Transcription Related Genetic Variation in Human Genome

Giovanni Scala

Dottorato in Biologia Computazionale e Bioinformatica
XVII Ciclo
Università degli Studi di Napoli "Federico II"

Tutor: Prof. Gennaro Miele

Co-tutor: Prof. Sergio Cocozza

March 2015

Table of contents

List of figures

Chapter 1

Introduction

1.1 Background

Single nucleotide polymorphism (SNP) is a DNA sequence variation in a single position of a genome, for which individuals of the same specie differ. Despite their simplicity (if compared with other structural variations like insertions and deletions), SNPs can have relevant functionl effects and, in some cases, can directly or indirectly cause phenotypic abnormalities or diseases. This kind of variation represents ~90% of human genetic variants and, consequently, is the main source of phenotypical differences among individuals. As many other structural variations, SNPs can be heritable, meaning that they can spread in a population and reach different frequencies in the course of time. As well as other functional and heritable variants, SNPs can be subject to evolutionary forces (e.g. natural selection) that alter their frequency by increasing (positive selection) or decreasing (purifying selection) their occurrences in a population of individuals.

Given the importance of these objects, researchers spent a lot of efforts investigating the mechanisms that generate SNPs in a genome. It is proved that SNPs, like several other genetic variants, mostly arise as consequence of mutational phenomena and, in particular, from damages to DNA that alter its original sequence. Such damages are usually repaired by specialised protein complexes that, in the majority of cases, are able to restore the original conformation of the damaged sequence. Nevertheless, there are cases in which such mechanisms fail to restore the original configuration of the damaged DNA sequence and, consequently, a different and permanent configuration (a mutation) arises. If the last described event happens in germline cells, or during the very early stage of cell differentiation, the resulting mutation has the possibility to be become heritable.

Several kind of mutational phenomena exists and they differ on various aspects, such as the DNA region where they occur and the particular mechanisms that triggered them. In particular, mutations can occur as a result of natural processes in the cell ("spontaneous mutations") or by the interaction of DNA with external agents or mutagens ("induced mutations"). Spontaneous mutations can arise during DNA replication events, when incorrect nucleotides are added by the DNA polymerase during DNA copy, or by base alteration and base damage events such as tautomerization, deamination or oxidation phenomena.

Past studies revealed that the mutational spectrum of human genome varies from region to region [1]. The main causes of this variability are related to the local DNA base composition (e.g. the CpG effect), the functional role, and the rate of occurrence of some cellular activities that can cause damages to DNA [1].

In general, mutations have a very low probability to have functional effects.

In fact, in the majority of the cases, in order to have functional effects, a mutation has to alter in a specific way the transcript of some gene and most importantly has to occur in a coding region of the genome.

In human these regions represent only a small fraction of the genome (about 1%). Nonetheless, the transcribed regions of a genome are of particular relevance in the investigation of genetic variation. These loci contain all the information needed by cellular processes to correctly build up proteins. For this reason, the vast majority of mutations with functional effects can be found in these regions.

Coding regions of the genome were shown to be characterised by a lower mutation rate [2], this is probably because natural selection has led to the formation of particular mechanism and/or DNA conformations that protect them from mutational events.

However, transcribed regions, compared to other genomic ones, are generally subject to a higher number of molecular activities, mainly related to DNA transcription. The nature of these molecular activities and the higher rates of occurrence that characterise them in these regions are taught to have a detrimental effect on the underlying DNA region over time [3].

This is particularly true for transcriptional activities, that in eukaryotes involve a series of molecular factors, each one acting on DNA with different tasks: from the unwinding of the double helix, to the strand elongation, to the cleavage of the nascent transcripts. These transcriptional units sometimes act concurrently on the same region with the possibility of collision events [3] and consequent DNA damages events like double strand breaks.

For these reasons, it is of actual interest to study the mutational spectrum of coding regions and how this variability is tolerated in an evolutionary view.

SNPs, and in particular their genomic distribution, represent a good starting point to investigate these aspects, mostly because they generally arise from mutations and can be considered as markers of past mutational events. The idea to analyse the mutational spectrum of a genomic region by looking at the SNPs distribution is not new [4] but it has some appealing advantages. First of all, SNPs tend to accumulate over time in a population, giving the possibility to observe mutational signatures even in those regions where mutational events are less frequent Second, SNPs that are characterised by different ages could be used as "snapshots" over time of their distribution in a population, hence giving hints on the evolutionary forces acting on them.

1.1.1 Transcript-centric mutations

The relationship between transcription and DNA mutation was studied by Cui et al. [4]. The authors of this paper investigated the relationship between sequence signatures of genetic variations in human and transcription processes and chromosomal structures. In this study, the distributions of single nucleotide polymorphisms from NCBI dbSNP (snp128) was analysed along the length of the transcript for a set of ubiquitous expression-invariable genes in human. For these regions, Cui et al. revealed the existence of a relationship between gene expression and genetic variation, by observing an high statistically significant correlation between gene expression and SNP density. Moreover, they show that the variant density in these regions was characterised by a periodical distribution that was, in turn, related with nucleosomal occupation. Interestingly, they found that a gradient of SNP density exists in these regions, with a peak in the distribution of SNPs near transcription start sites that tamper off toward 3' end. Notably, this is one of the first works pointing out the existence of a correlation between transcription related activity and the genetic variability of a genomic region.

The work of Cui. et al. opens the doors to a series of interesting analyses that can be performed to confirm their hypothesis and expand the current knowledge on the nature of genetic variation in coding regions. Two main advances in current technology can be exploited to this aim: the availability of new and more complete variation data and the availability of new datasets dedicated to the structural and functional analysis of DNA (e.g. ENCODE project). NCBI dbSNP now is released on version 141 (snp141) and with respect to the version used in [4] benefits of the variant data submitted by a massive, multi-population whole genome sequencing projects: the 1000 Genomes project.

1.1.2 1000 Genomes data

1000 Genomes project is a massive sequencing project that aims to provide a deep characterisation of human genome sequence variation. This was accomplished through the combination of low-coverage sequencing, array-based genotyping and deep targeted resequencing of coding regions in 2500 individuals from five large regions of the world (Europe, East Asia, South Asia and West Africa).

This project led to creation of a public reference database for DNA variants that is 98% complete (in the phase 1 release) for polymorphism that have allele frequency greater or equal to 1% in related populations. Variant data from this project was released in several phases:

- the “pilot phase”, whose goal was to develop and compare different strategies for genome-wide sequencing with high throughput platforms, includes data from 553 individuals;

- the “phase 1” includes data from 1092 individuals;

- the “final phase discovery” includes data from 2535 individuals;

- the “final release” represents a refinement on samples that do not progress on the final phase due to quality control or changing criteria with respect to complete samples and consists of 2504 samples.

1000 Genomes SNPs give the possibility to analyse the relationship between transcription and variation with an exceptionally high level of resolution. The main difference between 1000 Genomes project and studies that contributed to older projects on the same topic, relies in the characterisation of variants through a whole genome sequencing approach and in the high number of assayed samples.

An important aspect, deriving from the huge sample size, is that newly discovered variants are mostly classified as low-frequency SNPs (with a minor allele frequency < 0.05). These variants tend to be more recent in their origin than high-frequency ones and, hence, to be more closely related to the mutational process. This gives the possibility to investigate the distribution of novel mutations with a substantial reduction of the bias introduced by selection phenomena. Moreover, the huge number of available SNPs from 1000 Genomes, leads to the possibility to analyse separately variants that are characterised by different frequencies without loosing too much resolution. This is useful to investigate separately either the mutational forces, that affect mainly the distribution of low frequency variants, and the evolutionary ones, whose effects are better recognisable in the distributions of higher frequency variants.

1.1.3 Structural and functional DNA dataset

The availability of datasets dedicated to the structural and functional analysis of DNA, gives the possibility to study the relationship between genetic variations and several interesting structural aspects. These datasets can be broadly divided in two classes: those containing structural information and those containing information about functional roles of DNA sequences.

Structural information about DNA is mainly obtained in two ways: by direct assay of particular DNA related structures or by computational inference using existent experimental data. The Encyclopedia of DNA Elements (ENCODE) [5] project provides several examples of detection of DNA structural information through direct assay and the use of next generation technologies. Nucleosome positioning maps are obtained in ENCODE for several cell lines through the combination of chromatin immunoprecipitation (ChIP) with massively parallel DNA sequencing. This gives the possibility to analyse the local nucleosome landscape with a high level of resolution compared with the past.

Other structural features of interest are indirectly inferred by analysing and/or integrating exiting data, mainly obtained through next generation sequencing techniques. This is the case of human transcription start sites from Switch Gear Genomics database (http://www.switchgeargenomics.com) and polyadenylation sites from PolyA_DB [6]. These loci were inferred (along with a level of certainty on the prediction) by analysing cDNA/expressed sequence tags data using suitable prediction algorithms. Although these maps are not obtained by direct assay of the underlying molecular structure, they can be used in downstream analyses that do not require a strict exact matching of the predicted site with the real one.

Functional characterisation of DNA could be useful to investigate the effect of mutations in transcription related sites. Several datasets are currently available to investigate this point. Genomic Evolutionary Rate Profiling (GERP) is a method to evaluate the conservation evolutionary constraint intensity of a genomic locus. Through this dataset it is possible to study the distribution of variants from an evolutionary perspective, distinguishing regions where purifying selection is more active from those that evolved neutrally.

Combined Annotation Depletion Score (CADD) [7] is a novel method to infer the potential deleteriousness of a mutation in a genomic sites. This score gives the possibility to analyse the potential deleteriousness of mutations at single base resolution.

In summary, transcription start sites, polyadenylation sites, and 1000 Genomes variant datasets contain a good amount of information to study the localisation and distribution of transcription-related variants. Additionally, the described functional and structural datasets contain information to investigate on some of the most important forces that generate and shape variability around these sites.

1.2 Aims

The aim of this study was to develop a set of statistical tools to investigate, compare and correlate the distribution of genomic signals around a set of DNA loci related to the transcriptional process. The developed techniques are able to handle and define the distribution of two type of signals: those related to measurements of a value in a locus (quantitative signals) and those (binary signals) related to the fulfilment in a locus of a particular condition, such as the presence of a SNP. The generation and the analysis of these distributions will be the main topic of this thesis work.

Methods were defined to evaluate sub-regions where the analysed distributions deviate from suitably defined neutral models, or sub-regions where two distributions are statistically different. Finally, suitable procedures to compare and correlate the distribution of different signals were provided.

This set of techniques will be employed to refine and extend the knowledge on mutation and population variation around transcription-related sites in human. To this end, two sets of human genomic loci were considered: transcription start sites (TSSs) and polyadenylation sites (PSs).

These two sets of loci are of crucial importance with regard to the transcriptional activity. In order to capture as much as possible of the relevant phenomena related with genetic variation, a 10-kb region surrounding the above mentioned sites was considered in the following analyses.

The distribution of human SNPs around these two set of loci was studied by considering SNPs from 1000 Genomes phase 1. These variants were split into four frequency classes: rare, common and two other middle classes. The relationship with a structural factor, known to be implicated in mutagenesis, was studied analysing the distribution of nucleosome occupation in the same region and its relationship with the four SNP distributions. The eventual action of evolutionary forces, that shape the variation around transcription related sites, was investigated by analysing the distribution of the GERP conservation scores.

Finally, the potential pathogenic effects of the analysed variants were investigated by analysing the distribution of the associated deleteriousness scores from the CADD dataset.

1.3 Contents

In Chapter 2 the methodic developed to analyse genomic signals related with the DNA transcriptional process will be described. In particular, the procedure to map a set of genomic sites of interest to a corresponding set of genomic regions will be illustrated. These regions will be provided with a coordinate system that is based on the DNA transcription verse. The methods to derive the distribution of both quantitative and binary signals in the region of interest will be discussed. Next, it will be shown the procedure to search for subregions where the observed distribution has a statistically relevant deviation from a corresponding neutral expectation. Also procedures to investigate subregions of statistically significant deviation or to investigate correlation among distribution will be discussed.

In Chapter 3 the application of the developed methodic to human TSSs will be illustrated. In particular, the study will consist in a genome-wide analysis of the distribution of four classes of SNPs (rare, two intermediate classes, and common) inside the 10 kb region flanking human TSSs. It will be shown how the distribution of variants depends on their frequency and on their localisation relative to the TSS. Also, the relationship between variant distribution in transcription start sites and the presence of an enclosing CpG islands will be analysed. Following findings from previous studies, the relationship between the distribution of rare variants and nucleosome occupancy score will be investigated. Evolutionary (purifying selection) and non-evolutionary (biased gene conversion) forces will be related with SNP frequency and the potential pathogenicity of each class of variant will be analysed.

Finally, the results of this analysis will be discussed and related to previous findings in literature.

In Chapter 4 the application of the developed methodic to human polyadenylation sites will be described. As for TSSs, the study will be performed through a genome-wide analysis of the distribution of four groups of SNPs (rare, two intermediate classes, and common) inside the 10 kb region flanking human PSs. The relationship between the distribution of variants, their frequency and/or their localisation relative to the PS will be investigated. The relationship between rare variants and nucleosome occupancy will be also investigated. It will also be shown how evolutionary forces play a role in determining the relative SNP frequency nearby PSs. As for TSSs, an analysis of the potential pathogenicity of each class of variants will be performed. Finally, the results of this analysis will be discussed and related to previous findings in literature.

In Chapter 5 the overall results of the described work will be discussed and my conclusions will be presented. Possible future extensions of the presented results will be presented along with other interesting cases of study where the described techniques can be applied.

Chapter 2

Analysis of the distribution of signals around a genomic site

This chapter describes the general technique designed to study the distribution of signals in regions flanking a set of genomic sites of interest.

Firstly, it will be described the procedure to map, bin and orient (following the transcriptional process) the genomic regions flanking the sites under study. Then, it will be presented the procedure to obtain a signal (along with its confidence interval) representing the spatial distribution of a measurement (dichotomous or quantitative) in these regions. It will be also shown how to search for subregions where the considered signal is statistically different from a neutral model or for regions where two or more signals are statistically different. Finally, it will be shown how to search for possible cross-correlations between signals and how to treat with overlapping regions to avoid biases in the described computations.

2.1 Definition of the regions

In this section it will be shown how to map a set S of N human transcription-related genomic sites on the corresponding set R of K (bp) flanking genomic regions. The elements

s of S will be defined by tuples of four values such that:

$$\forall s \in S,$$

s takes values in

$$\{1,2,\ldots,22,X,Y\} \times \mathbb{N} \times \mathbb{N} \times \{-1,1\}$$

For each s, the first value of the tuple represents the chromosome, the second and the third respectively the position of the first and the last bp of s, and the fourth represents the DNA strand (coded as -1 for the minus strand and +1 for the plus strand).

Hereafter, for a given site s_i, we will use the notation $chr(s_i)$, $start(s_i)$, $stop(s_i)$ and $strand(s_i)$ to denote the values of the corresponding tuple.

Given S and the size K of the flanking region to analyse, we can define the corresponding set R of genomic loci representing the K bp flanking regions of the elements of S. In particular, $\forall i \in \{1,\ldots,N\}$, each element r_i of R will be defined as a tuple of four values as follows:

$$r_i = [chr(s_i), \max(0, start(s_i) - K), \min(len(chr(s_i)), stop(s_i) + K), strand(s_i)]$$

Where $len(chr(s_i))$ represents the size in bp of $chr(s_i)$.

2.2 Orientation and binning of the regions

The next step is to divide regions of R in bins of a fixed size and to define an ordering among these bins. The binning of the region is useful to smooth the resulting signal when the data is too sparse along the analysed region.

Let bs the chosen size of the bins, and let $nb = \lceil K/bs \rceil$. For each $s_i \in S$, we will assign a positive index $\{1,\ldots,nb\}$, representing the distance of the bin from the site s_i, to the nb bins located downstream of s_i (with regard to the transcription verse) and, likewise, negative index $\{-1,\ldots,-nb\}$ to the nb bins located upstream of s_i.

Thus, $\forall i \in \{1,\dots,N\}$ and $\forall j \in \{-nb,\dots,-1,1,\dots,nb\}$ we will identify the sub-region of *bs* bases in position *j* with respect to the site s_i as $Bin(i,j)$.

In particular, $Bin(i,j)$ is univocally defined as follows.

Let $l_1 = (|j|-1) * bs$ and $l_2 = |j| * bs$, then:

$$Bin(i,j) = \begin{cases} Null\ if\ (stop(s_i) + l_1 > stop(r_i)\ \&\ j * str(r_i) > 0) \\ Null\ if\ (start(s_i) - l_1 < start(r_i)\ \&\ j * str(r_i) < 0) \\ [stop(s_i) + l_1, \min(stop(r_i), stop(s_i) + l_2)]\ if\ j * str(r_i) > 0 \\ [\max(start(r_i), start(s_i) - l_2), start(s_i) - l_1]\ if\ j * str(r_i) < 0 \end{cases}$$

Finally, notice that the use of a bin-size equal to 1 leaves the possibility to perform the analysis at a base pair level.

2.3 Computation of the signals

Once that the regions of interest (R_i) and the granularity (bs) of the basic components are defined, a value $Sig(i,j)$, representing the value of the signal under study in the component $Bin(i,j)$, can be associated to each $Bin(i,j)$.

Each $Sig(i,j)$ value will be computed from a set $V(i,j)$ of values defined for bases inside the $Bin(i,j)$ region. In order to compute a value for each bin j, we require $V(i,j) \neq 0$ for at least one i.

For each $j \in \{-nb,\dots,-1,1,\dots,nb\}$ we can then compute a single value $Sig(j)$ by summarising the values $Sig(i,j)$, over all $\{1,\dots,N\}$. The whole list of $2nb$ $Sig(j)$ values will constitute the signal around the class of sites under study.

In the following paragraph, it will be described the computation of the value of $Sig(i,j)$ starting from values contained in each $Bin(i,j)$. In particular, the computation of $Sig(i,j)$ for binary and quantitative values will be covered.

2.3.1 Analysis of dichotomous signals

The case of binary values is representative of all the situations for which a base pair can be characterised or not by a given property (e.g. to be SNP or not).

In this case, for each $Bin(i, j)$, we set $Sig(i, j)$ equal to the number of bp, located inside $Bin(i, j)$, for which the property of interest holds.

For a fixed bin j, we define $Sig(j)$ by averaging the value of $Sig(i, j)$ over all considered regions:

$$Sig(j) = \frac{\sum_{i=1}^{N} Sig(i, j)}{N}$$

Along with each $Sig(j)$ value, we can compute the corresponding variance as follows:

$$Var(j) = \frac{\sum_{i=1}^{N} [Sig(i, j) - Sig(j)]^2}{N}$$

and the corresponding standard error:

$$Se(j) = \sqrt{\frac{Var(j)}{N}}$$

This latter will be used, in the following, to investigate differences between two or among more signals, and to test for deviations of a signal from a neutral expectation.

2.3.2 Analysis of quantitative signals

Quantitative values are treated in a slightly different way. In this case, we define $SIGS(i, j)$ as the set of quantitative values (e.g. conservation scores) associated to the base pairs of the region $Bin(i, j)$ and $|SIGS(i, j)|$ its cardinality.

For each fixed bin j, we can compute $Sig(j)$ as the average of $SIGS(i, j)$ values over all bins, namely:

$$Sig(j) = \frac{\sum_{i=1}^{N} \sum_{s \in SIGS(i,j)} s}{\sum_{i=1}^{N} |SIGS(i, j)|}$$

As for dichotomous values, we can compute, for each j, the variance and standard error associated to $Sig(j)$:

$$Var(j) = \frac{\sum_{i=1}^{N} \sum_{s \in SIGS(i,j)} (s - Sig(j))^2}{\sum_{i=1}^{N} |SIGS(i,j)|}$$

and

$$Se(j) = \sqrt{\frac{Var(j)}{\sum_{i=1}^{N} |SIGS(i,j)|}}$$

2.4 Test for positional effects

In the previous paragraph, it was shown the procedure to compute a series of average signals around a set of genomic sites of interest.

The list of $Sig(j)$ values can be interpreted as an approximation of the average distribution of the analysed values in the flanking regions of the sites under study.

In many cases it is important to evaluate if, where and how much, this distribution differs from a defined neutral expectation.

In general, to perform such a task we need:

1. a list of observed values $S(j)$;
2. an estimate $E(j)$ of the uncertainty associated to each value $S(j)$;
3. a neutral model $M(j)$ for the considered signal;
4. a suitable summary statistic $ss(j)$ of the studied signal and its distribution under $M(j)$.

Given this set of items, we can asses the statistical significance of each observed value $S(j)$ by evaluating the probability of observing more extreme values than $ss(j)$, corresponding to $S(j)$ under the model $M(j)$.

In this specific case, the signal $S(j)$ is obtained, for each bin j, by averaging values over all $Bin(i,j)$ regions.

Thus, given a neutral model that generates values $N(J)$, with neutral standard errors $SE_n(j)$, we can assess the statistical significance of $S(j)$ using two tailed Student's T-test over the values $S(j), Se(j), N(j)$ and $SE_n(j)$.

This procedure leads to execution of $2nb$ tests of the same null hypothesis that in turn generate a list of p-values $p(j)$ with $j \in \{1, \ldots, 2nb\}$.

We can use the resulting list of $p(j)$ values to:

- assess if the observed signal is generally significant from the neutral model;
- delimitate subregions of statistically strong deviation from neutrality.

In the first case we want to verify the hypothesis that the two signals (the observed and the neutral) are statistically different as a whole.

Since we can consider each $p(j)$ as the output of an independent test of the null hypothesis that $N(j)$ and $S(j)$ are sampling the same distribution, we can make use of the Fisher's combined probability test. This test combines p-values from each of the above mentioned tests, into one test statistic χ^2 using the formula:

$$\chi^2_{4nb} \sim -2 \sum_{j \in \{1 \ldots 2nb\}} \ln(p(j))$$

In the second case we want to identify contiguous regions of bins whose p-values are all under a significance threshold $sigT$.

To do this, we need to consider each p-value separately. Thus we first need to adjust $p(j)$ values into $p_{corr}(j)$ values with a suitable correction procedure (e.g. Bonferroni or False Discovery Rate). Once a threshold k on the minimum length for a sub-region of interest is chosen, we can define a sub-region of bins $j_1, j_2, \ldots, j_m$ as non neutral if:

$$\begin{cases} m >= k \\ j_{i+1} - j_i = 1, \forall i \in \{1, \ldots, m-1\} \\ p_{corr}(j_i) < sigT, \forall i \in \{1, \ldots, m\} \end{cases}$$

2.5 Compare the distribution of a signal among groups

In addition to test the difference between a signal and a corresponding neutral model, we may be also interested in the comparison of two or more signals $S_i(j)$, with $i \in \{1,\dots,k\}$ and $k >= 2$.

Following an analog approach as described for the neutrality test, we can test for differences among signals using their value $S_i(j)$ and the associated error $Se_i(j)$.

In the case of averages, let k be the number of signals to be compared.

For each bin j, we can test if the values $S_i(j)$ are samples of the same distribution using the Student's T-test if $k = 2$ or the ANOVA if $k > 2$. As before, this procedure will result in $2nb$ type 1 errors, that can be used to asses the general difference among the whole signals or to delimitate sub-regions of statistically significant difference.

In the first case, we will employ the Fisher's method. In the second case, we will adjust values for multiple testing and, given a suitable threshold, we will identify a set of sub-regions, characterised by distinct signal values, following the same procedure described in the previous section.

2.6 Correlation analyses between distributions

Finally, one may be also interested in the discovery of possible relationships between signals of different nature.

This can be accomplished by means of correlation analyses between the dependent signal S_1 and the independent one S_2. In particular, we can evaluate linear dependency between the two signals by computing Pearson product-moment correlation coefficient ρ.

We can then assess the statistical significance of this value by evaluating the following statistic:

$$t = \frac{\rho\sqrt{2nb-2}}{\sqrt{1-\rho^2}}$$

Under the null hypothesis of no-correlation between the two tested variables S_1 and S_2, t will be distributed as a Student's T with $2nb-2$ degrees of freedom.

2.7 Handling overlapping regions

In this last section, the possible procedures to treat overlapping genomic regions will be illustrated.

Analysing a set R of genomic intervals, generated from a set of N transcription related genomic sites S, we can find a subset $R_{over} \subseteq R$ such that $\forall r_i \in R_{over}\ \exists j$ characterised by:

$$\begin{cases} j \neq i \\ r_j \in R_{over} \\ \left[stop(r_i) > start(r_j)\ \&\ start(r_i) < stop(r_j)\right] \mid \left[stop(r_j) > start(r_i)\ \&\ start(r_j) < stop(r_i)\right] \end{cases}$$

In this case, we will have some bins $Bin(i,j)$ and $Bin(k,l)$ (with $i \neq k$) that cover the same genomic region, but represent different localisations relative to the sites s_i and s_k. Consequently, these sites will be considered multiple times, but in different locations, in the computation of the signal to study.

This can happen when the size k of the chosen flanking region results to be too large with respect to the genomic distances (in terms of bp) of the sites contained in S.

To avoid eventual biases deriving from overlapping regions, one can reduce k in order to reduce the size of the resulting set R_{over}. The drawback is that extremely small values of k could weaken the analysis power to detect interesting phenomena.

The golden rule to apply would be to choose a value for k that minimises the occurrence of this kind of phenomena and that does not shrink too much the considered region, thus preserving the explorative power of the analysis.

Once a good value for k is chosen, we can decide to treat regions in R_{over} by either:

- leaving the analysis run on R as it is, hence considering multiple times elements of R_{over};

- removing from R regions belonging to R_{over} before starting the analysis.

The first option is suitable when the number of overlapping regions is negligible or when we want to take in account the effect of the different relative positing of overlapping elements in R_{over}. For example, if the sites of interest are transcription start sites we can have these regions being localised on the 5' side of gene as well as on the 3' side of another gene, hence acting as both promoter or coding sites. Of course, in this case when calculating averages one must take into account the presence of this kind of regions and, consequently, of the same signal multiple times.

The second option is more conservative and avoids the possible noise generated by multiple different forces acting on the same site. Moreover it can be employed to quantify the effects of overlapping regions, in terms of value and variability of the resulting signal, when compared with results obtained using the first procedure.

Chapter 3

Genetic variation around human Transcription Start Sites

3.1 Description of the study

The transcription start site (TSS) is canonically defined as the first nucleotide of a transcribed DNA sequence where RNA polymerase begins synthesising the RNA transcript.

TSSs are surrounded by cis-acting regulatory sequences, including core and proximal promoter elements located within 1 kb of the TSS, as well as distal promoter elements.

Most mammalian promoters are enriched for GC-rich regions, also called CpG islands (CGIs), that serve as structural and functional punctuation marks for transcription [8]. The high GC content of sequences around the TSSs of genes suggests a functional relevance for GC-rich elements in higher eukaryotes [8].

Moreover, enrichment of GC-rich regions has been implicated in mutational bias, gene conversion bias and structural plasticity associated with transcription [9–13]. Indeed, GC-rich mammalian genes exhibit up to 100-fold greater transcription rates than orthologous GC-poor genes [14].

Populations harbour many genetic variations in promoter regions, most of which are single nucleotide polymorphisms [15]. In [4], by analysing the distribution of SNPs along the length of transcripts, Cui et al. found that SNPs were more abundant near TSSs, with their abundance tampering off toward the 3' end. Moreover, evolutionary studies show that different types of nucleotide substitutions occur with widely varying rates that may reflect biases intrinsic to mutation and repair mechanisms [16]. These mutations are thought to be strand-asymmetric and context-dependent (CpG vs. non-CpG) [17–19].

In this context, packaging of DNA into nucleosomes can limit the accessibility of regulatory proteins involved both in transcriptional regulation [20–22] and in replication and, more importantly, in DNA repair [23–25]. Also, in vivo studies with yeast and mammalian cells revealed that the DNA packaging into nucleosomes might impair the efficient repair of DNA damage [26–29], promoting the emergence of novel mutations.

Novel mutations act as both the substrate for evolution and the cause of genetic disease. Under a neutral model, all the new regulatory mutations should have an equal probability of fixation [30]. In general, evolutionary forces, such as natural selection, drive new alleles either towards a loss from the population (purifying selection), or towards an increasing frequency and possible fixation (positive selection). Given the particular genomic landscape that characterises promoters and given that most TSSs tend to be conserved between mammals [31, 32], one can imagine that these regions are generally subjected to selection forces, at least of a conservative kind.

In this chapter it will be discussed the first of the two cases of study where the techniques described in Chapter 2 are applied. In particular, human TSSs have been considered as the set of genomic sites of interest and the occurrence of SNPs in their 5Kb flanking region as the main signal to study. Also, the relationship of SNPs occurrence in human with their frequency and other genomic factors, namely the presence of a surrounding CpG island, nucleosome occupation, evolutionary conservation, biased gene conversion and potential deleterious effects of SNPs in the region was investigated.

3.2 Distribution of variant around TSSs

The distribution of genetic variants around human TSSs, was studied by analysing their frequency in the 10 kb region surrounding each start site. TSSs genomic coordinates were downloaded from the University of California, Santa Cruz (UCSC) database. To adopt a conservative approach, only TSSs having a confidence score ≥ 20 were selected, obtaining a total of 27,487 TSSs. Since previous studies showed that the CpG context of TSSs is related to a different mutagenic load and to different evolutionary constraints [33–35], TSSs were divided according to their location inside a CGI. In particular, the genomic coordinates of 27,718 unique CGIs were retrieved UCSC "CpgIslandExt" track and used to divide TSSs into two groups:

- CGI-TSSs, defined as TSSs that were inside a CGI (14,561, ~53% of the selected TSSs);
- nCGI-TSS, defined as TSSs located outside CGIs (12,926, ~47% of the selected TSSs).

CGI-TSSs were, on average, symmetrically located inside CGIs. No significant asymmetries were found in the distributions of the distances from the center of the CGI (D'Agostino skewness test).

Human genetic variant data were obtained from the dbSNP version 138 database. To achieve robust and comparable allele frequencies, only bi-allelic single nucleotide variants with available frequency values and detected by whole genome sequencing approach in the 1000 Genomes Phase 1 Variant Catalog, were selected.

These criteria led to 2,550,709 SNPs considering the only ones located in the 10 kb region surrounding the above selected TSSs.

This set of variants was then split, according to MAF values, into four classes: rare, mid1, mid2 and common. The reason behind this choice was to appreciate the eventual relationships between the population frequency and the relative positioning of a variant with respect to the transcriptional process.

Rare variants (~20% of all selected variants) were defined by a MAF less than or equal to 4.59×10^{-4}. This threshold corresponds to the lowest possible MAF value present in the 1000 Genomes Phase 1 release, and corresponds to a heterozygous variant being present in only one individual among all 1092 individuals from the Phase 1. Common variants (~33% of the selected variants) were defined according to the canonical criterion of a MAF value greater than 0.01. The remaining variants, with intermediate frequency between rare and common, were partitioned in two groups of approximately equal size and were referred as “mid1” (frequency range: $4.59 \times 10^{-4} - 0.0014$, ~23% of all selected variants) and “mid2” (frequency range: $0.0014 - 0.01$, ~24% of all selected variants).

Then, for each TSS, the surrounding 10 kb region was divided into 200 bins of 50 bp and for each SNP frequency class the corresponding normalised mean variant frequency (hereafter denoted by BVF) was calculated for each bin by following the technique described in Materials and Methods section.

Figure 3.1 shows confidence intervals of BVF values for the four frequency classes and for CGI-TSSs and nCGI-TSSs. Several peaks and/or depressions in the BVF distribution are present in several genomic positions. To evaluate if these possible positional effects on BVF values were statistically robust, BVF confidence intervals were compared with a simulated neutral model in which variants were uniformly distributed among different bins and different TSSs (see Materials and Methods).

A significant positional effect on BVF values was observed for all frequency classes in the regions around CGI-TSSs, with a significant deviation from the BVF neutral distribution when in close proximity to TSSs. In addition, for rare and mid1 classes only, an extended region of significant deviation is evident between approximately 2000 and 5000 bp downstream of the TSS was observed.

Looking in detail at the rare variant class, one can observe for CGI-TSSs a depression of BVF values in the near vicinity of the TSS, while in regions around nCGI-TSSs a significant positional effect was found for mid1 and common variants only, where a mild depression of variants is observed in the close proximity of the TSS. Figure 3.1 also shows that BVF signals of the four classes seem to deviate from the neutral model in different manners. This phenomenon is better shown in Figure 3.2, where all normalised signals are shown on the same graph.

The comparative analysis of normalised BVF signals showed that in some regions they almost overlap, while in other regions they are very diversely distributed. To investigate this point, for each bin, the BVF difference between the two extreme frequency classes (rare and common) was computed. This latter quantity (hereafter called ΔBVF) was then

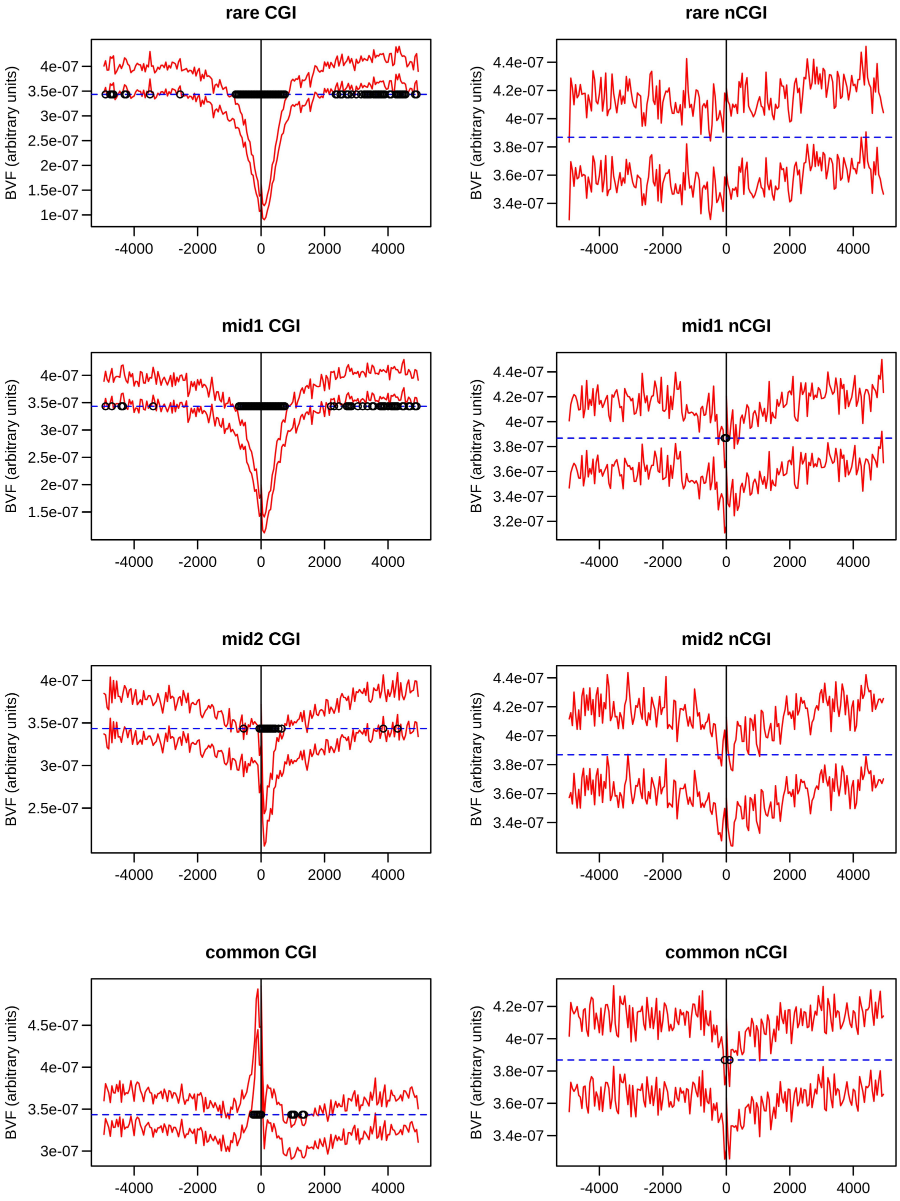

Fig. 3.1 **Positional effects of BVF values.** The two standard error confidence intervals for the observed normalised BVF values (red-dashed lines) are plotted along with its neutral expectation (blue-dashed line) for CGI-TSS frequency classes (left panel) and nCGI-TSS frequency classes (right panel). A dot is placed over the bins whose difference between the observed mean BVF value and the neutral expectation is statistically significant. On the x-axis is the position of the bin relative to the TSS.

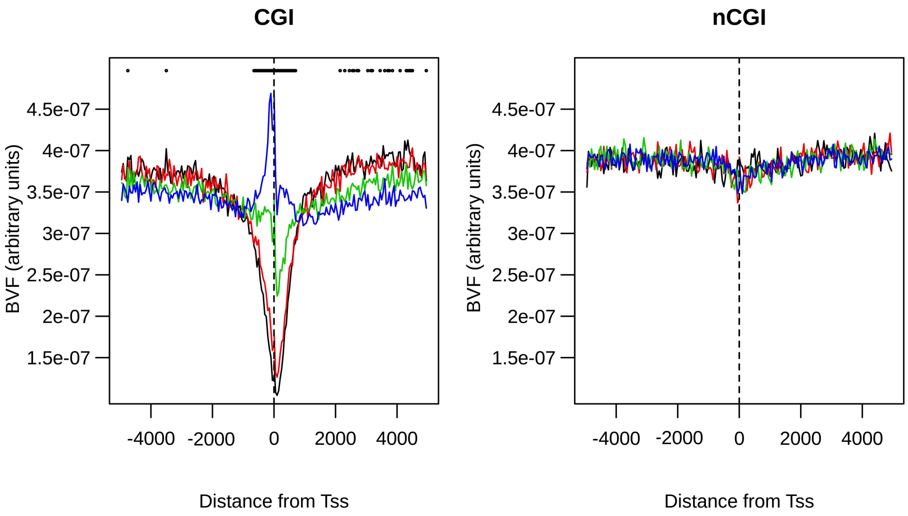

Fig. 3.2 **BVF distribution is different among classes.** Normalised BVF values for rare (black line), mid1 (red line), mid2 (green line) and common (blue line) variants are reported together on the same plot for CGI-TSSs (left panel) and nCGI-TSSs (right panel). A dot is placed over the bins where the difference of normalised BVF among the four classes is statistically significant. On the x-axis is the position of the bin relative to the TSS.

compared with the corresponding value derived from a neutral model (see Materials and Methods). In Figure 3.2, dots identify the regions in which the observed ΔBVF value was statistically different from that expected by the neutral model. This approach led, in CGI-TSSs, to the identification of two regions where the common variant BVF is statistically different from the rare variant BVF. The first region is in the near vicinity of TSSs, while the second one is downstream of TSSs (approximately between 1000 and 4500 bp). It should be noted that, in the first region, the local density of the common variant BVF was higher than that of the rare variants, while the opposite relationship was observed in the second region, where the common variant BVF value was lower than that of the rare variants. Figure 3.2 also shows that in both these regions mid1 and mid2 variants were distributed between rare and common classes, creating a frequency gradient.

In nCGI-TSSs, no statistically significant difference among the four distributions was found.

To quantify the potential bias introduced by the presence of bi-directional promoters and/or multiple closely located TSSs, the above analysis was repeated excluding regions that host two or more TSSs (see Chapter 2 for a discussion on this matter). Using this conservative filter, 58% of CGI-TSSs and 68% of nCGI-TSSs were retained. The results obtained using this reduced dataset do overlap with those shown for the entire dataset.

3.3 Relationship with nucleosomal occupancy

In general one can expect that variants belonging to different (frequency and/or CGI) classes will be differentially susceptible to the action of different evolutionary forces. It is likely that rare variants are more closely linked to the mutational process and that their frequency is influenced by the presence of mutational "hotspots".

On the other hand, stochastic and evolutionary events (such as drift and selection) can influence the localization of common variants. As a first step, the forces potentially affecting the distribution of rare variants were explored. It is known that the presence of DNA packaging structures, for example nucleosomes, can affect the emergence of novel mutations, thus influencing the presence of low frequency variants in a genomic region.

Therefore, possible relationships between nucleosome position and rare variant distribution were investigated. Nucleosome positioning scores of the Gm12878 cell line were downloaded from the UCSC "Stanf Nucleosome" track. By following an analog approach, as for BVF computation (see Materials and Methods), the "average nucleosome positioning score" for each bin (BNP) was calculated by averaging nucleosome scores for a fixed bin on all TSSs. As expected [36, 37], nucleosome positioning distribution around the TSS behaved differently for CGI-TSS and nCGI-TSS (Fisher p-value $< 1 \times 10^{-4}$), with CGI-TSSs being characterized by a marked depression in nucleosome density in the proximity of the TSS (Figure 3.3).

For each variant frequency class, Figure 3.4 reports the nucleosome positioning score against the variant density in the same bin and for the same TSS class. For all these plots, the correlation between the two signals was computed. In CGI-TSSs (Figure 3.4 upper panel), a very strong positive correlation was found for rarer variants. This latter decreases in higher frequency variant classes. In contrast, a weaker positive correlation was found

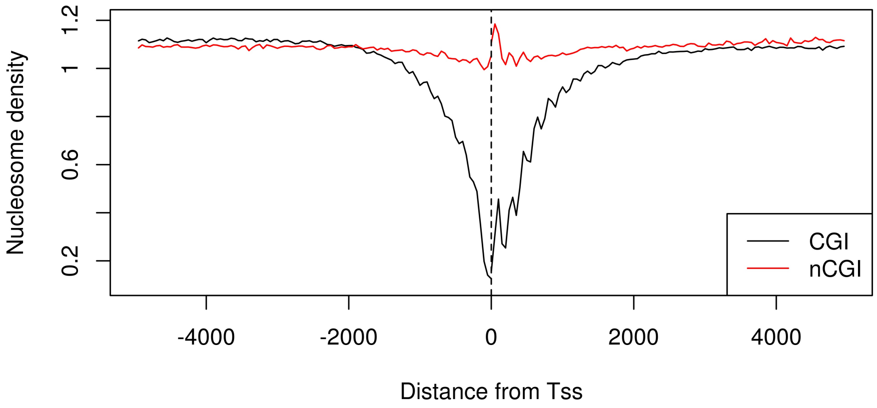

Fig. 3.3 **Nucleosome density distribution is different between CGI-TSSs and nCGI-TSSs.** The BNP values are plotted together for CGI-TSSs (black line) and nCGI-TSSs (red line). On the x-axis is the position of the bin relative to the TSS.

for nCGI-TSSs variants. The same analysis was conducted for the other available cell line, K562, with similar results (Figure 3.5).

3.4 Relationship with evolutionary and non-evolutionary forces

Next, possible forces affecting common variant distribution in the region were analysed. In particular, two forces that may affect allele frequencies within such a genomic area were considered: natural selection and GC-biased gene conversion (gBGC). As it is well known, natural selection acts when alleles differ in the resulting fitness of the individual. The effect of natural selection may drive the allele either towards a reduced frequency/loss in the population (purifying selection), or towards an increasing frequency/fixation (positive selection).

gBGC is a recombination-associated process that favours some alleles over others independently of the fitness conferred. gBGC is a process in which GC/AT (strong/weak)

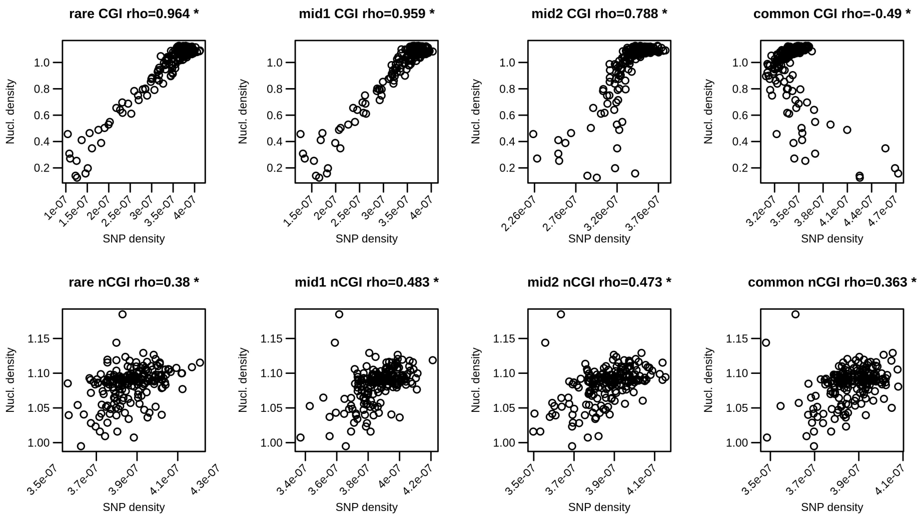

Fig. 3.4 **Nucleosome density correlation with SNP density values.** Pearson correlations between BNP values and BVF values are reported along with corresponding scatter plots for rare, mid1, mid2 and common variants (from left to right) and for the two TSS classes (CGI-TSSs on the top and nCGI-TSSs on the bottom). * indicates statistically significant correlations.

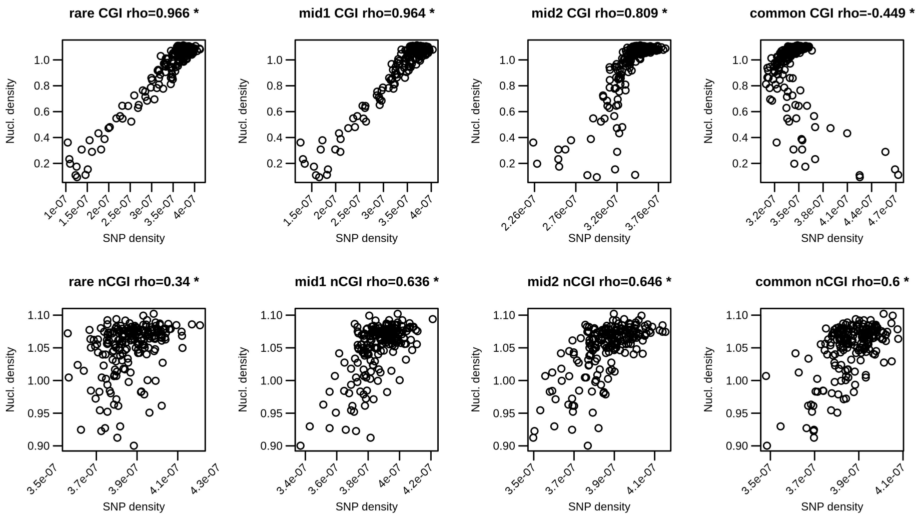

Fig. 3.5 **Correlation of nucleosome density with SNP density for the K562 cell line.** Pearson correlations between BNP and BVF values are reported along with corresponding scatter plots for rare, mid1, mid2 and common variants (from the left to right) and for the two TSS classes (CGI-TSSs top and nCGI-TSSs bottom). * indicates statistically significant correlations.

heterozygotes are preferentially resolved to the strong allele during gene conversion. This force could be particularly important in CpG rich regions near TSSs.

To identify possible signatures of natural selection, the conservation profiles of the analysed regions was studied using the GERP scores [38]. High values of this score indicate a lower level of substitutions among species (with respect to a neutral value derived by applying a maximum likelihood evolutionary rate estimation), hence indicating high evolutionary conservation.

To evaluate the possible presence of gBGC phenomena,“phastBias” gBGC track from UCSC was employed. By using this track, bases predicted to be influenced by GC-biased gene conversion (gBGC bases) were determined [39].

Next, the “bin average GERP score” (BGS) was evaluated by computing, for a fixed bin, the GERP values averaged over bin loci and over all TSSs (Figure 3.6).

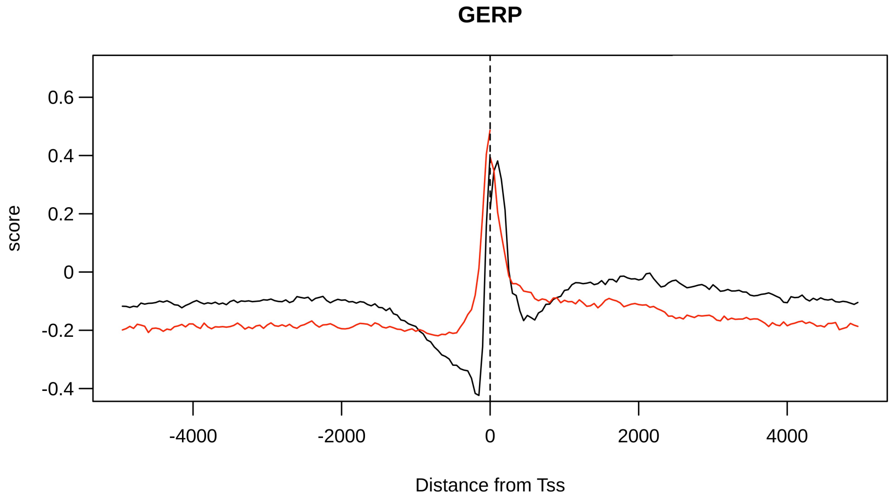

Fig. 3.6 **GERP distribution is different between CGI-TSSs and nCGI-TSSs.** The BGS values are plotted together for CGI-TSSs (black line) and nCGI-TSSs (red line). On the x-axis is the position of the bin relative to the TSS.

By using an analog process, the“bin average gBGC score” (BBS) was computed, for a fixed bin, as the average number of gBGC bases over all considered TSSs (Figure 3.7).

The BGS distribution resulted to be different between CGI-TSSs and nCGI-TSSs (Fisher p-value < 10^{-4}). For both TSS classes, a peak exists in the region ~100 bp upstream and

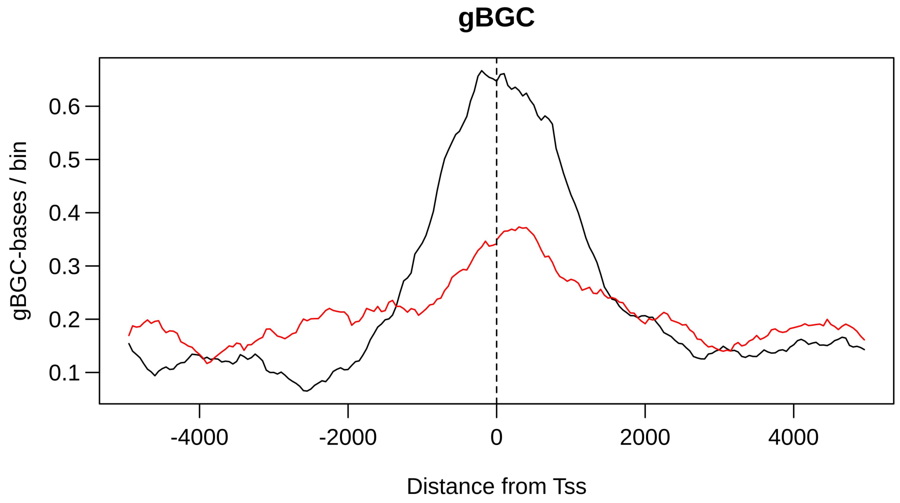

Fig. 3.7 **GERP distribution is different between CGI-TSSs and nCGI-TSSs.** The BGS values are plotted together for CGI-TSSs (black line) and nCGI-TSSs (red line). On the x-axis is the position of the bin relative to the TSS.

~200 bp downstream of the TSS. A region with negative BGS signal was found 200-700 bp upstream of the CGI-TSS only. Also, for the BBS signal different distributions for CGI-TSSs and nCGI-TSSs were found, with a peak in the region from approximately 2000 bp upstream to 2000 bp downstream of the CGI-TSSs.

Then, BGS and common variant BVF values were plotted on the same graph for nCGI-TSSs (Figure 3.8).

In this case, Figure 3.8 shows an apparent inverse correlation among the two variables. By testing these correlation, a negative correlation between common variant BVF and BGS values (Pearson correlation coefficient = -0.489, $p-value = 2.23 \times 10^{-13}$) was found (Figure 3.9).

In Figure 3.10 BBS, BGS and common variant BVF normalised values were plotted on the same graph.

In particular, a direct correlation appeared between common variant BVF and BBS in the vicinity of TSSs, whereas an inverse correlation seemed to be present between common variant BVF and BGS in the complementary distal regions.

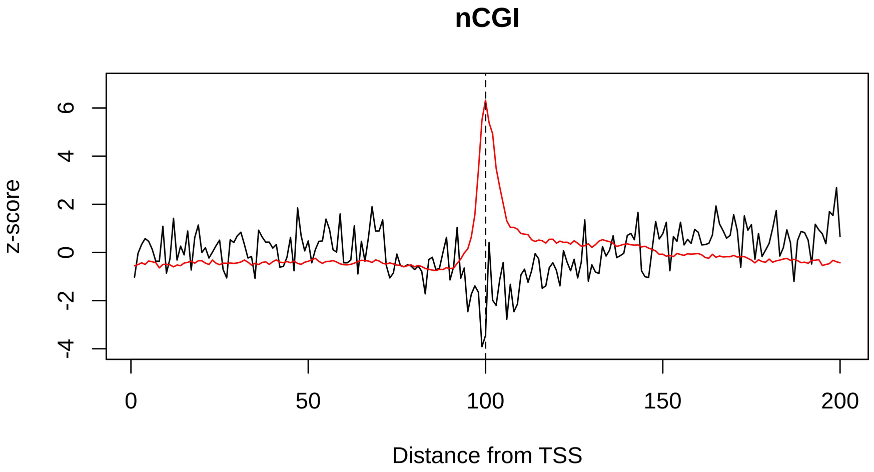

Fig. 3.8 **Overlapping normalised values of BGS and common variant BVF values for nCGI-TSSs.** The z-scores for BGS (red line) and common variant BVF (black line) are plotted for the same region. On the x-axis is the position of the bin relative to the TSS.

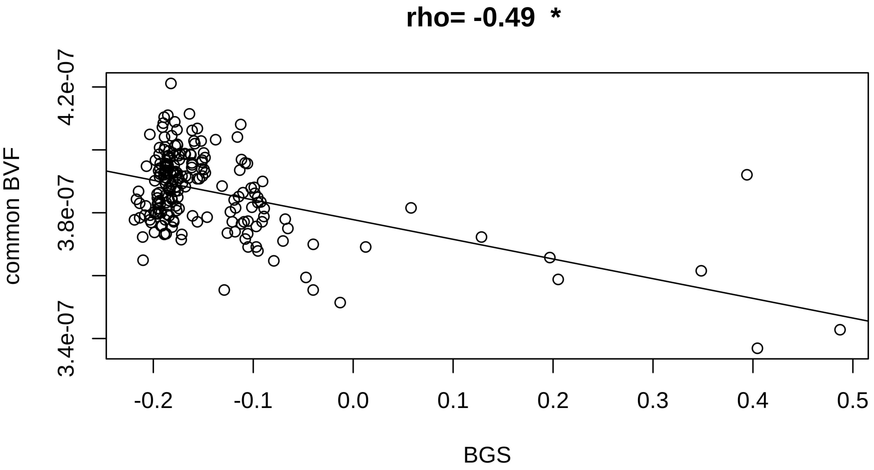

Fig. 3.9 **Common variant BVF correlation with GERP in nCGI-TSSs.** Pearson correlation between BGS and BVF values is reported along with the corresponding scatter plot for nCGI-TSSs. * indicates statistically significant correlations.

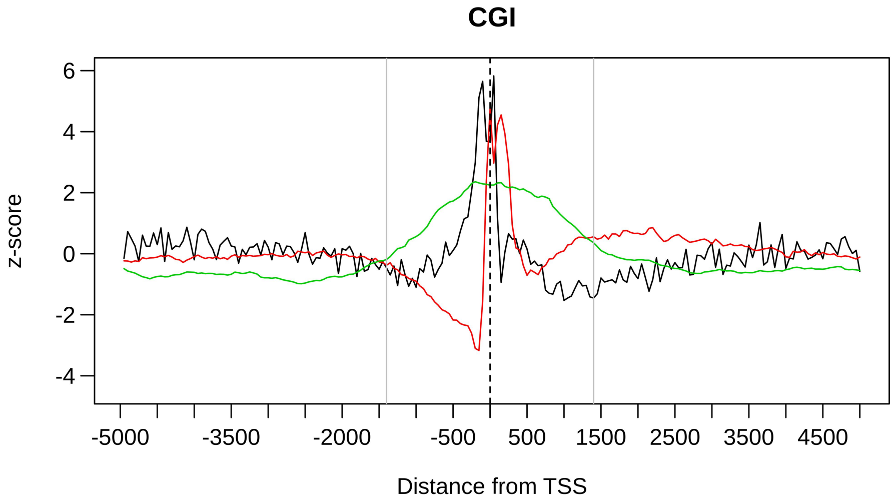

Fig. 3.10 **Overlapping normalised values of BGS, BBS and common variant BVF values for CGI-TSSs.** The z-scores for BBS (green line), BGS (red line) and common variant BVF (black line) are plotted for the same region. The grey lines delimit the region defined under strong gBGC influence. On the x-axis is the position of the bin relative to the TSS.

To better quantify such an involved patterns, a generic symmetric window around the TSS was considered and the correlations of BBS vs. common variant BVF and BGS vs. common variant BVF were computed as a function of the window size. Since the first correlation (BBS vs. common variant BVF), as an absolute value, was always significantly larger than the second one in proximal (inner) regions, it was decided to compute the correlation of BBS and common variant BVF in the inner region and between BGS and common variant BVF in the complementary one. By using a window-based approach (see Materials and Methods), the whole region was split into an inner one (approximately 1400 bp region flanking the TSS), where common variant BVF is mainly correlated with BBS, and an outer complementary one, where common variant BVF is mainly correlated with BGS. The analysis of the two regions showed a negative correlation ($\rho = -0.621, p-value = 1.25 \times 10^{-16}$) between BGS and common variant BVF in the external region (Figure 3.11, left panel) and, conversely, a positive correlation ($\rho = 0.581$, $p-value = 2.2 \times 10^{-6}$) in the inner region between common variant BVF and BBS (Figure 3.11, right panel).

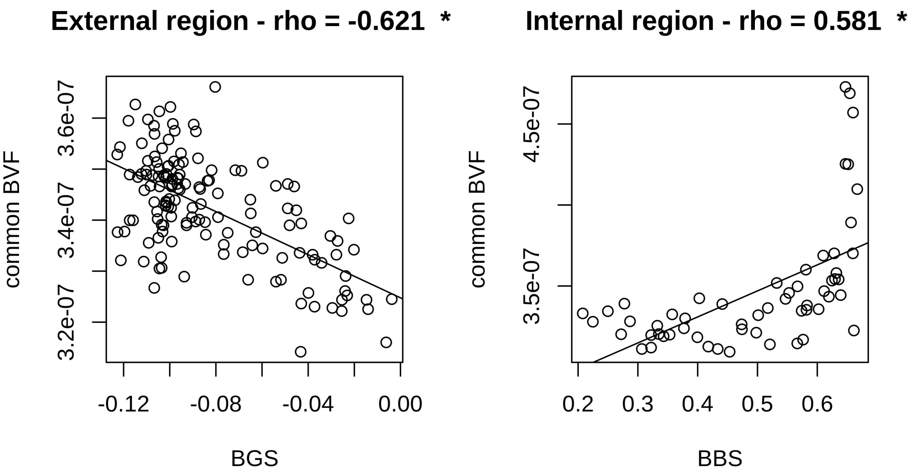

Fig. 3.11 **GERP and gBGC correlations for inner and outer regions.** Pearson correlations between BGS and common variant BVF values in the outer region (left panel) and between BBS and common variant BVF values in the inner region (right panel) are reported along with corresponding scatter plots for CGI-TSSs. * indicates statistically significant correlations.

3.5 Functional effects of variants around TSSs

As a last step, the potential pathogenicity of each class of variants was analysed.

This was accomplished by analysing the CADD (Combined Annotation Dependent Depletion) score [7]. High values of this signal characterise variants that are likely to have deleterious effects, namely their deleteriousness.

For both TSS classes and for each of the variant frequency classes, the “bin average CADD score” (BCS) was computed, for a fixed bin, as the average CADD value over bin variants and over TSSs (Figure 3.12).

As expected, a statistically significant difference (see Materials and Methods) was found among the four signals, with SNP deleteriousness values that generally decreased as the frequency of a variant increases. In all considered classes, deleteriousness increased moving toward the TSS from both sides.

Finally, for each frequency class, significantly higher values of deleteriousness were seen for CGI-TSSs compared with nCGI-TSSs in the region proximal (~1300 bp) to the TSS site (see Materials and Methods).

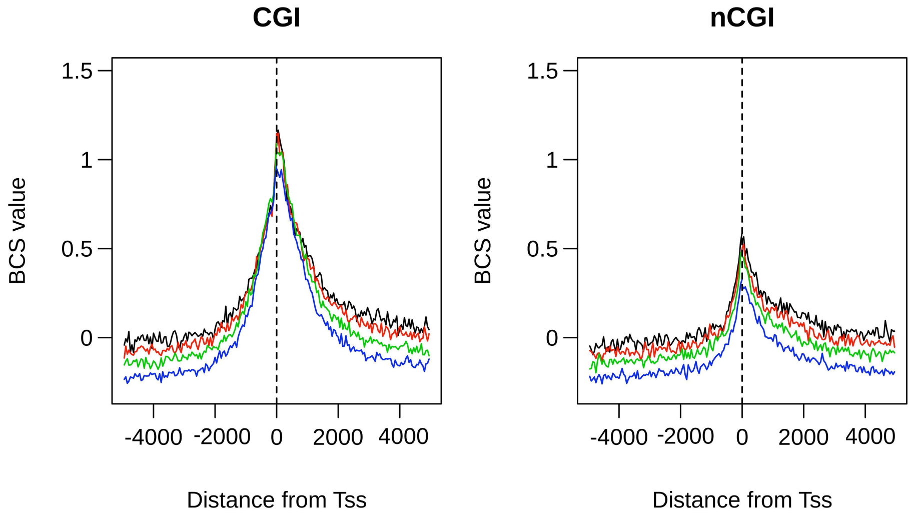

Fig. 3.12 **Deleteriousness scores among TSS classes.** BCS values are plotted on the same region for rare (black line), mid1 (red line), mid2 (green line) and common variants (blue line) for CGI-TSSs (left panel) and nCGI-TSSs (right panel). On the x-axis is the position of the bin relative to the TSS.

3.6 Discussion

In this chapter, the distribution of SNPs was analysed inside a 10 kb region flanking human TSSs. SNPs were divided into four classes according to their frequency (rare, two intermediate classes, and common) to explore in detail the genetic variability of the region, and to gain insight into the forces that generate and maintain this variability.

The distribution of variants in these regions depends on their frequency class, and on their localisation relative to TSSs.

Furthermore, splitting TSSs in CGI-TSSs (located inside a CGI) and nCGI-TSSs (not located inside a CGI), showed that the distribution of variants is generally different for the two subsets.

CGIs not only act as promoters, but are also associated with several other functionally-relevant genomic features, including recombination hotspots [40–44], the presence of transposable elements [45], domain organization and nuclear lamina interactions [46], origins of replication [47–51], and local mutational processes [33]. Furthermore, transcription-associated mutagenic processes and transcription-coupled repair are more active in

CGI-TSSs than in nCGI-TSSs. Indeed, CGI-promoters are generally associated with constitutively expressed genes in all cell types (housekeeping genes) [52], whose expression is necessary for the maintenance of cell physiology, while nCGI promoters are generally associated with highly tissue-specific genes and tend to have more restricted expression patterns.

Previous studies reported SNPs arising as a result of transcription-associated mutagenic processes [4, 16, 18] and a strong inhibitory effect of nucleosomes on mutation reparability by limiting the access of repair proteins [25–29].

Studying the relationship of the rare allele distribution with nucleosome occupancy score, a very strong positive correlation for CGI-TSS variants, and a weaker, but still significant, correlation for nCGI-TSS variants were found.

Interestingly, Higasa et. al. found a periodical distribution of SNPs around TSSs in regions associated with CGIs [34], and explained such result in terms of location of nucleosomes. The above presented data support the hypothesis that transcription-related mutational phenomena could be related to a reduced efficiency of the repair mechanisms in the regions occupied by nucleosomes [34, 53–55]. Indeed, damage within the nucleosome core is repaired at a rate of about 10% of that for naked DNA [26]. A strong reduction of the rare variants frequency in CGI-TSSs is also shown, but this is a known characteristic of CGIs in which the incidence of mutations is depressed. Indeed, if, on the one hand, transcriptional-related mutagenic processes are more active in CGIs, on the other hand, purifying selection might then counteract the loss of CpGs to preserve the existence of CGIs for regulatory processes [56]. As expected [36, 37], nucleosome distribution around TSSs is different for CGI-TSSs and nCGI-TSSs. In the case of CGI-TSSs, a marked depression was observed in nucleosome density in the 1 kb flanking regions. This is in agreement with the expectation that constitutive genes, such as those associated with CGI-promoters, exhibit a nucleosome-free region at their TSS to provide space for the assembly of the transcription machinery [57].

In contrast, in nCGI-TSSs a peak was observed just downstream of the TSS, corresponding to a high intrinsic nucleosome occupancy. This finding can be explained by the fact that CGI-promoters are associated with housekeeping genes [52], while nCGI promoters generally have more restricted expression patterns, depending, for example, on the developmental stage or cell type. In fact, in the absence of physiological requirements, it could be advantageous to keep nCGI regulatory sites masked with nucleosomes to minimize risks of inappropriate utilization and aberrant transcription.

A common characteristic of both CGI-TSSs and nCGI-TSSs is a neutral regime observed for regions more than ~1000 bp upstream of the TSS. This is expected because of the significant distance from the TSS and because it generally corresponds to a noncoding region.

In contrast, the presence of a gradient in the relative frequencies, with increasing values from common to rare variants, could reliably testify for purifying selection. In this specific case, it would preserve coding regions from the accumulation of deleterious mutations and prevent deleterious mutations from reaching common frequencies. This last observation is also supported by the correlation of the common variant BVF signal with GERP scores in regions where rare variants (younger) are more frequent than the common (oldest) ones.

Such features can be observed in the coding regions of CGI-TSSs. In particular, for these latter, this kind of selection seems to be at work in a restricted region in the near downstream vicinity of TSS, as indicated by a strong depletion of common variants with respect to the specular region upstream. Also for nCGI-TSSs a mild depression in common variants was found in the near vicinity of the TSSs where the conservation score resulted to be higher.

Interestingly, the gradient in relative frequencies extends far into the downstream region for CGI-TSSs only. This difference with the nCGI-TSSs could be related to the strong enrichment of housekeeping genes among CGI-TSSs. As described in [35], housekeeping genes evolve more slowly than tissue-specific genes in terms of both coding and core promoter sequences. Selective constraint differences could arise from differences in gene function and expression. As housekeeping genes play a key role in the maintenance of most cells, strong purifying selection acts to preserve their normal function, whereas for tissue-specific genes, which are expressed in few tissues, the impact of a deleterious mutation is less than that of housekeeping genes.

Interestingly, there is the intriguing presence of a sharp peak of common variants in the near vicinity of CGI-TSSs. This peak was completely absent in nCGI-TSSs and implies that forces influencing this pattern could be, at least in part, related to the CpG content of the region. A CpG content-related force, able to influence the distribution of allele frequencies within a population, is the GC-biased gene conversion (gBGC). According to such conversion phenomenon, GC/AT heterozygotes are preferentially resolved to GC/GC homozygotes during gene conversion. This may be seen as a kind of back mutation that tends to restore the intermediate gene frequencies causing a nontrivial stable distribution of gene frequencies at equilibrium [58]. In other words, the effect of back mutation interferes with the possible loss or fixation of variants, with the result of an enrichment of common frequency variants with respect to the case in which only genetic drift is at work (neutral scenario).

As observed by Polak and Arndt [40], CGIs contain a mutational signature of GC-biased gene conversion, which determines an enrichment of GC nucleotides in CGIs [59]. Also, according to Galtier and Duret [60] the BGC process is not a mutagenic process that introduces de novo mutations into the genome, but instead it increases the fixation probability of GC alleles over AT alleles [60, 61]. The results of this analysis seem to show the simultaneous action of drift, selective pressure and gBGC for CGI-TSSs. In particular, the depletion of common variants with respect to the specular upstream region could be explained by the presence of a purifying selection, whereas the larger relative frequency of common variants with respect to the rare variants testifies the action of gBGC. In this sense, gBGC could compete against, and slow the effects of purifying selection. In some cases, gBGC overcomes purifying selection and leads to the fixation of deleterious AT/GC mutations that would be eliminated in the absence of gBGC [60, 62], thus counteracting natural selection [63]. The limited extension of the region where gBGC phenomena dominate over selection is marked by the inversion point of the gradient. Remarkably, such a point almost corresponds to the average downstream end of the CGI.

Finally, when analysing the deleteriousness of each class of variants, it was found that it decreases as the frequency of variants increases, suggesting that rare variants are more deleterious than common ones. Furthermore, the deleteriousness is higher in CGI-TSSs than in nCGI-TSSs, in accordance with the structural and functional importance of CGIs, as explained above.

3.7 Materials and Methods

3.7.1 TSS selection

Genomic coordinates of human TSSs were downloaded from the UCSC track “switchDbTss”. This track reports data collected in the “switchDB”, an open-access online database of human TSSs from “SwitchGear Genomics” (http://www.switchgeargenomics.com), where the location of 131,780 TSSs throughout the human genome is determined by integrating experimental data. The database provides, for each site in the genome, a score that encodes the confidence to have a TSS placed in that position. To exclude unreliable data, only TSSs whose confidence score was greater than or equal to 20 were retained; $N_{TSS} = 27,487$ (~21% of the total).

3.7.2 Variant selection

The human genomic variants used in this study were downloaded from the “snp138” UCSC table. This track consists of all human genomic variants reported in dbSNP build 138: a massive collection of molecular variants (including SNPs, insertions, deletions) coming from heterogeneous studies. From dbSNP were selected the only variants that:

1. include the term “1000GENOMES” in the “submitters” field;
2. have the value of the field “class” equal to “single”;
3. have the value of the field “alleleFreqCount” equal to 2;
4. have both the two comma separated values of the field “alleleNs” greater than 0;

5. have the sum of the two comma separated values of the field “alleleNs” greater than 1000.

By applying this filter, 2,636,282 variants were obtained. Only the variants from low coverage set of 1000 Genomes phase 1 release were retained, by excluding the ones not having the value “LOWCOV “ for the field “SNPSOURCE” in the 1000 Genomes vcf file, available from ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20110521/ALL.wgs.phase1_release_v3.20101123.snps_indels_sv.sites.vcf.gz.

This last filter led to the final set of 2,550,709 variants used in this study. Since there is no explicit representation of the MAF in this dataset, it was computed by taking the minimum of the two-allele frequencies for each variant. This was computed using the two comma separated values reported in the dbSNP field “alleleNs”.

3.7.3 CpG island selection

The genomic coordinates of CGIs were obtained from the UCSC “CpgIslandExt” track. In this track CGIs were predicted by searching the sequence one base at a time, scoring each dinucleotide (+17 for CG and -1 for others) and identifying maximally scoring segments. In this dataset, to define a CGI the following criteria were used:

1. to have a GC content of 50% or greater;

2. to have a length greater than 200 bp;

3. to show a ratio greater than 0.6 of observed number of CG dinucleotides to the expected number, calculated on the basis of the number of Gs and Cs in the segment under analysis.

UCSC CGI files also contain data related to sequence for alternative haplotypes (present mainly in chromosome 6, for the inclusion of alternative versions of the MHC region). Of course, in this analysis the file was filtered to exclude these duplicated data.

3.7.4 Nucleosome positioning scores

Nucleosome localization data were downloaded for two cell lines, namely Gm12878 and K562, from the ENCODE UCSC tracks “wgEncodeSydhNsomeGm12878Sig” and “wgEncodeSydhNsomeK562Sig”, respectively [64]. These tracks contain density signal maps of nucleosome positions produced by the MNase-seq technique.

3.7.5 GERP evolutionary scores

GERP signal was used to evaluate the impact of conservative evolutionary forces. This approach identifies, by multiple alignments and maximum likelihood evolutionary rate estimation, genomic regions with a substitution rate lower than that expected under neutral hypothesis, suggesting that they are under functional constraint. Therefore GERP is able to detect signatures of past purifying selection. Conservation data for the regions of interest were downloaded from the “allHg19RS_BW” UCSC track. This track contains an evolutionary score obtained by GERP. Constraint intensity at each individual alignment position is quantified in terms of a “rejected substitutions” score, defined as the number of substitutions expected under neutrality minus the number of substitutions “observed” at the position [38]. Expected and observed substitutions were computed by alignment of hg19 to 35 other mammalian species.

3.7.6 gBGC scores

Human genomic regions predicted to be influenced by GC-biased gene conversion were downloaded from the UCSC “phastBias” track. This signal was derived by phastBias prediction model applied to the human-chimp lineages of the phylogenetic tree as described in [39]. This method, predicts gBGC tracts using a four-state hidden Markov model (HMM) (conserved, neutral, conserved with gBGC, neutral with gBGC). In particular, the phastBias track contains only regions for which the assigned posterior probability to be affected by gBGC is greater than 0.5.

3.7.7 CADD scores

CADD scores were generated using the method of Kircher et al.[7]. These scores are computed through a Support Vector Machine approach that is based on 63 different annotations spanning a wide range of data types. These include conservation metrics such as GERP, phastCons, and phyloP; functional genomic data such as DNase hypersensitivity and transcription factor binding; transcript information such as distance to exon-intron boundaries or expression levels in commonly studied cell lines; and protein-level scores such as Grantham, SIFT, and PolyPhen. In particular these scores can be interpreted as the extent to which the annotation profile for a given variant suggests that the variant has deleterious effects. Scores were pre-computed for all 8.6 billion possible single nucleotide variants of the reference genome, as well as variants from the 1000 Genomes and ESP variant releases. CADD data for all 1000 Genomes variants were downloaded from http://cadd.gs.washington.edu. The file contains, for each variant, the genomic position and two scores, namely the “raw” and “scaled” C-Scores. For the purposes of this work the “raw” score was employed .

3.7.8 Statistical Analysis

Definition of upstream and downstream regions

Upstream and downstream regions were defined on each strand for each TSS by considering the direction of the transcription process, following the procedure described in Chapter 2.

For each TSS, the upstream region was defined as the 5000 bp on the 5’ side if the TSS is located on the coding strand and as the 5000 bp region on the 3’ side if the TSS lies on the noncoding strand. Conversely, the downstream region was defined as the 5000 bp on the 3’ side of the TSS if this is located on the plus strand and as the 5000 bp on the 5’ side if the TSS lies on the minus strand. Let N_{TSS} be the number of analysed TSSs and let $TSS(i)$ be the position of the i-th TSS and $Bin(i,j)$ the bin in position j with respect to $TSS(i)$ as defined in Chapter 2.

Bin Variant Frequency calculation

Let us denote with $vars(i,j)$ the number of variants falling into the $Bin(i,j)$ region. By following the procedure described in the previous chapter, one can define the normalised mean variant frequency for the j-th Bin as:

$$BVF(j) = \frac{\sum_{i=1}^{N_{TSS}} vars(i,j)}{N_{TSS} * N_{HITS}} \tag{3.1}$$

with

$$N_{HITS} = \sum_{i=1}^{N_{TSS}} \sum_{j \in \{-100,...,-1,+1,...,+100\}} vars(i,j) \tag{3.2}$$

Positional effects on BFV values

For each considered subset of SNPs, a neutral model in which variants are uniformly distributed among bins and among all TSS regions was considered. In particular the mean of this null distribution was calculated as:

$$Null\_BVFmean = \frac{1}{N_{TSS} * 200} \tag{3.3}$$

The difference between the observed mean values $BVF(j)$ and the $Null\_BVFmean$ was then tested, for each bin j, by means of a two-sided t-test. All 200 p-values were finally corrected using the Bonferroni method.

To investigate the presence of different positional effects on BVF values among classes it was evaluated, for each bin, the difference between rare variant-normalised BVF (BVF_{rare}) and common variant-normalised BVF (BVF_{common}) values:

$$\Delta BVF(j) = BVF_{rare}(j) - BVF_{common}(j) \tag{3.4}$$

The statistical significance of the ΔBVF values was then evaluated by testing their difference from the neutral value 0. The above differences were tested by means of a two-sided t-test and the corresponding p-values were finally corrected using the Bonferroni procedure.

Bin Nucleosome Positioning calculation

Let $NUC(i,j)$ be the set of nucleosome density values from “wgEncodeSydhNsomeGm12878Sig” or “wgEncodeSydhNsomeK562Sig” (for Gm12878 and K562 cell lines, respectively), associated with sites placed inside $Bin(i,j)$, and let $|NUC(i,j)|$ be its cardinality.

If we denote with $score(i,j)$ the sum of all nucleosome scores from $NUC(i,j)$, we can compute the “average Nucleosome Positioning score” for the j-th Bin as:

$$BNP(j) = \frac{\sum_{i=1}^{N_{TSS}} score(i,j)}{\sum_{i=1}^{N_{TSS}} |NUC(i,j)|} \tag{3.5}$$

To evaluate the statistical difference of the whole BNP signal between CGI-TSSs and nCGI-TSSs, the BNP values of each pair of corresponding bins were compared by means of the two-sided t-test, thus obtaining 200 p-values. It was then tested if the observed local differences are representative of a global difference of the whole signals by means of Fisher’s method as described in Chapter 2. Correlations between BNP vs. BVF values were obtained by means of Pearson’s product moment correlation coefficient and tested as described in Chapter 2.

Evolutionary analysis

Let $GER(i,j)$ be the set of GERP score values from the “allHg19RS_BW” sites whose genomic position fall inside the $Bin(i,j)$ region, and let $|GER(i,j)|$ be its cardinality. By denoting with $score(i,j)$ the sum of all scores belonging to $GER(i,j)$, we can compute the “Bin average GERP Score” for the j-th Bin as:

$$BGS(j) = \frac{\sum_{i=1}^{N_{TSS}} score(i,j)}{\sum_{i=1}^{N_{TSS}} |GER(i,j)|} \tag{3.6}$$

Analogously, let us denote with $bgc(i,j)$ the number of genomic positions from the “phastBiasTracts3” UCSC track falling into the $Bin(i,j)$ region. The normalised Bin average gBGC Score for the j-th Bin was computed as:

$$BBS(j) = \frac{\sum_{i=1}^{N_{TSS}} bgc(i,j)}{N_{TSS}} \tag{3.7}$$

To evaluate the statistical difference of the whole BBS (BGS) signal between CGI-TSSs and nCGI-TSSs, the BBS (BGS) values of each pair of corresponding bins were compared by means of the two-sided t-test, thus obtaining 200 p-values.

The hypothesis that the observed local differences were representative of a global difference of the whole signals was tested using the same procedure described for BNP values.

For CGI-TSSs, the whole region was split into two subregions where the effects on common variant BVF of gBGC or BGS are respectively dominating. To this aim the following method was employed.

Starting from the TSS site, at each iteration i (starting with i = 2 till i = 98) it was considered the region defined by bins with an index of [-i,i] and the complementary one defined by [-100, -i[U]+i, +100]. At each iteration the correlation between BBS and common variant BVF in the region [-i,i] and between BGS and common variant BVF in the region [-100, -i[U]+i, +100] were evaluated. Let j be the index i for which the absolute value of the product of the two correlations is maximised, then [-j,+j] is chosen as the region under strong BGC influence and the complementary region (with respect to the whole analysed region) as the one under strong BGS influence (3.13).

We can safely compute the product of correlations since BGS and BBS signals are supposed to be independent. Correlations between BGS vs. common variant BVF were obtained by means of Pearson’s product moment correlation coefficient and tested as described in Chapter 2.

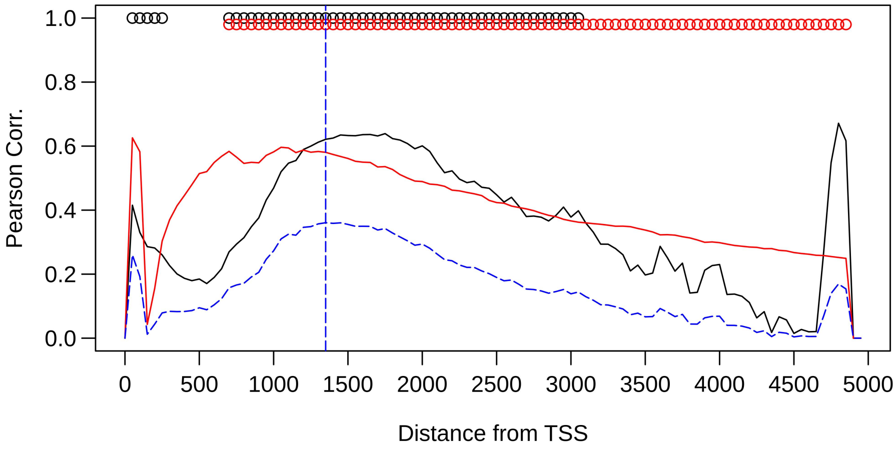

Fig. 3.13 **Regions under strong BBS and/or BGS influence.** Figure shows, for each pair of inner-outer regions defined by the distance reported on the x-axis, the absolute value of the Pearson correlation for BBS and common variant BVF in the inner region (red line), the absolute value of correlation for BGS and common variant BVF in the outer region (black line) and the product of the two correlations (blue-dashed line). Red dots are placed where the correlation between BBS and common variant BVF is statistically significant and black dots are placed where the correlation between BGS and common variant BVF is statistically significant. The vertical dashed line represent the distance for which the value of the product of the two correlations is maximized.

Deleteriousness analysis

Let $DEL(i,j)$ be the set of deleteriousness values (from the CADD-1000 Genomes dataset) associated with sites placed inside $Bin(i,j)$, and $|DEL(i,j)|$ its cardinality. If we denote with $score(i,j)$ the sum of all CADD scores belonging to $DEL(i,j)$, we can compute the "bin average CADD score" for the j-th Bin as:

$$BCS(j) = \frac{\sum_{i=1}^{N_{TSS}} score(i,j)}{\sum_{i=1}^{N_{TSS}} |DEL(i,j)|} \tag{3.8}$$

For each bin j, the difference in the mean BCS value between the four frequency classes was tested by means of a one-way analysis of variance. It was then tested if the observed local differences were representative of a global difference of the whole signals by means of Fisher's method as described in Chapter 2. For each bin j, the difference in the mean BCS value between CGI-TSSs and nCGI-TSSs of the four frequency classes was tested by means of a two-sided t-test and Bonferroni correction.

Statistical significance assessment

This study was conducted considering a p-value of 0.001 as statistically significant. All statistical analyses were performed using R ver. 2.10.1 software.

Chapter 4

Genetic variation around human Polyadenylation Sites

4.1 Description of the study

Despite the crucial relevance of transcription termination in the formation of correct RNA molecules, termination is the least understood step of the transcription process [65]. Polyadenylation sites (PSs) are canonically defined as the sites at which a poly(A) tail is added to a messenger RNA. Therefore, PSs basically mark the end of a gene. Correct transcription termination plays an important role in the stability, localisation and translation of the nascent mRNA molecule [66]. Moreover, correct transcription termination also ensures that a pool of RNA polymerases is available for re-initiation of new transcription, and prevents the formation of antisense RNAs that can interfere with pre-RNA production [67]. The genomic region near PSs provides signals that activate molecular processes to free RNA polymerase and release the mRNA from the transcriptional machinery. The occurrence of genetic variants in these sites may have relevant functional effects and, consequently, be subject to selective phenomena. Moreover, transcription-related events that characterise these loci, such as r-loop formation [68] and DNA-RNA polymerase collisions [69], could damage DNA, increasing the mutational rate in the surrounding region. In the literature, the distribution of SNPs in promoter [15, 54, 70], gene body [54], and termination sites [54] have been studied. In the previous chapter, it is shown how the distribution of SNPs around transcription start sites is affected by the distance from the TSS and characterised by different patterns when considering different SNP frequency classes. Moreover, in Chapter 3 it is shown, by means of a correlation analysis, that the insurgence of novel mutations around TSSs could be related to nucleosome occupation. Furthermore it is suggested that evolutionary-conservative forces could be at work in this region to "control" variability.

The aim of the study presented in this chapter is to extend this analysis to genomic regions near PSs.

4.2 Definition of PSs

As a first step, genomic coordinates of human mRNA PSs (poly(A) sites) were downloaded from the University of California, Santa Cruz (UCSC) database, and a total of 42,382 genomic loci were obtained.

4.3 Distribution of variants around PSs

To investigate the distribution of genetic variants around human PSs, the 10-kb region surrounding each polyadenylation site was considered.

Human genetic variant data were downloaded from the dbSNP version 141 database. To achieve robust and comparable allele frequencies, only bi-allelic single nucleotide variants with available frequency values and detected by whole genome sequencing approach in the 1000 Genomes Phase 1 Variant Catalog, were selected.

These criteria led to 2,720,492 SNPs considering only the ones located in the 10-kb region surrounding the above-selected PSs. Such subset was then split into four classes according to MAF values.

As in the previous chapter, these classes were named as: rare, mid1, mid2 and common. Rare variants (~21% of all selected variants) were defined by a MAF less than or equal to

4.59×10^{-4}. This threshold corresponds to a heterozygous variant being present in only one among all 1092 individuals from the phase 1 release. Common variants (~32% of the selected variants) were defined according to the canonical criterion of a MAF value greater than 0.01. The remaining variants, with intermediate frequency between rare and common, were partitioned in two groups of approximately equal size and are named "mid1" (frequency range: $4.59 \times 10^{-4} - 0.0023$, ~23% of all selected variants) and "mid2" (frequency range: $0.0023 - 0.01$, ~24% of all selected variants).

As a first step, for each PS, the surrounding 10-kb region was divided into 200 bins of 50 bp each, and the normalised mean bin variant frequency (BVF) was computed for each bin (see Materials and Methods).

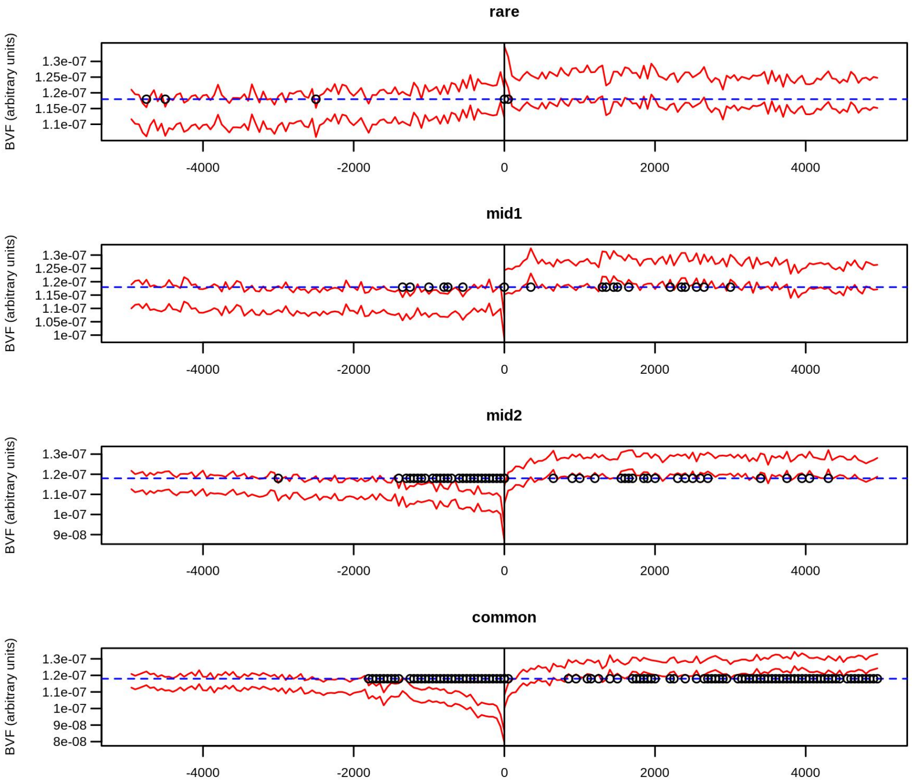

Fig. 4.1 **Positional effects of bin variant frequency (BVF) values in the four single nucleotide polymorphism (SNP) classes.** The two standard error confidence intervals for the observed normalized BVF values (red dashed lines) are plotted along with the neutral expectation (blue dashed line). A dot is placed over the bins whose difference between the observed mean BVF value and the neutral expectation is statistically significant. The position of the bin relative to the polyadenylation site (PS) is shown on the x-axis.

Figure 4.1 shows confidence intervals of BVF values for the four frequency classes (from rare to common variants going from the top to bottom panel). Several peaks and/or depressions were observed in the BVF distribution at several genomic positions. To evaluate if these possible positional effects on BVF values were statistically robust, BVF confidence intervals were compared with a simulated neutral model in which variants were uniformly distributed among different bins and different PSs (see Materials and Methods).

Significant positional effects on BVF values were observed for all frequency classes. In particular, an extended region (from about -1500 to +100 bp) of significant deviation from the BVF neutral distribution for the mid2 and common variants. In this region, SNPs belonging to these frequency classes showed a marked decrease in frequency values moving from the upstream side of the analysed region toward the polyadenylation site. Also, a relatively small increase in the rare variants density was observed for the first 100 bp downstream of the PS.

Figure 4.1 also shows that BVF signals of the four classes seem to deviate from the neutral model in different manners. This phenomenon is better shown in Figure 4.2, where all normalised signals are shown on the same graph.

The comparative analysis of normalised BVF signals showed that the signals overlap in some regions (for example, -5000 to -2000), whereas they are very diversely distributed in other regions. To investigate this point, the BVF difference between the two extreme frequency classes (rare and common) was calculated. This value (hereafter called ΔBVF) was then compared with the corresponding value derived from a neutral model (see Materials and Methods).

In Figure 4.2, dots identify the regions in which the observed ΔBVF value was statistically different from the neutral expectation. This approach led to the identification of an extended region where the common variant BVF is statistically different from the rare variant BVF. In this region, the value of the rare variant BVF was higher than that of the common one. Figure 4.2 also shows that in both these regions mid1 and mid2 variants

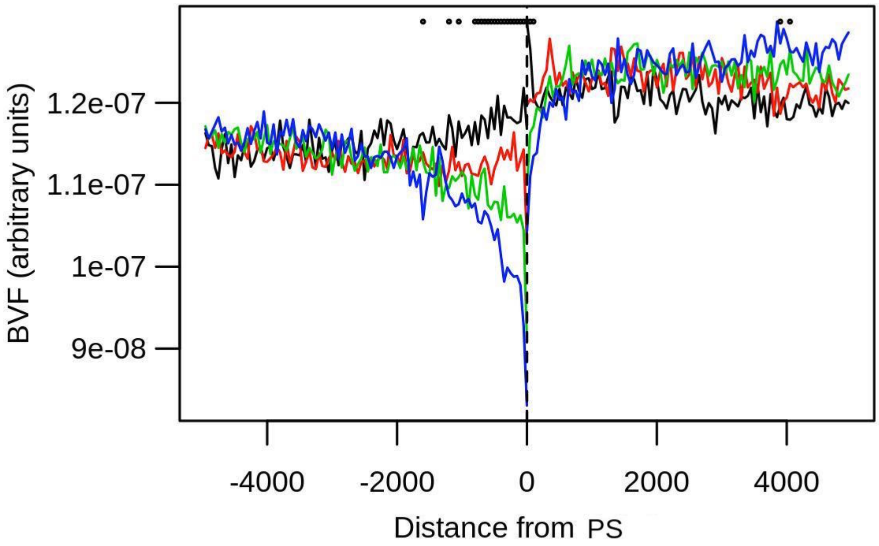

Fig. 4.2 **BVF distribution is different among frequency classes.** Normalized BVF values for rare (black line), mid1 (red line), mid2 (green line), and common (blue line) variants are reported together on the same plot. A dot is placed over the bins where the difference of the normalised BVF values among the four classes is statistically significant. The position of the bin relative to the PS is shown on the x-axis.

were distributed on average between rare and common classes in both of these regions, creating a frequency gradient.

4.4 Relationship with nucleosomal occupancy

In general, one can expect that variants belonging to diverse frequency classes will be differently affected by the action of evolutionary forces. In this respect, it is likely that rare variants are more closely linked to the mutational processes and that their frequency is influenced by the presence of mutational hotspots.

However, stochastic and evolutionary events (such as drift and selection) certainly influence the localisation of common variants.

As in Chapter 3, the analysis was first focused on the forces potentially affecting the distribution of rare variants. For this reason, the possible relationship between nucleosome positioning and rare variant distribution was investigated.

To this aim nucleosome positioning scores of the K562 and Gm12878 cell lines from the UCSC "Stanf Nucleosome" track were employed. By following an analog approach, as for the BVF computation (see Materials and Methods), the "average bin nucleosome positioning score" (BNP) was computed, for each bin, by averaging nucleosome scores on all PSs.

Figure 4.3 (left panel) shows the distribution of BNP values around PSs for the K562 cell line (black line). In the upstream region, the signal shows a monotonic behaviour in the proximity of PSs, whereas a marked depression of nucleosome occupation was found at ~100 bp from the PS. Finally, the nucleosome occupation was lower on average in upstream regions than in downstream regions.

In the same figure, the rare variant density (red line) is shown.

Figure 4.3 (left panel) also shows an apparent direct correlation between BNP and rare BVF values. This is better shown in the right panel, where the regression analysis between these two signals is reported.

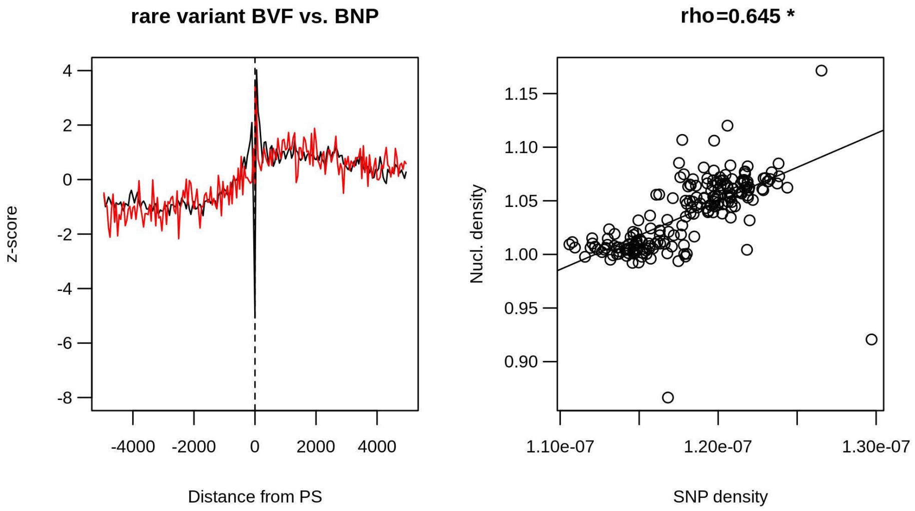

Fig. 4.3 **K562 nucleosome density correlation with SNP density values.** Left panel: overlapping normalised values of bin nucleosome positioning (BNP) and rare BVF values. Right panel: Pearson correlations between BNP and rare BVF values ; * indicates statistically significant correlations.

In particular, a statistically significant positive correlation (Pearson correlation coefficient = 0.65, $p-value < 2.2 \times 10^{-16}$) was found. While, for the other available cell line, Gm12878, a lower correlation value was found (Pearson correlation coefficient = 0.317, $p-value = 4.85 \times 10^{-6}$) (see Figure 4.4).

4.5 Relationship with evolutionary forces

Next, the analysis focused on possible forces affecting common variant distribution in the region.

To identify possible signatures of natural selection, the conservation profiles of the analysed regions were evaluated by means of GERP scores [71].

The bin average GERP score (BGS) was computed for each bin as the average GERP value over the bin loci and over all PSs (red line in Figure 4.5). In particular, enhanced conservation in the region ~2000 bp immediately upstream of the PSs was observed.

Since the distribution of common variants is likely to be more closely connected with the presence of selection forces than the rarer ones, mainly affected by mutagenic

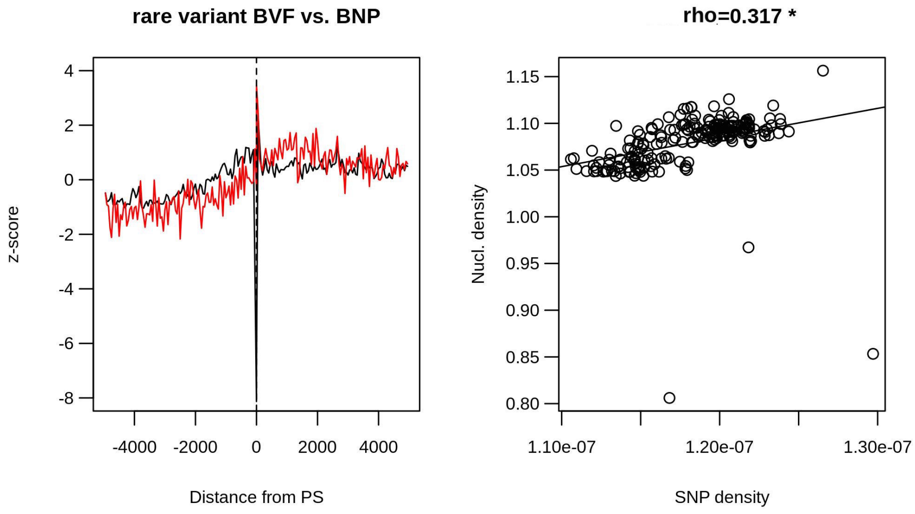

Fig. 4.4 **GM12878 nucleosome density correlation with SNP density values.** Left panel: overlapping normalised values of bin nucleosome positioning (BNP) and rare BVF values. Right panel: Pearson correlations between BNP and rare BVF values; * indicates statistically significant correlations.

processes, a correlation study was performed between BGS and common variant BVF. Figure 4.5 (left-panel) compares the common variant BVF and BGS scores (standardised values).

As in Figure 4.3 (right panel), a correlation analysis between the BGC and common variant BVF was performed. In Figure 4.5 (right panel), the black dots indicate a statistically significant anti-correlation (Pearson correlation coefficient=-0.849, $p - value < 2.2 \times 10^{-16}$).

4.6 Functional effects of variants around PSs

Finally, the potential pathogenicity of each class of variants was analysed by using the CADD score [7]. For each class of variants the "bin average CADD score" (BCS) was obtained by computing, for each bin, the CADD values averaged over bin variants and over PSs (Figure 4.6).

As expected, a statistically significant difference (see Materials and Methods) was found among the four signals, with SNP deleteriousness values that generally decrease as

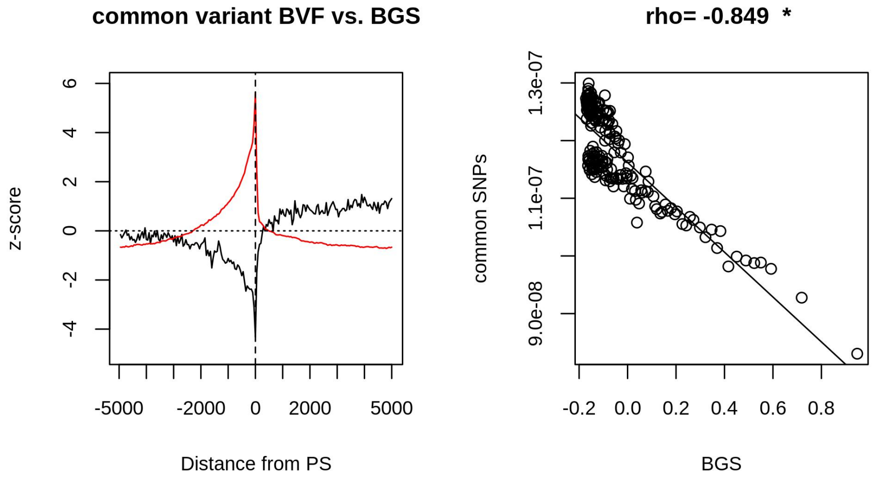

Fig. 4.5 **Genomic Evolutionary Rate Profiling (GERP) correlation with common SNP density values.** Left panel: overlapping normalised values of bin average GERP score (BGS) and common variant BVF values. Right panel: Pearson correlation between BGS and common BVF values with the corresponding scatter plot; * indicates statistically significant correlations.

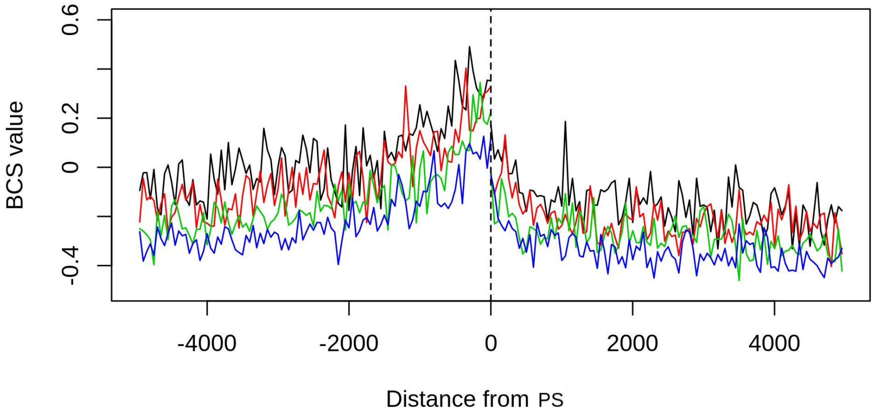

Fig. 4.6 **Deleteriousness scores among the four frequency classes.** Bin average Combined Annotation Dependent Depletion (CADD) score (BCS) values are plotted on the same region for rare (black line), mid1 (red line), mid2 (green line), and common variants (blue line). The position of the bin relative to the PS is shown on the x-axis.

the frequency of a variant increases. Moreover, for all considered classes, deleteriousness increases while moving toward the PS from both sides.

4.7 Discussion

In this chapter, the distribution of SNPs inside a 10 kb region flanking human PSs was analysed. To this aim, human polyadenylation sites were considered as transcription termination markers, and SNPs from the selected regions were split into four classes according to their population frequency (rare, two intermediate classes, and common). This allows to explore the genetic variability of the region by simultaneously taking into account the forces that generate and maintain this variability.

In the present analysis, it is shown that the distribution of variants depends on their frequency class and on their localisation relative to PSs.

The distribution of SNPs near human polyadenylation sites has been previously reported in [70]. In this study, the author used data from dbSNP release 130 to explore a 250-kb region flanking human poly(A) sites and found that the overall SNP frequency shows a local minimum and the conservation a peak 20 nt before the poly(A) site/transcript ends. In the present study, a larger region was analysed and variants were divided according to their frequencies. This approach allowed a more detailed description of the genetic variability of the region. Interestingly, the previously described local minimum very close to the PS (see the 1kb upstream region of Figure 4.1) is confirmed for all frequency classes except the rarest one.

Rare variants are younger than others [72] and, hence, more closely related to mutational processes. A relatively small peak of in this class of variants was observed in the first 100 bp downstream of the PS. This is not a sharp phenomenon and involves a very limited region, but transcription-related mutagenic events could be invoked to explain this increase. For example, PSs were shown [73] to be enriched with transcription-related by-products like R-loops. R-loops form during the invasion of the DNA duplex by a nascent transcript, which gives rise to a three-stranded nucleic acid structure formed by an RNA:DNA hybrid plus a displaced DNA strand. These structures were thought to be harmful with respect to genome integrity [73] and a possible source of different forms of genome instability including mutations, recombination, and chromosome rearrangements [69].

Moreover, the insurgence of novel variants, could be more o less active depending on the local conformation of the genomic region (i.e., the nucleosome occupancy). In fact, it has been reported that nucleosomes could have a strong inhibitory effect on mutation reparability, limiting the access of repair proteins [25–29]. Also, the authors of [26] reported that damage within the nucleosome core is repaired at a rate of about 10% of that of naked DNA.

To explore the eventual relationship between the rare allele distribution and the nucleosome occupancy, two cell lines (GM12878 and K562) were examined. For both cell lines, a positive correlation was found between the rare allele distribution and the nucleosome occupancy. This positive correlation supports the hypothesis that the slight increase of rare variants near PSs could be related, at least in part, to a reduced efficiency of the repair mechanisms in the regions occupied by nucleosomes [25–29].

It should be noted that the value of these correlations is in part influenced by the presence of some outlier values in the associated BNP distribution. Among these outlier values, it is interesting to note a strong depletion for nucleosomes in near vicinity of the PS. This phenomenon was also described in [74] where authors reported a depletion of nucleosomes near the site of cleavage/polyadenylation regardless of their relative position in the gene for different cell types. Nucleosomes distribution is mainly determined by the DNA sequence, indeed DNA sequences rich in AT disfavour core histones [75–77]. In addition, conservative sequence elements and the protein binding to them also play an important role in causing nucleosome depletion near polyA sites. The right nucleosome distribution near polyadenylation sites regulates the 3’ end processing of precursor messenger RNA (pre-mRNA), a key event in the transcription termination, mRNA stability, and export [78, 79].

Nevertheless, the two BNP distributions that were observed for the analysed cell lines correlate quite differently with rare variants distribution, with a lower value for the Gm12878 cell line. The analysis of nucleosomes occupation of more cell lines could help to better describe this phenomenon.

Considering common variants, the decreasing frequency values that was observed moving from coding regions to the PS could reliably explain purifying selection. The presence in the same region of a lower density of common variants (compared with rarer ones) seems to indicate that only a part of the genetic variability created by mutation can reach a high frequency in the population. This observation is also supported by the strong anti-correlation of the common variant BVF signal with GERP.

The processing of pre-mRNA 3' end is an essential step in the eukaryotic gene expression. Deregulation of the 3' end processing due to poly(A) signal mutations can have a pathogenic effect [80]. When the potential pathogenicity of each class of variants was analysed, it was found a decrease as the frequency of variants increases, probably for the effect of purifying selection. Moreover, for all frequency classes, data showed that deleteriousness increases as the distance from the PS decreases, confirming the relevance of regions surrounding PSs for correct transcription.

4.8 Materials and Methods

4.8.1 PS selection

Genomic coordinates of 43,186 human poly(A) sites were downloaded from the UCSC track "polyaDb". This track reports data from the polyA_DB server http://exon.umdnj.edu/polya_db, a web resource for analysis of pre-mRNA cleavage and polyadenylation sites.

Poly(A) sites contained in this database are genomic loci signalling the beginning of a poly(A) tail in a nascent mRNA transcript, and are predicted on the basis of expressed sequence tag/cDNA evidence [6]. The majority of these sites are represented by single DNA positions (57%), while 2% of them have length greater than 50 bp.

In order to exclude eventual biases in the analysis, deriving from inaccurate predictions, only the $N_{PS} = 42,382$ sites having length less or equal to 50 bp were retained.

4.8.2 Variant selection

The human genomic variants used in this study were downloaded from the “snp141” UCSC table. This track consists of all human genomic variants reported in dbSNP build 141, a massive collection of molecular variants (including SNPs, insertions and deletions) from heterogeneous studies. met the following conditions were selected:

1. include the term “1000GENOMES” in the “submitters” field;
2. have the value of the field “class” equal to “single”;
3. have the value of the field “alleleFreqCount” equal to 2;
4. have both of the two comma-separated values of the field “alleleNs” greater than 0; and
5. have the sum of the two comma-separated values of the field “alleleNs” greater than 1000.

By applying this filter, 38,120,928 variants were obtained.

Only the variants from low coverage set of 1000 Genomes phase 1 release were retained, by excluding the ones not having the value “LOWCOV” for the field “SNPSOURCE" in the 1000 Genomes vcf file available from ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20110521/ALL.wgs.phase1_release_v3.20101123.snps_indels_sv.sites.vcf.gz.

This last filter led to a final set of 2,720,492 variants considering only the ones located in the 10-kb region surrounding the above-selected PSs.

Because there is no explicit representation of the MAF in this dataset, it was computed by taking the minimum of the two allele frequencies for each variant.

4.8.3 Nucleosome positioning scores

As in the previous study, nucleosome localisation data were downloaded for two cell lines, namely Gm12878 and K562, from the ENCODE UCSC tracks “wgEncodeSydhNsomeGm12878Sig” and “wgEncodeSydhNsomeK562Sig”, respectively [64].

These tracks contain density signal maps of nucleosome positions produced by the micrococcal nuclease-sequencing technique (for a more accurate description of this signal see Materials and Methods section of Chapter 3).

4.8.4 GERP evolutionary scores

The impact of conservative evolutionary forces was evaluated using GERP. Conservation data for the regions of interest were downloaded from the “allHg19RS_BW” UCSC track.

The constraint intensity at each individual alignment position was quantified in terms of a “rejected substitutions” score, which is defined as the number of substitutions expected under neutrality minus the number of substitutions observed at the position [38]. Expected and observed substitutions were computed by aligning hg19 and 35 other mammalian species (for a more accurate description of this signal see Materials and Methods section of Chapter 3).

4.8.5 CADD scores

CADD scores were generated using the method of Kircher et al. [7]. CADD data for 1000 Genomes variants were downloaded from http://cadd.gs.washington.edu. The file contains the genomic position and the “raw” and “scaled” C-scores for each variant. For the purposes of this work the raw score was employed (for a more accurate description of this signal see Materials and Methods section of Chapter 3).

4.8.6 Statistical analysis

Definition of upstream and downstream regions

Upstream and downstream regions were defined on each strand for each PS by considering the direction of the transcriptional process following the procedure described in Chapter 2. For each PS, the upstream region was defined as the 5000-bp region on the 5' side if the PS is located on the coding strand and as the 5000-bp region on the 3' side if the PS lies on the noncoding strand. Conversely, the downstream region was defined as the 5000-bp region on the 3' side of the PS if this is located on the plus strand and as the 5000-bp region on the 5' side if the PS is on the minus strand. Let N_{PS} be the number of analysed PSs and let $PS(i)$ be the position of the i-th PS and $Bin(i,j)$ the bin in position j with respect to $PS(i)$ as defined in Chapter 2.

BVF calculation

Let us denote the number of variants falling into the $Bin(i,j)$ region as $vars(i,j)$. By following the same procedure described in the previous chapter, we can define the normalised mean variant frequency for the j-th bin as follows:

$$BVF(j) = \frac{\sum_{i=1}^{N_{PS}} vars(i,j)}{N_{PS} * N_{HITS}} \tag{4.1}$$

with

$$N_{HITS} = \sum_{i=1}^{N_{PS}} \sum_{j\in\{-100,\ldots,-1,+1,\ldots,+100\}} vars(i,j) \tag{4.2}$$

Positional effects on BFV values

For each considered subset of SNPs, a neutral model in which variants are uniformly distributed among bins and among all PS regions was considered.

In particular, under the null hypothesis, the following equation is true:

$$Null\_BVFmean = \frac{1}{N_{PS} * 200} \tag{4.3}$$

The difference between the observed mean values $BVF(j)$ and the $Null\_BVFmean$ was then tested using a two-tailed t-test. All 200 p-values were finally corrected using the Bonferroni method.

To investigate the presence of different positional effects on BVF values among classes, the difference between rare variant-normalized BVF (BVF_{rare}) and common variant-normalized BVF (BVF_{common}) values was calculated as follows:

$$\Delta BVF(j) = BVF_{rare}(j) - BVF_{common}(j) \tag{4.4}$$

The statistical significance of the ΔBVF values was the evaluated by testing their difference from the neutral value 0. The above differences were tested using a two-tailed t-test and the corresponding p-values were finally corrected using the Bonferroni procedure.

BNP calculation

Let $NUC(i, j)$ be the set of nucleosome density values from "wgEncodeSydhNsomeGm12878Sig" or "wgEncodeSydhNsomeK562Sig" (for the Gm12878 and K562 cell lines, respectively) associated with sites placed inside $Bin(i, j)$, and let $|NUC(i, j)|$ be its cardinality.

If we denote the sum of all nucleosome scores from $NUC(i, j)$ as $score(i, j)$, we can compute the average nucleosome positioning score for the j-th bin as follows:

$$BNP(j) = \frac{\sum_{i=1}^{N_{PS}} score(i, j)}{\sum_{i=1}^{N_{PS}} |NUC(i, j)|} \tag{4.5}$$

Correlations between the BNP and rare variant BVF were obtained using Pearson's product moment correlation coefficient and tested as described in Chapter 2.

Evolutionary analysis

Let $GER(i, j)$ be the set of GERP score values from the "allHg19RS_BW" sites whose genomic position falls inside the $Bin(i, j)$ region, and let $|GER(i, j)|$ be its cardinality. By denoting the sum of all scores belonging to $GER(i,j)$ as $score(i,j)$, we can compute the bin average GERP score for the j-th bin as follows:

$$BGS(j) = \frac{\sum_{i=1}^{N_{PS}} score(i,j)}{\sum_{i=1}^{N_{PS}} |GER(i,j)|} \tag{4.6}$$

Correlations between the BGS and common variant BVF were obtained by Pearson's product moment correlation coefficient and tested as described in Chapter 2.

Deleteriousness analysis

Let $DEL(i,j)$ be the set of deleteriousness values (from the CADD 1000 Genomes dataset) associated with sites placed inside $Bin(i,j)$, and $|DEL(i,j)|$ its cardinality.

If we denote the sum of all CADD scores belonging to $DEL(i,j)$ as $score(i,j)$, we can compute the "bin average CADD score" for the j-th bin as follows:

$$BCS(j) = \frac{\sum_{i=1}^{N_{PS}} score(i,j)}{\sum_{i=1}^{N_{PS}} |DEL(i,j)|} \tag{4.7}$$

For each bin j, the difference in the mean BCS value among the four frequency classes was tested by means of a "one-way analysis of variance" test. It was then tested if the observed local differences are representative of a global difference of the whole signals by means of Fisher's method as described in Chapter 2.

Statistical significance assessment

This study was conducted considering a p-value of 0.001 as statistically significant. All statistical analyses were performed using R ver. 3.1.2 software.

Chapter 5

Conclusions

5.1 Discussion

The advances in genomic technology in the last ten years led to a dramatic reduction in the costs for researchers to analyse and produce genomic data. As a consequence, the amount of produced genomic data has rapidly increased and it is actually growing and accumulating at exponential rates.

It is a matter of fact that we are producing more data than we can exhaustively analyse with current technologies and this is demonstrated by the increasing amount of “in silico” studies that are solely based on publicly available datasets.

The variety of biological features, for which genomic data have been produced, leads to the unprecedented possibility to analyse relationships between these features at a genome wide level for an ever-increasing number of samples. Specialised bioinformatics approaches are thus needed to approach this enormous amount of data and to investigate biological problems in a focused manner.

The main topic of this work was to develop techniques to analyse genomic signals related to the transcriptional process in human, by taking advantage of the current availability of public datasets.

This activity led to the creation of a series of statistical/bioinformatics procedures to extract and analyse several aspects of the distribution of genomic signals in a set of genomic loci of interest. These techniques were applied to investigate the structural variation around human transcription start sites and human polyadenylation sites as well as the influence of evolutionary and non-evolutionary forces in shaping this variability.

Several genomic structural signals, from single nucleotide polymorphisms to nucleosomes occupation and several genomic related characteristics (conservation, biased gene conversion and pathogenicity) were analysed.

The overall results showed that variants tend to distribute not at random manner around the analysed sites. The way in which variants distribute seems to be related to their frequency. Also, the analysis of the distribution of other genomic signals suggests that sequence composition, chromatin structure and natural selection play an active role on the insurgence and the maintenance of such diversity.

These two studies confirm the presence of strong purifying selection phenomena in regions around TSSs and PSs, and suggests a relationship of this force with the distribution of common variants in human. Also, peculiar conformation of nucleosome occupation density were found in the analysed regions and related with the distribution of rare variants, supporting the hypothesis that nucleosomes play an active role in the insurgence of mutations.

These points are actually intriguing but further studies are needed to confirm these hypothesis. In fact, despite of the great amount of variant data currently available, rarest variants are still far to be completely characterised at genome wide level. The ongoing reduction of sequencing costs and the future availability of datasets, that contain genome-wide variant characterisation for an even growing number of individuals, will allow a refinement of the current analysis. Moreover, experimental evidence will be also needed to confirm the investigated hypothesis.

The only evolutionary force analysed in these two studies (by means of the GERP scores) was purifying selection. This latter is generally considered an ancient selection phenomenon and, in fact, it is usually measured by cross-species genome comparisons. We know that also recent positive selection phenomena are able to alter the frequency of a genetic variant in a population of individuals. This point couldn't be investigated in the presented analyses mainly because of its local behaviour. In fact, recent positive selection phenomena in humans affect a very lower number loci and a lower number genes if compared with negative selection. Thus, their effects on the distribution of variants at a genomic scale are likely to be negligible.

Other DNA related structures and forces may be implied in the generation and evolution of variants in the analysed regions. Also, the transcriptional process is still far to be fully characterised and other factors could emerge in future studies to be implied in the mutagenesis of these regions.

5.2 Future directions

This work took into account two particular class of loci in the human genome: the transcription start sites and the polyadenylation sites. These two classes roughly represent the regions where RNA transcription machinery respectively begins and terminates its activity. These regions were chosen as the subject of these two analyses because of their importance in the transcription process.

Other genomic region could represent good candidates for this kind of analysis, as for example the gene body. Indeed, the analysis of the gene body would complete the picture about variants distribution in the coding regions of the genome. Also in this case, nucleosome-related mutagenic effects could be hypothesised and selective forces can be expected to be more intensive than in intergenic regions. The analysis of such kind of structure involves regions that can greatly differ in length, even of in orders of magnitude. This problem can be overcome by enhancing the developed techniques and considering relative distances between analysed elements instead of absolute ones.

Other kinds of DNA features, rather than sequence variation, could be investigated by using the presented approaches. Of particular, and actual, interest are epigenetic features like DNA methylation and histone modifications. The existence of a relationship between methylation levels and transcriptional activity has been extensively investigated. It was demonstrated that DNA methylation plays a functional role in the transcriptional process and that the methylation status of particular regions of gene are strongly associated with their expression levels. An intriguing question to investigate could be if transcriptional activity of genes has in turn an impact on the methylation signature of those genes. The analysis of the distribution of methylation levels in transcription related regions of the genome could give hints about the existence of such phenomena. Currently, epigenetic profiles for several tissues and samples are available to be analysed, as for example data from ENCODE project [5] or from the recent release of Roadmap Epigenomics project [81].

5.3 Conclusion

The main result of this work is the demonstration that genetic variants as well as DNA chromatin structures do have peculiar conformations around transcription-related sites in human. These conformations were shown to be strongly connected with several mutagenic, evolutionary and non-evolutionary forces. The analysis of transcription start sites and polyadenylation sites revealed that the density of SNPs in the corresponding surrounding regions varies with the distance from these sites. Furthermore, it is characterised by regular patterns, especially in the coding side of the those structures. These patterns were shown, in both classes of sites, to be dependent on the variants' frequency, suggesting the existence of forces that control the variability of these regions. Such forces were investigated and the distribution of genetic variants was related to the nucleosome occupation, the sequence context and the functional severity of the studied regions. The presented results need of course to be extended and experimentally validated, but represent a step forward in our understanding of how genetic variability arise and how it is maintained in such critical DNA regions.

The presented works were carried out in the Prof. Cocozza's lab under the supervision of Prof. Sergio Cocozza and of the Prof. Gennaro Miele. I want to acknowledge them for their constant support, the interesting discussions and for their illuminating insights.

References

[1] Michael Lynch. Evolution of the mutation rate. *Trends in Genetics*, 26(8):345–352, 2010.

[2] Mitchell Guttman, Manuel Garber, Joshua Z Levin, Julie Donaghey, James Robinson, Xian Adiconis, Lin Fan, Magdalena J Koziol, Andreas Gnirke, Chad Nusbaum, et al. Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincrnas. *Nat Biotechnol*, 28(5):503–510, 2010.

[3] Alessandra Montecucco and Giuseppe Biamonti. Pre-mrna processing factors meet the dna damage response. *Front. Genet.*, 4, 2013.

[4] Peng Cui, Qiang Lin, Feng Ding, Songnian Hu, and Jun Yu. The transcript-centric mutations in human genomes. *Genomics, Proteomics and Bioinformatics*, 10(1):11–22, 2012.

[5] The ENCODE Project Consortium. The encode (encyclopedia of dna elements) project. *Science*, 306(5696):636–640, 2004.

[6] H. Zhang. Polya_db: a database for mammalian mrna polyadenylation. *Nucleic Acids Research*, 33(Database issue):D116–D120, 2004.

[7] Martin Kircher, Daniela M Witten, Preti Jain, Brian J O'Roak, Gregory M Cooper, and Jay Shendure. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet*, 46(3):310–315, 2014.

[8] L. Zhang, S. Kasif, C. R. Cantor, and N. E. Broude. Gc/at-content spikes as genomic punctuation marks. *Proceedings of the National Academy of Sciences*, 101(48):16855–16860, 2004.

[9] Surajit Basak and Tapash Chandra Ghosh. On the origin of genomic adaptation at high temperature for prokaryotic organisms. *Biochemical and Biophysical Research Communications*, 330(3):629–632, 2005.

[10] Giorgio Bernardi. Isochores and the evolutionary genomics of vertebrates. *Gene*, 241(1):3–17, 2000.

[11] Laurence D Hurst and Elizabeth J.B Williams. Covariation of gc content and the silent site substitution rate in rodents: implications for methodology and for the evolution of isochores. *Gene*, 261(1):107–114, 2000.

[12] Nicolas Galtier. Gene conversion drives gc content evolution in mammalian histones. *Trends in Genetics*, 19(2):65–68, 2003.

[13] A. E. Vinogradov. Isochores and tissue-specificity. *Nucleic Acids Research*, 31(17):5212–5220, 2003.

[14] Grzegorz Kudla, Leszek Lipinski, Fanny Caffin, Aleksandra Helwak, and Maciej Zylicz. High guanine and cytosine content increases mrna levels in mammalian cells. *Plos Biol*, 4(6):e180, 2006.

[15] Y Guo and DC Jamison. The distribution of snps in human gene regulatory regions. *BMC Genomics*, 6:140, 2005.

[16] Phil Green, Brent Ewing, Webb Miller, Pamela J. Thomas, and Eric D. Green. Transcription-associated mutational asymmetry in mammalian evolution. *Nat Genet*, 33(4):514–517, 2003.

[17] D. G. Hwang and P. Green. Bayesian markov chain monte carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. *Proceedings of the National Academy of Sciences*, 101(39):13994–14001, 2004.

[18] P. Polak and P. F. Arndt. Transcription induces strand-specific mutations at the 5' end of human genes. *Genome Research*, 18(8):1216–1223, 2008.

[19] N. De Maio, C. Schlotterer, and C. Kosiol. Linking great apes genome evolution across time scales using polymorphism-aware phylogenetic models. *Molecular Biology and Evolution*, 30(10):2249–2262, 2013.

[20] Steven Henikoff. Nucleosome destabilization in the epigenetic regulation of gene expression. *Nat Rev Genet*, 9(1):15–26, 2008.

[21] Cizhong Jiang and B. Franklin Pugh. Nucleosome positioning and gene regulation: advances through genomics. *Nat Rev Genet*, 10(3):161–172, 2009.

[22] Dustin E. Schones, Kairong Cui, Suresh Cuddapah, Tae-Young Roh, Artem Barski, Zhibin Wang, Gang Wei, and Keji Zhao. Dynamic regulation of nucleosome positioning in the human genome. *Cell*, 132(5):887–898, 2008.

[23] A Gontijo. Repairing dna damage in chromatin. *Biochimie*, 85(11):1133–1147, 2003.

[24] Yeganeh Ataian and Jocelyn E. Krebs. Five repair pathways in one context: chromatin modification during dna repair. *Biochemistry and Cell Biology*, 84(4):490–494, 2006.

[25] Anja Groth, Walter Rocha, Alain Verreault, and Geneviève Almouzni. Chromatin challenges during dna replication and repair. *Cell*, 128(4):721–733, 2007.

[26] R. Hara, J. Mo, and A. Sancar. Dna damage in the nucleosome core is refractory to repair by human excision nuclease. *Molecular and Cellular Biology*, 20(24):9173–9181, 2000.

[27] Maria Meijer and Michael J. Smerdon. Accessing dna damage in chromatin: Insights from transcription. *Bioessays*, 21(7):596–603, 1999.

[28] F. Thoma. Light and dark in chromatin repair: repair of uv-induced dna lesions by photolyase and nucleotide excision repair. *The EMBO Journal*, 18(23):6585–6598, 1999.

[29] K. Ura. Atp-dependent chromatin remodeling facilitates nucleotide excision repair of uv-induced dna lesions in synthetic dinucleosomes. *The EMBO Journal*, 20(8):2004–2014, 2001.

[30] Motoo Kimura. *The neutral theory of molecular evolution*. Cambridge University Press, 1983.

[31] M. C. Frith. Evolutionary turnover of mammalian transcription start sites. *Genome Research*, 16(6):713–722, 2006.

[32] Martin S. Taylor, Chikatoshi Kai, Jun Kawai, Piero Carninci, Yoshihide Hayashizaki, and Colin A. M. Semple. Heterotachy in mammalian promoter evolution. *PLoS Genetics*, 2(4):e30, 2006.

[33] J.-C. Walser and A. V. Furano. The mutational spectrum of non-cpg dna varies with cpg content. *Genome Research*, 20(7):875–882, 2010.

[34] K Higasa and K Hayashi. Periodicity of snp distribution around transcription start sites. *BMC Genomics*, 7:66, 2006.

[35] L. Zhang. Mammalian housekeeping genes evolve more slowly than tissue-specific genes. *Molecular Biology and Evolution*, 21(2):236–239, 2003.

[36] Desiree Tillo, Noam Kaplan, Irene K. Moore, Yvonne Fondufe-Mittendorf, Andrea J. Gossett, Yair Field, Jason D. Lieb, Jonathan Widom, Eran Segal, and Timothy R. Hughes. High nucleosome occupancy is encoded at human regulatory sequences. *PLoS ONE*, 5(2):e9129, 2010.

[37] Tanya Vavouri and Ben Lehner. Human genes with cpg island promoters have a distinct transcription-associated chromatin organization. *Genome Biol*, 13(11):R110, 2012.

[38] Eugene V. Davydov, David L. Goode, Marina Sirota, Gregory M. Cooper, Arend Sidow, and Serafim Batzoglou. Identifying a high fraction of the human genome to be under selective constraint using gerp++. *PLoS Computational Biology*, 6(12):e1001025, 2010.

[39] John A. Capra, Melissa J. Hubisz, Dennis Kostka, Katherine S. Pollard, and Adam Siepel. A model-based analysis of gc-biased gene conversion in the human and chimpanzee genomes. *PLoS Genetics*, 9(8):e1003684, 2013.

[40] P. Polak and P. F. Arndt. Long-range bidirectional strand asymmetries originate at cpg islands in the human genome. *Genome Biology and Evolution*, 1(0):189–197, 2009.

[41] L. Ponger and D. Mouchiroud. Cpgprod: identifying cpg islands associated with transcription start sites in large genomic mammalian sequences. *Bioinformatics*, 18(4):631–633, 2002.

[42] J. Meunier. Recombination drives the evolution of gc-content in the human genome. *Molecular Biology and Evolution*, 21(6):984–990, 2004.

[43] Augustine Kong, Daniel F. Gudbjartsson, Jesus Sainz, Gudrun M. Jonsdottir, Sigurjon A. Gudjonsson, Bjorgvin Richardsson, Sigrun Sigurdardottir, John Barnard, Bjorn Hallbeck, Gisli Masson, et al. A high-resolution recombination map of the human genome. *Nat. Genet.*, 2002.

[44] G. A. T. McVean. The fine-scale structure of recombination rate variation in the human genome. *Science*, 304(5670):581–584, 2004.

[45] Arian FA Smit. Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Current Opinion in Genetics and Development*, 9(6):657–663, 1999.

[46] Lars Guelen, Ludo Pagie, Emilie Brasset, Wouter Meuleman, Marius B. Faza, Wendy Talhout, Bert H. Eussen, Annelies de Klein, Lodewyk Wessels, Wouter de Laat, et al. Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature*, 453(7197):948–951, 2008.

[47] S. Delgado. Initiation of dna replication at cpg islands in mammalian chromosomes. *The EMBO Journal*, 17(8):2426–2435, 1998.

[48] Sequeira-Mendes et al. Transcription initiation activity sets replication origin efficiency in mammalian cells. *PLoS Genetics*, 5(4):e1000446, 2009.

[49] C. Cayrou, P. Coulombe, Vigneron, et al. Genome-scale analysis of metazoan replication origins reveals their organization in specific but flexible sites defined by conserved features. *Genome Research*, 21(9):1438–1449, 2011.

[50] M. Huvet, S. Nicolay, M. Touchon, B. Audit, Y. d'Aubenton Carafa, A. Arneodo, and C. Thermes. Human gene organization driven by the coordination of replication and transcription. *Genome Research*, 17(9):1278–1285, 2007.

[51] A. Necsulea, C. Guillet, J.-C. Cadoret, M.-N. Prioleau, and L. Duret. The relationship between dna replication and human genome organization. *Molecular Biology and Evolution*, 26(4):729–741, 2009.

[52] Jiang Zhu, Fuhong He, Songnian Hu, and Jun Yu. On the nature of human housekeeping genes. *Trends in Genetics*, 24(10):481–484, 2008.

[53] Feng Gong, YoungHo Kwon, and Michael J. Smerdon. Nucleotide excision repair in chromatin and the right of entry. *DNA Repair*, 4(8):884–896, 2005.

[54] Michael Y Tolstorukov, Natalia Volfovsky, Robert M Stephens, and Peter J Park. Impact of chromatin structure on sequence variability in the human genome. *Nat Struct Mol Biol*, 18(4):510–515, 2011.

[55] John M. Hinz. Role of homologous recombination in dna interstrand crosslink repair. *Environ. Mol. Mutagen.*, pages NA–NA, 2010.

[56] J. Majewski. Distribution and characterization of regulatory elements in the human genome. *Genome Research*, 12(12):1827–1836, 2002.

[57] Anna Portela and Manel Esteller. Epigenetic modifications and human disease. *Nat Biotechnol*, 28(10):1057–1068, 2010.

[58] Motoo Kimura. *Diffusion models in population genetics*. Methuen, 1964.

[59] NM Cohen et al. Primate cpg islands are maintained by heterogeneous evolutionary regimes involving minimal selection. *Cell*, 145(5):773–786, 2011.

[60] Nicolas Galtier and Laurent Duret. Adaptation or biased gene conversion? extending the null hypothesis of molecular evolution. *Trends in Genetics*, 23(6):273–277, 2007.

[61] T. Nagylaki. Evolution of a finite population under gene conversion. *Proceedings of the National Academy of Sciences*, 80(20):6278–6281, 1983.

[62] Nicolas Galtier et al. Gc-biased gene conversion promotes the fixation of deleterious amino acid changes in primates. *Trends in Genetics*, 25(1):1–5, 2009.

[63] T. R. Dreszer, G. D. Wall, D. Haussler, and K. S. Pollard. Biased clustered substitutions in the human genome: The footprints of male-driven biased gene conversion. *Genome Research*, 17(10):1420–1430, 2007.

[64] Anton Valouev, Steven M. Johnson, Scott D. Boyd, Cheryl L. Smith, Andrew Z. Fire, and Arend Sidow. Determinants of nucleosome organization in primary human cells. *Nature*, 474(7352):516–520, 2011.

[65] Krishanpal Anamika, Gyenis, et al. How to stop. *Transcription*, 4(1):7–12, 2013.

[66] J. D. Lewis, S. I. Gunderson, and I. W. Mattaj. The influence of 5' and 3' end structures on pre-mrna metabolism. *Journal of Cell Science*, 1995(Supplement 19):13–19, 1995.

[67] P. Richard and J. L. Manley. Transcription termination by nuclear rna polymerases. *Genes and Development*, 23(11):1247–1269, 2009.

[68] Anne Helmrich, Monica Ballarino, Evgeny Nudler, and Laszlo Tora. Transcription-replication encounters, consequences and genomic instability. *Nat Struct Mol Biol*, 20(4):412–418, 2013.

[69] Andres Aguilera and Tatiana Garcia-Muse. R loops: From transcription byproducts to threats to genome stability. *Molecular Cell*, 46(2):115–124, 2012.

[70] John C. Castle. Snps occur in regions with less genomic sequence conservation. *PLoS ONE*, 6(6):e20660, 2011.

[71] G. M. Cooper. Distribution and intensity of constraint in mammalian genomic sequence. *Genome Research*, 15(7):901–913, 2005.

[72] G.A. Watterson and H.A. Guess. Is the most frequent allele the oldest? *Theoretical Population Biology*, 11(2):141–160, 1977.

[73] Konstantina Skourti-Stathaki, Kinga Kamieniarz-Gdula, and Nicholas J. Proudfoot. R-loops induce repressive chromatin marks over mammalian gene terminators. *Nature*, 516(7531):436–439, 2014.

[74] Huan HUANG, Hongde LIU, and Xiao SUN. Nucleosome distribution near the 3' ends of genes in the human genome. *Bioscience, Biotechnology and Biochemistry*, 77(10):2051–2055, 2013.

[75] Noam Kaplan, Irene K. Moore, Yvonne Fondufe-Mittendorf, Andrea J. Gossett, Desiree Tillo, Yair Field, Emily M. LeProust, Timothy R. Hughes, Jason D. Lieb, Jonathan Widom, et al. The dna-encoded nucleosome organization of a eukaryotic genome. *Nature*, 458(7236):362–366, 2008.

[76] Travis N. Mavrich, Cizhong Jiang, Ilya P. Ioshikhes, Xiaoyong Li, Bryan J. Venters, Sara J. Zanton, Lynn P. Tomsho, Ji Qi, Robert L. Glaser, Stephan C. Schuster, et al. Nucleosome organization in the drosophila genome. *Nature*, 453(7193):358–362, 2008.

[77] G. S. Chang, A. A. Noegel, T. N. Mavrich, R. Muller, L. Tomsho, E. Ward, M. Felder, C. Jiang, L. Eichinger, G. Glockner, et al. Unusual combinatorial involvement of poly-a/t tracts in organizing genes and chromatin in dictyostelium. *Genome Research*, 22(6):1098–1106, 2012.

[78] Melissa J. Moore and Nick J. Proudfoot. Pre-mrna processing reaches back totranscription and ahead to translation. *Cell*, 136(4):688–700, 2009.

[79] D. F. Colgan and J. L. Manley. Mechanism and regulation of mrna polyadenylation. *Genes and Development*, 11(21):2755–2766, 1997.

[80] Sven Danckwardt, Matthias W Hentze, and Andreas E Kulozik. 3' end mrna processing: molecular mechanisms and implications for health and disease. *The EMBO Journal*, 27(3):482–498, 2008.

[81] Anshul Kundaje, Wouter Meuleman, Jason Ernst, Misha Bilenky, Angela Yen, Alireza Heravi-Moussavi, Pouya Kheradpour, Zhizhuo Zhang, Jianrong Wang, Michael J. Ziller, et al. Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317–330, 2015.