorphism upstream of IRGM associated with altered IRGM expression and Crohn's disease. *Nat Genet* **40**: 1107-1112.

Meissner A, Mikkelsen TS, Gu H, Wernig M, Hanna J, Sivachenko A, Zhang X, Bernstein BE, Nusbaum C, Jaffe DB, Gnirke A, Jaenisch R, Lander ES. (2008) Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature* **454**: 766-770.

Messer PW, Arndt PF. (2007) The majority of recent short DNA insertions in the human genome are tandem duplications. *Mol Biol Evol* **24**: 1190-1197.

Millard JT, Weidner MF, Kirchner JJ, Ribeiro S, Hopkins PB. (1991) Sequence preferences of DNA interstrand crosslinking agents: quantitation of interstrand crosslink locations in DNA duplex fragments containing multiple crosslinkable sites. *Nucleic Acids Res* **19**: 1885-1891.

Miller JH. (1985) Mutagenic specificity of ultraviolet light. *J Mol Biol* **182**: 45-65.

Mills RE. (2006) An initial map of insertion and deletion (INDEL) variation in the human genome. *Genome Res* **16**: 1182-1190.

Mills RE, Bennett EA, Iskow RC, Devine SE. (2007) Which transposable elements are active in the human genome? *Trends Genet* **23**: 183-191.

Mitchell DL, Jen J, Cleaver JE. (1991) Relative induction of cyclobutane dimers and cytosine photohydrates in DNA irradiated in vitro and in vivo with ultraviolet-C and ultraviolet-B light. *Photochem Photobiol* **54**: 741-746.

Mitchell-Olds T, James RV, Palmer MJ, Williams PH. (1995) Genetics of Brassica rapa (syn. campestris). 2. Multiple disease resistance to three fungal pathogens: Peronospora parasitica, Albugo candida and Leptosphaeria maculans. *Heredity* **75 (Pt 4)**: 362-369.

Mouse Genome Sequencing Consortium. (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**: 520-562.

Muller HJ. (1927) Artificial transmutation of the gene. *Science* **66**: 84-87.

Muller HJ. (1928a) The Measurement of Gene Mutation Rate in Drosophila, Its High Variability, and Its Dependence upon Temperature. *Genetics* **13**: 279-357.

Muller HJ. (1928b) The problem of genic modification. *Z. VererbLehre*: 234-260.

Muller HJ, Altenburg E. (1930) The Frequency of Translocations Produced by X-Rays in Drosophila. *Genetics* **15**: 283-311.

Muramatsu M, Kinoshita K, Fagarasan S, Yamada S, Shinkai Y, Honjo T. (2000) Class switch recombination and hypermutation require activation-induced cytidine deaminase (AID), a potential RNA editing enzyme. *Cell* **102**: 553-563.

Murnane JP. (1996) Role of induced genetic instability in the mutagenic effects of chemicals and radiation. *Mutat Res* **367**: 11-23.

Musumeci L, Arthur JW, Cheung FS, Hoque A, Lippman S, Reichardt JK. (2010) Single nucleotide differences (SNDs) in the dbSNP database may lead to errors in genotyping and haplotyping studies. *Hum Mutat* **31**: 67-73.

Myers S, Bottolo L, Freeman C, McVean G, Donnelly P. (2005) A fine-scale map of recombination rates and hotspots across the human genome. *Science* **310**: 321-324.

Nakken S, Rødland EA, Rognes T, Hovig E. (2009a) Large-scale inference of the point mutational spectrum in human segmental duplications. *BMC Genomics* **10**: 43.

Nakken S, Rognes T, Hovig E. (2009b) The disruptive positions in human G-quadruplex motifs are less polymorphic and more conserved than their neutral counterparts. *Nucleic Acids Res* **37**: 5749-5756.

Nelson MR, Marnellos G, Kammerer S, Hoyal CR, Siepel M, Cantor CR, Braun A. (2004) Large-scale validation of single nucleotide polymorphisms in gene regions. *Genome Res* **14**: 1664-1668.

Ng PC, Henikoff S. (2006) Predicting the effects of amino acid substitutions on protein function. *Annu Rev Genomics Hum Genet* **7**: 61-80.

Nielsen R. (2005) Molecular signatures of natural selection. *Annu Rev Genet* **39**: 197-218.

Nielsen R, Hellmann I, Hubisz M, Bustamante C, Clark AG. (2007) Recent and ongoing selection in the human genome. *Nat Rev Genet* **8**: 857-868.

Nohmi T, Masumura K. (2005) Molecular nature of intrachromosomal deletions and base substitutions induced by environmental mutagens. *Environ Mol Mutagen* **45**: 150-161.

Oakes CC, La Salle S, Smiraglia DJ, Robaire B, Trasler JM. (2007) A unique configuration of genome-wide DNA methylation patterns in the testis. *Proc Natl Acad Sci USA* **104**: 228-233.

Ogura Y, Bonen DK, Inohara N, Nicolae DL, Chen FF, Ramos R, Britton H, Moran T, Karaliuskas R, Duerr RH, Achkar JP, Brant SR, Bayless TM, Kirschner BS, Hanauer SB et al. (2001) A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease. *Nature* **411**: 603-606.

Ohno M, Miura T, Furuichi M, Tominaga Y, Tsuchimoto D, Sakumi K, Nakabeppu Y. (2006) A genome-wide distribution of 8-oxoguanine correlates with the preferred regions for recombination and single nucleotide polymorphism in the human genome. *Genome Res* **16**: 567-575.

Okano M, Bell DW, Haber DA, Li E. (1999) DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. *Cell* **99**: 247-257.

Olsen A-K, Duale N, Bjørås M, Larsen CT, Wiger R, Holme JA, Seeberg EC, Brunborg G. (2003) Limited repair of 8-hydroxy-7,8-dihydroguanine residues in human testicular cells. *Nucleic Acids Res* **31**: 1351-1363.

Olsen AK, Bjørtuft H, Wiger R, Holme J, Seeberg E, Bjørås M, Brunborg G. (2001) Highly efficient base excision repair (BER) in human and rat male germ cells. *Nucleic Acids Res* **29**: 1781-1790.

Otake M, Schull WJ, Neel JV. (1990) Congenital malformations, stillbirths, and early mortality among the children of atomic bomb survivors: a reanalysis. *Radiat Res* **122**: 1-11.

Otsuka K, Suzuki T, Shibata H, Kato S, Sakayori M, Shimodaira H, Kanamaru R, Ishioka C. (2003) Analysis of the human APC mutation spectrum in a saccharomyces cerevisiae strain with a mismatch repair defect. *Int J Cancer* **103**: 624-630.

Paeschke K, Simonsson T, Postberg J, Rhodes D, Lipps HJ. (2005) Telomere end-binding proteins control the formation of G-quadruplex DNA structures in vivo. *Nat Struct Mol Biol* **12**: 847-854.

Park H, Zhang K, Ren Y, Nadji S, Sinha N, Taylor JS, Kang C. (2002) Crystal structure of a DNA decamer containing a cis-syn thymine dimer. *Proc Natl Acad Sci U S A* **99**: 15965-15970.

Pastinen T, Hudson TJ. (2004) Cis-acting regulatory variation in the human genome. *Science* **306**: 647-650.

Pearson CE, Nichol Edamura K, Cleary JD. (2005) Repeat instability: mechanisms of dynamic mutations. *Nat Rev Genet* **6**: 729-742.

Petronzelli F, Riccio A, Markham GD, Seeholzer SH, Genuardi M, Karbowski M, Yeung AT, Matsumoto Y, Bellacosa A. (2000) Investigation of the substrate spectrum of the human mismatch-specific DNA N-glycosylase MED1 (MBD4): fundamental role of the catalytic domain. *J Cell Physiol* **185**: 473-480.

Peyret N, Seneviratne PA, Allawi HT, SantaLucia J. (1999) Nearest-neighbor thermodynamics and NMR of DNA sequences with internal A.A, C.C, G.G, and T.T mismatches. *Biochemistry* **38**: 3468-3477.

Pfeifer GP. (2006) Mutagenesis at methylated CpG sequences. *Curr Top Microbiol Immunol* **301**: 259-281.

Pfeifer GP, Besaratinia A. (2009) Mutational spectra of human cancer. *Hum Genet* **125**: 493-506.

Pfeifer GP, Denissenko MF, Olivier M, Tretyakova N, Hecht SS, Hainaut P. (2002) Tobacco smoke carcinogens, DNA damage and p53 mutations in smoking-associated cancers. *Oncogene* **21**: 7435-7451.

Platzer M, Hiller M, Szafranski K, Jahn N, Hampe J, Schreiber S, Backofen R, Huse K. (2006) Sequencing errors or SNPs at splice-acceptor guanines in dbSNP? *Nat Biotechnol* **24**: 1068-1070.

Pleasance ED, Cheetham RK, Stephens PJ, McBride DJ, Humphray SJ, Greenman CD, Varela I, Lin M-L, Ordóñez GR, Bignell GR, Ye K, Alipaz J, Bauer MJ, Beare D, Butler A et al. (2010) A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* **463**: 191-196.

Prakash S, Johnson RE, Prakash L. (2005) Eukaryotic translesion synthesis DNA polymerases: specificity of structure and function. *Annu Rev Biochem* **74**: 317-353.

Prendergast JG, Campbell H, Gilbert N, Dunlop MG, Bickmore WA, Semple CA. (2007) Chromatin structure and evolution in the human genome. *BMC Evol Biol* **7**: 72.

Pruitt KD, Tatusova T, Maglott DR. (2005) NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* **33**: D501-504.

Qin J, Calabrese P, Tiemann-Boege I, Shinde D, Yoon S-R, Gelfand D, Bauer K, Arnheim N. (2007) The molecular anatomy of spontaneous germline mutations in human testes. *PLoS Biol* **5**: e224.

Qu HQ, Lawrence SG, Guo F, Majewski J, Polychronakos C. (2006) Strand bias in complementary single-nucleotide polymorphisms of transcribed human sequences: evidence for functional effects of synonymous polymorphisms. *BMC Genomics* **7**: 213.

Rabbitts TH. (1994) Chromosomal translocations in human cancer. *Nature* **372**: 143-149.

Radany EH, Dornfeld KJ, Sanderson RJ, Savage MK, Majumdar A, Seidman MM, Mosbaugh DW. (2000) Increased spontaneous mutation frequency in human cells expressing the phage PBS2-encoded inhibitor of uracil-DNA glycosylase. *Mutat Res* **461**: 41-58.

Rajski SR, Jackson BA, Barton JK. (2000) DNA repair: models for damage and mismatch recognition. *Mutat Res* **447**: 49-72.

Rakyan VK, Down TA, Thorne NP, Flicek P, Kulesha E, Graf S, Tomazou EM, Backdahl L, Johnson N, Herberth M, Howe KL, Jackson DK, Miretti MM, Fiegler H, Marioni JC et al. (2008) An integrated resource for genome-wide identification and analysis of human tissue-specific differentially methylated regions (tDMRs). *Genome Res* **18**: 1518-1529.

Ramsahoye BH, Biniszkiewicz D, Lyko F, Clark V, Bird AP, Jaenisch R. (2000) Non-CpG methylation is prevalent in embryonic stem cells and may be mediated by DNA methyltransferase 3a. *Proc Natl Acad Sci U S A* **97**: 5237-5242.

Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, Shapero MH, Carson AR, Chen W, Cho EK, Dallaire S, Freeman JL, Gonzalez JR, Gratacos M et al. (2006) Global variation in copy number in the human genome. *Nature* **444**: 444-454.

Richardson FC, Richardson KK. (1990) Sequence-dependent formation of alkyl DNA adducts: a review of methods, results, and biological correlates. *Mutat Res* **233**: 127-138.

Ripley LS. (1982) Model for the participation of quasi-palindromic DNA sequences in frameshift mutation. *Proc Natl Acad Sci USA* **79**: 4128-4132.

Roach JC, Glusman G, Smit AF, Huff CD, Hubley R, Shannon PT, Rowen L, Pant KP, Goodman N, Bamshad M, Shendure J, Drmanac R, Jorde LB, Hood L, Galas DJ. (2010) Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science* **328**: 636-639.

Rochette PJ, Lacoste S, Therrien J-P, Bastien N, Brash DE, Drouin R. (2009) Influence of cytosine methylation on ultraviolet-induced cyclobutane pyrimidine dimer formation in genomic DNA. *Mutat Res* **665**: 7-13.

Rosche WA, Trinh TQ, Sinden RR. (1997) Leading strand specific spontaneous mutation corrects a quasipalindrome by an intermolecular strand switch mechanism. *J Mol Biol* **269**: 176-187.

Rosenstein BS, Ducore JM. (1983) Induction of DNA strand breaks in normal human fibroblasts exposed to monochromatic ultraviolet and visible wavelengths in the 240-546 nm range. *Photochem Photobiol* **38**: 51-55.

Rudiger HW. (1991) Clinical, genetic and regulatory consequences of exposure to mutagens. *Ann Genet* **34**: 173-178.

Rydberg B, Lindahl T. (1982) Nonenzymatic methylation of DNA by the intracellular methyl group donor S-adenosyl-L-methionine is a potentially mutagenic reaction. *EMBO J* **1**: 211-216.

Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, Xie X, Byrne EH, McCarroll SA, Gaudet R, Schaffner SF, Lander ES, Consortium IH, Frazer KA, Ballinger DG et al. (2007) Genome-wide detection and characterization of positive selection in human populations. *Nature* **449**: 913-918.

Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, Marth G, Sherry S, Mullikin JC, Mortimore BJ, Willey DL, Hunt SE, Cole CG, Coggill PC, Rice CM, Ning Z et al. (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**: 928-933.

Sakata SF, Shelly LL, Ruppert S, Schutz G, Chou JY. (1993) Cloning and expression of murine S-adenosylmethionine synthetase. *J Biol Chem* **268**: 13978-13986.

Sammalkorpi H, Alhopuro P, Lehtonen R, Tuimala J, Mecklin JP, Jarvinen HJ, Jiricny J, Karhu A, Aaltonen LA. (2007) Background mutation frequency in microsatellite-unstable colorectal cancer. *Cancer Res* **67**: 5691-5698.

Sasaki S, Mello CC, Shimada A, Nakatani Y, Hashimoto S, Ogawa M, Matsushima K, Gu SG, Kasahara M, Ahsan B, Sasaki A, Saito T, Suzuki Y, Sugano S, Kohara Y et al. (2009) Chromatin-associated periodicity in genetic variation downstream of transcriptional start sites. *Science* **323**: 401-404.

Schaefer CB, Ooi SKT, Bestor TH, Bourc'his D. (2007) Epigenetic decisions in mammalian germ cells. *Science* **316**: 398-399.

Schones DE, Cui K, Cuddapah S, Roh TY, Barski A, Wang Z, Wei G, Zhao K. (2008) Dynamic regulation of nucleosome positioning in the human genome. *Cell* **132**: 887-898.

Schones DE, Zhao K. (2008) Genome-wide approaches to studying chromatin modifications. *Nat Rev Genet* **9**: 179-191.

Sebat J, Lakshmi B, Malhotra D, Troge J, Lese-Martin C, Walsh T, Yamrom B, Yoon S, Krasnitz A, Kendall J, Leotta A, Pai D, Zhang R, Lee Y-H, Hicks J et al. (2007) Strong association of de novo copy number mutations with autism. *Science* **316**: 445-449.

Seibert E, Ross JBA, Osman R. (2002) Role of DNA flexibility in sequence-dependent activity of uracil DNA glycosylase. *Biochemistry* **41**: 10976-10984.

Serre D, Hudson TJ. (2006) Resources for genetic variation studies. *Annu Rev Genomics Hum Genet* **7**: 443-457.

Sethupathy P, Giang H, Plotkin JB, Hannenhalli S. (2008) Genome-wide analysis of natural selection on human cis-elements. *PLoS ONE* **3**: e3137.

Shah SN, Hile SE, Eckert KA. (2010) Defective mismatch repair, microsatellite mutation bias, and variability in clinical cancer phenotypes. *Cancer Res* **70**: 431-435.

Shao H, Burrage LC, Sinasac DS, Hill AE, Ernest SR, O'Brien W, Courtland HW, Jepsen KJ, Kirby A, Kulbokas EJ, Daly MJ, Broman KW, Lander ES, Nadeau JH. (2008) Genetic architecture of complex traits: large phenotypic effects and pervasive epistasis. *Proc Natl Acad Sci U S A* **105**: 19910-19914.

Sharp AJ, Cheng Z, Eichler EE. (2006) Structural variation of the human genome. *Annu Rev Genomics Hum Genet* **7**: 407-442.

She X, Liu G, Ventura M, Zhao S, Misceo D, Roberto R, Cardone MF, Rocchi M, Green ED, Archidiacano N, Eichler EE. (2006) A preliminary comparative analysis of primate segmental duplications shows elevated substitution rates and a great-ape expansion of intrachromosomal duplications. *Genome Res* **16**: 576-583.

Shen JC, Rideout WM, 3rd, Jones PA. (1994) The rate of hydrolytic deamination of 5-methylcytosine in double-stranded DNA. *Nucleic Acids Res* **22**: 972-976.

Shendure J, Ji H. (2008) Next-generation DNA sequencing. *Nat Biotechnol* **26**: 1135-1145.

Sherry ST, Ward M, Sirotkin K. (1999) dbSNP-database for single nucleotide polymorphisms and other classes of minor genetic variation. *Genome Res* **9**: 677-679.

Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* **29**: 308-311.

Shibutani S, Takeshita M, Grollman AP. (1991) Insertion of specific bases during DNA synthesis past the oxidation-damaged base 8-oxodG. *Nature* **349**: 431-434.

Sibghat-Ullah, Gallinari P, Xu YZ, Goodman MF, Bloom LB, Jiricny J, Day RS. (1996) Base analog and neighboring base effects on substrate specificity of recombinant human G:T mismatch-specific thymine DNA-glycosylase. *Biochemistry* **35**: 12926-12932.

Sigurdsson MI, Smith AV, Bjornsson HT, Jonsson JJ. (2009) HapMap methylation-associated SNPs, markers of germline DNA methylation, positively correlate with regional levels of human meiotic recombination. *Genome Res* **19**: 581-589.

Simonsson T. (2001) G-quadruplex DNA structures--variations on a theme. *Biol Chem* **382**: 621-628.

Singer B. (1975) The chemical effects of nucleic acid alkylation and their relation to mutagenesis and carcinogenesis. *Prog Nucleic Acid Res Mol Biol* **15**: 219-284.

Singer B, Kusmierek JT. (1982) Chemical mutagenesis. *Annu Rev Biochem* **51**: 655-693.

Siva N. (2008) 1000 Genomes project. *Nat Biotechnol* **26**: 256.

Smela ME, Currier SS, Bailey EA, Essigmann JM. (2001) The chemistry and biology of aflatoxin B(1): from mutational spectrometry to carcinogenesis. *Carcinogenesis* **22**: 535-545.

Spencer CCA. (2006) Human polymorphism around recombination hotspots. *Biochem Soc Trans* **34**: 535-536.

Spencer CCA, Deloukas P, Hunt S, Mullikin JC, Myers S, Silverman B, Donnelly P, Bentley DR, McVean G. (2006) The influence of recombination on human genetic diversity. *PLoS Genet* **2**: e148.

Stamatoyannopoulos JA, Adzhubei I, Thurman RE, Kryukov GV, Mirkin SM, Sunyaev SR. (2009) Human mutation rate associated with DNA replication timing. *Nat Genet* **41**: 393-395.

Stefansson H, Helgason A, Thorleifsson G, Steinthorsdottir V, Masson G, Barnard J, Baker A, Jonasdottir A, Ingason A, Gudnadottir VG, Desnica N, Hicks A, Gylfason A, Gudbjartsson DF, Jonsdottir GM et al. (2005) A common inversion under selection in Europeans. *Nat Genet* **37**: 129-137.

Stephens PJ, McBride DJ, Lin M-L, Varela I, Pleasance ED, Simpson JT, Stebbings LA, Leroy C, Edkins S, Mudie LJ, Greenman CD, Jia M, Latimer C, Teague JW, Lau KW et al. (2009) Complex landscapes of somatic rearrangement in human breast cancer genomes. *Nature* **462**: 1005-1010.

Stoltzfus A. (2008) Evidence for a predominant role of oxidative damage in germline mutation in mammals. *Mutat Res* **644**: 71-73.

Strathern JN, Shafer BK, McGill CB. (1995) DNA synthesis errors associated with double-strand-break repair. *Genetics* **140**: 965-972.

Stratton MR, Campbell PJ, Futreal PA. (2009) The cancer genome. *Nature* **458**: 719-724.

Streisinger G, Okada Y, Emrich J, Newton J, Tsugita A, Terzaghi E, Inouye M. (1966) Frameshift mutations and the genetic code. This paper is dedicated to Professor Theodosius Dobzhansky on the occasion of his 66th birthday. *Cold Spring Harb Symp Quant Biol* **31**: 77-84.

Subramanian S, Mishra RK, Singh L. (2003) Genome-wide analysis of microsatellite repeats in humans: their abundance and density in specific genomic regions. *Genome Biol* **4**: R13.

Sundquist WI, Klug A. (1989) Telomeric DNA dimerizes by formation of guanine tetrads between hairpin loops. *Nature* **342**: 825-829.

Suter B, Thoma F. (2002) DNA-repair by photolyase reveals dynamic properties of nucleosome positioning in vivo. *J Mol Biol* **319**: 395-406.

Sutton WS. (1903) The chromosomes in heredity. *Biol. Bull. mar. biol. Lab.* **4**: 231-248.

Suzuki MM, Bird A. (2008) DNA methylation landscapes: provocative insights from epigenomics. *Nat Rev Genet* **9**: 465-476.

Tammi MT, Arner E, Kindlund E, Andersson B. (2003) Correcting errors in shotgun sequences. *Nucleic Acids Res* **31**: 4663-4672.

Taylor J, Tyekucheva S, Zody M, Chiaromonte F, Makova KD. (2006) Strong and weak male mutation bias at different sites in the primate genomes: insights from the human-chimpanzee comparison. *Mol Biol Evol* **23**: 565-573.

Tessmer I, Yang Y, Zhai J, Du C, Hsieh P, Hingorani MM, Erie DA. (2008) Mechanism of MutS searching for DNA mismatches and signaling repair. *J Biol Chem* **283**: 36646-36654.

Thomas DJ, Trumbower H, Kern AD, Rhead BL, Kuhn RM, Haussler D, Kent WJ. (2007) Variation resources at UC Santa Cruz. *Nucleic Acids Res* **35**: D716-720.

Tommasi S, Denissenko MF, Pfeifer GP. (1997) Sunlight induces pyrimidine dimers preferentially at 5-methylcytosine bases. *Cancer Res* **57**: 4727-4730.

Tomso DJ, Bell DA. (2003) Sequence context at human single nucleotide polymorphisms: overrepresentation of CpG dinucleotide at polymorphic sites and suppression of variation in CpG islands. *J Mol Biol* **327**: 303-308.

Trasler JM. (2009) Epigenetics in spermatogenesis. *Mol Cell Endocrinol* **306**: 33-36.

Tschermak EV. (1900) Über künstliche Kreuzung bei *Pisum sativum*. *Ber. dt. bot. Ges.* **18**: 232-239.

van Beneden E. (1883) Recherches sur la matruation de l'oeuf, le fécondation. *Archs Biol., Paris* **4**: 265-638.

van Noort V, Worning P, Ussery DW, Rosche WA, Sinden RR. (2003) Strand misalignments lead to quasipalindrome correction. *Trends Genet* **19**: 365-369.

Venter JC, Adams MD, Li PW, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH et al. (2001) The sequence of the human genome. *Science* **291**: 1304-1351.

Voight BF, Kudaravalli S, Wen X, Pritchard JK. (2006) A map of recent positive selection in the human genome. *PLoS Biol* **4**: e72.

Walker FO. (2007) Huntington's disease. *Lancet* **369**: 218-228.

Walser JC, Furano AV. (2010) The mutational spectrum of non-CpG DNA varies with CpG content. *Genome Res*.

Walsh CP, Xu GL. (2006) Cytosine methylation and DNA repair. *Curr Top Microbiol Immunol* **301**: 283-315.

Wang D, Johnson AD, Papp AC, Kroetz DL, Sadee W. (2005) Multidrug resistance polypeptide 1 (MDR1, ABCB1) variant 3435C>T affects mRNA stability. *Pharmacogenet Genomics* **15**: 693-704.

Wang H, Yang Y, Schofield MJ, Du C, Fridman Y, Lee SD, Larson ED, Drummond JT, Alani E, Hsieh P, Erie DA. (2003) DNA bending and unbending by MutS govern mismatch recognition and specificity. *Proc Natl Acad Sci USA* **100**: 14822-14827.

Wang J, Pitarque M, Ingelman-Sundberg M. (2006) 3'-UTR polymorphism in the human CYP2A6 gene affects mRNA stability and enzyme expression. *Biochem Biophys Res Commun* **340**: 491-497.

Wang Z, Moult J. (2001) SNPs, protein structure, and disease. *Hum Mutat* **17**: 263-270.

Ward JF. (1985) Biochemistry of DNA lesions. *Radiat Res Suppl* **8**: S103-111.

Ward JF. (1988) DNA damage produced by ionizing radiation in mammalian cells: identities, mechanisms of formation, and reparability. *Prog Nucleic Acid Res Mol Biol* **35**: 95-125.

Ward JF. (1990) The yield of DNA double-strand breaks produced intracellularly by ionizing radiation: a review. *Int J Radiat Biol* **57**: 1141-1150.

Warnecke T, Batada NN, Hurst LD. (2008) The impact of the nucleosome code on protein-coding sequence evolution in yeast. *PLoS Genet* **4**: e1000250.

Washietl S, Machne R, Goldman N. (2008) Evolutionary footprints of nucleosome positions in yeast. *Trends Genet* **24**: 583-587.

Watson J, Crick F. (1953) Molecular structure of nucleic acids. *Nature* **171**: 737-738.

Webster MT, Smith NGC, Ellegren H. (2002) Microsatellite evolution inferred from human-chimpanzee genomic sequence alignments. *Proc Natl Acad Sci USA* **99**: 8748-8753.

Weissman D, Schmidt SC, Kakol JM, Stein LD, Sherry ST, Mortimore BJ, Willey DL, Hunt S, Cole CG, Coggill PC, Rice CM, Ning Z, Rogers J, Bentley DR, Kwok P-Y et al. (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**: 928-933.

Wellcome Trust Case Control Consortium. (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**: 661-678.

Whitehouse HLK. (1973) *Towards an Understanding of the Mechanism of Heredity*. Edward Arnold Publishers, Ltd., London.

Widlak P, Pietrowska M, Lanuszewska J. (2006) The role of chromatin proteins in DNA damage recognition and repair. *Histochem Cell Biol* **125**: 119-126.

Witt KL, Bishop JB. (1996) Mutagenicity of anticancer drugs in mammalian germ cells. *Mutat Res* **355**: 209-234.

Wyszynski M, Gabbara S, Bhagwat AS. (1994) Cytosine deaminations catalyzed by DNA cytosine methyltransferases are unlikely to be the major cause of mutational hot spots at sites of cytosine methylation in Escherichia coli. *Proc Natl Acad Sci U S A* **91**: 1574-1578.

Xu G, Spivak G, Mitchell DL, Mori T, McCarrey JR, McMahan CA, Walter RB, Hanawalt PC, Walter CA. (2005) Nucleotide excision repair activity varies among murine spermatogenic cell types. *Biol Reprod* **73**: 123-130.

Yao X, Buermeyer AB, Narayanan L, Tran D, Baker SM, Prolla TA, Glazer PM, Liskay RM, Arnheim N. (1999) Different mutator phenotypes in Mlh1- versus Pms2-deficient mice. *Proc Natl Acad Sci U S A* **96**: 6850-6855.

Ye N, Holmquist GP, O'Connor TR. (1998) Heterogeneous repair of N-methylpurines at the nucleotide level in normal human cells. *J Mol Biol* **284**: 269-285.

Yeo GS, Farooqi IS, Aminian S, Halsall DJ, Stanhope RG, O'Rahilly S. (1998) A frameshift mutation in MC4R associated with dominantly inherited human obesity. *Nat Genet* **20**: 111-112.

Ying H, Epps J, Williams R, Huttley G. Evidence that localized variation in primate sequence divergence arises from an influence of nucleosome placement on DNA repair. *Mol Biol Evol* **27**: 637-649.

Yu K, Herr AB, Waksman G, Ornitz DM. (2000) Loss of fibroblast growth factor receptor 2 ligand-binding specificity in Apert syndrome. *Proc Natl Acad Sci U S A* **97**: 14536-14541.

Yung C, Suzuki T, Okugawa Y, Kawakami A, Loakes D, Negishi K, Negishi T. (2007) Nucleotide incorporation against 7,8-dihydro-8-oxoguanine is influenced by neighboring base sequences in TLS DNA polymerase reaction. *Nucleic Acids Symp Ser (Oxf)*: 49-50.

Yung C-W, Okugawa Y, Otsuka C, Okamoto K, Arimoto S, Loakes D, Negishi K, Negishi T. (2008) Influence of neighbouring base sequences on the mutagenesis induced by 7,8-dihydro-8-oxoguanine in yeast. *Mutagenesis* **23**: 509-513.

Zhang F, Gu W, Hurles ME, Lupski JR. (2009) Copy number variation in human health, disease, and evolution. *Annu Rev Genomics Hum Genet* **10**: 451-481.

Zhang J, Feuk L, Duggan GE, Khaja R, Scherer SW. (2006) Development of bioinformatics resources for display and analysis of copy number and other structural variants in the human genome. *Cytogenet Genome Res* **115**: 205-214.

Zhao Y, Du Z, Li N. (2007) Extensive selection for the enrichment of G4 DNA motifs in transcriptional regulatory regions of warm blooded animals. *FEBS Lett* **581**: 1951-1956.

Zhao Z, Boerwinkle E. (2002) Neighboring-nucleotide effects on single nucleotide polymorphisms: a study of 2.6 million polymorphisms across the human genome. *Genome Res* **12**: 1679-1686.

Zhao Z, Jiang C. (2007) Methylation-dependent transition rates are dependent on local sequence lengths and genomic regions. *Mol Biol Evol* **24**: 23-25.

Zhao Z, Zhang F. (2006) Sequence context analysis of 8.2 million single nucleotide polymorphisms in the human genome. *Gene* **366**: 316-324.

I

BMC Genomics

BioMed Central

Research article

Open Access

Large-scale inference of the point mutational spectrum in human segmental duplications

Sigve Nakken[1], Einar A Rødland[2], Torbjørn Rognes[1,2] and Eivind Hovig*[2,3,4]

Address: [1]Centre for Molecular Biology and Neuroscience, Institute of Medical Microbiology, Rikshospitalet University Hospital, NO-0027 Oslo, Norway, [2]Department of Informatics, University of Oslo, PO Box 1080 Blindern, NO-0316 Oslo, Norway, [3]Department of Tumor Biology, Institute for Cancer Research, Norwegian Radium Hospital, Montebello, 0310 Oslo, Norway and [4]Department of Medical Informatics, Norwegian Radium Hospital, Montebello, 0310 Oslo, Norway

Email: Sigve Nakken - sigve.nakken@medisin.uio.no; Einar A Rødland - einarro@ifi.uio.no; Torbjørn Rognes - torbjorn.rognes@rr-research.no; Eivind Hovig* - ehovig@radium.uio.no

* Corresponding author

Published: 22 January 2009

BMC Genomics 2009, **10**:43 doi:10.1186/1471-2164-10-43

Received: 2 September 2008
Accepted: 22 January 2009

This article is available from: http://www.biomedcentral.com/1471-2164/10/43

Abstract

Background: Recent segmental duplications are relatively large (≥ 1 kb) genomic regions of high sequence identity (≥ 90%). They cover approximately 4–5% of the human genome and play important roles in gene evolution and genomic disease. The DNA sequence differences between copies of a segmental duplication represent the result of various mutational events over time, since any two duplication copies originated from the same ancestral DNA sequence. Based on this fact, we have developed a computational scheme for inference of point mutational events in human segmental duplications, which we collectively term duplication-inferred mutations (DIMs). We have characterized these nucleotide substitutions by comparing them with high-quality SNPs from dbSNP, both in terms of sequence context and frequency of substitution types.

Results: Overall, DIMs show a lower ratio of transitions relative to transversions than SNPs, although this ratio approaches that of SNPs when considering DIMs within most recent duplications. Our findings indicate that DIMs and SNPs in general are caused by similar mutational mechanisms, with some deviances at the CpG dinucleotide. Furthermore, we discover a large number of reference SNPs that coincide with computationally inferred DIMs. The latter reflects how sequence variation in duplicated sequences can be misinterpreted as ordinary allelic variation.

Conclusion: In summary, we show how DNA sequence analysis of segmental duplications can provide a genome-wide mutational spectrum that mirrors recent genome evolution. The inferred set of nucleotide substitutions represents a valuable complement to SNPs for the analysis of genetic variation and point mutagenesis.

Background

Single point mutations represent a fundamental driving force for the evolution of any vertebrate genome. Mutations create DNA sequence variation that may alter gene function as well as DNA conformation and protein binding [1,2]. The spectrum of nucleotide substitutions occurring in human DNA sequences is the result of actions of various mutational sources of both endogenous and exogenous origin. An increasing body of evidence supports the idea that the majority of mutations are generated by error-

prone intracellular processes that operate in a DNA sequence-dependent manner [3,4]. Examples of endogenous mutagenic processes are DNA replication (i.e. polymerase fidelity and replication slippage), post-replicative DNA mismatch repair and methylation-mediated deamination of cytosines in CpG dinucleotides [5-9]. The sequence dependence of these processes is reflected in a biased distribution of point mutations and their sequence neighbourhoods, as shown by previous analyses of pseudogene mutations, germline disease mutations and single nucleotide polymorphisms (SNPs) [3,10-13]. The nature of the observed point mutational bias is by far dominated by the hypermutability of the CpG dinucleotide [14,15]. The extent of CpG depletion in mammalian DNA attributable by methylation-mediated mutation (i.e. $5^{m}C \rightarrow T$) is however a matter of debate [16-21]. The deficiency of CpG seen in unmethylated vertebrate DNA viruses and observations that CpG sequences are favored targets for specific exogenous mutagens suggest that other mutational and selectional mechanisms might contribute to CpG depletion [22-24]. With the exception of CpG mutations, linking the observed non-randomness of human mutations to known sequence-dependent mutational mechanisms remains challenging.

So far, large-scale genome-wide analyses of the DNA context of point mutational events have relied on either disease-causing mutations or SNP data from NCBI's dbSNP [25]. As of February 2007, dbSNP contains more than 9 million polymorphic (biallelic only) positions in the human genome. However, studies have shown that a substantial fraction of entries in dbSNP have been erroneously submitted (e.g. as a result of DNA sequencing errors), and are most likely monomorphic alleles in human populations [26,27]. Comprehensive computational analyses of SNPs may thus easily get corrupted unless a careful discrimination between validated and non-validated entries in dbSNP is undertaken.

A valuable source of information on vertebrate point mutagenesis that to our knowledge has not been thoroughly investigated is contained within human segmental duplications. Recent segmental duplications are large (≥ 1 kb) regions of high sequence identity (≥ 90%) that constitute all types of genomic elements, such as high-copy repeats and gene sequences with exon-intron structures [28-31]. Approximately 4–5% of the human genome is covered with recent duplications, being enriched in pericentromeric and subtelomeric regions of the chromosomes [32-34]. Owing to their high degree of sequence identity, a large number of mutational events can be inferred with high confidence using only pairwise DNA sequence alignments. Knowing that duplications were once identical during evolution, point mutational events correspond to mismatches in the aligned sequences. This simple approach is thus powerful for detection of a mutational spectrum in recent mammalian evolution. A proper classification of the allelic fate of newly derived alleles in segmental duplications is a different matter, however. An allele created by a point mutation in one duplication copy may be subject to a number of genetic processes that determines its allelic state in duplicated DNA. Allelic drift can take the newly derived allele through a polymorphic state (that is, SNP in a duplication) and further to fixation, in which the new allele and its counterpart in the other duplication copy are termed paralogous sequence variants (PSVs) [29,35-37]. At the same time, the newly derived allele can be distributed into multiple sequence copies by duplication or gene conversion [38-40]. The latter mechanisms take the initial mutational event into a complex type of sequence variation coined multisite variation (MSV) by Fredman and colleagues [41]. Mutational events in segmental duplications thus result in a mosaic of different genotype patterns. Altogether, data on duplication-inferred mutations generated by our approach both enriches the available pool of known mutational events within recent mammalian evolution and complements the data on disease mutations and SNPs for a contextual DNA sequence analysis of single nucleotide substitutions in humans.

We have developed a computational pipeline for inference of mutational events in segmental duplications in the human genome. The analysis of duplication-inferred mutations (DIMs) was restricted to intergenic regions of duplications, focusing on the mutational spectrum in regions that are believed to be more neutral with respect to selection forces. With the aim of detecting mutational hotspots of DIMs, we conducted a computational analysis of the local DNA sequence context of DIMs. A comparative analysis with a large set of high-quality, intergenic SNPs from dbSNP provides insights into similarities and differences between duplication-inferred variation and ordinary allelic variation in unique regions of the genome. We have also investigated the overlap between reference SNPs in segmental duplications and computationally, duplication-inferred variants. Initial reports concerning the high density of SNPs in duplications suggested that this was due to paralogous variation being misinterpreted as SNPs [28,36,37]. A following experimental study of a limited set of SNPs in duplications found that only 23% of the SNPs were consistent with paralogous variation [41], and that multisite variation appeared to be a common type of variation in these regions. We used a computational, *in silico* approach for the discovery of positional and allelic overlap between SNPs and DIMs. Our data pinpoints a large number of recorded SNPs in segmental duplications that mimic variation between paralogous sequences, and these may consequently give rise to strange patterns during traditional SNP genotyping.

Results

Distribution of nucleotide substitutions

A total of 343,864 human duplication-inferred mutations from intergenic regions of segmental duplications satisfied the criteria we established for reliable DIM inference in DNA duplicon sequence alignments (see Figure 1 and Methods). These DIMs were subject to a comparative analysis with 1,115,692 intergenic HapMap-validated SNPs in non-duplicated regions of the human genome. The nucleotide composition of the two regions in which substitutions originated displayed a difference in GC content. Overall, intergenic regions of segmental duplications had a GC content of 41.7%, while the corresponding regions of non-duplicated DNA contained 39.6% (χ = 42,336, df = 1, $p < 0.00001$). When considering GC content in duplications of different levels of sequence identity, we observed a higher content at all levels (Figure 2). Figure 3 illustrates the distribution of substitution types and how the proportions of SNPs compared with inferred DIMs. Here, each type of substitution combines the nucleotide change for both directions (e.g. A/G represents the sum of all A→G and G→A substitutions) because the directions of SNPs and DIMs in our dataset are generally unknown. The histogram shows that the proportions of A/C and G/T, as well as A/G and C/T, were close to identical for SNPs. This observation reflects the complementary strand symmetry in DNA sequences as reported in previous studies on SNPs [13]. Similarly, equal proportions of complementary substitutions were observed for DIMs.

As Figure 3 shows, overall DIMs and SNPs shared similar characteristics in terms of the distribution of substitution types. The two transition substitutions, A/G and C/T, account for approximately two-thirds of all substitutions for both SNPs and DIMs. Among transversions, we discovered that DIMs increased most relative to SNPs for substitutions between C and G (1.37%). The observed differences between SNPs and DIMs in terms of transition bias were noteworthy. We found that DIMs display a much smaller overall ratio of transitions over transversions than SNPs (2.11 for SNPs vs. 1.70 for all DIMs, χ = 576.7, df = 1, $p < 0.00001$). Estimating the transition bias

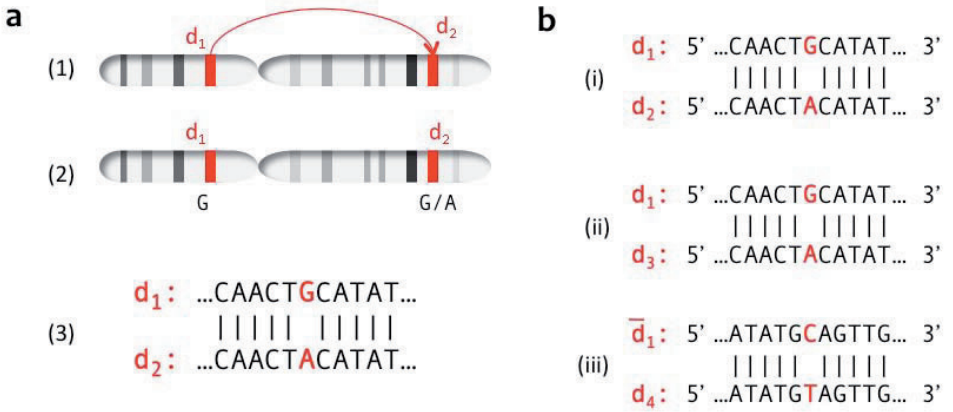


Figure 1

Evolution of segmental duplications and principles for duplication-inferred mutations. **A:** (1) An intrachromosomal duplication event occurs during evolution, followed by (2) a mutation in one of the duplication copies, causing a G/A transition. (3) A historical mutation event is inferred from an alignment of the present-day duplication copy sequences. **B:** Sequences d_1, d_2, d_3 and d_4 are all segmental duplication copies of the same ancestral DNA sequence. (i) A single base mismatch observed in the alignment between duplication copies d_1 and d_2 lead to the inference of a G/A (A/G) point mutational event within this DNA sequence context. (ii) An identical mismatch in the same position as observed between d_1 and d_2 was observed between d_1 and d_3. Such instances were not recorded twice in the set of mutational events, as the mismatch most likely is a propagation of the result in (i). (iii) A C/T (T/C) base mismatch in the same position as observed between d_1 and d_2 was observed in the alignment of d_1 (reverse strand) and d_4. Since the complementary mutation has been recorded in (i), we did not record this mismatch as a mutational event, as it most likely was the result of propagation by duplication.

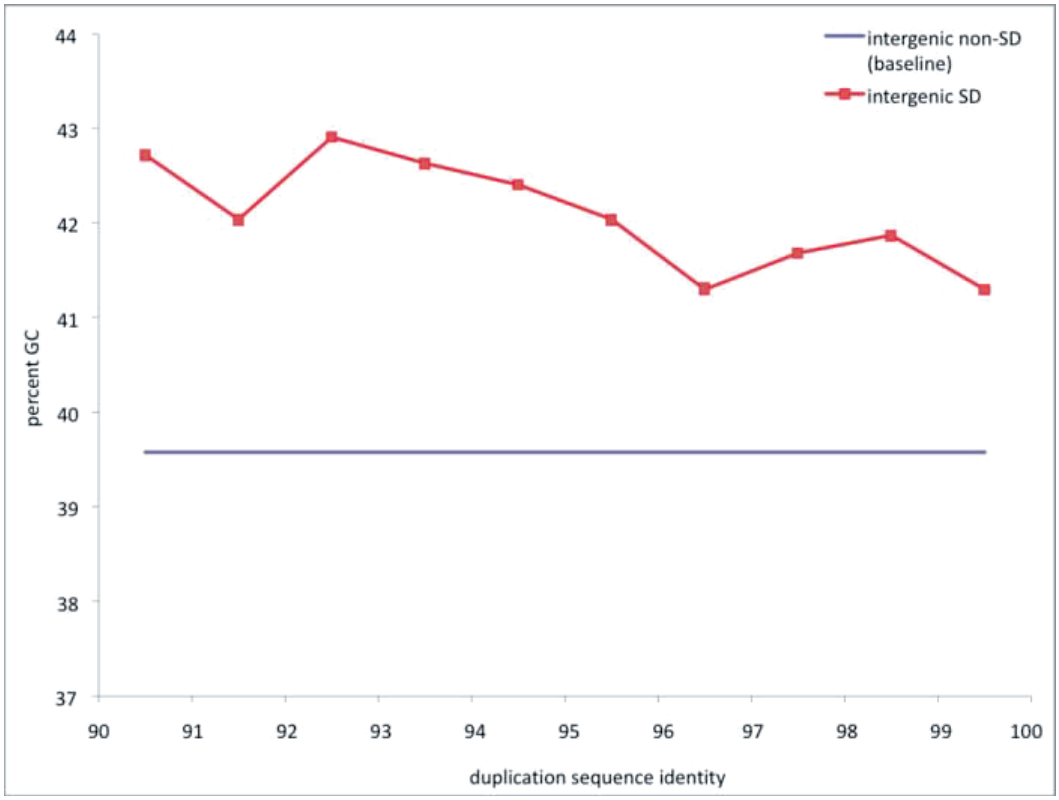


Figure 2
GC content in segmental duplications compared to non-duplicated genomic regions. GC content in intergenic regions of human segmental duplications (intergenic SD) at different levels of sequence divergence. The average GC content in non-duplicated regions is also drawn (intergenic non-SD).

with our approach ignores a potential substitution rate variation among sites, and may thus underestimate the extent of the bias. Henceforth, we emphasize the observed difference we found for SNPs and DIMs rather than the bias in itself. Furthermore, when considering the ratio of DIM transitions over DIM transversions at different levels of duplication sequence divergence, we observed a trend in which the transition bias increased when duplication divergence decreased (Figure 4). In other words, we found that DNA sequences from duplication events in most recent time had the highest degree of transition bias. To gain further insight, we established a subset of the DIMs in which transitions at the CpG dinucleotide were excluded. It is generally accepted that the CpG dinucleotide mutates at a high rate in the human genome due to deamination of 5-methylcytosine (5mC) to thymine, although this phenomenon has not yet been shown within the context of paralogous sequence variation. When excluding transitions at CpG dinucleotides, the trend towards increased transition bias in recent duplications was not as evident as when considering all DIMs (Figure 4).

We next determined the overall distribution of substitutions within the context of CpG dinucleotides (Table 1). Substitution frequencies were obtained in two different regions of the genome; important regulatory regions clustered with unmethylated CpGs known as CpG islands, and regions outside CpG islands. The density of CpG islands was higher in segmental duplications (1.11%) than in nonduplicated regions (0.87%, $\chi = 69{,}257$, df = 1, $p < 0.00001$). As indicated in Table 1, the fraction of

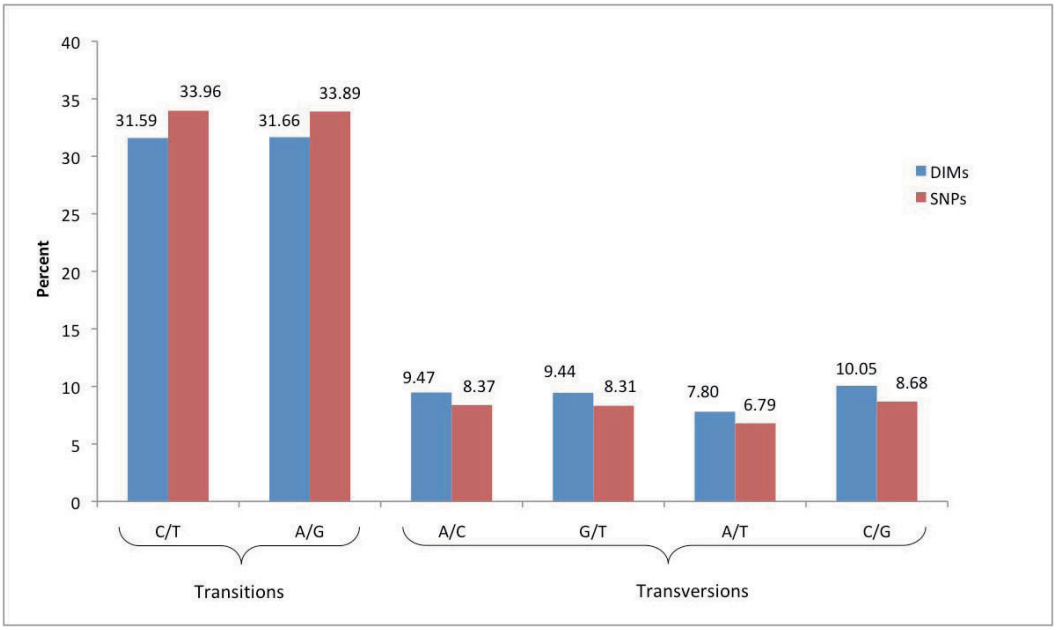


Figure 3
Distribution of substitution types for DIMs and SNPs in intergenic regions. Distribution of substitution types for 1,115,692 high-quality, genome-wide intergenic SNPs and 343,864 intergenic DIMs inferred from all human segmental duplications with sequence identity ≥ 90%. Substitution types do not carry direction (i.e. the fraction of A/G substitutions is the sum of A→G substitutions and G→A substitutions).

methylation-related transitions at CpG dinucleotides was much higher outside of CpG islands than within CpG islands. This was evident both in segmental duplications (DIMs) and non-duplicated regions (SNPs).

Any observed difference between high-quality SNPs and DIMs could potentially be a result of ascertainment biases between the two mutation sets. The SNP set was ascertained using HapMap allele frequencies, excluding potential false positives originating from DNA sequencing errors. The DIM set was established using sequence alignments only, thus there is a greater chance that DIMs contain false positive mutations arising from either alignment artefacts or sequencing errors in duplications. We assessed the potential impact of noise among the duplication-inferred mutations in two different ways. First, we established two subsets of DIMs using different alignment criteria for DIM calling. The estimated, overall transition to transversion ratios in these two sets were 1.72 and 1.73. Second, we looked at three different sequence contexts that account for many false positive SNPs arising from DNA sequencing errors [27]. Having excluded all false positive SNPs in our high-quality set, we assume that the fractions of these sequence contexts in the SNP set resembles expected numbers in a human mutational spectrum. Compared with high-quality SNPs, the fraction of DIMs that occurred in these error-prone sequencing contexts increased with approximately 0.1–0.4% (Table 2).

Sequence contexts of DIMs

We obtained DNA oligomer frequencies at SNPs and DIMs and in their corresponding reference regions to address whether both types of mutations were subject to similar mutational hotspots. Under the assumption that the middle nucleotide of an odd-length oligomer is independent of its surrounding sequence, we computed expected numbers for all oligomers. Finally, we compared the actual number for each oligomer with its expected number, defined as overrepresentation (see Methods).

Figure 5 compares the overrepresentation of DNA oligomers of length five (five-mers) at DIMs and SNPs. The plot illustrates similar levels of overrepresentation for the majority of five-mers at DIMs and SNPs. Oligomers where

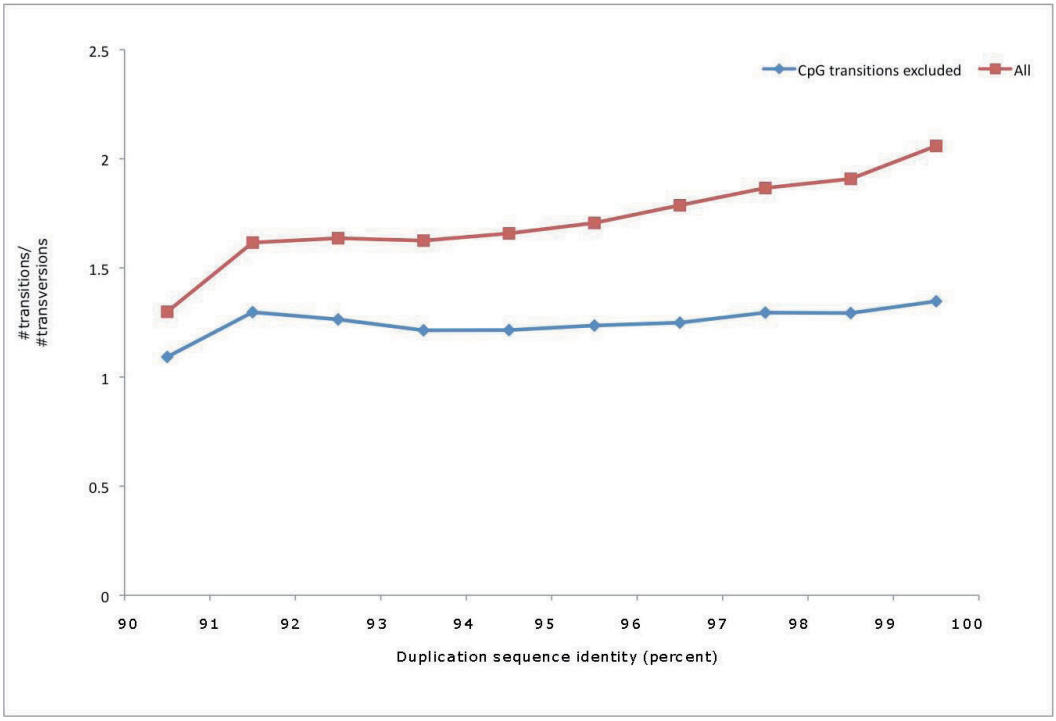


Figure 4
Transitions to transversions ratio among duplication-inferred mutations. The ratio of transitions to transversions among inferred mutations in segmental duplications at different levels of sequence divergence. The uppermost line in the plot illustrates transitions to transversions ratios when all inferred mutations were included. The lowermost line show computed ratios when transitions at the CpG dinucleotide were excluded.

substitutions occur within the CpG dinucleotide (CpG at center) did not distribute quite as evenly between DIMs and SNPs, however. In the majority of these oligomers, SNPs were slightly overrepresented. The opposite was observed for DIMs, which occurred less frequently than expected in most of these sequences. Oligomers where substitutions take place before or after a CpG dinucleotide (CpG in surroundings) did not show any notable differences in abundance levels between SNPs and DIM.

We next compared the distribution of five-mers in the reference regions of SNPs and DIMs, illustrated in Figure 6. The figure shows that the set of five-mers containing no CpG, as well five-mers with CpG in surroundings were roughly equally abundant in intergenic regions of segmental duplications as in intergenic, nonduplicated regions of the genome. These five-mers were distributed close to the identity line. Five-mers affected by the CpG effect was however strongly underrepresented in both duplicated regions and non-duplicated regions. Furthermore, the degree of underrepresentation of these oligomers was slightly stronger in the regions where SNPs originate compared to the regions where DIMs originate.

Overlap between SNPs and DIMs in segmental duplications

Previous analyses of SNPs in segmental duplications have reported an uncertainty about the validity of this particular set of SNPs [28,36,37]. The observed SNP enrichment was initially viewed as duplication-induced, representing paralogous rather than allelic variation. More advanced techniques have later shown that the spectrum of sequence variation in duplications appears as a complex combination of PSVs, SNPs in duplications and MSVs

Table 1: Substitution frequencies at the CpG dinucleotide context

Substitution context	*SNPs*	*DIMs*
Non-CpG island: (A/C)G	4.69 (14,399/307,128)	5.73 (4,320/75,433)
Non-CpG island: (C/G)G	4.65 (14,270/307,128)	6.28 (4,665/75,433)
Non-CpG island: (C/T)G	**40.71** (125,040/307,128)	**37.69** (28,589/75,433)
Non-CpG island: C(A/G)	**40.62** (124,746/307,128)	**37.91** (28,713/75,433)
Non-CpG island: C(C/G)	4.58 (14,078/307,128)	4.58 (4,765/75,433)
Non-CpG island: C(G/T	4.75 (14,595/307,128)	4.75 (4,381/75,433)
CpG island: (A/C)G	9.80 (206/2,103)	9.02 (304/3,371)
CpG island: (C/G)G	13.03 (274/2,103)	14.74 (497/3,371)
CpG island: (C/T)G	**27.29** (574/2,103)	**26.49** (893/3,371)
CpG island: C(A/G)	**26.82** (564/2,103)	**26.76** (902/3,371)
CpG island: C(C/G)	13.41 (282/2,103)	14.06 (474/3,371)
CpG island: C(G/T)	9.65 (203/2,103)	8.93 (301/3,371)

Nucleotide substitution percentages at the CpG dinucleotide context are shown for intergenic DIMs and SNPs, within and outside CpG islands. The percentages of substitutions are shown along with raw counts in parentheses. Differences between islands and non-island regions for methylation-related substitutions (in boldface) are statistically significant ($p < 0.00001$) by Chi-square analysis.

[41]. In this work, we have quantified the number of inferred mutational events in segmental duplications that overlap with reference SNPs in segmental duplications.

We retrieved a total of 458,811 SNPs from dbSNP that mapped within intergenic regions of segmental duplications. Of these SNPs, 301,968 (65.8%) were non-validated. The remaining 156,843 (34.2%) SNPs had been validated according to different criteria (see Methods). In comparison with the complementary, non-duplicated regions of the genome, which contained 31.3% nonvalidated SNPs, segmental duplications were significantly enriched for non-validated SNPs ($\chi = 24{,}952$, df = 1, $p < 0.00001$). We then established a procedure to test whether SNPs in intergenic regions of segmental duplications coincided with DIMs. The procedure matched SNP and DIM alleles at chromosomal positions where SNPs had been identified and inferred DIMs had been recorded. Overall, we found that the chromosomal positions of 83,987 SNPs matched either a target or a source position of our inferred DIMs. Among these 83,987 SNPs, the alleles of 80,856 (96.3%) SNPs matched perfectly with corresponding DIM bases. Thus, we discovered that 17.6% of all reported intergenic SNPs in segmental duplications (80,856 SNPs out of total of 458,811) mirror sequence variation found among inferred DIMs. Although the majority of the 80,856 SNPs that overlapped with DIMs were non-validated (56,425 SNPs, 69.8%), our findings also revealed a substantial fraction of DIM overlap for validated SNPs (24,431 SNPs, 30.2%). These observations suggest that many reference SNPs in duplications most likely represent paralogous sequence variation, induced by signals from paralogous sequences in the genome. The subset of DIM-overlapping SNPs was inferred from segmental duplications that displayed a distribution dominated by duplications with 97–100% DNA sequence identity (Figure 7). All SNP entries that we found to coincide with mutational events in segmental duplications are available as supplementary material http://snp.uio.no/dim/.

Fredman et al. conducted an experimental study in which they genotyped predicted SNPs in segmental duplications from fully homozygous genomes of complete hydatidiform moles (CHMs) [41]. They discovered that only 23% gave patterns indicative of PSVs, and 28% behaved differently than SNPs and PSVs, being the sum of individual genotyping signals from similar-sequence duplication copies. They termed the latter category multisite variants (MSV). Among 105 SNPs being targeted in their study, 64 SNPs mapped to the intergenic regions of duplications used in our analysis, of which they experimentally verified 11 as PSVs. We observed all 11 variants among our computationally inferred DIMs. An additional overlap was

Table 2: Substitution frequencies at sequence contexts associated with DNA sequencing errors

Substitution context	*SNPs*	*DIMs*
A(G/H)N	14.24 (158,901/1,115,692)	14.62 (50,277/343,864)
C(A/Y)C	0.38 (4,250/1,115,692)	0.45 (1,542/343,864)
G(A/C)C	1.14 (12,697/1,115,692)	1.42 (4,866/343,864)

A comparison of nucleotide substitution percentages in DIMs and high-quality SNPs at three sequence contexts previously shown to be overrepresented in false positive SNPs [27]. The percentages of substitutions are shown along with raw counts in parentheses. H stands for A, C or T, Y stands for C or T and N stands for any base.

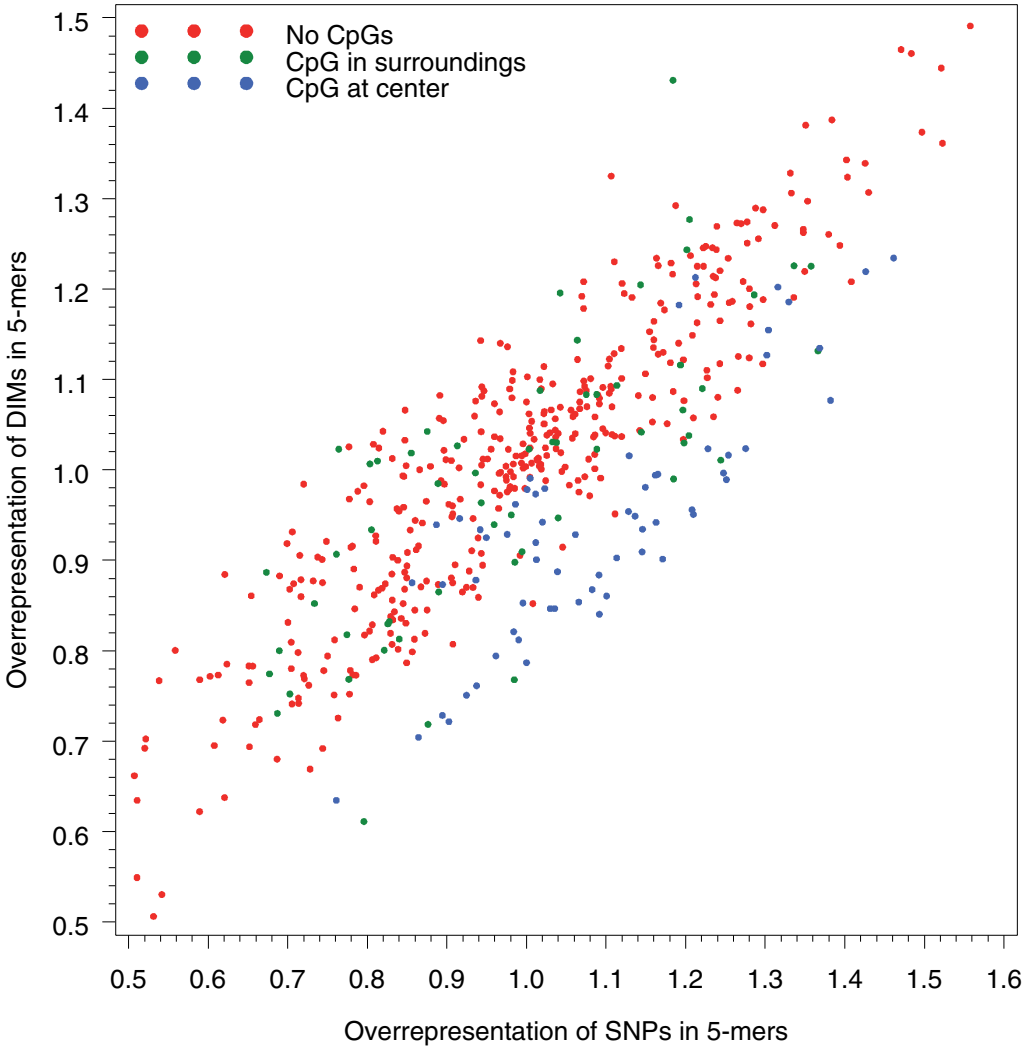


Figure 5
Overrepresentation of DNA oligomers (five-mers) at sites of SNPs and recorded DIMs. The plot compares the abundance of five-mers at substitution sites between SNPs and DIMs. Overrepresentation for a given five-mer is defined as the ratio between the number of observed five-mers with the expected number of five-mers. Five-mers are further divided into three groups; five-mers with CpG in their surroundings (e.g. CGAAT/ATTCG), those with CpG in the center (e.g. AGCGA/TCGCT) and five-mers with no CpGs.

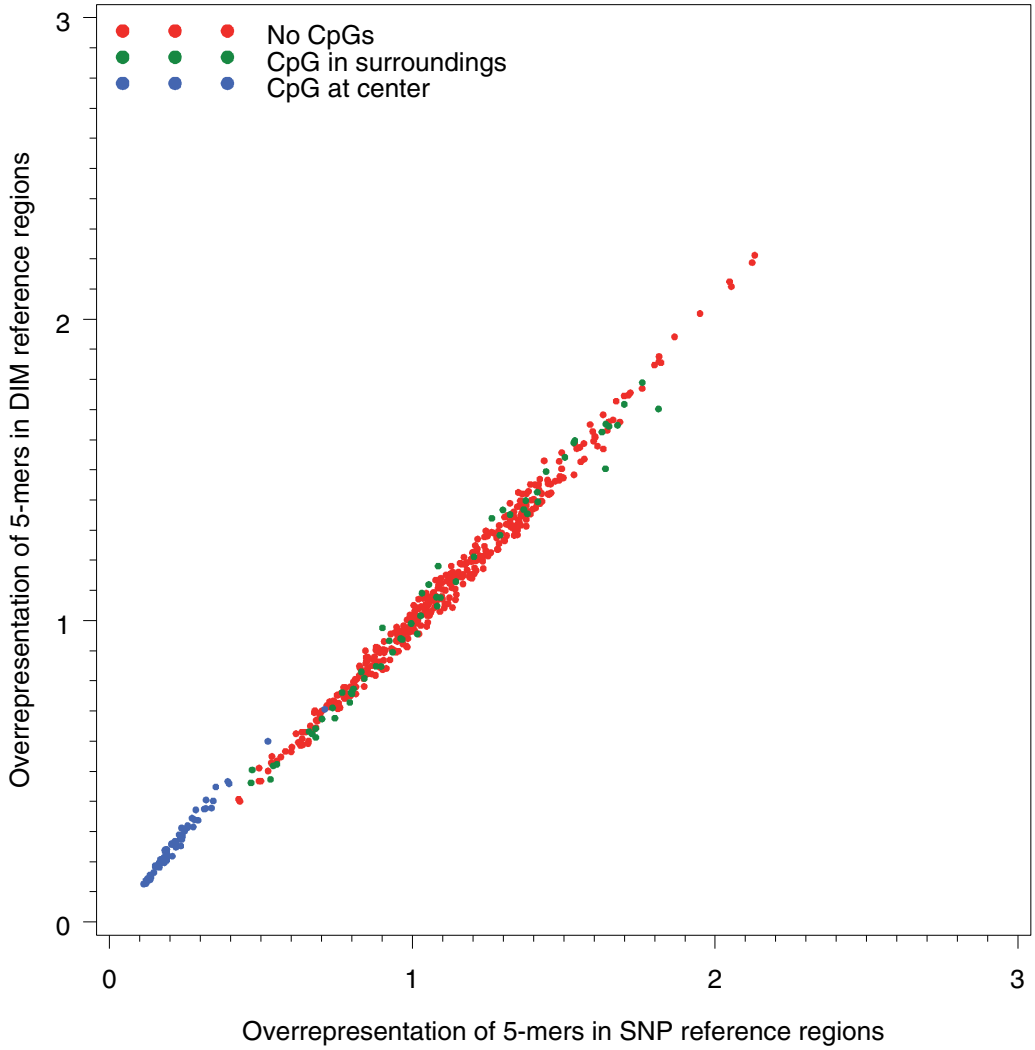


Figure 6
Overrepresentation of DNA oligomers (five-mers) in duplicated and non-duplicated genomic regions. Five-mers are divided into three groups; five-mers with CpG in their surroundings (e.g. CGAAT/ATTCG), those with CpG in the center (e.g. AGCGA/TCGCT) and five-mers with no CpGs.

observed for 25 inferred DIMs that were designated as MSVs by Fredman and colleagues.

Discussion

An understanding of the contextual patterns of nucleotide substitutions in the vertebrate genome is important for several reasons. The spectrum of mutational events reflect how the genome has been shaped during evolution, the mechanism of substitution mutagenesis, and it can also shed light on fundamental cellular processes such as genome stability, DNA replication and repair.

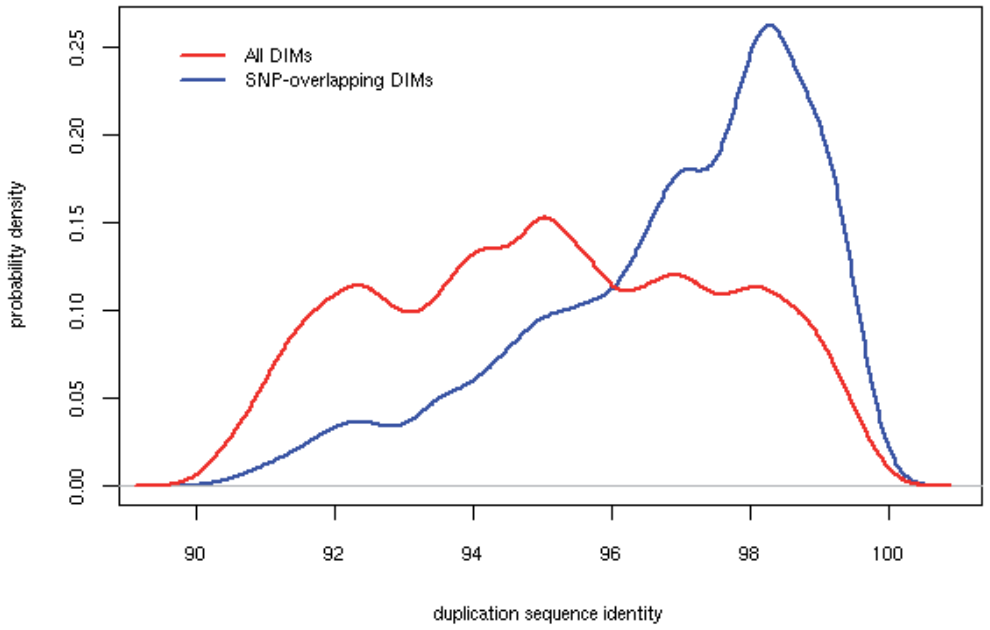


Figure 7
DIMs overlap with SNPs in segmental duplications at different levels of sequence divergence. Density distribution of the sequence divergence in which inferred mutations overlap (in terms of alleles and position) with predicted SNPs in segmental duplications. The distribution of all DIMs is shown for comparative purposes.

In this study, we have inferred a large set of point mutations originating within segmental duplications in the human genome. These point mutations were compared with a genome-wide collection of high-quality SNPs to assess whether these two datasets of mutational events show similar patterns in terms of distribution and surrounding sequence contexts. We initially recognized that regions of the genome covered by segmental duplications had a higher GC content than the grand average in non-duplicated regions. A previous study also reported a positive correlation between GC content and segmental duplications [34]. However, the biological interpretation of the strong association between GC content and segmental duplications is not obvious. One part may by attributed to the increased gene density in duplications [41], as regions containing genes are known to be GC-rich. Biased gene conversion may in addition play a role, a process in which repair of mismatches in heteroduplex recombination intermediates favour the fixation of G and C alleles [42,43]. Also, duplications are particularly enriched in subtelomeric regions of the chromosomes that are directly linked with GC-rich isochors [44].

The distribution of nucleotide substitutions observed in segmental duplications displays a pattern that in general is similar to SNPs. Both sets of mutations display an excess of transitional substitutions, a common phenomenon in vertebrate genomes. Among the four different transversions, the greatest difference between SNPs and DIMs were found for C/G substitutions. This finding suggests a potential association between the nucleotide composition of duplications and the frequency of substitutions, given the high GC content found in segmental duplications. Moreover, we observed a notable difference in the overall

ratio of transitions to transversions between DIMs and SNPs, and an increased ratio in recently occurring DIMs. These results may reflect the evolutionary time window in which the two sets of substitutions were sampled, as well as differences in nucleotide composition between duplicated and non-duplicated DNA. Substitutions within recent segmental duplications comprise mutational events potentially originating 35–40 million years ago (≥ 90% sequence identity) up until today (100% sequence identity), and will thus include a substitution spectrum beyond the human lineage. SNPs should on the other hand represent point mutational events within the human lineage only, as they represent genetic variation between humans. If one assumes that the rate of transversions and transitions varies over time [45], one would therefore expect to see stronger long-term effects within the DIM dataset than in the SNP set. Previous studies have shown that the rate of $5^{m}C$ deamination is limited by local GC content [46,47]. Thus, the GC richness of segmental duplications may be partly responsible for the fewer observed transitions relative to transversions.

The majority of DNA oligomers at DIM and SNP sites, respectively, displayed similar levels of abundance. This observation implies in essence that the majority of SNPs and DIMs appear to be generated by similar mutational mechanisms. We confirmed the latter in oligomers drawn from reference regions, that is intergenic regions of segmental duplications and intergenic, non-duplicated regions. However, we also discovered that many oligomers that contain substitutions at the CpG dinucleotide are overrepresented at SNPs while underrepresented at DIMs. In the reference regions, these oligomers were less underrepresented in duplications than in nonduplicated regions. As mentioned above, different effects at the CpG dinucleotide may be caused by differences in GC content, which in turn lead to different 5 mC deamination rates. Furthermore, when looking at the total mutational spectrum at the CpG dinucleotide, we observed that the frequency of methylation-related transitions differed significantly in CpG islands and non-island regions (Table 1). Our results henceforth imply that mutational events drawn from paralogous sequences exhibit the same suppression of methylation-dependent deamination in CpG islands as SNPs have been shown to do [15].

During large-scale computational identification of SNPs, many single nucleotide differences between genomic clones are taken as evidence of allelic variation and submitted to dbSNP. Without proper validation by other means, this form of SNP discovery will inevitably lead to spurious results caused by the duplication content of the human genome [26,48]. To address this issue, we systematically examined predicted SNP alleles in segmental duplications and mutations inferred from duplication alignments. Our approach revealed that nearly one out of five SNPs in duplications bear resemblance of paralogous sequence variation. Whether these SNPs behave like ordinary SNPs, MSVs or fixed PSVs is yet to be determined. Nonetheless, we suspect that traditional genotyping of the majority of these SNPs will produce misleading allele frequencies and genotype patterns since they will receive additional signals from paralogous sequences. Further, we discovered that SNPs that mirror mutational events in duplications are most prominent in duplications of high (≈97–100%) sequence identity, an observation for which we have no obvious explanation at present. In a comparative analysis with a small set of previously experimentally verified PSVs, we found all designated paralogous sequence variants among our computationally inferred mutations. In addition, we observed an overlap with computationally inferred DIMs and sites that were determined to be MSVs. The type of polymorphisms represented by MSVs involves a variation in duplication copy-number, and presumably indicates that much multisite variation may have originated from point mutational events in paralogous sequences.

Our approach does have some inherent limitations that could affect the reliability of the results obtained. These limitations involve the data source, i.e. detection of segmental duplications and reliability of DNA sequence alignments, the approach for inference of mutational events, and the sample effect. With respect to the source of segmental duplications, we relied on data provided by HGSDB [36]. The detection scheme employed by HGSDB uses BLAST for pairwise comparisons of all assembled chromosomes. Detected duplications will thus depend on the overall quality of the genome assembly, and inferred mutations will rely on correctly determined consensus sequences in the assembly. We reduced some potential assembly (and sequencing) errors by excluding high-copy repeats from the analysis, as assembly programs may fail to distinguish single base differences between repeat copies from erroneous base calls [49,50]. Since the degree of sequence divergence between duplications in HGSBB are all less than 10%, the resulting alignments are highly significant. Also, we placed restrictions on the alignment window around candidate DIMs to exclude potential alignment artefacts. Altering the alignment restrictions for DIM calling in two other DIM sets did not change the distribution of DIM substitutions to a large extent. In error-prone DNA sequencing contexts we observed a small increase of DIMs relative to high-quality SNPs, suggesting a minor impact of random noise in the DIM set. Altogether, we believe that the sequence alignments did not cause any serious errors.

Computational inference of mutational events leading to DIMs also has limitations. First of all, the directionality of

the mutations was not inferred with our approach, i.e. an A→T mutation could not be distinguished from a T→A mutation. Thus, an observed (C/T)G substitution may not necessarily reflect the deamination of a methylated thymine, but rather correspond to a thymine to cytosine transition. A recent study of the directionality of SNPs indicated that most substitutions in intergenic regions have roughly the same amounts of substitutions in either direction [11]. Whether DIMs display the same characteristics is unknown. Secondly, when the same mutational events were found propagated in several duplications (Figure 1B), we excluded them as individual events under the assumption of no multiple substitutions at a single site. This assumption is not likely to be violated in DNA sequences that show as low degree of sequence divergence as recent segmental duplications.

The sample of inferred DIMs were, as mentioned above, retrieved from all human chromosomes in regions where duplications have been found to exist. The total number of DIMs sampled was so large (≈344,000) that we believe they can provide a general pattern of substitutions in segmental duplications. In contrast to unique DNA sequences, duplicated sequences frequently undergo homology-driven mutation when involved in either non-allelic homologous recombination or gene conversion [28,42]. In the latter process, DNA repair of nucleotide mismatches in heteroduplex DNA intermediates has been shown to be GC-biased, providing a direct link to the GC-richness of duplications [51]. Investigating the relationship between biased repair and the observed distribution of DIMs requires further work, considering that base mispairs are corrected with different efficiencies and specificities in mammals [52]. The inferred point mutational spectrum was restricted to intergenic regions, excluding all DIMs located within RefSeq transcripts. Among all DIMs inferred, we thus omitted nearly 31.5% in our analyzed sample, as they all originated within UTRs, exons and introns residing in segmental duplications. As shown in early studies of molecular evolution, regions under functional constraints (i.e. human transcripts) show different patterns and rates of substitutions from selectively neutral sequences such as pseudogenes [53,54]. In order to establish a neutral pattern of point mutations in segmental duplications, minimized with the confounding effects of natural selection, we excluded any mutational event in which either of the nucleotides were found inside RefSeq transcripts. Since the point mutational spectrum in coding regions of segmental duplications may display different characteristics than what we found in intergenic regions, we suggest that these nucleotide substitutions should be explored in further work.

Most important, our computational analysis of segmental duplications in the human genome suggests that they can be utilized as a novel data source for the analysis of vertebrate point mutagenesis. There are essentially two different observations that support this claim. First, the distribution and context of computationally inferred DIMs and a set of high-quality set of SNPs in intergenic regions of the genome were largely similar (Figures 3, 5 and 6). Second, we found that a large fraction of the inferred DIMs overlap with verified SNPs, which provides evidence that our inference strategy is able to retrieve actual mutational events that lead to genetic variation. Moreover, our inferred set of nucleotide substitutions originates from regions in all human chromosomes, as segmental duplications are not restricted to any particular chromosome, but rather distributed in a genome-wide fashion. We believe that the inferred dataset of point mutations may be a valuable complement to SNPs for the analysis of human genetic variation.

Methods

Segmental duplication data

The Human Genome Segmental Duplication Database (HGSDB, http://projects.tcag.ca/humandup) has been reported to contain chromosomal coordinates of all segmental duplications (length ≥ 5 kb and sequence identity ≥ 90%) in the human genome, based on a computational detection scheme [36]. In total, 12589 unique pairwise sequence alignments of duplication copies were downloaded from HGSDB (build hg17). The two sequences in any pairwise sequence alignment of duplications were denoted as source and target sequences. 6587 alignments had both source and target sequences located on the same chromosome (intrachromosomal duplications), the remaining 6002 alignments had their duplication copies on nonhomologous chromosomes (interchromosomal duplications). Several regions were involved in both inter- and intrachromosomal duplications. The average alignment length was approximately 20.5 kb. The total nonredundant content of recent segmental duplications was found to be 133.9 Mb, comprising 4.7% of the non-gap length (2851.3 Mb) of the human genome. Chromosomal coordinates of RefSeq transcripts and CpG islands annotated to hg17 were downloaded as flat files from the UCSC genome browser http://genome.ucsc.edu and mapped to segmental duplications from HGSDB. High-copy repeats in segmental duplications were identified as lower-case nucleotides (output from RepeatMasker) within alignments downloaded from HGSDB.

Inference of mutational events in segmental duplications

Mutational events were inferred using DNA sequence alignments from HGSDB only. Figure 1A illustrates the basic inference principle. Since no other mammalian genome was used in our analysis, we did not attempt to infer the directionality of the mutational events or separate events that originated within different vertebrate lin-

eages. We merely inferred that mutational events had occurred since the duplication event took place. Two other factors related to the nature of segmental duplications had further impact on how DIMs were recorded (see Figure 1B). We wrote software for the traversal of pairwise DNA sequence alignments and recording of all mutational events along with their neighbouring sequence context (total entries n = 800,649). The dataset was reduced by excluding DIMs occurring in RefSeq transcripts as well as high-copy repeats as masked by RepeatMasker (n = 548,088). An alignment window of length 40 around each candidate DIM was extracted. To ensure that inferred DIMs were results of actual point mutational events rather than alignment artefacts, we only kept DIMs where the 40 bp alignment window satisfied the following criteria: (1) maximum four mismatches, (2) maximum two gaps (indels) and (3) no mismatches in the three immediate positions upstream and downstream of the candidate DIM site. With these criteria, the total number of intergenic inferred DIMs was 343,864. To test whether these alignment criteria induced any bias in the distribution of DIM types, we established two control sets in which DIMs were inferred in a stricter manner. In the first control set, we required a minimum of seven non-variant bases upstream and downstream of the candidate DIM site (258,612 DIMs), and in the second control set we increased this number to fifteen (108,117 DIMs).

To ensure that substitutions were sampled consistently across alignments with different sequence identity, we calculated the overall transition to transversion ratio for DIMs as a weighted sum of ten different bin ratios. DIMs were initially put in ten bins according to the sequence alignment identity in which they originated (i.e. 90 to 100), and a ratio for each bin was calculated without weighting. Each bin was then assigned a weight, representing the expected fraction of all substitutions that originated from alignments in the given bin. The expected number of substitutions in an alignment was estimated as alignment length multiplied by the fraction of nonidentical bases (the expected number in a bin was found by summing over all bin alignments).

SNP data

The human dbSNP database (build 126) was downloaded as XML files and parsed with Perl scripts for retrieval of biallelic RefSNP entries (reference SNPs). We established two different sets of SNP data. The first set contained a high-quality set of SNPs in non-duplicated regions of the genome, used for a comparative sequence context analysis with DIMs. The second set contained all reference SNPs in segmental duplications.

In the high-quality set of SNPs, we decided to only keep entries that were validated within the HapMap project [55]. We excluded all ambiguously mapped SNPs, that is, polymorphic sites where the flanking sequences did not map to a unique region in the genome with an alignment identity of at least 99% (total entries n = 2,160,150). The fraction of SNPs where allele frequencies in none of the four HapMap populations satisfied the basic SNP definition, that is, minor allele frequency ≥ 1%, were also omitted (as these may not mirror true SNP sites). The number of SNPs was further reduced by excluding SNPs that mapped within RefSeq transcripts (n = 1,337,235), SNPs where the flanking sequence (100 bp) fell inside high-copy repeats as masked by RepeatMasker (n = 1,131,893), and finally SNPs inside segmental duplications (n = 1,115,692).

A second set of SNPs was established by fetching all reference SNPs located within intergenic regions of segmental duplications, both validated and nonvalidated (n = 458,811). A SNP was classified as nonvalidated within dbSNP if it did not satisfy any of the following criteria: 1) allele frequencies in a given population, 2) multiple independent submissions, or 3) both alleles seen in at least two chromosomes. An overlap between a SNP and a DIM was considered valid if the chromosomal position of the SNP matched either the source or the target position associated with the DIM, and that the alleles at the SNP and DIM site matched (either directly or in a complementary manner if the SNP and DIM were recorded on different strands).

Sequence context of nucleotide substitutions

We determined whether similar mutational mechanisms act upon segmental duplications as in non-duplicated genomic regions by quantifying the frequencies of DNA oligomers at DIMs and high-quality SNPs. For comparison, we counted reference oligomer frequencies in the surrounding regions of DIMs (intergenic, duplicated DNA) and SNPs (intergenic, non-duplicated DNA).

Let uxv represent a k-mer where x is the middle nucleotide and u and v are surrounding nucleotides, and $u[xy]v$ represent a k-mer where the middle nucleotide is a substitution pair x/y. Let $n(uxv)$ count the number of k-mers that are either uxv or its reverse complement, and define $n(u[xy]v)$ similarly. We count SNPs and DIMs separately.

The nucleotide and substitution pair probabilities are $p(x) = n(x)/n$ and $p([xy]) = n([xy])/n$ for reference region and substitution respectively, with n the corresponding total number of nucleotides. Note that by this definition f.ex. $p(A) = p(T)$ is the probability of any nucleotide being either A or T, each with a 1/2 probability of being on either strand. If the middle nucleotide, x or $[xy]$, is independent of the surrounding nucleotides, the expected numbers in the reference regions are

$$\mu(uxv) = n(u^*v) \cdot p(x)/2$$

where $n(u^*v)$ is the sum of all $n(uxv)$ for different x, and the division by two is because there is a 1/2 chance that x is on the same strand as u^*v. For substitutions,

$$\mu(u[xy]v) = n(u[**]v) \cdot p(x)p(y)/2$$

except that if either $[xy]$ or u^*v are their own reverse complements one should not divide by 2. The overrepresentation (or abundance) is defined as $R(uxv) = n(uxv)/\mu(uxv)$, and $R(u[x^*]v) = n(u[x^*]v)/\mu(u[x^*]v)$ where $u[x^*]v$ indicates the sum over all matching $u[xy]v$ for n and μ.

Authors' contributions

EH conceived the study and outlined data analysis tasks. SN performed data retrieval and statistical analysis with help from EAR and TR. EH and TR provided feedback on results obtained. SN and EH drafted the manuscript together. All authors read and approved the final manuscript.

Acknowledgements

We thank various people (Razi Khaja, Jeff MacDonald and Dr. Steven W. Scherer) working at the Centre for Applied Genomics in Toronto for assistance with data from the Human Genome Segmental Duplication Database [36]. We also wish to thank Dr. Anthony J. Brookes and Dr. Evan E. Eichler for providing the rsIDs of SNPs analyzed in their study of sequence variation in segmental duplications [41].

References

1. Gartenberg MR, Crothers DM: **DNA sequence determinants of CAP-induced bending and protein binding affinity.** *Nature* 1988, **333(6176):**824-829.
2. Sims J, Rabbitts TH, Estess P, Slaughter C, Tucker PW, Capra JD: **Somatic mutation in genes for the variable portion of the immunoglobulin heavy chain.** *Science* 1982, **216(4543):**309-311.
3. Krawczak M, Ball EV, Cooper DN: **Neighboring-nucleotide effects on the rates of germ-line single-base-pair substitution in human genes.** *Am J Hum Genet* 1998, **63(2):**474-488.
4. Thilly WG: **Have environmental mutagens caused oncomutations in people?** *Nat Genet* 2003, **34(3):**255-259.
5. Aquilina G, Bignami M: **Mismatch repair in correction of replication errors and processing of DNA damage.** *J Cell Physiol* 2001, **187(2):**145-154.
6. Ehrlich M, Wang RY: **5-Methylcytosine in eukaryotic DNA.** *Science* 1981, **212(4501):**1350-1357.
7. Kunkel TA: **Misalignment-mediated DNA synthesis errors.** *Biochemistry* 1990, **29(35):**8003-8011.
8. Kunkel TA, Loeb LA: **Fidelity of mammalian DNA polymerases.** *Science* 1981, **213(4509):**765-767.
9. Lindahl T, Nyberg B: **Heat-induced deamination of cytosine residues in deoxyribonucleic acid.** *Biochemistry* 1974, **13(16):**3405-3410.
10. Blake RD, Hess ST, Nicholson-Tuell J: **The influence of nearest neighbors on the rate and pattern of spontaneous point mutations.** *J Mol Evol* 1992, **34(3):**189-200.
11. Jiang C, Zhao Z: **Mutational spectrum in the recent human genome inferred by single nucleotide polymorphisms.** *Genomics* 2006, **88(5):**527-534.
12. Zhao Z: **Neighboring-Nucleotide Effects on Single Nucleotide Polymorphisms: A Study of 2.6 Million Polymorphisms Across the Human Genome.** *Genome Res* 2002, **12(11):**1679-1686.
13. Zhao Z, Zhang F: **Sequence context analysis of 8.2 million single nucleotide polymorphisms in the human genome.** *Gene* 2006, **366(2):**316-324.
14. Cooper DN, Youssoufian H: **The CpG dinucleotide and human genetic disease.** *Hum Genet* 1988, **78(2):**151-155.
15. Tomso DJ, Bell DA: **Sequence context at human single nucleotide polymorphisms: overrepresentation of CpG dinucleotide at polymorphic sites and suppression of variation in CpG islands.** *J Mol Biol* 2003, **327(2):**303-308.
16. Bird AP: **DNA methylation and the frequency of CpG in animal DNA.** *Nucleic Acids Res* 1980, **8(7):**1499-1504.
17. Cooper DN, Krawczak M: **Cytosine methylation and the fate of CpG dinucleotides in vertebrate genomes.** *Hum Genet* 1989, **83(2):**181-188.
18. Duret L, Galtier N: **The covariation between TpA deficiency, CpG deficiency, and G+C content of human isochores is due to a mathematical artifact.** *Mol Biol Evol* 2000, **17(11):**1620-1625.
19. Jabbari K, Bernardi G: **Cytosine methylation and CpG, TpG (CpA) and TpA frequencies.** *Gene* 2004, **333:**143-149.
20. Simmen MW: **Genome-scale relationships between cytosine methylation and dinucleotide abundances in animals.** *Genomics* 2008, **92(1):**33-40.
21. Sved J, Bird A: **The expected equilibrium of the CpG dinucleotide in vertebrate genomes under a mutation model.** *Proc Natl Acad Sci USA* 1990, **87(12):**4692-4696.
22. Karlin S, Doerfler W, Cardon LR: **Why is CpG suppressed in the genomes of virtually all small eukaryotic viruses but not in those of large eukaryotic viruses?** *J Virol* 1994, **68(5):**2889-2897.
23. Pfeifer GP: **Mutagenesis at methylated CpG sequences.** *Curr Top Microbiol Immunol* 2006, **301:**259-281.
24. Shackelton LA, Parrish CR, Holmes EC: **Evolutionary basis of codon usage and nucleotide composition bias in vertebrate DNA viruses.** *J Mol Evol* 2006, **62(5):**551-563.
25. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K: **dbSNP: the NCBI database of genetic variation.** *Nucleic Acids Res* 2001, **29(1):**308-311.
26. Nelson MR, Marnellos G, Kammerer S, Hoyal CR, Shi MM, Cantor CR, Braun A: **Large-scale validation of single nucleotide polymorphisms in gene regions.** *Genome Res* 2004, **14(8):**1664-1668.
27. Platzer M, Hiller M, Szafranski K, Jahn N, Hampe J, Schreiber S, Backofen R, Huse K: **Sequencing errors or SNPs at splice-acceptor guanines in dbSNP?** *Nat Biotechnol* 2006, **24(9):**1068-1070.
28. Bailey JA, Gu Z, Clark RA, Reinert K, Samonte RV, Schwartz S, Adams MD, Myers EW, Li PW, Eichler EE: **Recent segmental duplications in the human genome.** *Science* 2002, **297(5583):**1003-1007.
29. Bailey JA, Yavor AM, Massa HF, Trask BJ, Eichler EE: **Segmental duplications: organization and impact within the current human genome project assembly.** *Genome Res* 2001, **11(6):**1005-1017.
30. Eichler EE: **Recent duplication, domain accretion and the dynamic mutation of the human genome.** *Trends Genet* 2001, **17(11):**661-669.
31. Samonte RV, Eichler EE: **Segmental duplications and the evolution of the primate genome.** *Nat Rev Genet* 2002, **3(1):**65-72.
32. Linardopoulou EV, Williams EM, Fan Y, Friedman C, Young JM, Trask BJ: **Human subtelomeres are hot spots of interchromosomal recombination and segmental duplication.** *Nature* 2005, **437(7055):**94-100.
33. She X, Jiang Z, Clark RA, Liu G, Cheng Z, Tuzun E, Church DM, Sutton G, Halpern AL, Eichler EE: **Shotgun sequence assembly and recent segmental duplications within the human genome.** *Nature* 2004, **431(7011):**927-930.
34. Zhang L, Lu HH, Chung WY, Yang J, Li WH: **Patterns of segmental duplication in the human genome.** *Mol Biol Evol* 2005, **22(1):**135-141.
35. Bailey JA, Eichler EE: **Primate segmental duplications: crucibles of evolution, diversity and disease.** *Nat Rev Genet* 2006, **7(7):**552-564.
36. Cheung J, Estivill X, Khaja R, MacDonald JR, Lau K, Tsui LC, Scherer SW: **Genome-wide detection of segmental duplications and potential assembly errors in the human genome sequence.** *Genome Biol* 2003, **4(4):**R25.
37. Estivill X, Cheung J, Pujana MA, Nakabayashi K, Scherer SW, Tsui LC: **Chromosomal regions containing high-density and ambigu-**

ously mapped putative single nucleotide polymorphisms (SNPs) correlate with segmental duplications in the human genome. *Hum Mol Genet* 2002, **11(17):**1987-1995.
38. Bosch E, Hurles ME, Navarro A, Jobling MA: **Dynamics of a human interparalog gene conversion hotspot.** *Genome Res* 2004, **14(5):**835-844.
39. Hurles ME: **Gene conversion homogenizes the CMT1A paralogous repeats.** *BMC Genomics* 2001, **2(1):**11.
40. Rozen S, Skaletsky H, Marszalek JD, Minx PJ, Cordum HS, Waterston RH, Wilson RK, Page DC: **Abundant gene conversion between arms of palindromes in human and ape Y chromosomes.** *Nature* 2003, **423(6942):**873-876.
41. Fredman D, White SJ, Potter S, Eichler EE, Den Dunnen JT, Brookes AJ: **Complex SNP-related sequence variation in segmental genome duplications.** *Nat Genet* 2004, **36(8):**861-866.
42. Chen JM, Cooper DN, Chuzhanova NA, Férec C, Patrinos GP: **Gene conversion: mechanisms, evolution and human disease.** *Nat Rev Genet* 2007, **8(10):**762-775.
43. Galtier N: **Gene conversion drives GC content evolution in mammalian histones.** *Trends Genet* 2003, **19(2):**65-68.
44. Costantini M: **An isochore map of human chromosomes.** *Genome Res* 2006, **16(4):**536-541.
45. Kimura M: **A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences.** *J Mol Evol* 1980, **16(2):**111-120.
46. Fryxell KJ, Moon W: **CpG mutation rates in the human genome are highly dependent on local GC content.** *Mol Biol Evol* 2005, **22(3):**650-658.
47. Zhao Z, Jiang C: **Methylation-dependent transition rates are dependent on local sequence lengths and genomic regions.** *Mol Biol Evol* 2007, **24(1):**23-25.
48. Reich DE, Gabriel SB, Altshuler DA: **Quality and completeness of SNP databases.** *Nat Genet* 2003, **33(4):**457-458.
49. Batzoglou S, Jaffe D, Stanley K, Butler J, Gnerre S, Mauceli E, Berger B, Mesirov JP, Lander E: **ARACHNE: a whole-genome shotgun assembler.** *Genome Res* 2002, **12(1):**177-189.
50. Tammi MT, Arner E, Kindlund E, Andersson B: **Correcting errors in shotgun sequences.** *Nucleic Acids Res* 2003, **31(15):**4663-4672.
51. Marais G: **Biased gene conversion: implications for genome and sex evolution.** *Trends Genet* 2003, **19(6):**330-338.
52. Brown TC, Jiricny J: **Different base/base mispairs are corrected with different efficiencies and specificities in monkey kidney cells.** *Cell* 1988, **54(5):**705-711.
53. Gojobori T, Li WH, Graur D: **Patterns of nucleotide substitution in pseudogenes and functional genes.** *J Mol Evol* 1982, **18(5):**360-369.
54. Imanishi T, Gojobori T: **Patterns of nucleotide substitutions inferred from the phylogenies of the class I major histocompatibility complex genes.** *J Mol Evol* 1992, **35(3):**196-204.
55. The International HapMap Consortium: **The International HapMap Project.** *Nature* 2003, **426(6968):**789-796.

II

Published online 17 July 2009

Nucleic Acids Research, 2009, Vol. 37, No. 17 ***5749–5756***
doi:10.1093/nar/gkp590

The disruptive positions in human G-quadruplex motifs are less polymorphic and more conserved than their neutral counterparts

Sigve Nakken[1,*], Torbjørn Rognes[1,2] and Eivind Hovig[2,3,4]

[1]Centre for Molecular Biology and Neuroscience, Institute of Medical Microbiology, Oslo University Hospital, Rikshospitalet, NO-0027, Oslo, [2]Department of Informatics, University of Oslo, PO Box 1080 Blindern, NO-0316, Oslo, [3]Department of Tumor Biology, Institute for Cancer Research and [4]Department of Medical Informatics, Oslo University Hospital, Norwegian Radium Hospital, Montebello, NO-0310, Oslo, Norway

Received May 18, 2009; Revised June 25, 2009; Accepted June 26, 2009

ABSTRACT

Specific guanine-rich sequence motifs in the human genome have considerable potential to form four-stranded structures known as G-quadruplexes or G4 DNA. The enrichment of these motifs in key chromosomal regions has suggested a functional role for the G-quadruplex structure in genomic regulation. In this work, we have examined the spectrum of nucleotide substitutions in G4 motifs, and related this spectrum to G4 prevalence. Data collected from the large repository of human SNPs indicates that the core feature of G-quadruplex motifs, 5′-GGG-3′, exhibits specific mutational patterns that preserve the potential for G4 formation. In particular, we find a genome-wide pattern in which sites that disrupt the guanine triplets are more conserved and less polymorphic than their neutral counterparts. This also holds when considering non-CpG sites only. However, the low level of polymorphisms in guanine tracts is not only confined to G4 motifs. A complete mapping of DNA three-mers at guanine polymorphisms indicated that short guanine tracts are the most under-represented sequence context at polymorphic sites. Furthermore, we provide evidence for a strand bias upstream of human genes. Here, a significantly lower rate of G4-disruptive SNPs on the non-template strand supports a higher relative influence of G4 formation on this strand during transcription.

INTRODUCTION

Human genomic DNA usually exists in the double-stranded conformation, but during denaturation, single strands containing tandemly repeated sequences can assemble into higher order DNA structures. In repetitive and guanine-rich sequences of the genome, single-stranded DNA can adopt four-stranded structures known as G-quadruplexes or G4 DNA (1). The G-quadruplex comprises a stack of G-tetrads, which are planar arrays of four guanines connected by Hoogsteen hydrogen bonds (2). G-quadruplexes are rapidly stabilized in the presence of monovalent cations, and their folding topology is influenced by the length and composition of short-sequence loops that link the stacked G-tetrads together (3–6). The first *in vitro* observations of G-quadruplex formation came from the single-stranded overhang at human telomeres (7,8), a sequence characterized by tandem repeats of TTAGGG. This finding was later followed by studies that demonstrated the existence of G-quadruplexes *in vivo* (9–11). The hypothesized role of G-quadruplex formation in living cells has received further support from the recognition of conserved factors that selectively bind and unwind G4 (12–15). However, the relative impact of G-quadruplex formation in the context of gene regulation and genome stability is still unclear.

Computational algorithms have been used to scan the human genome for the G4 consensus motif, which is a sequence containing at least four runs of at least three guanines (G-tracts) (16–18). These scans have identified enrichment in a number of chromosomal regions of biological importance, including the ribosomal DNA (19), the immunoglobulin heavy chain switch regions (20), telomeres (21) and transcriptional regulatory regions (22,23). With respect to gene transcription, different modes of G4-mediated regulation have been proposed. In one scenario, the formation of G4 is thought to increase the rate of transcription by preventing renaturation of double-stranded DNA (23). Others have though shown experimentally how small compounds can stabilize a promoter G-quadruplex and thereby decrease the expression rate (24). The idea that G-quadruplexes may act as regulators

*To whom correspondence should be addressed. Tel: +47 22 84 47 86; Fax: +47 22 84 47 82; Email: sigve.nakken@medisin.uio.no

of gene expression has been strengthened by multiple observations of G-quadruplex formation in human promoters, including the proto-oncogenes c-MYC (24,25) and c-KIT (26), as well as muscle-specific genes (27). Moreover, G4 motifs appear to be enriched in the promoters of other warm-blooded animals (28). Within motifs, there is a considerable preference for single-nucleotide loops between the consecutive guanine runs, and this is also characteristic of the experimentally derived structures that are most stable (22,29,30). The latter studies showed how a correlation between common sequence features of G4 motifs and observations *in vitro* might aid the interpretation of G4 prevalence. An important set of data that remains to be explored in this respect is the spectrum of common nucleotide polymorphisms in G4 motifs, and how this spectrum relates to findings from recent kinetic and spectroscopic studies of mutated G4 (31,32). The studies of single-base mutated G-quadruplexes have demonstrated a strong relationship between quadruplex stability and the mutation position, with the central guanines of G-tracts being most critical for stable quadruplex folds. Thus, if the G-quadruplexes exhibit biological activity in genomic regions, one would expect to see a relatively lower rate of polymorphic bases at critical sites of the G4 motif, as a consequence of negative selection. Taking into account the non-randomness of point mutagenesis, in which both base composition and DNA sequence contexts influence substitution rates (33–36), it is therefore of importance to see how the different sites in G4 motifs relate to known genetic variation in the form of human single nucleotide polymorphisms (SNPs). The collection of DNA polymorphisms in G4 motifs also represents an additional dimension in the identification of genomic regions undergoing G4 selection. In particular, the relative rate of G4-disruptive SNPs could indicate the extent of selection for the G-quadruplex structure in different genomic regions.

Here, we report a genome-wide analysis of SNPs in human G-quadruplex motifs, with an emphasis towards their occurrences in gene and regulatory sequences. We have used a large collection of validated SNPs from dbSNP as our data source of nucleotide substitutions (37). Overall, the results demonstrate a non-random pattern of nucleotide polymorphism in G-quadruplex motifs. In particular, we show that the internal sites of guanine runs are well protected from polymorphisms in the human genome, indicating a relationship between sequence-dependent mutagenesis of guanine and the prevalence of guanine tracts.

MATERIALS AND METHODS

SNP data

dbSNP (build 129, released on 18 April 2008) was downloaded in XML format from ftp://ftp.ncbi.nlm.nih.gov/snp/. We included SNPs that (i) were biallelic, (ii) had been uniquely mapped to the human genome with an alignment accuracy of at least 99%, (iii) had been validated by at least one of NCBI's validation criteria (that is, 'by-frequency', 'byCluster', 'by2Hit2Allele' or 'byOtherPop') and (iv) if genotyped by the HapMap project, had a minor allele frequency of at least 1% in minimum one of the sampled populations. A total of 5 717 575 SNPs satisfied the criteria above.

Sequence and annotation data

We used the *quadparser* algorithm to retrieve all sequences in the human genome (NCBI build 36.3) capable of forming a G-quadruplex, identified by the sequence motif $G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}$, where G is guanine and N is any nucleotide (16). This simple consensus was inferred after several biophysical experiments had investigated the sequence basis for stable quadruplex folds (3,4), and represents the most common approach to map the grand total of potential G-quadruplex forming sequences. From the *quadparser* output, we extracted each putative G-quadruplex motif, regardless of any potential overlap with a neighboring motif [this corresponds to the 'un-restricted' set of G4 motifs, as defined by Todd *et al.* (17)]. Motifs with guanine tracts of length greater than six were excluded. The choice of overlapping motifs allowed us to evaluate the context and effect of a SNP for each individual putative G-quadruplex-forming structure. We only considered SNPs that mapped to G4 motifs present in the reference genome; SNPs that potentially introduced new G4 motifs were not analyzed.

The genomic coordinates of 24 243 protein-coding RefSeq genes were downloaded from ftp://ftp.ncbi.nih.gov/refseq (NCBI build 36.3) and used for the annotation of G4 motifs. CpG islands and 28-way vertebrate MultiZ alignments were obtained from the UCSC genome browser (38), available at http://genome.ucsc.edu. Motifs located in four defined genomic regions were subsequently analyzed: 5′ gene regions, 3′ gene regions, the first gene intron and intergenic regions. In order to target regulatory G4 sequences involved in gene transcription, we set the limits of the 5′ region of genes to 2-kb upstream of the transcription start site (TSS) and 1-kb downstream of the TSS. Only non-coding sequences (i.e. UTR) were targeted downstream of the TSS (Figure 1a), since coding sequences exhibit a significant depletion of G4 (39). We are aware that downstream of the TSS, the 5′ region will encompass G4 motifs that could be involved in both transcription and RNA processing. Ideally, one should thus evaluate the upstream and downstream regions of the TSS separately. However, having limited our analysis to the transcriptional aspect of G4, we considered it appropriate to combine the contributions by pre-transcription regulatory G4 (upstream of the TSS) and transcription regulatory G4 (downstream of the TSS). The 3′ end of genes was defined in the same manner as the 5′ end, encompassing 1-kb within 3′ UTR and 2-kb downstream of the transcription stop site. We included an analysis of G4 in the first intron (restricted to the first thousand bp), since this genomic region has shown a particular enrichment of G4 (40). Last, for control purposes, we included G4 motifs located in intergenic regions of the human genome.

Genomic G4 motifs that were found within high-copy repeats (as identified by RepeatMasker and Tandem Repeats Finder) were excluded from the analysis.

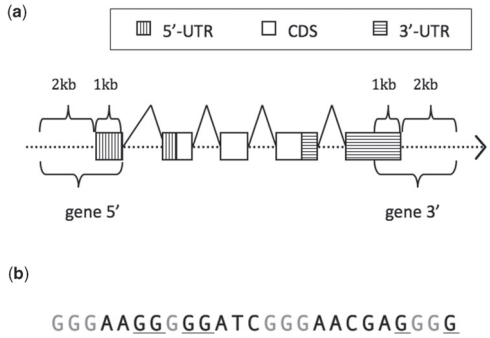


Figure 1. (**a**) A simplified illustration of a human gene, showing how the gene 5′ and gene 3′ regions were defined. (**b**) An example of a G4 sequence motif. The G4-disruptive sites are in grey colour, while the G4-neutral sites are in black. The underlined guanines are guanines within tracts that, when mutated, will not disrupt the G4 consensus.

There were several reasons for this decision. First of all, in the genomic regions of interest (regulatory sequences), the frequency of G4 within unique sequence is nearly twice as that of G4 within repeats. Second, reliable (i.e. validated) SNPs are under-represented in repeats; whereas 51.1% of all reference SNPs in dbSNP are mapped to repeats, only 45.1% of the validated SNPs are located within repeats. Third, in the vertebrate MultiZ alignments, we noted that the availability of reliable alignments for G4 in repeats was poor compared to unique G4.

Non-G4 control sequences

In a search for characteristic patterns of substitutions in the G-rich G4 motifs, we established a set of non-G4 control sequences. The selected non-G4 sequences had the same high GC content as the G4 sequences, but did not match the G4 consensus. This approach enabled us to target differences between G4 and non-G4 unrelated to CpG dinucleotides, since the rate of the most common substitution at CpG dinucleotides (i.e. transition caused by spontaneous hydrolytic deamination of 5-methylcytosine) are dependent on GC content (34,41).

We next provide a short description of the stepwise procedure. For each genomic region analyzed, we created a large library of non-G4 sequence fragments (length 20–28 bp; average length of G4 motifs) that originated either outside or within CpG islands. All fragments were subsequently binned according to GC content. We randomly picked sequence fragments within each bin, the number of fragments being dictated by the probability distribution of G4 motifs with respect to GC content and CpG islands. The SNP density in the total collection of non-G4 fragments was then calculated. This procedure was repeated fifty times for each genomic region and averaged.

RESULTS AND DISCUSSION

Previous studies have demonstrated the importance of computational analyses for the understanding of G4 enrichment in vertebrate genomes (16,17,22,23,28,40, 42–45). In this work, we investigate G4 prevalence from a single nucleotide substitution perspective.

We calculated the density of SNPs in G4 motifs by querying the dbSNP database at the locations of 282 501 motifs in non-repetitive regions of the human genome. Due to the overlapping nature of many G4 motifs (and also some overlapping gene annotations), a number of SNPs were counted more than once in the overall count of SNPs. We checked that this approach did not influence our findings by performing an alternative analysis allowing only one count per SNP in non-overlapping motifs (data not shown). The strandedness of G4 motifs was ignored at this point, and we thus combined the total G4 formation potential involved in either DNA replication or gene transcription.

A total of 10 794 validated SNPs mapped to G4 motifs in the human genome, with an overall density of 1.97 SNPs/kb. With an estimated density of 2.00 SNPs/kb in the genomic background, it was apparent that the level of polymorphism in G4 motifs reflected the genome average. This finding seemed intuitively somewhat unexpected, considering the 2-fold enrichment of hypermutable CpG dinucleotides in G4 compared to the genomic background (Table 1). However, there are two important characteristics of G4 motifs that impose a relatively lower rate of SNPs at CpGs in these sequences. The first feature is the high GC content of G4, since 5-methylcytosine deamination rates are inversely correlated with local GC content (34). Second, there is an extensive overlap between G4 and CpG islands, that is genomic regions in which the cytosines of CpG dinucleotides preferentially remain unmethylated (45,46). Specifically, the coverage density of G4 inside CpG islands was several-fold higher than outside islands (Table 1). The latter observation implies that many G4 CpGs inevitably appear unmethylated in the genome, and this will likely reduce their overall mutagenic potential.

We next sought to identify mutational patterns of G4 motifs that were not related to CpG. To do so, we compared them with a set of randomly picked non-G4 sequences that matched the GC distribution of G4 (see 'Materials and Methods' section). Sampling non-G4 sequences in this manner enabled us to target non-CpG types of pattern in G4, since the mutational characteristics of CpG were approximately equalized between G4 and non-G4. We observed that the SNP density in G4 was consistently lower than in non-G4 sequences, although to a varying extent in the different genomic regions (Figure 2). Since the primary sequence difference between G4 and the random non-G4 fragments was the density of guanine triplets, we hypothesized a suppression of nucleotide polymorphisms in the G4 tetrad regions (i.e. guanine triplets), and that this phenomenon would influence the relative low rate of G4 SNPs.

Critical sites of G4 motifs display low levels of polymorphism

We next investigated whether loop and tetrad (i.e. G-tracts) regions of G4 motifs are subject to different

Table 1. Density of SNPs and CpG dinucleotides in G4 motifs

	Number of G4 motifs	Number of SNPs[a]	SNPs/CpG[b]	CpG island coverage[c]	CpG/kb[d]
Genome	282 501	–	–	–	–
First introns	17 926 (0.33 Mb)	555 (441)	0.00413 (0.014)	0.093 (0.014)	58.4 (28.7)
Gene 5′	31 694 (0.55 Mb)	1157 (874)	0.0052 (0.012)	0.044 (0.010)	57.7 (34.9)
Gene 3′	17 458 (0.30 Mb)	906 (639)	0.0190 (0.038)	0.048 (0.008)	29.5 (13.6)
Intergenic	103 911 (2.01 Mb)	5001 (4096)	0.023 (0.064)	0.036 (0.002)	22.6 (7.6)

[a]Total number of SNPs that map to G4 motifs. The number of unique (non-redundant) SNPs is given in parentheses.
[b]The density estimate of SNPs at G4-CpGs included only C/T and A/G SNPs, since the majority of substitutions occurring at the hypermutable CpG are methylation-dependent transitions. A similar density estimate of SNPs at CpGs in the genomic background is given in parentheses.
[c]Coverage is defined as the fraction of island bases covered by G4 bases. Coverage of G4 outside CpG islands is given in parentheses.
[d]CpG density in genomic background is given in parentheses.

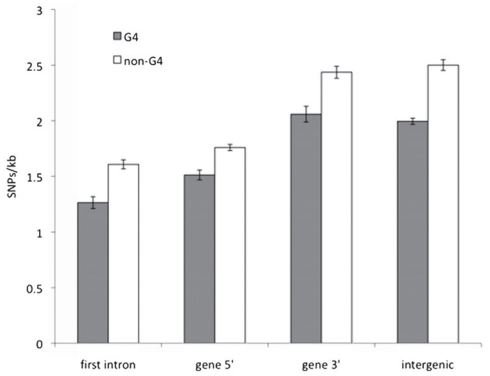


Figure 2. SNP density in G4 sequences versus randomly picked non-G4 sequences. The set of non-G4 sequences were drawn such that their GC-richness was equivalent to that of G4.

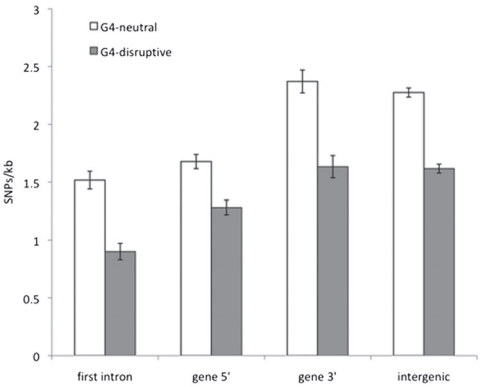


Figure 3. SNP density in G4-disruptive sites versus G4-neutral sites (see Figure 1b for a definition of G4-disruptive and G4-neutral).

mutational pressures. The two distinct G4 regions are important for quadruplex formation and stability, the G-tracts that make up the tetrad planes being critical for formation and folding (32). It is worth noting that, in the G-tracts of G4 motifs, not all substitutions of guanine will disrupt the potential to form a quadruplex structure. For example, if a motif contains a run of four guanines, substitutions at either end of the run will not disrupt the required triplet and could therefore, in principle, preserve the quadruplex-forming potential. On the basis of this reasoning, we classified each position in G4 motifs as either 'G4-disruptive' or 'G4-neutral' (Figure 1b). In all genomic regions analyzed, we found a significantly lower rate of SNPs in G4-disruptive positions relative to the G4-neutral positions (Figure 3). However, since hypermutable CpGs are more frequent at neutral positions than disruptive positions by a factor of nearly three, we performed an additional analysis where CpG sites were masked (Table 2). The difference in SNP density between neutral and disruptive G4 positions decreased when considering non-CpG sites only, though disruptive sites still displayed a significantly lower level of sequence polymorphism. We elaborated on this finding with comparative genomics data, assessing the level of sequence conservation within the two classes of G4 sites. This was accomplished by constructing a four-species multiple sequence alignment (human, monkey, dog and mouse) of G4 motifs from the 28-way vertebrate MultiZ alignments. The disruptive sites of CpG-masked G4 motifs showed consistently higher levels of mammalian sequence conservation than non-disruptive sites (Figure 4).

The evident conserved nature and suppressed level of polymorphisms at G4-disruptive sites could, intuitively, be interpreted as if the G-quadruplex consensus sequence is under functional constraints in the genome. The basic rationale for this argument comes from two recent studies of mutated G-quadruplexes, which demonstrated that their conformational dynamics strongly depends on the position of the mutated guanine (31,32). In an analysis that applied single-molecule FRET spectroscopy on telomeric G4 motifs, the G-quadruplex was severely destabilized when a central guanine was substituted with thymine. Substitutions at the end of a guanine tract also produced less stable structures, though with a far less dramatic effect than the central ones (31). In accordance with these data, we observed a tendency in which the critical guanines of human G4 motifs are less polymorphic than their neutral counterparts. However, we found that this characteristic

Table 2. Density of SNPs in disruptive and neutral sites of G4 sequence motifs

	G4-neutral			G4-disruptive	
	CpGs/kb	SNPs/kb[a]	P[b]	CpGs/kb	SNPs/kb[a]
First introns	166.5	1.52 (1.38)	<0.00001	73.1	0.90 (0.91)
Gene 5′	166.1	1.68 (1.48)	<0.05	71.4	1.28 (1.29)
Gene 3′	83.9	2.37 (1.72)	<0.05	36.1	1.63 (1.45)
Intergenic	65.0	2.28 (1.65)	<0.001	25.6	1.62 (1.46)

[a]Density of SNPs in non-CpG sites are given in parentheses.
[b]Difference in SNP density between G4-disruptive and G4-neutral sites (non-CpG) by Chi-squared analysis.

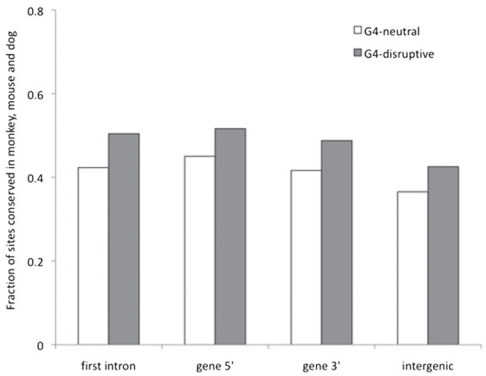


Figure 4. Sequence conservation in G4-disruptive sites versus G4-neutral sites. Shown is the fraction of conserved (i.e. all bases identical) sites at G4-disruptive and G4-neutral sites, as extracted from MultiZ sequence alignments of human G4 with monkey (rheMac2), dog (canFam2) and mouse (mm8). Only non-CpG sites were probed for conservation.

feature of G4 motifs occurred genome-wide, in a strand-independent manner, and also among G4 motifs in intergenic regions. These latter observations suggested that the phenomenon occurs as an effect of intrinsic mutation or DNA repair mechanisms rather than as a consequence of selection for the G4 consensus.

General under-representation of SNPs in guanine tracts

The distribution of SNPs in G4 motifs revealed that nucleotide polymorphisms in G4 DNA would more likely alter the loop conformation than the quadruplex-forming potential. We next asked whether this pattern of guanine substitutions is occurring in a genome-wide fashion, not restricted to the G-tracts of G4 motifs. More specifically, we estimated the relative over-representation of each DNA three-mer at polymorphic guanines by comparing its frequency at polymorphic sites versus non-polymorphic sites, adopting the approach used by Tomso and co-workers (41). For each polymorphic site, two centered three-mers were recorded, one for each allele. Importantly, since the SNP data does not provide any

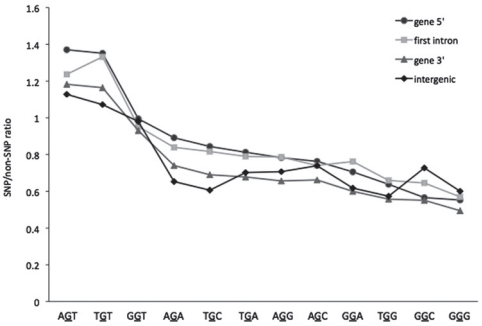


Figure 5. The ratio of DNA three-mers at polymorphic to non-polymorphic sites. Only non-CpG three-mers have been plotted, and each three-mer ratio constitutes the combined ratio of the forward and reverse complementary context. Only SNPs that were proven polymorphic by the HapMap project were used in the calculation.

information as to which strand the original mutational event occurred, we cannot distinguish between a context and its reverse complementary context. We thus ignored strandedness and pooled reverse complementary three-mers together. We confirmed previous observations that CpG-containing three-mers are the most over-represented sequence contexts at human SNPs (36,41). At the opposite end, we observed that a guanine surrounded by other guanines (i.e. 5′-GGG-3′/5′-CCC-3′, polymorphic site underlined), is among the DNA sequence contexts that is most under-represented at polymorphic sites (Figure 5). In fact, it was the most under-represented sequence context among polymorphisms within first introns, at the 5′ end of genes, and at the 3′ end of genes. Our data thus indicate that SNPs with the highest probability of disrupting G-tracts represent the most under-represented SNP context in regulatory gene sequences. We also noted that for sequence contexts at both ends of G-tracts, which for three-mers constitute the 5′-NGG-3′/5′-CCN-3′ and 5′-GGN-3′/5′-NCC-3′ contexts, the frequencies of polymorphisms were generally low. An exception was the 5′-GGT-3′/5′-ACC-3′ context (and the CpG-containing 5′-CGG-3′/5′-CCG-3′, not shown in Figure 5).

Which biological mechanisms could underlie the low rate of polymorphisms inside guanine/cytosine tracts? The phenomenon was not only evident in regulatory regions, but also appeared to occur in intergenic regions, where the modulation of mutational output by natural selection is believed to be weaker. The latter suggests that the observed pattern of SNPs reflects a context-dependency in the mechanisms underlying human mutation. The mutational input to polymorphisms in DNA is considered to be base damage or incorporation of incorrect bases by polymerases during replication, followed by no or error-prone DNA repair. Both the frequency of damages, and the efficiency and fidelity of DNA replication and repair are probably dependent on the sequence context. It is clear that a very significant source of mutations is due to deamination of 5-methylcytosine (5mC) in

CpG dinucleotides. An important additional source of mutations is due to lacking or error-prone repair of 7,8-dihydro-8-oxo-guanine (8-oxoG) in the DNA. It may be caused by UV radiation or oxidative damage to guanine. Several DNA repair systems targets this type of damage, including base excision repair and mismatch repair, but they are not perfect. The damage may occur either to guanines in the nucleotide pool or directly to the guanines in the DNA. In the former case, 8-oxoG may subsequently be incorporated into the DNA unless degraded by the NUDT1 hydrolase (47). If 8-oxoG in the DNA is not removed by the OGG1 glycosylase (48,49), subsequent replication may lead to an adenine being incorrectly incorporated opposite the 8-oxoG instead of a cytosine. If the adenine is not removed by the MUTYH glycosylase (50) before the next round of replication, this process may result in a G:C to T:A transversion. McCulloch *et al.* (51) has recently studied the efficiency and fidelity of DNA in 8-oxoG bypass by polymerases, and their work may indicate a slight dependency on the sequence context for the human polymerase η. Further work is necessary to determine, in detail, the context dependency of polymerases and if this can be a basis for sequence-dependent mutation rates.

Imbalance in the nucleotide precursor pool represents another potential source of mutations. In a mammalian model system that induced thymidine mutations by pool perturbation, it was shown that guanine residues flanked on their 3′ side by other guanine residues are severalfold less mutable than guanine residues flanked on their 3′ side by a different base (52). The underlying mechanism for this pattern was not examined. The authors do, however, argue that differential repair of misincorporated thymidines could be involved. Nonetheless, it is intriguing to see how well these patterns of induced mutations fit with the spectrum we observed for guanine SNPs.

Could systematic DNA-sequencing errors among the polymorphisms collected from dbSNP account for the observed pattern? It has been shown that a few sequence contexts are particularly prone to sequencing errors (one of them being C(A/Y)C), and that these are over-represented among non-validated SNPs (53). However, our strategy to pick SNPs from dbSNP was designed in a conservative manner (see 'Materials and Methods' section), thereby excluding the majority of false-positive SNPs. Also, we imposed even stricter requirements in the analysis of SNP three-mers, in which we only considered SNPs that were proven polymorphic by HapMap genotyping.

A G4 strand bias for disruptive SNPs

In the previous analyses of SNPs in G4 motifs, we considered general G4 formation potential during DNA denaturation, thereby ignoring the strand orientation of motifs. If we regard G4-regulated gene transcription as a separate process, the potential for regulation lies primarily within motifs on the nontemplate strand, which has shown a significant enrichment relative to the template strand (40,42). We therefore undertook an additional analysis of SNPs in G4 that incorporated strandness of motifs.

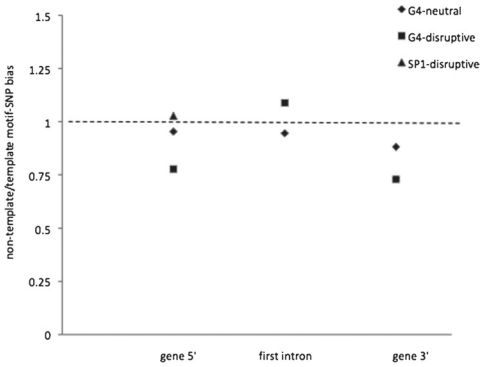


Figure 6. The ratio of SNP density (non-CpG) in nontemplate motifs to the SNP density in template motifs. The dashed line indicates a similar rate of SNPs with respect to the strandedness of the motif, i.e. no strand bias.

The extent of G4 strand bias was defined as the ratio of SNP density (non-CpG) in G4 on the non-template strand to the SNP density in G4 on the template strand, where a ratio of 1 implies no strand bias. Interestingly, we observed a marked strand bias for G4-disruptive SNPs in regulatory sequences, while negligible biases were observed among the neutral G4 SNPs (Figure 6). For disruptive SNPs, it was evident that their density in G4 motifs on the non-template strand was lower than on the template strand. This bias was significant at the 5′ end of genes ($P < 0.02$, $\chi^2 = 5.77$, df = 1) and at the 3′ end of genes ($P < 0.02$, $\chi^2 = 5.82$, df = 1). The result was not an artefact of the overlapping G4 motifs (and SNPs), since the count of unique SNPs in non-overlapping G4 motifs also produced significant strand biases at a significance level of 0.05 (data not shown). As a means to validate the observation at the 5′ end, and to test whether the result was a mere consequence of general suppression of polymorphisms in guanine tracts on the nontemplate strand, we carried out a similar type of analysis with a related sequence element, the SP1 transcription factor (5′-GGGCGG-3′) (44). More specifically, we asked whether there is a strand bias (with respect to SP1) for nucleotide polymorphisms that disrupt the SP1 motif at positions 2 or 3 (two non-CpG sites). The level of SP1 disruption did not differ significantly between the two strands at the 5′ end ($P = 0.855$, $\chi^2 = 0.03$, df = 1), although the set of polymorphisms that mapped to the SP1 motif was considerably smaller than the G4 set (524 SP1 polymorphisms versus 1157 G4 polymorphisms).

The low rate of human SNPs in G4-disruptive positions on the non-template strand support a higher relative importance for this strand in G4-mediated gene regulation. When present on this strand downstream of the TSS, the G-quadruplex may form as part of the pre-mRNA and/or potentially the mRNA, and it may thus serve as multiple targets for regulation (40). The formation of G4 on the template strand would on the other hand hinder the progression of the RNA polymerase, and is

therefore less desirable (23). We also showed that another G-rich element, the SP1 transcription factor, did not display any strand bias with respect to disruptive SNPs at the 5′ end. It may thus seem as if the pattern supports a specific biological importance for G4 motifs on the non-template strand at the 5′ end of genes.

CONCLUSION

The recent genome-wide scans of G4 motifs in the human genome have identified enrichment in gene regulatory sequences, and the same tendency has been shown when searching the genomes of chimpanzee, rat and mouse (16,17,45). The prevalence of G4 motifs upstream of mammalian genes has been interpreted as a sign of selection for G4, and consequently implicated the G-quadruplex structure as a potential mechanism for regulating gene expression (23,39). On the basis of sequence data only, it is nonetheless impossible to determine the extent of quadruplex formation *in vivo*, although it seems most likely that only a low percentage of the G4 motifs will adopt structures during denaturation.

Here, a close examination of the context-dependent pattern of guanine polymorphisms has provided an additional perspective on G4 prevalence. It shows how the aspect of sequence mutagenesis could impact the evolution of guanine tracts, the key component in G4 motifs. Although significant patterns emerged, our results are limited by the approximately 11 000 SNPs that map to G4 motifs in the human genome. Following next-generation sequencing and collaborative efforts such as the 1000 Genome Projects (54), more data should be available for studying the nature of G4 sequence polymorphism. An interesting extension of our analysis, which requires more validated SNPs available, is to relate the directionality of each SNP (i.e. by determining the ancestral and derived allele) to G4 evolution. Nevertheless, in light of our current findings, we warrant a closer examination of the relationship between G4 and other factors that might constrain the nearest-neighbour sequence patterns in DNA, an example being the physical requirements needed for the dense packing of DNA around nucleosomes.

FUNDING

Research Council of Norway. Funding for open access charge: the EU FP7 contract 223367.

Conflict of interest statement. None declared.

REFERENCES

1. Sen,D. and Gilbert,W. (1988) Formation of parallel four-stranded complexes by guanine-rich motifs in DNA and its implications for meiosis. *Nature*, **334**, 364–366.
2. Gellert,M., Lipsett,M.N. and Davies,D.R. (1962) Helix formation by guanylic acid. *Proc. Natl Acad. Sci. USA*, **48**, 2013–2018.
3. Hazel,P., Huppert,J., Balasubramanian,S. and Neidle,S. (2004) Loop-length-dependent folding of G-quadruplexes. *J. Am. Chem. Soc.*, **126**, 16405–16415.
4. Risitano,A. and Fox,K.R. (2004) Influence of loop size on the stability of intramolecular DNA quadruplexes. *Nucleic Acids Res.*, **32**, 2598–2606.
5. Burge,S., Hazel,P. and Todd,A.K. (2006) Quadruplex DNA: sequence, topology and structure. *Nucleic Acids Res.*, **34**, 5402–5415.
6. Rachwal,P.A., Findlow,I.S., Werner,J.M., Brown,T. and Fox,K.R. (2007) Intramolecular DNA quadruplexes with different arrangements of short and long loops. *Nucleic Acids Res.*, **35**, 4214–4222.
7. Sundquist,W.I. and Klug,A. (1989) Telomeric DNA dimerizes by formation of guanine tetrads between hairpin loops. *Nature*, **342**, 825–829.
8. Williamson,J.R., Raghuraman,M.K. and Cech,T.R. (1989) Monovalent cation-induced structure of telomeric DNA: the G-quartet model. *Cell*, **59**, 871–880.
9. Schaffitzel,C., Berger,I., Postberg,J., Hanes,J., Lipps,H.J. and Pluckthun,A. (2001) In vitro generated antibodies specific for telomeric guanine-quadruplex DNA react with Stylonychia lemnae macronuclei. *Proc. Natl Acad. Sci. USA*, **98**, 8572–8577.
10. Duquette,M.L., Handa,P., Vincent,J.A., Taylor,A.F. and Maizels,N. (2004) Intracellular transcription of G-rich DNAs induces formation of G-loops, novel structures containing G4 DNA. *Genes Dev.*, **18**, 1618–1629.
11. Paeschke,K., Simonsson,T., Postberg,J., Rhodes,D. and Lipps,H.J. (2005) Telomere end-binding proteins control the formation of G-quadruplex DNA structures *in vivo*. *Nat. Struct. Mol. Biol.*, **12**, 847–854.
12. Bachrati,C.Z. and Hickson,I.D. (2006) Analysis of the DNA unwinding activity of RecQ family helicases. *Methods Enzymol.*, **409**, 86–100.
13. Sun,H., Karow,J.K., Hickson,I.D. and Maizels,N. (1998) The Bloom's syndrome helicase unwinds G4 DNA. *J. Biol. Chem.*, **273**, 27587–27592.
14. Wu,Y., Shin-ya,K. and Brosh,R.M. Jr. (2008) FANCJ helicase defective in Fanconia anemia and breast cancer unwinds G-quadruplex DNA to defend genomic stability. *Mol. Cell Biol.*, **28**, 4116–4128.
15. Fry,M. (2007) Tetraplex DNA and its interacting proteins. *Front. Biosci.*, **12**, 4336–4351.
16. Huppert,J.L. and Balasubramanian,S. (2005) Prevalence of quadruplexes in the human genome. *Nucleic Acids Res.*, **33**, 2908–2916.
17. Todd,A.K., Johnston,M. and Neidle,S. (2005) Highly prevalent putative quadruplex sequence motifs in human DNA. *Nucleic Acids Res.*, **33**, 2901–2907.
18. Kikin,O., D'Antonio,L. and Bagga,P.S. (2006) QGRS Mapper: a web-based server for predicting G-quadruplexes in nucleotide sequences. *Nucleic Acids Res.*, **34**, W676–W682.
19. Hanakahi,L.A., Sun,H. and Maizels,N. (1999) High affinity interactions of nucleolin with G-G-paired rDNA. *J. Biol. Chem.*, **274**, 15908–15912.
20. Dempsey,L.A., Sun,H., Hanakahi,L.A. and Maizels,N. (1999) G4 DNA binding by LR1 and its subunits, nucleolin and hnRNP D, A role for G-G pairing in immunoglobulin switch recombination. *J. Biol. Chem.*, **274**, 1066–1071.
21. Wang,Y. and Patel,D.J. (1993) Solution structure of the human telomeric repeat d[AG3(T2AG3)3] G-tetraplex. *Structure*, **1**, 263–282.
22. Huppert,J.L. and Balasubramanian,S. (2007) G-quadruplexes in promoters throughout the human genome. *Nucleic Acids Res.*, **35**, 406–413.
23. Du,Z., Zhao,Y. and Li,N. (2008) Genome-wide analysis reveals regulatory role of G4 DNA in gene transcription. *Genome Res.*, **18**, 233–241.
24. Siddiqui-Jain,A., Grand,C.L., Bearss,D.J. and Hurley,L.H. (2002) Direct evidence for a G-quadruplex in a promoter region and its targeting with a small molecule to repress c-MYC transcription. *Proc. Natl Acad. Sci. USA*, **99**, 11593–11598.
25. Simonsson,T., Pecinka,P. and Kubista,M. (1998) DNA tetraplex formation in the control region of c-myc. *Nucleic Acids Res.*, **26**, 1167–1172.
26. Fernando,H., Reszka,A.P., Huppert,J., Ladame,S., Rankin,S., Venkitaraman,A.R., Neidle,S. and Balasubramanian,S. (2006)

A conserved quadruplex motif located in a transcription activation site of the human c-kit oncogene. *Biochemistry*, **45**, 7854–7860.
27. Yafe,A., Etzioni,S., Weisman-Shomer,P. and Fry,M. (2005) Formation and properties of hairpin and tetraplex structures of guanine-rich regulatory sequences of muscle-specific genes. *Nucleic Acids Res.*, **33**, 2887–2900.
28. Zhao,Y., Du,Z. and Li,N. (2007) Extensive selection for the enrichment of G4 DNA motifs in transcriptional regulatory regions of warm blooded animals. *FEBS Lett.*, **581**, 1951–1956.
29. Bugaut,A. and Balasubramanian,S. (2008) A sequence-independent study of the influence of short loop lengths on the stability and topology of intramolecular DNA G-quadruplexes. *Biochemistry*, **47**, 689–697.
30. Kumar,N., Sahoo,B., Varun,K.A. and Maiti,S. (2008) Effect of loop length variation on quadruplex-Watson Crick duplex competition. *Nucleic Acids Res.*, **36**, 4433–4442.
31. Lee,J.Y. and Kim,D.S. (2009) Dramatic effect of single-base mutation on the conformational dynamics of human telomeric G-quadruplex. *Nucleic Acids Res.*, **37**, 3625–3634.
32. Gros,J., Rosu,F., Amrane,S., De Cian,A., Gabelica,V., Lacroix,L. and Mergny,J.L. (2007) Guanines are a quartet's best friend: impact of base substitutions on the kinetics and stability of tetramolecular quadruplexes. *Nucleic Acids Res.*, **35**, 3064–3075.
33. Blake,R.D., Hess,S.T. and Nicholson-Tuell,J. (1992) The influence of nearest neighbors on the rate and pattern of spontaneous point mutations. *J. Mol. Evol.*, **34**, 189–200.
34. Fryxell,K.J. and Moon,W.J. (2005) CpG mutation rates in the human genome are highly dependent on local GC content. *Mol. Biol. Evol.*, **22**, 650–658.
35. Hodgkinson,A., Ladoukakis,E. and Eyre-Walker,A. (2009) Cryptic variation in the human mutation rate. *PLoS Biol.*, **7**, e27.
36. Krawczak,M., Ball,E.V. and Cooper,D.N. (1998) Neighboring-nucleotide effects on the rates of germ-line single-base-pair substitution in human genes. *Am. J. Hum. Genet.*, **63**, 474–488.
37. Sherry,S.T., Ward,M.H., Kholodov,M., Baker,J., Phan,L., Smigielski,E.M. and Sirotkin,K. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–311.
38. Karolchik,D., Baertsch,R., Diekhans,M., Furey,T.S., Hinrichs,A., Lu,Y.T., Roskin,K.M., Schwartz,M., Sugnet,C.W., Thomas,D.J. *et al.* (2003) The UCSC Genome Browser Database. *Nucleic Acids Res.*, **31**, 51–54.
39. Eddy,J. and Maizels,N. (2009) Selection for the G4 DNA motif at the 5′ end of human genes. *Mol. Carcinog.*, **48**, 319–325.
40. Eddy,J. (2007) Conserved elements with potential to form polymorphic G-quadruplex structures in the first intron of human genes. *Nucleic Acids Res.*, **36**, 1321–1333.
41. Tomso,D.J. and Bell,D.A. (2003) Sequence context at human single nucleotide polymorphisms: overrepresentation of CpG dinucleotide at polymorphic sites and suppression of variation in CpG islands. *J. Mol. Biol.*, **327**, 303–308.
42. Eddy,J. and Maizels,N. (2006) Gene function correlates with potential for G4 DNA formation in the human genome. *Nucleic Acids Res.*, **34**, 3887–3896.
43. Huppert,J.L., Bugaut,A., Kumari,S. and Balasubramanian,S. (2008) G-quadruplexes: the beginning and end of UTRs. *Nucleic Acids Res.*, **36**, 6260–6268.
44. Todd,A.K. and Neidle,S. (2008) The relationship of potential G-quadruplex sequences in cis-upstream regions of the human genome to SP1-binding elements. *Nucleic Acids Res.*, **36**, 2700–2704.
45. Verma,A., Halder,K., Halder,R., Yadav,V.K., Rawal,P., Thakur,R.K., Mohd,F., Sharma,A. and Chowdhury,S. (2008) Genome-wide computational and expression analyses reveal G-quadruplex DNA motifs as conserved cis-regulatory elements in human and related species. *J. Med. Chem.*, **51**, 5641–5649.
46. Bird,A.P. (1986) CpG-rich islands and the function of DNA methylation. *Nature*, **321**, 209–213.
47. Sakumi,K., Furuichi,M., Tsuzuki,T., Kakuma,T., Kawabata,S., Maki,H. and Sekiguchi,M. (1993) Cloning and expression of cDNA for a human enzyme that hydrolyzes 8-oxo-dGTP, a mutagenic substrate for DNA synthesis. *J. Biol. Chem.*, **268**, 23524–23530.
48. Bjoras,M., Luna,L., Johnsen,B., Hoff,E., Haug,T., Rognes,T. and Seeberg,E. (1997) Opposite base-dependent reactions of a human base excision repair enzyme on DNA containing 7,8-dihydro-8-oxoguanine and abasic sites. *EMBO J.*, **16**, 6314–6322.
49. Nash,H.M., Bruner,S.D., Scharer,O.D., Kawate,T., Addona,T.A., Spooner,E., Lane,W.S. and Verdine,G.L. (1996) Cloning of a yeast 8-oxoguanine DNA glycosylase reveals the existence of a base-excision DNA-repair protein superfamily. *Curr. Biol.*, **6**, 968–980.
50. Slupska,M.M., Baikalov,C., Luther,W.M., Chiang,J.H., Wei,Y.F. and Miller,J.H. (1996) Cloning and sequencing a human homolog (hMYH) of the Escherichia coli mutY gene whose function is required for the repair of oxidative DNA damage. *J. Bacteriol.*, **178**, 3885–3892.
51. McCulloch,S.D., Kokoska,R.J., Garg,P., Burgers,P.M. and Kunkel,T.A. (2009) The efficiency and fidelity of 8-oxo-guanine bypass by DNA polymerases {delta} and {eta}. *Nucleic Acids Res.*, **37**, 2830–2840.
52. Kresnak,M.T. and Davidson,R.L. (1992) Thymidine-induced mutations in mammalian cells: sequence specificity and implications for mutagenesis in vivo. *Proc. Natl Acad. Sci. USA*, **89**, 2829–2833.
53. Platzer,M., Hiller,M., Szafranski,K., Jahn,N., Hampe,J., Schreiber,S., Backofen,R. and Huse,K. (2006) Sequencing errors or SNPs at splice-acceptor guanines in dbSNP? *Nat. Biotechnol.*, **24**, 1068–1070.
54. Siva,N. (2008) 1000 Genomes project. *Nat. Biotechnol.*, **26**, 256.

III

Impact of DNA physical properties on local sequence bias of human mutation

Sigve Nakken[1], Einar A. Rødland[2,3] and Eivind Hovig[1,2,4,]*

[1]Department of Tumor Biology, Institute for Cancer Research, Oslo University Hospital - Norwegian Radium Hospital, Norway

[2]Department of Informatics, University of Oslo, Norway

[3]Centre for Cancer Biomedicine, University of Oslo, Norway

[4]Department of Medical Informatics, Oslo University Hospital – Rikshospitalet, Norway

*Corresponding author: Prof. Eivind Hovig
Department of Tumor Biology
Institute for Cancer Research
Oslo University Hospital, Norwegian Radium Hospital
Montebello, 0310 Oslo, Norway
Email: ehovig@ifi.uio.no
Telephone: +47 22 78 17 78
Fax: +47 22 52 24 21

Abstract

In selectively neutral regions of the human genome, nucleotide substitutions do not occur at random with respect to the local DNA sequence neighbourhood. Sequence specificities present in DNA replication, damage and repair factors are considered to be the main underlying causes of this non-randomness. However, apart from the hypermutability of methylated CpG dinucleotides, which can explain the overrepresentation of nucleotide transitions in this context, the sequence-specific factors underlying the point mutation bias remain largely to be determined, both in nature and in quantitative impact. One hypothesis suggests that the physical characteristics of a DNA context could have a modulating effect on its mutability, adjusting the impact of damage or the efficiency of repair. Here, we report a genome-wide computational test of this hypothesis, in which we utilize a constrained set of human non-CpG SNPs as the source of selectively neutral germline mutations. Through a simple statistic that compares the frequency of a DNA *k*-mer at SNP sites with its frequency in the genomic background, we estimate the relative impact of context on nucleotide substitution rate. Interestingly, we observe that the quantitative context-dependencies of some substitution types display significant associations to measures of local structural topography and helix stability in DNA. Most prominently, we find that the local sequence bias of transition mutations is significantly associated with the sequence-dependent level of helix instability imposed by the underlying DNA mismatches. The associations are supported in an independent analysis of fixed nucleotide substitutions inferred in the human lineage. The results of our work indicate the extent to which DNA physical properties could have shaped the recent point mutational spectrum in the human genome.

Key Words: single nucleotide polymorphisms; mutagenesis; sequence context; hydroxyl radical cleavage; local DNA stability; DNA bendability

Introduction

In selectively neutral regions of the human genome, two primary forces shape the spectrum of nucleotide substitutions. These are the mutational input, represented by base damage and spontaneously occurring base misincorporations during DNA synthesis, and the DNA repair machinery that acts on this input, producing the net mutational output. Considering the joint contributions by DNA mutagens that target specific sequence contexts [Shibutani et al., 1999; Cloutier et al., 2001a; Cloutier et al., 2001b; Rochette et al., 2003], and DNA repair enzymes with marked sequence specificity [Eftedal et al., 1993; Sibghat-Ullah et al., 1996; Mazurek et al., 2009], it is not entirely unexpected that the spectrum of human germline mutations exhibits biases with respect to its local DNA sequence neighbourhood [Krawczak et al., 1998; Zhao, 2002; Tomso and Bell, 2003; Jiang and Zhao, 2006; Zhao and Zhang, 2006]. In addition to the well-known mutation bias at CpG dinucleotides, which primarily has been explained by a high incidence of spontaneous deamination of 5-methylcytosine [Duncan and Miller, 1980; Ehrlich and Wang, 1981], subtle, yet notable mutation biases have been observed also for other contexts [Hess et al., 1994; Siepel and Haussler, 2004]. The underlying biological factors that generate local sequence biases for non-CpG substitutions have nevertheless been hard to identify.

Recently, it was discovered that DNA cleavage patterns from the hydroxyl radical (·OH) correlated significantly with context-dependent mutation rates in mammals [Stoltzfus, 2008]. Since hydroxyl radicals represent a major DNA damage source in human cells, it was argued that oxidative damage could be a predominant factor underlying the sequence bias of human mutations. However, damage is merely a step towards mutation, and can thus only explain parts of the observed mutation bias. Interestingly enough, one should note that ·OH cleavage patterns are tightly linked to the local shape of the DNA molecule, as it mirrors the solvent-accessible surface of the DNA backbone [Balasubramanian et al., 1998]. One could in principle hypothesize that, to some extent, it is the intrinsic thermodynamical and conformational properties of different DNA sequence contexts that dictate the impact of damage or the efficiency of DNA repair [Rajski et al., 2000; Isaacs and Spielmann, 2004]. With respect to damage, it has been shown that varying levels of thermal motion within the double helix determine the extent of fluctuational basepair openings, making normally buried groups accessible for interaction with proteins, chemicals and potential mutagens [Frank-Kamenetskii, 1987; Gueron et al., 1987]. Perhaps more important is the observation that the different base mismatches in DNA, occurring naturally as a result of base misincorporation during replication, display significant context-dependency with respect to helix stability

[Allawi and SantaLucia, 1997; Allawi and SantaLucia, 1998b; Allawi and SantaLucia, 1998c; Allawi and SantaLucia, 1998a; Peyret et al., 1999]. Thus, if local helix stability affects the efficiency of mismatch recognition or repair, we would expect to see a trend in which the likelihood of mutation varies according to context stability. The role of local thermodynamics in the context of primer-template misalignment propensity during DNA synthesis has furthermore been highlighted in recent NMR studies [Chi and Lam, 2007; Chi and Lam, 2009].

Another important physical aspect of DNA in the context of damage and repair is its bending propensity. Different sequence segments of DNA exhibit varying levels of bendability, with obvious implications for DNA packaging, and perhaps also the interaction with several DNA repair proteins [Drew and Travers, 1985]. In fact, it has been shown that local DNA flexibility could modulate the efficiency of both base excision repair as well as mismatch repair enzymes [Seibert et al., 2002; Seibert et al., 2003; Wang et al., 2003]. Finally, there is reason to believe that the different physical aspects of DNA will influence each other, for instance that the local bendability of DNA relates to the solvent-accessible area of the backbone. To our knowledge, however, the complexity and biological impact of sequence-dependent physical properties of DNA remains to be fully explored, particularly within the context of mutation bias in the human genome.

The aim of this study was to add more quantitative information to the potential sources of human mutation bias. A genome-wide collection of strictly filtered human single nucleotide polymorphisms (SNPs) was used as the data source of selectively neutral germline mutations. For each of the different types of nucleotide substitutions, we applied a simple statistic to estimate the relative impact of local sequence context on the incidence of substitution. In order to focus our analysis towards mutation biases not attributable to CpG, we excluded this dinucleotide in our computations. Further, we utilized available experimental data with respect to DNA physics to estimate three quantities for any given sequence context: 1) mean relative intensity of hydroxyl radical cleavage, 2) thermodynamic stability (ΔG), and 3) mean relative DNA bendability. Our data indicates that the local sequence biases of particular substitution types are significantly associated with DNA physical properties. Importantly, we replicate our main findings in an independent set of computationally inferred, fixed nucleotide substitutions in the human lineage.

Materials and Methods

SNP and sequence data

A neutral spectrum of nucleotide substitutions in the recent human genome was created based on 1) a strictly defined set of presumably neutrally evolving genomic regions, and 2) a strictly filtered set of SNPs within these regions.

To approximate neutrally evolving regions of the human genome, we targeted genomic regions that are not associated with known functional constraints. For this, we utilized the combination of publicly available genome annotation tracks at UCSC [Karolchik et al., 2003], and the Galaxy data management tool [Giardine et al., 2005]. Using the full human genome sequence as a starting point, we excluded constrained regions in several consecutive filtering steps. With respect to genes, we used the RefSeq database as our data source (NCBI build 37.1). Regions excluded were coding exons, potential regulatory regions (5kb upstream and downstream of transcription start sites), untranslated regions, and intronic splice sites (30bp). We furthermore excluded highly conserved elements [Siepel et al., 2005], CpG islands [Gardiner-Garden and Frommer, 1987], and transcriptional enhancers [Pennacchio et al., 2006]. Considering that recent recombination hotspots in the genome are associated with distinct patterns of nucleotide substitutions and particular mechanisms of mutation fixation (e.g. biased gene conversion [Spencer et al., 2006]), we decided to exclude also these regions in our main analysis. A total of 32,991 recombination hotspots inferred from genotype data in phase II of the HapMap project were thus excluded [International HapMap Consortium, 2007]. Finally, we excluded segmental duplications [Bailey et al., 2002], as SNPs residing in these regions are potentially unreliable [Fredman et al., 2004; Musumeci et al., 2009; Nakken et al., 2009]. Within the final set of filtered genomic regions, we defined high-copy repeats to be all sequences that were identified by RepeatMasker and Tandem Repeats Finder (with period of 12 or less), a definition we adopted from the UCSC Genome Browser resource [Karolchik et al., 2003].

dbSNP (build 130, released on April 30, 2009) was downloaded in XML format from NCBI (ftp://ftp.ncbi.nlm.nih.gov/snp/). All biallelic SNPs were extracted along with their neighbouring (+/- 50bp) DNA sequence context (we used the *liftOver* tool in order to work with annotations in the hg19 release of the human genome). We included SNPs that *i)* had been uniquely mapped to the genome with an alignment accuracy of at least 99%, *ii)* had been validated by at least one of NCBI's validation criteria (that is, "by-frequency", "byCluster", "by2Hit2Allele" or "byOtherPop"), and *iii)* if genotyped by the HapMap project, had a minor

allele frequency of at least 1% in minimum one of the sampled populations. We used available data from the UCSC genome browser to infer the ancestral and derived allele of each SNP (i.e. providing directionality of the nucleotide substitution). Specifically, we compared the SNP alleles with the orthologous chimpanzee base (panTro2), employing a maximum parsimony principle with the orangutan (ponAbe2) as outgroup [Karolchik et al., 2003; Jiang and Zhao, 2006]. We excluded SNPs that were annotated with ambiguous or unknown substitution directionalities, and those that originated within an ancestral CpG context.

Fixed substitutions in the human lineage

We established an independent, genome-wide set of human nucleotide substitutions by means of a comparative genomics approach. Specifically, within the neutrally evolving regions defined as above, we downloaded MultiZ sequence alignments of human (hg19), chimpanzee (panTro2) and orangutan (ponAbe2) [Blanchette et al., 2004]. We then inferred fixed nucleotide substitutions in the human lineage through parsimony, using the orangutan sequence as outgroup. We furthermore required a minimum of two bases flanking each side of a putative substitution to have a quality Phred score of 20 or higher in both the chimpanzee and orangutan draft genomes (a requirement adopted from [Kvikstad et al., 2007]). Inferred substitutions that overlapped with known polymorphic sites in dbSNP were excluded, as well as those originating within an ancestral CpG context. Finally, we restricted the set of fixed nucleotide substitutions to those that occurred within 1kb sequence windows centred at the main set of non-CpG SNPs. The latter step was done to ensure that SNPs and the independent set of fixed substitutions were comparable, in the sense that both sets were sampled from the same genomic environments, associated with the same approximate mutation rate. Only the sequences that had a quality Phred score of 20 or higher in both the chimpanzee and orangutan draft genomes were used for measuring background frequencies in the estimation of mutation bias (see below).

Estimation of local sequence bias of human mutation

In order to measure the relative impact of local sequence context on the rate of nucleotide substitutions, we compared the frequencies of short DNA sequence contexts (*k*-mers) at substitution sites with their corresponding frequencies in the genomic background. This general approach was applied to two independent sets of nucleotide substitutions: those inferred from human SNPs (main analysis), and those inferred through comparative genomics (validation analysis).

If a nucleotide substitution occurs independently of its neighbouring nucleotides, we would expect the ratio of substitution to background frequency to be approximately constant for any given context (i.e. any combination of neighbouring nucleotides). As an assessment of relative context-dependency, we defined *rmf (relative mutation fraction)* formally as follows:

$$rmf\left(b_1[b \rightarrow b']b_2\right) = \left(\frac{n\left(b_1[b \rightarrow b']b_2\right)}{n\left(b_1 b b_2\right)} \Big/ v \right)$$

Here, b represents the base that undergoes mutation to a base other than b (b'), b_1 represents the upstream sequence context, and b_2 represents the downstream sequence context. $n\left(b_1 b b_2\right)$ denotes the estimated background frequency of the $b_1 b b_2$ sequence context, and $n\left(b_1[b \rightarrow b']b_2\right)$ denotes the total frequency of $b \rightarrow b'$ substitutions observed in the $b_1 b b_2$ sequence context. The estimated background frequencies (i.e. $n\left(b_1 b b_2\right)$) were found using a sliding window procedure in 1kb sequence windows centred at the substitution sites. Each *rmf* ratio was normalized by a constant factor v, which denotes the ratio of total substitution frequencies to total background frequencies (i.e. global mutation fraction). The *rmf* estimate of a given context was thus expressed relative to the global mutation fraction. We estimated the variance in each *rmf* point estimate through $n\left(b_1[b \rightarrow b']b_2\right)$, which we assume is following a Poisson distribution.

As for the length of the local sequence context, we analyzed both DNA trimers ($length(b_1) = length(b_2) = 1$, $k = 3$), and DNA pentamers ($length(b_1) = length(b_2) = 2$, $k = 5$). Ideally, one should investigate a larger and more complete range of *k*-mers, since the extent of DNA context-dependence for human mutations is not completely clear [Krawczak et al., 1998; Siepel and Haussler, 2004]. Note however that the number of possible *k*-mers increases substantially with the size of *k*, and that many *rmf* point estimates thus will become unreliable

due to few observed substitutions per *k*-mer. An analysis of context-dependence at the level of DNA heptamers (*k*=7*)* will consequently require a larger set of substitutions. The primary reason for choosing trimers (and pentamers) in this study was nevertheless that the experimentally determined estimates of DNA mismatch stability, local structural topography, and DNA bendability all were defined at the level of trimers.

Generally, we did not distinguish between a sequence context and its reverse complementary context. We thus merged reverse complementary contexts (and substitutions) during analysis. To test for potential differences in the relative mutation fraction of a context and its reverse complementary context, we mapped all intronic substitutions to the nontemplate strand of their corresponding transcripts and calculated the *rmf* correlation between reverse complementary contexts (for each pair of reverse complementary substitutions). When measuring *rmf* in introns, background *k*-mers were sampled from the nontemplate strand only.

Statistical analysis

All *rmf* point estimates were transformed to a logarithmic scale, in that sense transforming relative differences in mutation fractions to an additive scale, and providing more symmetric *rmf* distributions. Pearson's product-moment correlation coefficient was used as the measure of correlation in our analyses.

Analysis of variance (ANOVA) with respect to the logarithm of *rmf* was used to determine how much of the *rmf* variance originated within substitution types versus between substitution types. It should be emphasized that statistical tests were not based on uncertainties in *rmf* point estimates, but rather on general variance in *rmf* between *k*-mers. The statistical tests for ANOVA and regression analyses assume that residuals are independent, but this is most likely not entirely the case as similar contexts must be expected to be correlated beyond what the models capture. We analyzed the variance for *rmf* at the DNA pentamer level by including the effects of the central trimers in the model.

Multiple linear regression models were used to determine the dependency of *rmf* on different combinations of DNA physical properties. All possible submodels, with four covariates, that make 15 models, were analyzed. In order to account for multiple testing, Simes' procedure was used [Simes, 1986]. Where Simes' p-value is statistically significant, the best model is cited. In our case, the significant Simes' p-values were identical to the Bonferroni corrected p-values, although Simes' procedure is in general less conservative.

Local stability and bendability properties of DNA

In double-stranded DNA, two factors govern the stability of the helix, Watson-Crick base pairs and the base stacking interactions along the helical axis. Specifically, it has been demonstrated that a nearest-neighbour (NN) model is sufficient to accurately predict the thermodynamic properties of a DNA oligomer duplex. Thus, by using an experimentally established thermodynamic library of the ten different DNA nearest-neighbour pairs, one can approximate the local free energy (stabilization energy, ΔG) of a DNA sequence context as the sum of the different nearest-neighbour interactions in the context. Importantly, the stabilization energy is not merely a function of GC content, e.g. 5'-*GCA*-3' is more stable than 5'-*GGG*-3'. Here, we used the unified library published by SantaLucia et al. to estimate ΔG for DNA *k*-mers [SantaLucia, 1998]. Similarly, we used libraries from five different studies to estimate ΔG for DNA *k*-mers with internal mismatches [Allawi and SantaLucia, 1997; Allawi and SantaLucia, 1998b; Allawi and SantaLucia, 1998c; Allawi and SantaLucia, 1998a; Peyret et al., 1999]. We ignored the initiation and symmetry contributions, and normalized ΔG by the number of nearest-neighbour interactions in the *k*-mer. All values of ΔG used in the analysis are thus of type kcal/mol pr. nearest-neighbour interaction (kcal/mol/nn).

We estimated the relative bendability of DNA three-mers from a consensus scale that was derived from DNase 1 cutting and nucleosome positioning experiments [Gabrielian and Pongor, 1996]. For a given DNA five-mer, the relative bendability was judged as the average of the bendability from its three overlapping three-mers. Relative bendabilities were transformed to the range [0,1], where 0 corresponded to the most rigid and 1 to the most bendable.

Hydroxyl radical cleavage intensity in DNA

Data from the ORCHiD (OH Radical Cleavage Intensity Database) database reveals consistent patterns of hydroxyl radical cleavage for DNA three-mers. A sliding-window algorithm was sufficient for highly accurate (R^2 = 0.91) predictions of hydroxyl radical cleavage intensities in genomic DNA [Greenbaum et al., 2007]. We downloaded the hydroxyl radical cleavage intensities that were predicted from ORCHiD in the ENCODE regions, and calculated the mean intensity of cleavage for the central nucleotide of all DNA three-mers and five-mers (normalized to the range [0,1]). Reverse complementary contexts were combined into one category, its intensity of cleavage corresponding to the average of the two contexts.

We assessed the quantitative relationship between *rmf* and hydroxyl radical cleavage

intensity within ENCODE regions before extrapolation genome-wide. The low number of SNPs in neutrally evolving ENCODE regions (n=28,300) did however make *rmf* estimation unreliable here. The results obtained (not shown) did nevertheless indicate that ENCODE regions were representative of the full genome.

Results and Discussion

The local sequence bias of human germline mutations has been recognized in many different studies [Krawczak et al., 1998; Zhao, 2002; Tomso and Bell, 2003; Jiang and Zhao, 2006; Zhao and Zhang, 2006]. With the exception of the hypermutable CpG dinucleotide, it has nevertheless been challenging to pinpoint factors that might explain the observed biases. Here, we attempt to provide additional insight into this topic by exploring the relationships between local topographical patterns in DNA, the stability and bendability properties of DNA, and the relative incidence of human SNPs.

We established a selectively neutral mutation spectrum in the human genome by querying the dbSNP database at a large set of genomic coordinates that were filtered for functionally constrained sequences. The collection of validated SNPs that mapped to the neutrally evolving regions was further annotated with ancestral and derived alleles, thereby identifying directionalities of the underlying nucleotide substitutions. The relative mutation fraction (*rmf*) of different sequence contexts were next estimated based on the combination of *k*-mer counts in the targeted genomic regions and the distribution of nucleotide substitution contexts inferred from SNPs.

Initially, we observed that the density of validated, non-CpG SNPs was significantly lower in the defined high-copy repeats compared to non-repetitive sequence (1.71 SNPs/kb versus 2.12 SNPs/kb, χ^2=49340, df=1, p<2.2e-16). We decided to limit our analysis to non-repetitive sequence only. The reason for this choice was not only the low density of SNPs in repeats, but also because repeats are likely to have a skewed sequence composition compared to unique regions. In addition, repeats are relatively difficult to assemble, making these regions less suitable for sequence analysis. In total, we analyzed 2,174,291 validated non-CpG SNPs with inferred substitution directionality, extracted from approximately 1,025 Mb of non-repetitive sequence in the human genome. The CpG filter excluded approximately 12.5% of all validated SNPs that mapped to the non-repetitive target sequences in the neutrally evolving regions.

The box-whisker plots in Figure 1 illustrate the distributions of *rmf* for the six

different types of nucleotide substitutions. First, note that the distributions of *rmf* for the two transition types generally lie to the right of transversion types, confirming that nucleotide transitions dominate the human mutational spectrum. It should also be mentioned that all substitutions of cytosine or guanine (C>*/G>*) had fewer context observations (~25%) than substitutions of adenine or thymine (A>*/T>*), since we excluded the CpG contexts in our analysis. The box-whisker plots support the general point that the incidence of nucleotide substitution varies according to local sequence context. We next carried out several analyses using ANOVA to further characterize the observed *rmf* distributions.

Variation in *rmf*, i.e. var[log(*rmf*)] computed per substitution type, was not found to differ between substitution types (Levene's test, p=0.10). We next discovered that the relative mutation fraction of the 28 non-CpG DNA trimers (e.g. AAA, AAT etc.) was primarily determined by the rate of the different types of nucleotide substitutions. Specifically, we found that 81.8 % of the *rmf* variance for DNA trimers could be attributed to differences in rates between the substitution types. This number contrasts to the 2% of *rmf* variance that could be explained by general differences in terms of influence by neighbouring bases (i.e. upstream and downstream) on mutability. The remaining ~16% of *rmf* variance could be explained by specific interactions between the substitution type and its two neighbouring bases. The variance caused by the estimation uncertainty in log(*rmf*) was aggregated and compared to the sum of squares from the ANOVA, and indicated that this uncertainty could be ignored in the analyses (e.g. 0.01% of total variance for DNA trimers). For DNA pentamers, we assessed whether the two outermost bases contributed significantly to *rmf* (given the central trimer), and found that this contribution was non-significant (p=0.68). Thus, although some pentamer contexts could be of particular importance in terms of mutation bias, we will emphasize results at the DNA trimer level in the remaining part of the discussion.

During *rmf* estimation, we did not distinguish between a context and its reverse complementary context. Two reverse complementary contexts were combined into one single category. This is generally unproblematic if there are no significant differences in damage intensity or repair efficiency between the two DNA strands, as presumably is the case in most intergenic regions of the human genome. However, 28.8% (295 Mb) of the analyzed regions originated within gene transcripts (i.e. introns) and harboured 26.0% (565,314 SNPs) of all SNP-inferred nucleotide substitutions. Several studies have demonstrated that there exist mutational strand asymmetries in transcriptional regions, and it is believed that these occur due to transcription-coupled repair [Green et al., 2003; Mugal et al., 2009]. However, it is less certain if the observed substitution rate asymmetries also yield asymmetries with respect to

the sequence contexts that undergo mutation. In particular, a rate asymmetry does not necessarily imply asymmetry at the context level. To check for potential differences in mutation fractions of reverse complementary contexts, we mapped 565,314 SNPs in human gene introns to the nontemplate strand, and correlated *rmf* estimates (DNA trimers) of reverse complementary substitution contexts. We generally found strong positive associations for most substitution types (Supplementary Table 1). The reverse complementary contexts of C>T(G>A) transitions stood clearly out and displayed the relatively weakest and least significant correlation (r=0.79, p=2.1e-3). This could suggest that some of the mechanisms that underlie strand asymmetries for this transition type, being either asymmetric damage intensities or repair efficiencies (or both), are context-specific. We surmised nevertheless that on a genome-wide scale, mutational strand asymmetries in transcripts did not generate substantial differences in the *rmf* of reverse complementary DNA trimers.

Role of local helix stability

Table 2 shows how the relative mutation fraction of different DNA sequence contexts correlated to DNA physical property scales. Considering all possible combinations of substitution-specific *rmf* estimates and DNA physical properties, there were generally few significant findings for DNA trimers (Table 1). Some substitution types displayed nevertheless significant associations between *rmf* and DNA physical properties, and were noteworthy.

Overall, the local stabilization energy of the ancestral context (i.e. stability of the DNA sequence targeted for mutation, $(\Delta G_{ancestral})$) showed few significant associations to the relative mutation fraction of DNA trimers. An exception was for C>G (G>C) substitutions, in which we discovered a positive association between the stabilization energy of the ancestral context and *rmf* (r = 0.63, p=2.8e-2, see Figure 2). This association was also significant, though not as strong, when we considered the surrounding DNA five-mer (r = 0.40, p=7.8e-8). Based on these data, it thus appears as if local DNA stability of the target sequence context may explain some of the non-random distribution of C>G (G>C) transversions. We are however not aware of particular types of DNA damage sources or repair proteins that are specifically associated with C>G(G>C) transversions, and whose mode of operation is dependent upon local helix stability.

A model of physical polymerase slippage, in which the presence of repetitive sequences causes transient misalignment of the primer and template strands, may explain some nucleotide misincorporations that occur during DNA replication [Kunkel and

Alexander, 1986]. Studies have further suggested that the propensity of slippage could be influenced by thermodynamic stability in the local primer-template alignment [Chi and Lam, 2006; Chi and Lam, 2009]. We examined whether there were associations between the *rmf* of slippage-like mutation contexts (i.e. contexts in which the newly introduced base is identical to either of the two neighbouring bases) and local stability. No significant relationships of such kind (r = -0.19, p=0.26, all substitution types combined) were found. One potential reason for the lack of correlation is simply that the impact of *in vivo* replication slippage could be more relevant to the spectrum of insertions/deletions than single point mutations [Cooper and Krawczak, 1993].

Role of local DNA mismatch stability

Mismatched nucleotides in DNA can occur as a result of replication errors, heteroduplex formation during homologous recombination, or from spontaneous base deamination events [Hunter et al., 1987; Jiricny, 1998]. Thus, the relative number of nucleotide substitutions originating as non-Watson-Crick basepairs may have been substantial in recent evolution of the human genome.

The possible base mispairs that can form within DNA display significant differences with respect to thermodynamic stability, and it has been suggested that these physical differences could influence mismatch recognition, and thereby the local incidence of mutation [Rajski et al., 2000; Isaacs and Spielmann, 2004]. In an effort to test this hypothesis, we used available thermodynamic libraries to estimate the context-dependent level of destabilization caused by a DNA mispair. As an example, for a nucleotide substitution context A[A>G]A (reverse complementary T[T>C]T), we measured the absolute change in stabilization energy between the ancestral Watson-Crick trimer context (AAA/TTT), and the average of the two potentially underlying mispair contexts (A<u>G</u>A/T<u>T</u>T and T<u>C</u>T/A<u>A</u>A, (mispair underlined)). We found significant negative associations between this measure *($\Delta G_{mispair}$* in Table 1*)* and *rmf* of both transition types: A>G/T>C (r = -0.76, p=6.7e-4) and C>T/G>A (r = -0.63, p=3.0e-2). Note that both types of transition types are associated with the same combination of potentially underlying mispairs, that is T:G and A:C. If we represent the detected statistical relationships in terms of linear regression models, the explanatory value of mispair destabilization to *rmf* variance was significant (R^2=0.57 and R^2=0.39, see Figure 2). Thus, our data could be interpreted as if the context-dependent level of helix instability induced by DNA mismatches may explain some 40-60% of the local sequence bias of nucleotide transitions. This relationship may have significant impact on mutation bias in general,

considering that transitions make up 64.2% of all nucleotide substitutions in the inferred non-CpG mutational spectrum.

The impact of native replication errors on the human mutational spectrum is jointly influenced by i) the probability of initial misinsertion by the DNA polymerase, ii) efficiency of exonucleolytic proofreading activity, and iii) the efficiency of mismatch recognition and repair by the mismatch repair pathway (MMR) [Maki, 2002]. In contrast to the other primary underlying source of nucleotide substitutions, that is chemically altered bases (e.g. by oxidation or methylation), naturally occurring mispairs consist of ordinary bases only, which challenges the operation of recognition and repair. Intuitively, the strong negative correlation between relative mutation fractions of nucleotide transitions and the local helix instability imposed by the underlying mispairs may support a role of DNA thermodynamics in mismatch recognition and repair. Specifically, it suggests that the most thermodynamically unstable mispair contexts are more readily repaired than those that reduce stability to minor extents. The structural perturbations by both T:G and A:C mismatches have been shown to be small and very localized in the double helix [Hunter et al., 1986; Hunter et al., 1987]. However, both T:G and A:C mispairs appear to be the most efficiently repaired mismatched basepairs, as observed *in vitro* in monkey kidney cells [Brown and Jiricny, 1988]. Even though mispair repair efficiencies *in vivo* could be quantitatively different, we suggest that an alternative explanation of the observed association may come from the mismatch extension process by the DNA polymerase. Some mispairs are more readily extended than others (notably G:T), and it may be that the levels of imposed instability influence the likelihood of extension [Goodman et al., 1993; Kunkel, 2004].

DNA mismatches frequently form at sites of meiotic recombination. First of all, the necessity for double-strand breaks (DSB) requires subsequent DNA synthesis, which may introduce errors and thus novel substitutions [Strathern et al., 1995]. Secondly, the pairing of strands from homologous chromosomes that occur during DSB repair can lead to mismatches at heterozygous loci. A biased repair of the latter mismatches may influence the fate of existing mutations, a mechanism known as gene conversion [Duret and Galtier, 2009]. In order to test for the impact of mispair stability on sequence bias of recombination-associated nucleotide transitions, we conducted an exclusive analysis of *rmf* in 32,991 recent recombination hotspots in the human genome. These hotspots were excluded in our main set of neutrally evolving regions. Estimates based on 197,431 SNPs from neutrally evolving sequences of recombination hotspots produced roughly the same associations to trimer

mispair instability that were observed genome-wide (A>G(T>C): r = -0.74 (p=9.4e-3); C>T(G>A): r = -0.65 (p=2.3e-2).

Role of local DNA topography – intensity of cleavage by the hydroxyl radical

For the majority of nucleotide substitution types, we found that *rmf* was positively associated with ·OH cleavage intensity, though weakly and primarily significant for DNA five-mers (Table 1). Our data are thus in agreement with the main finding of a recent study [Stoltzfus, 2008], which suggested that the intensity of oxidative damage could have shaped much of the local sequence bias observed in mammalian mutation spectra. This association also highlights a link between local DNA conformation and the incidence of mutation, considering that the solvent-accessible area of the DNA backbone modulates the level of attack from the hydroxyl radical [Balasubramanian et al., 1998]. For our data, we observed the strongest positive association when we considered the combined contribution by all substitution types (r=0.60, p=7.5e-4, *DNA trimers*). This association was somewhat weaker than the result reported by Stoltzfus *et al.* ($r \approx 0.70$, p=2.4e-5). A couple of methodological factors could potentially underlie this discrepancy. As opposed to Stoltzfus *et al*., who calculated the correlation using context-dependent mutation rates inferred from evolutionary divergence of mammalian non-coding regions [Hwang and Green, 2004], our results were derived solely from the local bias of SNPs in the human lineage. Also, in order to analyze both DNA trimers and pentamers, we utilized the predicted intensity patterns in the ENCODE regions rather than the raw experimental data from ORCHiD.

Role of DNA bendability

The bendability of a sequence context represents an important property of DNA, and appears central in several aspects related to human mutation. One aspect concerns the sequences that resist the sharp bending required by the nucleosome [Segal and Widom, 2009]. A negative preference for nucleosome occupancy would potentially leave these rigid sequences more accessible to DNA mutagens. Studies in fish have on the other hand suggested that the linker regions between nucleosome cores exhibit a decrease in the mutation rate [Sasaki et al., 2009]. It is also clear that local DNA bendability may be an important modulator of DNA repair efficiency. Within the family of MutS proteins, which recognize base-base mismatches in DNA, it has been shown that specific mismatch recognition appears in structural conformations where the DNA is unbent [Wang et al., 2003]. Another study has shown that the sequence-dependent efficiency of uracil DNA glycosylase could be explained in terms of

the DNA flexibility surrounding the substrate [Seibert et al., 2002]. In order to explore the role of bendability with respect to mutation bias, we quantified the relationship between local bendability and relative mutation fractions. Since bendability in principle could be important both with respect to the ancestral context (i.e. damage) and the derived context (i.e. repair), we looked at the average bendability of the ancestral and derived context in relation to *rmf*. At the DNA trimer level, we found no significant associations between this property and the context biases of specific nucleotide substitutions (Table 1). For DNA pentamers, we observed the most significant associations for the two nucleotide transition types, which displayed a weak positive association between *rmf* and relative DNA bendability. Overall, our data did not indicate a predominant role of bendability in shaping the sequence bias of human mutations. A caveat of our analysis is that the bendability of native DNA contexts may not accurately reflect the true bendability of pre-mutagenic lesions, which is the primary substrate of repair enzymes. There is in fact some evidence that both endogenous DNA lesions and mismatched sites will alter the local DNA flexibility [Marathias et al., 1999].

Validation of associations within fixed lineage substitutions

Although the SNP set was carefully filtered, we assumed that it nevertheless could contain some minor unknown, technological biases contributed for instance by sequencing errors [Nelson et al., 2004]. Such biases may lead to spurious statistical associations that do not reflect the underlying biology of germline mutagenesis. We therefore considered it important to validate the observed SNP associations using an independent set of nucleotide substitutions.

We inferred a total of 2,066,656 fixed nucleotide substitutions in the human lineage from sequence alignments of human, chimpanzee and orangutan. These substitutions were drawn in close proximity of the SNP set to ensure comparability (see Material and Methods). We initially measured the correlation between *rmf* estimates for the two different sets of nucleotide substitutions, and generally found that all substitution types displayed strong and significant positive associations at the DNA trimer level (Table 2). The weakest correlation between SNPs and fixed substitutions was observed for *rmf* of A>T(T>A) transversions (r=0.88, p=7.5e-6). One should note that the relative fraction of each substitution type displayed some differences between SNPs and fixed lineage substitutions (as can be seen by comparing Table 1 with Table 2), potentially influencing the *rmf* correlations between the two sets.

We replicated the most significant associations between *rmf* and DNA physical properties that were discovered within the SNP set (Table 3). As we observed for SNPs, the relative mutation fraction of nucleotide transitions displayed a strong negative association with context-dependent levels of instability induced by the underlying T:G and A:C mismatches. We further confirmed that the global *rmf* distribution (i.e. by combining contributions by all substitution types) inferred from fixed substitutions was significantly associated to the intensity of cleavage by the hydroxyl radical, yet somewhat weaker than for SNPs (r=0.48, p=1.1e-2, k = 3). Some discrepancies did however exist between the SNP set and the set of fixed substitutions with respect to significant associations between physical properties. An example of such discrepancy concerned the relationship between *rmf* of C>G(G>C) transversions and local helix stability, which was not significant (at the 0.05 level) when it was analyzed within the fixed set of substitutions.

A combined impact of physical properties

The analyses above indicated that each of the different DNA physical properties correlated significantly with the relative mutation fraction of short DNA sequence contexts, though to different extents. However, the three scales that measure physical aspects of DNA are not quantitatively independent, but appear significantly correlated (see Supplementary Figures 1-3). For instance, we found significant negative correlation between ·OH cleavage intensities of DNA *k*-mers and the corresponding stabilization energies, with Pearson's r in the range of 0.5 to 0.68 (Supplementary Figures 2AB). Consequently, the observed bias of mutation is likely to result from the combined impact of these three properties, whose relative contributions, however, remain to be determined. We therefore developed multiple linear regression models that could indicate a stronger explanatory value for combinations of DNA physical properties, as opposed to the individual impacts outlined above.

For each of the six different nucleotide substitution types, we modelled the relationship between relative mutation fraction and all possible combinations of DNA physical properties. For the four different nucleotide transversion types, no model appeared significant at the 5% level after correction for multiple testing (data not shown). For each of the two transitions, however, we obtained submodels that were significant after multiple testing correction. For A>G(T>C), the best submodel was given solely by the level of underlying mispair stability, as depicted in Figure 1 (adjusted R^2=0.54, Simes' p=0.01). We further observed the most significant result for the C>T/G>A transition, in which the combination of mispair destabilization and intensity of cleavage by the hydroxyl radical could

explain close to 90% of the variance in relative mutation fractions (adjusted R^2=0.87, Simes' p=5.8e-4).

Final remarks

In summary, we show that sequence context biases observed in the recent human mutational spectrum display some notable significant associations with DNA physical properties, and that the associations are limited to particular substitution types. Our study thus highlights the physical characteristics of DNA as a potentially significant factor underlying local sequence bias of human mutation.

The results of the correlation analyses are clearly dependent upon the estimates of local mutation bias, i.e. the relative mutation fractions (*rmf*) inferred from SNPs. The intention underlying the *rmf* expression was to provide reliable figures of how the local DNA sequence imposes constraints on its mutability. We are aware though, that some factors could have introduced biases within the *rmf* estimates. First, *rmf* will be sensitive to the quality of the SNP data being used as a germline mutation source. We do believe, however, that our method for retrieving SNPs from dbSNP was designed in a conservative fashion, by that means minimizing the fraction of false positive SNPs used in the analysis. We furthermore replicated our main findings within a fixed set of nucleotide substitutions in the human lineage, suggesting a nonsignificant impact of noise in the SNP set. Secondly, we are aware that estimating mutation bias in terms of DNA trimers and pentamers may not provide the optimal resolution for capturing the complete spectrum of local context dependencies in germline mutagenesis. Specifically, it may well be that some substitution types are primarily dependent upon one neighbouring nucleotide rather than a joint dependency of two nucleotides upstream and downstream. Such instances will consequently yield *rmf* trimer estimates that are not completely independent, since two nearly identical trimers may both contain the dominant dinucleotide dependency.

Another important aspect of our analysis concerns the exclusion of the CpG dinucleotide context. This dinucleotide is heavily underrepresented in vertebrate genomes, primarily due to the high rate of spontaneous deamination at methylated cytosines, leading to C>T/G>A transitions [Duncan and Miller, 1980]. The intention for excluding CpG in our computations was thus to provide *rmf* estimates that primarily reflected non-CpG mutational mechanisms. However, since the underrepresentation of CpG is linked to an excess of TpG/CpA dinucleotides (though orders of magnitude less than the underrepresentation of

CpG), the *rmf* contexts containing these dinucleotides could be biased owing to the CpG effect [Simmen, 2008]. On the other hand, studies have also reported that CpG may be a favoured target for mutations not explicable in terms of the spontaneous deamination process, and more importantly, that TpG/CpA are slightly overrepresented even in unmethylated animal genomes [You et al., 1999; Pfeifer, 2006; Simmen, 2008].

It is important to emphasize that the impact of local DNA sequence context is merely one piece of the puzzle, and that several other factors jointly influence the incidence of mutation. For instance, it is clear that macroscopic features of the genome, such as the density of GC-rich isochores and level of chromatin packaging, will affect the substitution rate [Hellmann et al., 2005; Sasaki et al., 2009]. These factors are however likely to exert their influence at a lower resolution than at the level of DNA trimers and pentamers.

References

Allawi HT, SantaLucia J. 1998a. Nearest neighbor thermodynamic parameters for internal G.A mismatches in DNA. Biochemistry 37:2170-9.

Allawi HT, SantaLucia J. 1998b. Nearest-neighbor thermodynamics of internal A.C mismatches in DNA: sequence dependence and pH effects. Biochemistry 37:9435-44.

Allawi HT, SantaLucia J. 1998c. Thermodynamics of internal C.T mismatches in DNA. Nucleic Acids Res 26:2694-701.

Allawi HT, SantaLucia J, Jr. 1997. Thermodynamics and NMR of internal G.T mismatches in DNA. Biochemistry 36:10581-94.

Bailey JA, Gu Z, Clark RA, Reinert K, Samonte RV, Adams MD, Myers EW, Li PW, Eichler EE. 2002. Recent segmental duplications in the human genome. Science 297:1003-7.

Balasubramanian B, Pogozelski WK, Tullius TD. 1998. DNA strand breaking by the hydroxyl radical is governed by the accessible surface areas of the hydrogen atoms of the DNA backbone. Proc Natl Acad Sci U S A 95:9738-43.

Blanchette M, Kent WJ, Riemer C, Elnitski L, Smit AF, Roskin KM, Baertsch R, Rosenbloom K, Clawson H, Green ED, Haussler D, Miller W. 2004. Aligning multiple genomic sequences with the threaded blockset aligner. Genome Res 14:708-15.

Brown TC, Jiricny J. 1988. Different base/base mispairs are corrected with different efficiencies and specificities in monkey kidney cells. Cell 54:705-11.

Chi LM, Lam SL. 2006. NMR investigation of DNA primer-template models: structural insights into dislocation mutagenesis in DNA replication. FEBS Lett 580:6496-500.

Chi LM, Lam SL. 2007. NMR investigation of primer-template models: structural effect of sequence downstream of a thymine template on mutagenesis in DNA replication. Biochemistry 46:9292-300.

Chi LM, Lam SL. 2009. NMR investigation of DNA primer-template models: guanine templates are less prone to strand slippage upon misincorporation. Biochemistry 48:11478-86.

Cloutier JF, Castonguay A, O'Connor TR, Drouin R. 2001a. Alkylating agent and chromatin structure determine sequence context-dependent formation of alkylpurines. J Mol Biol 306:169-88.

Cloutier JF, Drouin R, Weinfeld M, O'Connor TR, Castonguay A. 2001b. Characterization and mapping of DNA damage induced by reactive metabolites of 4-(methylnitrosamino)-1-(3-pyridyl)-1-butanone (NNK) at nucleotide resolution in human genomic DNA. J Mol Biol 313:539-57.

Cooper DN, Krawczak M. 1993. Human Gene Mutation: BIOS Scientific Publishers Limited.

Drew HR, Travers AA. 1985. DNA bending and its relation to nucleosome positioning. J Mol Biol 186:773-90.

Duncan BK, Miller JH. 1980. Mutagenic deamination of cytosine residues in DNA. Nature 287:560-561.

Duret L, Galtier N. 2009. Biased gene conversion and the evolution of mammalian genomic landscapes. Annu Rev Genomics Hum Genet 10:285-311.

Eftedal I, Guddal PH, Slupphaug G, Volden G, Krokan HE. 1993. Consensus sequences for good and poor removal of uracil from double stranded DNA by uracil-DNA glycosylase. Nucleic Acids Res 21:2095-101.

Ehrlich M, Wang RY. 1981. 5-Methylcytosine in eukaryotic DNA. Science 212:1350-7.

Frank-Kamenetskii M. 1987. DNA chemistry. How the double helix breathes. Nature 328:17-8.

Fredman D, White SJ, Potter S, Eichler EE, Den Dunnen JT, Brookes AJ. 2004. Complex SNP-related sequence variation in segmental genome duplications. Nat Genet 36:861-6.

Gabrielian A, Pongor S. 1996. Correlation of intrinsic DNA curvature with DNA property periodicity. FEBS Lett 393:65-8.

Gardiner-Garden M, Frommer M. 1987. CpG islands in vertebrate genomes. J Mol Biol 196:261-82.

Giardine B, Riemer C, Hardison RC, Burhans R, Elnitski L, Shah P, Zhang Y, Blankenberg D, Albert I, Taylor J, Miller W, Kent WJ, Nekrutenko A. 2005. Galaxy: a platform for interactive large-scale genome analysis. Genome Res 15:1451-5.

Goodman MF, Creighton S, Bloom LB, Petruska J. 1993. Biochemical basis of DNA replication fidelity. Crit Rev Biochem Mol Biol 28:83-126.

Green P, Ewing B, Miller W, Thomas PJ, Program NCS, Green ED. 2003. Transcription-associated mutational asymmetry in mammalian evolution. Nat Genet 33:514-7.

Greenbaum JA, Pang B, Tullius TD. 2007. Construction of a genome-scale structural map at single-nucleotide resolution. Genome Res 17:947-53.

Gueron M, Kochoyan M, Leroy JL. 1987. A single mode of DNA base-pair opening drives imino proton exchange. Nature 328:89-92.

Hellmann I, Prufer K, Ji H, Zody MC, Paabo S, Ptak SE. 2005. Why do human diversity levels vary at a megabase scale? Genome Res 15:1222-31.

Hess ST, Blake JD, Blake RD. 1994. Wide variations in neighbor-dependent substitution rates. J Mol Biol 236:1022-33.

Hunter WN, Brown T, Anand NN, Kennard O. 1986. Structure of an adenine-cytosine base pair in DNA and its implications for mismatch repair. Nature 320:552-5.

Hunter WN, Brown T, Kneale G, Anand NN, Rabinovich D, Kennard O. 1987. The structure of guanosine-thymidine mismatches in B-DNA at 2.5-A resolution. J Biol Chem 262:9962-70.

Hwang DG, Green P. 2004. Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. Proc Natl Acad Sci USA 101:13994-4001.

International HapMap Consortium. 2007. A second generation human haplotype map of over 3.1 million SNPs. Nature 449:851-61.

Isaacs RJ, Spielmann HP. 2004. A model for initial DNA lesion recognition by NER and MMR based on local conformational flexibility. DNA Repair (Amst) 3:455-64.

Jiang C, Zhao Z. 2006. Mutational spectrum in the recent human genome inferred by single nucleotide polymorphisms. Genomics 88:527-534.

Jiricny J. 1998. Replication errors: cha(lle)nging the genome. EMBO J 17:6427-36.

Karolchik D, Baertsch R, Diekhans M, Furey TS, Hinrichs A, Lu YT, Roskin KM, Schwartz M, Sugnet CW, Thomas DJ, Weber RJ, Haussler D, Kent WJ. 2003. The UCSC Genome Browser Database. Nucleic Acids Res 31:51-4.

Krawczak M, Ball EV, Cooper DN. 1998. Neighboring-nucleotide effects on the rates of germ-line single-base-pair substitution in human genes. Am J Hum Genet 63:474-88.

Kunkel TA. 2004. DNA replication fidelity. J Biol Chem 279:16895-8.

Kunkel TA, Alexander PS. 1986. The base substitution fidelity of eucaryotic DNA polymerases. Mispairing frequencies, site preferences, insertion preferences, and base substitution by dislocation. J Biol Chem 261:160-6.

Kvikstad EM, Tyekucheva S, Chiaromonte F, Makova KD. 2007. A macaque's-eye view of human insertions and deletions: differences in mechanisms. PLoS Comput Biol 3:1772-82.

Maki H. 2002. Origins of spontaneous mutations: specificity and directionality of base-substitution, frameshift, and sequence-substitution mutageneses. Annu Rev Genet 36:279-303.

Marathias VM, Jerkovic B, Bolton PH. 1999. Damage increases the flexibility of duplex DNA. Nucleic Acids Res 27:1854-8.

Mazurek A, Johnson CN, Germann MW, Fishel R. 2009. Sequence context effect for hMSH2-hMSH6 mismatch-dependent activation. Proc Natl Acad Sci U S A 106:4177-82.

Mugal CF, von Grünberg H-H, Peifer M. 2009. Transcription-induced mutational strand bias and its effect on substitution rates in human genes. Mol Biol Evol 26:131-42.

Musumeci L, Arthur JW, Cheung FS, Hoque A, Lippman S, Reichardt JK. 2009. Single Nucleotide Differences (SNDs) in the dbSNP Database May Lead to Errors in Genotyping and Haplotyping Studies. Hum Mutat.

Nakken S, Rodland EA, Rognes T, Hovig E. 2009. Large-scale inference of the point mutational spectrum in human segmental duplications. BMC Genomics 10:43.

Nelson MR, Marnellos G, Kammerer S, Hoyal CR, Siepel M, Cantor CR, Braun A. 2004. Large-scale validation of single nucleotide polymorphisms in gene regions. Genome Res 14:1664-8.

Pennacchio LA, Ahituv N, Moses AM, Prabhakar S, Nobrega MA, Shoukry M, Minovitsky S, Dubchak I, Holt A, Lewis KD, Plajzer-Frick I, Akiyama J, De Val S, Afzal V, Black BL, Couronne O, Eisen MB, Visel A, Rubin EM. 2006. In vivo enhancer analysis of human conserved non-coding sequences. Nature 444:499-502.

Peyret N, Seneviratne PA, Allawi HT, SantaLucia J, Jr. 1999. Nearest-neighbor thermodynamics and NMR of DNA sequences with internal A.A, C.C, G.G, and T.T mismatches. Biochemistry 38:3468-77.

Pfeifer GP. 2006. Mutagenesis at methylated CpG sequences. Curr Top Microbiol Immunol 301:259-81.

Rajski SR, Jackson BA, Barton JK. 2000. DNA repair: models for damage and mismatch recognition. Mutat Res 447:49-72.

Rochette PJ, Therrien J-P, Drouin R, Perdiz D, Bastien N, Drobetsky EA, Sage E. 2003. UVA-induced cyclobutane pyrimidine dimers form predominantly at thymine-thymine dipyrimidines and correlate with the mutation spectrum in rodent cells. Nucleic Acids Res 31:2786-94.

SantaLucia J, Jr. 1998. A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. Proc Natl Acad Sci U S A 95:1460-5.

Sasaki S, Mello CC, Shimada A, Nakatani Y, Hashimoto S, Ogawa M, Matsushima K, Gu SG, Kasahara M, Ahsan B, Sasaki A, Saito T, Suzuki Y, Sugano S, Kohara Y, Takeda H, Fire A, Morishita S. 2009. Chromatin-associated periodicity in genetic variation downstream of transcriptional start sites. Science 323:401-4.

Segal E, Widom J. 2009. What controls nucleosome positions? Trends Genet 25:335-43.

Seibert E, Ross JB, Osman R. 2002. Role of DNA flexibility in sequence-dependent activity of uracil DNA glycosylase. Biochemistry 41:10976-84.

Seibert E, Ross JB, Osman R. 2003. Contribution of opening and bending dynamics to specific recognition of DNA damage. J Mol Biol 330:687-703.

Shibutani S, Fernandes A, Suzuki N, Zhou L, Johnson F, Grollman AP. 1999. Mutagenesis of the N-(deoxyguanosin-8-yl)-2-amino-1-methyl-6-phenylimidazo[4, 5-b]pyridine DNA adduct in mammalian cells. Sequence context effects. J Biol Chem 274:27433-8.

Sibghat-Ullah, Gallinari P, Xu YZ, Goodman MF, Bloom LB, Jiricny J, Day RS, 3rd. 1996. Base analog and neighboring base effects on substrate specificity of recombinant human G:T mismatch-specific thymine DNA-glycosylase. Biochemistry 35:12926-32.

Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, Weinstock GM, Wilson RK, Gibbs RA, Kent WJ, Miller W, Haussler D. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res 15:1034-50.

Siepel A, Haussler D. 2004. Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. Mol Biol Evol 21:468-88.

Simes RJ. 1986. An Improved Bonferroni Procedure for Multiple Tests of Significance Biometrika 73:751-4.

Simmen MW. 2008. Genome-scale relationships between cytosine methylation and dinucleotide abundances in animals. Genomics 92:33-40.

Spencer CCA, Deloukas P, Hunt S, Mullikin JC, Myers S, Silverman B, Donnelly P, Bentley DR, McVean G. 2006. The influence of recombination on human genetic diversity. PLoS Genet 2:e148.

Stoltzfus A. 2008. Evidence for a predominant role of oxidative damage in germline mutation in mammals. Mutat Res 644:71-3.

Strathern JN, Shafer BK, McGill CB. 1995. DNA synthesis errors associated with double-strand-break repair. Genetics 140:965-72.

Tomso DJ, Bell DA. 2003. Sequence context at human single nucleotide polymorphisms: overrepresentation of CpG dinucleotide at polymorphic sites and suppression of variation in CpG islands. J Mol Biol 327:303-8.

Wang H, Yang Y, Schofield MJ, Du C, Fridman Y, Lee SD, Larson ED, Drummond JT, Alani E, Hsieh P, Erie DA. 2003. DNA bending and unbending by MutS govern mismatch recognition and specificity. Proc Natl Acad Sci U S A 100:14822-7.

You YH, Li C, Pfeifer GP. 1999. Involvement of 5-methylcytosine in sunlight-induced mutagenesis. J Mol Biol 293:493-503.

Zhao Z. 2002. Neighboring-Nucleotide Effects on Single Nucleotide Polymorphisms: A Study of 2.6 Million Polymorphisms Across the Human Genome. Genome Res 12:1679-1686.

Zhao Z, Zhang F. 2006. Sequence context analysis of 8.2 million single nucleotide polymorphisms in the human genome. Gene 366:316-324.

Tables

Table 1 – Pearsons's product-moment correlation coefficient *r* between physical properties of DNA *k*-mers and their relative mutation fraction inferred from 2,174,291 human SNPs. Correlation coefficients are provided for all six types of nucleotide substitutions, in trimer contexts (*k*=3) and pentamer contexts (*k*=5). Significant associations (p<=0.05) are shown in bold font. The thermodynamic stabilities of DNA mispairs have only been estimated for DNA trimers (hence no correlation coefficients for pentamers given (indicated as 'n/a')). n.s. denotes non-significant associations.

	Substitution type	*c →t / g →a*	*a →g / t →c*	*a →c / t →g*	*c →a / g →t*	*a →t / t →a*	*c →g / g →c*
	Fraction of all substitutions	*27.3%*	*36.9%*	*9.3%*	*9.1%*	*8.1%*	*9.3%*
k = 3	ΔG *ancestral*	0.01 (n.s.)	0.27 (n.s.)	-0.16 (n.s.)	-0.23 (n.s.)	0.09 (n.s.)	**0.63 (p=2.8e-2)**
	ΔG *mispair*	**-0.63 (p=3.0e-2)**	**-0.76 (p=6.7e-4)**	0.47 (n.s.)	0.01 (n.s.)	-0.03 (n.s.)	-0.32 (n.s.)
	·OH cleavage intensity	0.32 (n.s.)	0.26 (n.s.)	-0.48 (n.s.)	0.49 (n.s.)	**0.58 (p=1.8e-2)**	-0.17 (n.s.)
	DNA bendability	0.32 (n.s.)	0.28 (n.s.)	-0.35 (n.s.)	-0.14 (n.s.)	0.27 (n.s.)	0.05 (n.s.)
k = 5	ΔG *ancestral*	-0.04 (n.s.)	0.09 (n.s.)	-0.11 (p=n.s.)	0.02 (n.s.)	0.05 (n.s.)	**0.40 (7.8e-8)**
	ΔG *mispair*	n/a	n/a	n/a	n/a	n/a	n/a
	·OH cleavage intensity	**0.19 (p=1.3e-2)**	**0.37 (p=1.4e-8)**	**-0.21 (p=1.6e-3)**	**0.23 (3.3e-3)**	**0.44 (p=3.9e-12)**	-0.03 (n.s.)
	DNA bendability	**0.31 (p=5.3e-5)**	**0.31 (p=1.5e-6)**	-0.07 (n.s.)	**-0.21 (p=7.0e-3)**	-0.01 (n.s.)	-0.15 (n.s.)

Table 3 – Pearsons's product-moment correlation coefficient *r* between physical properties of DNA trimers and their relative mutation fraction inferred from 2,066,656 fixed substitutions in the human lineage. Correlation coefficients are provided for all six types of nucleotide substitutions in trimer contexts (*k*=3). Significant associations (p<=0.05) are shown in bold font. n.s. denotes non-significant associations.

	Substitution type	$c \rightarrow t / g \rightarrow a$	$a \rightarrow g / t \rightarrow c$	$a \rightarrow c / t \rightarrow g$	$c \rightarrow a / g \rightarrow t$	$a \rightarrow t / t \rightarrow a$	$c \rightarrow g / g \rightarrow c$
	Fraction of all substitutions	26.8%	*39.6%*	*9.8%*	*8.1%*	*7.1%*	*8.6%*
k = 3	$\Delta G_{ancestral}$	0.09 (n.s.)	0.12 (n.s.)	0.06 (n.s.)	-0.31 (n.s.)	0.28 (n.s.)	0.53 (n.s.)
	$\Delta G_{mispair}$	**-0.68 (p=1.6e-2)**	**-0.74 (p=1.1e-3)**	0.42 (n.s.)	0.06 (n.s.)	-0.39 (n.s.)	-0.25 (n.s.)
	·OH cleavage intensity	0.21 (n.s.)	0.27 (n.s.)	**-0.51 (p=4.2e-2)**	0.55 (n.s.)	0.22 (n.s.)	0.03 (n.s.)
	DNA bendability	0.34 (n.s.)	0.40 (n.s.)	-0.24 (n.s.)	-0.12 (n.s.)	0.02 (n.s.)	0.10 (n.s.)

Table 2 – Pearsons's product-moment correlation coefficient *r* between *rmf* estimates from SNPs and *rmf* estimates from fixed lineage substitutions (DNA trimers)

Substitution	*Correlation coefficient*
$c \rightarrow t\,/\,g \rightarrow a$	r = 0.92 (p=2.6e-5)
$a \rightarrow g\,/\,t \rightarrow c$	r = 0.98 (p=2.6e-11)
$a \rightarrow c\,/\,t \rightarrow g$	r = 0.92 (p=3.3e-7)
$c \rightarrow a\,/\,g \rightarrow t$	r = 0.97 (p=1.1e-7)
$a \rightarrow t\,/\,t \rightarrow a$	r = 0.88 (p=7.5e-6)
$c \rightarrow g\,/\,g \rightarrow c$	r = 0.96 (p=1.3e-6)

Figure Legends

Figure 1 – Box-whisker plots of the relative mutation fraction (*rmf*) for DNA trimers and DNA pentamers. Transitions are depicted with blue boxes, transversions with green boxes. Only non-CpG contexts have been considered.

Figure 2 – Association between relative mutation fraction (*rmf*) of DNA trimer contexts and the local helix stability. Left plot shows association between *rmf* of C>G(G>C) transversions and thermodynamic stability of the ancestral context. The center plot shows association between *rmf* of A>G(T>C) transitions and the thermodynamic instability induced by the underlying mispairs (i.e. T:G and A:C). The right plot shows association between *rmf* of C>T(G>A) transitions and the thermodynamic instability induced by the underlying mispairs (also T:G and A:C). Error bars were omitted in the plots due to negligible levels of *rmf* point estimate uncertainty.

Supplementary Material

Table 1 – Correlation in relative mutation fraction of reverse complementary contexts in transcribed (intronic) sequences. The underlying *rmf* estimates are based on 565,314 intronic SNPs mapped to the non-template strand (see Methods & Results). Pearson's correlation coefficient *r* is given as the measure of association.

Reverse complementary substitutions	*rmf correlation (DNA trimers)*
$c \rightarrow t$ *versus* $g \rightarrow a$	r = 0.79 (p=2.1e-3)
$a \rightarrow g$ *versus* $t \rightarrow c$	r = 0.99 (p=3.6e-12)
$a \rightarrow c$ *versus* $t \rightarrow g$	r = 0.92 (p=5.1e-7)
$c \rightarrow a$ *versus* $g \rightarrow t$	r = 0.97 (p=1.5e-7)
$a \rightarrow t$ *versus* $t \rightarrow a$	r = 0.96 (p=2.6e-9)
$c \rightarrow g$ *versus* $g \rightarrow c$	r = 0.94 (p=4.9e-6)

Figure 1A - Correlation between local thermodynamic stability and relative DNA bendability (DNA trimers)

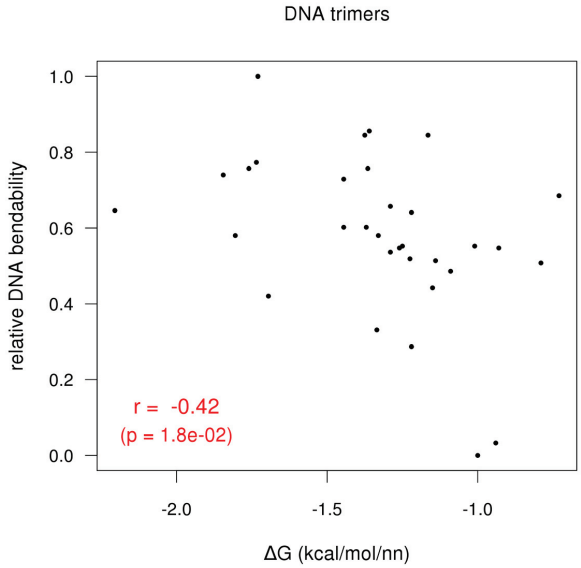


Figure 1B - Correlation between local thermodynamic stability and relative DNA bendability (DNA pentamers)

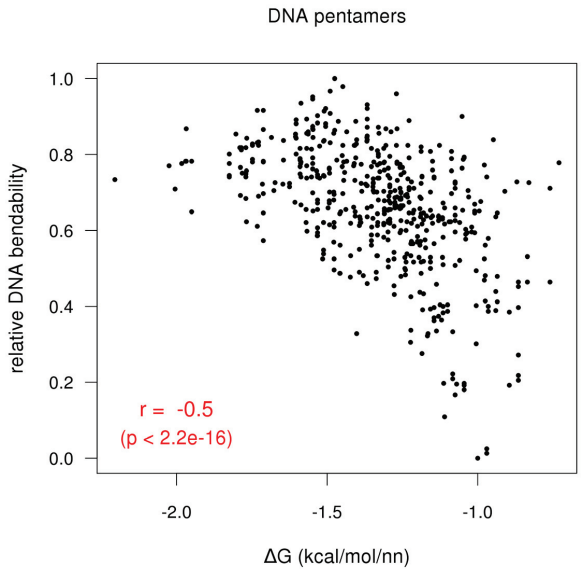


Figure 2A - Correlation between local thermodynamic stability and relative intensity of cleavage by the hydroxyl radical (DNA trimers)

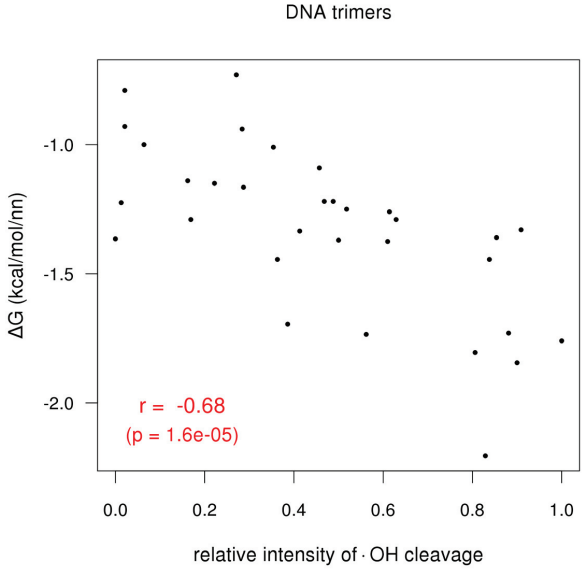


Figure 2B - Correlation between local thermodynamic stability and relative intensity of cleavage by the hydroxyl radical (DNA pentamers)

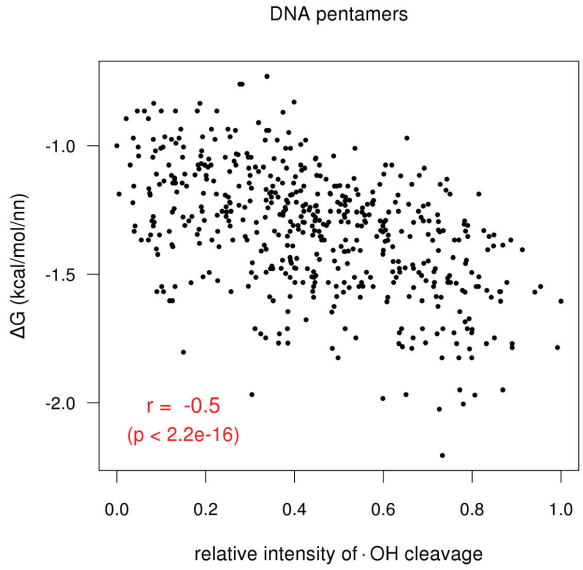


Figure 3A - Correlation between relative DNA bendability and relative intensity of cleavage by the hydroxyl radical (DNA trimers)

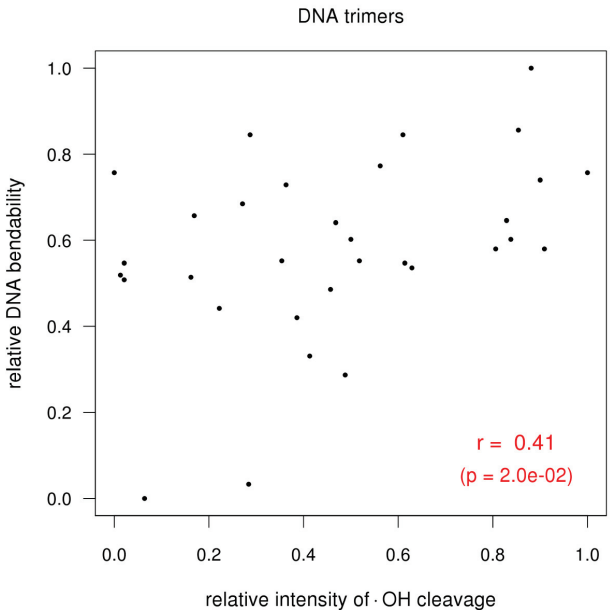
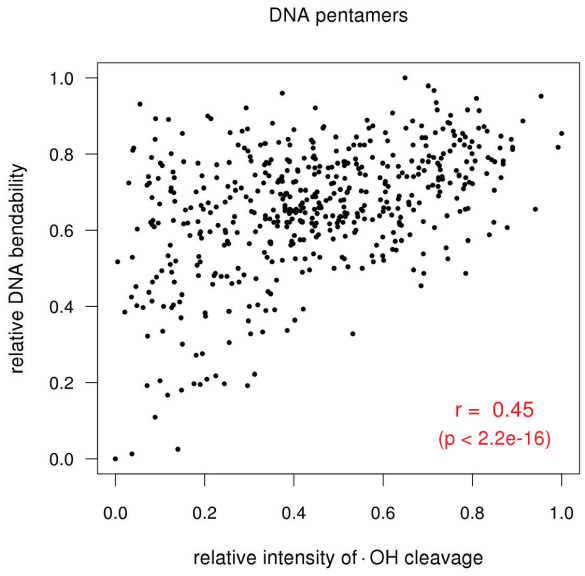


Figure 3B - Correlation between relative DNA bendability and relative intensity of cleavage by the hydroxyl radical (DNA pentamers)



IV

J Mol Evol (2010) 70:266–274
DOI 10.1007/s00239-010-9328-0

Unstable DNA Repair Genes Shaped by Their Own Sequence Modifying Phenotypes

Daniel S. Falster · Sigve Nakken · Marie Bergem-Ohr · Einar Andreas Rødland · Jarle Breivik

Received: 21 January 2010 / Accepted: 10 February 2010 / Published online: 6 March 2010

Abstract The question of whether natural selection favors genetic stability or genetic variability is a fundamental problem in evolutionary biology. Bioinformatic analyses demonstrate that selection favors genetic stability by avoiding unstable nucleotide sequences in protein encoding DNA. Yet, such unstable sequences are maintained in several DNA repair genes, thereby promoting breakdown of repair and destabilizing the genome. Several studies have therefore argued that selection favors genetic variability at the expense of stability. Here we propose a new evolutionary mechanism, with supporting bioinformatic evidence, that resolves this paradox. Combining the concepts of gene-dependent mutation biases and meiotic recombination, we argue that unstable sequences in the DNA mismatch repair (MMR) genes are maintained by their own phenotype. In particular, we predict that human MMR maintains an overrepresentation of mononucleotide repeats (monorepeats) within and around the MMR genes. In support of this hypothesis, we report a 31% excess in monorepeats in 250 kb regions surrounding the seven MMR genes compared to all other RefSeq genes (1.75 vs. 1.34%, $P = 0.0047$), with a particularly high content in PMS2 (2.41%, $P = 0.0047$) and MSH6 (2.07%, $P = 0.043$). Based on a mathematical model of monorepeat frequency, we argue that the proposed mechanism may suffice to explain the observed excess of repeats around MMR genes. Our findings thus indicate that unstable sequences in MMR genes are maintained through evolution by the MMR mechanism. The evolutionary paradox of genetically unstable DNA repair genes may thus be explained by an equilibrium in which the phenotype acts back on its own genotype.

Keywords DNA repair · Microsatellites · Genetic instability · Cancer · DNA mismatch repair · Recombination · Mutation bias

Electronic supplementary material The online version of this article (doi:10.1007/s00239-010-9328-0) contains supplementary material, which is available to authorized users.

D. S. Falster · M. Bergem-Ohr · J. Breivik (✉)
Institute of Basic Medical Science, University of Oslo,
P.O. Box 1018 Blindern, 0315 Oslo, Norway
e-mail: jbreivik@medisin.uio.no

Present Address:
D. S. Falster
Department of Biological Sciences, Macquarie University,
Sydney, Australia

S. Nakken
Centre for Molecular Biology and Neuroscience,
Institute of Medical Microbiology, Rikshospitalet
University Hospital, 0027 Oslo, Norway

Present Address:
S. Nakken
Bioinformatics Core Facility, Institute of Medical Informatics,
Rikshospitalet, 0310 Oslo, Norway

E. A. Rødland
Department of Informatics and Center for Cancer Biomedicine,
University of Oslo, 0316 Oslo, Norway

E. A. Rødland
Norwegian Computing Center, 0314 Oslo, Norway

Introduction

DNA mismatch repair (MMR) is an enzymatic mechanism that recognizes and corrects single nucleotide and

insertion–deletion mismatches in DNA (Lyer et al. 2006; Marti et al. 2002). It thereby maintains the overall stability of the genome and is central to the prevention of cancer (Lynch et al. 2006; Peltomaki 2005). MMR is particularly important in stabilizing the length of microsatellites (also known as short tandem repeats or simple sequence repeats), and MMR deficiency is recognized as microsatellite instability throughout the genome (Ellegren 2004). Concurrently, several of the MMR genes, in human and other eukaryotes, contain microsatellites within their own coding sequence (Chang et al. 2001). These monorepeats make MMR genes particularly susceptible to deactivation by frame-shift mutation and a mutational target in cancer development (Venkatesan et al. 2006; Ohmiya et al. 2001; Perucho 1996). Thus, the very genes that protect against genetic instability and cancer are themselves unstable. In this article, we provide a mechanistic explanation for this seeming evolutionary paradox.

Chang et al. (2001) previously proposed that the unstable sequences in the MMR genes have been selected because they provide genetic variability. This idea of selection for variability has been proposed to explain a number of biological phenomena (Kashi and King 2006; Li et al. 2004), but evidence for this interpretation is limited. Other authors have therefore argued that although instability is not selected per se, unstable sequences may spread when linked to other favorable properties (Sniegowski et al. 2000; Baer et al. 2007). In general, however, full genome analyses demonstrate that selection favors stability by avoiding nucleotide repeats in coding sequences (Ackermann and Chao 2006; Wanner et al. 2008). The question thus remains: Why are unstable microsatellites overrepresented in the very MMR genes responsible for maintaining microsatellite stability?

Another relationship between MMR and microsatellites gives hint of a possible solution. Numerous studies show that MMR not only stabilizes microsatellites, but can also induce different types of mutation biases in such sequences (Burt and Trivers 2006; Sleckman 2005; Ellegren 2002; Pearson et al. 2005; Shah et al. 2010). As a primary example, wild-type MSH2 promotes expansion of trinucleotide repeats related to inheritance and progression of neurodegenerative disorders in mouse models (Subramanian et al. 2003; Manley et al. 1999), whereas the homologous gene in *Drosophila melanogaster* (Spel1) causes genome-wide contraction of dinucleotide repeats (Harr et al. 2002).

In humans, mutation of MSH2 and other MMR genes is related to the Lynch syndrome (Lynch et al. 2006; Felton et al. 2007). This condition, with an incidence of approximately 1:1000 in the general population (de la Chapelle 2005), is characterized by early development of tumors with microsatellite instability. The affected individual is generally heterozygous, and MMR deficiency arise as a consequence of somatic inactivation of the normal allele. The instability is particularly evident in monorepeats (Lynch et al. 2006; Peltomaki 2001), and the mutated repeats show a strong overrepresentation (89%) of contractions (Sammalkorpi et al. 2007; Zhou et al. 1997), implying that MMR proficiency maintains the length and stability of monorepeats.

Microsatellite instability in Lynch syndrome is generally confined to the tumor cells, and little is known about the effect of MMR mutations through the human germline. Still, evidence from animal studies and cell lines, show that even heterozygous MMR mutations may produce an increase in mutation rate (Zhang et al. 2002; Alazzouzi et al. 2005; Bouffler et al. 2000), and such haploinsufficiency has also been detected in the germline (Larson et al. 2004; Gurtu et al. 2002; Baida et al. 2003).

Summing up, there are two different connections between MMR and monorepeats. First, several of the MMR genes are destabilized by monorepeats within their own coding regions (Chang et al. 2001). Second, MMR activity introduces a mutation bias that maintains the length and stability of monorepeats in somatic cells, and probably also through the germline. These observations led us to propose a mechanism that links these two phenomena. More specifically, we predict that the paradoxical occurrence of unstable monorepeats within the MMR genotypes is maintained by the mutation bias of the MMR phenotype.

Proposed Evolutionary Mechanism

The evidence summarized above indicates that the length of monorepeats is determined by a dynamic balance between expansion and contraction of repeat sequences, and that this equilibrium is influenced by different MMR phenotypes. Specifically, it suggests that the homozygous wild-type maintains the length and stability of long monorepeats, whereas the heterozygous mutant show a tendency for contraction due to haploinsufficiency.

For a random region of the genome, rearranged with new MMR alleles every generation, the state of equilibrium will be determined by the relative strength and frequency of the different MMR phenotypes in the population. For a wild-type MMR allele itself, however, this point of equilibrium will be shifted toward expansion. The reason may be illustrated by a Mendelian crossing scheme (Fig. 1). In brief, due to meiotic recombination through the course of evolution, an MMR allele will be more exposed to its own phenotype than to the phenotypes of the alternative alleles. Accordingly, an allele whose phenotype promotes a particular composition of nucleotides should in general contain more of such sequence elements than other sequences of the genome.

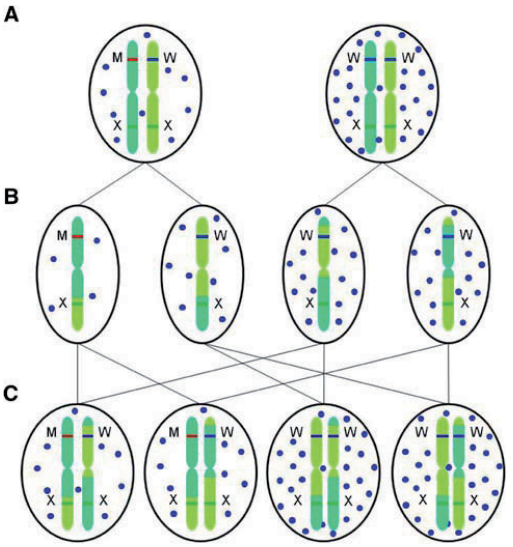


Fig. 1 Proposed mechanism by which an MMR protein (*blue dots*) selectively affects its own genotype. To illustrate the evolutionary dynamics we regard the crossing between a homozygous wild-type, W/W, and a heterozygous mutant, W/M (A). The W/W phenotype maintains the length and stability of monorepeats, whereas the insufficient phenotype (W/M) leads to contraction of these sequences. Regarding possible offspring (C), a random allele in the genome, X, is exposed to the insufficient phenotype in 4 of 8 cases (50%), whereas the W allele is exposed to this phenotype in 2 of 6 cases (33%). Regarding the haploid gametes (B), the W allele is physically separated from the M allele and may involve a differentiated mutagenic effect in the early stages of development. Combined, these effects of meiotic recombination suggest that an allele should be more influenced by its own phenotype than by the phenotype of alternative alleles. Or more specifically, a wild-type MMR allele should maintain longer monorepeats than other regions of the genome (Color figure online)

From this deduction we thus made the following predictions: (1) Wild-type MMR alleles, which maintain the stability of monorepeats, should have more monorepeats than other regions of the genome; (2) This effect should be seen throughout the haplotype block (McVean et al. 2004), not just as individual repeats in coding sequences (Chang et al. 2001); and (3) The amount of repeats in an MMR allele should correlate to the strength and frequency of its mutator phenotype (Marti et al. 2002).

Sequence Analysis

To test the hypotheses outlined above we performed a complete mapping of monorepeats in the human genome. Sequence data comprising 21,958 defined RefSeq gene sequences (hg19, NCBI Build 37.1) were analyzed for monorepeats. The MMR system was defined by the seven genes *MSH2*, *MSH3*, *MSH6*, *PMS1*, *PMS2*, *MLH1*, and *MLH3* (Marti et al. 2002). Comparisons were made between standardized genomic regions of 250 kb centered to the defined gene sequences, thus spanning the average length of haplotype blocks in the human genome, which is approximately 200 kb (McVean et al. 2004).

The dataset confirmed previous reports that monorepeats are overrepresented in the human genome compared to expectations based on random nucleotide sequences with similar base compositions (Subramanian et al. 2003; Borstnik and Pumpernik 2002). In particular, there was a marked deviation for long repeat lengths, starting from about 7 bp (Fig. 2). This pattern of deviation was matched by the 250 kb regions for all genes and for those comprising the MMR genes. The observed pattern is also consistent with experimental studies showing that there exists a threshold length about which monorepeats become intrinsically unstable and subject to the stabilizing effect of MMR (Lai and Sun 2003). Therefore, we considered only repeats of length 7 bp or longer in subsequent analyses.

To test for differences in the cumulative number of repeats among sequences, we calculated the proportion of

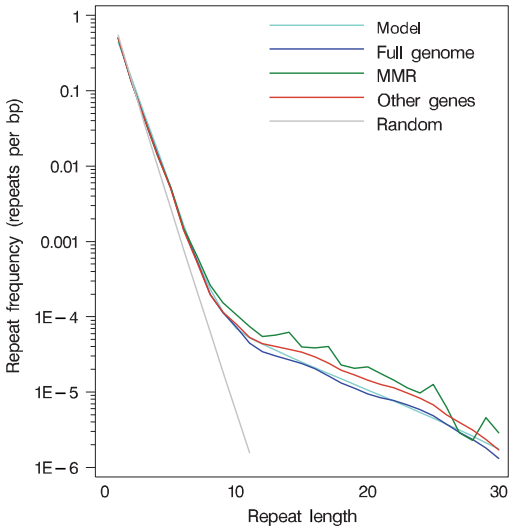


Fig. 2 Frequency of monorepeats in MMR genes and the genome. The frequency of monorepeats of increasing length was predicted based on the assumption of random distribution of nucleotides (*gray line*) (Borstnik and Pumpernik 2002), as well as, the presented mathematical model (*light blue line*). These predictions were then plotted against the observed frequency in the full genome (*blue line*), MMR gene regions (*green line*), and all other 250 kb gene regions (*red line*). MMR gene regions show a general excess of repeat lengths of 7 bp and longer compared to all other gene regions and the genome in general (Color figure online)

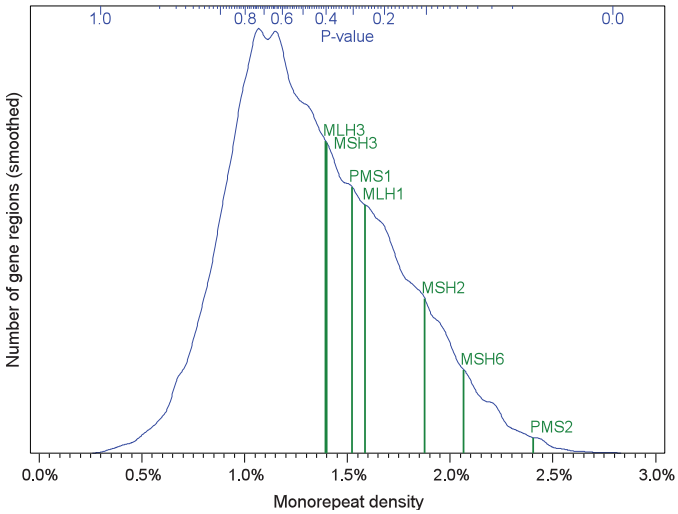


Fig. 3 Distribution of repeat content. The graph illustrates the distribution of all 250 kb gene regions relative to their content of monorepeat (7 bp and longer). Positions of the seven MMR regions are indicated. Top scale represents *P*-values for the distribution. The PMS2 and MSH6 regions each had a significant overrepresentation of repeats. All seven regions had above median repeat content and scored significantly as a group (Table 1)

the 250 kb gene regions made up of monorepeats (hereafter called repeat content, %) and compared the repeat content of MMR regions to the remaining gene regions. Repeat content varied greatly with respect to chromosome position (supporting information, Fig. S1) and showed a non-normal distribution (Fig. 3). Accordingly, statistical comparison of monorepeats between MMR and other gene regions were performed using Wilcoxon rank-sum test (one-sided, $\alpha = 0.05$).

The primary results are summarized in Table 1. Combined, the MMR regions had a 31% higher content of monorepeats than other gene regions (1.75 vs. 1.34%, $P = 0.0047$), with the excess of repeats distributed evenly across repeat lengths (Fig. 2). The seven MMR regions varied in repeat content from 1.39 to 2.41%. Two of the MMR regions differed significantly from the other gene regions when analyzed individually, *PMS2* (2.41%, $P = 0.0047$) and *MSH6* (2.07%, $P = 0.043$). All MMR regions scored above median repeat content (Fig. 3).

An excess of monorepeats in MMR coding sequences has previously been reported (Chang et al. 2001). Our results confirmed these findings, with a repeat content of 0.26% in protein coding parts of the 250 kb in MMR regions compared to 0.13% for other genes. Still, coding sequences had a lower repeat content than the non-coding sequences (0.26 vs. 1.79% for MMR regions, 0.13 vs. 1.38% in other gene regions) and contributed only 0.39% of the monorepeats in the 250 kb regions around the MMR genes. The contribution of the protein coding repeats, known prior to our analysis (Chang et al. 2001), was thus negligible for the overall repeat content of the MMR regions.

Analyses of Potential Confounding Factors

We found that monorepeat density varied between chromosomes ($P < 0.0001$, Kruskal–Wallis test). Moreover, we found that it was correlated (using Spearman correlation) with the GC content of the region (corr = 0.13), the fraction of region that was protein coding (corr = 0.26) and the level of gene expression (only available for 71% of genes; corr = 0.20), all highly significant ($P < 0.0001$). There was also a weak correlation to codon bias (corr = −0.012, $P = 0.068$).

In order to check if these factors could explain the observed density of monorepeats within and around the MMR genes, we applied a general linear model. Because repeat density had a slightly skewed distribution, we ran these analyses on the square root of the repeat density, which was less skewed. We then fitted a linear model using the above listed factors, with log-transformed gene expression values. Since we only had gene expression data for 71% of the genes, we first did the analyses without accounting for gene expression level, then an additional analysis including this factor.

The residuals from these analyses, i.e., the difference between the actual value and the value predicted by the linear model, were used as a measure of over- or under-representation of monorepeats corrected for chromosome differences and correlations. Wilcoxon analyses were then performed on these residuals comparing the MMR regions against the remaining.

The GLM model, with all factors included except gene expression level, explained 11.0% of the variance in repeat density, strengthening the difference between MMR

Table 1 Characteristics of MMR and other genes

Gene regions	Genomic location	Repeat content[a]	GC content[b]	Coding content[c]	Codon bias[d]	Expression[e]
MSH2	2p22-p21	1.88 ($P = 0.10$)	45.16	1.86	0.56	290.30
MSH3	5q11-q12	1.40 ($P = 0.40$)	38.36	1.37	0.56	401.65
MSH6	2p16	2.07 ($P = 0.043$)	42.39	2.46	0.57	2016.15
PMS1	2q31.1	1.52 ($P = 0.30$)	36.92	2.58	0.54	2517.20
PMS2	7p22.2	2.41 ($P = 0.0047$)	46.52	4.11	0.57	62.35
MLH1	3p21.3	1.59 ($P = 0.26$)	40.96	2.47	0.57	1883.90
MLH3	14q.24	1.39 ($P = 0.40$)	45.37	4.90	0.55	133.90
All MMR		1.75 ($P = 0.0047$)	42.24	2.82	0.56	1043.64
All other genes		1.34	44.75	2.96	0.60	1259.61

[a,b,c] Repeat content, GC content, and coding content are given as percentages. *P*-values were computed using a one-sided Wilcoxon rank-sum test

[d] Codon bias was computed using the *B measure* (Karlin et al. 1998)

[e] Gene expression data from testis germ cells were collected from Gene Atlas v2 (Su et al. 2004), and are given as gcRMA-condensed intensities

regions and control regions slightly (to $P = 0.0046$). When gene expression levels were included, all seven MMR genes, but only 71% of the other genes could be included in the analyses. This increased the explained variance to 14.0% and weakened the difference between MMR regions and control regions somewhat (to $P = 0.0102$). However, even when controlling for the effects of confounding factors, the differences between MMR genes and the remainder of the genome remained statistically significant. Thus, we may conclude that these factors, although contributing somewhat to observed differences, cannot explain the differences in repeat content between MMR genes and the rest of the genome. Further details are available as Supplementary Information.

Mathematical Model of Monorepeat Frequency

Our bioinformatic analyses support the hypothesis that differential exposure of MMR and other genes to MMR activity has led to differences in repeat content. In this section, we consider what size difference in expansion and contraction mutation rates are needed to explain these differences.

To assess the impact of varying mutation rate on repeat content, we modelled a stochastic process describing the evolution of repeat content due to slippage and point mutations. Our approach is based on the model presented by Lai and Sun (2003), which describes the effects of slippage mutation (contractions and expansions) on equilibrium repeat frequency. However, their model only treats the evolution of repeats after they have arisen, not the processes by which short repeats are created by point mutations. We therefore extended their model to include the processes by which point mutations maintain a background frequency of short monorepeats such as that expected in a purely random sequence.

The model is described in brief here; a full mathematical description is given in Supplementary Information. The genome was considered as a sequence of monorepeats and repeat evolution modeled as a stochastic process. The ordering of monorepeats was not modeled explicitly, only the frequency of repeats of different length. Repeat frequencies are influenced by point and slippage mutations, which extend, contract, join, or split existing repeats. Slippage mutations were assumed to expand or contract existing repeats by a single nucleotide, with mutation rates for expansion and contraction mutation increasing exponentially with repeat length. The effect of point mutations depends on their location within a repeat: point mutations can split an existing repeat, extend an existing repeat by a single base pair, or by join nearby repeats of similar type. The effects of slippage and point mutations combine to give transition probabilities for each repeat length. To simplify the dynamics, we assumed that sizes of neighboring repeats were independent. We then solved for the equilibrium length distribution (see Supporting Information for more details).

With relatively few parameters, the model described gave a good fit to the observed repeat distribution in the whole genome for repeats of length 2–30 bp (Fig. 2). To achieve this fit, we used a combination of observed mutation spectra and empirical fitting. The frequency of short repeats (2–5 bp) was influenced primarily by the probability that a point mutation extends a neighboring repeat sequence. This parameter was empirically fitted to match the observed repeat distribution. Based on data from Kelkar et al. (2008), the slippage mutation rate was set to increase exponentially with repeat length, starting at approximately 1000 times the point mutation ratio for

11-repeats and increasing by a factor 10 for every 15 nucleotides of length added. The ratio of expansion to contraction was adjusted to fit the observed repeat distribution. In order to get a reasonable fit for repeats of intermediate length, a correction term was needed to reduce the slippage mutation rate for repeats of less than 11 bp.

To explore influence of different levels of MMR activity on repeat content, we varied expansion and contraction rates across a range of values around the fitted values and assessed the effect on repeat content of the genome (Fig. 4). These adjustments represent possible effects of going from the general mutation rates experienced by the genome, to the mutation rates experienced by proficient MMR alleles. The results from the model indicate that small changes in rate of contraction mutation can alter mean repeat content in line with observed data. In particular, a 31% increase in repeat content, as observed in the MMR regions, might be explained by as little as a 3.4% reduction in the contraction frequency. An 81% increase in repeat content, as observed in the PMS2 region, requires only a 6.1% reduction in contraction frequency.

If MMR activity reduces expansion as well as contraction mutations, then a proportionately larger effect on contractions is needed to generate the observed repeat content. For example, if 89% of the slippage mutations caused by a defective MMR allele are contractions (Sammalkorpi et al. 2007), a 3.8% reduction in contraction rates and 0.5% reduction in expansion rate will again give 31% increase in repeat content. Similarly, increasing the rate of contraction mutation (as occurs in MMR deficient cells) caused a decrease in repeat content, as occurs in genetically unstable tumors and cell lineages.

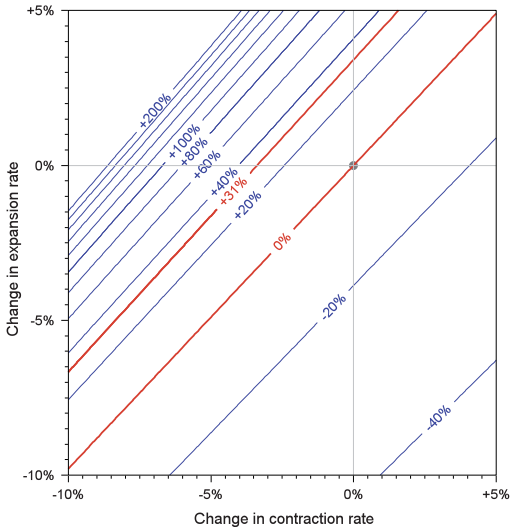


Fig. 4 Influence of expansion and contraction mutation rates on equilibrium repeat content predicted from stochastic model of repeat evolution. The contours show the change in repeat content (7 bp and longer) when contraction rates (*X* axis) and expansion rates (*Y* axis) are modified. The 31% change contour corresponds to the difference between MMR genes and other genes

Dunlop et al. (2000) have estimated the carrier frequency of MLH1 and MSH2 mutations to approximately 1:3139. Based on the approximate 1:1000 incidence of Lynch syndrome (de la Chapelle 2005), of which 40% are related to MSH2 (Peltomaki 2005) with a penetrance of 54% (Choi et al. 2009), we estimate the carrier frequency of mutated MSH2 to 1:1350 and the allele frequency to 1:2700. In order to get an overall increase of 3.4%, the mutated alleles must then increase the contraction rates $\sim$100-fold ($2700 \times 0.034 = 91.8$) to explain the observed differences in repeat content. Note that these numbers are very approximate, and merely serve to indicate the order of magnitude.

Discussion

Combining gene-dependent mutation biases with Mendelian inheritance (Fig. 1), we have deduced that an allele should be more affected by its own mutation bias than should other sequences of the genome. In particular, we predicted that the stabilizing effect of MMR on monorepeats has promoted an excess of such repeats within the MMR haplotype blocks. Confirming this prediction, we found a general expansion of monorepeats in 250 kb regions surrounding the MMR genes. This finding was based on a conservative statistical assessment controlling for the overrepresentation and uneven distribution of monorepeats in the genome. Furthermore, controlling for covariation of repeat density with protein coding content, GC content, codon bias or level of expression did not have significant influence on the results. The evolutionary dynamic proposed thus provides a novel explanation for the prevalence of unstable sequences in several MMR genes.

In accordance with previous analyses (Subramanian et al. 2003), we found a general overrepresentation of monorepeats longer than 7 bp in the human genome (Fig. 2), indicating a mechanism that promotes such sequences through the course of evolution. The same pattern was mirrored in the MMR regions, suggesting that the 31% excess of monorepeats is caused by the same mechanism that promotes such sequences throughout the genome. The statistical analysis and the pattern of repeat lengths thus support our hypothesis that the MMR proteins promote expansion of monorepeats in the human germline, and that this effect is particularly strong within and around their own nucleotide sequence.

Looking at the individual MMR regions, the highest content of monorepeats was found for *PMS2* and *MSH6*, followed by *MSH2* and *MLH*. These four genes cooperate in the recognition of small DNA loops that frequently arise in monorepeats during DNA replication (Lyer et al. 2006; Marti et al. 2002). Correspondingly, loss of function of any of these genes has been related to a particularly high degree of instability in monorepeats, whereas the other MMR genes have a limited effect (Lyer et al. 2006; Marti et al. 2002). *MLH1*, *MSH2*, *MSH6*, and *PMS2* are also the genes of which mutated alleles are related to the Lynch syndrome (Lynch et al. 2006), with an incidence of 1:1000 in the general population. Moreover, all four genes are expressed in oocytes and embryos of rhesus monkeys (Zheng et al. 2005), indicating a key function also in the human germline (Jaroudi and SenGupta 2007). In line with our predictions, we thus found that the MMR genes, which reportedly have the strongest effect on monorepeat stability, also contain the largest amount of such sequences. These findings contrast the conclusion of Chang et al. that monorepeats are particularly related to the "minor" components of MMR (Chang et al. 2001).

Our hypothesis also predicts that mutated MMR alleles should experience their own contraction bias more often than other regions of the genome. This effect of MMR deficiency has been extensively demonstrated in cancer cells (Sammalkorpi et al. 2007). In particular, MMR deficiencies have been directly related to contractions of the *BAT-26* microsatellite marker (also a monorepeat) located within *MSH2* (Boyer et al. 2002; de Leeuw et al. 2001; Zhou et al. 1997; Hoang et al. 1997). However, as homozygous and heterozygous germline mutations in MMR involve strong risk for early cancer, such alleles are probably short-lived in the population (Desai et al. 2000; Sun et al. 2005; Felton et al. 2007). A germline effect of the contraction bias on deficient MMR alleles may thus be hard to detect and has not been tested for in this study, as full genomic sequences of mutated MMR alleles are presently unavailable.

Chang et al. (2001) have argued that "the exceptional density of microsatellites in the minor MMR genes represents a genetic switch that allows the adaptive mutation rate to be modulated over evolutionary time." This hypothesis cannot explain the excess in monorepeats in non-coding regions within and around MMR genes, several of which have a major role in the prevention of genetic instability and cancer. Nor can it explain the striking association between the mutation bias of the MMR phenotype and repeat content in the MMR genotype. Based on the proposed evolutionary mechanism, we therefore argue that the overrepresentation of monorepeats within and around the MMR genes is maintained by the MMR mechanism.

The population frequency of MMR deficient alleles, including complete as well as partial loss of function, is unknown as we generally only recognize the polymorphisms that cause disease. Nor do we know the effect of human MMR on the germline mutation rate. However, based on the presented model, we argue that the high repeat content in MMR regions may be explained by less than 100-fold difference in microsatellite mutation rate between the MMR wild-type and the heterozygous mutant. This level of instability is in the lower range of that observed in MMR deficient tumors (Lynch et al. 2006; Sammalkorpi et al. 2007) and in the germline of MMR deficient and insufficient mice (Larson et al. 2004; Gurtu et al. 2002).

Most interestingly, the study by Larson et al. (2004) suggests that embryos formed from PMS2-deficient eggs have a strong increase in monorepeat mutation rate limited to the earliest stages of development. Heterozygous MMR mutations may thus have significant effect on germline mutation rate, even though the resulting offspring is phenotypically normal. It is therefore interesting to speculate that a similar maternal effect occurs in the human germline.

Moreover, the proposed evolutionary mechanism might be related to the phenomenon of genetic anticipation in Lynch syndrome, i.e., the observation that the disease occurs at an earlier age in successive generations (Nilbert et al. 2009). As the MMR proteins maintain the length of monorepeats within their own nucleotide sequences, they establish a network of self-sustaining loops propagating through the generations. Although the high content of monorepeats makes the MMR genes vulnerable to MMR deficiency, the interdependency of gene and protein may be understood as a stable evolutionary strategy. When a loop is broken, however, it triggers a cascade of events leading to accumulated breakdown of the regulatory network and increasing cancer risk through the generations.

In conclusion, we demonstrate an overrepresentation of monorepeats within and around the MMR genes, and provide an evolutionary and mechanistic explanation to this paradox. In brief, we argue that the MMR proteins have shaped the sequence composition of their own alleles. This concept challenges the dogma that flow of information is unidirectional from DNA to protein (Thieffry and Sarkar 1998; Crick 1970), but is based on simple deduction from well-established molecular mechanisms. In theory, the concept is applicable to any protein that either directly or indirectly affects the nucleotide composition. Other DNA repair genes may also induce mutation biases leading to accumulation of particular sequences within the genome (Pearson et al. 2005; Burt and Trivers 2006). Further testing of the hypothesis will thus require a systematic mapping of sequence-modifying phenotypes and their respective genotypes.

Acknowledgments J.B. conceived and developed the theoretical model, initiated the project, interpreted results, and wrote the manuscript. D.S.F. and E.A.R. developed the theoretical model and the methodology for testing it, performed bioinformatics analysis, interpreted results and wrote the manuscript. E.A.R. developed the mathematical model for monorepeat evolution. M.B.-O. contributed to developing the theoretical model and interpreted results. S.N. developed methodology and performed bioinformatic analysis. All authors discussed the results and commented on the manuscript. We thank Andrés Ögmundsson for technical assistance and Eivind Hovig for insightful comments on the manuscript.

References

Ackermann M, Chao L (2006) DNA sequences shaped by selection for stability. PLoS Genet 2:224–230

Alazzouzi H, Domingo E, Gonzalez S, Blanco I, Armengol M, Espin E, Plaja A, Schwartz S, Capella G, Schwartz SJ (2005) Low levels of microsatellite instability characterize MLH1 and MSH2 HNPCC carriers before tumor diagnosis. Hum Mol Genet 14:235–239

Baer CF, Miyamoto MM, Denver DR (2007) Mutation rate variation in multicellular eukaryotes: causes and consequences. Nat Rev Genet 8:619–631

Baida A, Lopez A, Marcos R, Velazquez A (2003) Germline mutations at microsatellite loci in homozygous and heterozygous mutants for mismatch repair and PCNA genes in Drosophila. DNA Repair 2:827–833

Borstnik B, Pumpernik D (2002) Tandem repeats in protein coding regions of primate genes. Genome Res 12:909–915

Bouffler SD, Hofland N, Cox R, Fodde R (2000) Evidence for Msh2 haploinsufficiency in mice revealed by MNU-induced sister-chromatid exchange analysis. Br J Cancer 83:1291–1294

Boyer JC, Yamada NA, Roques CN, Hatch SB, Riess K, Farber RA (2002) Sequence dependent instability of mononucleotide microsatellites in cultured mismatch repair proficient and deficient mammalian cells. Hum Mol Genet 11:707–713

Burt A, Trivers R (2006) Genes in conflict: the biology of selfish genetic elements. Belknap Press, Cambridge

Chang DK, Metzgar D, Wills C, Boland CR (2001) Microsatellites in the eukaryotic DNA mismatch repair genes as modulators of evolutionary mutation rate. Genome Res 11:1145–1146

Choi YH, Cotterchio M, McKeown-Eyssen G, Neerav M, Bapat B, Boyd K, Gallinger S, McLaughlin J, Aronson M, Briollais L (2009) Penetrance of colorectal cancer among MLH1/MSH2 carriers participating in the colorectal cancer familial registry in Ontario. Hered Cancer Clin Pract 7

Crick F (1970) Central dogma of molecular biology. Nature 227:561–563

de la Chapelle A (2005) The incidence of Lynch syndrome. Fam Cancer 4:233–237

de Leeuw WJF, van Puijenbroek M, Merx R, Wijnen JT, Brocker-Vriends AHJT, Tops C, Vasen H, Cornelisse CJ, Morreau H (2001) Bias in detection of instability of the (C)8 mononucleotide repeat of MSH6 in tumours from HNPCC patients. Oncogene 20:6241–6244

Desai DC, Lockman JC, Chadwick RB, Gao X, Percesepe A, Evans DGR, Miyaki M, Yuen ST, Radice P, Maher ER, Wright FA, de la Chapelle A (2000) Recurrent germline mutation in MSH2 arises frequently de novo. J Med Genet 37:646–652

Dunlop MG, Farrington SM, Nicholl I, Aaltonen L, Petersen G, Porteous M, Carothers A (2000) Population carrier frequency of hMSH2 and hMLH1 mutations. Br J Cancer 83:1643–1645

Ellegren H (2002) Mismatch repair and mutational bias in microsatellite DNA. Trends Genet 18:552

Ellegren H (2004) Microsatellites: simple sequences with complex evolution. Nat Rev Genet 5:435–445

Felton KEA, Gilchrist DM, Andrew SE (2007) Constitutive deficiency in DNA mismatch repair. Clin Genet 71:483–498

Gurtu VE, Verma S, Grossmann AH, Liskay RM, Skarnes WC, Baker SM (2002) Maternal effect for DNA mismatch repair in the mouse. Genetics 160:271–277

Harr B, Todorova J, Schlotterer C (2002) Mismatch repair-driven mutational bias in *D. melanogaster*. Mol Cell 10:199–205

Hoang JM, Cottu PH, Thuille B, Salmon RJ, Thomas G, Hamelin R (1997) *BAT-26*, an indicator of the replication error phenotype in colorectal cancers and cell lines. Cancer Res 57:300–303

Jaroudi S, SenGupta S (2007) DNA repair in mammalian embryos. Mutat Res 635:53–77

Karlin S, Mrazek J, Campbell AM (1998) Codon usages in different gene classes of the *Escherichia coli* genome. Mol Microbiol 29:1341–1355

Kashi Y, King DG (2006) Simple sequence repeats as advantageous mutators in evolution. Trends Genet 22:253–259

Kelkar YD, Tyekucheva S, Chiaromonte F, Makova KD (2008) The genome-wide determinants of human and chimpanzee microsatellite evolution. Genome Res 18:30–38

Lai Y, Sun F (2003) The relationship between microsatellite slippage mutation rate and the number of repeat units. Mol Biol Evol 20:2123–2131

Larson JS, Stringer SL, Stringer JR (2004) Impact of mismatch repair deficiency on genomic stability in the maternal germline and during early embryonic development. Mutat Res 556:45–53

Li YC, Korol AB, Fahima T, Nevo E (2004) Microsatellites within genes: structure, function, and evolution. Mol Biol Evol 21:991–1007

Lyer RR, Pluciennik A, Burdett V, Modrich PL (2006) DNA mismatch repair: functions and mechanisms. Chem Rev 106:302–323

Lynch HT, Boland CR, Gong G, Shaw TG, Lynch PM, Fodde R, Lynch JF, de la CA (2006) Phenotypic and genotypic heterogeneity in the Lynch syndrome: diagnostic, surveillance and management implications. Eur J Hum Genet 14:390–402

Manley K, Shirley TL, Flaherty L, Messer A (1999) Msh2 deficiency prevents in vivo somatic instability of the CAG repeat in Huntington disease transgenic mice. Nat Genet 23:471–473

Marti TM, Kunz C, Fleck O (2002) DNA mismatch repair and mutation avoidance pathways. J Cell Physiol 191:28–41

McVean GA, Myers SR, Hunt S, Deloukas P, Bentley DR, Donnelly P (2004) The fine-scale structure of recombination rate variation in the human genome. Science 304:581–584

Nilbert M, Timshel S, Bernstein I, Larsen K (2009) Role for genetic anticipation in Lynch syndrome. J Clin Oncol 27:360–364

Ohmiya N, Matsumoto S, Yamamoto H, Baranovskaya S, Malkhosyan SR, Perucho M (2001) Germline and somatic mutations in *hMSH6* and *hMSH3* in gastrointestinal cancers of the microsatellite mutator phenotype. Gene 272:301–313

Pearson CE, Edamura KN, Cleary JD (2005) Repeat instability: mechanisms of dynamic mutations. Nat Rev Genet 6:729–742

Peltomaki P (2001) DNA mismatch repair and cancer. Mutat Res 488:77–85

Peltomaki P (2005) Lynch syndrome genes. Fam Cancer 4:227–232

Perucho M (1996) Microsatellite instability: the mutator that mutates the other mutator. Nat Med 2:630–631

Sammalkorpi H, Alhopuro P, Lehtonen R, Tuimala J, Mecklin JP, Jarvinen HJ, Jiricny J, Karhu A, Aaltonen LA (2007) Background mutation frequency in microsatellite-unstable colorectal cancer. Cancer Res 67:5691–5698

Shah SN, Hile SE, Eckert KA (2010) Defective mismatch repair, microsatellite mutation bias, and variability in clinical cancer phenotypes. Cancer Res 70:431–435

Sleckman BR (2005) Lymphocyte antigen receptor gene assembly—multiple layers of regulation. Immunol Res 32:253–258

Sniegowski PD, Gerrish PJ, Johnson T, Shaver A (2000) The evolution of mutation rates: separating causes from consequences. Bioessays 22:1057–1066

Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G, Cooke MP, Walker JR, Hogenesch JB (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. Proc Natl Acad Sci USA 101:6062–6067

Subramanian S, Mishra RK, Singh L (2003) Genome-wide analysis of microsatellite repeats in humans: their abundance and density in specific genomic regions. Genome Biol 4:R13

Sun S, Greenwood CMT, Thiffault I, Hamel N, Chong G, Foulkes WD (2005) The HNPCC associated MSH2*1906G>C founder mutation probably originated between 1440 CE and 1715 CE in the Ashkenazi Jewish population. J Med Genet 42:766–768

Thieffry D, Sarkar S (1998) Forty years under the central dogma. Trends Biochem Sci 23:312–316

Venkatesan RN, Bielas JH, Loeb LA (2006) Generation of mutator mutants during carcinogenesis. DNA Repair 5:294–302

Wanner RM, Guthlein C, Springer B, Bottger EC, Ackermann M (2008) Stabilization of the genome of the mismatch repair deficient *Mycobacterium tuberculosis* by context-dependent codon choice. BMC Genom 9:294

Zhang SL, Lloyd R, Bowden G, Glickman BW, de Boer JG (2002) Msh2 deficiency increases the mutation frequency in all parts of the mouse colon. Environ Mol Mutagen 40:243–250

Zheng P, Schramm RD, Latham KE (2005) Developmental regulation and in vitro culture effects on expression of DNA repair and cell cycle checkpoint control genes in rhesus monkey oocytes and embryos. Biol Reprod 72:1359–1369

Zhou XP, Hoang JM, Cottu P, Thomas G, Hamelin R (1997) Allelic profiles of mononucleotide repeat microsatellites in control individuals and in colorectal tumors with and without replication errors. Oncogene 15:1713–1718