Computational Reconstruction of Enhancer-Gene Regulatory Networks Altered in Cancer

Thesis

COMPUTATIONAL RECONSTRUCTION OF ENHANCER-GENE REGULATORY NETWORKS ALTERED IN CANCER

JUDITH MARY HARIPRAKASH

G4581026

Supervisors:

Francesco Ferrari

Florian Markowetz

Vincenzo Costanzo

The Open University PhD Program

Fundamentals of Cancer Biology

IFOM The FIRC Institute of Molecular Oncology

1. ABSTRACT

Enhancers are *cis*-acting non-coding regulatory elements that regulate the transcriptional output of target genes. Their dysregulation has been associated with various diseases including cancer. However, identification and characterisation of non-coding mutations that are relevant for tumorigenesis and prognosis remains a major challenge. We hypothesised that non-coding mutations in enhancers could significantly influence cancer prognosis and patient survival and thus can be exploited as novel prognostic biomarkers for better patient stratification and targeted therapy in lung cancer. Here we present the detection and characterisation of enhancer mutations in a genome-wide analysis of 159 lung cancer samples. To define enhancers across the genome we leverage the epigenomic signatures incorporating histone marks (H3K27ac) and chromatin accessibility (DNase I sensitivity or ATAC-seq) from 8 lung cell lines and primary tissue. We observe that the mutation burden at enhancers, promoters and exons is similar, whereas the mutation signature at these genomic locations varies significantly. We observe recurrent mutations in enhancers at base pair level and show their impact on target genes. We also demonstrate that genes have more than one enhancer and when they are mutated, the gene expression is altered. We also observe pathway-level aggregated enhancer mutations in cancer patients. These results contribute to a new approach towards the functional validation of non-coding mutations in cancer.

ACKNOWLEDGEMENT

My first and most sincere thanks to Francesco for having been a wonderful guide. You have never failed to appreciate every small achievement. Your guidance, words of encouragement, and advice kept me motivated all through the PhD. I want to take this opportunity to express my gratitude for everything.

I would like to thank my external supervisor Florian for his suggestions and input. You have been extremely kind to host me in your lab, have always been available for discussing my work. I would also like to thank Vincenzo Costanzo, my internal supervisor for his support. I am also thankful to my third part monitor Angela Bachi for her wise counsel during my PhD.

My sincere thanks to all my lab mates for a lively and interactive environment to work in. I would like to specially thank Elisa for all the interesting discussions on the work, and for making me understand the gravity of statistics. Thanks to Cristiano, Giulia, Giovanni, Ludovico and Koustav for all the insightful and invigorating discussions. I extend my thanks to Ilario, Endre, Raquel and Federica for your support in the project.

I would like to thank Federica La Mastra for her tremendous support in the experimental work and for teaching me to do ATAC-seq. I would like to thank Stefano Casola for all his inputs and suggestions for my project. Though it started as a collaboration, without the support from you and your lab, I could not have done the experiments. I would like to profusely thank you all for that. From helping me find reagents to troubleshooting errors, you have been extremely accommodating. Special thanks to Federica Mainoldi and Rachele Niccolai.

I would like to sincerely thank Chiara Lanzuolo and Francesca Gorini for sharing their expertise and helping me with the ChIP-seq experiments.

I would like to thank Luca Roz for his suggestions on the cell line of choice and his timely help in providing us with the HBEC cells for experiments.

I am obliged to everyone at IFOM cell culture facility especially Ilaria, Giuseppe, Cinzia and Stefania. From teaching me all the tips and tricks, to answering all my doubts, to helping

me through my bouts of contamination, and all the while encouraging me, you have been a great support.

I would like to thank Simone, Mirko and Claudia at the IFOM Genomics facility for your timely service. You have always been very supportive of all my requests, and questions, and have taken the time to explain things and help me with a lot of technical difficulties. I would like to thank Mario Cinquanto for your help in the genome editing work.

I would like to thank Francesco Iorio and Martin Schaefer for their feedbacks and suggestions on the project.

Words are not enough to thank Marina and Mio for going above and beyond to make me feel at home in Italy. Since the day I got selected, you have been with me through all the steps. Thank you for your tremendous help.

I would like to thank everyone at the secretary's office, grant's office, personal office, and procurement office. You have been extremely welcoming and swift in aiding us all.

Sincere thanks to my buddies Guru, Adhil, Ram and Yathish for all the care and love. I would have been a mess without you all.

My heartfelt thanks to all my friends who inspired and motivated me to pursue research, especially Vaishu for instilling this desire in me.

I would like to thank Rose amma, Patti, my parents and family for believing in me and enabling me to dream big. I would like to thank my sister and all my cousins for always encouraging me. Special thanks to my niece Elsha for her unconditional love.

Last but not the least, I'm grateful to have Parashar, my pillar of strength. You have inspired me to be a better researcher. Thank you for always seeing the best in me, and all the while making me even better.

TABLE OF CONTENTS

FIGURE INDEX

PUBLICATIONS BY THE CANDIDATE

Hariprakash JM, Ferrari F. Computational Biology Solutions to Identify Enhancers-target Gene Pairs. Comput Struct Biotechnol J. 2019 Jun 14; 17:821-831. doi: 10.1016/j.csbj.2019.06.012. PMID: 31316726; PMCID: PMC6611831.

Salviato E, Djordjilović V, Hariprakash JM, Tagliaferri I, Pal K, Ferrari F. Leveraging three-dimensional chromatin architecture for effective reconstruction of enhancer-target gene regulatory interactions. Nucleic Acids Res. 2021 Sep 27;49(17): e97. doi: 10.1093/nar/gkab547. PMID: 34197622; PMCID: PMC8464068.

(Note: This thesis contains figures and contents from the above published papers)

ABBREVIATIONS

Table 1: Abbreviations

3D	Three dimensional
AdapT	Adaptive P-value thresholding procedure
APOBEC	apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like
ATAC	Assay for Transposase-Accessible Chromatin
ATCC	American Type Culture Collection
BAM	Binary Alignment Map
BWA	Burrows Wheel Algorithm
CAGE	Cap Analysis Gene Expression
ChIP	Chromatin Immuno Precipitation
COSMIC	Catalogue Of Somatic Mutations In Cancer
CT	Computed Tomography
DFI	Disease Free Interval
DHS	DNase Hypersensitivity Site
DNA	Deoxyribonucleic acid
EGA	European Genome Archive
ENCODE	Encyclopaedia of DNA Elements
eQTL	expression Quantitative Trait Loci
eRNAs	enhancer-templated RNAs
ETG	Enhancer Target Gene
FANTOM	Functional Annotation of the Mammalian genome
FBS	Foetal Bovine Serum
FDA	Food and Drug Administration
FDR	False Discovery Rate
FIMO	Find Individual Motif Occurrences
GATK	Genome Analysis Toolkit
GO	Gene Ontology
GRO-seq	Global run-on sequencing
GSEA	Gene Set Enrichment Analysis
GTEX	Genotype Tissue Expression
HACER	Human active enhancer
IGV	Integrative Genomics Viewer

INDEL	Insertion and Deletions
KEGG	Kyoto Encyclopaedia of Genes and Genomes
LARVA	Large-scale Analysis of Recurrent Variants in noncoding Annotations
LUAD	Lung Adenocarcinoma
LUSC	Lung Squamous Cell Carcinoma
MEME	Multiple Em for Motif Elicitation
MRI	Magnetic Resonance Imaging
MSigDB	Molecular Signatures Database
NC	Non-coding
NET-CAGE	Native Elongating Transcript–Cap Analysis of Gene Expression
NET-seq	Native Elongating Transcript sequencing
NGS	Next Generation Sequencing
NSCLC	Non-Small Cell Lung Cancer
PET	Positron Emission Tomography
POLE	DNA Polymerase Epsilon
PRO-seq	Precision Run-On Sequencing
qRT-PCR	Quantitative Real Time- Polymerase Chain Reaction
RNA	Ribonucleic acid
RNAPII	RNA polymerase II
SBS	Single Base Substitution
SCLC	Small Cell Lung Cancer
SNP	Single Nucleotide Polymorphism
SNV	Single Nucleotide Variation
TAD	Topologically Associating Domains
TCGA	The Cancer Genome Atlas
TFBS	Transcription Factor Binding Site
TNM	Tumour-Nodes-Metastasis Classification of Malignant Tumours
TPM	Transcripts per Million
TRANSFAC	TRANScription FACtor database
TSS	Transcription start site
UTR	Untranslated region
UV	Ultraviolet
WGS	Whole genome sequencing
WT	Wild type

2. INTRODUCTION

Cancer is a disease of the genome. Over the years, enormous research has been done towards understanding the genomics of cancer, and achieved major milestones in identifying the genetic roots of various cancers. These discoveries have also been translated towards the betterment of the livelihoods of the patients. In spite of these bench to bedside transitions, we still face delayed diagnosis, poor prognosis and low survival rates. This indicates a need for alternative or additional approaches for an improved understanding of cancer.

When we look closely at the research done so far, the predominant focus of the field of cancer genetics has been towards the coding genome[1–3], which contributes to only 2% of the total human genome. In the past two decades, the focus has also included the vast and unexplored non-coding regions of the genome[4–7]. As the tumorigenic effect of a non-coding mutation is likely affected by *cis* change in gene expression, *cis*-acting regulatory regions such as promoters, and enhancers need to be studied in association with the coding genes to get a better perspective[8]. However, elucidating the function of a non-coding region is still a major challenge.

This thesis will elucidate a strategy to characterise non-coding regulatory mutations in cancer. The following section will give an overview of the non-coding regulatory regions, how they function and the mechanisms that alter them.

2.1. CHROMATIN THREE - DIMENSIONAL ARCHITECTURE

The organization of chromatin within the nucleus is not only an efficient way to package the enormous amount of information within the genome but is also a key mechanism in gene regulation[9]. It is achieved through a multi-layered, hierarchical structural arrangement **(Figure 1**). The genetic information of eukaryotic cells is stored in extremely long DNA molecules, which are then packed in to nucleosomes[10]. Nucleosomes are formed by approximately two turns of DNA (146 bp) wrapped around a histone octamer; this association is enabled by the electrostatic charges of the corresponding molecules. Nucleosomes are the basic structural unit of DNA packaging in eukaryotes[11]. The packaging of DNA into nucleosomes shortens the fibre length about sevenfold[10]. Beyond the nucleosomes, at the next scale of organisation is the nucleosome-nucleosome interactions forming the 30nm chromatin fibres[12,13].

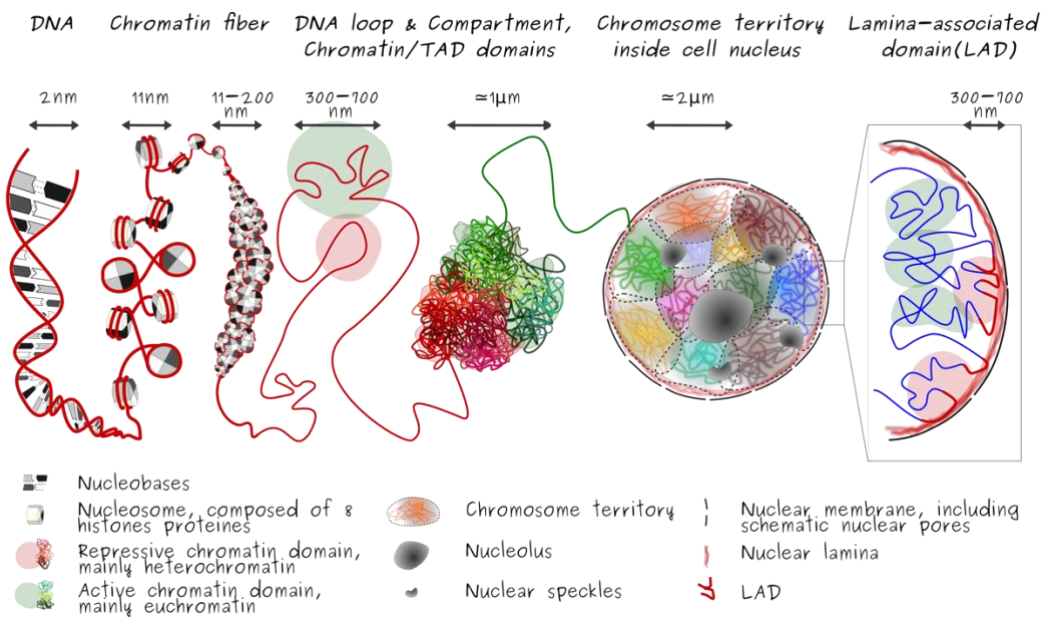


Figure 1: Hierarchical organisation of the eukaryotic genome. *Schematic representation of DNA folding to chromatin fibre and its 3D organization and architecture inside the nucleus of a eukaryotic cell. Adapted from https://abrunet.com/*

These nucleosome fibres are further organized in loops, spanning large distances along the genome[14]. Chromatin loops can be either transient or are stabilized by specific proteins such as structural maintenance of chromosomes (SMC) protein, cohesin and the CCCTC-binding factor (CTCF)[15–21]. The dynamics of loop extrusion imposes functional organization enabling relatively distant cis-regulatory elements to interact with their target genes by bringing them in close spatial proximity[22–25]. Chromatin loops are characteristic of fine scale interactions, whereas at the global scale eukaryotic genomes are organized into sub-megabase scale domains often called topologically associating domains (TADs)[26,27]. Regions within the same TAD interact with each other much more frequently than with regions located in adjacent domains. TADs are thought to be conserved between different cell types and across species[28–30] . TADs are also known to exhibit a nested structed in mammals, wherein large TADs can be further subdivided in to smaller domains called subTADs[31–33].

At the next level of organisation is the compartmentalisation of megabase-scale chromatin. Long range interactions between TADs can be observed in mammals, that show preferential interaction with each other.[23] Two types of compartments, often called A and B compartments, were initially identified as domains mostly interact with each other[34]. Recent advances in the field have suggested that the two major compartments can be further subdivided into sub-compartments[31,35]. Compartment A associated with active regions can be subdivided further into two sub-compartments and the inactive B compartment into

four[36]. Unlike TADs compartments are not conserved between different cell types[36]. TADs can switch between the compartments in a cell type specific manner[37].

At even higher level, we observe chromatin organized into individual chromosome territories that seldom intermix[26]. Chromosomes occupy their own territory and adopt a radial position. Large chromosomes are often found at the nuclear periphery and smaller ones at the interior[38]. This arrangement is also modulated by other cellular organelles such as nucleoli, splicing speckles, and the nuclear envelope, by acting as tethering points for chromatin[39].

2.1.1. CHROMOSOME CONFORMATION CAPTURE TECHNOLOGIES

Early studies of genomic conformation were largely based on cytological techniques, such as fluorescence in situ hybridization[40] (FISH), which enabled direct evaluation of the proximity between genetic loci using probes. In the recent years, chromosome conformation capture[41] (3C) and 3C-based techniques using high-throughput sequencing data are widely used to understand the spatial topology of the genome[42–45] (**Figure 2**). Hi-C quantifies the frequencies of contacts between distal DNA segments in cell populations to map the genomic architecture[46].

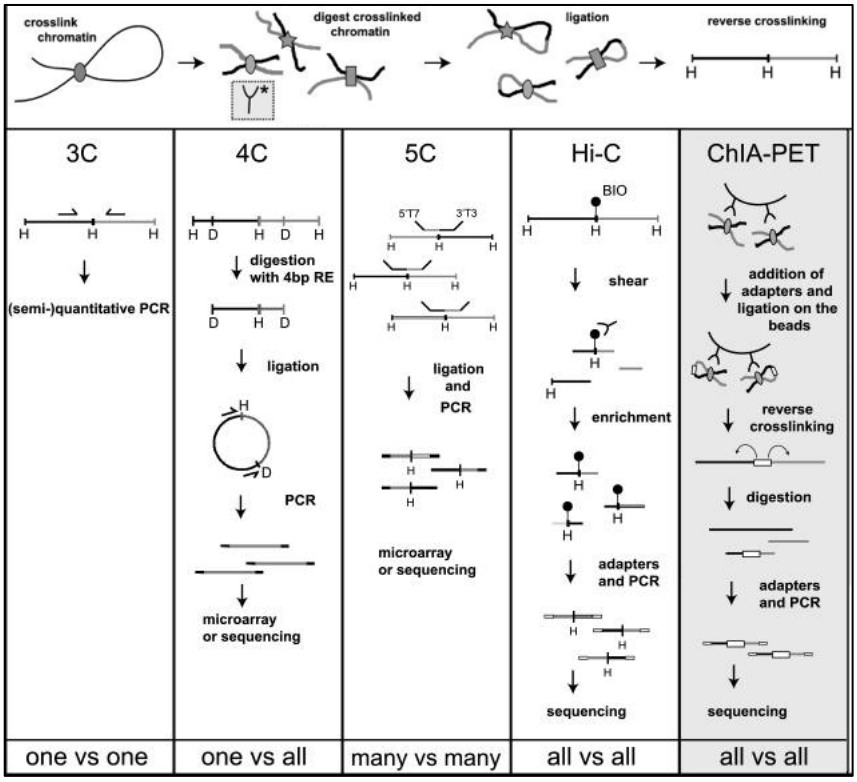


Figure 2: Overview of 3C-derived methods. *Common steps of all the 3c-derived techniques viz., Cross-linking, digestion and ligation are depicted in the horizontal panel. Steps specific to individual methods are depicted in the vertical panel. Modified and adapted from* [47].

The protocol involves the formaldehyde fixation of chromatin, followed by restriction digestion using a 6bp cutters (HindIII, BglII, SacI, BamHI or EcoRI) or more frequent cutters (AciO or DpnII). Digested chromatin is religated under diluted conditions, or in intact nuclei in the "in situ Hi-C" protocol[31]. Finally, the number of ligation events between non neighbouring sites are quantified to determine the DNA contact frequencies[48]. In 3C, a semiquantitative or quantitative PCR amplification of selected ligation junctions is implemented to determine one vs one contact, whereas in circular chromosome conformation capture (4C) microarrays or high-throughput sequencing are used to analyse the contacts of a selected genomic site[49,50]. (one vs all).

Chromosome conformation capture carbon copy (5C) allows concurrent detection of interactions between multiple sequences, thus getting its name many vs many. 5C is implemented using a combination of oligonucleotides with overlapping restriction sites at the locus[51,52]. Oligonucleotides for the interacting fragments are juxtaposed, ligated together, following which they are amplified. Junctions are then quantified either on a microarray or by high throughput sequencing[51].

The high throughput version of the 3C based technology is the Hi-C, an all versus all method incorporating next generation sequencing technology[45]. In Hi-C, before ligation of the 3C template, restriction ends are filled in with biotin labelled nucleotides. Followed by blunt end ligation, DNA purification and shearing, a biotin pull-down is performed to ensure that only ligation junctions are selected and are sequenced[48]. A combination of chromatin immunoprecipitation (ChIP) with 3C-technology offers the possibility to analyse the chromatin interactions exclusive to regions bound by protein of interest through ChIP-loop[53], however this technique is seldom used. Widely used high throughput genome-wide version of ChIP-loop is the Chromatin Interaction Analysis by Paired-End Tag sequencing (ChIA-PET) technology[54].

In-situ Hi-C is a variant of Hi-C wherein DNA-DNA ligation events generated via proximity ligation is performed in intact nuclei[31,55]. Micro-C is another derivative of Hi-C technique which uses micrococcal nuclease (MNase) for genomic digestion rather than the restriction enzymes[56]. Micro-C is efficient in detecting short-range interactions as opposed to long range interactions effectively captured by Hi-C[57,58]. A low cost and high-resolution alternative to obtain interaction data for a set of regions is the Capture Hi-C[59–62]. It is a combination of 3C or Hi-C with a capture enrichment step. Promoter capture Hi-C is enriched for interactions around a set of promoters[59,63,64].

2.1.2. *Hi-C dataset processing*

Hi-C data analysis involves multiple steps that can be classified in to pre-processing and downstream analysis. At the pre-process stage, FASTQ files of paired-end sequencing reads are aligned to the reference genome. As the Hi-C sequencing reads are expected to align to different unrelated genomic regions, the reads are aligned separately. Although, standard alignment tools such as bowtie[65] or BWA[66] can be employed to map Hi-C paired end reads, the complexity of the chimeric reads arising from ligation junctions demands more curated approaches[67]. The current pipelines implement varying strategies of iterative mapping to align chimeric reads to the genome[68–71].

Following which, spurious reads arising from experimental artifacts are removed. The next step is to count reads. Although, the reads are mapped on individual restriction fragment ends, read counts are summarized at the level of genomic bins. The choice of the bin size determines the final resolution of the analysis results. The final pre-processing step is normalisation to obtain contact matrix as final output, which is often performed simultaneously with read counts binning. Normalization can be performed either using explicit or implicit normalization methods. Explicit methods account for biases arising from fragment length, GC content and mappability[72]. Whereas implicit methods or matrix-balancing normalization methods relies on the assumption of equal visibility in all genomic regions[31,68,70,73,74]. **Figure 3** summarises the Hi-C data pre-processing steps.

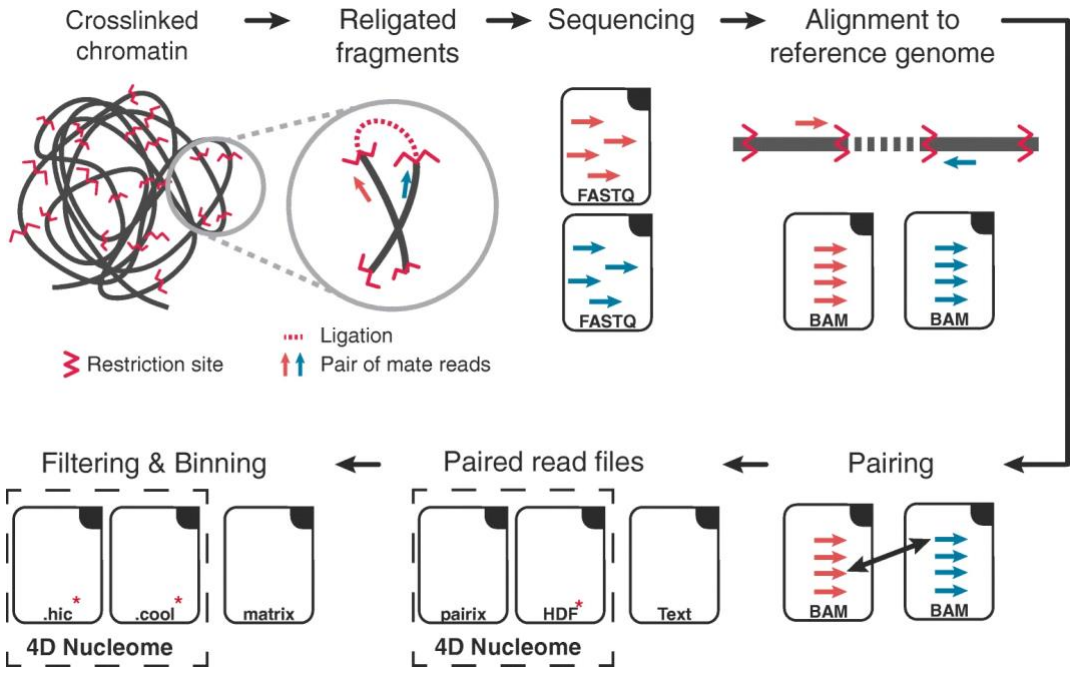


Figure 3: Hi-C data, from generation to contact matrix. *Schematic representation of Hi-C data analysis. Adapted from* [67]

Downstream analysis includes methods that extract meaningful 3D genome information at various levels of resolutions in the form of (a) genomic compartments (b) TADs (c) point interactions through loop calling[75]. Compartments are often observed in the Hi-C map as a plaid pattern by implementing Pearson correlation of the distance normalized map[45]. The sign of the first principal component is used to define the active A and inactive B compartments.

TADs are visible along the diagonal of the contact matrix as blocks of self-interacting regions[28,76]. One-dimensional scores such as directionality index and insulation scores were the first methods proposed for TAD calling. Directionality index (DI), a signed chi-squared statistic, is calculated by binning the genome into equal size bins and computing the degree of upstream and downstream biases. A positive value indicates that more read pairs lie downstream than upstream, and a negative value indicates the reverse. Whereas, the insulation score quantifies the interactions in each genomic bin and uses the local minima to define the boundaries. Other proposed methods use clustering algorithms to identify the best partitioning of the contact matrix in TADs[71,77,78]. Increased resolution of the Hi-C data highlights the existence of hierarchical structure of TADs and newer methods to call for multiscale TADs are now proposed.

Point interactions are the points of contact between distant chromatin regions. For the interaction identification, a background is first estimated using either local distribution or modelled at global chromosome-wide or genome-wide scales, contacts with higher frequency than the expected are discerned from this background[79–81]. In principle, Hi-C data can be used to identify point interactions such as enhancer promoter loops[82]. However, these interactions are analysed by binning the read counts at a resolution of few kilobases, with the maximum resolution of 1kb with the high coverage datasets[31,67,83]. Thus, other methods to identify enhancer target gene pairing approaches are required.

2.2. ENHANCERS – THE DISTAL NON-CODING REGULATOR

Enhancers are non-coding regulatory regions that are distant from their cognate promoter along the linear sequence of DNA[84,85]. They play a crucial role in regulating gene expression, and often in their absence transcription of its target gene is weak[86]. Enhancers are analogous to promoters from various points of view such as their chromatin accessibility, involvement in transcriptional activity, transcription factor binding. However the relative location of an enhancer, with respect to its target gene can be greatly variable[87]. An enhancer can be present

in the vicinity of its target gene, but need not necessarily regulate the closest one[24,88]. They do not have univocal sequence motif for their genome-wide identification. Enhancers contribute additively and partly redundantly to their target gene's expression[89,90].

Enhancers are extremely cell type-specific, and are considered as the genomic feature that is most variable across tissues and cell types in terms of their activation[91]. Even though a specific gene may be active in multiple cell types, its activation can be regulated by different enhancers in different tissues. For instance, the first discovered mammalian enhancer - the immunoglobulin heavy chain (IgH) associated enhancer is present downstream of the IgH gene. It is known to exhibit enhancer activity only in lymphocyte-derived cell lines and during B lymphocyte differentiation[92]. Such cell type-specific enhancer activity is mainly a result of variable accessibility of an enhancer region in different cell types; on the other hand, promoters generally are open and nucleosome-free chromatin regardless of the cell type. Thus, the chromatin landscape forms a critical barrier for cell type-specific gene regulation by modulating enhancer activity[93].

2.2.1. ENHANCER FUNCTIONING

Proper functioning of enhancers is dependent on the accessibility of local chromatin. Open chromatin leads to the exposure of short DNA motifs contained within the enhancer[94]. These motifs are sequences that are recognized by transcription factors (TFs) and facilitate their binding to the enhancer. Following such binding, other mediator proteins are recruited[89,95–97]. The combined regulatory signal of all bound proteins primarily determines the activity of the enhancer. The interaction between enhancers and promoters is facilitated by the higher order chromatin structure leading to physical proximity of the interacting pairs[94,98]. **Figure 4** illustrates the enhancer-promoter chromosomal looping showing how a distal enhancer element physically interacts with and activates a gene promoter.

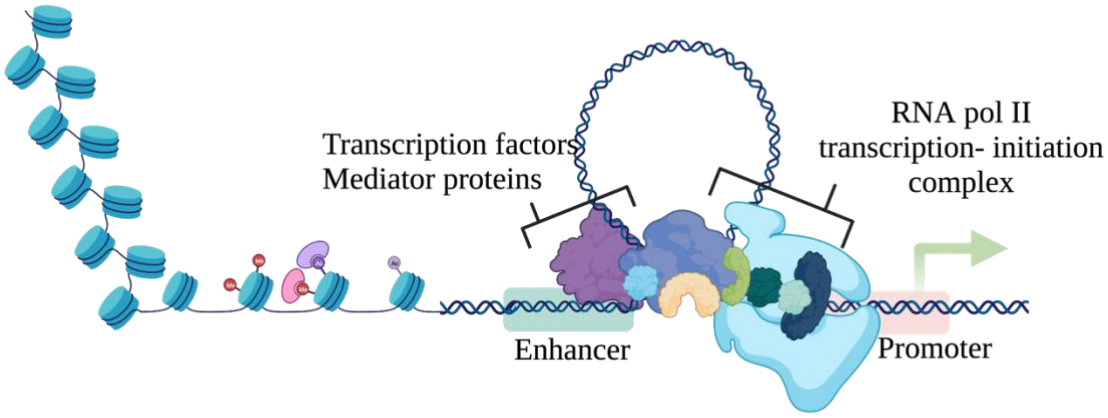


Figure 4: Regulation of gene expression patterns by genomic enhancers. *Illustration of enhancer-promoter chromosomal looping that allows distal enhancer elements to physically interacts with and activates gene promoters. (Figure generated using Biorender).*

Gene expression through mRNA transcription is a tightly regulated process carried out in three phases *viz*., i) initiation ii) elongation and iii) termination. In the initiation phase RNA polymerase II (RNAPII) recognizes and binds to the gene promoter, and thereafter RNA synthesis occurs in the elongation phase[99,100]. Following a productive elongation, RNA pol II is released in the termination phase[101,102]. In higher eukaryotes, the polymerase is paused at promoter-proximal regions before active elongation in a signal-integration step[103,104]. Enhancers play important role in all three phases of transcription[86,105,106]. They help recruit RNAPII to promoters and are known to recruit pioneer factors and lineage-specific transcription factors[106]. Additionally, enhancers are involved in promoter-proximal pause-release and transcription elongation[94,107]. Distal enhancers are known to form contacts with their cognate promoters during the entire interval of elongation[108].

2.2.2. GENOME-WIDE DEFINITION OF ENHANCERS

The lack of a sequence grammar for the genome-wide identification of enhancers based on sequence information can be addressed by leveraging the epigenomic features of enhancers. Chromatin Immunoprecipitation followed by high-throughput sequencing (ChIP-seq) targeting specific histone marks associated with enhancers are typically adopted in enhancer definitions. Active enhancers, are acetylated at lysine 27 of histone H3 (H3K27ac) and are enriched in histone H3 lysine 4 mono methylation (H3K4me1)[109–111]. H3K4me1 along with H3K4me3 is usually found also at promoter regions, but the relative enrichment of the two marks is expected to be different at enhancers and promoters[112]. The presence of high levels of H3K4me1 along with H3K27ac are characteristic of enhancer regions[109–111]. However, H3K4me1 are observed in enhancers that are not active in a specific cell type annotated as

poised enhancers[113]. Hence, H3K27ac is usually the preferred chromatin mark for the genome-wide identification of active enhancers[114].

Additionally, active enhancers are characterised by chromatin with high accessibility[115] and thus, can be distinguished by genomics methods that probe chromatin accessibility such as DNase-seq and Assay for Transposase-Accessible Chromatin followed by high-throughput sequencing (ATAC-seq). DNase-seq is based on the partial digestion with DNA nucleases like DNase I that frequently cut the regions with higher chromatin accessibility[116]. Such regions are known as DNase hypersensitivity sites (DHS)[116]. ATAC-seq leverages differential sensitivity of open and close chromatin regions to transposase activity[117].

Active enhancers are known to produce enhancer-templated RNAs (eRNAs) that are short, unstable, unspliced, non-polyadenylated, and noncoding RNAs expressed at low abundance levels[118,119]. Due to the lack of polyadenylation, they are subject to exosome-mediated degradation and hence are unstable in nature[106]. Hence using eRNA as a marker for defining enhancers remains a difficult challenge. In recent years a number of sequencing protocols have been developed to detect such nascent transcripts, these include CAGE-seq[120–122], GRO-seq[123], NET-seq[124], NET-CAGE[125] among others. Nevertheless, the use of eRNA based enhancer definition is not widely used.

2.2.3. COMPENDIUM OF ENHANCERS

The plummeting cost of sequencing technologies and advancement in epigenomic profiling technologies, along with the knowledge on the chromatin features associated with enhancers, has fuelled a large number of efforts to identify enhancers across genome in various cell and tissue types[111,126–128]

The Encyclopaedia of DNA Elements (ENCODE)[129] provides a registry of candidate *cis*-regulatory elements by integrating high-quality DNase-seq and H3K4me3, H3K27ac, and CTCF ChIP-seq data produced by the ENCODE and Roadmap Epigenomics Consortia[130]. The current version (version 2) comprises 926,535 human *cis*-regulatory elements. The atlas of active enhancers provided by the FANTOM5[131] project is based on 808 human Cap Analysis of Gene Expression (CAGE) experiments. VISTA Enhancer database[132] consists of a collection of *in-vivo* validated enhancers based on transgenic mice reporter assays in 23 tissues of mouse embryos. EnhancerAtlas[131] is a comprehensive database housing 13,494,603 enhancers based on 16,055 genome-wide profiling datasets covering 586

tissue/cell types across nine species. GeneHancer[133] has a collection of more than one million regulatory elements obtained from seven genome-wide databases such as ENCONDE, FANTOM5, VISTA and dbSUPER. Similarly, HACER[134], an atlas of Human ACtive Enhancer to interpret Regulatory variants, catalogues and annotates 1,676,284 enhancers from 265 human cell lines by integrating FANTOM5 CAGE profiles and reprocessing publicly available Global run-on sequencing (GRO-seq) and Precision Run-On Sequencing (PRO-Seq) data. Some of the other enhacer databases that are currently available are SEdb[135], RAEdb[136], HEDD[137], DENdb[138] and dbSUPER[139].

Despite efforts by large epigenomic consortia such as ENCODE and FANTOM for enhancer identification, the dynamic and cell type-specific nature of enhancer activity results in an inability to create an exhaustive reference list of enhancers. The enhancer databases such as SEdb and dbSUPER are more catered to super-enhancers rather than typical enhancers, while others are more focused on disease-related enhancers such as HEDD. Furthermore, the limiting factor for the exploitation of enhancer databases is the availability of fewer number of annotated enhancers for specific cell types.

2.2.4. ENHANCER-TARGET GENE PAIRING

It is known that more than one enhancer can regulate the same gene and also an enhancer can regulate more than one gene. Also, the location of an enhancer relative to its target gene is highly variable. This many-to-many arrangement of enhancer-target gene association and the variability in distance makes the relation between enhancers and their cognate promoter very complex. Thus, making the enhancer-target gene (ETG) pairing elusive.

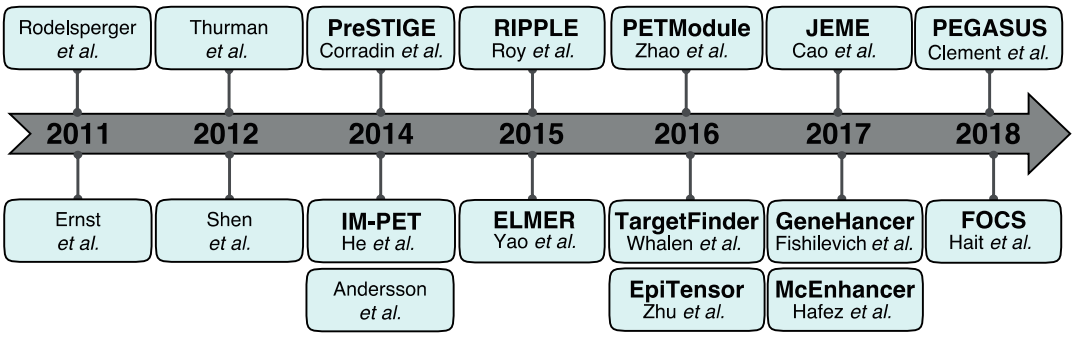


Figure 5: Timeline of the enhancer-target gene pairing algorithms. *The main methods described in our previous review (tool name in bold, if defined) are listed to highlight the timeline of their publication over the years (horizontal axis) (Published in Hariprakash and Ferrari, Comput Struct Biotechnol J. 2019).*

A number of algorithms and computational tools have been developed for the pairing of enhancers and their target genes. We have thoroughly discussed such methods in our

previously published review[140] (**Figure 5**). Owing to the inherent complexity of the problem an integrative, consensus-based approach is generally implemented in most of these methods. Multiple genomic features such as chromatin accessibility, epigenomic features, gene ontology, sequence information, methylation, genomic distance and expression information have been used either individually or in combination (**Figure 6**).

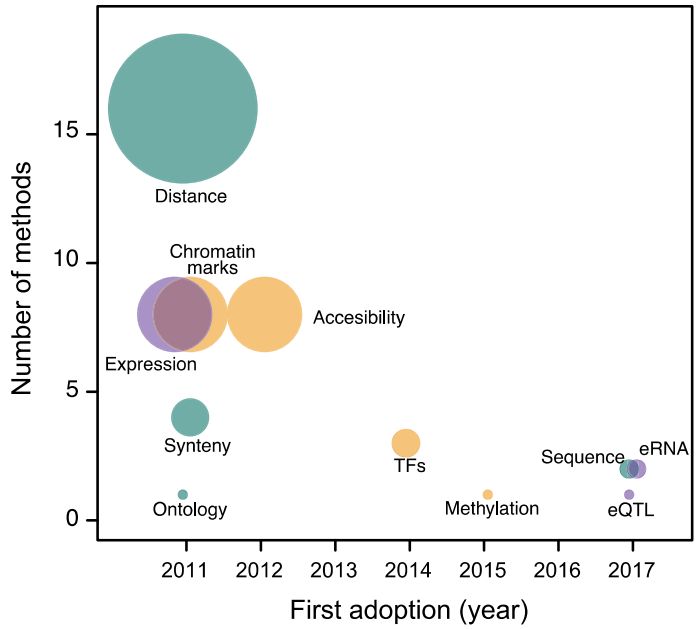


Figure 6: Features used in ETG pairing tools. *The figure summarizes the main types of features used to define ETG pairs by the tools discussed in our previous review. For each feature, its respective frequency (y-axis, number of methods) and first adoption by the tools discussed in this review (x-axis, year) is reported. The size of each dot is also proportional to the frequency (number of methods). The colours represent the category of the data: genomic annotations independent to cell type (dark green); epigenomics data (orange); transcriptomic data (mauve). (Published in Hariprakash and Ferrari, Comput Struct Biotechnol J. 2019).*

We proposed to classify the ETG pairing approaches into four major categories namely 1) Correlation-based 2) Supervised learning-based 3) Regression-based and 4) Score-based methods, centred on the methodology adopted (**Table 2** and **Figure 7**).

Correlation-based methods (**Figure 7a**) rely on the rationale that the activity of enhancers and their cognate genes will correlate across multiple cell and tissue types. Thus, these algorithms use a large panel of epigenomic or transcriptomic data across multiple conditions to correlate the ETG pairs. Correlation-based methods can identify multiple targets of an enhancer and can also be implemented to measure the correlation of potential ETG pairs within short genomic distances thus achieving high spatial resolution. Examples of algorithms that implement this approach include Shen *et al*[141] in mouse cell types, Thurman *et al*[142] in 79 cell types and ELMER[143,144] in cancer samples. These methods are limited by

the availability of genomics data over a large panel of cells with comparable quality and resolution.

Supervised learning-based methods (**Figure 7b**) build a predictor that is based on a known set of true positive and true negative ETG pairs, which can then be applied to call ETG pairs in other independent cell types. However, the caveat in this approach is the lack of a universal set of true and false ETG pairs. Tools such as IM-PET[145], McEnhnacer[146], PETModule[147] and TargetFinder[148] use this approach.

Table 2: Enhancer - Target Gene pairing methods*. The table lists the various ETG algorithms. Their grouping into four main classes is specified: correlation-based (C), prediction-based (P), regression-based (R), score-based (S). Methods with mixed features are specified (e.g., P+R or C+R). C* is for a method conceptually related to correlation-based solutions. Details on each method and features adopted for ETG pairing are also listed.*

Name	Class	Method details	Features
Correlation-based methods			
Thurman *et al*[142]	C	Pearson correlation	DNase-seq
Shen *et al*	C	Spearman correlation	ChIP-seq for Pol2 and H3K4me1
PreSTIGE[149]	C*	Shannon entropy to select cell type-specific patterns	RNA-seq, ChIP-seq for H3K4me1
InTAD[150]	C	Pearson, Kendal or Spearman correlation	RNA-seq, ChIP-seq for H3K27ac, TAD regions
Prediction-based methods			
Rodelsperger *et al*[151]	P	Random forest	Distance, conserved synteny, gene ontology, protein-protein interactions
Ernst *et al*[152]	P	Logistic regression	Gene expression (microarrays), ChIP-seq for 3 histone marks
IM-PET[145]	P	Random forest	Distance, conserved synteny, correlation between enhancer (CSI-ANN score on 3 histone marks) and target promoter (RNA-seq) activity, TFs binding (sequence motifs) and target promoter correlation
PETModule[147]	P	Random forest	Distance, conserved synteny, DNase-seq

TargetFinder[148]	P	Ensemble of boosted decision trees	DNase-seq, FAIRE-seq, DNA methylation, RNA-seq, ChIP-seq for 32 histone marks, in addition to TFs and architectural proteins
McEnhancer[146]	P	Third-order interpolated Markov chain model in a semi-supervised learning setup via the expectation maximization algorithm	Sequence motifs
EAGLE[153]	P	ensemble boosting algorithm "AdaBoost"	RNA-seq, TF binding, STARR-seq, CAGE, FAIRE-seq, DNA-seq/ATAC-seq, ChIP-seq for histone marks and p300
Regression-based methods			
Andersson *et al*[131]	C+R	Pearson correlation, then linear models and lasso shrinkage	DNase-seq
RIPPLE[154]	P+R	Random forest and group lasso	DNase-seq, RNA-seq, ChIP-seq for 8 histone marks and 15 TFs.
JEME[155]	R+P	Multiple linear regression and lasso shrinkage	DNase-seq, RNA-seq, ChIP-seq for 3 histone marks
FOCS[156]	R	Ordinary least squares regression	DNase-seq, CAGE-seq
Score-based methods			
EpiTensor[157]	S	Higher-order tensors decomposition	DNase-seq, RNA-seq, ChIP-seq for 16 histone marks
GeneHancer[133]	S	Additive score with custom weights and data transformations for each quantitative	Distance, TFs co-expression, eRNAs, eQTLs, capture Hi-C

PEGASUS[158,159]	S	Score reflecting the evolutionary sequence and synteny conservation	Conserved synteny and sequence conservation

Regression-based methods (**Figure 7c**) work on the rationale that multiple enhancers can act on a single gene and hence use a combinatorial approach. In addition to identifying significant enhancer target gene pairs, regression-based methods also assess the strength of the impact of multiple enhancers on their target. JEME, RIPPLE and FOCS are some of the regression-based methods. In principle, these methods have the ability to determine the relative influence of one or more predictor variables.

Score-based methods (**Figure 7d**) are those that have implemented a custom quantitative score to define the strength of association between enhancers and genes. This enables flexible prioritisation of ETG pairs by adjusting the threshold on the score, and all possible interacting pairs of genes and enhancers can be obtained. The limitations with both regression-based and score-based methods are that they rely on arbitrarily chosen parameters. Tools like GeneHancer, EpiTensor and PEGASUS implement a score to associate target genes to enhancers.

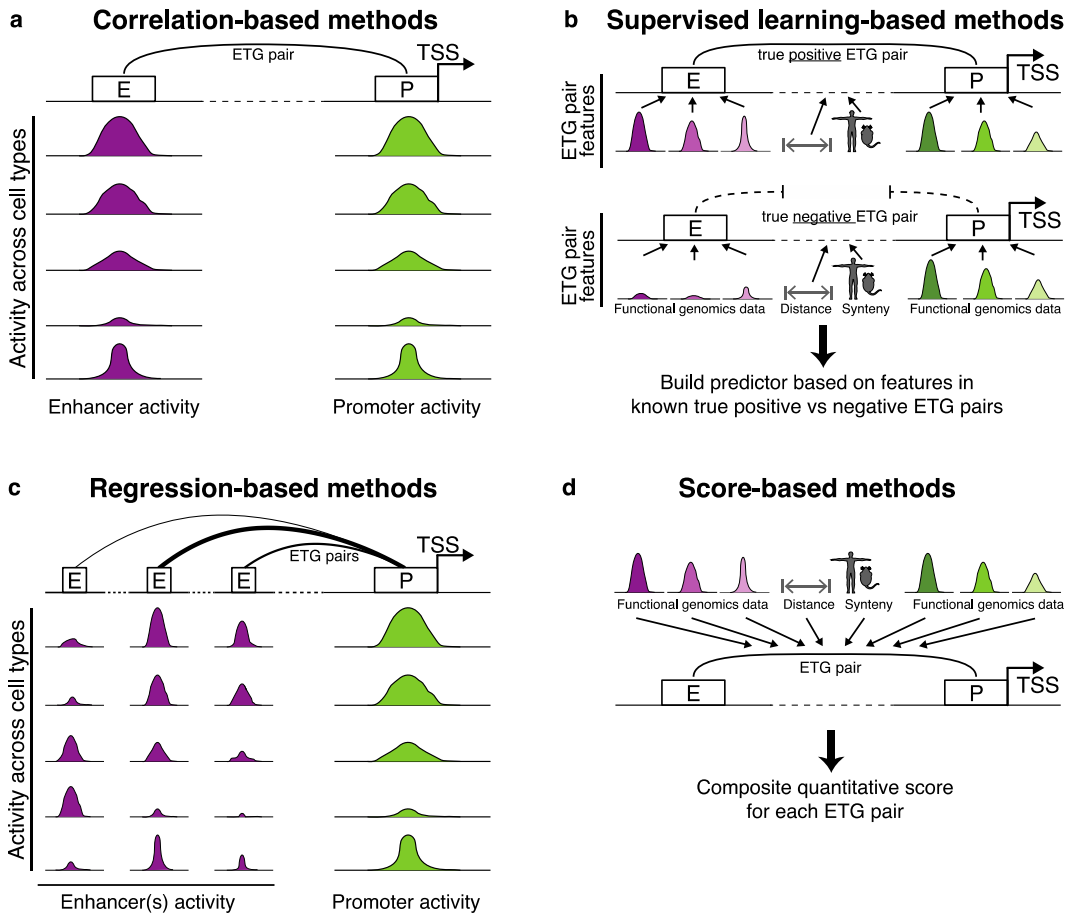


Figure 7: Main classes of ETG pairing methods*. The cartoon highlights the main principles underlying the four main classes of ETG pairing methods (a) Correlation-based methods (b) Supervised learning-based methods (c) Regression-based methods (d) Score-based methods. (Published in Hariprakash and Ferrari, Comput Struct Biotechnol J. 2019).*

2.2.5. ENHANCERS IN DISEASES

Enhancer dysregulation can cause abnormal gene expression and is involved in diseases that can be collectively called enhanceropathies[160–162]. Initial evidence of this phenomenon was observed when DNA translocation caused mis-regulation of the human β-globin gene in β-thalassemia patients[163]. Thenceforth, multiple evidences of disruption of enhancer function through genetic, structural and/or epigenetic mechanisms such as mutations in enhancers, enhancer hijacking, TAD boundary removal and differential methylation have been reported[164].

Genetic alterations in enhancer dysregulation:

Genome-wide association studies (GWAS) have identified that alpha – synuclein (SNCA) as one of the strongest risk alleles associated with sporadic Parkinson's disease[165]. Additionally, studies have also shown that Parkinson's disease associated risk variants leads to increase in SNCA gene expression leading to the development of the disease[166,167]. With

this knowledge, Soldner *et al*[168]., have demonstrated that the mutations in the distal enhancers of SNCA gene modulate the target gene's expression. They identified seven risk variants localized to two distal enhancers (intron-4 and 3'UTR regions), of which the SNP rs356168 in the intron-4 enhancer was a transcription factor binding hotspot. Using CRISPR-CAS9 mediated genome editing, they demonstrate that the alteration of the enhancer sequence at rs356168 with a G allele results in an increased expression of SNCA gene.

Another study that demonstrates the role of enhancer mutations in disease is Michael *et al*[169]., in which the authors have uncovered six different recessive mutations in Pancreas Associated Transcription Factor 1a (PTF1A) enhancer that are associated with isolated pancreatic agenesis. Although the genetic players of pancreatic agenesis are well known[170–172], most cases of isolated, non-syndromic pancreatic agenesis remained unexplained. This study identified a recessive variant in an enhancer ~25kb downstream of PTF1A gene in 7 out of 10 individuals with non-syndromic pancreatic agenesis. Using chromatin conformation capture, they demonstrate that the enhancer region establishes direct interactions with the PTF1A promoter. Additionally, they also demonstrate that the mutations prevent enhancer activity by disrupting the transcription factor binding.

Structural modifications in enhancer dysregulation:

Structural variations (SV) such as insertions or deletions, occur frequently in cancer genomes, and have been shown to play a crucial role in tumorigenesis. In addition to affecting the protein coding regions, SVs can impact through non-coding mechanisms such as altering copy number or position of non-coding regulatory elements or by reshuffling higher order chromatin structures[173]. Enhancer hijacking or enhancer adoption are events resulting from structural variations in which enhancers are juxtaposed to key cancer genes inducing their aberrant expression[174]. For example, in neuroblastoma cell lines, chromosomal translocations leading to juxtaposition of enhancers to MYC gene have been observed with increased MYC gene expression[175]. Similarly, in primary gastric adenocarcinoma, enhancer-based SVs targeting Cyclin E1(CCNE1) gene were identified by Ooi *et al*[176]. In this study, the authors report frequent juxta positioning of diverse distal enhancers to CCNE1 proximal regions in 8% of gastric cancer patients leading to high CCNE1 expression[176]. Contrarily, in group 3 and group 4 medulloblastomas, a series of spatially clustered somatic genomic SVs result in the juxtaposition of Growth Factor Independent 1b Transcriptional Repressor (GFI1b) to DNA elements which are located several hundred kilobases upstream, enabling the activation of GFI1b oncogenes[177].

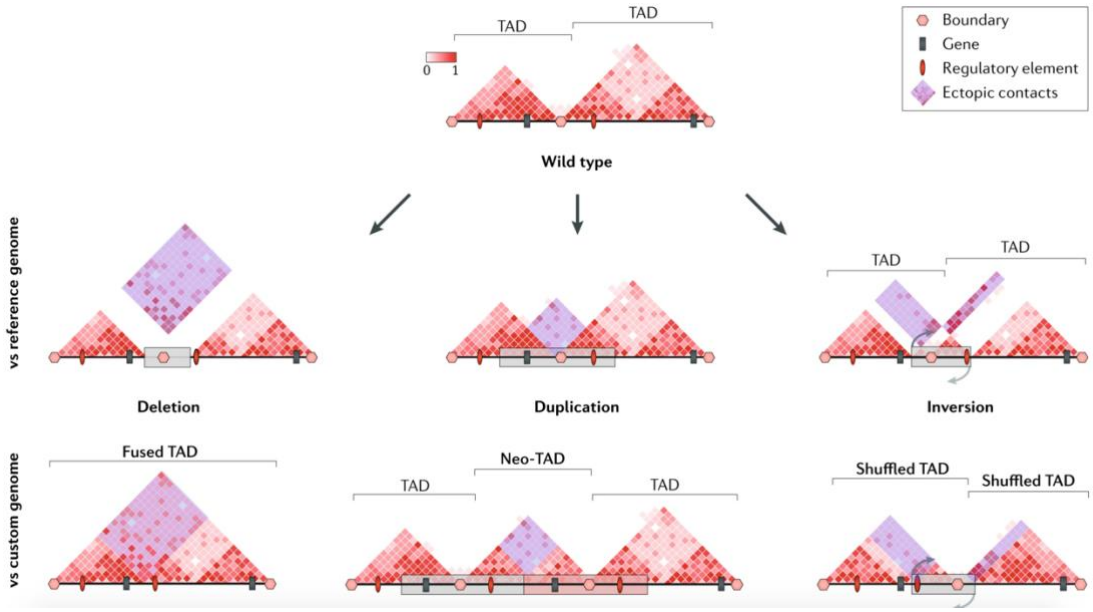


Figure 8: Structural variation in the 3D genome. *Figure depicts the various pathomechanics, such as TAD fusion (deletion), neo-TAD formation (duplication) or TAD shuffling (inversion) that can arise from structural variations at the topologically associating domains. Adapted from* [173].

Structural variations that cross a TAD boundary can lead to (i) fusion of TADs, thus enabling interactions between enhancers and genes outside the previous boundary, (ii) creation of a neo-TAD, thus restricting previously existing interactions and (iii) reshuffle TADs, resulting in altered interactions (**Figure 8**). In adult-onset demyelinating leukodystrophy, a ~660kb heterozygous deletion upstream of lamin B1 (LMNB1) promoter is observed to overexpress LMNB1 gene. This deletion results in the removal of a TAD boundary at the LMNB1 locus leading to overexpression of LMNB1 gene through enhancer adoption[178].

The pathomechanics of how duplications can result in the formation of new TADs was demonstrated by Franke *et al*[179]. In this study, the authors investigated two major TADs, one containing SRY-Box Transcription Factor 9 (SOX9) and the other containing, two potassium channels KCNJ2 and KCNJ16. Duplication events spanning the two TADs result in Cooks syndrome, a congenital limb malformation characterized by aplasia of nails and short digits[180]. The inter-TAD duplication event resulted in the formation of a new interaction domain covering the duplicated KCNJ2 gene causing misexpression of KCNJ2 gene under the control of duplicated SOX9 enhancers, which results in limb malformations.

Inversions spanning TAD boundaries can result in the fusion of two regulatory domains – TAD reshuffling. For example, in F-syndrome, a limb malformation syndrome characterized by severe and complex polydactyly, an inversion of ~1.1MB at the Ephrin Type-A Receptor

4 (EPHA4) locus results in relocation of an EPHA4 associated enhancer cluster into the vicinity of WNT6 (Wingless-Type MMTV Integration Site Family, Member 6) gene. This change results in the activation of WNT6 leading to abnormal limb development. [181]

Epigenetic modifications in enhancer dysregulation:

Global and local epigenetic changes such as chromatin remodelling and DNA modifications can mis-regulate enhancers. For example, Souren *et al*[182]., identified seven multiple sclerosis (MS) associated differentially methylated positions including the promoter of TMEM232 (Transmembrane Protein 232) gene and ZBTB16 (Zinc Finger and BTB Domain Containing 16) enhancer in MS-discordant monozygotic twins. Similarly, patients with hip and knee osteoarthritis, have differentially methylated enhancers compared to healthy individuals, in addition to organ source-dependent differences in enhancer methylation[183]. Similarly, in Wilson disease, differentially methylated regions specifically identifying patients were enriched in liver specific enhancers[184]. Alternatively, pulmonary endothelial cells from pulmonary arterial hypertension patients display an altered H3K27ac pattern especially in enhancers[185,186].

2.3. SOMATIC MUTATIONS AND CANCER

Mutations are changes in the DNA sequence often resulting from errors in replication or due to external factors such as exposure to chemicals and radiation. Germline mutations are inherited mutations that are present in all tissues of an individual. In addition to the germline mutations, over the course of an organism's life spontaneously occurring mutations called somatic mutations steadily accumulate in the cells[187,188]. Mutations can be of three different kinds *viz*., base substitutions, insertions and deletions. Single base substitutions are called point mutations and can be subdivided in to transitions and transversions. Transition occurs when a purine is substituted with another purine or when a pyrimidine is substituted with another pyrimidine, whereas transversion is when a purine is substituted for a pyrimidine or a vice versa. Insertions and deletions (indels) are additions or deletions of one or more nucleotides in DNA sequence. While most of these mutations do not have any significant impact, some mutations may affect a gene or a regulatory element and can lead to alterations in key cellular functions even leading to cancer[2,189].

Cancer results from the clonal expansion of single abnormal cells, mutations in the cells can confer selective advantages[187]. Mutations that are advantageous to individual expanding

clones are termed driver mutations[190]. Often, mutations that present no selective advantage may be carried over by expanding clones, such mutations are termed passenger mutations[1].

2.3.1. NON-CODING MUTATIONS IN CANCER

The discovery of driver mutations in various disease contexts has been focused only towards the protein-coding region[191–193]. This focality was primarily because of two reasons namely 1. the relevance of non-coding regions was largely unknown in the realm of diseases especially cancer, 2. use of exome sequencing rather than whole-genome sequencing (WGS) owing to large cost difference. However, in the past decade with the advancement of next-generation sequencing (NGS) both the reasons have vanished. Despite these changes, a major setback prevails in the annotation of the relevance of non-coding mutations with respect to cancer. This is due to the lack of a linear method to corroborate non-coding mutations and their function.

The various ways in which non-coding mutations exert effects include alterations in transcription regulation[194], disruption of chromatin domain structure[181], changes in mRNA stability[195,196]. The creation of a *de novo* TF binding site is yet another way in which a non-coding mutation can lead to functional outcomes. This was demonstrated in the landmark discoveries involving TERT core promoter mutations in melanoma[197,198]. In these studies, two mutually exclusive mutations in TERT promoter were identified, where both mutations lead to the creation of *de novo* predicted binding sites for E-twenty-six (ETS) family transcription factors, which in turn regulate TERT promoter activity. Thenceforth, TERT promoter mutations have been reported to be present in more than 50 tumour types, and in many of these they are the most frequently occurring driver alteration[199–201].

2.3.2. EXPLOITING MUTATIONS FOR CLINICAL ADVANTAGE

Even though the role of driver mutations in cancer is irrefutable, the genome-wide landscape of passenger mutations is important for understanding the complexity of oncogenesis and tumour evolution[202]. Moreover, recent studies have also shown that passengers can increase tumour immunogenicity and can correlate with improved clinical outcomes or reduced cell proliferation[203]. This knowledge has altered the paradigm of cancer diagnosis and treatment through the development of early diagnostic markers centred on mutational signatures, treatments based on targetable oncogenic alterations and better patient classifiers[204,205].

2.3.3. MUTATION BURDEN AND SPECTRA

Mutation burden can vary depending on the cancer type. Lower mutation rates are found in paediatric and haematological cancer while higher rates exist in cancers with environmental mutagens such as melanoma and lung cancer[206,207]. Also, within a cancer type, the degree of exposure to an environmental mutagen can greatly affect the mutation burden[208]. In addition, mutations in mismatch repair genes often lead to an accumulation of somatic mutations in human tissues[209].

Complex variation of mutation burden can also be observed in different regions of the genome, for example, late replicating and non-transcribed regions of the genome are often more mutated than early replicating regions and highly expressed genes[210–212]. Functional genomic elements such as exons, transcription factor binding sites and chromatin architectural elements are also known to have variable mutation burden compared to other genomic counterparts[5,213,214]. This variability can be explained by the accessibility of chromosomal areas to DNA repair, variable repair of DNA mismatches, nucleosome occupancy[209,212,215].

Different mutagens induce different mutational spectra (the proportion of various kinds such as transitions, transversions, deletions) and leave characteristic patterns of mutations, termed mutational signatures[216,217]. Signatures reveal mutation aetiology as a mutational process will cause only one type of somatic mutation. Examples include, the carcinogen aristolochic acid that causes an A>T substitution, while UV light exposure is associated with C>T mutations resulting from the erroneous repair of UV-induced pyrimidine dimers[218]. Additionally, the rates of different mutational processes also vary among cancer types and the mutations in an individual cancer genome can be a result of multiple mutational processes[219]. Thus, to systematically characterise and annotate the mutational processes, a number of mathematical models have been developed. Studies on multiple cancers have identified more than 30 single-base substitution (SBS) signatures, 11 doublet-base substitutions (DBS) signatures, and 17 Indel signatures[216,220–222].

For SBS, the mutation signature identification involves a classification comprising 96 classes[223]. This is constituted by 6 base substitutions C>A, C>G, C>T, T>A, T>C, T>G, plus the flanking 5’ and 3’ bases. The profile of each signature is displaced using these six substitution subtypes (**Figure 9**). Mutation signatures are reported based on the observed trinucleotide frequency which represents the relative proportion of mutations generated by

each signature to the reference genome[216,220]. Some of the known aetiologies of these signatures include exposures to external mutagens such as UV light[218], aflatoxin[224], aristolochic acid[225] and alkylating agents[226]; intrinsic factors such as defective repair mechanisms[227], altered polymerase E (POLE) activity[228] and APOBEC (apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like) activity[229].

COSMIC (catalogue of somatic mutations in cancer)[230] is a comprehensive repertoire of curated census of signatures, including the mutational profile, proposed aetiology and tissue distribution that shed light on the process of mutagenesis in cancer. The current version (v3.2) of the database hosts 67 SBS mutational signatures and version 2 consists of 30 signatures.

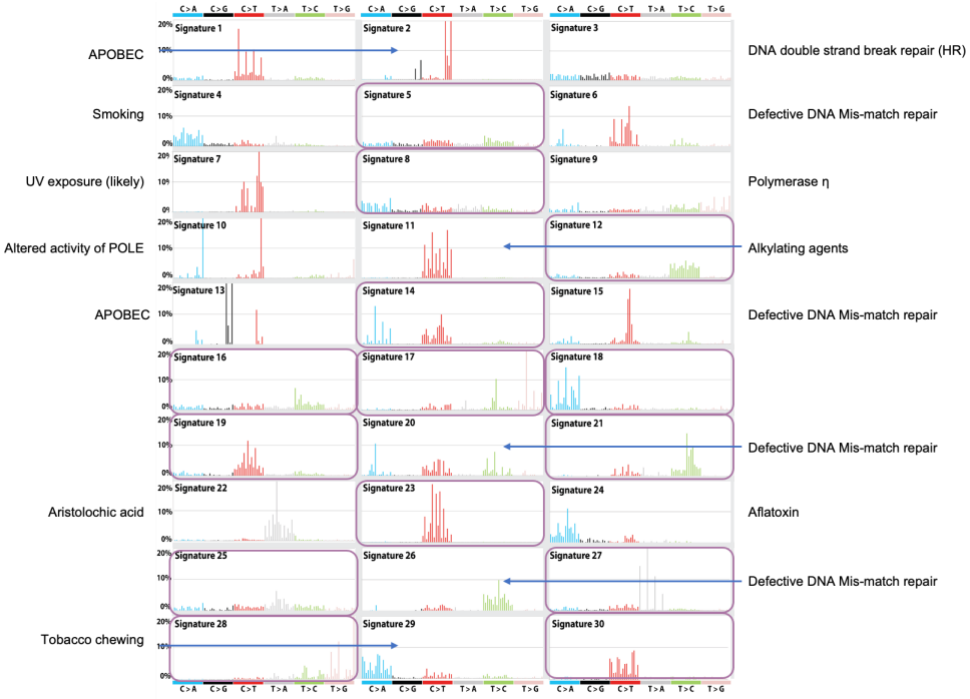


Figure 9: COSMIC mutation signatures with corresponding annotations. *Squared purple boxes indicate the signatures with unknown aetiology. (Modified and adapted from Alexandrov et al; Nature 2013).*

2.4. MUTATIONS AT PATHWAY LEVEL

Extensive efforts by large-scale sequencing projects have unravelled that the mutation load in cancer is abundant and heterogeneous[206,231]. These mutations do not independently cause cancer but rather in coalitions within various signalling and regulatory complexes[232]. Hence to study the effect of mutations in cancer progression, one of the rational approaches adopted is the inclusion of *a priori* knowledge of the cellular mechanisms and biological pathways of the mutated genes. This strategy not only reduces the dimensionality of the data but also

clusters genes into more interpretable and possibly clinically or experimentally actionable groups[233].

The main goals of this approach are (a) a better patient stratification to achieve optimal therapeutic outcomes; (b) development of pathway-targeted therapies and (c) generation of diagnostic signatures. To this end, multiple methodologies and computational solutions have been implemented to identify cancer driver pathways that disrupt the normalcy of a cell, hallmarked by poor regulation of critical functions such as growth, proliferation, and metabolism[191,234,235]. It is observed that even though a large number of driver genes are mutated in cancers, only a smaller number of pathways are targeted[236]. Among the pathways that are targeted in multiple cancers, RTK/-RAS signalling pathways is one of the most altered across all cancer types[237,238]. Other key pathways that are likely to be cancer drivers (functional contributors) are cell cycle, p53 and signalling pathways such as Hippo, Myc, Notch, PI-3-Kinase, TGFb, b-catenin/Wnt and Nrf2[239–244]. In each of these pathways multiple members are known to be mutated in more than one cancer type.

Genes in key pathways are not altered at equal frequencies, certain genes are recurrently mutated, while others are rarely or seldom mutated[237]. Numerous combinations of driver mutations can perturb a pathway important for cancer[245]. Consequently, even genes with infrequent mutation rates are relevant to cancer progression based on their pathway membership, physical or regulatory interactions with recurrently mutated genes[235]. Furthermore, mutations affecting a single pathway have shown synergistic effects with mutations deregulating alternative signalling pathways in the same tumour. Additionally, mutual exclusivity of mutations within a single pathway is noted in various cancers wherein it is rare for multiple cancer genes to be mutated in a single pathway, in a single tumorous tissue[246,247].

Phosphatidylinositol-3-kinase (PI3K-AKT) signalling is one of the key intracellular pathways which plays a role in cell growth, motility, survival, metabolism and angiogenesis[248,249]. PI3K-AKT pathway is found to be deregulated in almost all human cancers[250,251]. Hyperactivity of PI3K signalling is correlated with tumour progression and inhibitors targeting the signalling are used as therapeutic agents in various cancers.

2.4.1. TARGETED THERAPY

Traditional chemotherapeutic methods implement cytotoxic drugs to interfere with mitosis in a rapidly dividing cell[252]. Due to the generality of chemotherapy, side effects often result from the death of highly proliferative normal cells in the gut and immune system[253]. The new generation of cancer treatment drugs are designed to interfere with the molecular targets such as proteins that are critical in tumour growth or progression[254–256]. Such drugs fall in the category called targeted therapy.

An example of the inhibition of growth promoting pathways found in cancers is the usage of small molecule kinase inhibitor drugs on tumours with mutations in genes encoding protein kinases. For instance, in chronic myeloid leukaemia (CML), constitutive activation of Abl kinase in the Bcr-Abl fusion protein is inhibited by a small molecule tyrosine-kinase inhibitor imatinib mesylate (Gleevec)[257–260]. Similarly in gastrointestinal stromal tumours, the same drug inhibits platelet derived growth factor receptor (PDGFR)-α kinases[261–265]. FDA approved PI3K inhibitors include Copanlisib for follicular lymphoma (FL) treatment; Duvelisib for chronic lymphocytic leukaemia (CLL), small lymphocytic lymphoma (SLL) and FL; Idelalisib for CLL; Alpelisib for HR-positive and HER2/neu-negative breast cancer; and Umbralisib for marginal zone lymphoma (MZL) and FL[251]. Molecular targets based on mutations also include VEGF inhibitors, anti-BRAF target drugs, mTOR inhibitors, and cancer-specific fusion proteins. **Figure 10** lists the timeline of various small molecule anticancer drugs.

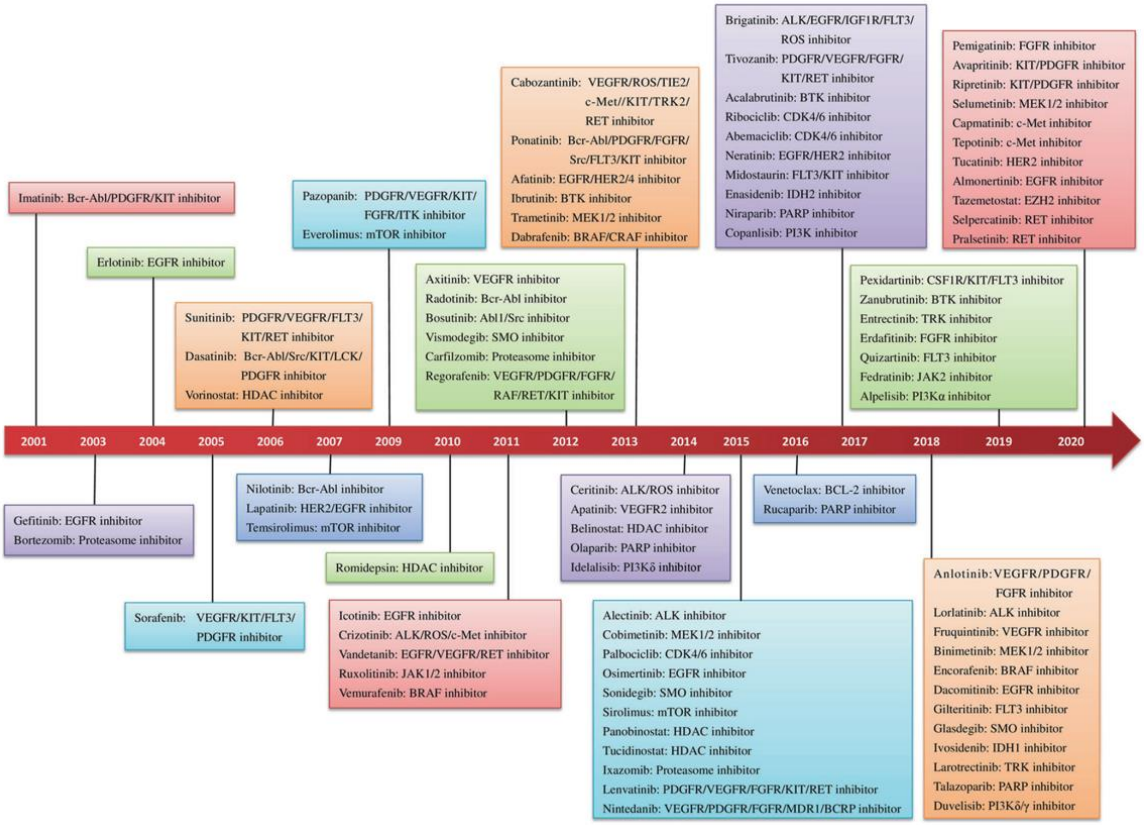


Figure 10: Timeline of small-molecule targeted anti-cancer drugs. *The figure shows various small-molecule anti-cancer drugs approved by the US FDA and National Medical Products Administration (NMPA) of China since 2001. (Adapted from Zhong et al; 2021*[266]*)*

2.5. RESPIRATORY SYSTEM AND LUNG ANATOMY

The human pulmonary respiratory system is broadly divided into airways and lung parenchyma[267,268]. The airways consist of the trachea that bifurcates into the right and left bronchus that divides into bronchioles and then further into alveoli. The lung parenchyma includes the alveoli, alveolar ducts and bronchioles. Lungs have a light, porous and spongy texture, whereas the surface is smooth and shining. Lungs are highly elastic in nature. From a pinkish white colour at birth, lungs turn to a dark slaty grey colour mottled in black patches with age. Anatomically, each lung is a conical shaped organ with an apex, three borders, three surfaces and a base.

2.5.1. CELLS OF THE RESPIRATORY TRACT

The respiratory tract is a complex system with multiple cell types found in precise numbers and positions to create the architectural features enabling the functioning of the organ (**Figure 11**). Diverse mesenchymal cells, namely fibroblasts, smooth muscle cells, endothelial cells, lipofibroblasts, myofibroblasts and bone marrow-derived cells are involved in the construction of this architecture[269]. Distinct epithelial cell types, primarily

basal and ciliated cells and fewer secretory cells, line the tubules of the airways and alveolar saccules. A varying number of secretory cells such as brush, goblet, club and neuroepithelial cells are present in the airways and submucosal glands[270]. Submucosal glands made of myoepithelial, basal, ciliated, goblet and other secretory cells make the cartilaginous airways. The alveolar region provides a vast epithelial lined surface, covered primarily by alveolar type 1 cells, which are in close contact with endothelial cells of the pulmonary capillaries[270].

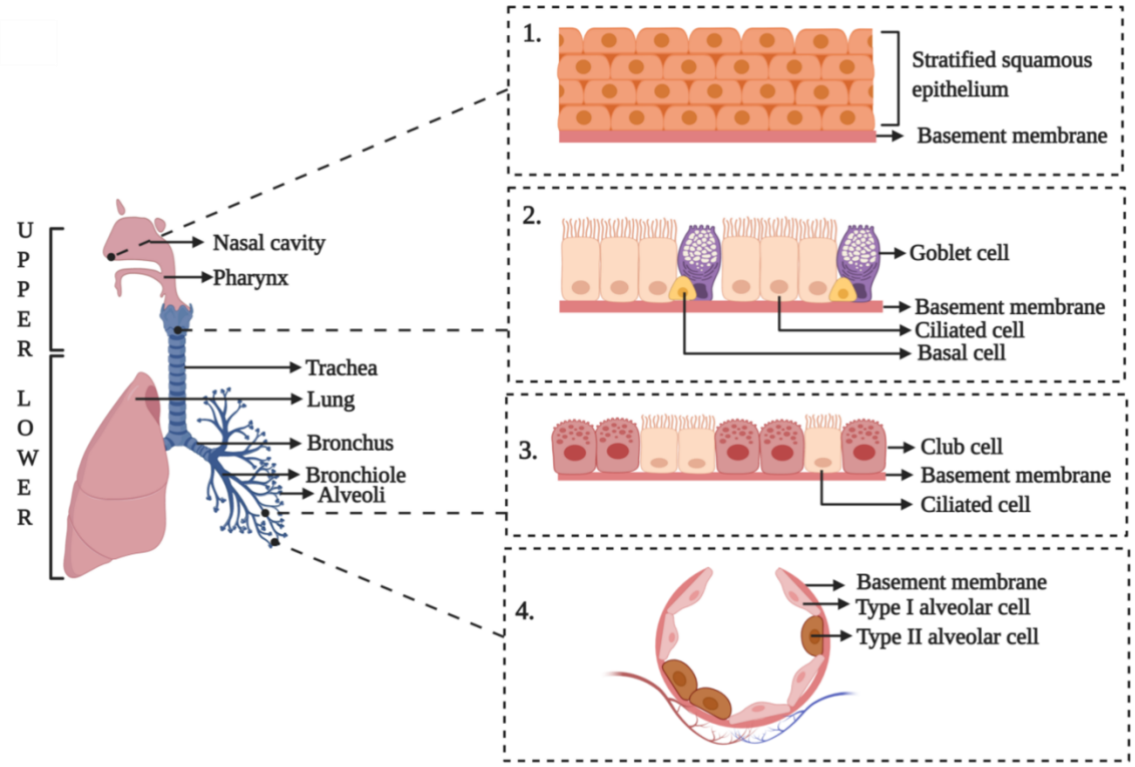


Figure 11: Cells of the respiratory tract. *Figure depicts the lung anatomy and the various cells that compose the respiratory tract. Adapted from* [271]

2.5.2. LUNG CANCER

Lung cancer is a malignant tumour characterised by uncontrolled cell growth in tissues of the lung. It is the leading cause of cancer related deaths worldwide[272]. The majority of lung cancers are a result of long-term tobacco smoking[273–275]. Other common causes include exposure to radon[276–278], arsenic[279,280], chromium,[281] and nickel[282,283]. Increased lung cancer rates are also associated with pre-existing non-malignant lung diseases such as chronic obstructive pulmonary disease, idiopathic pulmonary fibrosis and tuberculosis[284,285]. Lung cancer is categorised in to two main histopathological groups: non-small cell lung cancer (NSCLC) and small cell lung cancer (SCLC)[286]. This classification also reflects in the prognostic, and therapeutic implications of the disease.

NSCLC is the most common type accounting for 80-85% of cases and can be grouped into three main types namely 1. lung adenocarcinoma (LUAD), 2. lung squamous cell carcinoma (LUSC) and 3. large cell undifferentiated carcinoma (LCUD)[286]. Amongst these lung adenocarcinoma is the most common form accounting for 30% of all cases[285]. Adenocarcinomas of the lung are found in the glands that secrete mucus[287]. Squamous cell carcinomas are usually found in cells that cover the surface of the airways near the centre of the lung[288]. While large cell undifferentiated carcinoma can be found anywhere in the lung. SCLC often starts in the bronchi and is known to spread to other parts[289]. However, the exact cellular origin of lung cancer is not evident.

2.5.3. DIAGNOSIS AND PROGNOSIS

Signs and symptoms of lung cancer vary depending on the type and the extent of the tumour. The diagnostic evolution of suspected patients includes tissue diagnosis, staging and a functional patient evaluation. Sputum cytology, thoracentesis, accessible lymph node biopsy, bronchoscopy, thoracoscopy and thoracotomy are various ways employed by physicians for histologic diagnosis. Chest radiographs, chest computed tomography (CT), positron emission tomography (PET), fluoroscopy, Magnetic resonance imaging (MRI) and video assisted thoracoscopy are other modalities that are used for diagnostic or staging purposes[290].

Staging of lung cancer is based on the tumour-node-metastases (TNM) which describes the extent of the disease in terms of the size, location and extent of the primary tumour (T), the presence and location of lymph node involvement (N) and the presence or absence of distant metastasis (M). Based on the T, N and M descriptors lung cancer is categorised into IA, IB, IIA, IIB, IIIA, IIIB and IV stages[291,292]. Stage IV lung cancer is the most advanced form and is characterised by the (a) presence of cancer in both lungs, (b) if the cancer is spread outside the chest to a lymph node or to an organ or (c) more than one organ.

Treatment for lung cancer primarily depends on the type and stage of lung cancer. Surgical resection is the standard treatment for stage I-II NSCLC patients[293], IIIA patients in addition to surgery undergo pre- or post-operative radiation and/or chemotherapy[294]. Chemotherapy is the standard treatment for patients with inoperable tumours and advanced stages of cancer[295,296]. Patients with SCLC are treated with single-agent or combination chemotherapy with radiation therapy[297–299]. Disease free-survival time varies greatly in patients despite the considerations in the treatment regime based on the known predictors of variability[290].

2.5.4. *GENOMICS OF LUNG CANCER*

Lung cancer has the second highest reported mutation rate with a mean of 12.9 mutations per megabase for smokers[300]. Mutations in the Tumor Protein P53 (TP53), epidermal growth factor (EGFR), anaplastic lymphoma kinase (ALK), Erb-B2 Receptor Tyrosine Kinase 2 (ERBB2), B-Raf proto-oncogene (BRAF) and kirsten rat sarcoma virus (KRAS) genes are commonly observed in lung adenocarcinoma patients[301]. In squamous cell carcinoma, Fibroblast Growth Factor Receptor 1 (FGFR1) amplification, ERBB2 amplification, Discoidin Domain Receptor Tyrosine Kinase 2 (DDR2) mutation are some of the commonly observed alterations[288]. Amongst lung cancer patients, KRAS is the most frequent driver mutation seen in 25% of patients with adenocarcinoma[302]. In spite of the high number of mutations, KRAS has proven to be a challenging targetable alteration. One of the important developments from bench to bedside in lung cancer studies is the molecular-guided precision therapy targeting EGFR mutations in NSCLC patients. Inhibition of EGFR kinase activity by EGFR tyrosine inhibitors such as gefitinib or erlotinib are known to have significant results in the effective treatment of patients with NSCLC and with EGFR mutations[303]. Similarly, NSCLC patients with translocations of ALK and c-ros oncogene 1 (ROS1) genes are known to respond well to crizotinib, a tyrosine kinase inhibitor[304].

2.5.5. *MARKERS FOR TARGETED THERAPY*

Currently, patients with advanced lung cancer are tested for two biomarkers: EGFR and ALK[305,306]. EGFR mutation analysis is the best predictive marker for the use of EGFR tyrosine kinase inhibitors (EGFR-TKI) therapy, the first line of treatment in NSCLC.[307] Whereas, EGFR mutations have not been identified in SCLC. Deletions in exon 19 and a SNP (L858R) in exon 21 of the EGFR gene has been associated with a 70% response rate to EGFR-TKI therapy[305]. It is important to note that irrespective of the treatment, it is observed that the prognosis of EGFR mutated NSCLC is better than the wild-type NSCLC. Sanger sequencing is most widely used for the mutation detection, followed by NGS based sequencing to reach higher analytical sensitivity. However, for NGS the amount of input DNA could be difficult to obtain on bronchial biopsies. Alternatively, immunohistochemistry (IHC) with EGFR mutation-specific antibodies to identify the deletion in exon 19 and SNP in exon 21 have be suggested, but due to lower sensitivity are not used in predictive testing[308].

In NSCLC patients, a small inversion within chromosome 2p results in a fusion gene with portions of EML4 (echinoderm microtubule associated protein like 4) and ALK gene[309].

Fluorescent in-situ hybridization (FISH), IHC or Real time- polymerase chain reaction (RT-PCR) is employed in the detection of ALK gene fusion. Patients with ALK-positive lung cancer undergo a second line of treatment with crizotinib[304,305].

Additionally, mutations in specific genes including Her2, BRAF, NUT, MET, ROS1, DDR2, FGFR1, KRAS, and PTEN might provide information for clinical decision making, the association between mutation and clinical response is not straightforward[305].

Based on the knowledge that non-coding regulatory regions such as enhancers are cell type-specific and they play a crucial role in regulating the expression of their target genes, we hypothesised that mutations in enhancers could significantly influence cancer prognosis or patient survival and thus be exploited for better patient stratification and as novel prognostic biomarkers. In this thesis, we explore strategies to prioritise non-coding mutations characterising their functional importance in cancer. We leverage the epigenomic information on enhancers for their genome-wide identification in various lung tissue and cell types. We use state of the art enhancer target gene mapping by accounting for the three-dimensional architecture information; implement an ensemble approach for the calling of non-coding somatic mutations. We identify non-coding regulatory mutations that are relevant to lung cancer prognosis by exploring recurrently mutated enhancers and mutations that aggregate at biologically relevant pathways.

3. MATERIALS AND METHODS

3.1. CELL LINES AND CELL CULTURE

Human NCI-H460 cells were obtained from ATCC and were cultured in RPMI 1640 medium (catalogue no: BE12-167F, Lonza) containing 10% foetal bovine serum (FBS) (catalogue no. 10270-106 Life Technologies), Glutamine (catalogue no: LOBE17605F, Euroclone).

HBEC-3KT (Human Bronchial Epithelial cells immortalized with CDK4 and hTERT) cells were obtained from Voden and also available through our collaborators at INT (Dr Luca Roz) and were cultured in Keratinocyte-SFM with L-glutamine (catalogue no: 17005034 Thermofisher Scientific) with Keratinocyte-SFM supplements: human recombinant epidermal growth factor (EGF 1-53) and bovine pituitary extract (BPE) (catalogue no:37000015 ThermoFisher Scientific).

All cells were cultured at 37 °C in 5% CO_2 and were regularly tested for mycoplasma contamination

3.2. GROWTH MEDIA AND BUFFER COMPOSITION

Table 3: Growth media and buffers used.

TE	Tris-HCl 10mM (pH 8), EDTA 1mM (filter with 0.2µM)
PBS	137mM NaCl, 10mM PO_4 (pH 7.4), 2.7mM KCl (filter with 0.2µM)
Resuspension Buffer	10mM Tris -HCl (pH 7.4), 10mM NaCl, 3mM MgCl2, 0.1% Tween-20
Lysis Buffer	10mM Tris -HCl (pH 7.4), 10mM NaCl, 3mM MgCl2, 0.1% Tween-20, 0.1% NP40, 0.01%Digitonin
2x TD Buffer	20 mM Tris-HCl (pH 7.6), 10 mM MgCl2, 20% Dimethyl Formamide
Transposition Buffer	25 ul 2x TD buffer, 2.5 ul transposase (100nM final), 16.5 ul PBS, 0.5 ul 1% digitonin, 0.5 ul 10% Tween-20, 5 ul H2O
Fixing solution	50mM Hepes-KOH (pH7.5), 100mM NaCl, 1mM EDTA, 0.5mM EGTA, 11% formaldehyde - in H2O

Sonication buffer	10 mM Tris-HCl (pH 8.0), 2 mM EDTA, 0.25% SDS, 1X PMSF, 1X protease inhibitors
Equilibration buffer	10mM Tris-HCl (pH 8.0), 233mM NaCl, 1.66% Triton X-100, 0.166% DOC, 1mM EDTA, 1X PMSF, 1X protease inhibitors
IP buffer	10 mM Tris-HCl (pH 8.0), 140 mM NaCl, 1 mM EDTA, 0.1% SDS, 0.1% DOC, 1% Triton X-100 1X PMSF, 1X protease inhibitors
High-salt IP buffer	10 mM Tris-HCl (pH 8.0), 500 mM NaCl, 1 mM EDTA, 0.1% SDS, 0.1% DOC, 1% Triton X-100 1X PMSF, 1X protease inhibitors
RIPA-LiCl buffer	10 mM TrisHCl (pH 8.0), 1 mM EDTA, 250 mM LiCl, 0.5% DOC, 0.5% NP-40, 1X PMSF, 1X protease inhibitors
Elution buffer	10 mM Tris-HCl (pH 8.0), 5 mM EDTA, 300mM NaCl, 0.4% SDS

3.3. REAGENTS AND INSTRUMENTS

Table 4: Reagents and instruments used.

Qiagen Mini elute kit
Zymo DNA Clean and Concentrator-5 Kit (Cat No. D4014)
Covaris Sonication E220 evolution
microTUBE AFA Fiber Pre-Slit Snap-Cap 130 µl (Part. No. 520045)
Agilent 2100 Bioanalyzer Instrument
NanoDrop™ 2000/2000c
Qubit 4 Fluorometer (Cat No. Q33238)
Qubit™ 1X dsDNA HS Assay Kit (Cat No. Q33230)
Illumina Nextseq 550 System Next Generation Sequencer
NextSeq 500/550 High Output Kit v2.5 (150 Cycles) (Cat No. 20024907)
37% Formaldehyde solution Sigma (Cat No. 47608)

Protease Inhibitor Cocktail EDTA-free
Phenol:Chloroform:Isoamyl Alcohol 25:24:1, Saturated with 10mM Tris, pH 8.0, 1mM EDTA (Cat No. P3803)
SYBR select master mix (Invitrogen, 4472908)
Dynabeads Protein G (Invitrogen 10004D).
IGEPAL CA-630 (Sigma Aldrich I3021)

3.4. GENOME-WIDE REGULATORY REGION IDENTIFICATION

3.4.1. CHIP -SEQ

Cells were cross-linked in 1% fixing solution for 10 min at room temperature, lysed and chromatin sheared. 5% of chromatin was saved as input. Immunoprecipitation (IP) was performed overnight on a wheel at 4 °C with H3K27ac antibody (Catalogue No. AB4729) or control IgG (AB37415). The following day, antibody-chromatin immune-complexes were loaded onto Dynabeads Protein G (Invitrogen 10004D).

The bound complexes were washed twice in IP buffer, twice in High Salt Solution followed by RIPA-LiCl buffer (twice) and once in 10mM Tris-HCl (pH 8). Crosslinking was reversed at 65 °C overnight in Elution Buffer, DNA was purified by standard phenol/chloroform extraction, precipitated and resuspended in 30 µl of 10 mM Tris-HCl (pH 8). ChIP efficiency was tested by qPCR reactions, performed in triplicate using the SYBR select master mix (Invitrogen, 4472908) on a StepOnePlus™ Real-Time PCR System (Applied Biosystems) on CDH13 promoter (positive control) and on the gene body of RARRES2P9 (negative control, **Table 5**). Relative enrichment was calculated as the IP/Input ratio. The Input and IP samples were ligated with illumina barcodes and amplified using the Kapa library amplification kit, followed by size-selection with AMPure XP Beads. ASPRI cleanup with a 1.5× AMPure XP Bead: DNA ratio was performed and final libraries were eluted and sequenced using Illumina Nextseq 550 System with NextSeq 500/550 High Output Kit v2.5. For each sample (IP and Input) approximately 90 million paired-end reads were obtained.

Table 5: ChIP-qPCR Primers

Primers	**Name**	**Sequence**	**Significance**
CDH13	Forward	5’- TGTGTCTGCCCATCATCTGT -3’	

Promoter	Reverse	5'- TGAATTGTGGTTACATGGAGGT-3'	Positive control
RARRES2P9 gene body	Forward	5'- AGCTGTGGTATCCTCACCG-3'	Negative control
	Reverse	5'- GACTGCCTTACAGAGACGC-3'	

3.4.2. ATAC-SEQ

This method is adapted from Omni-ATAC protocol[310]. 50,000 cells were transferred in 500µl ice-cold PBS 1X and spun at 500 x g for 10 minutes. To avoid losing cells during the nuclei prep, the supernatant from the pellet after each centrifugation was carefully pipetted away. Cells were washed with 500 µl of ice-cold PBS 1X, centrifuged at 500 x g for 10 minutes. Cells were resuspended in 300 µl of Lysis buffer (freshly prepared and chilled) and incubated on ice for 15 minutes. Immediately after lysis, nuclei were spun at 500 x g for 10 minutes and the nuclei pellet was washed twice with 300 of Lysis buffer. Pellet was resuspended in the Transposase reaction mix. For 10 K cells, prepare a mix with 2 µl of Tn5 in a total volume of 50 µl and incubate for 30 mins at 37°C in a thermal-mixer at 1000 rpm. Samples were purified by Qiagen Mini elute kit according to the manufacturer's protocol (elution in 21 µl of elution buffer) and amplified for 5 cycles using NEBNext 2x MasterMix.

Table 6: PCR profile for adapter incorporation

Temperature	Time	Cycles
72 ºC	5 min	1
98 ºC	30 sec	
98 ºC	10 sec	5
63 ºC	30 sec	
72 ºC	1 min	
4 ºC	Inf	

Tubes were removed from the thermocycler and stored on ice. Using 5 ul (10%) of the pre-amplified mixture qPCR amplification was performed to determine additional cycles.

Table 7: qRT-PCR profile for additional cycle computation

Temperature	Time	Cycles
98 ºC	30 sec	1
98 ºC	10 sec	20
63 ºC	30 sec	

72 °C	1 min	
4 °C	Inf	

After qPCR amplification, the amplification profiles were manually assessed and the required number of additional cycles to amplify were determined as described in [117]. Using the remainder of the pre-amplified DNA, the required number of additional cycles were run without additional reagents. PCR products were purified using Zymo DNA Clean and Concentrator and eluted in 20 ul H_2O. The libraries were then size selected with AMPure XP Bead. ASPRI cleanup with a 1.5× AMPure XP Bead: DNA ratio was performed and final libraries were eluted and sequenced using Illumina Nextseq 550 System with NextSeq 500/550 mid Output Kit v2.5. For each sample approximately 70 million paired-end reads were obtained.

3.4.3. ChIP-seq and ATAC-seq data analysis

Paired-end raw reads were filtered based on the quality value obtained from FastQC[311,312] v0.11.9 (-q 10 and -p 30) using the Trim Galore![313] software v0.6.4_dev. The filtered reads were aligned to the hg19 reference genome using BWA[314] v0.7.17-r1188 to produce the alignment file (BAM). The PCR duplicates were removed from the aligned BAM files using PICARD tools[315] v2.23.1. The BAM files were sorted and indexed for peak calling using SAMtools[316]. The BedGraph files were generated by comparing BAM files of IP and input (IP read coverage/input read coverage) resulting in a ratio for every base across the whole genome using bamCompare from deepTools[317] v3.4.3. To call the peaks MACS2[318] v2.2.7.1 tool was used. This framework was implemented using nfcore[319]/chipseq v1.2.1 or nfcore/atacseq v1.2.1 pipeline for ChIP and ATAC sequencing data respectively. The bed and bedgraph files obtained from the analysis were visualized using the IGV[320] browser and further processed using custom made R and Python scripts.

3.4.4. Enhancer definition

For the definition of lung-specific enhancers across the genome we leveraged the epigenetic markers of open chromatin such as H3K27ac and DNase sensitivity. We downloaded uniformly processed H3K27ac ChIP-seq and DNAse-seq files in bigbed format for six lung tissue/cell types with replicates from ENCODE3. First, we filtered the results for subsequent analyses considering only peaks with strong significant enrichment, i.e. $-\log10(\text{adj.P-value}) \geq 2$. Second, we merged peak genomic coordinates across replicates and defined consensus peaks as merged peaks that overlapped individual replicate peaks in greater than 50% of

replicates. In addition, we also performed H3K27ac ChIP-seq and ATAC-seq on two lung cell lines viz., HBEC-3KT and NCI-H460. We intersected the peaks of the replicates for the in-house data to obtain a cell type-specific list.

To obtain a comprehensive list of *cis*-regulatory elements we conducted a two-step procedure. First, for each of the eight-lung cell/tissue type (*viz*., lung, IMR-90 and PC-9, A549, AG04450, fibroblast, NCI-H460 and HBEC-3KT), the intersection between H3K27ac and accessible peaks (ATAC or DNase-seq based) with overlapping regions (≥6 bps) were used to define cell-specific enhancers. Both the number and size of DNase-seq and H3K27ac ChIP-seq peaks vary across cell and tissue types. Namely, accessible peaks were 213,793 on average, with average size of 466 bps, H3K27ac peaks were 88,898 on average with size-1,168 bps resulting in 55,103 putative enhancers with an average of 405 bps.

Additional filters were applied *ex-post*, such as the removal of interval portions overlapping annotated exons (for both coding and non-coding genes). The resulting regions were further annotated with respect to the transcription start site (TSS) as promoter-proximal (within 3.5 kb upstream and 1.5 kb downstream of TSS) or distal, and only the promoter-distal ones were retained for the following steps.

Secondly, cell type-specific enhancers with overlapping intervals across different cell types were merged (union) together to define a consensus set of enhancer regions. This set was filtered based on size to remove intervals larger than 2.5 kb. Non-canonical and Y chromosomes were excluded. The merged regions were also filtered based on position to include only non-coding promoter-distal regions similar to the previous step to obtain the reference list of lung-specific enhancers. (N = 187,206). This is meant to be a comprehensive reference set of enhancer regions in at least one of the lung cell types considered.

3.4.5. PROMOTER DEFINITION

We defined reference promoters as 2 kb regions (1.5 kb upstream and 0.5 kb downstream) around the transcription start site (TSS) of annotated protein-coding genes, based on RefSeq[321] annotations in (hg19.ncbiRefseq.gtf.gz; May 2019, hg19 genome assembly). Non-canonical and Y chromosomes were excluded. To create a more comprehensive list of promoters, in case of multiple alternative transcripts for the same gene the promoter for each

transcript was considered barring overlapping regions with exons and 5'UTR of another transcript.

3.5. MUTATIONS

3.5.1. MUTATION CALLING AND MAPPING

High coverage whole-genome sequencing data were downloaded from TCGA (105 samples) and EGA (59 samples), in the form of tumour and matched normal BAM files for multiple subtypes of lung cancer (Error! Reference source not found.**, Appendix**). For the uniform processing of the samples, the sequence data were realigned on hg19 using BWA following the GATK[322,323] best practices. Mutations as SNPs and small indels were called across the whole genome using Freebayes[324], Mutect[325], Scalpel[326], Vardict[327] and Varscan[328]. Mutations present in the low complexity regions as defined by *Li*[314] were removed. Finally, for determining a somatic variant we used the intersection of a variant call by at least two tools. A custom pipeline built on BC-BIO[329] was used to perform all the operations. The mutation list of each sample was then mapped on the lung-specific enhancers and promoters using pybedtools[330].

3.5.2. REGION SPECIFIC MUTATION BURDEN

To identify somatic mutation enrichment of various regions of the genome, we computed the burden of somatic mutations in enhancers, exons, promoters and non-coding regions for each sample. Non-coding regions was defined as the whole genome devoid of exons and enhancers.

$$mutation\ frequency\ of\ region\ x = \frac{\#mutations\ in\ region\ x}{\sum size\ of\ region\ x}$$

Where x $\in$ {enhancer, exons and non-coding regions, promoters.}

The mutation burden of each sample was plotted in a scatter plot in various comparison scenarios and the slopes of each linear regression were estimated.

3.5.3. MUTATION SIGNATURE

To obtain an approximate estimate of the contribution of different known mutational signatures to each sample we used the MutationalPatterns[331] Bioconductor package. As a reference set of mutational signatures, we used a table with the relative frequency of each of the 96 trinucleotide substitutions across 30 known mutation signatures from COSMIC

database version 2. Mutation signatures were estimated for the whole genome and the relative frequency of each signature was plotted in a heatmap.

To assess the variations in mutation signature between coding and non-coding regions in LUAD, LUSC and SCLC samples, we computed the difference in the relative contribution of the frequency and the significance was assessed using p-value obtained with the Wilcoxon rank-sum test[332] using the scipy.stats.ranksums function in SciPy[333]. Mutation signature was deemed significantly different between the two categories with a p-value lesser than 0.05. Further, the mutation signatures prevalent in enhancers, promoters and exons were computed and the relative contribution of the signatures across samples were used to plot box and whiskers plot.

3.6. HI-C DATASET PROCESSING

We processed eleven Hi-C datasets covering different cell lines and primary tissues from a compendium of public datasets[31,334–338]. For each Hi-C dataset we retrieved the raw FASTQ files from the NIH SRA database. The sequencing reads were aligned with the iterative mapping procedure (single-end mode) as implemented in hiclib (https://github.com/mirnylab/hiclib-legacy) (version from gitHub commit d38f198, date: 28 September 2017) using botwie2 (version 2.3.4.3) aligner (https://github.com/BenLangmead/bowtie2)[65]. The uniquely mapped reads information was stored in a HDF5 (Hierarchical Data Format) file for each FASTQ file. We filtered out events originating from non-canonical enzyme activity or non-enzymatic physical breakage. The distance cut-off was estimated for each dataset based on the frequency distribution of distances and the expected fragment length. We further removed duplicated read pairs, as well as read pairs derived from unligated or circularized fragments.

Finally, the genome was binned at 10 kb bin size, and the raw read counts were summarized in a Hi-C contact matrix for each chromosome, accounting for intra-chromosomal interactions. Chromosome-wise iterative correction (ICE)[68] with default parameters was applied using cooler[339] (version 0.8.5, https://github.com/open2c/cooler) to correct for technical biases and to enable comparability among all tissues and cell types. A balanced matrix of relative contact probabilities was obtained. The output files (cool format) were converted to txt files and compressed.

3.6.1. HIERARCHICAL CONTACT SCORE

We devised a score proportional to the likelihood of enhancer–promoter pairs co-localization, named Hierarchical Contact (HC) score to account for 3D spatial proximity of regulatory elements including the TADs hierarchical structure across multiple tissue and cell types. For HC definition we used on the Local Score Differentiator (LSD)[340] TAD borders calling procedure, as implemented in the HiCBricks (version 1.8.0) bioconductor package[341]. We defined TADs as regions between two consecutive domain boundaries.

3.7. ENHANCER TARGET GENE PREDICTION.

Target genes of enhancers were obtained using our in-house ETG prediction framework (**Figure 12**)[82]. We provided the tool with 180,852 lung-specific enhancers and 18,027 promoters.

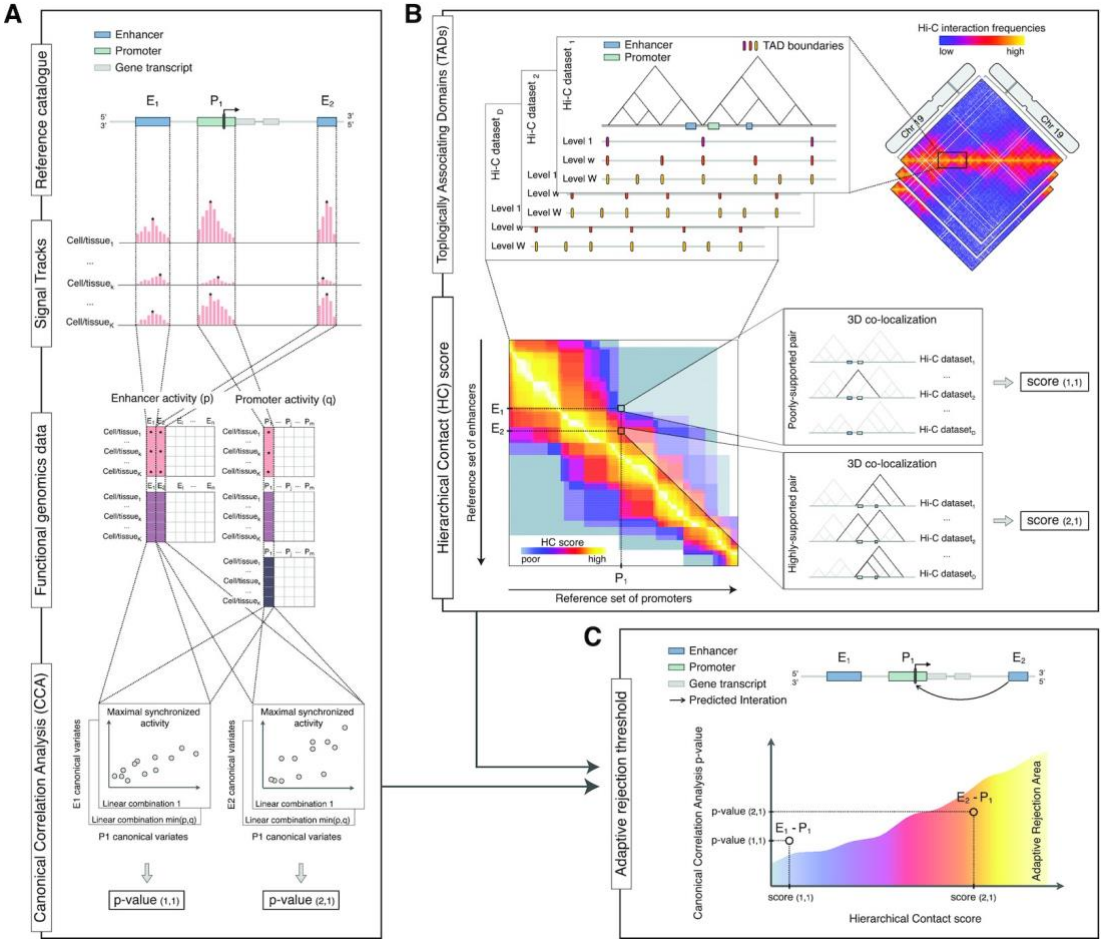


Figure 12: Enhancer target gene prediction. *Schematic illustration of the workflow of enhancer target gene prediction algorithm.* ***(A)*** *Correlation Analysis (CCA) is used to investigate the synchronized activity of each enhancer–promoter (EP) pair across k cell and tissue types.* ***(B)*** *Computation of Hierarchical Contact (HC) score based on the 3-dimensional localization.* ***(C)*** *The 3D co-localization information encoded in the HC score is used to estimate an adaptive rejection threshold to control for FDR in the multiple testing hypothesis of EP pairs synchronization. Published in* [82]

3.7.1. *Enhancer-promoter pairs synchronization analysis with Canonical Correlation*

Enhancer and promoter regions were considered separately and their respective activity statuses were measured using two sets of epigenetic marks: enrichment of DNase-seq and H3K27ac ChIP-seq was used for enhancers and DNase-seq, H3K27ac and H3K4me3 for the promoters. Consolidated fold-change enrichment signal tracks in bigwig format from the Roadmap Epigenomics consortium for 44 cell and tissue types were used as the source data. We then computed the maximum signal of the corresponding epigenomic marks for each enhancer and promoter region. Canonical correlation analysis [342] was adopted to investigate the inter-set correlation patterns to quantify the strength of each enhancer-promoter pair. We then computed a p-value for the overall dependence between each promoter and enhancer. Based on the correlation of enhancers and promoters, we obtained 1,809,529 pairs.

3.7.2. *3D architectural integration in the enhancer-promoter pairs FDR control*

To control over the number of false discoveries due to multiple hypothesis testing, we implemented the Adaptive P-value thresholding procedure (AdaPT[343]) by considering relevant contextual three-dimensional co-localization information in the form of HC-score. Upon implementing a 0.01 P-value thresholding based on AdaPT we obtained 48,829 enhancer-promoter pairs.

3.8. Tissue-specific expression quantification

For studying the tissue-specific expression levels, we obtained the gene TPMs from the Genotype-Tissue Expression (GTEx[344]) project database v8, (GTEx_Analysis_2017-06-05_v8_RNASeQCv1.1.9_gene_tpm.gct.gz). We compared the expression levels of genes with more than 25 enhancers (arbitrary cut-off) with the genes with fewer lung specific enhancers. The test gene set is composed of genes with at least 25 enhancers, (n=130). For the background gene set, we bootstrapped 10 sets from the genes with less than 25 enhancers with approximately the same size. The mean expression levels of the genes in each group were quantified and log2 fold change was computed. Mann–Whitney U test[345] was implemented to assess the significance, and Bonferroni correction[346] for multiple hypothesis testing was used to obtain adjusted p-value. Gene expression values for all the tissues were represented in box plots.

3.9. GENE ONTOLOGY ANALYSIS

For the gene ontology analysis, we used the target genes of mutated enhancers (n>3). The list of target genes (n =7,102) were then used to obtain gene ontology- molecular function and biological process through the g:profiler[347] tool.

3.10. GENE EXPRESSION ANALYSIS

RNA sequencing data was obtained for the patients (with high coverage WGS data) from TCGA (105 samples) and EGA (30 samples). The quantification of the transcripts was obtained using kallisto[348] as Transcripts per Kilobase Million (TPM) based on the hg19 reference genome. Samples were stratified into mutated and non-mutated based on the presence of enhancer mutation (at least one), and their corresponding gene's expression as TPM values were compared. The unpaired two-samples-Wilcoxon tests[332] was used to assess the significance of the difference.

3.11. PROMOTER METHYLATION

Methylation data for the TCGA samples were obtained as beta values of the Illumina Human Methylation 450 array. To assess the methylation status of a promoter, mean methylation beta values of the probes present in the promoter (2 kb around TSS) were computed.

3.12. STRUCTURAL VARIATION

Copy number alteration information for the TCGA samples was obtained as GISTIC 2[349] gene-level copy number scores. We also applied Meerkat[350] (v0.189), a somatic structural variations tool to understand the mechanism of complex structural variations at specific loci. The tool was implemented using a custom-made Singularity image, that consisted of appropriate Ubuntu system libraries, BioPerl[351] v1.7.2, BWA v0.6.2, NCBI-BLAST[352] v2.2.24 and samtools v0.1.19.

3.13. SURVIVAL ANALYSIS

Clinical features such as sex, vital status, TNM stage and smoke exposure were also obtained from TCGA for the patients. Event-free survival probabilities were calculated by using the Kaplan-Meier[353] method (survminer R package). Log-rank test[354] was used to assess the statistical significance of the different groups.

3.14. CANDIDATE ENHANCER VALIDATION

For the experimental validation of the role of the CDH13 intronic enhancer in regulating its gene expression, several lung cancer cell lines were screened for the expression of CDH13 gene and the sequence of the enhancer region was assessed using sanger sequencing. We also determined the copy number of CDH13 genes in the cell lines by comparing with the GAPDH gene in a cell line with known copy number of GAPDH.

3.14.1. RNA ISOLATION AND QRT-PCR ANALYSIS

RNA was isolated using Qiagen AllPrep DNA/RNA Mini Kit following the manufacturer's protocols. For qRT-PCR, 500 ng of RNA was reverse-transcribed using superscript III following the manufacturer's protocol. Quantitative RT-PCR (qRT-PCR) was performed with TB Green® Premix Ex Taq™ (Tli RNase H Plus) using Roche LightCycler 96. PCR amplification parameters were 98°C (30s), and 35 cycles of 98°C (10s), 65°C (30s), 72°C (10s) and 72°C (2min). Primer sequences are listed below.

Table 8: qRT- PCR primer sequences for gene expression quantification

Primers	**Name**	**Sequence**
CDH13_all	Forward	5'-AAAGCCTGGCTCCCACGGAAAATA-3'
	Reverse	5'-CGGCTGCATTTTGTCCGACTAGAA-3'
CDH13 4iso	Forward	5'-GACATTGTCACTGTTGTGTCACCTG-3'
	Reverse	5'-CCGTGCCTGTTAATCCAACATC-3'
Beta-actin	Forward	5'-TGGCACCCAGCACAATGAA-3'
	Reverse	5'-CTAAGTCATAGTCCGCCTAGAAGCA-3'

3.14.2. ENHANCER SEQUENCE DETERMINATION

DNA was isolated using Qiagen AllPrep DNA/RNA Mini Kit following the manufacturer's protocols. CDH13 enhancer region was amplified and run on a 1.5% agarose gel. All the bands were purified with Qiagen PCR purification kit, according to the manufacturer's instructions. Samples were eluted in 30 µl and sequenced.

Table 9: CDH13 enhancer - primer sequence

Primers	**Name**	**Sequence**
	Forward	5'-CCCTCGGGATTCATGCCTCATAAA-3'

CDH13-enhancer	Reverse	5’-CCTCTAAGGGTTGCAGGAAGGATT-3’

3.14.3. HOMOZYGOUS DELETION OF CDH13 ENHANCER

For evaluating the role of intronic enhancer in the CDH13 gene, we employed CRISPR based genome editing in the NCI-H460 cell line. We designed 2 pairs of guide RNAs namely T1-T3 and T2-T3 targeting the CDH13 intronic enhancer (**Figure 13**). With the help of the IFOM genome editing facility, we performed the deletion of the enhancer region.

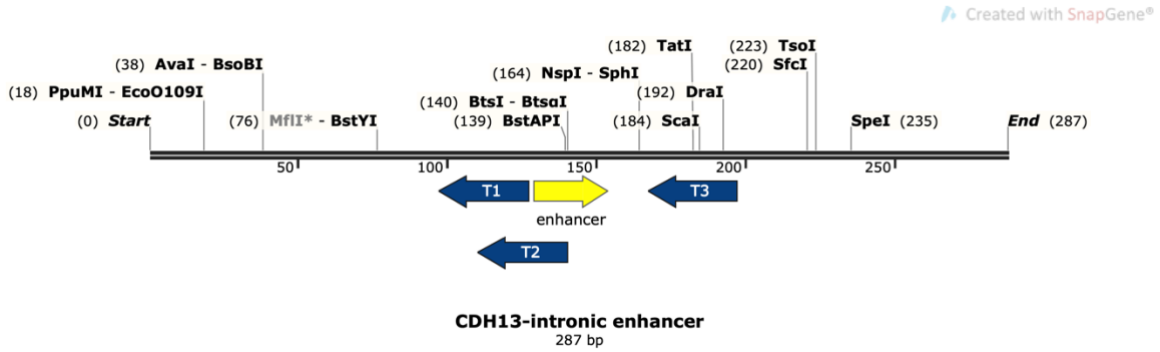


Figure 13: Map of the CDH13 enhancer locus with CRISPR Cas9 shRNA guides. *Yellow arrow indicates the enhancer region, blue arrows indicate the Cas9 shRNA guides. Figure generated using SnapGene.*

We identified the clones with successful removal of the enhancer region using qPCR (**Figure 14**). Following which we determined the sequence of the clones using sanger sequencing **(Appendix Figure 2**). We selected 5 clones with homozygous deletion for the expression quantification of the CDH13 gene. CDH13 expression was quantified in the homozygous clones and WT NCI-H460 cells using qRT-PCR in comparison to beta actin as described in section 3.14.1.

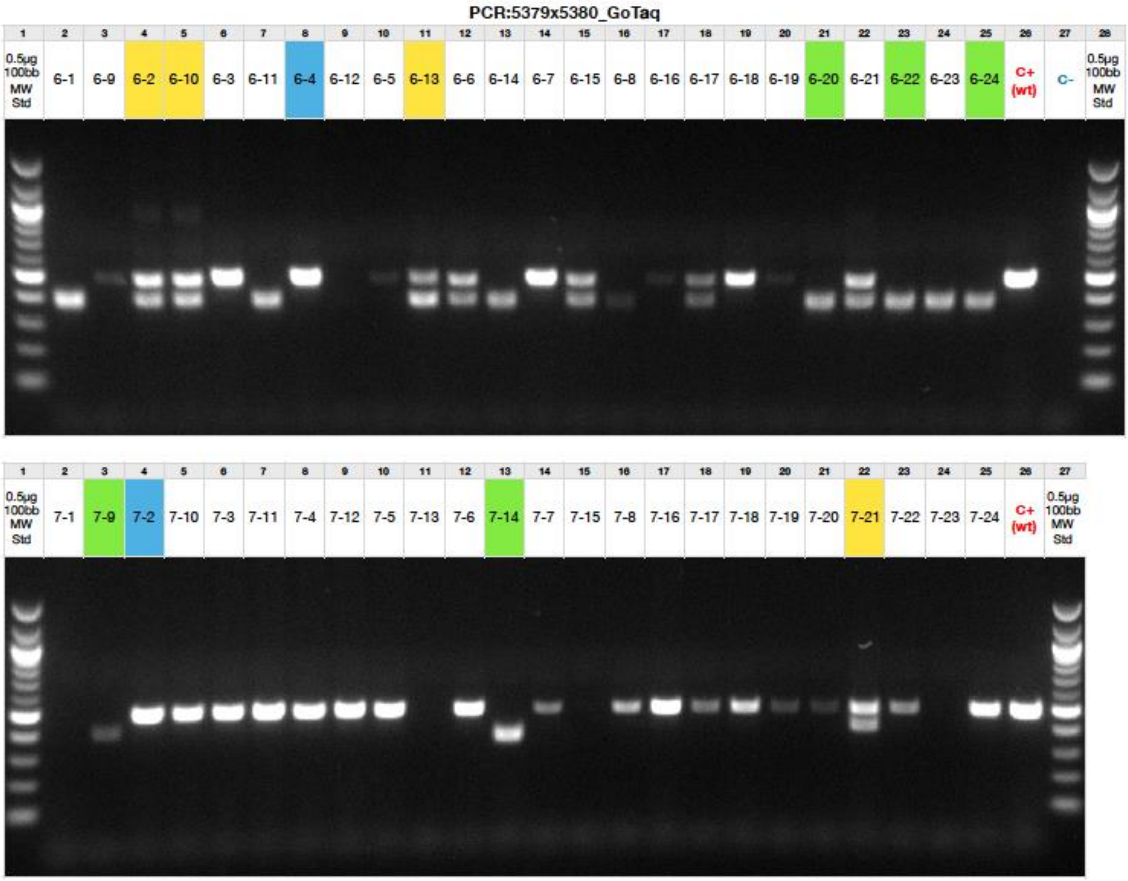


Figure 14: Screening for CDH13 enhancer deletion*. Each lane is a clone of NCI-H460 cell line for CRISPR deletion screening. The two rows denote the two different combination of guide RNAs used in the CRISPR experiment. WT clones have a size estimate of 503bp, and the deleted allele is ~400bp. The lane colours yellow: heterozygous deletion, green: homozygous deletion and blue: no deletion. The first and the last lane in both the rows show the 100bp molecular weight marker. Penultimate lane in the top row (c-) is the negative control and c+ lane in top and bottom row show the positive control (wild type NCI-H460 without CRISPR deletion)*

3.15. TRANSCRIPTION FACTOR BINDING SITE ANALYSIS

To calculate the presence of motifs in enhancer cores, we used FIMO from the MEME suite with a custom library of all TRANSFAC (and Jaspar) motifs at a q-value threshold (FDR – Benjamini-Hochberg multiple testing correction) of 0.05. The sequence motif alteration upon mutation in enhancer core was assessed by identifying motifs in the enhancer core with reference alleles and mutant alleles. The loss of a motif or the gain of a new motif at any locus were given different scores and plotted as a stacked bar plot.

3.16. PATHWAY LEVEL ENRICHMENT ANALYSIS

For the pathway level enrichment analysis, we used target genes of mutated enhancers (n>3). A total of 7,102 genes were used as the query in g:profiler tool to identify KEGG pathways that were significantly enriched (p adjusted <0.01).

3.17. GENE-SET ENRICHMENT ANALYSIS

Target genes of enhancers with at least 12 mutations (n= 466) were used for the gene-set enrichment analysis. Two different datasets of the MSigDB resources C2 (curated gene sets) and C6 (oncogenic gene sets) were used through the GSEA[355]. Gene-sets with a p-value <0.01 were considered significantly enriched and the results were plotted using custom Python scripts.

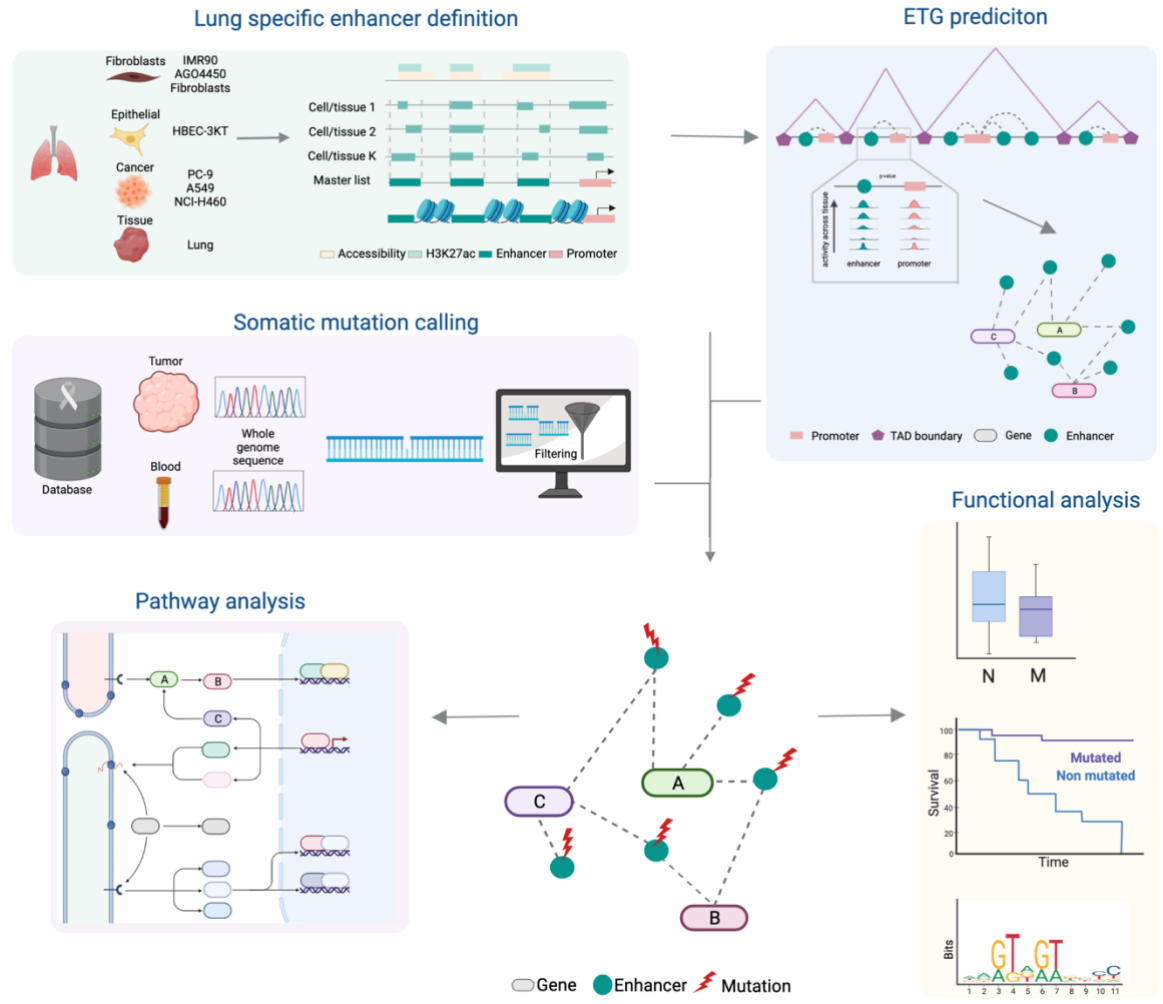


Figure 15: Schematic illustration of the methodology. *Figure depicts the various aspects of the methodologies namely: enhancer definition, somatic mutation calling, enhancer target gene prediction, pathway and functional analysis. Figure was generated using Biorender tool.*

4. RESULTS

4.1. GENOME-WIDE DEFINITION OF ENHANCER

Enhancers are distal regulatory elements known to regulate the transcriptional output of their target genes. Enhancer elements have higher cell type-specificity than the genes they regulate. Hence defining enhancers that are specific to a given cell type of interest is a crucial prerequisite for the annotation of non-coding regulatory mutation. A comprehensive definition of enhancers in lung based on functional data would ideally require gathering and analysing all the lung cell types. Despite enormous efforts from large scale collaborative projects such as ENCODE and ROADMAP, the feasibility of the task is nearly impossible. Moreover, due to the lack of knowledge on the exact cell of origin of lung cancer, this becomes an even more challenging task.

To address this issue, we created a repertoire of lung data based on eight different cell and tissue types including lung fibroblasts, epithelial, adenocarcinoma, large cell cancer and primary lung tissue (**Table 10**). Accordingly, we primarily relied on ENCODE consortium data as it (A) covers a broad range of cells and tissue types, (B) produces high-quality data, and (C) analyses the data in an integrative fashion. We obtained the peaks called by ENCODE for H3K27ac ChIP-seq and DNase-seq to define enhancers in six lung cell and tissue types. Further, we also performed H3K27ac ChIP-seq and ATAC-seq in two lung cell lines for the lung repertoire. To our knowledge, this is the first epigenomic profile for active chromatin marks (H3K27ac ChIP-seq and ATAC-seq) of normal bronchial epithelial cells (HBEC).

Table 10: Cell lines used in the definition of enhancers.

Name	Type	H3K27ac ChIP-seq	Accessible peaks (Source)
NCI-H460	large cell lung cancer	In-house	ATAC-seq (In-house)
HBEC-3KT	bronchial epithelial cells	In-house	ATAC-seq (In-house)
A549	lung carcinoma	ENCODE	DNase-seq (ENCODE)
Lung	primary tissue	ENCODE	DNase-seq (ENCODE)
AG04450	lung fibroblast	ENCODE	DNase-seq (ENCODE)
PC9	lung adenocarcinoma	ENCODE	DNase-seq (ENCODE)
Fibroblast	lung primary fibroblast	ENCODE	DNase-seq (ENCODE)
IMR-90	lung fibroblast	ENCODE	DNase-seq (ENCODE)

The average number of cell type-specific enhancers is 49,017 with an average size 402 bps (**Figure 16**). Their pairwise comparison showed an average of 44% similarity (Jaccard Index); this reflects the cell-specific nature of enhancers and the commonness observed at the tissue level. To define a comprehensive list, we considered the union of the cell-specific enhancers resulting in 180,852 enhancers with an average size of 456 bps. The number of enhancers per chromosome ranged between 2,213 in chromosome 21 to 16,710 in chromosome 1, in line with the size of the chromosome and the gene density.

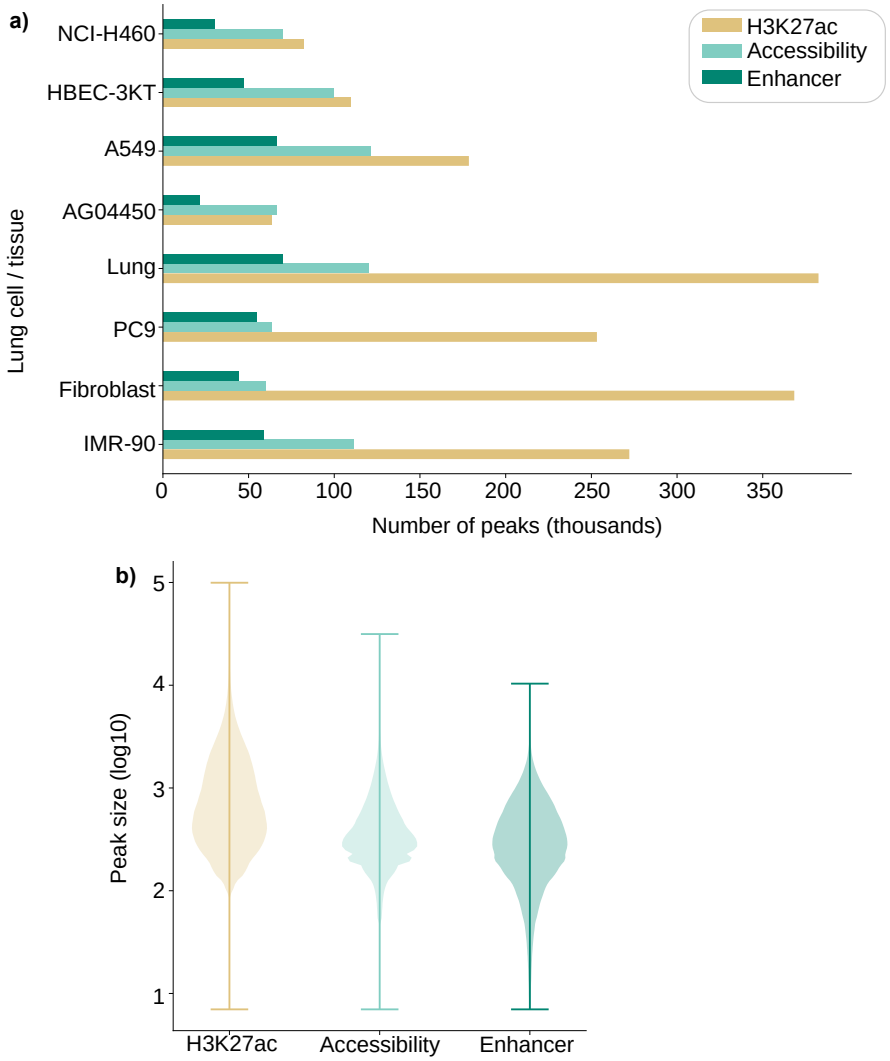


Figure 16: Definition of the lung-specific enhancer catalogue. *(a) Number of cell type-specific enhancer regions (dark cyan) resulting from the intersection of chromatin accessible regions (Light cyan) and H3K27ac ChIP-seq (light yellow) in a selected set of eight lung cell and tissue types. (b) length of the regions is represented as peak sizes using a violin plot for H3K27ac ChIP-seq (light yellow), chromatin accessible regions (light cyan) and enhancers (dark cyan).*

4.2. MUTATION MAPPING

For understanding the effect of non-coding mutations in lung cancer we obtained WGS data from three different lung cancer types. The cohort consists of 55 lung adenocarcinoma (LUAD)[356], 50 lung squamous cell carcinoma (LUSC)[357] and 54 small cell lung cancer

(SCLC)[358] (**Table 11**). Samples in the cohorts were chosen based on the availability of high coverage WGS (>30x) data from paired tumour and normal samples, in addition to RNA-seq data for functional validation (**Appendix Table 1**, **Appendix Figure 1**).

Table 11: Lung cancer sample cohort with the number of samples.

Type	Acronym	Source	Colour code	samples
Lung adenocarcinoma	LUAD	TCGA		55
Lung squamous cell carcinoma	LUSC	TCGA		50
Small Cell Lung Cancer	SCLC	Univ. of cologne		54

Detection of somatic mutations in tumour sequencing data is challenging due to various clinical aspects such as tumour purity, clonal mutation frequency, tumour ploidy, and chemical interference during sample fixation. In addition to these, detection of non-coding mutations poses an extra challenge because of the presence of large repetitive regions. Due to the lack of a single mutation caller that is ideal in all scenarios, we adopted an ensemble approach that combines the results of four complementary callers to balance sensitivity and specificity. We obtained high-confidence somatic SNVs and indels retaining only the ones called by at least two somatic mutation calling tools for each tumour sample against the matched normal (**Figure 17**).

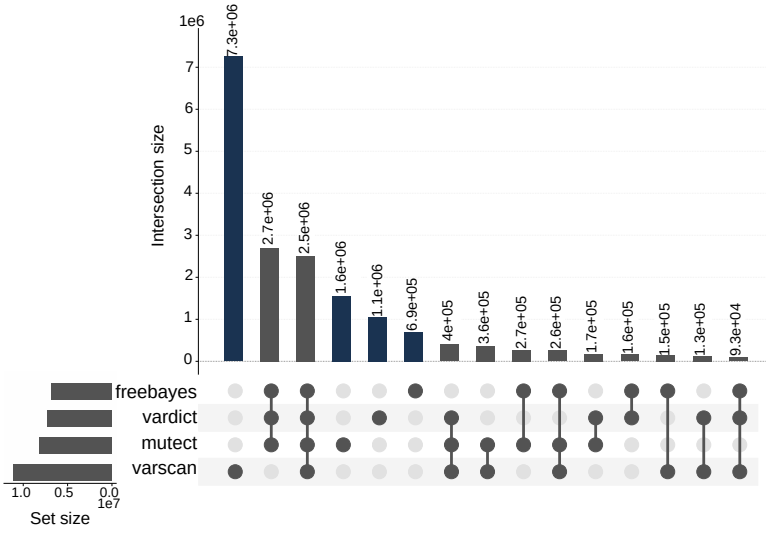


Figure 17: Ensemble mutation calling. *UpsetR plot of the various mutation callers. The left horizontal bars show the number of somatic mutations called by each variant caller considered. The vertical bars show the number of variants in each intersection of sets, specified by dark circles. Variations called by only one tool (Dark blue bars) were removed from further analysis.*

In total, we observed 6,937,213 mutations in the lung cancer cohort, with a number of variations from samples that range from 2,926 to 288,853 (mean = 52,956; median = 48,292). We observed that the regulatory mutations and exon mutations are spread across the genome (**Figure 18**). On average 830 enhancers and 437 promoters are mutated per sample (**Figure 19**).

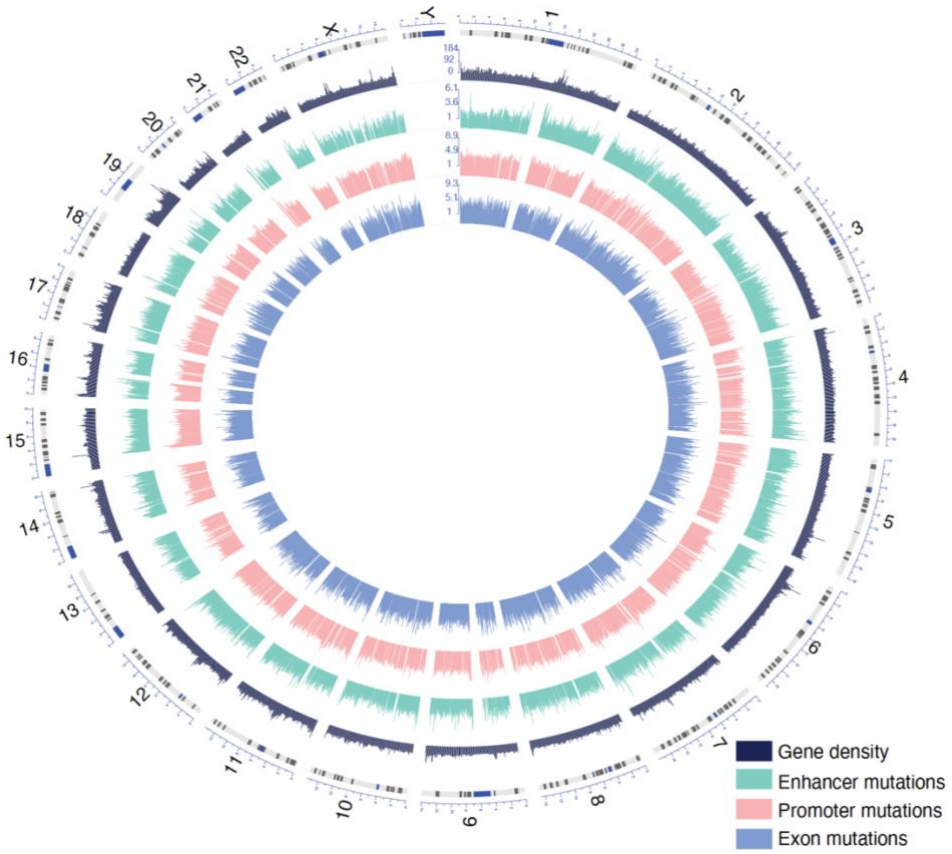


Figure 18: Circos plot of the global landscape of mutations in lung cancer patients. *Chromosomes are shown on the outer most circle. The next circle is a bar graph of gene density obtained by binning the genome in 1Mbp windows. The next circles from periphery to centre are the bar graphs of enhancer (dark cyan), promoter (salmon pink) and exon (powder blue) mutations in log scale. The scale each bar graph is represented at the start of chromosome1. Mutations in non-canonical chromosomes such as chromosome Y was removed from the analysis.*

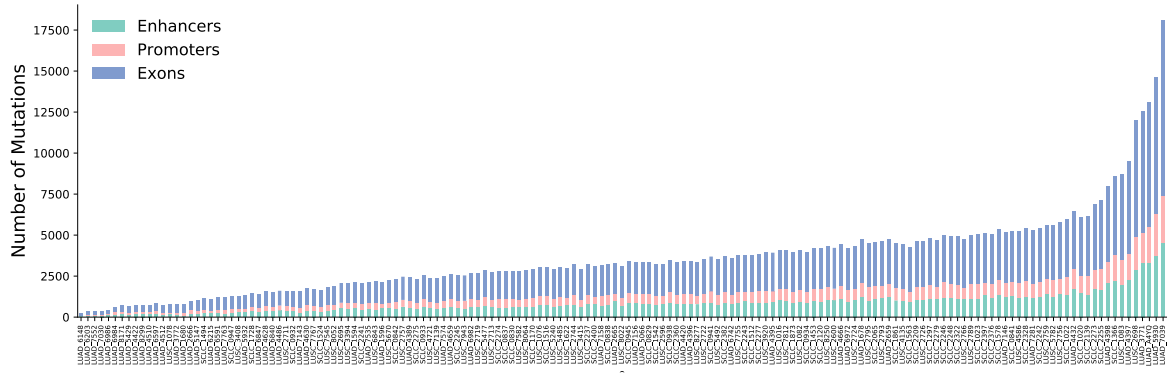


Figure 19: Non-coding regulatory mutations. *Stacked barplot depicting the number of mutations in non-coding regulatory regions. Each bar represents the total number of mutations in exons (powder blue), promoters (salmon pink) and enhancers (dark cyan) for a patient. Samples are sorted based on the total number of mutations in enhancers (x-axis).*

4.3. MUTATION BURDEN

The spectrum of somatic SNVs observed in patient samples hugely varies along the genome. The local somatic mutation burden is influenced by a number of factors such as the histone marks for open and closed chromatin, replication time and gene expression. We hypothesised that genomic regions have varying degree of occurrence of somatic mutations. To test this, we computed the region-specific somatic mutation burden in promoters, exons, enhancers and the non-coding regions. We observed that for all the samples enhancers have a similar mutation burden with respect to exons and promoters (**Figure 20 a** and **b**). Whereas the mutation burden of the non-coding regions was very high compared to the enhancer regions (**Figure 20c**). The similar propensity for mutation burden in regulatory and coding regions of the genome can be suggestive of a functional relevance.

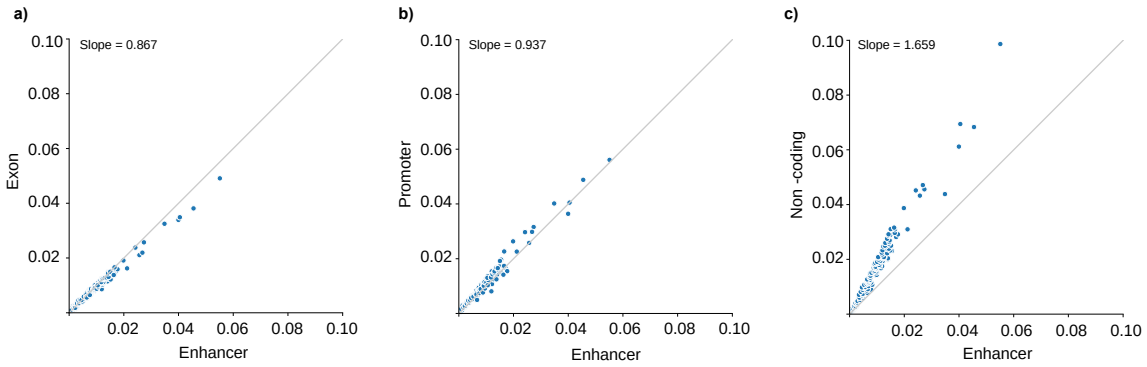


Figure 20: Mutation burden comparisons. *Scatter plots showing the mutation burden comparison (per MB) between enhancers and (a) exons, (b) promoters, (c) non-coding regions devoid of enhancers. Each dot in the plot represents a lung cancer sample. Grey line represents the bisectors. Slope of the regression for each comparison is mentioned in the plot.*

4.4. MUTATION SIGNATURE

The mutational process leaves a specific pattern that can be detected using a mutation signature[220]. To identify the mutational signatures, present in our cohort we computed the mutational profiles from the whole genome sequencing data. We then compared them with the COSMIC single base substitution v2 profile, to identify the prevailing signatures and their relative contribution in each sample (**Figure 21** and **Figure 9**). We observe the most prevalent signature in our cohort is associated with smoking (signature 4), together with three of unknown aetiology (Signature 8, 16 and 5).

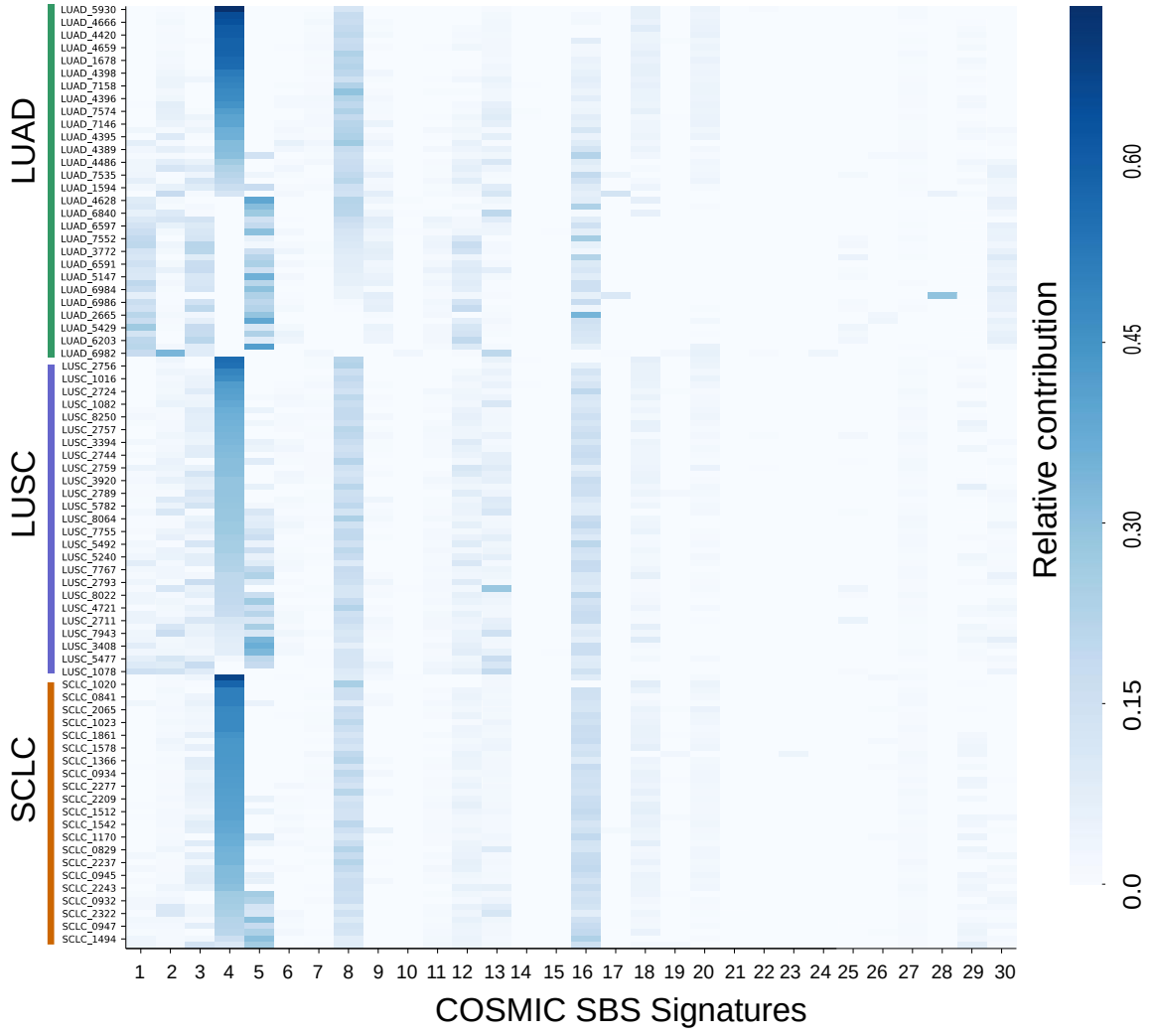


Figure 21: Mutation signatures in lung cancer cohort*. Heatmap of the relative contribution of each COSMIC single base substitutions (SBS) signature for each sample. The samples are grouped based on the lung cancer subtype indicated by the colour band (orange – SCLC; Purple-LUSC; and Green -LUAD). The aetiology of each signature is reported in Figure 5.*

Mutational processes can be unevenly distributed across the genome, as observed through the burden of mutations in different genomic regions. As previously reported, DNA replication machinery associated with transcription or other genomic transactions may have a different impact across distinct regions[359]. Furthermore, the DNA regions bound by

transcription factors may impact the DNA repair mechanism thus leading to altered mutation frequency and modalities[215]. Hence, we hypothesised that the different functional regions of the genome have different mutational patterns. To ascertain this difference, we computed the relative frequency of mutational signatures in the coding and non-coding genome. We identified a significant difference in 18 out of 30 signatures. Among them, signatures 5, 8 and 16 were prevalent in coding regions compared to non-coding regions. On the contrary signatures associated with defective DNA mis match repair (signature 6), likely UV exposure (signature 7), Aflatoxin exposure (signature 24) and APOBEC activity (signature 1) were significantly higher in non-coding regions (**Figure 22**).

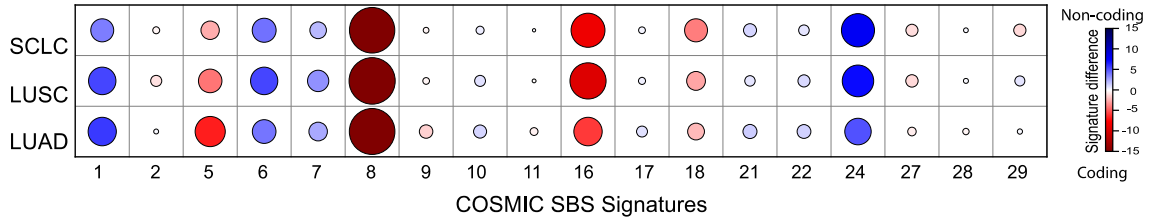


Figure 22: Mutation signature difference in coding and non-coding genome. *Comparison of underlying signature distribution between coding and non-coding regions in LUAD, LUSC and SCLC for a subset of COSMIC SBS signatures. For a given signature, the size of a dot corresponds to the percent increase or decrease in their contribution to describe coding compared to non-coding mutations. Blue and red coloured dots represent non-coding vs coding signature differences, respectively. Only the subset of signatures which had significant contribution differences (p value < 0.05, Wilcoxon rank-sum test) are reported.*

To further understand the difference in the mutational profile in specific functional regions, we compared the relative frequencies of mutation signatures in samples between enhancers, promoters and exons. We observe that 1, 3, 6, 10, 12, 18, 24 and 25 signature profiles are significantly different among all the three regions. Signatures 5, 7 and 16 are significantly different between enhancers and promoters, and promoters and exons (**Figure 23**). Among them, notably the signature associated with defective DNA mismatch repair (signature 6) is higher in promoters compared to enhancer and exons; whereas the signature associated with failure of double strand break repair is higher in enhancer (signature 3) compared to promoters and exons. Signature 3 is also characteristic of insertions and deletions with overlapping microhomology at breakpoint junctions. We also observed a significant difference in the signature associated with activity of error-prone polymerase POLE (signature 10). Even though the mutation burden in the enhancers, promoters and exons are similar (**Figure 20**), the differences in signatures clearly show the different mutagenic processes occurring in different regions of the genome. Also the presence of signatures associated with failures in double strand break repair corroborates with previous reports[360]. Additionally, we also checked for the prevalent signatures in exons, enhancers and

promoters using Sparsesignatures (**Appendix Figure 3**). We observe significantly different patterns in all three genomic regions, although their aetiology was not determined.

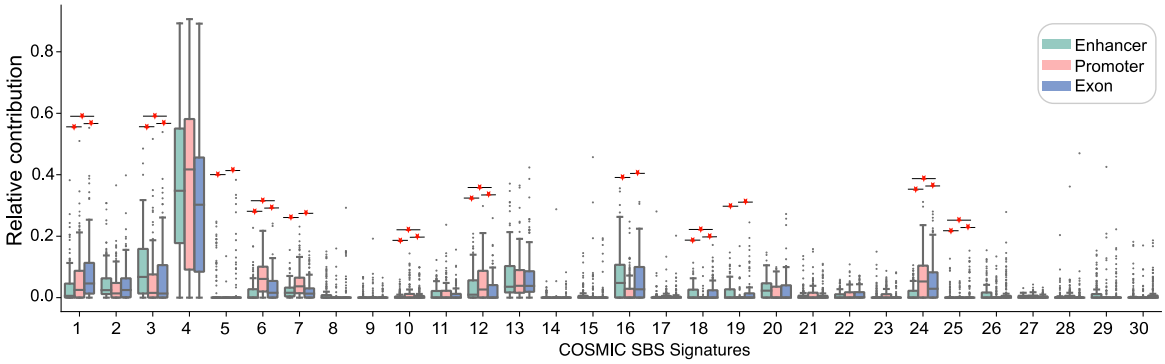


Figure 23: Mutation signature associated with different genomic regions*. Box and whiskers plot of the relative contribution of mutation signature in enhancers (dark cyan), promoters (salmon pink) and exons (powder blue). Statistical significance of comparisons (p-value <0.05) is presented as star marks. Each point of the boxplots represents a sample.*

4.5. ENHANCER TARGET GENE PREDICTION

The study of the role of enhancers is not just limited by their genome-wide identification, but more importantly in identifying their target genes. The enhancer-target gene (ETG) network is characterised by complex many-to-many relationships. Namely, more than one enhancer can regulate a gene and more than one gene can potentially be regulated by an enhancer[24,84]. This non-bijective association is essential for the cell type-specificity of the enhancers thus increasing the complexity of the network. Also, studies have shown that an enhancer does not necessarily regulate its nearest gene[85]. Over the years, a number of computational prediction tools and multiple genomic data have been employed to elucidate this relationship. As reviewed in [361], we found that the information on the three-dimensional architecture of the genome has not been effectively integrated in the prediction of the enhancer-target gene pairs. We adopted the approach developed in our lab[82], that incorporates the 3D genome information in the form of Topologically Associated Domains (TADs) for accurate ETG pairing.

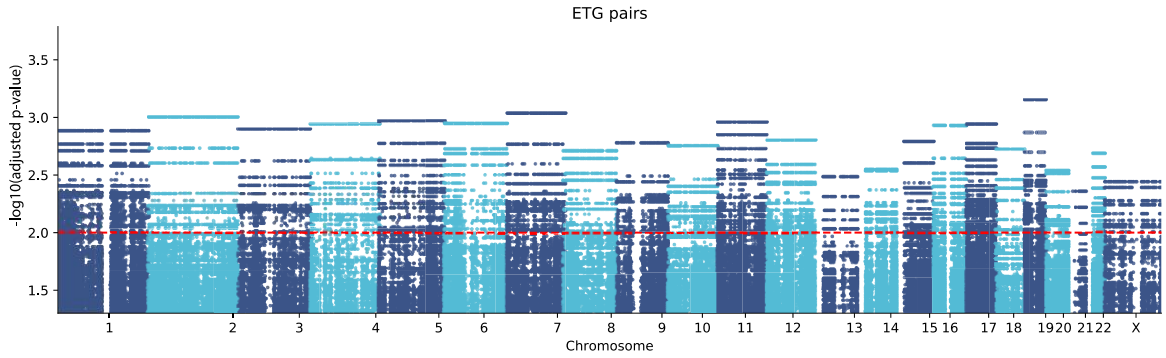


Figure 24: Enhancer-target gene pairs. *Manhattan plot representing all the candidate ETG pairs. Each dot represents an enhancer-target gene grouped by chromosome (x-axis), and its adjusted (AdaPT method) p-values (y-axis), quantifying the strength of their synchronized activity measured across different cell and tissue types. The red line distinguishes the significant pairs (adjusted p-value <=0.01, n= 48,829) from the non-significant ones.*

The inclusion of the information of the hierarchical structure of TADs for the ETG pairing has never been done before. Consistently with the knowledge that more than one enhancer can regulate a promoter, we found a similar scenario with our ETG prediction. We obtained 48,829 enhancer-target gene pairs (adjusted p-value <= 0.01, **Figure 24**). The predictions resulted in 10,709 genes with at least one enhancer. With the inclusion of 3D genome information, we are able to predict enhancer-target gene pairs within 1kb to even 500kb apart (**Figure 25a**). On an average each gene has 5 enhancers associated, while in the median, each enhancer regulates only one gene (**Figure 25b**). The COL1A1 gene has a maximum of 75 enhancers and an enhancer in chromosome 5 is associated with 18 genes.

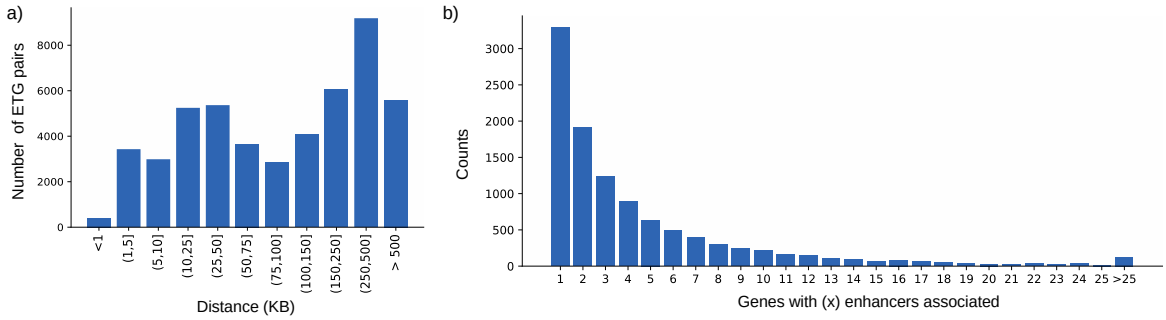


Figure 25: Enhancers target gene pairs. (a) Distance between enhancer and target gene*. X-axis denotes the distance in kb between enhancer and the predicted target gene, y-axis denotes the number of ETG pairs in the distance range.* ***(b)*** *Number of enhancers to a gene. X-axis denotes the number of enhancers associated to a gene, and the y-axis denotes the count of genes with x number of enhancers.*

4.5.1. ENHANCER MUTATION AND ASSOCIATED GENES.

We mapped mutations on the lung specific enhancers previously defined using the epigenomic marks (**Chapter 4.1**). We observed that 10,425 genes had at least one enhancer mutated. The genes with many enhancer mutations were genes with a high number of enhancers. However, the number of genes with at least 25 enhancers mutated are 46 genes

compared to 126 genes associated with at least 25 enhancers. **Figure 26** depicts the number of enhancers associated to a gene, and the number of mutated samples in enhancers of the gene. We observe varying trends of mutation spectrum with respect to enhancers associated i.e., genes with (a) many enhancers associated having more mutations; (b) many enhancers associated having fewer mutations; (c) few enhancers associated having fewer samples mutated and (d) few enhancers associated having more mutations. The opposing trends of (b) and (d) show negative and positive selections in enhancer mutations associated to genes.

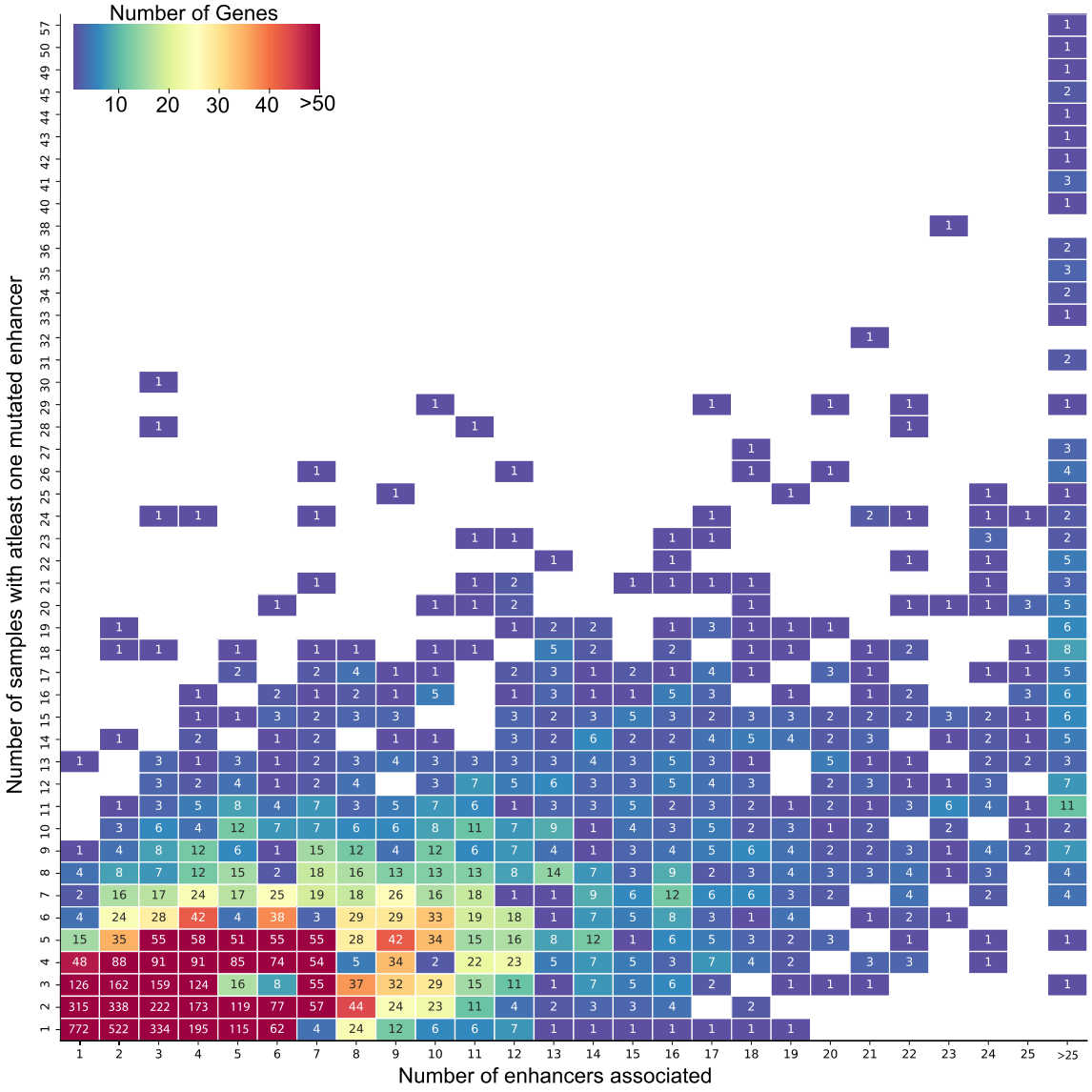


Figure 26: Number of enhancers vs number of mutations. *Heatmap showing the number of enhancers associated with a gene (x-axis) compared to the number of enhancers mutated (y-axis). Colour of the square indicates the number of genes with x number of enhancers and y number of mutated samples.*

As we observed that certain genes had more lung specific enhancers than other genes, we wanted to know if these genes had different expression patterns in lung tissue compared to

other genes with fewer associated enhancers. To estimate this, we selected genes that had at least 25 enhancers (n = 126) and compared their tissue-specific expression with a background set of genes by bootstrapping genes with few enhancers (n ~ 130). We observed that the genes with higher number of enhancers had significantly higher expression in the lung than any other tissue (p value <0.01) **(Figure 27)**. This highlights the importance of the multimodal nature of enhancer – gene regulation for cell type-specific gene expression.

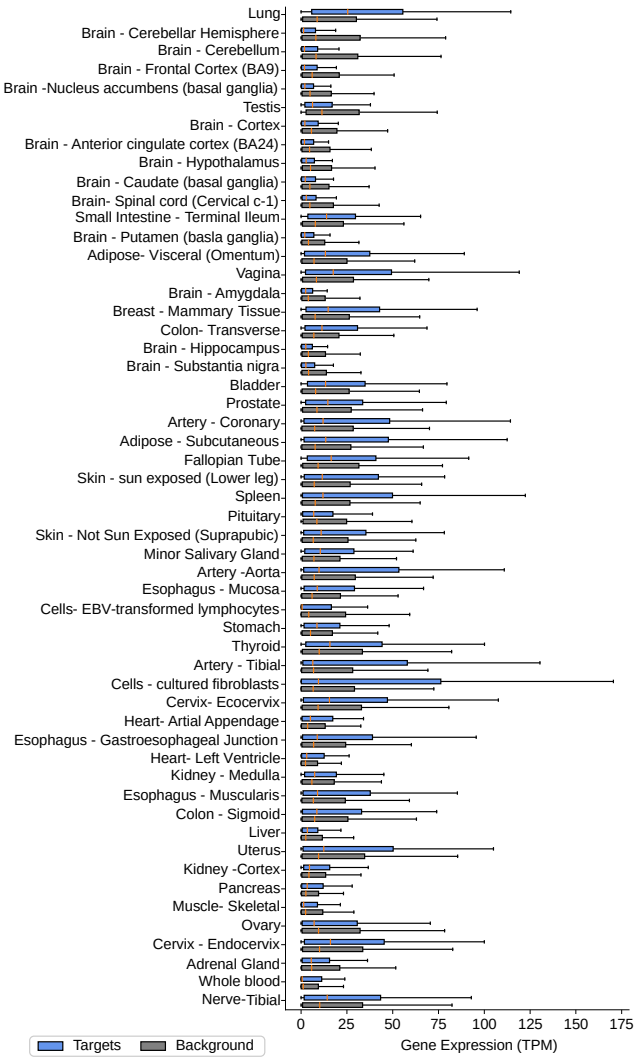


Figure 27: Tissue specific gene expression. *Box plot representing the expression in TPM in GTEX tissue for genes with at least 25 lung specific enhancers (blue) to a set of background genes with fewer enhancers (grey).*

4.6. FUNCTIONAL ANALYSIS

4.6.1. REGULATORY MUTATIONS AND GENE EXPRESSION

Mutations in the coding regions often impact the function of the gene by altering the protein sequence thereby resulting in either a gain of function or loss of function. Whereas mutations in the non-coding regions such as promoters and enhancers often lead to alterations in the expression levels of their target genes[362–364]. To assess the impact of enhancer mutations, we compared the expression level of the target genes in patients stratified as mutated and non-

mutated with reference to enhancer mutations respectively. We observed that 79 genes had a significant impact upon enhancer mutation, of which 4 genes were significantly upregulated (>2-fold increment) and 11 genes with at least two-fold down regulation. (**Figure 28**)

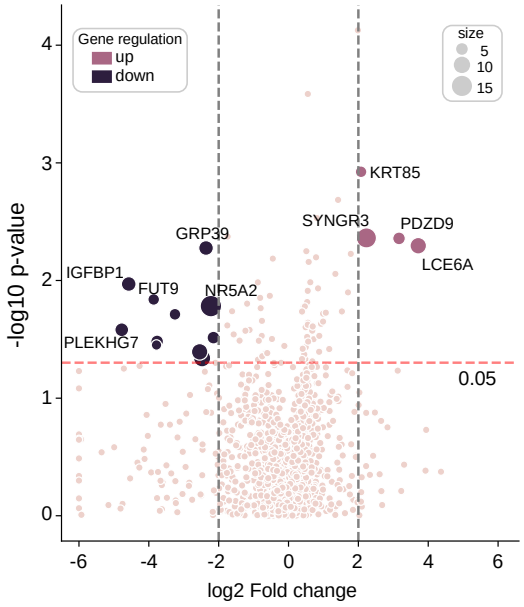


Figure 28: Gene expression changes between genes with enhancer mutations*. Volcano plot displays the log2 fold change in expression in samples stratified individually for a gene with and without enhancer mutations. Transcripts with log2 fold change ≥2 are highlighted in pink and ≤ −2 are highlighted in violet. The red line marks the P ≤ 0.05 value significance. The size of the up and downregulated genes indicates the number of associated enhancers mutated.*

As the activity of enhancers and promoters are synchronised, we also compared the expression of LY6K gene in the mutated and non-mutated samples stratified based on the presence of a mutation in LY6K enhancers and promoter (***Figure 29***). Diseases associated with LY6K include lung, breast and bladder cancers[365–367]. It is a therapeutic target due to its involvement in invasion and metastasis[367]. Higher expression of LY6K gene has been associated with poor overall survival and shorter disease-free in various cancers[368]. In our study cohort, we observe that the presence of enhancer and promoter mutation significantly increases the LY6K expression in patients compared to non-mutated patients.

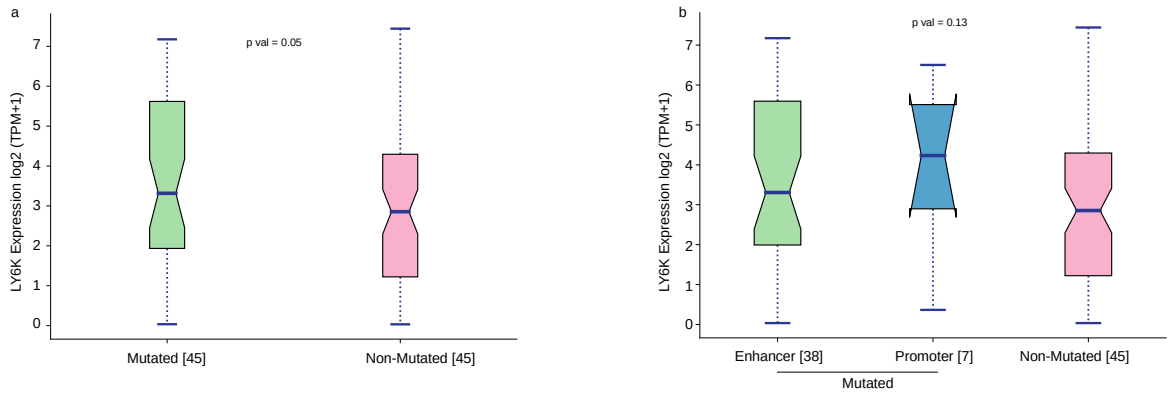


Figure 29: Mutations in regulatory regions affects gene expression. *Box plot show the log2 expression of LY6K gene in (a)mutated and non-mutated samples stratified based on the presence of mutation in regulatory regions of LY6K gene. (b) enhancer mutated, promoter mutated and non-mutated samples. The median is marked with a line across each box. Number of patients in each category is mentioned in square brackets.*

4.6.2. TRANSCRIPTION FACTOR BINDING SITE AT ENHANCERS

Enhancers interact with their cognate promoters and regulate the target gene expression with the help of transcription factors (TF) and mediator proteins. Enhancers serve as operational platforms to recruit TFs through DNA motifs to regulate transcription[369]. The binding affinity of a TF is dependent on its DNA-binding domain and the specific sequence of nucleotides known as transcription factor binding sites (TFBS) or consensus motifs [370]. The typical length of these motifs is 6 to 10 bps[371–374], and a TF protein usually can recognize a set of similar DNA sequences with varying degrees of binding affinity[375,376]. Changes in the motifs at the DNA can alter the affinity or completely hamper the binding of TFs at enhancers[369,377–379].

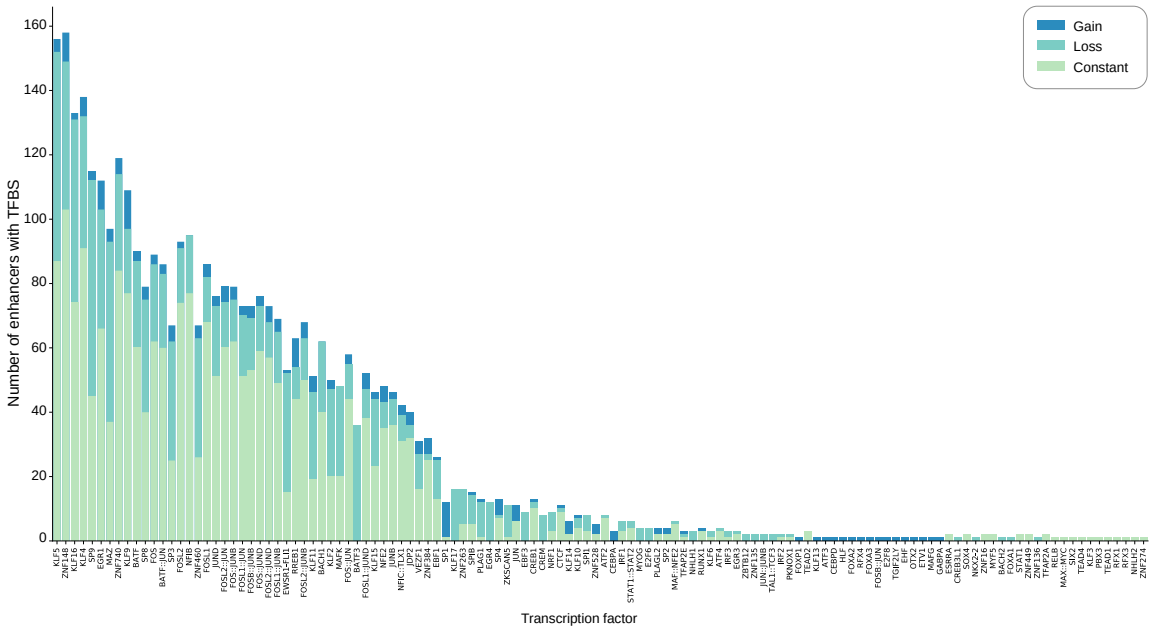


Figure 30: Transcription factor binding sites at enhancers. *Stacked bar plot shows the effect of mutation on the TFBS. Each bar represents the number of enhancers that have gain (dark blue), loss (light blue) and no change (light green) in motif sequence for the given TF (x-axis).*

With the aim of studying the transcription factor binding motifs at enhancers, we leveraged DNase I footprinting information to identify high-resolution TFBS. We call these regions as enhancer cores. To identify the extent of motif alteration at enhancer cores, we first identified the TF motifs at enhancer cores with reference allele. We then identified the motifs in the altered sequence based on the somatic mutations of our patient cohort. Consequently, we compared the TFBS at enhancer cores before and after mutation and observed that the changes in the sequence of enhancer core, did result in both gain and loss of TFBS (**Figure 30Error! Reference source not found.**). Although, several of the sites did not have a change in the motifs. Moreover, we observe a higher number of TFBS loss compared to gain of motifs **(Figure 31)**.

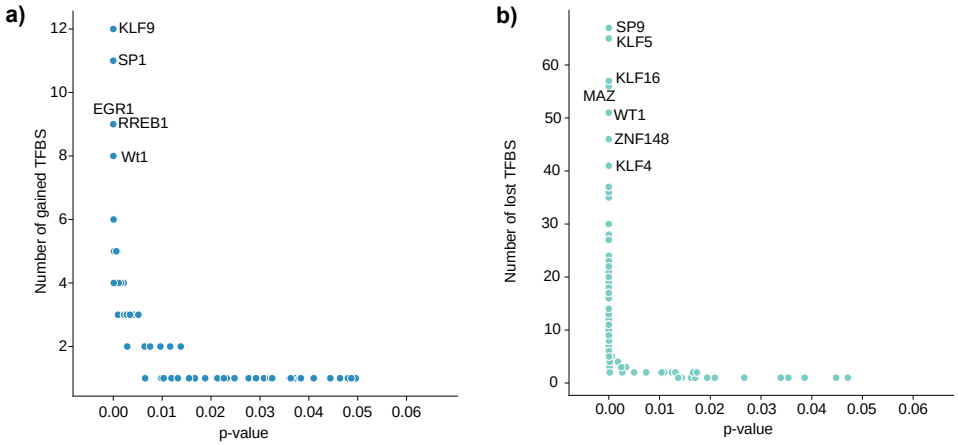


Figure 31: Gain and loss of transcription factor motifs. *Scatter plot shows the (a)gain and (b) loss of motifs. s. Each dot represents a TF, the y-axis represents the number of enhancers with the predicted motifs of that particular TF, and the x-axis represents the significance of the motif computed based on position-specific scoring matrices using FIMO.*

4.6.3. RECURRENCE OF ENHANCER MUTATIONS

Recurrence of mutation has proven to be a powerful tool for the identification of new cancer genes[380]. With the aim to annotate the biological relevance of enhancer mutations, we explored the recurrently mutated enhancers with the same base alterations across multiple samples. We observed a peculiar mutation in an enhancer of CDH13 gene (**Figure 32 a**). CDH13 has known tumour suppressor activity and its down-regulation has been associated with poor prognosis in various carcinomas namely lung, ovarian, cervical and prostate cancer[381].

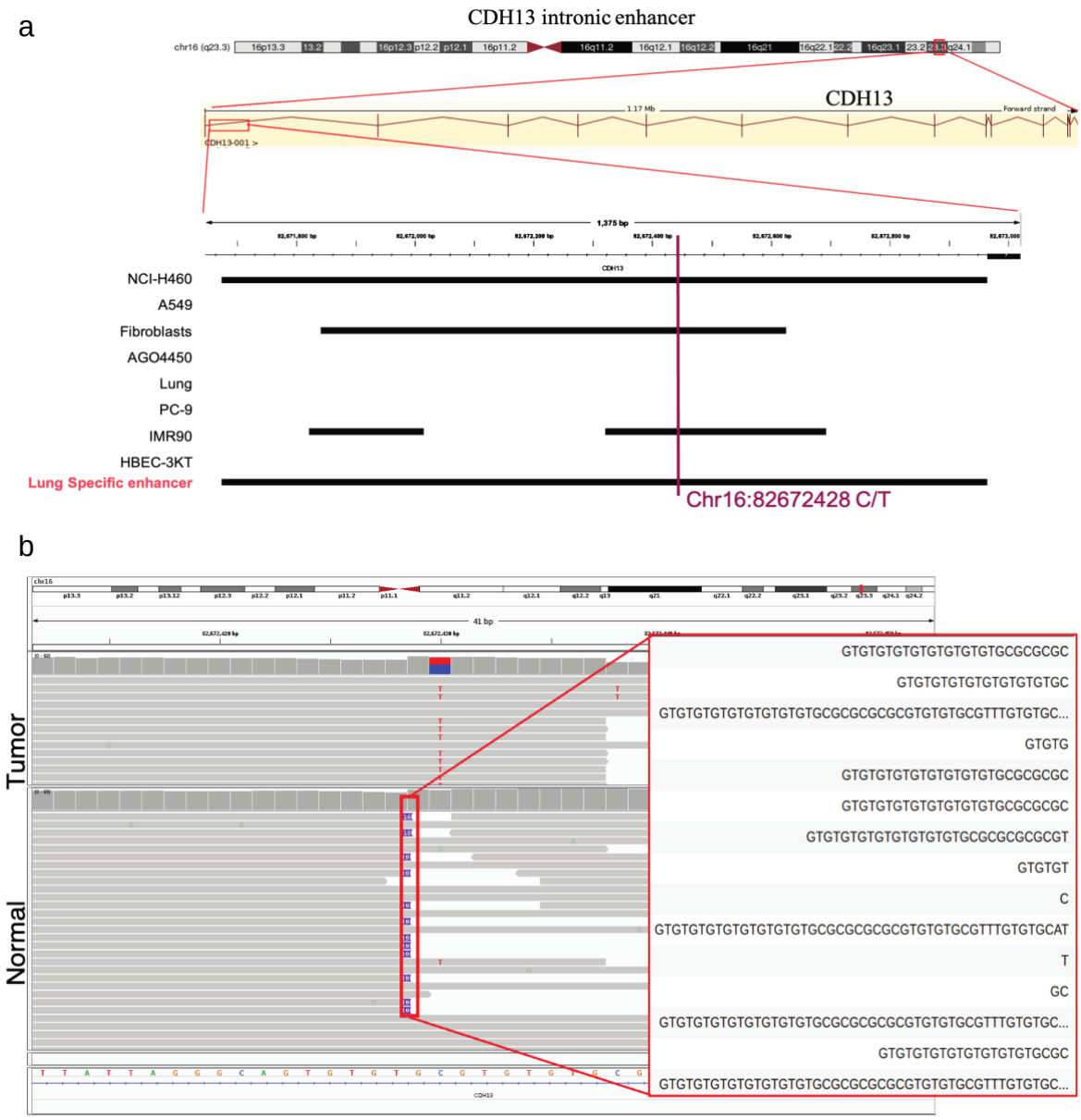


Figure 32: CDH13 insertion variation. *(a) CDH13 enhancer loci at chromosome 16, in the first intron of the gene. The black horizontal lines indicate the region of open chromatin determined by H3K27ac and chromatin accessibility data in respective cell lines. The vertical maroon line indicates the SNP at position 82672428 in the intronic enhancer. (b) Snapshot of IGV viewer at enhancer loci with whole genome sequencing data of a tumour tissue and matched normal of an individual. Each grey line indicates the reads from whole-genome sequence data corresponding to the region. The top panel shows the WGS of the tumour tissue of a patient with the SNP marked with red and blue square. The insertion of interest in the matched normal [purple square] and its sequence in the individual reads aligned to the region is highlighted in red box.*

The somatic mutation of interest is present in the enhancer located in the first intron of CDH13 gene. This intronic mutation is annotated as a somatic SNV conversion from C to T at locus chr16:82672430 (hg19). However, upon closer inspection, we noted that in the matched normal of the same patient, we observe a germline insertion with respect to the reference genome assembly at locus 8,267,2428. The sequence of insertion is a stretch of GT with a tail of GC bases (**Figure 32 b**). Due to the complexity of the insertion sequence, in most of the samples, the region was removed as a putative sequencing artefact. However, upon careful reprocessing of the samples individually, we observe the presence of this

mutation in different combinations in different patients. A) Presence of insertion in tumour and normal samples B) presence in tumour samples and absence in their matched normal, C) *vice versa* (**Figure 33**).

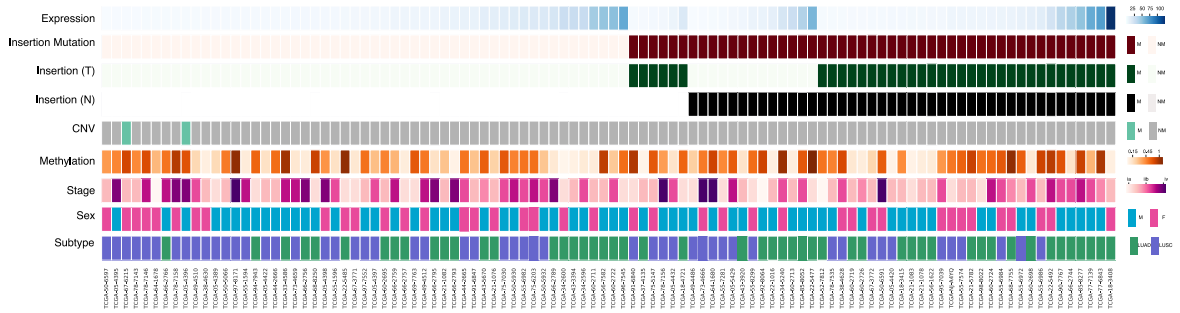


Figure 33: CDH13 insertion variation and patient clinical information. *Co-mutation plot shows CDH13 expression (TPM), CDH13 enhancer insertion variant, presence of insertion in tumour tissue, presence of insertion in matched normal, copy number alteration, promoter methylation, TNM staging of the cancer, sex of the patient and the lung cancer subtype are represented by indicated colours.*

In order to perform the experimental validation on the candidate enhancer, we screened lung cancer cell lines for a) Sequence of the enhancer loci b) expression of CDH13 gene c) Copy number alteration of CDH13 gene. We also observed the presence of insertion in CDH13 enhancers in lung cancer cell lines **(Figure 34 a)**. The expression of CDH13 gene was quantified for all the isoforms of the gene in 10 lung cancer cell lines in comparison to WI38 (normal lung fibroblast cell line and BJ (normal skin fibroblast cell line). We observed that cell lines NCI-H460 and MSTO-21H had CDH13 expression comparable to the normal cell lines (**Figure 34 b**). Upon quantification of copy number of CDH13 gene, through relative qPCR, we observe that NCI-H460 had a diploid copy number (**Figure 34 c**). Based on these assessments, we chose NCI-H460 as our cell line of choice for the experimental validation of CDH13 gene.

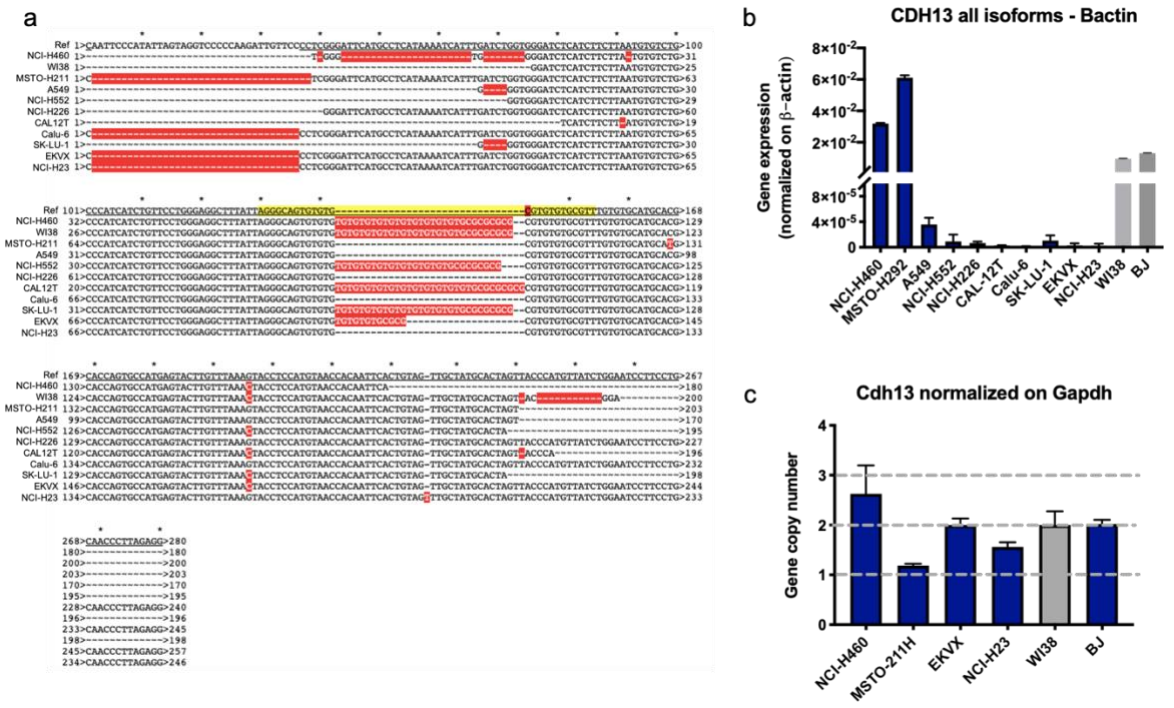


Figure 34: Cell line characterization for experimental validation. *a) Sanger sequencing results of CDH13 enhancer in lung cancer cell lines. The first row shows the reference sequence, followed by the lung cell lines viz., NCI-h460, WI38, MSTO-H211, A549, NCI-H552, NCI-H226, CAL-12T, Calu-6, SK-LU-1, EKVX and NCI-H23. Each group has a stretch of ~100bps, corresponding to a total of 280bps.* ***b)*** *Bar-plot representing the expression of CDH13 gene normalised to beta-actin in lung cancer (blue) and normal cell lines (grey).* ***c)*** *Copy-number of CDH13 determined by PCR in lung cancer cells with respect to GAPDH. The grey lines indicate the ploidy of the cells for the gene.*

To understand the role of the CDH13 intronic enhancer with such complexity of a structural variation and a SNP, we employed CRISPR based deletion of the enhancer in NCI-H460 lung cancer cell line. Following which the clones with deletion was assessed using sanger sequencing to determine the region of deletion. We then quantified the expression of the CDH13 gene in the clones with the homozygous deletion of the enhancer region. We observe that upon the homozygous deletion of the enhancer CDH13 gene is significantly downregulated (**Figure 35**).

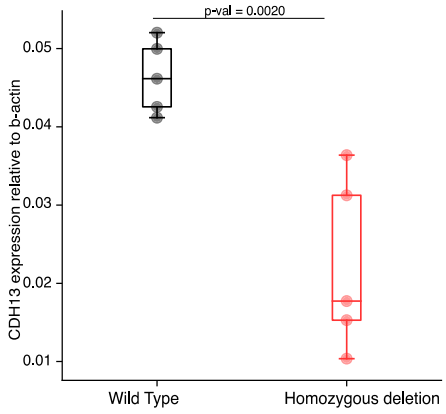


Figure 35: CDH13 expression upon enhancer deletion. *CDH13 gene expression relative to beta actin in wild type and homozygous deletion of enhancer in NCI-H460 cell line. The dots represent biological replicates (n = 5).*

The deletion of the enhancer and corresponding downregulation of CDH13 confirmed that the locus is an enhancer. To understand the role of the insertion sequence variant, we compared the CDH13 gene expression in patients with and without insertion. We observed that the presence of insertion sequence variant resulted in a higher expression compared to the latter, not -significant (**Figure 36**).

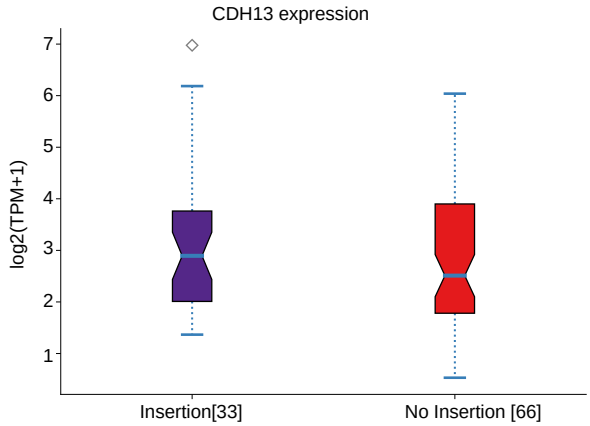


Figure 36: CDH13 expression in samples with enhancer insertion mutation. *Box plot shows the expression of CDH13 gene as transcripts per million (log scale) in lung cancer samples stratified based on the presence or absence of insertion. Median expression is marked with a line across each box. Number of patients in the categories are represented in square brackets.*

We further evaluated the effect of the insertion sequence variant on the survival probabilities. We observed that the patients with the insertion mutation, had better progression free-survival (**Figure 37 a- c**), however not significant. Similarly, we also observe the disease-free survival to be better in patients with insertion than those without (**Figure 37 d** and **f**). Whereas, when considering the disease-free survival, we observed, that in a small cohort of lung squamous cell carcinoma (based on data availability), the patients with insertion sequence variant in the CDH13 enhancer had a hundred percent disease-free survival probability (**Figure 37 e**).

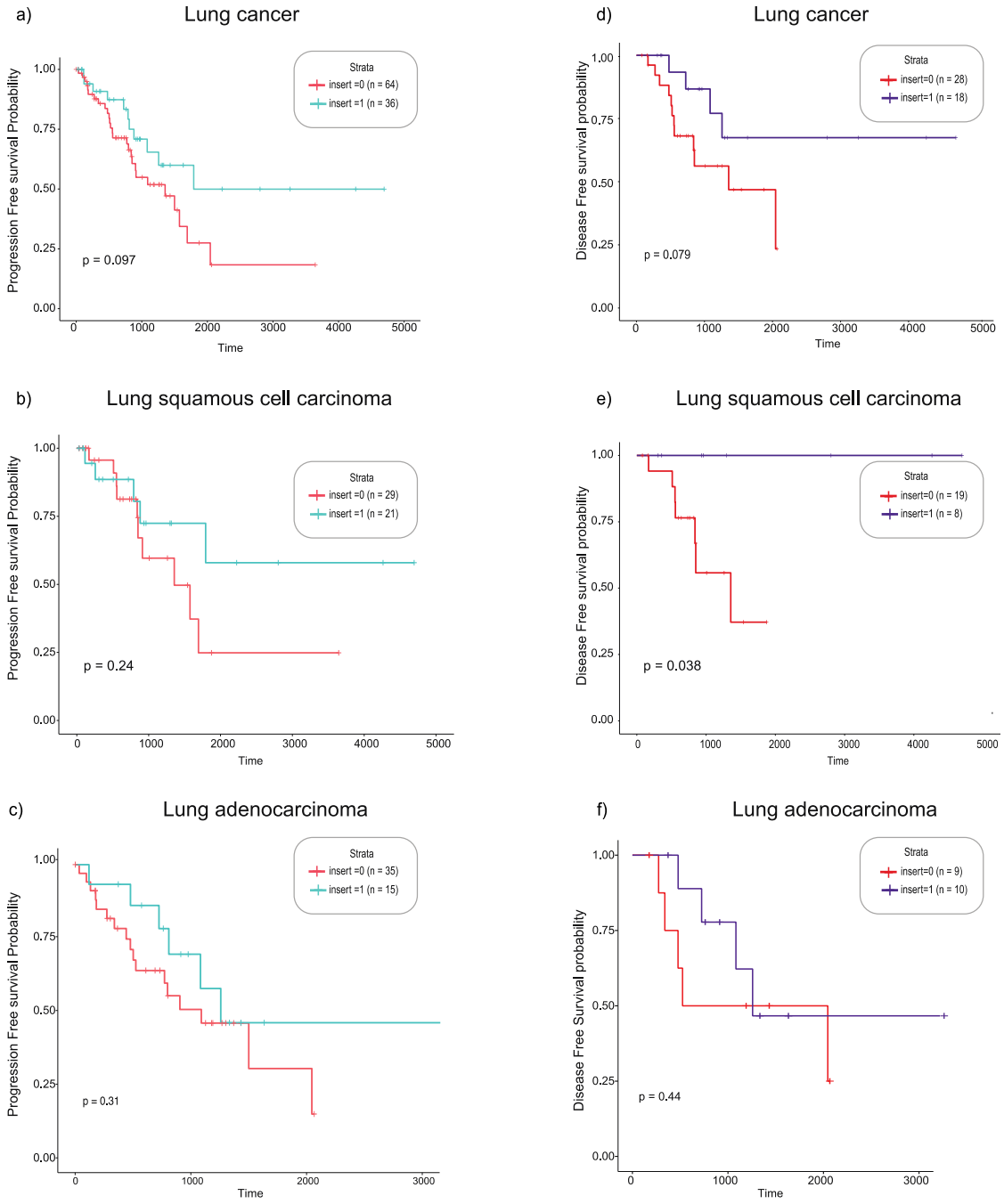


***Figure 37**: **Survival probabilities.** Kaplan Meier Curves depicting the progression-free survival interval (PFI) probability in (a) all lung patients (LUAD+LUSC) (b) LUSC (c) LUAD and disease-free survival interval (DFI) probability in (d) all lung patients (LUAD+LUSC) (e) LUSC (f) LUAD. Patients stratified based on the presence of insertion sequence variant in CDH13 enhancer. For PFI: cyan – present, orange – absent and for DFI: purple– present, Red – absent. Differences between two groups were evaluated using a log-rank test.*

We further examined the transcription factor motif alteration at CDH13 enhancer with and without insertion mutation. CDH13 enhancer core of interest houses three motifs for the transcription factors: EGR1, KLF9 and ZSCAN4 (**Figure 38 a**). Insertion mutation at the locus results in the creation of seven new motifs in addition to the previous motifs. The new motifs that were created include HES1, HES2, ZBTB14, EGR4, TCFL5, NRF1 and RREB1.

We also observe that several of the motifs are present at more than once within the enhancer core (**Figure 38 b**). Sequence of the motifs are represented in **Figure 38 c**.

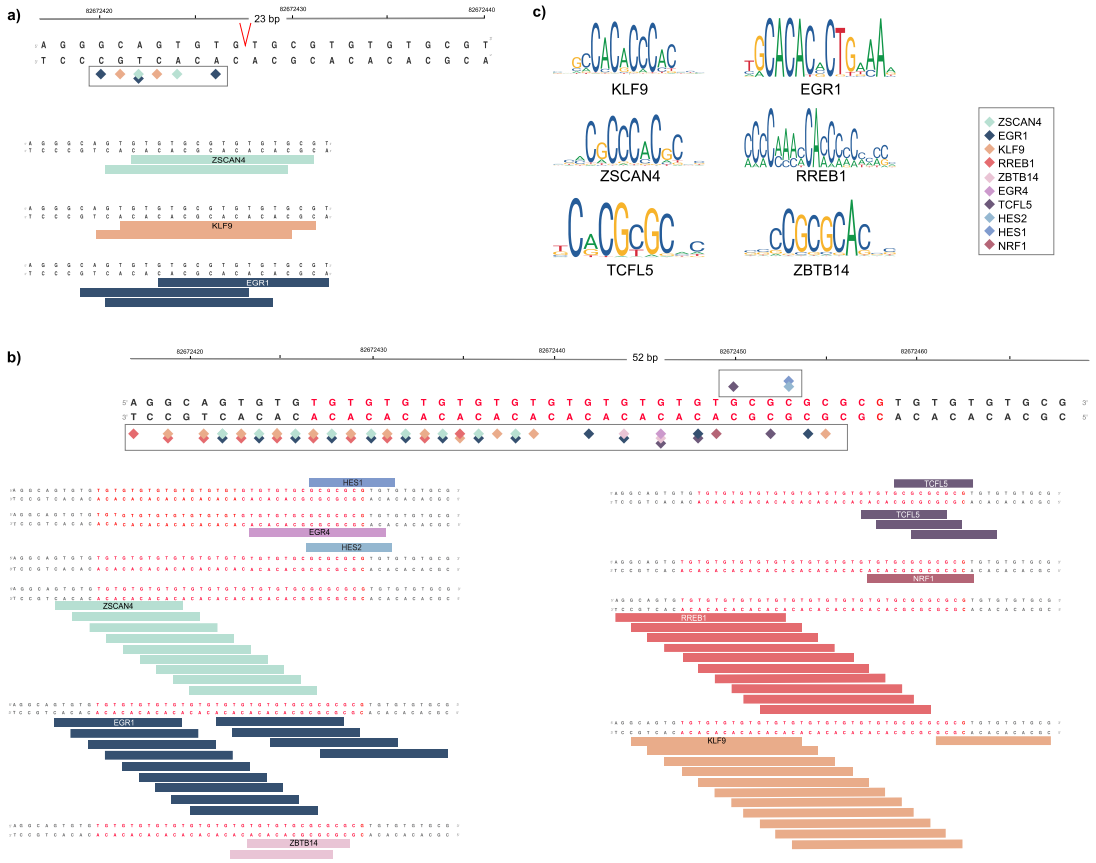


Figure 38: Transcription factor motif alteration at CDH13 enhancer. (a) *Figure representing the TF motifs observed at the CDH13 enhancer with reference sequence. Red V in the reference sequence represents the location of insertion.* ***(b)*** *TF motifs present at CDH13 enhancer with insertion mutation (red section within the sequence). Summary of the various motifs are represented using small color-coded squares beneath the top sequence in (a) and (b). The exact match of the motifs is represented for individual transcription factors below.* ***(c)*** *Sequence logo of the transcription factors motifs with more than one binding site.*

We further explored, if these transcription factors are expressed in our lung cancer patient cohort to enable successful binding and observed that except for EGR4 and ZSCAN4 whose motifs are present in both the reference and altered allele sequence, the rest of the transcription factors were expressed (median TPM >1) in the lung cancer cohort (**Figure 39**).

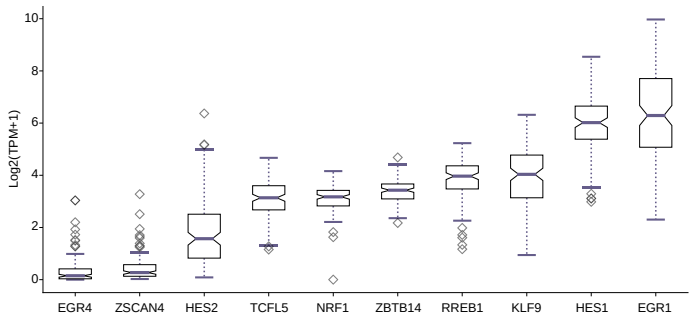


Figure 39: Expression of transcription factors with predicted TFBS in CDH13 enhancer. *Box plot showing the expression of the predicted transcription factors as transcripts per million (log scale) in lung cancer cohort. The median is marked with a line across each box.*

As an independent control, we explored if this phenomenon was also observed in breast cancer (a cohort of TCGA high coverage WGS samples n= 112) wherein CDH13 is reported to be downregulated, and found that only 9% of the breast cancer samples had the insertion sequence variant in tumour or normal tissue (**Figure 40 a**), in contrast to 45% in lung cancer samples. We also explored the effect of the insertion mutation in the survival of the breast cancer samples with and without enhancer insertion, and found poor progression free (**Figure 40 b**) and disease free-survival (**Figure 40 c**)

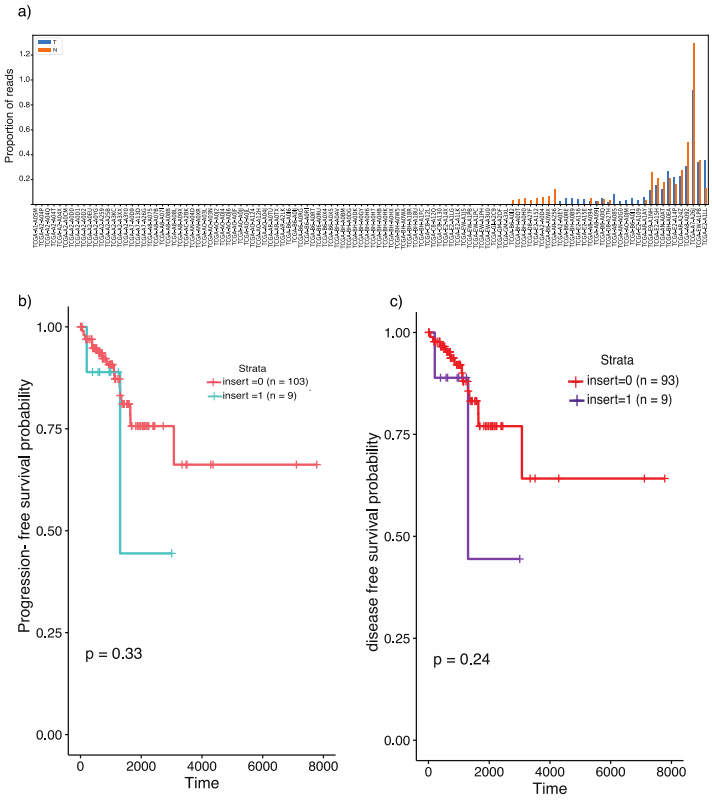


Figure 40: Breast cancer- CDH13 insertion analysis. a) *Bar-plot showing the proportion of reads corresponding to CDH13 insertion mutation in tumour (blue) and normal (orange) WGS data of breast cancer samples. Kaplan Meier Curves depicting the (**b**) progression-free survival interval (PFI) probability and (**c**) disease-free survival interval (DFI) probability in breast cancer samples.*

4.6.4. PATHWAY LEVEL AGGREGATION OF ENHANCER MUTATIONS

Genes work in coalitions and their activities depend on and/or impact on each other. Genes are co-expressed, co-regulated and they co-operate with each other. Furthermore, the enhancer gene regulatory network is non-bijective, that is an enhancer may regulate more than one gene and a gene can have more than one enhancer associated with them. The complex enhancer- target gene association is represented in **Figure 41**.

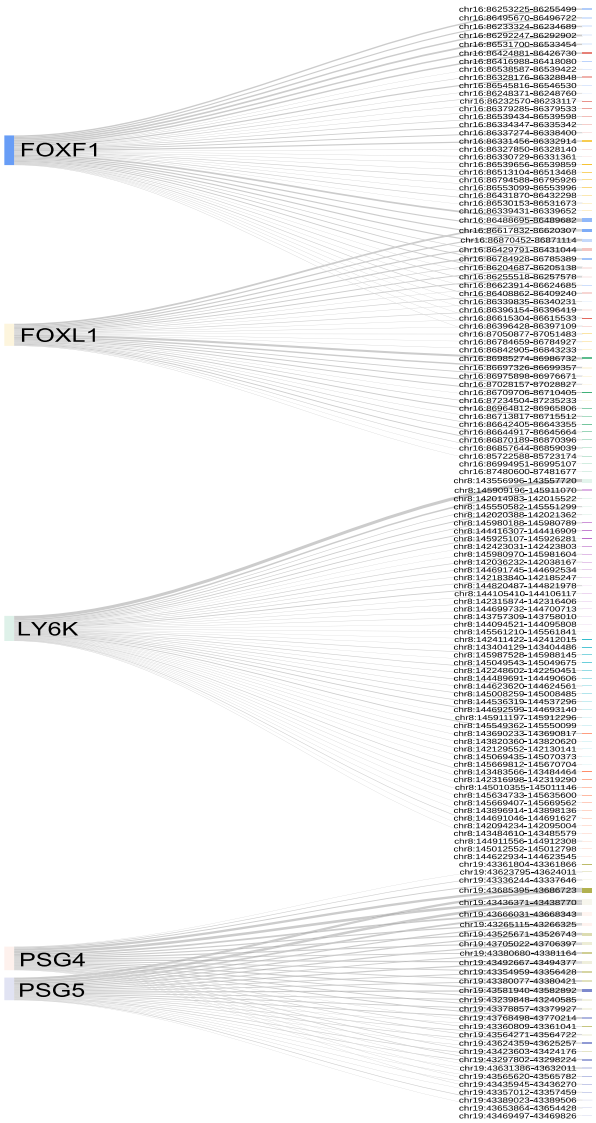


Figure 41: Multimodal enhancer gene association. *Sankey plot showing the mutated enhancers and the predicted target gene, the thickness of the line indicates the number of samples with mutation in the enhancer.*

The best way to understand the role of multiple interacting genes is to study biological pathways that they belong to. Hence, we performed an over-representation analysis of target genes with enhancer mutations to identify pathways that are enriched with these genes. We

found key cancer pathways to be significantly enriched (**Figure 42**), including, PI3K-AKT signalling pathway, focal adhesion and regulation of actin cytoskeleton pathway.

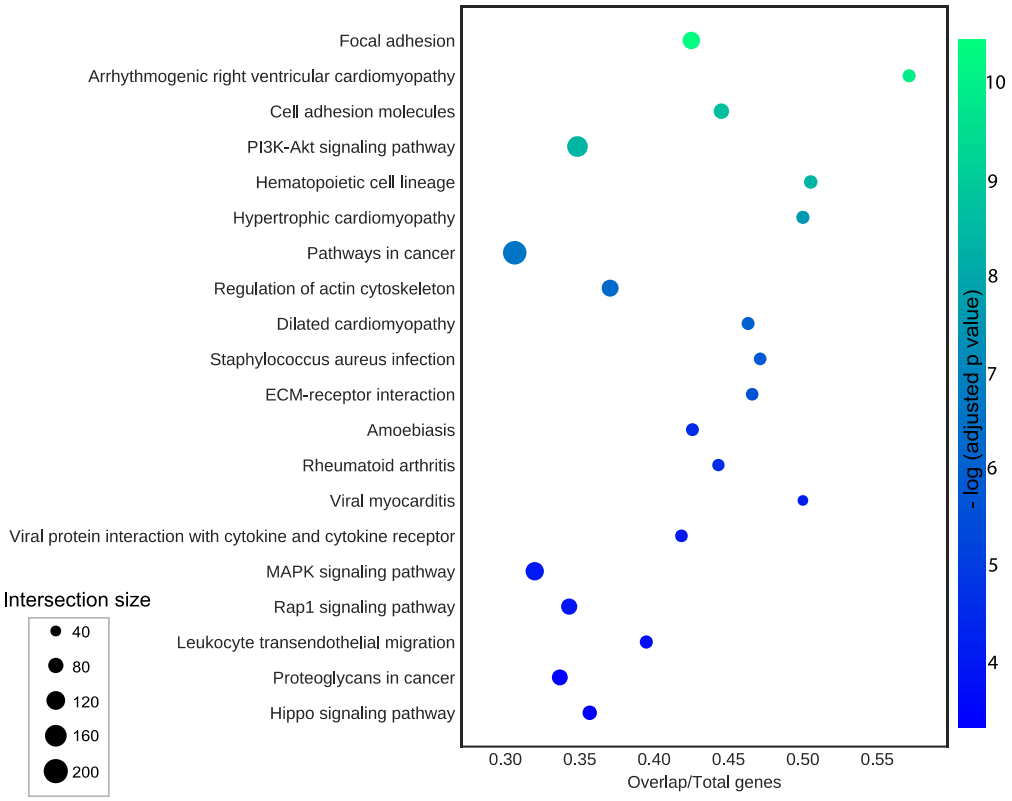


***Figure 42: Pathway level enrichment of enhancer mutations**. Scatter plot shows the over-representation of genes with enhancer mutations in KEGG pathway. X axis represents the ratio of the overlapping genes to total number of genes in the pathway. The size of the circle denotes the number of genes in overlap and the colour shows the negative logarithmic adjusted p-value.*

Thus, we further explored the mutational landscape of PI3K-AKT signalling pathway, as it is one of the cancer driver pathways and have been leveraged for therapeutic targets. We observed that enhancer mutations, promoter mutations and the exon mutations in PI3K-AKT pathway genes have a complementary behaviour in patients, *i.e.*, an individual gene of the pathway is targeted by either of the three categories of mutation in a patient and a combination of two or all three mutations was not observed (**Figure 43**).

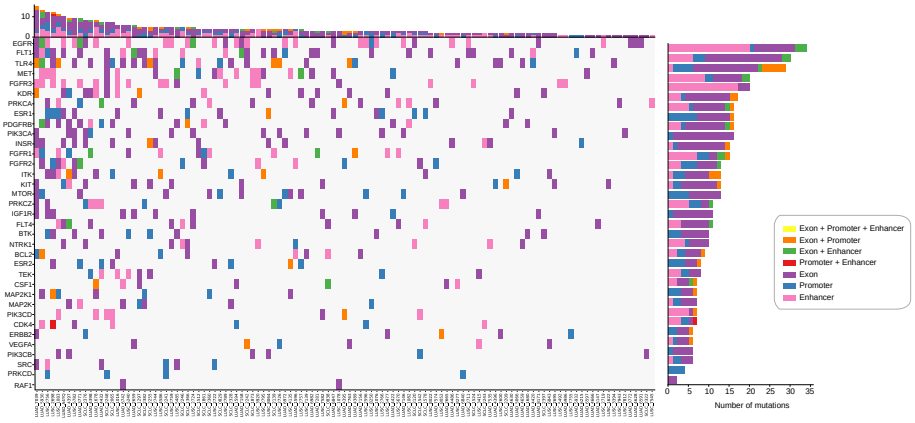


Figure 43: Mutational landscape of PI3K-AKT pathway. *Co-mutation plot showing druggable PI3k-AKT signalling pathway genes (y-axis) affected in lung cancer samples (x axis) by mutations in enhancer (pink), promoter (blue), exon (purple), promoter and enhancer (red), exon and enhancer (green), exon and promoter (orange), exon, promoter and enhancer (yellow). The top stacked bar plot shows the number of mutations in each sample and the gene wise mutations rate is displayed on the right.*

Additionally, we explored if the genes with enhancer mutations have a significant overlap with other gene sets that have been curated from literature. To this aim, we performed the gene set enrichment with curated gene sets from MSigDB database. We observed significant overlap with gene sets associated with various invasive tumours, stemness, extracellular matrix organisation and focal adhesion (**Figure 44**).

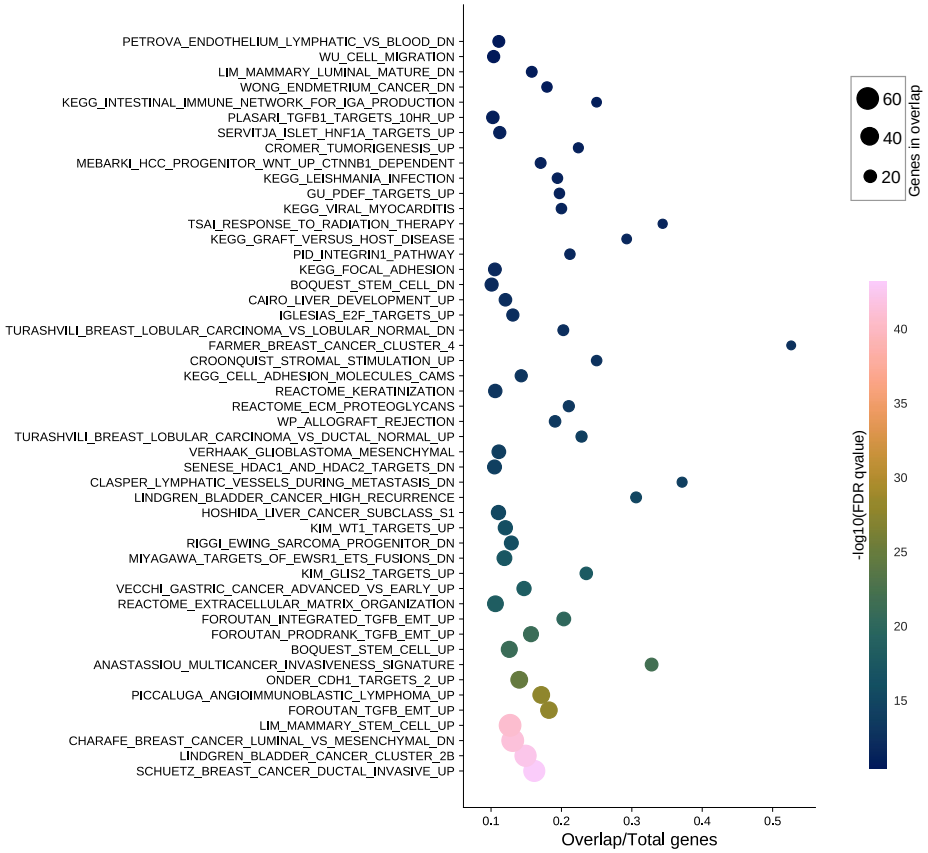


Figure 44: Gene set enrichment analysis of genes with enhancer mutations. *Scatter plot shows the genset enrichment of genes with enhancer mutations in MSigDB C2 curated gene sets($p<0.0001$). X axis represents the ratio of the overlapping genes to total number of genes in the gene sets. The size of the circle denotes the number of genes in overlap and the colour shows the negative logarithmic adjusted p-value.*

As we observed several cancer related gene sets in the enrichment analysis, we further explored specifically the overlap with the oncogenic gene sets in MSigDB (**Figure 45**). We observed numerous gene sets associated with perturbations in PCGF2, KRAS, RAF1, MAPK and TP5.

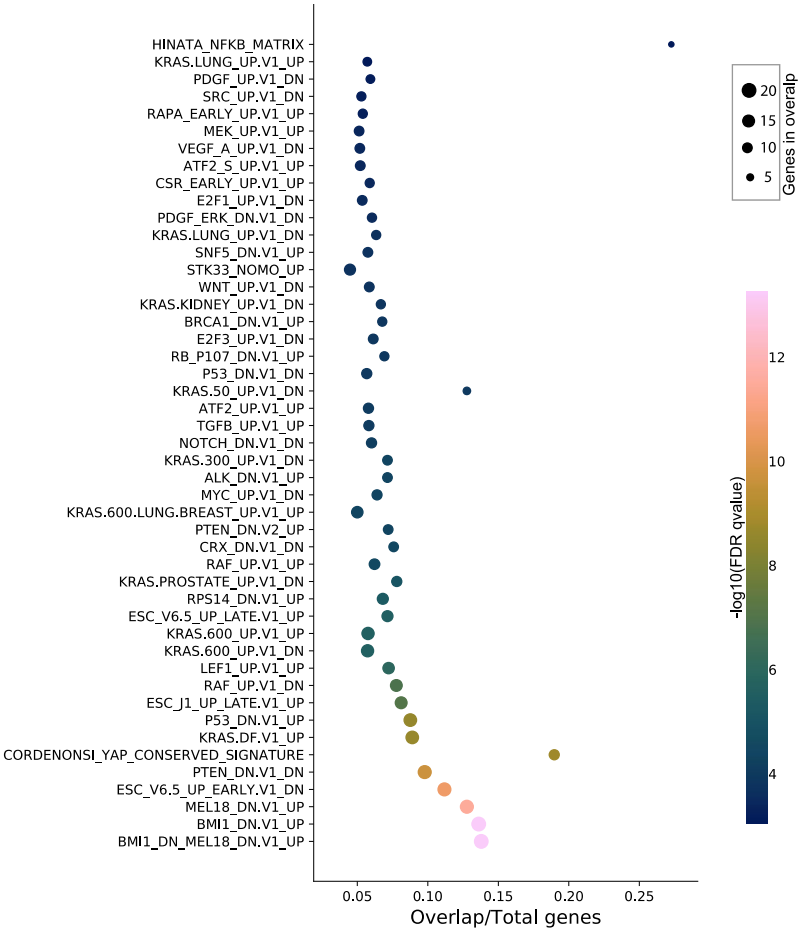


Figure 45: Gene set enrichment analysis (Oncogenic – gene sets). *Scatter plot shows the genset enrichment of genes with enhancer mutations in MSigDB C6 oncogenic gene sets (p<0.01). X axis represents the ratio of the overlapping genes to total number of genes in the gene sets. The size of the circle denotes the number of genes in overlap and the colour shows the negative logarithmic adjusted p-value.*

Gene ontology enrichment revealed that the genes with enhancer mutations were found to be mainly involved in the RNA polymerase II transcription factor activity and other DNA binding related molecular functions (**Figure 46**).

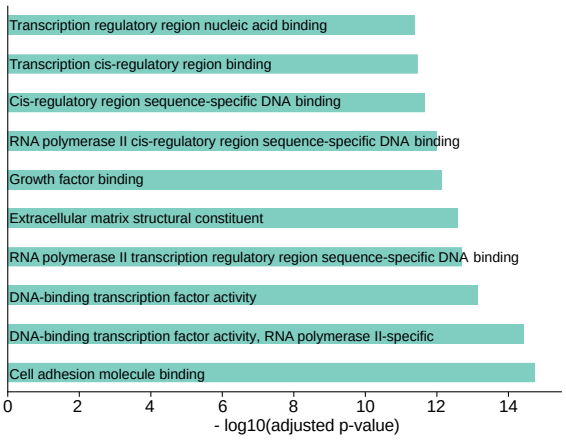


***Figure 46**: **Molecular function of genes with enhancer mutation.** Enriched Gene Ontology (GO) molecular function terms for the target genes associated with the mutated enhancers*

We observed that the genes with mutated enhancers converge on biological processes such as positive regulation of kinase activity, regulation of protein phosphorylation, positive regulation of MAPK cascade, mononuclear cell differentiation, negative regulation of transcription by RNA pol II and angiogenesis (**Figure 47**).

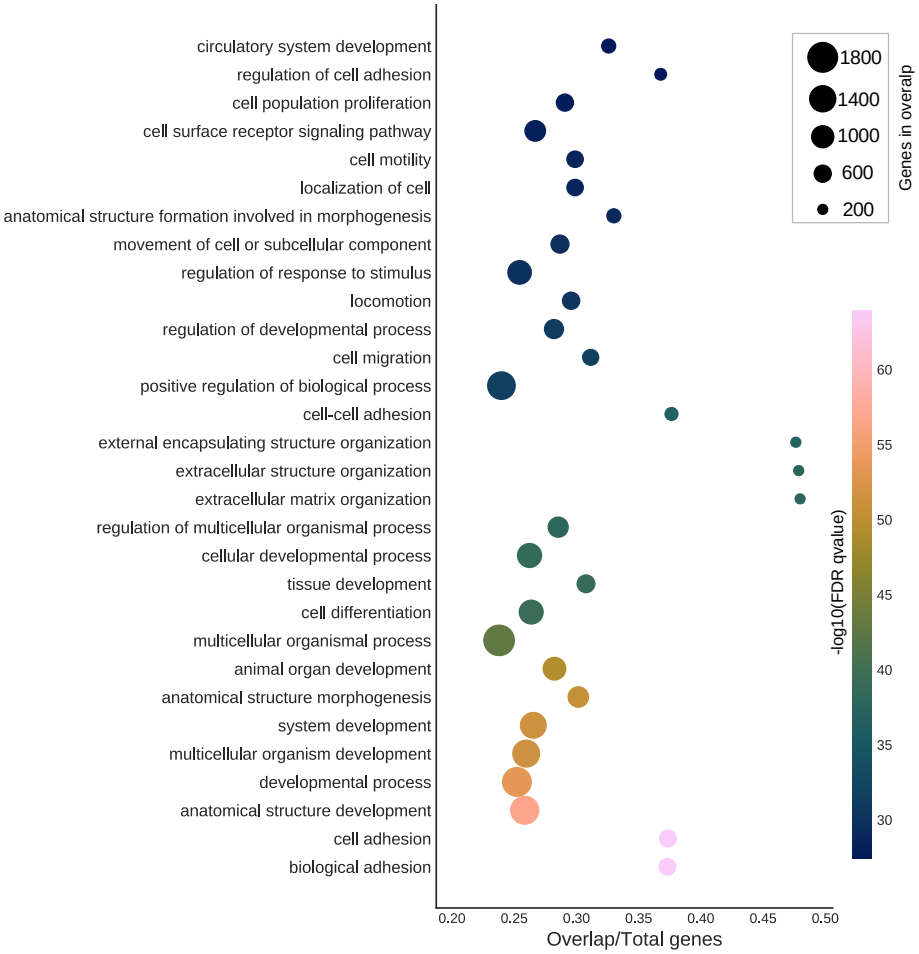


Figure 47: Gene Ontology: Biological Process enrichment analysis of genes with mutated enhancers. *Scatter plot shows the genset enrichment of genes with enhancer mutations in Gene Ontology biological process. X axis represents the ratio of the overlapping genes to total number of genes in the gene sets. The size of the circle denotes the number of genes in overlap and the colour shows the negative logarithmic adjusted p-value.*

5. DISCUSSION

The coding genome has been extensively studied in cancer to identify potential driver mutations and therapeutic targets. Despite these enormous efforts, there is a sizeable gap with respect to the prognosis and patient stratification for better treatment opportunities. We hypothesised that non-coding mutations in regulatory regions such as enhancers and promoters could significantly contribute to cancer prognosis or predisposition and hence can be exploited as novel prognostic biomarkers for better patient stratification and treatment. To validate this hypothesis, we present two different strategies to identify functionally relevant non-coding mutations. Namely, 1. Recurrence of non-coding mutations affecting an enhancer; 2. Aggregation of enhancer mutations in cancer associated pathways.

We present a comprehensive analysis framework for characterizing non-coding regulatory mutations. We worked on various challenges in this venture namely, defining lung-specific enhancers, enhancer target gene prediction, mutation mapping and functional analysis. Defining enhancers has been a major challenge due to the lack of an exhaustive list of enhancers in the literature for all the cell types. Enhancers are cell type-specific and hence list of enhancers from one cell type do not represent the whole lung tissue. Moreover, the cell of origin of lung cancers is not well-defined. Hence, we opted for a comprehensive list of lung-specific enhancers defined from a collection of lung cell lines and primary tissue comprising of fibroblast cell lines, epithelial cell lines, cancer cell lines and primary lung tissue and physiological fibroblast. This may seem a counter-intuitive solution as opposed to directly using only one cell type for enhancer definition. However, we reasoned that as the tumour of the patients are highly heterogeneous and our cohort of samples is a mix of different lung cancer sub-types, stages and other clinical features, we went for a comprehensive approach to include all the active lung enhancers. Although, the ideal solution would be to use the patient derived data for enhancer definition, but due to lack of data, we have used the cell-line based enhancers.

The general consensus from the literature about the epigenetic markers of enhancers are H3K27ac and accessibility obtained via DNase or transposase activity. Although, H3K4me1 is found at enhancers, it is often reported to be present also in poised or weak enhancers. Hence for identifying active enhancers in lung, we opted to use H3K27ac and chromatin accessibility (DHS and ATAC). In addition to the information from ENCODE3 data, we also performed ChIP-seq and ATAC-seq in two cell lines for enriching the repertoire.

Although, the cell of origin of lung cancer is not evident, normal lung epithelium is widely accepted in the field. As previously reported in Polak *et al.*,[382] the lack of epigenomic data from normal lung epithelial cells hampered the association of chromatin organization in the cell-of origin and the mutational landscape in lung cancer. Hence, we ensured the inclusion of Immortalized Human Bronchial Epithelial Cells (HBEC-3KT) in our repertoire by performing the experiments in-house.

Somatic mutation calling is challenging due to the hurdles posed by various factors including tumour heterogeneity, clonal mutations and tumour ploidy. Additionally, somatic mutations calling can be highly impacted by the sensitivity of the tool. To overcome these concerns, we employed a custom pipeline implementing an ensemble approach and used the concordance of at least two variant callers to ascertain a variant. One of the caveats in this approach is that when using the concordance of two tools, a single tool with a lot of false positives can confound the results. We observed a similar situation with the earlier version of the variant callers used, which included samtools in the ensemble. Due to a very high proportion of variants supported by samtools, we removed the tool from the current version.

Another limitation that we may encounter because of ensemble approach is the loss of variations that are called because of the sensitivity of the tool. However, as lung cancer has high mutation burden our priority was to ensure reduced false positives.

Furthermore, the identification of somatic mutations in non-coding regions is hampered by the intrinsic complexity of the regions due to repeats. Correspondingly, the availability of whole genome sequencing data is limited compared to exome sequencing data. We thus, opted for high coverage whole genome sequencing data for efficient non-coding mutation identification. We present here a comprehensive analysis of whole genome sequencing data of 159 individuals with lung cancers to characterise the landscape of non-coding mutations. We included three different cohorts covering three distinct lung cancer subtypes, viz adenocarcinoma, squamous cell carcinoma and small cell lung cancer.

Lung cancer is reported to have a high mutation burden. Hence understanding the burden of mutations in the non-coding regulatory regions in comparison to coding regions and the rest of the non-coding regions is crucial. We observe the mutation burden at enhancers is lower compared to the rest of the non-coding genome. Moreover, enhancers, promoters and exons have a comparable mutation burden. We speculate that this lower mutation burden in regulatory elements and exons could be attributed to a combination of negative selection[383].

These comparisons sheds light on the biological relevance of the mutation in all the regions. In our comparison of mutation burden between non-coding regulatory and coding mutations, we did not separate the coding mutations into synonymous or non-synonymous mutations as we reasoned that the non-coding mutations also have variable impact based on the location of the mutation with respect to the TFBS. As we have not given any weight to the various mutations in the non-coding regions, we did not stratify the coding mutations as well. Our aim was to ascertain, if the various genomic regions of interest have a similar tendency to be mutated.

Mutations in the genome occur because of various mutagens and are rectified by several repair mechanisms. To shed light on the process of mutagenesis, we compared the mutation signatures at enhancers, promoters and exons. Signature 4 associated with smoking was prevalent in all of the genomic regions compared in concordance with its association to lung cancer. When looking at the mutation signatures with an altered propensity at different regions, we observe mutation signatures associated with defective DNA mismatch repair and DNA double strand break repair to be higher in regulatory regions compared to coding regions. These results corroborate the recent literature on the accumulation of single and double strand breaks at regulatory regions[360,384]. The role of the mismatch repair system for maintaining genome stability is well characterised and their role in activating gene enhancers in cancer is emerging[385,386].

Enhancers are distal regulatory elements, and hence the location of an enhancer with respect to its cognate promoter is farther in linear sequence. Whereas in the three-dimensional space looping of the chromatin positions enhancers and promoters proximal to one another. TAD boundaries demarcate these dynamics and can help identify possible interacting pairs. So far, information on 3D genomics has not been implemented by the algorithms predicting enhancer and promoter pairs[145,147,148,155]. Hence, we developed a prediction methodology leveraging the three-dimensional chromatin architecture for effective reconstruction of enhancer -target gene regulatory interactions. Our approach integrates the information from genome-wide profiles of epigenetic marks for 44 cell and tissue types along with multi-scale TAD calls derived from 11 high coverage Hi-C datasets. We quantified the gene activity using the epigenetic marks at promoters and associated them with the activity of enhancers. We used the prior-knowledge on chromatin 3D organization to quantify the physical proximity by incorporating TAD information. This information was used to adjust the P-values for each enhancer promoter pair.

In our approach, we used the chromatin 3D information from Hi-C data to score the predicted ETG pairs rather than to directly identifying ETG loops. In principle, Hi-C data could be used for the genome-wide identification of specific points of contacts. However, Hi-C data is generally analysed by binning read counts at a resolution of few kbs, this resolution level is lower for mapping ETG pairs when present close to each other[31,67,83]. Additionally, Hi-C point interact calling algorithms have been shown to yield variable results even across biological replicates[75]. It is worth remarking that in our ETG reconstruction methodology, we have quantified the gene activity using epigenetic marks at promoters, as opposed to the mRNA expression. The rationale behind this solution is that the transcript abundance depends on multiple levels of co-transcriptional and post-transcriptional regulations, such as polymerase pausing, splicing, mRNA decay etc. Hence, we opted for promoter activity, which is also a common choice in literature of this field.

Based on our prediction of enhancer target gene pairs, we observed that a gene has more than one associated enhancer. The convergence of several enhancers for the regulation of a single gene are reported in various studies[387–389]. We further explored, if the genes with many lung specific enhancers are more relevant to lung tissue. We observed that the genes with many enhancers were highly expressed in lung, compared to other tissues. These observations highlight a possible network of enhancer-enhancer interactions for orchestrated gene expression.

The effect of coding mutation can be directly corroborated by its functional consequence on the protein sequence or structure. Whereas the effect of non-coding mutations is usually not as straightforward. Although regulatory mutations affect the expression of their target gene, other factors like the hypermethylation of promoters or copy number alterations can be confounding. Hence to understand if the enhancer mutations have an impact on the expression of its target gene, we compared the gene expression of patients stratified as mutated and non-mutated based on enhancer mutations. We observe significant changes in the expression of the genes both in the positive and negative direction.

Yet another way to predict non-coding mutations is by characterizing tissue-specific binding sites of transcription factors. Transcription factors have DNA-binding domains that give them the ability to bind to specific sequences of DNA at enhancers and promoters. Regions of the enhancer that are actively bound by a transcription factor can be identified through DNase-seq footprinting. Hence, we incorporated the DNase-seq information of IMR-90 cell line to identify regions of active TF binding within an enhancer. We call these enhancer

cores. We were limited by the availability of DNase footprinting data only in IMR-90 cell line. With the hypothesis that mutations in the enhancer cores can alter sequence motifs thereby impacting the binding of TFs, we identified TF motifs on reference sequence and altered sequence and found significant alterations in the motifs. We predicted the possible loss or gain of a TFBS at the enhancers based on somatic mutations, however, the impact of a mutation at the TFBS can also alter the propensity of binding. Additionally, the presence of a TF motif is not a definitive proof for the binding of a transcription factor.

Recurrence of mutation has been a powerful tool to identify biologically relevant mutations in coding genome so far. The sheer volume of the non-coding genome further lowers the chance recurrences of mutations. Hence, we adopted recurrence as a key characteristic for identifying a biologically meaningful mutation. We found an enhancer within CDH13 gene to be mutated with an insertion. Upon the CRISPR deletion of the insertion mutation, we observe the downregulation of the gene. While in the patient samples, the presence of the insertion, resulted in the increase in gene expression (not significant). Although the mutation did not significantly alter the gene expression, we observe that in a small cohort of LUSC samples with insertion mutation, had hundred percent disease free survival. Due to the non-availability of clinical information from the SCLC patients, we could not confirm the clinical relevance of CDH13 insertion sequence variant in SCLC.

At the CDH13 enhancer we observe creation of several transcription factors that are expressed in the patients. We also observe the presence of theses motifs at multiple adjacent locations within the region. This result corroborates literature that at the gene regulatory regions there is an accumulation of potential TF binding sites in regions and the presence of multiple degenerate or weakly competing binding sites could accelerate the TF search for its target gene.

Genes are part of a larger network of multiple interacting pathways. Alterations in the expression or function of a gene affects other genes. Hence, the role of non-coding mutations is not limited to gene, but in extension the pathway where the genes belong. Hence, we aggregated the enhancer mutations at pathway level. We observe mutual exclusivity of regulatory and coding mutations. We speculate that this could be a result of cancer evolution of the patient, *i.e.*, when the genes expression is altered by a regulatory mutation, mutation in the coding region to invoke a functional change is not relevant. We also observe a significant overlap between genes with enhancer mutations and gene sets associated with perturbations in cancer drivers like KRAS, RAF1, MAPK and p53, indicating the relevance

of enhancer mutations in cancer progression. We have shown the enhancer mutations converging at pathway level; however, we have not assessed the impact of these mutations.

In conclusion, here we present two different strategies to identify functionally relevant non-coding mutations. Namely 1. Recurrence of non-coding mutations affecting the core of an enhancer 2. Aggregation of enhancer mutations in cancer associated pathways. We have shown that enhancer mutations can impact expression of target genes and that the patients with recurrent enhancer mutations have an effect in survival probability. Finally, we also highlight how mutations in enhancers can impact key cancer pathways. These results show that, non-coding regulatory mutations can be exploited for patient stratification.

6. REFERENCES

1. Stratton, M. R., Campbell, P. J. & Futreal, P. A. The cancer genome. *Nature* **458**, 719–724 (2009).

2. Martincorena, I. & Campbell, P. J. Somatic mutation in cancer and normal cells. *Science (80-.).* **349**, 1483–1489 (2015).

3. Martincorena, I. *et al.* Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell* **173**, 1823 (2018).

4. Rheinbay, E. *et al.* Recurrent and functional regulatory mutations in breast cancer. *Nature* **547**, 55–60 (2017).

5. Mathelier, A. *et al.* Cis-regulatory somatic mutations and gene-expression alteration in B-cell lymphomas. *Genome Biol.* **16**, 1–17 (2015).

6. Melton, C., Reuter, J. A., Spacek, D. V. & Snyder, M. Recurrent Somatic Mutations in Regulatory Regions of Human Cancer Genomes. *Nat. Genet.* **47**, 710–716 (2016).

7. Zhang, W. *et al.* A global transcriptional network connecting noncoding mutations to changes in tumor gene expression. *Nat. Genet.* (2018) doi:10.1038/s41588-018-0091-2.

8. Lange, M., Begolli, R. & Giakountis, A. Non-coding variants in cancer: Mechanistic insights and clinical potential for personalized medicine. *Non-coding RNA* **7**, (2021).

9. Hayes, J. J. & Hansen, J. C. Nucleosomes and the chromatin fiber. *Curr. Opin. Genet. Dev.* **11**, 124–129 (2001).

10. Jan, B. *et al.* Nucleosomes, linker DNA, and linker histone form a unique structural motif that directs the higher-order folding and compaction of chromatin. *Proc. Natl. Acad. Sci.* **95**, 14173–14178 (1998).

11. Kornberg, R. D. Chromatin structure: a repeating unit of histones and DNA. *Science* **184**, 868–871 (1974).

12. Olins, D. E. & Olins, A. L. Chromatin history: our view from the bridge. *Nature reviews. Molecular cell biology* vol. 4 809–814 (2003).

13. Ricci, M. A., Manzo, C., García-Parajo, M. F., Lakadamyali, M. & Cosma, M. P. Chromatin Fibers Are Formed by Heterogeneous Groups of Nucleosomes In Vivo.

Cell **160**, 1145–1158 (2015).

14. Bascom, G. & Schlick, T. Linking Chromatin Fibers to Gene Folding by Hierarchical Looping. *Biophys. J.* **112**, 434–445 (2017).

15. Hagstrom, K. A. & Meyer, B. J. Condensin and cohesin: more than chromosome compactor and glue. *Nat. Rev. Genet.* **4**, 520–534 (2003).

16. Merkenschlager, M. & Odom, D. T. CTCF and cohesin: linking gene regulatory elements with their targets. *Cell* **152**, 1285–1297 (2013).

17. Phillips, J. E. & Corces, V. G. CTCF: master weaver of the genome. *Cell* **137**, 1194–1211 (2009).

18. Thadani, R., Uhlmann, F. & Heeger, S. Condensin, chromatin crossbarring and chromosome condensation. *Curr. Biol.* **22**, R1012-21 (2012).

19. Fudenberg, G. *et al.* Formation of Chromosomal Domains by Loop Extrusion. *Cell Rep.* **15**, 2038–2049 (2016).

20. Gabriele, M. *et al.* Dynamics of CTCF- and cohesin-mediated chromatin looping revealed by live-cell imaging. *Science (80-.).* **376**, 496–501 (2022).

21. Sanborn, A. L. *et al.* Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proc. Natl. Acad. Sci.* **112**, E6456–E6465 (2015).

22. Lieberman-Aiden, E. *et al.* Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science (80-.).* **326**, 289–293 (2009).

23. Gibcus, J. H. & Dekker, J. The Hierarchy of the 3D Genome. *Mol. Cell* **49**, 773–782 (2013).

24. Sanyal, A., Lajoie, B. R., Jain, G. & Dekker, J. The long-range interaction landscape of gene promoters. *Nature* **489**, 109–113 (2012).

25. Dekker, J. & Misteli, T. Long-Range Chromatin Interactions. *Cold Spring Harb. Perspect. Biol.* **7**, a019356–a019356 (2015).

26. Bonev, B. & Cavalli, G. Organization and function of the 3D genome. *Nat. Rev. Genet.* **17**, 661–678 (2016).

27. Rowley, M. J. & Corces, V. G. Organizational principles of 3D genome architecture. *Nat. Rev. Genet.* **19**, 789–800 (2018).

28. Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376–380 (2012).

29. Nora, E. P. *et al.* Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* **485**, 381–385 (2012).

30. Dixon, J. R. *et al.* Chromatin architecture reorganization during stem cell differentiation. *Nature* **518**, 331–336 (2015).

31. Rao, S. S. P. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).

32. Weinreb, C. & Raphael, B. J. Identification of hierarchical chromatin domains. *Bioinformatics* **32**, 1601–1609 (2016).

33. Fraser, J. *et al.* Hierarchical folding and reorganization of chromosomes are linked to transcriptional changes in cellular differentiation. *Mol. Syst. Biol.* **11**, 852 (2015).

34. Nichols, M. H. & Corces, V. G. Principles of 3D compartmentalization of the human genome. *Cell Rep.* **35**, 109330 (2021).

35. Liu, Y. *et al.* Systematic inference and comparison of multi-scale chromatin sub-compartments connects spatial organization to cell phenotypes. *Nat. Commun.* **12**, 2439 (2021).

36. Hildebrand, E. M. & Dekker, J. Mechanisms and Functions of Chromosome Compartmentalization. *Trends Biochem. Sci.* **45**, 385–396 (2020).

37. McArthur, E. & Capra, J. A. Topologically associating domain boundaries that are stable across diverse cell types are evolutionarily constrained and enriched for heritability. *Am. J. Hum. Genet.* **108**, 269–283 (2021).

38. Wit, E. De & Laat, W. De. A decade of 3C technologies-insights into nuclear organization. *Genes Dev.* 11–24 (2012) doi:10.1101/gad.179804.111.GENES.

39. Briand, N. & Collas, P. Lamina-associated domains: peripheral matters and internal affairs. *Genome Biol.* **21**, 85 (2020).

40. Bayani, J. & Squire, J. A. Fluorescence in situ Hybridization (FISH). *Curr. Protoc.*

cell Biol. **Chapter 22**, Unit 22.4 (2004).

41. Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. Capturing Chromosome Conformation. *Science (80-.).* **295**, 1306–1311 (2002).

42. de Laat, W. & Dekker, J. 3C-based technologies to study the shape of the genome. *Methods (San Diego, Calif.)* vol. 58 189–191 (2012).

43. Schmitt, A. D., Hu, M. & Ren, B. Genome-wide mapping and analysis of chromosome architecture. *Nat. Rev. Mol. Cell Biol.* **17**, 743–755 (2016).

44. Kempfer, R. & Pombo, A. Methods for mapping 3D chromosome architecture. *Nat. Rev. Genet.* **21**, 207–226 (2020).

45. Lieberman-aiden, E. *et al.* Comprehensive Mapping of Long-Range Interactions Revelas Folding Principles of the Human Genome. *Science* **326**, 289–294 (2009).

46. Lieberman-Aiden, E. *et al.* Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science (80-.).* **326**, 289–293 (2009).

47. Li, G. *et al.* Chromatin Interaction Analysis with Paired-End Tag (ChIA-PET) sequencing technology and application. *BMC Genomics* **15**, S11 (2014).

48. van Berkum, N. L. *et al.* Hi-C: a method to study the three-dimensional architecture of genomes. *J. Vis. Exp.* 1869 (2010) doi:10.3791/1869.

49. Simonis, M. *et al.* Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nat. Genet.* **38**, 1348–1354 (2006).

50. van de Werken, H. J. G. *et al.* 4C technology: protocols and data analysis. *Methods Enzymol.* **513**, 89–112 (2012).

51. Dostie, J. *et al.* Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome Res.* **16**, 1299–1309 (2006).

52. Phillips-Cremins, J. E. *et al.* Architectural protein subclasses shape 3D organization of genomes during lineage commitment. *Cell* **153**, 1281–1295 (2013).

53. Kim, T. H. & Dekker, J. Generation of ChIP-Loop Libraries. *Cold Spring Harb.*

Protoc. **2018**, (2018).

54. Zhang, J. *et al.* ChIA-PET analysis of transcriptional chromatin interactions. *Methods* **58**, 289–299 (2012).

55. Johanson, T. M. & Allan, R. S. In Situ HiC. *Methods Mol. Biol.* **2458**, 333–343 (2022).

56. Hsieh, T.-H. S. *et al.* Mapping Nucleosome Resolution Chromosome Folding in Yeast by Micro-C. *Cell* **162**, 108–119 (2015).

57. Burgess, D. J. Chromosome structure at micro-scale. *Nat. Rev. Genet.* **21**, 337 (2020).

58. de Souza, N. Micro-C maps of genome structure. *Nat. Methods* **12**, 812 (2015).

59. Mifsud, B. *et al.* Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nat. Genet.* **47**, 598–606 (2015).

60. Dryden, N. H. *et al.* Unbiased analysis of potential targets of breast cancer susceptibility loci by Capture Hi-C. *Genome Res.* **24**, 1854–1868 (2014).

61. Javierre, B. M. *et al.* Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. *Cell* **167**, 1369-1384.e19 (2016).

62. Davies, J. O. J. *et al.* Multiplexed analysis of chromosome conformation at vastly improved sensitivity. *Nat. Methods* **13**, 74–80 (2016).

63. Schoenfelder, S. *et al.* The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements. *Genome Res.* **25**, 582–597 (2015).

64. Schoenfelder, S., Javierre, B.-M., Furlan-Magaril, M., Wingett, S. W. & Fraser, P. Promoter Capture Hi-C: High-resolution, Genome-wide Profiling of Promoter Interactions. *J. Vis. Exp.* (2018) doi:10.3791/57320.

65. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).

66. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *bioinformatics* **25**, 1754–1760 (2009).

67. Pal, K., Forcato, M. & Ferrari, F. Hi-C analysis: from data generation to integration.

Biophys. Rev. 67–78 (2018) doi:10.1007/s12551-018-0489-1.

68. Imakaev, M. *et al.* Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat. Methods* **9**, 999–1003 (2012).

69. Wingett, S. *et al.* HiCUP: pipeline for mapping and processing Hi-C data. *F1000Research* **4**, 1310 (2015).

70. Servant, N. *et al.* HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol.* **16**, 259 (2015).

71. Serra, F. *et al.* Automatic analysis and 3D-modelling of Hi-C data using TADbit reveals structural features of the fly chromatin colors. *PLoS Comput. Biol.* **13**, e1005665 (2017).

72. Ay, F. & Noble, W. S. Analysis methods for studying the 3D architecture of the genome. *Genome Biol.* **16**, 183 (2015).

73. Cournac, A., Marie-Nelly, H., Marbouty, M., Koszul, R. & Mozziconacci, J. Normalization of a chromosomal contact map. *BMC Genomics* **13**, 436 (2012).

74. Knight, P. A. & Ruiz, D. A fast algorithm for matrix balancing. *IMA J. Numer. Anal.* **33**, 1029–1047 (2013).

75. Forcato, M. *et al.* Comparison of computational methods for Hi-C data analysis. *Nat. Methods* **14**, 679–685 (2017).

76. Seaman, L. & Rajapakse, I. 4D nucleome Analysis Toolbox: analysis of Hi-C data with abnormal karyotype and time series capabilities. *Bioinformatics* **34**, 104–106 (2018).

77. Lévy-Leduc, C., Delattre, M., Mary-Huard, T. & Robin, S. Two-dimensional segmentation for analyzing Hi-C data. *Bioinformatics* **30**, i386-92 (2014).

78. Oluwadare, O. & Cheng, J. ClusterTAD: an unsupervised machine learning approach to detecting topologically associated domains of chromosomes from Hi-C data. *BMC Bioinformatics* **18**, 1–14 (2017).

79. Ay, F., Bailey, T. L. & Noble, W. S. Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts. *Genome Res.* **24**, 999–1011 (2014).

80. Mifsud, B. *et al.* GOTHiC, a probabilistic model to resolve complex biases and to

identify real interactions in Hi-C data. *PLoS One* **12**, e0174744 (2017).

81. Hwang, Y. C. *et al.* HIPPIE: A high-throughput identification pipeline for promoter interacting enhancer elements. *Bioinformatics* **31**, 1290–1292 (2015).

82. Salviato, E. *et al.* Leveraging three-dimensional chromatin architecture for effective reconstruction of enhancer–target gene regulatory interactions. *Nucleic Acids Res.* **49**, e97–e97 (2021).

83. Bonev, B. *et al.* Multiscale 3D Genome Rewiring during Mouse Neural Development. *Cell* **171**, 557-572.e24 (2017).

84. Laat, W. De & Duboule, D. Topology of mammalian developmental enhancers and their regulatory landscapes. *Nature* **502**, 499–506 (2013).

85. Pennacchio, L. a, Bickmore, W., Dean, A., Nobrega, M. a & Bejerano, G. Enhancers: five essential questions. *Nat. Rev. Genet.* **14**, 288–295 (2013).

86. Shlyueva, D., Stampfel, G. & Stark, A. Transcriptional enhancers: From properties to genome-wide predictions. *Nature Reviews Genetics* vol. 15 272–286 (2014).

87. Kim, T.-K. & Shiekhattar, R. Architectural and Functional Commonalities between Enhancers and Promoters. *Cell* **162**, 948–959 (2015).

88. Visel, A. *et al.* Functional autonomy of distant-acting human enhancers. *Genomics* **93**, 509–513 (2009).

89. Ong, C. T. & Corces, V. G. Enhancer function: New insights into the regulation of tissue-specific gene expression. *Nat. Rev. Genet.* **12**, 283–293 (2011).

90. Kagey, M. H. *et al.* Mediator and cohesin connect gene expression and chromatin architecture. *Nature* **467**, 430–435 (2010).

91. Yen, A. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).

92. Banerji, J., Rusconi, S. & Schaffner, W. Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. *Cell* **27**, 299–308 (1981).

93. Magnani, L., Eeckhoute, J. & Lupien, M. Pioneer factors: Directing transcriptional regulators within the chromatin environment. *Trends Genet.* **27**, 465–474 (2011).

94. Hu, Z. & Tee, W.-W. W. Enhancers and chromatin structures: regulatory hubs in

gene expression and diseases. *Biosci. Rep.* **37**, BSR20160183 (2017).

95. Yip, K. Y. *et al.* Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors. *Genome Biol.* **13**, R48 (2012).

96. Joshi, A. Mammalian transcriptional hotspots are enriched for tissue specific enhancers near cell type specific highly expressed genes and are predicted to act as transcriptional activator hubs. *BMC Bioinformatics* **15**, 412 (2014).

97. Allen, B. L. & Taatjes, D. J. The Mediator complex: A central integrator of transcription. *Nat. Rev. Mol. Cell Biol.* **16**, 155–166 (2015).

98. Schoenfelder, S. & Fraser, P. Long-range enhancer-promoter contacts in gene expression control. *Nat. Rev. Genet.* **20**, 437–455 (2019).

99. Juven-Gershon, T., Hsu, J.-Y., Theisen, J. W. & Kadonaga, J. T. The RNA polymerase II core promoter - the gateway to transcription. *Curr. Opin. Cell Biol.* **20**, 253–259 (2008).

100. Haberle, V. & Stark, A. Eukaryotic core promoters and the functional basis of transcription initiation. *Nat. Rev. Mol. Cell Biol.* **19**, 621–637 (2018).

101. Kuehner, J. N., Pearson, E. L. & Moore, C. Unravelling the means to an end: RNA polymerase II transcription termination. *Nat. Rev. Mol. Cell Biol.* **12**, 283–294 (2011).

102. Mischo, H. E. & Proudfoot, N. J. Disengaging polymerase: terminating RNA polymerase II transcription in budding yeast. *Biochim. Biophys. Acta* **1829**, 174–185 (2013).

103. Adelman, K. & Lis, J. T. Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans. *Nat. Rev. Genet.* **13**, 720–731 (2012).

104. Dollinger, R. & Gilmour, D. S. Regulation of Promoter Proximal Pausing of RNA Polymerase II in Metazoans. *J. Mol. Biol.* **433**, 166897 (2021).

105. Tippens, N. D., Vihervaara, A. & Lis, J. T. Enhancer transcription: What, where, when, and why? *Genes Dev.* **32**, 1–3 (2018).

106. Meng, H. & Bartholomew, B. Emerging roles of transcriptional enhancers in chromatin looping and promoter-proximal pausing of RNA polymerase II. *J. Biol.*

Chem. **293**, 13786–13794 (2018).

107. Li, W., Notani, D. & Rosenfeld, M. G. Enhancers as non-coding RNA transcription units: Recent insights and future perspectives. *Nat. Rev. Genet.* **17**, 207–223 (2016).

108. Lee, K., Hsiung, C. C.-S., Huang, P., Raj, A. & Blobel, G. A. Dynamic enhancer-gene body contacts during transcription elongation. *Genes Dev.* **29**, 1992–1997 (2015).

109. Bonn, S. *et al.* Tissue-specific analysis of chromatin state identifies temporal signatures of enhancer activity during embryonic development. *Nat. Genet.* **44**, 148–156 (2012).

110. Rada-Iglesias, A. *et al.* A unique chromatin signature uncovers early developmental enhancers in humans. *Nature* **470**, 279 (2010).

111. Heintzman, N. D. *et al.* Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* **459**, 108–112 (2009).

112. Bae, S. & Lesch, B. J. H3K4me1 Distribution Predicts Transcription State and Poising at Promoters. *Front. cell Dev. Biol.* **8**, 289 (2020).

113. Ernst, J. & Kellis, M. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat. Biotechnol.* **28**, 817–25 (2010).

114. Creyghton, M. P. *et al.* Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 21931–21936 (2010).

115. Tsompana, M. & Buck, M. J. Chromatin accessibility: a window into the genome. *Epigenetics Chromatin* **7**, 33 (2014).

116. Boyle, A. P. *et al.* High-resolution mapping and characterization of open chromatin across the genome. *Cell* **132**, 311–322 (2008).

117. Buenrostro, J. D., Wu, B., Chang, H. Y. & Greenleaf, W. J. ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. *Curr. Protoc. Mol. Biol.* **109**, 21.29.1-21.29.9 (2015).

118. Kim, T.-K. *et al.* Widespread transcription at neuronal activity-regulated enhancers. *Nature* **465**, 182–187 (2010).

119. De Santa, F. *et al.* A large fraction of extragenic RNA pol II transcription sites overlap enhancers. *PLoS Biol.* **8**, e1000384 (2010).

120. Kodzius, R. *et al.* CAGE: cap analysis of gene expression. *Nat. Methods* **3**, 211–222 (2006).

121. Takahashi, H., Lassmann, T., Murata, M. & Carninci, P. 5' end-centered expression profiling using cap-analysis gene expression and next-generation sequencing. *Nat. Protoc.* **7**, 542–561 (2012).

122. Valen, E. *et al.* Genome-wide detection and analysis of hippocampus core promoters using DeepCAGE. *Genome Res.* **19**, 255–265 (2009).

123. Core, L. J., Waterfall, J. J. & Lis, J. T. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science* **322**, 1845–1848 (2008).

124. Churchman, L. S. & Weissman, J. S. Native elongating transcript sequencing (NET-seq). *Curr. Protoc. Mol. Biol.* **Chapter 4**, Unit 4.14.1-17 (2012).

125. Hirabayashi, S. *et al.* NET-CAGE characterizes the dynamics and topology of human transcribed cis-regulatory elements. *Nat. Genet.* **51**, 1369–1379 (2019).

126. Calo, E. & Wysocka, J. Modification of Enhancer Chromatin: What, How, and Why? *Mol. Cell* **49**, 825–837 (2013).

127. Buecker, C. & Wysocka, J. Enhancers as information integration hubs in development: Lessons from genomics. *Trends Genet.* **28**, 276–284 (2012).

128. Maston, G. A. *et al.* Non-canonical TAF complexes regulate active promoters in human embryonic stem cells. *Elife* **1**, e00068 (2012).

129. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).

130. Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–329 (2015).

131. Andersson, R. *et al.* An atlas of active enhancers across human cell types and tissues. *Nature* **507**, 455–461 (2014).

132. Visel, A., Minovitsky, S., Dubchak, I. & Pennacchio, L. A. VISTA Enhancer

Browser--a database of tissue-specific human enhancers. *Nucleic Acids Res.* **35**, D88-92 (2007).

133. Fishilevich, S. *et al.* GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. *Database* **2017**, 1–17 (2017).

134. Wang, J. *et al.* HACER: an atlas of human active enhancers to interpret regulatory variants. *Nucleic Acids Res.* **47**, D106–D112 (2019).

135. Jiang, Y. *et al.* SEdb: a comprehensive human super-enhancer database. *Nucleic Acids Res.* **47**, D235–D243 (2019).

136. Cai, Z. *et al.* RAEdb: a database of enhancers identified by high-throughput reporter assays. *Database* (2019) doi:10.1093/database/bay140.

137. Wang, Z. *et al.* HEDD: Human Enhancer Disease Database. *Nucleic Acids Res.* **46**, D113–D120 (2018).

138. Ashoor, H., Kleftogiannis, D., Radovanovic, A. & Bajic, V. B. DENdb: database of integrated human enhancers. *Database (Oxford).* **2015**, bav085 (2015).

139. Khan, A. & Zhang, X. dbSUPER: a database of super-enhancers in mouse and human genome. *Nucleic Acids Res.* **44**, D164–D171 (2016).

140. Hariprakash, J. M. & Ferrari, F. Computational Biology Solutions to Identify Enhancers-target Gene Pairs. *Comput. Struct. Biotechnol. J.* **17**, (2019).

141. Shen, Y. *et al. A map of the cis-regulatory sequences in the mouse genome*. *Nature* vol. 488 (2012).

142. Thurman, R. E. *et al.* The accessible chromatin landscape of the human genome. *Nature* **489**, 75–82 (2012).

143. Yao, L., Shen, H., Laird, P. W., Farnham, P. J. & Berman, B. P. Inferring regulatory element landscapes and transcription factor networks from cancer methylomes. *Genome Biol.* **16**, 1–21 (2015).

144. Silva, T. C. *et al.* ELMER v.2: an R/Bioconductor package to reconstruct gene regulatory networks from DNA methylation and transcriptome profiles. *Bioinformatics* (2018) doi:10.1093/bioinformatics/bty902.

145. He, B., Chen, C., Teng, L. & Tan, K. Global view of enhancer-promoter interactome

in human cells. *Proc. Natl. Acad. Sci.* **111**, E2191–E2199 (2014).

146. Hafez, D. *et al.* McEnhancer: Predicting gene expression via semi-supervised assignment of enhancers to target genes. *Genome Biol.* **18**, 1–21 (2017).

147. Zhao, C., Li, X. & Hu, H. PETModule: A motif module based approach for enhancer target gene prediction. *Sci. Rep.* **6**, 1–10 (2016).

148. Whalen, S. *et al.* Enhancer-promoter interactions are encoded by complex genomic signatures on looping chromatin. *Nat. Genet.* **48**, 488–496 (2016).

149. Corradin, O. *et al.* Combinatorial effects of multiple enhancer variants in linkage disequilibrium dictate levels of gene expression to confer susceptibility to common traits. *Genome Res.* **24**, 1–13 (2014).

150. Okonechnikov, K., Erkek, S., Korbel, J. O., Pfister, S. M. & Chavez, L. InTAD: chromosome conformation guided analysis of enhancer target genes. *BMC Bioinformatics* **20**, 60 (2019).

151. Rödelsperger, C. *et al.* Integrative analysis of genomic, functional and protein interaction data predicts long-range enhancer-target gene interactions. *Nucleic Acids Res.* **39**, 2492–2502 (2011).

152. Ernst, J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**, 43–49 (2011).

153. Gao, T. & Qian, J. EAGLE: An algorithm that utilizes a small number of genomic features to predict tissue/cell type-specific enhancer-gene interactions. *PLOS Comput. Biol.* **15**, e1007436 (2019).

154. Roy, S. *et al.* A predictive modeling approach for cell line-specific long-range regulatory interactions. *Nucleic Acids Res.* **43**, 8694–8712 (2015).

155. Cao, Q. *et al.* Reconstruction of enhancer-target networks in 935 samples of human primary cells, tissues and cell lines. *Nat. Genet.* **49**, 1428–1436 (2017).

156. Hait, T. ., Amar, D., Shamir, R. & Elkon, R. FOCS: a novel method for analyzing enhancer and gene activity Map, an extensive enhancer–promoter. *Genome Biol.* **19**, 1–14 (2018).

157. Zhu, Y. *et al.* Constructing 3D interaction maps from 1D epigenomes. *Nat. Commun.* **7**, 1–11 (2016).

158. Clément, Y., Torbey, P., Gilardi-Hebenstreit, P. & Crollius, H. R. Enhancer-gene maps in the human and zebrafish genomes using evolutionary linkage conservation. *Nucleic Acids Res.* **48**, 2357–2371 (2020).

159. Naville, M. *et al.* Long-range evolutionary constraints reveal cis-regulatory interactions on the human X chromosome. *Nat. Commun.* **6**, 6904 (2015).

160. Han Chen, Chunyan Li, Xinxin Peng, Zhicheng Zhou, John N. Weinstein, The Cancer Genome Atlas Research Network, H. L. A Pan-Cancer Analysis of Enhancer Expression in Nearly 9000 Patient Samples. *Cell* 386–399 (2018) doi:10.1016/j.cell.2018.03.027.

161. Smith, E. & Shilatifard, A. Enhancer biology and enhanceropathies. *Nat. Struct. Mol. Biol.* **21**, 210–219 (2014).

162. Murakawa, Y. *et al.* Enhanced Identification of Transcriptional Enhancers Provides Mechanistic Insights into Diseases. *Trends Genet.* **32**, 76–88 (2016).

163. Kioussis, D., Vanin, E., deLange, T., Flavell, R. A. & Grosveld, F. G. β-Globin gene inactivation by DNA translocation in γβ-thalassaemi. *Nature* **306**, 662–666 (1983).

164. Claringbould, A. & Zaugg, J. B. Enhancers in disease: molecular basis and emerging treatment strategies. *Trends Mol. Med.* **27**, 1060–1073 (2021).

165. Nalls, M. A. *et al.* Large-scale meta-analysis of genome-wide association data identifies six new risk loci for Parkinson's disease. *Nat. Genet.* **46**, 989–993 (2014).

166. Miller, D. W. *et al.* Alpha-synuclein in blood and brain from familial Parkinson disease with SNCA locus triplication. *Neurology* **62**, 1835–1838 (2004).

167. Devine, M. J., Gwinn, K., Singleton, A. & Hardy, J. Parkinson's disease and α-synuclein expression. *Mov. Disord.* **26**, 2160–2168 (2011).

168. Soldner, F. *et al.* Parkinson-associated risk variant in distal enhancer of α-synuclein modulates target gene expression. *Nature* **533**, 95–99 (2016).

169. Weedon, M. N. *et al.* Recessive mutations in a distal PTF1A enhancer cause isolated pancreatic agenesis. *Nat. Genet.* **46**, 61–64 (2014).

170. Sellick, G. S. *et al.* Mutations in PTF1A cause pancreatic and cerebellar agenesis. *Nat. Genet.* **36**, 1301–1305 (2004).

171. Tutak, E. *et al.* A Turkish newborn infant with cerebellar agenesis/neonatal diabetes mellitus and PTF1A mutation. *Genet. Couns.* **20**, 147–152 (2009).

172. Al-Shammari, M., Al-Husain, M., Al-Kharfy, T. & Alkuraya, F. S. A novel PTF1A mutation in a patient with severe pancreatic and cerebellar involvement. *Clinical genetics* vol. 80 196–198 (2011).

173. Spielmann, M., Lupiáñez, D. G. & Mundlos, S. Structural variation in the 3D genome. *Nat. Rev. Genet.* **19**, 453–467 (2018).

174. Spielmann, M., Lupiáñez, D. G. & Mundlos, S. Structural variation in the 3D genome. *Nat. Rev. Genet.* **19**, 453–467 (2018).

175. Zimmerman, M. W. *et al.* MYC Drives a Subset of High-Risk Pediatric Neuroblastomas and Is Activated through Mechanisms Including Enhancer Hijacking and Focal Enhancer Amplification. *Cancer Discov.* **8**, 320–335 (2018).

176. Ooi, W. F. *et al.* Integrated paired-end enhancer profiling and whole-genome sequencing reveals recurrent CCNE1 and IGF2 enhancer hijacking in primary gastric adenocarcinoma. *Gut* **69**, 1039–1052 (2020).

177. Northcott, P. A. *et al.* Enhancer hijacking activates GFI1 family oncogenes in medulloblastoma. *Nature* **511**, 428–434 (2014).

178. Giorgio, E. *et al.* A large genomic deletion leads to enhancer adoption by the lamin B1 gene: a second path to autosomal dominant adult-onset demyelinating leukodystrophy (ADLD). *Hum. Mol. Genet.* **24**, 3143–3154 (2015).

179. Franke, M. *et al.* Formation of new chromatin domains determines pathogenicity of genomic duplications. *Nature* **538**, 265–269 (2016).

180. Kurth, I. *et al.* Duplications of noncoding elements 5’ of SOX9 are associated with brachydactyly-anonychia. *Nature genetics* vol. 41 862–863 (2009).

181. Lupiáñez, D. G. *et al.* Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell* **161**, 1012–1025 (2015).

182. Souren, N. Y. *et al.* DNA methylation signatures of monozygotic twins clinically discordant for multiple sclerosis. *Nat. Commun.* **10**, 2094 (2019).

183. Lin, X. *et al.* Genome-wide analysis of aberrant methylation of enhancer DNA in human osteoarthritis. *BMC Med. Genomics* **13**, 1 (2020).

184. Mordaunt, C. E. *et al.* Epigenomic signatures in liver and blood of Wilson disease patients include hypermethylation of liver-specific enhancers. *Epigenetics Chromatin* **12**, 10 (2019).

185. Reyes-Palomares, A. *et al.* Remodeling of active endothelial enhancers is associated with aberrant gene-regulatory networks in pulmonary arterial hypertension. *Nat. Commun.* **11**, 1673 (2020).

186. Rhodes, C. J. *et al.* Genetic determinants of risk in pulmonary arterial hypertension: international genome-wide association studies and meta-analysis. *Lancet Respir. Med.* **7**, 227–238 (2019).

187. Greenman, C. *et al.* Patterns of somatic mutation in human cancer genomes. *Nature* **446**, 153–158 (2007).

188. Vijg, J. Somatic mutations, genome mosaicism, cancer and aging. *Curr. Opin. Genet. Dev.* **26**, 141–149 (2014).

189. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: The next generation. *Cell* **144**, 646–674 (2011).

190. Greaves, M. & Maley, C. C. Clonal evolution in cancer. *Nature* **481**, 306–313 (2012).

191. Bailey, M. H. *et al.* Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell* **174**, 1034–1035 (2018).

192. Zack, T. I. *et al.* Pan-cancer patterns of somatic copy number alteration. *Nat. Genet.* **45**, 1134–1140 (2013).

193. Lawrence, M. S. *et al.* Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **505**, 495–501 (2014).

194. Lee, T. I. & Young, R. A. Transcriptional regulation and its misregulation in disease. *Cell* **152**, 1237–1251 (2013).

195. Elliott, K. & Larsson, E. Non-coding driver mutations in human cancer. *Nat. Rev. Cancer* **21**, 500–509 (2021).

196. Diederichs, S. *et al.* The dark matter of the cancer genome: aberrations in regulatory elements, untranslated regions, splice sites, non-coding RNA and synonymous mutations. *EMBO Mol. Med.* **8**, 442–457 (2016).

197. Huang, F. W. *et al.* Highly recurrent TERT promoter mutations in human melanoma. *Science* **339**, 957–959 (2013).

198. Horn, S. *et al.* TERT promoter mutations in familial and sporadic melanoma. *Science* **339**, 959–961 (2013).

199. Nonoguchi, N. *et al.* TERT promoter mutations in primary and secondary glioblastomas. *Acta Neuropathol.* **126**, 931–937 (2013).

200. Giedl, J. *et al.* TERT Core Promotor Mutations in Early-Onset Bladder Cancer. *J. Cancer* **7**, 915–920 (2016).

201. Heidenreich, B. *et al.* Telomerase reverse transcriptase promoter mutations in primary cutaneous melanoma. *Nat. Commun.* **5**, 3401 (2014).

202. Salvadores, M., Mas-Ponte, D. & Supek, F. Passenger mutations accurately classify human tumors. *PLoS Comput. Biol.* **15**, (2019).

203. McFarland, C. D. *et al.* The Damaging Effect of Passenger Mutations on Cancer Progression. *Cancer Res.* **77**, 4763–4772 (2017).

204. Van Hoeck, A., Tjoonk, N. H., van Boxtel, R. & Cuppen, E. Portrait of a cancer: mutational signature analyses for cancer diagnostics. *BMC Cancer* **19**, 457 (2019).

205. Xue, Y. & Wilcox, W. R. Changing paradigm of cancer therapy: precision medicine by next-generation sequencing. *Cancer Biol. Med.* **13**, 12–18 (2016).

206. Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214–218 (2013).

207. Polak, P. *et al.* mutational landscape of cancer. *Nature* **518**, 360–364 (2015).

208. Chalmers, Z. R. *et al.* Analysis of 100,000 human cancer genomes reveals the landscape of tumor mutational burden. *Genome Med.* **9**, 34 (2017).

209. Simpson, A. J. The natural somatic mutation frequency and human carcinogenesis. *Adv. Cancer Res.* **71**, 209–240 (1997).

210. Cui, P. *et al.* Distinct contributions of replication and transcription to mutation rate variation of human genomes. *Genomics. Proteomics Bioinformatics* **10**, 4–10 (2012).

211. Gonzalez-Perez, A., Sabarinathan, R. & Lopez-Bigas, N. Local Determinants of the

Mutational Landscape of the Human Genome. *Cell* **177**, 101–114 (2019).

212. Sima, J. & Gilbert, D. M. Complex correlations: replication timing and mutational landscapes during cancer and genome evolution. *Curr. Opin. Genet. Dev.* **25**, 93–100 (2014).

213. Kaiser, V. B., Taylor, M. S. & Semple, C. A. Mutational Biases Drive Elevated Rates of Substitution at Regulatory Sites across Cancer Types. *PLOS Genet.* **12**, e1006207 (2016).

214. Lee, C. A., Abd-Rabbo, D. & Reimand, J. Functional and genetic determinants of mutation rate variability in regulatory elements of cancer genomes. *Genome Biol.* **22**, 133 (2021).

215. Sabarinathan, R., Mularoni, L., Deu-Pons, J., Gonzalez-Perez, A. & Lopez-Bigas, N. Nucleotide excision repair is impaired by binding of transcription factors to DNA. *Nature* **532**, 264–267 (2016).

216. Alexandrov, L. B. & Stratton, M. R. Mutational signatures: the patterns of somatic mutations hidden in cancer genomes. *Curr. Opin. Genet. Dev.* **24**, 52–60 (2014).

217. Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).

218. Brash, D. E. UV signature mutations. *Photochem. Photobiol.* **91**, 15–26 (2015).

219. Paul, P., Malakar, A. K. & Chakraborty, S. The significance of gene mutations across eight major cancer types. *Mutat. Res. Mutat. Res.* **781**, 88–99 (2019).

220. Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).

221. Alexandrov, L. B. *et al.* Clock-like mutational processes in human somatic cells. *Nat. Genet.* **47**, 1402–1407 (2015).

222. Nik-Zainal, S. *et al.* Mutational processes molding the genomes of 21 breast cancers. *Cell* **149**, 979–993 (2012).

223. Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Campbell, P. J. & Stratton, M. R. Deciphering Signatures of Mutational Processes Operative in Human Cancer. *Cell Rep.* **3**, 246–259 (2013).

224. Supawadee, C. *et al.* Mutational spectra of aflatoxin B1 in vivo establish biomarkers of exposure for human hepatocellular carcinoma. *Proc. Natl. Acad. Sci.* **114**, E3101–E3109 (2017).

225. Nik-Zainal, S. *et al.* The genome as a record of environmental exposure. *Mutagenesis* **30**, 763–770 (2015).

226. Kucab, J. E. *et al.* A Compendium of Mutational Signatures of Environmental Agents. *Cell* **177**, 821-836.e16 (2019).

227. Meier, B. *et al.* Mutational signatures of DNA mismatch repair deficiency in C. elegans and human cancers. *Genome Res.* **28**, 666–675 (2018).

228. Haradhvala, N. J. *et al.* Distinct mutational signatures characterize concurrent loss of polymerase proofreading and mismatch repair. *Nat. Commun.* **9**, 1746 (2018).

229. Chan, K. *et al.* An APOBEC3A hypermutation signature is distinguishable from the signature of background mutagenesis by APOBEC3B in human cancers. *Nat. Genet.* **47**, 1067–1072 (2015).

230. Tate, J. G. *et al.* COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res.* **47**, D941–D947 (2019).

231. Rübben, A. & Araujo, A. Cancer heterogeneity: converting a limitation into a source of biologic information. *J. Transl. Med.* **15**, 190 (2017).

232. Vogelstein, B. & Kinzler, K. W. Cancer genes and the pathways they control. *Nat. Med.* **10**, 789–799 (2004).

233. Iorio, F. *et al.* Pathway-based dissection of the genomic heterogeneity of cancer hallmarks' acquisition with SLAPenrich. *Sci. Rep.* **8**, 6713 (2018).

234. Colaprico, A. *et al.* Interpreting pathways to discover cancer driver genes with Moonlight. *Nat. Commun.* **11**, 69 (2020).

235. Reyna, M. A. *et al.* Pathway and network analysis of more than 2500 whole cancer genomes. *Nat. Commun.* **11**, 729 (2020).

236. Iengar, P. Identifying pathways affected by cancer mutations. *Genomics* **110**, 318–328 (2018).

237. Sanchez-Vega, F. *et al.* Oncogenic Signaling Pathways in The Cancer Genome

Atlas. *Cell* **173**, 321-337.e10 (2018).

238. Gimple, R. C. & Wang, X. RAS: Striking at the Core of the Oncogenic Circuitry . *Frontiers in Oncology* vol. 9 (2019).

239. Seth, R. *et al.* Concomitant mutations and splice variants in KRAS and BRAF demonstrate complex perturbation of the Ras/Raf signalling pathway in advanced colorectal cancer. *Gut* **58**, 1234–1241 (2009).

240. Rajagopalan, H. *et al.* Tumorigenesis: RAF/RAS oncogenes and mismatch-repair status. *Nature* **418**, 934 (2002).

241. Alsina, J. *et al.* Detection of mutations in the mitogen-activated protein kinase pathway in human melanoma. *Clin. cancer Res. an Off. J. Am. Assoc. Cancer Res.* **9**, 6419–6425 (2003).

242. Borràs, E. *et al.* Clinical pharmacogenomic testing of KRAS, BRAF and EGFR mutations by high resolution melting analysis and ultra-deep pyrosequencing. *BMC Cancer* **11**, 406 (2011).

243. Kinno, T. *et al.* Clinicopathological features of nonsmall cell lung carcinomas with BRAF mutations. *Ann. Oncol. Off. J. Eur. Soc. Med. Oncol.* **25**, 138–142 (2014).

244. Pao, W. & Girard, N. New driver mutations in non-small-cell lung cancer. *Lancet. Oncol.* **12**, 175–180 (2011).

245. Vandin, F., Upfal, E. & Raphael, B. J. De novo discovery of mutated driver pathways in cancer. *Genome Res.* **22**, 375–385 (2012).

246. Ciriello, G., Cerami, E., Sander, C. & Schultz, N. Mutual exclusivity analysis identifies oncogenic network modules. *Genome Res.* **22**, 398–406 (2012).

247. Cisowski, J. & Bergo, M. O. What makes oncogenes mutually exclusive? *Small GTPases* **8**, 187–192 (2017).

248. García, Z., Kumar, A., Marqués, M., Cortés, I. & Carrera, A. C. Phosphoinositide 3-kinase controls early and late events in mammalian cell division. *EMBO J.* **25**, 655–661 (2006).

249. Engelman, J. A., Luo, J. & Cantley, L. C. The evolution of phosphatidylinositol 3-kinases as regulators of growth and metabolism. *Nat. Rev. Genet.* **7**, 606–619 (2006).

250. Asati, V., Mahapatra, D. K. & Bharti, S. K. PI3K/Akt/mTOR and Ras/Raf/MEK/ERK signaling pathways inhibitors as anticancer agents: Structural and pharmacological perspectives. *Eur. J. Med. Chem.* **109**, 314–341 (2016).

251. Yang, J. *et al.* Targeting PI3K in cancer: mechanisms and advances in clinical trials. *Mol. Cancer* **18**, 26 (2019).

252. Chabner, B. A. Antineoplastic agents. *Goodman Gilman's pharmacyological basis Ther.* (1996).

253. Johnstone, R. W., Ruefli, A. A. & Lowe, S. W. Apoptosis: A Link between Cancer Genetics and Chemotherapy. *Cell* **108**, 153–164 (2002).

254. Lee, Y. T., Tan, Y. J. & Oon, C. E. Molecular targeted therapy: treating cancer with specificity. *Eur. J. Pharmacol.* **834**, 188–196 (2018).

255. Bedard, P. L., Hyman, D. M., Davids, M. S. & Siu, L. L. Small molecules, big impact: 20 years of targeted therapy in oncology. *Lancet (London, England)* **395**, 1078–1088 (2020).

256. Wilkes, G. M. Targeted Therapy: Attacking Cancer with Molecular and Immunological Targeted Agents. *Asia-Pacific J. Oncol. Nurs.* **5**, 137–155 (2018).

257. Buchdunger, E. *et al.* Inhibition of the Abl protein-tyrosine kinase in vitro and in vivo by a 2-phenylaminopyrimidine derivative. *Cancer Res.* **56**, 100–104 (1996).

258. Smith, K. M., Yacobi, R. & Van Etten, R. A. Autoinhibition of Bcr-Abl through its SH3 domain. *Mol. Cell* **12**, 27–37 (2003).

259. Van Etten, R. A. c-Abl regulation: a tail of two lipids. *Curr. Biol.* **13**, R608-10 (2003).

260. Maekawa, T., Ashihara, E. & Kimura, S. The Bcr-Abl tyrosine kinase inhibitor imatinib and promising new agents against Philadelphia chromosome-positive leukemias. *Int. J. Clin. Oncol.* **12**, 327–340 (2007).

261. Debiec-Rychter, M. *et al.* Use of c-KIT/PDGFRA mutational analysis to predict the clinical response to imatinib in patients with advanced gastrointestinal stromal tumours entered on phase I and II studies of the EORTC Soft Tissue and Bone Sarcoma Group. *Eur. J. Cancer* **40**, 689–695 (2004).

262. Debiec-Rychter, M. *et al.* KIT mutations and dose selection for imatinib in patients

with advanced gastrointestinal stromal tumours. *Eur. J. Cancer* **42**, 1093–1103 (2006).

263. Heinrich, M. C. *et al.* Kinase mutations and imatinib response in patients with metastatic gastrointestinal stromal tumor. *J. Clin. Oncol.* **21**, 4342–4349 (2003).

264. Corless, C. L. *et al.* PDGFRA mutations in gastrointestinal stromal tumors: frequency, spectrum and in vitro sensitivity to imatinib. *J. Clin. Oncol.* **23**, 5357–5364 (2005).

265. Heinrich, M. C. *et al.* Primary and secondary kinase genotypes correlate with the biological and clinical activity of sunitinib in imatinib-resistant gastrointestinal stromal tumor. *J. Clin. Oncol.* **26**, 5352 (2008).

266. Zhong, L. *et al.* Small molecules in targeted cancer therapy: advances, challenges, and future perspectives. *Signal Transduct. Target. Ther.* **6**, 201 (2021).

267. Meyerholz, D. K., Suarez, C. J., Dintzis, S. M. & Frevert, C. W. 9 - Respiratory System. in (eds. Treuting, P. M., Dintzis, S. M. & Montine, K. S. B. T.-C. A. and H. (Second E.) 147–162 (Academic Press, 2018). doi:https://doi.org/10.1016/B978-0-12-802900-8.00009-9.

268. Greeley, M. A. Chapter 4 - Respiratory System. in (eds. Parker, G. A. & Picut, C. A. B. T.-A. of H. of the J. R.) 89–125 (Academic Press, 2016). doi:https://doi.org/10.1016/B978-0-12-802682-3.00004-5.

269. Haschek, W. M., Witschi, H. R. & Nikula, K. J. 28 - Respiratory System. in (eds. HASCHEK, W. M., ROUSSEAUX, C. G. & WALLIG, M. A. B. T.-H. of T. P. (Second E.) 3–83 (Academic Press, 2002). doi:https://doi.org/10.1016/B978-012330215-1/50029-6.

270. Whitsett, J. A., Kalin, T. V., Xu, Y. & Kalinichenko, V. V. Building and regenerating the lung cell by cell. *Physiol. Rev.* **99**, 513–554 (2019).

271. Adivitiya, Kaushik, M. S., Chakraborty, S., Veleri, S. & Kateriya, S. Mucociliary Respiratory Epithelium Integrity in Molecular Defense and Susceptibility to Pulmonary Viral Infections. *Biology* vol. 10 (2021).

272. Sung, H. *et al.* Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA. Cancer J. Clin.* **71**, 209–249 (2021).

273. Youlden, D. R., Cramb, S. M. & Baade, P. D. The International Epidemiology of Lung Cancer: geographical distribution and secular trends. *J. Thorac. Oncol. Off. Publ. Int. Assoc. Study Lung Cancer* **3**, 819–831 (2008).

274. Hecht, S. S. Tobacco smoke carcinogens and lung cancer. *JNCI J. Natl. Cancer Inst.* **91**, 1194–1210 (1999).

275. Hackshaw, A. K., Law, M. R. & Wald, N. J. The accumulated evidence on lung cancer and environmental tobacco smoke. *Bmj* **315**, 980–988 (1997).

276. Yao, S. X. *et al.* Exposure to radon progeny, tobacco use and lung cancer in a case-control study in southern China. *Radiat. Res.* **138**, 326–336 (1994).

277. Krewski, D. *et al.* Residential radon and risk of lung cancer: a combined analysis of 7 North American case-control studies. *Epidemiology* **16**, 137–145 (2005).

278. Wilcox, H. B. *et al.* Case-control study of radon and lung cancer in New Jersey. *Radiat. Prot. Dosimetry* **128**, 169–179 (2008).

279. Wei, S., Zhang, H. & Tao, S. A review of arsenic exposure and lung cancer. *Toxicol. Res. (Camb).* **8**, 319–327 (2019).

280. Putila, J. J. & Guo, N. L. Association of arsenic exposure with lung cancer incidence rates in the United States. *PLoS One* **6**, e25886–e25886 (2011).

281. Gibb, H. J., Lees, P. S., Pinsky, P. F. & Rooney, B. C. Lung cancer among workers in chromium chemical production. *Am. J. Ind. Med.* **38**, 115–126 (2000).

282. Grimsrud, T. K., Berge, S. R., Haldorsen, T. & Andersen, A. Exposure to different forms of nickel and risk of lung cancer. *Am. J. Epidemiol.* **156**, 1123–1132 (2002).

283. Shen, H. M. & Zhang, Q. F. Risk assessment of nickel carcinogenicity and occupational lung cancer. *Environ. Health Perspect.* **102**, 275–282 (1994).

284. Alberg, A. J., Brock, M. V, Ford, J. G., Samet, J. M. & Spivack, S. D. Epidemiology of lung cancer: Diagnosis and management of lung cancer, 3rd ed: American College of Chest Physicians evidence-based clinical practice guidelines. *Chest* **143**, e1S-e29S (2013).

285. Dela Cruz, C. S., Tanoue, L. T. & Matthay, R. A. Lung cancer: epidemiology, etiology, and prevention. *Clin. Chest Med.* **32**, 605–644 (2011).

286. Travis, W. D. Classification of Lung Cancer. *Semin. Roentgenol.* **46**, 178–186 (2011).

287. Myers, D. J. & Wallen, J. M. Lung Adenocarcinoma. in (2022).

288. Sabbula, B. R. & Anjum, F. Squamous Cell Lung Cancer. in (2022).

289. Raso, M. G., Bota-Rabassedas, N. & Wistuba, I. I. Pathology and Classification of SCLC. *Cancers (Basel).* **13**, (2021).

290. Yang, P. Epidemiology of lung cancer prognosis: Quantity and quality of life. *Methods Mol. Biol.* **471**, 469–486 (2009).

291. Woodard, G. A., Jones, K. D. & Jablons, D. M. Lung Cancer Staging and Prognosis. *Cancer Treat. Res.* **170**, 47–75 (2016).

292. Akhurst, T. Staging of Non-Small-Cell Lung Cancer. *PET Clin.* **13**, 1–10 (2018).

293. Lang-Lazdunski, L. Surgery for nonsmall cell lung cancer. *Eur. Respir. Rev.* **22**, 382–404 (2013).

294. Yoon, S. M., Shaikh, T. & Hallman, M. Therapeutic management options for stage III non-small cell lung cancer. *World J. Clin. Oncol.* **8**, 1–20 (2017).

295. Gressen, E. L. & Curran, W. J. J. Inoperable localized stage I and stage II non-small-cell lung cancer. *Curr. Treat. Options Oncol.* **3**, 75–83 (2002).

296. Provencio, M., Isla, D., Sánchez, A. & Cantos, B. Inoperable stage III non-small cell lung cancer: Current treatment and role of vinorelbine. *J. Thorac. Dis.* **3**, 197–204 (2011).

297. Yang, S., Zhang, Z. & Wang, Q. Emerging therapies for small cell lung cancer. *J. Hematol. Oncol.* **12**, 47 (2019).

298. van Meerbeeck, J. P., Fennell, D. A. & De Ruysscher, D. K. M. Small-cell lung cancer. *Lancet (London, England)* **378**, 1741–1755 (2011).

299. Bernhardt, E. B. & Jalal, S. I. Small Cell Lung Cancer. *Cancer Treat. Res.* **170**, 301–322 (2016).

300. Testa, U., Castelli, G. & Pelosi, E. Lung Cancers: Molecular Characterization, Clonal Heterogeneity and Evolution, and Cancer Stem Cells. *Cancers (Basel).* **10**, 248 (2018).

301. Imielinski, M. *et al.* Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. *Cell* **150**, 1107–1120 (2012).

302. Judd, J. *et al.* Characterization of KRAS Mutation Subtypes in Non–small Cell Lung Cancer. *Mol. Cancer Ther.* **20**, 2577–2584 (2021).

303. Karachaliou, N. *et al.* KRAS mutations in lung cancer. *Clin. Lung Cancer* **14**, 205–214 (2013).

304. Reck, M. & Rabe, K. F. Precision diagnosis and treatment for advanced non–small-cell lung cancer. *N. Engl. J. Med.* **377**, 849–861 (2017).

305. Thunnissen, E., van der Oord, K. & den Bakker, M. Prognostic and predictive biomarkers in lung cancer. A review. *Virchows Arch.* **464**, 347–358 (2014).

306. Ahmadzada, T. *et al.* An Update on Predictive Biomarkers for Treatment Selection in Non-Small Cell Lung Cancer. *J. Clin. Med.* **7**, (2018).

307. Nan, X., Xie, C., Yu, X. & Liu, J. EGFR TKI as first-line treatment for patients with advanced EGFR mutation-positive non-small-cell lung cancer. *Oncotarget* **8**, 75712–75726 (2017).

308. Schvartzman, J. M., Thompson, C. B. & Finley, L. W. S. Metabolic regulation of chromatin modifications and gene expression. *J. Cell Biol.* jcb.201803061 (2018) doi:10.1083/jcb.201803061.

309. Thunnissen, E. *et al.* EML4-ALK testing in non-small cell carcinomas of the lung: a review with recommendations. *Virchows Arch.* **461**, 245–257 (2012).

310. Corces, M. R. *et al.* An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. *Nat. Methods* **14**, 959–962 (2017).

311. Andrews, S. FastQC: a quality control tool for high throughput sequence data. 2010. (2017).

312. Andrews, S. FastQC: a quality control tool for high throughput sequence data. (2010).

313. Krueger, F. Trim galore. *A wrapper tool around Cutadapt FastQC to consistently apply Qual. Adapt. trimming to FastQ files* **516**, (2015).

314. Li, H. & Wren, J. Toward better understanding of artifacts in variant calling from

high-coverage samples. *Bioinformatics* **30**, 2843–2851 (2014).

315. Institute, B. Picard tools. (2016).

316. Li, H. *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

317. Ramírez, F., Dündar, F., Diehl, S., Grüning, B. A. & Manke, T. deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Res.* **42**, W187–W191 (2014).

318. Fu, S. *et al.* Differential analysis of chromatin accessibility and histone modifications for predicting mouse developmental enhancers. *Nucleic Acids Res.* **46**, 11184–11201 (2018).

319. Ewels, P. A. *et al.* The nf-core framework for community-curated bioinformatics pipelines. *Nat. Biotechnol.* **38**, 276–278 (2020).

320. Thorvaldsdóttir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.* **14**, 178–192 (2013).

321. Pruitt, K. D., Tatusova, T. & Maglott, D. R. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* **35**, D61–D65 (2007).

322. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).

323. Van der Auwera, G. A. & O'Connor, B. D. *Genomics in the cloud: using Docker, GATK, and WDL in Terra*. (O'Reilly Media, 2020).

324. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. *arXiv Prepr. arXiv1207.3907* (2012).

325. Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* **31**, 213–219 (2013).

326. Fang, H. *et al.* Indel variant analysis of short-read sequencing data with Scalpel. *Nat. Protoc.* **11**, 2529–2548 (2016).

327. Lai, Z. *et al.* VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. *Nucleic Acids Res.* **44**, e108 (2016).

328. Koboldt, D. C. *et al.* VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* **22**, 568–576 (2012).

329. Chapman, B. & Core, B. Interoperable community developed variant calling with bcbio and the Common Work ow Language. (2016).

330. Dale, R. K., Pedersen, B. S. & Quinlan, A. R. Pybedtools: A flexible Python library for manipulating genomic datasets and annotations. *Bioinformatics* **27**, 3423–3424 (2011).

331. Blokzijl, F., Janssen, R., van Boxtel, R. & Cuppen, E. MutationalPatterns: comprehensive genome-wide analysis of mutational processes. *Genome Med.* **10**, 1–11 (2018).

332. Collings, B. J. & Hamilton, M. A. Estimating the power of the two-sample Wilcoxon test for location shift. *Biometrics* 847–860 (1988).

333. Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).

334. Lawlor, N. *et al.* Multiomic Profiling Identifies cis-Regulatory Networks Underlying Human Pancreatic β Cell Identity and Function. *Cell Rep.* **26**, 788-801.e6 (2019).

335. Jin, F. *et al.* A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature* **503**, 290–294 (2013).

336. Barutcu, A. R. *et al.* Chromatin interaction analysis reveals changes in small chromosome and telomere clustering between epithelial and breast cancer cells. *Genome Biol.* **16**, 214 (2015).

337. Bunting, K. L. *et al.* Multi-tiered Reorganization of the Genome during B Cell Affinity Maturation Anchored by a Germinal Center-Specific Locus Control Region. *Immunity* **45**, 497–512 (2016).

338. Schmitt, A. D. *et al.* A Compendium of Chromatin Contact Maps Reveals Spatially Active Regions in the Human Genome. *Cell Rep.* **17**, 2042–2059 (2016).

339. Abdennur, N. & Mirny, L. A. Cooler: scalable storage for Hi-C data and other

genomically labeled arrays. *Bioinformatics* **36**, 311–316 (2020).

340. Pal, K. *et al.* Global chromatin conformation differences in the Drosophila dosage compensated chromosome X. *Nat. Commun.* **10**, 5355 (2019).

341. Pal, K., Tagliaferri, I., Livi, C. M. & Ferrari, F. HiCBricks: building blocks for efficient handling of large Hi-C datasets. *Bioinformatics* **36**, 1917–1919 (2019).

342. Mardia, K. V. *Multivariate analysis*. (1979).

343. Lei, L. & Fithian, W. AdaPT: an interactive procedure for multiple testing with side information. *J. R. Stat. Soc. Ser. B (Statistical Methodol.* **80**, 649–679 (2018).

344. Lonsdale, J. *et al.* The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).

345. McKnight, P. E. & Najab, J. Mann-Whitney U Test. *Corsini Encycl. Psychol.* 1 (2010).

346. Weisstein, E. W. Bonferroni correction. *https//mathworld. wolfram. com/* (2004).

347. Raudvere, U. *et al.* g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Res.* **47**, W191–W198 (2019).

348. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525 (2016).

349. Mermel, C. H. *et al.* GISTIC2. 0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.* **12**, 1–14 (2011).

350. Yang, L. *et al.* Diverse mechanisms of somatic structural variations in human cancer genomes. *Cell* **153**, 919–929 (2013).

351. Stajich, J. E. *et al.* The Bioperl toolkit: Perl modules for the life sciences. *Genome Res.* **12**, 1611–1618 (2002).

352. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).

353. Kaplan, E. L. & Meier, P. Nonparametric Estimation from Incomplete Observations. *J. Am. Stat. Assoc.* **53**, 457–481 (1958).

354. Peto, R. & Peto, J. Asymptotically Efficient Rank Invariant Test Procedures. *J. R. Stat. Soc. Ser. A* **135**, 185–207 (1972).

355. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci.* **102**, 15545 LP – 15550 (2005).

356. Collisson, E. A. *et al.* Comprehensive molecular profiling of lung adenocarcinoma. *Nature* **511**, 543–550 (2014).

357. Hammerman, P. S. *et al.* Comprehensive genomic characterization of squamous cell lung cancers. *Nature* **489**, 519–525 (2012).

358. George, J. *et al.* Comprehensive genomic profiles of small cell lung cancer. *Nature* **524**, 47–53 (2015).

359. Gómez-González, B. & Aguilera, A. Transcription-mediated replication hindrance: a major driver of genome instability. *Genes Dev.* **33**, 1008–1026 (2019).

360. Wu, W. *et al.* Neuronal enhancers are hotspots for DNA single-strand break repair. *Nature* **593**, 440–444 (2021).

361. Hariprakash, J. M. & Ferrari, F. Computational Biology Solutions to Identify Enhancers-target Gene Pairs. *Comput. Struct. Biotechnol. J.* **17**, 821–831 (2019).

362. Khurana, E. *et al.* Role of non-coding sequence variants in cancer. *Nat. Rev. Genet.* **17**, 93–108 (2016).

363. Weinhold, N., Jacobsen, A., Schultz, N., Sander, C. & Lee, W. Genome-wide analysis of noncoding regulatory mutations in cancer. *Nat. Genet.* **46**, 1160–1165 (2014).

364. Scacheri, C. A. & Scacheri, P. C. Mutations in the noncoding genome. *Curr. Opin. Pediatr.* **27**, 659–664 (2015).

365. Matsuda, R. *et al.* LY6K is a novel molecular target in bladder cancer on basis of integrate genome-wide profiling. *Br. J. Cancer* **104**, 376–386 (2011).

366. Ishikawa, H. *et al.* Phase I clinical trial of vaccination with LY6K-derived peptide in patients with advanced gastric cancer. *Gastric Cancer* **17**, 173–180 (2014).

367. Kong, H. K., Yoon, S. & Park, J. H. The regulatory mechanism of the LY6K gene

expression in human breast cancer cells. *J. Biol. Chem.* **287**, 38889–38900 (2012).

368. Kong, H. K. *et al.* Epigenetic activation of LY6K predicts the presence of metastasis and poor prognosis in breast carcinoma. *Oncotarget* **7**, 55677–55689 (2016).

369. Spitz, F. & Furlong, E. E. M. Transcription factors: from enhancer binding to developmental control. *Nat. Rev. Genet.* **13**, 613–626 (2012).

370. Zogopoulos, V. *et al.* TFBSPred: A functional transcription factor binding site prediction webtool for humans and mice. *Int. J. Epigenetics* **1**, 1–11 (2021).

371. Castellanos, M., Mothi, N. & Muñoz, V. Eukaryotic transcription factors can track and control their target genes using DNA antennas. *Nat. Commun.* **11**, 540 (2020).

372. Bryne, J. C. *et al.* JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. *Nucleic Acids Res.* **36**, D102–D106 (2007).

373. Badis, G. *et al.* Diversity and complexity in DNA recognition by transcription factors. *Science* **324**, 1720–1723 (2009).

374. Vinson, C., Chatterjee, R. & Fitzgerald, P. Transcription factor binding sites and other features in human and Drosophila proximal promoters. *Subcell. Biochem.* **52**, 205–222 (2011).

375. Kribelbauer, J. F., Rastogi, C., Bussemaker, H. J. & Mann, R. S. Low-Affinity Binding Sites and the Transcription Factor Specificity Paradox in Eukaryotes. *Annu. Rev. Cell Dev. Biol.* **35**, 357–379 (2019).

376. Rogers, J. M. & Bulyk, M. L. Diversification of transcription factor-DNA interactions and the evolution of gene regulatory networks. *Wiley Interdiscip. Rev. Syst. Biol. Med.* e1423–e1423 (2018) doi:10.1002/wsbm.1423.

377. Kasowski, M. *et al.* Variation in transcription factor binding among humans. *Science* **328**, 232–235 (2010).

378. Kilpinen, H. *et al.* Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription. *Science* **342**, 744–747 (2013).

379. Spivakov, M. *et al.* Analysis of variation at transcription factor binding sites in Drosophila and humans. *Genome Biol.* **13**, R49 (2012).

380. Stobbe, M. D. *et al.* Recurrent somatic mutations reveal new insights into consequences of mutagenic processes in cancer. *PLOS Comput. Biol.* **15**, e1007496 (2019).

381. Palmisano, W. A. *et al.* Aberrant promoter methylation of the transcription factor genes PAX5 α and β in human cancers. *Cancer Res.* **63**, 4620–4625 (2003).

382. Polak, P. *et al.* Cell-of-origin chromatin organization shapes the mutational landscape of cancer. *Nature* **518**, 360–364 (2015).

383. Bejerano, G. *et al.* Ultraconserved Elements in the Human Genome. *Science (80-.).* **304**, 1321 LP – 1325 (2004).

384. Pessina, F. *et al.* Functional transcription promoters at DNA double-strand breaks mediate RNA-driven phase separation of damage-response factors. *Nat. Cell Biol.* **21**, 1286–1299 (2019).

385. Hung, S. *et al.* Mismatch repair-signature mutations activate gene enhancers across human colorectal cancer epigenomes. *Elife* **8**, e40760 (2019).

386. Hazan, I., Monin, J., Bouwman, B. A. M., Crosetto, N. & Aqeilan, R. I. Activation of Oncogenic Super-Enhancers Is Coupled with DNA Repair by RAD51. *Cell Rep.* **29**, 560-572.e4 (2019).

387. Lin, C. Y. *et al.* Active medulloblastoma enhancers reveal subgroup-specific cellular origins. *Nature* **530**, 57–62 (2016).

388. Chepelev, I., Wei, G., Wangsa, D., Tang, Q. & Zhao, K. Characterization of genome-wide enhancer-promoter interactions reveals co-expression of interacting genes and modes of higher order chromatin organization. *Cell Res.* **22**, 490–503 (2012).

389. Chen, H. *et al.* Landscape of Enhancer-Enhancer Cooperative Regulation during Human Cardiac Commitment. *Mol. Ther. - Nucleic Acids* **17**, 840–851 (2019).

APPENDIX

Appendix Table 1: Samples used in the analysis with WGS coverage information and availability of RNA-seq data.

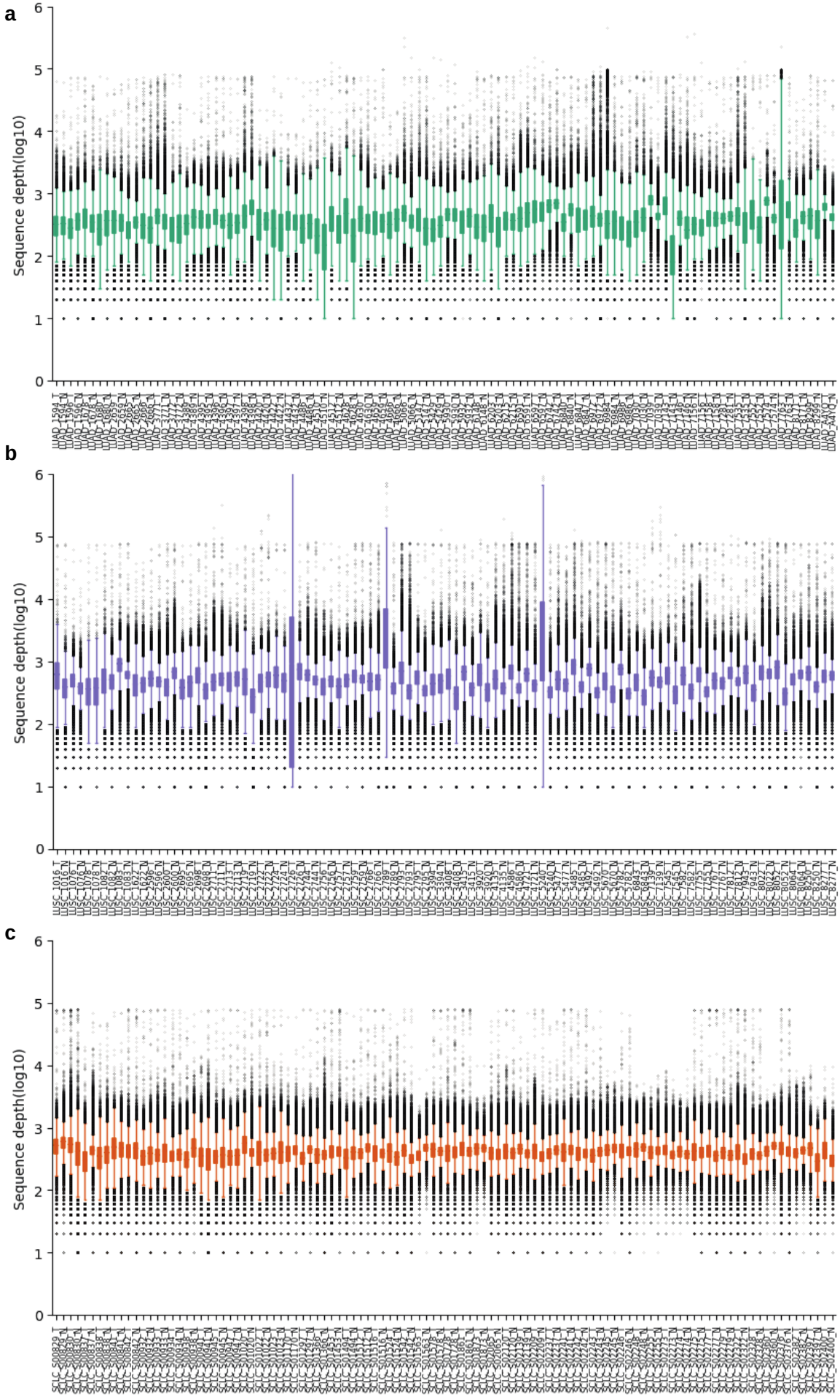
Sample Name	Subtype	Coverage (Tumour)	Coverage (Normal)	RNA seq Availability
SCLC_S00837	SCLC	32	43	-
SCLC_S01297	SCLC	32	44	Y
SCLC_S00938	SCLC	34	53	-
SCLC_S00945	SCLC	34	39	-
SCLC_S00932	SCLC	34	40	-
SCLC_S01563	SCLC	35	46	-
SCLC_S00947	SCLC	35	37	-
TCGA-55-6984	LUAD	35	43	Y
SCLC_S01494	SCLC	36	42	-
SCLC_S00934	SCLC	36	39	-
TCGA-78-7156	LUAD	36	42	Y
TCGA-05-4389	LUAD	36	45	Y
SCLC_S01366	SCLC	36	36	Y
SCLC_S02243	SCLC	36	40	Y
SCLC_S02328	SCLC	37	35	Y
TCGA-64-1680	LUAD	37	46	Y
TCGA-05-4397	LUAD	37	36	Y
SCLC_S02237	SCLC	38	44	-
SCLC_S00933	SCLC	38	45	-
SCLC_S00941	SCLC	38	39	-
SCLC_S01023	SCLC	38	47	-
TCGA-49-4486	LUAD	38	38	Y
SCLC_S02242	SCLC	38	40	Y
TCGA-05-5429	LUAD	38	38	Y
SCLC_S02382	SCLC	38	43	Y
SCLC_S00838	SCLC	38	41	Y
TCGA-91-6840	LUAD	38	61	Y
SCLC_S01516	SCLC	39	39	-
SCLC_S01512	SCLC	39	46	Y

TCGA-75-7030	LUAD	39	47	Y
SCLC_S02255	SCLC	39	42	Y
TCGA-05-4432	LUAD	39	37	Y
TCGA-73-4666	LUAD	39	41	Y
SCLC_S01170	SCLC	40	44	-
SCLC_S02279	SCLC	40	36	-
TCGA-49-4510	LUAD	40	30	Y
TCGA-05-4420	LUAD	40	38	Y
TCGA-73-4659	LUAD	40	39	Y
SCLC_S02065	SCLC	40	39	Y
SCLC_S02274	SCLC	41	37	-
TCGA-34-2600	LUSC	41	70	Y
TCGA-97-8171	LUAD	41	43	Y
TCGA-55-1596	LUAD	41	55	Y
TCGA-50-5932	LUAD	41	47	Y
SCLC_S02273	SCLC	42	36	-
SCLC_S01728	SCLC	42	43	-
SCLC_S02241	SCLC	42	44	Y
TCGA-55-7281	LUAD	42	46	Y
SCLC_S02322	SCLC	42	33	Y
SCLC_S02360	SCLC	42	50	Y
SCLC_S01453	SCLC	43	42	-
TCGA-44-2659	LUAD	43	37	Y
SCLC_S02209	SCLC	43	35	Y
SCLC_S02248	SCLC	43	47	Y
TCGA-05-4395	LUAD	43	40	Y
TCGA-44-2666	LUAD	43	58	Y
TCGA-78-7146	LUAD	43	47	Y
SCLC_S02245	SCLC	44	47	-
SCLC_S02277	SCLC	44	37	-
SCLC_S01578	SCLC	44	42	Y
SCLC_S01542	SCLC	44	32	Y
TCGA-44-6148	LUAD	44	45	Y
SCLC_S02246	SCLC	44	42	Y

SCLC_S01873	SCLC	44	46	Y
SCLC_S01524	SCLC	44	36	Y
TCGA-55-6986	LUAD	44	38	Y
TCGA-67-3772	LUAD	44	52	Y
SCLC_S01022	SCLC	45	38	-
SCLC_S02400	SCLC	45	33	-
SCLC_S02275	SCLC	45	42	-
SCLC_S00842	SCLC	45	48	-
SCLC_S02120	SCLC	45	40	Y
TCGA-78-7158	LUAD	45	42	Y
TCGA-67-6215	LUAD	46	44	Y
TCGA-05-4396	LUAD	46	39	Y
SCLC_S01861	SCLC	46	41	Y
SCLC_S02376	SCLC	46	41	Y
TCGA-78-7535	LUAD	46	42	Y
TCGA-50-6591	LUAD	47	55	Y
SCLC_S02139	SCLC	47	38	Y
TCGA-66-2756	LUSC	47	55	Y
TCGA-64-1678	LUAD	47	37	Y
TCGA-55-6972	LUAD	47	40	Y
TCGA-91-6847	LUAD	48	44	Y
TCGA-50-5930	LUAD	48	47	Y
TCGA-67-3771	LUAD	49	53	Y
TCGA-34-5240	LUSC	49	35	Y
SCLC_S02397	SCLC	50	36	Y
TCGA-66-2757	LUSC	50	59	Y
TCGA-55-1594	LUAD	50	50	Y
TCGA-56-1622	LUSC	51	48	Y
SCLC_S00829	SCLC	52	59	Y
TCGA-60-2711	LUSC	52	58	Y
TCGA-05-4398	LUAD	52	68	Y
TCGA-75-5147	LUAD	53	38	Y
TCGA-21-1078	LUSC	53	51	Y
TCGA-60-2695	LUSC	53	55	Y

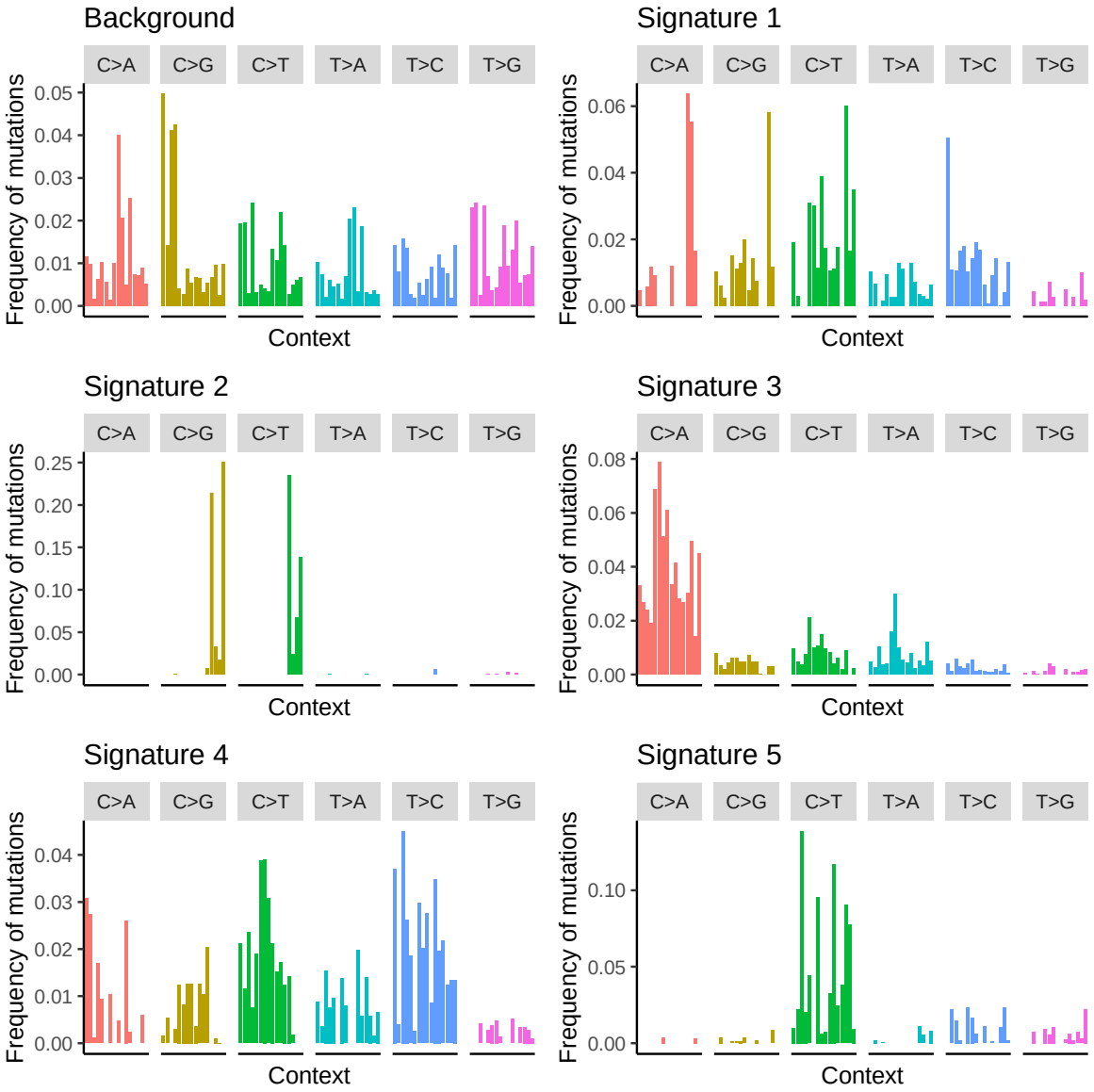
TCGA-44-2665	LUAD	53	55	Y
TCGA-05-4422	LUAD	53	42	Y
TCGA-43-3394	LUSC	54	54	Y
TCGA-60-2713	LUSC	54	60	Y
TCGA-38-4630	LUAD	55	41	Y
TCGA-66-2766	LUSC	55	57	Y
SCLC_S00830	SCLC	56	48	-
SCLC_S00841	SCLC	56	51	-
TCGA-34-2596	LUSC	57	49	Y
TCGA-NJ-A4YQ	LUAD	57	29	Y
TCGA-77-6843	LUSC	57	31	Y
SCLC_S01020	SCLC	58	41	-
TCGA-22-5477	LUSC	58	46	Y
TCGA-95-7039	LUAD	58	31	Y
TCGA-21-1076	LUSC	58	42	Y
TCGA-43-5670	LUSC	58	37	Y
TCGA-37-4135	LUSC	59	40	Y
TCGA-55-7574	LUAD	59	30	Y
TCGA-92-8064	LUSC	59	68	Y
TCGA-60-2722	LUSC	59	61	Y
TCGA-90-7767	LUSC	60	51	Y
TCGA-97-7552	LUAD	61	31	Y
TCGA-55-8299	LUAD	61	52	Y
TCGA-49-4512	LUAD	61	44	Y
TCGA-21-1082	LUSC	62	54	Y
TCGA-66-2795	LUSC	65	37	Y
TCGA-77-7139	LUSC	65	53	Y
TCGA-66-2759	LUSC	66	55	Y
TCGA-60-2698	LUSC	69	37	Y
TCGA-85-8277	LUSC	69	66	Y
TCGA-60-2724	LUSC	70	56	Y
TCGA-60-2719	LUSC	70	34	Y
TCGA-66-2744	LUSC	70	54	Y
TCGA-75-6203	LUAD	71	45	Y

TCGA-18-4721	LUSC	72	47	Y
TCGA-68-8250	LUSC	72	43	Y
TCGA-49-6742	LUAD	74	75	Y
TCGA-50-5066	LUAD	74	42	Y
TCGA-18-3408	LUSC	74	36	Y
TCGA-33-4586	LUSC	74	40	Y
TCGA-56-7582	LUSC	74	38	Y
TCGA-50-6597	LUAD	74	70	Y
TCGA-55-6982	LUAD	75	40	Y
TCGA-52-7812	LUSC	75	53	Y
TCGA-98-8022	LUSC	76	70	Y
TCGA-78-7143	LUAD	76	42	Y
TCGA-66-2793	LUSC	77	37	Y
TCGA-69-7763	LUAD	77	53	Y
TCGA-96-7545	LUSC	79	37	Y
TCGA-18-3415	LUSC	80	41	Y
TCGA-68-7755	LUSC	80	36	Y
TCGA-43-3920	LUSC	80	39	Y
TCGA-22-1016	LUSC	81	45	Y
TCGA-94-7943	LUSC	83	41	Y
TCGA-22-5492	LUSC	85	34	Y
TCGA-21-5782	LUSC	86	34	Y
TCGA-66-2789	LUSC	87	41	Y
TCGA-38-4628	LUAD	91	38	Y
TCGA-21-1083	LUSC	94	68	Y
TCGA-85-8052	LUSC	98	34	Y
TCGA-22-5485	LUSC	99	43	Y
TCGA-60-2726	LUSC	103	86	Y

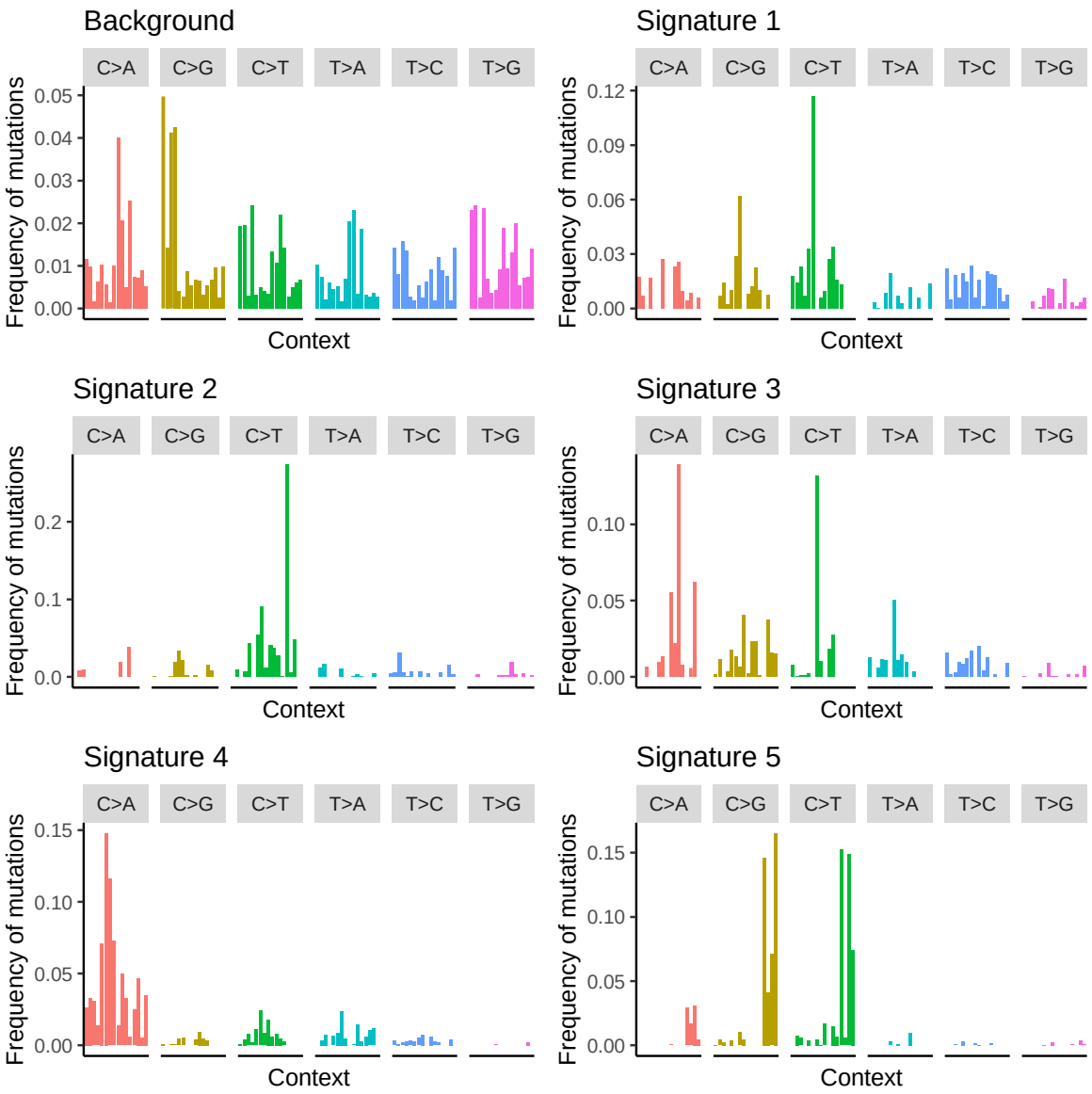

a
b
c
Sequence depth(log10)
0
1
2
3
4
5
6

Appendix Figure 1: Whole genome sequence coverage plot of (a) LUAD (b) LUSC (c) SCLC samples. *Box and whiskers plot show the sequence coverage at the sequence variations called in the samples.*

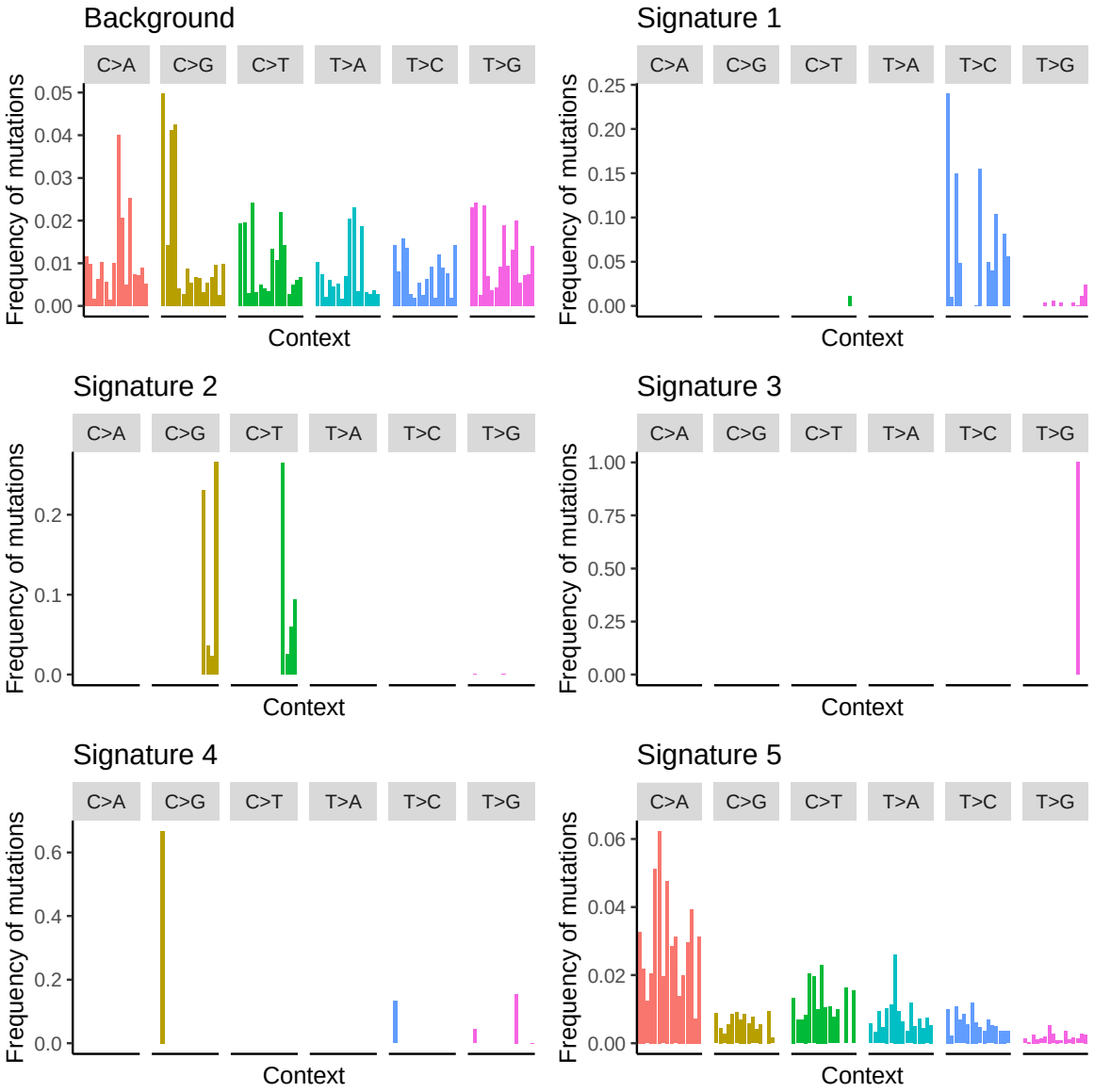
Appendix Figure 2: Sanger sequencing of clones selected after CRISPR deletion.



Appendix Figure 3: Mutation signature present in exons.



Appendix Figure 4:Mutation signatures present in promoters



Appendix Figure 5: Mutation signature present at enhancers