## Genomic neighbourhood and the regulation of gene expression

### Genomic neighbourhood and transcriptional regulation

Subhajyoti De and M. Madan Babu

*MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 0QH, UK
Address for correspondence: SD (sde@mrc-lmb.cam.ac.uk) and MMB (madanm@mrc-lmb.cam.ac.uk)

### Manuscript information

- Pages: 11
- Figures: 2
- Box: 1
- Words in abstract: 170
- Words in main text: 4,735
- References: 100

### Table of contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td>1</td>
</tr>
<tr>
<td>Abbreviations</td>
<td>1</td>
</tr>
<tr>
<td>1. Introduction</td>
<td>1</td>
</tr>
<tr>
<td>2. Genomic neighbourhood and its influence on gene regulation</td>
<td>1</td>
</tr>
<tr>
<td>2.1 The eukaryotic genome is organised into genomic neighbourhoods</td>
<td>1</td>
</tr>
<tr>
<td>2.2 Genomic neighbourhood influences gene expression</td>
<td>2</td>
</tr>
<tr>
<td>3. Mechanisms that alter the genomic neighbourhood of genes</td>
<td>2</td>
</tr>
<tr>
<td>4. Implications for evolution</td>
<td>3</td>
</tr>
<tr>
<td>4.1 Species evolution</td>
<td>3</td>
</tr>
<tr>
<td>4.2 Population variation</td>
<td>4</td>
</tr>
<tr>
<td>5. Implications for development and disease</td>
<td>4</td>
</tr>
<tr>
<td>5.1 Germ-line and somatic mosaicism</td>
<td>4</td>
</tr>
<tr>
<td>5.2 Developmental disorders and cancer</td>
<td>5</td>
</tr>
<tr>
<td>6. Conclusions and outlook</td>
<td>6</td>
</tr>
<tr>
<td>Acknowledgements</td>
<td>6</td>
</tr>
<tr>
<td>Figures and figure legend</td>
<td>7</td>
</tr>
<tr>
<td>Figure 2: Consequences and implications of altering the genomic...</td>
<td>7</td>
</tr>
<tr>
<td>Box 1</td>
<td>8</td>
</tr>
<tr>
<td>References and recommended reading</td>
<td>9</td>
</tr>
</tbody>
</table>
Abstract
Recent studies have defined ‘genomic neighbourhoods’ or ‘domains’ within the eukaryotic genome. These are segments of genetic material with specific characteristics associated with them (e.g., epigenetic modifications, physical interaction with the nuclear lamina, etc). A common theme that emerges from these studies is genes that reside within these ‘neighbourhoods’ tend to display distinct transcriptional activity. Several mechanisms and factors are now known to alter the neighbourhood of genes. Indeed, genetic manipulation or natural mutations that alter the neighbourhood of a gene have been shown to affect its expression and have been suggested to contribute to speciation, introduce diversity in a population, result in germ-line and somatic mosaicism, and cause specific diseases. Thus, where exactly a gene lies in the genome and its genomic neighbourhood influences the regulation of gene expression. In this review, we discuss our current understanding on all of the above aspects. Understanding the role of genomic neighbourhood on gene regulation has fundamental implications for evolution, development, disease, and in applications such as gene therapy and genetic engineering.

Abbreviations
H3K27me3, histone H3 lysine 27 trimethylation; H2AK5ac, histone H2A lysine 5 acetylation; H3K9Me2, histone H3 lysine 9 dimethylation; CTCF, CCCTC-binding factor; CNV, copy number variation

1. Introduction
Given that the eukaryotic genetic material is organised in a complex and hierarchical manner, regulation of gene expression in eukaryotes is an intricate process that involves multiple levels of control [1]. While several studies have investigated transcriptional regulation at the level of individual genes, a number of recent studies have examined transcriptional regulation of groups of genes that reside within specific genomic regions. These studies have reported the existence of “domains” or “genomic neighbourhood”, which are typically defined as genomic regions that have specific characteristics (e.g., epigenetic marks, physical interaction with nuclear lamina, etc) associated with them. The development of appropriate experimental and computational approaches has helped us to appreciate the fundamental role of genomic neighbourhood for the regulation of gene expression. The new insights gained have major implications for a better understanding of the regulation of gene expression, thus making it timely to discuss the recent advances in this area.

In the first part of this review, we will describe recent studies which have shown that eukaryotic chromosomes are organised into distinct “domains” or “genomic neighbourhoods”. We will then describe studies which have shown that the genes within these “domains” have distinct transcriptional activity associated with them and that altering the neighbourhood of a gene can affect its expression level. We will then discuss the current understanding of mechanisms that result in the alteration of the neighbourhood of a gene and how it may influence regulation of gene expression. In the second part of this review, we will discuss the implications of alteration in genomic neighbourhood and the associated change in gene expression for speciation and population divergence. We will then consider the implications for development and disease. Finally, we will conclude this review by discussing the major experimental and computational challenges for future research and highlight how the new understanding can be exploited in specific applications such as genetic engineering and gene therapy.

2. Genomic neighbourhood and its influence on gene regulation
Various factors are known to affect the expression of one or several genes simultaneously. This may involve cis-regulatory elements such as transcription factor binding sites (spanning few base pairs) or organisation of chromosomes into territories within the nucleus (spanning an entire chromosome). In addition to the above two modes of regulation, it is becoming increasingly clear that “genomic neighbourhoods” or “chromosomal domains” (spanning several megabases), provide an important level of regulation of gene expression.

2.1 The eukaryotic genome is organised into genomic neighbourhoods
A number of studies have shown that eukaryotic chromosomes are organised into chromosomal territories within the nucleus and that the genome forms extensive and dynamic physical interactions between different segments, resulting in chromosomal loops and bridges [2-7]. Such interactions may be inter- [8,9] or intra-chromosomal [10,11], and can contribute to the silencing and activation of genes within the three-dimensional context of the nuclear architecture [2,7,12]. Thus understanding the organisation of the genome within the nucleus and its impact on gene regulation and other cellular processes such as replication and repair has been an important goal in the post-genomic era [13]. In this direction, a recent study by Lieberman-Aiden et al., probed the three-dimensional architecture of whole genomes by coupling proximity-based ligation with massively parallel sequencing [14]. This study confirmed the presence of chromosomal territories and the spatial proximity of small, gene-rich chromosomes. In addition, the study also demonstrated that the open and closed chromatin form two genome-wide compartments which are spatially segregated. In another study, Berger et al. developed a computational imaging approach that creates high-resolution probabilistic maps of subnuclear domains occupied by individual loci in yeast by analysing thousands of live cells. The authors found that genomic loci are confined to ‘gene territories’, which are much smaller than the nucleus and that this can be remodelled during transcriptional activation [15].
Transcriptionally active and repressed regions are often associated with characteristic molecular signatures. Guelen et al. investigated the architecture of the human chromosomes in the interphase nuclei of fibroblasts and revealed the presence of ∼1300 lamina-associated domains (LADs) of 0.1-10 megabases in size [16]. The borders of LADs were demarcated by the insulator binding protein CTCF and genes within these LADs typically had low gene-expression levels, suggesting that LADs represent a repressive environment. Another study by Cuddapah et al. investigated three different cell types (CD4+, Jurkat and HeLa) and described regions that demarcated active and repressive domains by profiling the global binding of CTCF [17]. Specifically, the authors reported the existence of a complementary pattern between the repressive H3K27me3 and the active H2AK5ac regions and this was shown to be separated by CTCF. Along similar lines, Pauler et al. analysed the repressive H3K27me3 mark and defined broad local enrichments (BLOCs) of histone modifications using their BLOC algorithm [18]. The genes residing in these BLOCs were found to be transcriptionally repressed. In another study, Wen et al. investigated H3K9Me2 and revealed the existence of large regions (up to 4.9 Mb) of organized chromatin K9 modifications (LOCKs) [19]. These regions were shown to be dependent on the G9a histone methyltransferase, highly conserved between human and mouse and were differentiation-state specific. For instance, such regions only covered ∼4% in undifferentiated mouse embryonic stem (ES) cells, compared to 31% in differentiated ES cells. Apart from such regions, other studies have defined Genomic Regulatory Blocks (GRB), which contain highly conserved noncoding elements that affect expression of target genes which are several kb away from the GRBs [20]. It has been shown that the target genes influenced by GRBs have transcriptional features that are distinct from the non-targets that reside in the vicinity [21]. Taken together, these studies demonstrate that the human genome is divided into large, discrete “domains” or “genomic neighbourhoods” that are highly dynamic, cell-type and differentiation state specific and associated with specific transcriptional activity.

2.2 Genomic neighbourhood influences gene expression

Position effect refers to situations where the expression level of a gene is affected by an alteration in its genomic neighbourhood or its chromosomal environment, while maintaining an intact transcription unit. This phenomenon was initially extensively studied in Drosophila [22] and was subsequently observed to be the cause for a number of human diseases [23-25]. In the 1990s, it was noted that identical reporter gene constructs (inducible by tetracycline) integrated at different sites displayed different levels of expression to the same amount of inducer in transgenic mice. This suggested that where exactly a gene was integrated in the genome influenced its ability to be expressed [26,27]. In line with this, a recent study by Flagfeldt et al. systematically investigated the effect of inserting identical lacZ constructs in twenty different integration sites and observed up to 8.7 fold differences between the sites of lowest and highest expression in yeast [28]. Another systematic study was performed by Gierman and co-workers where identical green fluorescent protein reporter constructs were integrated at 90 different chromosomal positions in the human genome [29]. The integrated gene was observed to obtain expression levels that correspond to the activity of the domains of integration. These domains were found to be up to 80 genes long and to exert an 8 fold effect on the expression levels of the integrated genes.

Ebisuya et al. studied if it is possible to pinpoint and activate a particular locus without perturbing the expression of the neighbouring genes [30]. The authors found that transcription at one locus frequently spills over into its neighbouring loci. Specifically, it was seen that rapid induction of immediate-early genes in response to growth factor stimulations was accompanied by co-upregulation of their neighbouring genes. Based on these findings, the authors suggest that transcriptional activation has a ripple effect, which may be advantageous for coordinated expression. In another study, Becskei et al. showed that the positioning of the genes along the chromosomes is correlated with stochastic gene expression in a clonal cell population [31]. Indeed, it has been proposed by Batada and Hurst that selection for reduced gene expression stochasticity might have been a factor driving the organisation of essential genes on chromosomes during evolution [32]. Taken together these studies demonstrate that in addition to cis-regulatory elements or cognate transcription factors that affect the expression level of a gene, where exactly a gene resides in the genome may influence the regulation (level and stochasticity) of its expression (Figure 1). This phenomenon, i.e., the influence of genomic neighbourhood on gene expression, also appears to be a general principle that is conserved from yeast to human.

3. Mechanisms that alter the genomic neighbourhood of genes

Tandem duplication and mobile element based mechanisms drive insertion, deletion, transposition and inversion of genetic material, leading to alteration in genomic neighbourhood of one or several genes. These are also referred to as structural alterations of the genome. Tandem duplications are thought to occur mainly through non-allelic homologous recombination and replication error [33,34] and are more prevalent in mice, chickens and rats than primates [35]. Mobile elements, such as DNA transposons, autonomous retrotransposons (e.g., L1 element), and non-autonomous retrotransposons (e.g., Alu element) play a key role in genome reorganisation [36]. Transposons are generally excised from one genomic site and integrated into another site adjacent to the parental locus by a “cut and paste” mechanism, while retrotransposons are transcribed into RNA, and then reverse transcribed and reintegrated into the genome, thereby duplicating the genomic region (“copy and paste”). Among these mobile elements, Alu elements are very active in human, covering ~11% of the human genome. Furthermore,
mobile elements such as L1s and Alus provide material for DNA mispairing and unequal crossing over, leading to deletion or segmental duplication (blocks of 200 – 400kbp) of sequences between the repeats [36]. All these aspects therefore contribute to alterations in the genomic neighbourhood within the eukaryotic genome.

Elliot et al. reported that when double-strand breaks were introduced adjacent to identical Alu elements, translocations occurred at high frequency and this predominantly arose from repair by the single-strand annealing pathway [37]. In another recent study, Rosenfeld et al. reported a mechanism of androgen receptor (AR) dependent translocations of genetic material. The authors showed that binding of AR to intronic regions first introduces inter- and intra-chromosomal interactions. By recruiting activation-induced cytidine deaminase and the L1 encoded ORF2 endonuclease, AR introduces double strand breaks in the juxtaposed loci. These segments are ligated by nonhomologous end joining pathway, thereby resulting in a translocation event [38]. Taken together, these studies demonstrate that selfish elements alone, or together with host factors, can contribute to introducing alterations in genomic neighbourhoods.

The alteration of genomic neighbourhood of a gene can affect its transcription environment in several ways. For instance, insertion or deletion of genetic material (e.g., L1 elements) around the gene can result in altered expression level [39]. This may happen via (1) introduction of epigenetic modifications in a genomic region initiated by the presence of transposable elements [40], (2) introduction of new regulatory elements, carried through retrotransposable elements [41,42] or (3) creation of new microRNAs from retrotransposable elements [43]. Alternatively, a change in genomic neighborhood may lead to incorporation of the gene into a completely different region (e.g., euchromatic or heterochromatinic region), thereby affecting its expression [44]. This may involve (1) incorporation of a gene into a constitutive heterochromatic or euchromatic region, giving rise to the classical position effect, (2) translocation of a gene next to another gene, thereby resulting in competition for the same regulatory element or juxtaposition with an enhancer element, (3) separation of the promoter from a distal tissue-specific regulatory element such as an enhancer, locus control region, or an insulator [24]. As we will discuss in the following sections, the mechanisms leading to alteration in genomic neighbourhood, thereby resulting in altered expression level of some genes, are prevalent in germ and somatic cells and have important consequences for evolution, development and disease (Figure 2).

4. Implications for evolution
Alteration of key regulatory genes can give rise to morphological variation. For example, changes in the expression level of Bmp4 and calmodulin have been shown to underlie much of the variation in beak morphology in a population of Darwin’s finches, thereby fueling evolution of new species [45]. Given that L1 elements may alter genomic neighbourhood, and altered neighbourhood may influence gene expression levels, it is possible that such events may contribute to genetic and phenotypic diversity in a population of individuals. This variation, which may have beneficial or detrimental effect on the fitness of an individual, would serve as a substrate for natural selection and may ultimately contribute to speciation.

4.1 Species evolution
Alteration of genomic neighbourhood of a gene can modify its transcriptional environment, leading to changes in the expression level of genes. For example, Xiao et al. showed that a position effect induced change in the expression of the Sun gene underlies much of the variation in the shape and size of tomatoes [44]. Recently, by comparing the genomic neighbourhood of orthologous genes in human and chimpanzee with the expression levels of their transcripts from several equivalent tissues, De et al. demonstrated that genes with altered neighborhood undergo more expression divergence (i.e., alteration in expression level of orthologous genes) than genes with conserved neighborhood [46]. They also found enrichment for genes with altered neighborhood to be expressed in a tissue-specific manner in the human brain. Based on these observations, it was proposed that in addition to single nucleotide changes that affect promoter regions, expression divergence induced by this mechanism could have contributed to the phenotypic differences between human and chimpanzee [46]. In this context, a recent study by Tian et al. showed that single nucleotide mutation rate increases close to sites of insertions and deletions in eukaryotes [47]. De and Babu showed that orthologous genes with significant alterations in their genomic neighbourhood displayed an increased sequence divergence in protein coding and non-coding regions and that this contributes to further genome-level divergence between related species [48]. In Box 1, we show that such genomic alteration induced single nucleotide divergence in promoter regions may further contribute to expression divergence. In other words, we show that orthologous genes with significant alteration in genomic neighbourhood and increased promoter divergence appear to display a much larger divergence in their expression levels between human and chimpanzee (Box 1). This observation is consistent with Kleinjanz et al., who investigated the genomic neighbourhood and expression levels of the duplicated gene copies of Pax6, which is an important developmental regulator [49]. The authors showed that the duplicated gene copy show divergence in cis-regulatory elements, diverged expression and evidence for sub-functionalization of the duplicated gene copies [49].
Along similar lines, Yanai and Hunter compared the temporal expression profiles of one-to-one orthologs in conserved or altered genomic positions in *Caenorhabditis elegans* and *Caenorhabditis briggsae*, two related species that share a near-identical developmental program [50]. The authors showed that coexpression of gene neighbours in either species is highly divergent in the other when the neighbourhood is not conserved [50], further supporting the notion that neighbouring genes within chromosomal “domains” display similar transcriptional activity [50-52]. Given that genomic neighbourhood affects gene expression, one can imagine that the genes regulated by specific transcription factors may not be randomly distributed on eukaryotic chromosomes. Indeed a study by Janga *et al.* presented evidence for the existence of a higher-order organization of genes across and within chromosomes that is constrained by transcriptional regulation [53]. In particular, the authors revealed that the target genes of several transcription factors for yeast, human and mouse are encoded in an ordered manner both across and within the different linear chromosomes. Based on these observations, they concluded that genomic alterations leading to specific organization of genes that allowed for efficient control of transcription might have been selected during evolution [53].

4.2 Population variation

Structural variation, defined as alterations in the genome involving insertion, deletion, inversion or translocation of genomic segments, and their functional consequences in a population have by far been most studied for the human genome [54]. Functional consequences of such alterations include gene dosage, gene disruption, fusion and position effect and may have fundamental implications for healthy and disease conditions [55]. For instance, large, rare chromosomal deletions have been shown to be associated with severe early-onset obesity [56]. A recent study, Conrad and co-workers investigated the functional impact of copy number variation (CNV) and generated a comprehensive map of 11,700 CNVs with segments over 400 bp [57]. The authors showed that the predominant mutational mechanisms differ among CNV size classes. In addition, they showed that retrotransposition has duplicated and inserted both coding and non-coding DNA segments randomly around the genome. The authors also identified 30 loci with CNVs that are candidates for influencing disease susceptibility [57].

Xing *et al.* have reported a detailed analysis of structural variants in the complete DNA sequence of an individual (HuRef) and showed that mobile elements such as Alu, L1, and SVA play an important role in generating inter-individual structural variation [58]. A systematic investigation of CNVs and associated repeat elements by Kim *et al.* showed that approximately 40 million years ago, during the "Alu burst" in retrotransposition activity, non-allelic homologous recombination, first mediated by Alus and then the by newly formed CNVs themselves, was the main driver of genome rearrangements. However, its relative importance was found to be decreased markedly since then, with proportionally more events now stemming from other repeats and from non-homologous end-joining [59]. This study thereby provides an evidence for a change in the process of formation of CNVs in recent evolutionary history of humans [59].

A recent copy number variation (CNV) study in macaque showed that many copy number variable regions are shared between human and macaque, underscoring their inherent genomic instability [60]. By comparing three human genomes, Alkan *et al.* estimate that on average, 73-87 genes vary in copy number between any two individuals and find that these genic differences overwhelmingly correspond to segmental duplications [61]. A combined genome and transcriptome analysis in wild and inbred mice strains reported that CNVs not only influence the expression of genes within these regions, but also that of the genes in their neighbourhood and that the effect extends up to ~0.5Mb [62]. The authors also demonstrated that dosage alterations of brain-expressed genes are relatively rare and are buffered by tighter transcriptional regulation [62]. Taken together, these studies reveal that structural alteration induced by transposable elements is highly prevalent in the human genome and that such mechanisms introduce both genotypic and phenotypic variation in a population by affecting the expression level of genes.

5. Implications for development and disease

Alteration of genomic neighbourhood of the genes can occur in somatic and germ cells within the lifetime of an individual and can be propagated when cells divide. While the benign effects can result in somatic and germ-line mosaicism, perturbation of the neighbourhood of key genes (e.g., regulators such as transcription factors and signalling genes) has been associated with several developmental diseases and cancer.

5.1 Germ-line and somatic mosaicism

A number of studies increasingly support the idea that there is considerable heterogeneity among isogenic cell population in terms of large structural variations and copy number variations that contribute to germ-line and somatic mosaicism. A recent study reported significant copy number variation among tissues collected from the same human individuals, suggesting that somatic mosaicism arose during development [63]. Another study using single cell analysis, reported a high copy number variation between sperms obtained from the same donor, thereby presenting evidence for mosaicism in the germ-line [64]. Many of the molecular mechanisms for structural variation described in the previous section, such as L1 retrotransposition, are especially active during development and in the germ-line and contribute to somatic and germ-line mosaicism [65]. While the consequences of such mutagenic events are likely to be buffered by Hsp90 chaperone family proteins (which have been
proposed to act as a phenotypic buffer and as a capacitor for morphological evolution [66]), partial inhibition of its function can uncover cryptic mutations, thereby leading to mosaicism in the expression of a phenotype [67]. Yeyati et al. showed in zebrafish that mild perturbation of Hsp90 function at critical developmental stages may underpin the variable penetrance and expressivity of many developmental anomalies [68]. In this context it is interesting to note a recent study by Specchia et al. where the authors provide a possible explanation by which Hsp90 may buffer cryptic mutations. They showed that Hsp90 generally prevents phenotypic variation by suppressing the mutagenic activity of transposons. In other words, mutations in Hsp90 tend to introduce phenotypic variation by facilitating the mutagenic activity of transposons in individual cells during Drosophila development, possibly leading to somatic mosaicism and alteration in phenotype [69].

Recently, Kano et al. report that the RNA of L1 element is abundant in both germ cells and embryos, and showed that L1 RNA transcribed in male or female germ cells can be carried over through fertilization and integrate in daughter cells during embryogenesis, thereby resulting in somatic mosaicism [70]. The observation that L1 RNA can be carried over through fertilization and integrate during embryogenesis is an interesting example of heritability of RNA independent of its encoding DNA [70]. Coufal et al. and Muotri et al. showed that de novo L1 retrotransposition events occur in neural progenitor cells isolated from human fetal brain and derived from human embryonic stem cells, and that in principle, such events have the potential to contribute to somatic mosaicism [71,72]. Another study that reports copy number variation between monozygotic twins highlights that somatic mosaicism can operate during early embryogenesis and stressed their importance in design and analysis of genotype-phenotype studies [73]. While many of these studies primarily reported genetic heterogeneity between isogenic cell populations, one can imagine that such genomic heterogeneity contributes to heterogeneity in gene expression levels and may contribute to the “altered” phenotype of the relevant subpopulation of cells.

5.2 Developmental disorders and cancer

The molecular and genetic basis of several human diseases has been revealed in the last couple of decades. While many diseases have been linked with mutations or deletions in specific causative genes, it is less often appreciated that healthy condition also depends on the spatially, temporally, and quantitatively correct expression of genes [6,23]. Several studies have shown that alteration of genomic neighbourhood induced change in gene expression level of key regulatory genes (e.g., transcription factors or signalling genes) can cause specific developmental defects. For example, mutations that result in deletion or insertion of neighbouring genomic regions thereby resulting in altered expression of the regulatory proteins Sox9 and REEP3 has been linked with campomelic dysplasia [74] and autism [75]. For excellent reviews and a comprehensive list of position effect dependent developmental disorders and diseases associated with CNVs, please see [6,23-25,55,76]).

In addition to causing developmental disorders, several mutations that alter genomic neighbourhood have been associated with cancer [77]. Recent advances in sequencing technology has permitted the systematic characterisation of the different mutations on a genomic scale [78-83]. Campbell et al. investigated individuals with lung cancer and identified several germ-line and somatic rearrangements. Importantly, they showed that somatic rearrangements involved tandem duplication of genomic segments whereas germ-line rearrangements involved repeat elements (e.g., Alu and L1) [83]. Pleasance et al described a comprehensive catalog of somatic mutations from malignant melanomas and lymphoblastoid cell lines and identified several different types of inter- and intra-chromosomal alterations that involved deletion, inversion and duplication [82]. Stephens et al. used paired-end sequencing and showed that there are more rearrangements, most of which are intra-chromosomal, in breast cancer samples than previously appreciated [84]. They observed a significant preference for tandem duplications and rearrangements to be mediated by non-homologous end joining [84]. Maher and co-workers investigated chronic myelogenous leukaemia cell lines [78] and prostate tumor cell lines [78,79] and reported the discovery of several novel translocations and gene fusions with altered expression levels. Taken together, these studies reveal that alterations in the genomic neighbourhood could have consequences for gene expression and such events can often result in developmental disease and cancer.

In a recent exciting study, Navin et al. studied breast carcinomas by developing a new method called Sector-Ploidy-Profiling [85]. This involved macro-dissecting tumors, flow-sorting genomic subpopulations by DNA content, and profiling structural changes in genomes using comparative genomic hybridization. The authors described two classes of genomic structural variation: (1) monogenicomic and (2) polygenomic. Monogenicomic tumors were shown to contain a single major clonal subpopulation with a highly stable chromosome structure. Polygenomic tumors contained multiple clonal tumor subpopulations, which may occupy the same sectors, or separate anatomic locations. This heterogeneity was further ascribed to a few clonal subpopulations, rather than a series of gradual intermediates. By comparing multiple subpopulations from different anatomic locations, the authors inferred pathways of cancer progression and the organization of tumor growth. In another landmark study, Mullighan et al. explored the genetic basis of relapse by performing genome-wide DNA copy number analyses on matched diagnosed and relapsed samples from patients with acute lymphoblastic leukemia (ALL) [86]. They found that the diagnosis and relapse samples typically showed different patterns of genomic copy number abnormalities (CNAs), with the CNAs acquired at relapse preferentially affecting genes in cell cycle regulation and B cell development. Most relapse samples were found to lack some of the CNAs present at diagnosis, suggesting that the cells responsible for
relapse are ancestral to the primary leukemia cells. Backtracking studies by the authors revealed that cells corresponding to the relapse clone were often present as minor subpopulations at diagnosis. These studies together highlight how profiling genomic alterations could provide insights into cancer progression and relapse.

6. Conclusions and outlook

Although it is clear that the eukaryotic genome is highly organised, it is only recently that genomic neighbourhoods or “domains” are being defined objectively on a genomic scale. In addition, several recent studies have provided insights into how genomic neighbourhood could influence the expression level of a gene, with important consequences for evolution, development and disease. Future experiments aimed at systematically altering genomic neighbourhoods and investigating alteration in gene expression will likely provide a better understanding of the real impact of this phenomenon. In this direction, several experimental tools that permit alteration of genomic neighbourhoods are already available [87-89] to perform such studies. Indeed, Kobuku et al. have made a first attempt to survey the regulatory landscape in mice using such a transposon-based chromosomal engineering method [90].

While experimental advances in sequencing is providing an avalanche of information about various aspects of genome organisation and gene expression, one of the fundamental challenges for the future will be in developing the conceptual framework and computational methods to tackle and integrate this data from different model organisms, cell types and samples from healthy and disease states. Even for defining genomic neighbourhoods or domains robustly, only a few computer programs currently exist. Some recent computer programs that permits identification of these domains from genome and ChIP-seq data include SICER: Statistical approach for the Identification of ChIP-Enriched Regions [91], BLOC: Broad LOCal enrichments [18], ChipDiff [92], ChromaSig [93] and Synorth [94]. For a practical review of computational tools to identify conserved regulatory regions in mammals, please see Fredman et al. [95]. Another major challenge, given the enormous data on cancer genome sequencing that is being currently generated (see COSMIC [96]), would be to develop robust computational approaches that would differentiate between driver (causal) and passenger (bystander) alterations in the genome.

This is definitely an exciting time for experimental and computational biologists who aim to understand the role of genomic neighbourhood in the regulation of gene expression. An important direction for future research will be to develop concepts that integrate spatial information of genomic segments from large-scale Hi-C experiments [14] (e.g., to model them as a network) with gene expression data to better understand the spatial network of interactions between distinct genomic regions and their influence on gene regulation. Given that genomic neighbourhood can influence gene expression level between individual members in a clonal population of cells [31] and minor changes in expression levels of key regulators can have important consequences during cellular differentiation and development [97-99]; an improved understanding of this phenomenon will have important implications in gene therapy, regenerative medicine and genetic engineering as it would permit rational identification of genomic loci to incorporate reporter genes during the production of transgenic cells. Finally, while exciting advances in experimental and computational approaches are being made to understand the influence of genomic neighbourhood on gene regulation, we believe that a combined and iterative approach involving both experimental and computational work will rapidly allow for a better understanding of this phenomenon and its role in evolution, development and disease.

Acknowledgements

SD and MMB acknowledge the MRC Laboratory of Molecular Biology for support. MMB thanks Darwin College, Cambridge and SD thanks King’s College, Cambridge for generous support. The authors thank A Klug, AJ Venkatakrishnan, B Lang, B Lenhard, D Hebenstreit, D Kleinjan, G Chalancon, J Su, P Yeyati, S Janga, S Teichmann and V van Heyningen for reading this manuscript.
**Figures and figure legend**

**Figure 1:** In addition to cis-elements and transcription factors, genomic neighbourhood influences the expression level of a gene. The green circle and rectangle denotes the gene of interest; the wavy orange line denotes a chromosome. Two genomic neighbourhoods are shown schematically. The region containing blue circles denotes genes in a heterochromatinic domain and the region containing red circles denotes genes in a euchromatinic domain.

**Figure 2:** Consequences and implications of altering the genomic neighbourhood of genes for evolution, development, disease and practical applications.
Box 1

We performed an analysis to estimate the combined effect of promoter divergence and alteration in genomic neighbourhood on the expression divergence (i.e., difference in the expression level) of orthologous genes between human and chimpanzee.

We collected data on Kp, a measure of promoter sequence divergence, and gene expression divergence for orthologous genes between human and chimpanzee from Khaitovich et al. [100]. In this study, the authors compared mRNA expression levels of orthologous genes from six humans and five chimpanzees using probes that are identical between the two organisms from five different tissues, i.e., heart, lung, kidney, liver, brain, and testis. We also obtained the CGN score, a measure of alteration of genomic neighbourhood, for one-to-one orthologous genes between human and chimpanzee from De et al. [46]. In short, the CGN score of a given gene in human is simply the fraction of genes within a window (of size w) surrounding it, which are orthologous and are also present in an equivalent window around the chimpanzee ortholog [46]. For 7, 241 genes, we had Kp, CGN score and expression divergence data for at least one organ. We then grouped the genes in a 4 by 4 table according to the quartile distribution of their CGN scores and Kp values so that we have roughly comparable number of genes (n~ 450 genes) in each bin. For each combination of Kp and CGN score interval, we calculated median expression divergence value for all genes in that bin.

For each tissue, we observed that gene expression divergence between human and chimpanzee generally increases with (a) an increase in promoter divergence and also (b) an increase in the extent of alteration in their genomic neighbourhood (see Figure 3). Consistent with previous findings, these observations suggest that gene expression divergence may be driven by rapid promoter evolution alone or due to alterations in genomic neighbourhood alone [46,100]. In particular, we observed that the genes that have both altered neighbourhood and high promoter divergence displayed higher expression divergence than expected by chance (see Figure 3 below). These findings suggest that a combination of alteration in genomic neighbourhood and promoter divergence leads to high gene expression divergence. We find a similar trend if we only consider the maximal expression divergence value for a gene across all tissues (see Figure 3 below). Further, the trends reported above are statistically significant (Chi square test) for all the organs examined except testis. Our observations that a combination of alteration in genomic neighbourhood and increased promoter divergence is linked with high expression divergence, suggest a complex trajectory for transcriptome evolution.

Figure 3: Heatmap representation of the median expression divergence value for different sets of genes. The scale-bar of the log of expression divergence is shown above each heatmap (red and blue: high and low expression divergence, respectively). The promoter divergence value is measured by the Kp value [100] and the genomic neighbourhood is measured by the CGN score [46].
References and recommended reading
Papers of particular interest have been highlighted as:
* of special interest
** of outstanding interest


