Universidade Federal de Minas Gerais
Instituto de Ciências Biológicas
Departamento de Bioquímica e Imunologia
Laboratório de Genética Bioquímica

# FC-R2: A comprehensive atlas of human long non-coding RNA expression using a standardized pipeline

Eddie Luidy Imada
Advisor: Prof. Dr. Glória Regina Franco
Co-Adivisor: Prof. Dr. Luigi Marchionni

Thesis submitted to the Department of Biochemistry and Immunology of the Biological Sciences Institute from Universidade Federal de Minas Gerais as a pre-requisite to obtaining the PhD title in bioinformatics

Belo Horizonte
May, 2019

# ACKNOWLEDGMENTS

This document is the end of a journey. When I first started it, I didn't know where it would lead me, nor the adversities I would face along it. I quickly learned that it takes more than a passion for doing science to go down this pathway. It takes the support of amazing people that I'm truly thankful to have had in my life. Without them, I couldn't have gone this far.

First, none of this would be possible without my family. To my parents, Celso and Nanami, I am forever thankful. All accomplishments in my life are the sheer reflection of all your efforts and dedication. To my brother and sister, Rafael and Tami, thank you for your unconditional friendship.

To my advisor, Glória, I thank you for always being understanding and not only sharing your knowledge but also your affection with me. It takes a good heart and true caring to put other people needs before your own, and you have done it countless times. Without your support, I wouldn't be the person I'm today.

I thank you my co-advisor, Luigi, for giving me the opportunity to join your lab and experience life and science abroad for the first time. What I learned from you, goes beyond professional skills, and I cannot stress enough how thankful I'm for all your efforts and support you gave me during these years.

To our collaborators, Leonardo, Chris, and Ben, for their Tamara and Emmanuel for their help in our projects. Without your help, this work wouldn’t be possible.

To my friends, both old and new, a big thank you! From my undergrad and Paraná state: Deco, Alan, Mayara, and Rafael to name a few, I thank you for all these years of friendship. Who would have guessed we would come this far since we first laid feet on the red earth of UEL. From my early years in the 'Gloriosos' group: André, Isadora, Dani Chame, Dani de Laet, Michele, Nayara, Willian, Tarcisio, Heron, and Tiago Bruno, I thank you for all the good times we shared both in- and outside work. To the 'Gloriosos' that I couldn’t spend much time since going abroad: Stella, Rafael, Thomas, Lúcio, and Gabriel, thank you for the, although brief, quality time we shared in the lab. From Baltimore: Barbara, Michelle, and Mayara thank you for your friendship and company during the year we stayed abroad. Even miles away, I will never forget our travels and time shared during this period. To Marina and Alex, a special thank you! You were the first friends I made in Baltimore and were always caring and supporting me. Without you, my experience

abroad would not be the same. Also, from Baltimore: Diego, Wikum and Alex, it was always a pleasure working with you guys.

To all members of the laboratory of genetics and biochemistry, for their valuable insights during our meetings.

To the Bioinformatics graduate program, for providing such great environment that was essential in my training. A special thank you to Sheila and Tiago for always being kind and helpful whenever I needed their help.

Last, I would like to thank the supporting agencies CAPES, CNPq, and FAPEMIG for investing in my professional training.

To everyone I mentioned and those that I probably forgot to mention (sorry), this achievement is yours!

# Resumo

Recentemente, estudos à fundo das funções e estrutura de genomas revelou que RNAs não codificadores desempenham um papel essencial no controle e regulação de processos biológicos e celulares através da regulação da expressão gênica. Estes mecanismos também começaram a ser elucidados em doenças humanas, destacando a importância da caracterização dos papéis desempenhados pelos RNAs não-codificadores em doenças, como o câncer. Neste trabalho, nós construímos um atlas de expressão gênica do transcriptoma humano contendo mais de 100.000 genes fazendo uso de dois recursos públicos: o transcriptoma associado à CAGE do projeto FANTOM (do inglês FANTOM-CAT) e o *recount2,* denominado FC-R2. O FANTOM-CAT é uma meta-montagem completa do transcriptoma humano contendo ambos genes codificadores e não-codificadores, incluindo promotores, *enhancers* e RNAs não codificadores longos. *Recount2* é a maior coleção disponível de dados de RNA-seq humano processados e quantificados utilizando um pipeline unificado contendo mais de 4,4 trilhões de bases e mais de 70.000 amostras humanas derivadas do SRA e dos projetos TCGA e GTEx. Utilizando dados do GTEx derivados do FC-R2, nós validamos nossa abordagem ao reproduzir diversas descobertas importantes descritas recentemente pelo projeto FANTOM e do Pan-cancer atlas do TCGA. Em dois estudos de caso, nós também demonstramos a utilidade e capacidade do FC-R2 em recuperar novos RNAs não-codificadores longos potencialmente envolvidos em fenótipos de importância clínica. Concluindo, nós disponibilizamos o atlas FC-R2 como uma ferramenta publica para permitir que outros pesquisadores sejam capazes de identificar novos RNAs não-codificadores em fenótipos de interesse.

# Abstract

In recent years, in depth exploration of genomes structure and function has revealed a central role for non-coding RNAs (ncRNAs) in orchestrating key biological and cellular processes through the fine tuning of gene expression regulation. Most importantly, the understanding of the role for ncRNAs has also started to emerge in human disease pathogenesis. This further speaks to the importance of an in-depth characterization of ncRNA involvement in diseases, including cancer. In this work, we have built a comprehensive atlas of gene expression, named FC-R2, across the human transcriptome containing over 100,000 genes by leveraging two publicly available resources: the FANTOM CAGE Associated Transcriptome (FANTOM-CAT), and recount2. The FANTOM-CAT is a comprehensive meta-assembly of the human transcriptome encompassing coding and non-coding genes, including promoters, enhancers, and lncRNAs. recount2 is the largest, available collection of human RNA-seq data processed and quantified using a unified pipeline, containing over 4.4 trillion reads from over 70,000 human samples from the SRA, GTEx and TCGA projects. Using FC-R2 gene expression summaries across human tissue samples from the GTEx project, we validated our approach by reproducing key findings recently described by the FANTOM consortium and the TCGA Pan-Cancer atlas. We also demonstrated the power and usability of the FC-R2 by performing two case studies in prostate cancer highlighting potential "novel" lncRNAs players involved in the clinically relevant prostate cancer phenotype. Finally, we make the FC-R2 atlas available as a public tool to empower other researchers to study important biological and clinical phenotypes and identify new candidate ncRNAs for further investigation.

Table of Contents

## Table of figures

## Table of Tables

# Abbreviations List

| | |
|---|---|
| Androgen Receptor | **AR** |
| Chromatin Immunopreciptation | **ChIP** |
| Copy-Number-Variation | **CNV** |
| Differentially Methylated Regions | **DMR** |
| Divergent lncRNA | **d-lncRNA** |
| Enhancer lncRNAs | **e-lncRNA** |
| Epithelial-Mesenchymal Transition | **EMT** |
| Estrogen Receptor | **ER** |
| Expectation Maximization | **EM** |
| False Discovery Rate | **FDR** |
| Gene Set Enrichment Analysis | **GSEA** |
| Genomic Data Commons | **GDC** |
| Health Professionals Follow-up Study | **HPFS** |
| Imunohistochemistry | **IHC** |
| Intergenic lncRNA promoter | **i-lncRNA** |
| Metastatic Castration-Resistant Prostate Cancer | **mCRPC** |
| Mutation Annotation Files | **MAF** |
| Natural History | **NH** |
| Polypoly-Adenosine Diphosphate Ribose Polymerase | **PARP** |
| Programmed Death 1 | **PD-1** |
| Programmed Death Ligand 1 | **PD-L1** |
| Prostate Adenocarcinoma | **PRAD** |
| Prostate Cancer | **PCa** |
| Surrogate Variable Analysis | **SVA** |
| Transcription Start Site | **TSS** |

# Bibliographic revision

## The human genome

The human genome project (HGP) was a publicly funded project initiated in 1990 with the goal of determining the entire euchromatic regions of the human genome within 15 years. The sequencing of the human genome was an incredibly challenging task at the time since sequencing technologies were limited and expensive. The HGP opted for a hierarchical shotgun approach to sequence the human genome: in this approach the genome is first broken into large chunks and ligated into bacterial artificial chromosomes (BACs), were each BAC is later sequenced using the shotgun method and then assembled. The larger chunks are used to assembly the chromosomes and aid the assembly of smaller pieces[1]. In 1998, Celera Genomic owned by the researcher Craig Venter announced a competing effort of assembling the human genome with a different approach. The whole genome shotgun sequencing with paired-end sequencing approach used by Celera Genomics advanced at a quicker pace since it relied on data already released by the earlier project. In 2001 the first drafts of the human genome were published by both groups, following improved drafts in 2003 and 2005[1–3].

The projects revealed that only a small fraction of the human genome was 'functional' defined by the central dogma of molecular biology of proteins being the core of biological processes. The project estimated that around 20,000-25,00 protein coding genes were presented in the genome[2]. As technologies advances the actual number of protein coding genes is still in debate[4]. As of the latest version of GENCODE, a database run by the European Bioinformatic Institute, 19,940 are included as protein coding genes.

## Trancriptomics studies and RNA sequencing

A transcriptome represents the entire repertory of RNA molecules of one cell, tissue or organism and it is extremely dynamic. The identification of the set of transcripts in a sample is crucial to perform a comprehensive transcriptomic study. Transcriptomic studies first started with the idea of making DNA copies of mRNAs *in vitro* to amplifying a library of bacterial plasmids in 1979[5]. In 1983 Putney et al.[6] published the first study originated from this idea were a cDNA (herein referred as RNA to make it clearer) library of rabbit muscle was sequenced. The term Expressed Sequence Tags (ESTs) was later coined by Adams et al. in 1991[7]. Despite being the state of the art at the time, ESTs throughput were low, laborious, and limited by the sequencing technology of the time - the Sanger sequencing. Concomitantly with the advent of the expressed sequence tag (EST) technique, in 1995 the first study using DNA microarray (herein referred just as microarrays) to study gene expression was published by Schena et al.[8]. Microarrays provided a straightforward method to query expression of known genes with increased throughput at the time. The technique became increasingly popular, and the technology quickly evolved to become the state-of-the-art for gene expression studies during the 90's and the first decade of the 21th century, enabling relatively high-throughput expression screening.

**Table 1 - Advantages of the RNA-sequencing technique over existing transcriptomics methods. Original table from Wang et al. (2009)[9]**

| Technology | Tiling microarray | cDNA or EST sequencing | RNA-Seq |
|---|---|---|---|
| *Technology specifications* | | | |
| Principle | Hybridization | Sanger sequencing | High-throughput sequencing |
| Resolution | From several to 100 bp | Single base | Single base |
| Throughput | High | Low | High |
| Reliance on genomic sequence | Yes | No | In some cases |
| Background noise | High | Low | Low |
| *Application* | | | |
| Simultaneously map transcribed regions and gene expression | Yes | Limited for gene expression | Yes |
| Dynamic range to quantify gene expression level | Up to a few-hundredfold | Not practical | >8,000-fold |
| Ability to distinguish different isoforms | Limited | Yes | Yes |
| Ability to distinguish allelic expression | Limited | Yes | Yes |
| *Practical issues* | | | |
| Required amount of RNA | High | High | Low |
| Cost for mapping transcriptomes of large genomes | High | High | Relatively low |

With the advent of the Next Generation Sequencing (NGS) transcriptomics studies took off with the massive parallel sequencing of RNAs, in a technique latter coined as RNA-sequencing (RNA-seq). The RNA-seq consists of transcripts sequencing by NGS, which presented an enormous increase in throughput compared to the original ESTs technique. This increased throughput allowed for the detection of lowly expressed transcripts and the screening of the entire RNA repertoire. Even though modern microarrays already provided high-throughput screening of RNAs, the RNA-seq presented significantly advantages over it such as independence of a reference sequence and base pair resolution, allowing for the discovery of new genes and isoforms which turned the RNA-seq into the state-of-the-art approach for transcriptomics studies until the date[9,10] (Table 1).

Briefly, the RNA-seq consists of three steps: 1) RNA extraction; 2) library preparation and sequencing and 3) expression quantification. The first step is crucial as the quality and the nature of the data yielded depend on choices made in this step. Since rRNA is the most abundant class of RNA in a cell, an enrichment for other types of RNA is performed to avoid rRNAs dominating the sequencer capacity. This is usually done by either poly-A capture which effectively enriches the

sample for mRNAs or rRNA depletion by targeted degradation. The latter has the advantage of maintaining RNAs without poly-A tails which encompass several classes of ncRNAs. The enrichment can also be done with a specific aim in mind, such as the study of small RNAs were the sample is separated by size before the next step[11].

Following the RNA extract, libraries must be built before being sequenced. There are many variations of RNA-seq that relies on specific protocols in this step. With that aside, some choices always need to be made such as choosing between a single- or paired-ended and stranded or unstranded library. Paired-end libraries offers significant advantages over single-ended ones. Since both ends of the RNA fragment are sequenced in a paired manner, it improves the odds of correctly mapping the transcript to the reference genome as both pairs must map together in a given region and it can provide insightful information about the transcript architecture. Similarly, stranded libraries provide information about the strand of origin of the fragment sequenced. This is especially useful when handling complex genomes with gene dense regions where overlaps of genes across strands is often observed. Once the libraries are built, they are sequenced to a desired depth. The sequencing depth amounts to the number of fragments sequenced, therefore it is directly correlated with the sensitivity to detect low expressed transcripts. The final result of a sequencing process is generally a FASTQ file containing the fragments (also known as reads) sequences[11].

Finally, the last step is mapping the reads to a reference genome. Mapping reads to a reference genome, also known as alignment step, is performed by an alignment software of which dozens are available (e.g. HISAT2, STAR, RailRNA)[12–14]. Although recent efforts have been made to improve the speed and resources used in this step, it is still a relatively resource intensive step. Once reads are accurately mapped, they can be counted on basis of the overlapping genomic feature. Some software's allow this step to be bypassed by directly quantifying gene counts without

the alignment step[15,16], known as alignment-free quantification. The final object for an expression study is an expression matrix, containing expression estimates for each feature, that is used in downstream analysis[11].

**Non-coding RNAs**

For a long time, non-coding regions in the genome were seen as "junk DNA" which played no role in the biology of organisms[17]. As genomics studies evolved, it became increasingly clear that coding genes could not explain the complexity of more complex organisms since the number of coding genes in the lower branches of evolutive tree were not much different from the organisms at the top of the tree. Researchers started to notice that the 'junk' DNA could not be junk after all, and that the amount of non-coding regions tracks together with the evolutionary tree[18–20] (Figure 1).

For a long time since Francis Crick coined the central dogma of molecular biology in 1958 were, he stated that the genetic information flowed from DNA to RNA to proteins, RNA molecules were believed to be mostly intermediates components between a gene and its protein product. This notion heavily biased the discovery of new genes towards protein coding genes. It was only in the last two decades that the non-coding RNAs (ncRNAs) came to light revealing a best a complex variety of regulatory RNA molecules which had been neglected until then[18,21]. In its third iteration, the FANTOM consortium started to develop a new technique, named Cap Analysis of Gene Expression (CAGE)[22] which was designed to accurately map promoter regions and their usage. Remarkably, CAGE analysis not only identified promoters and quantified their activity and the expression of known RNAs, but also revealed that there were many more RNAs in the mammalian

transcriptome than previously thought. They showed for the first time, that over 63% of the genome produced some kind of transcript, many being non-coding[23].

ncRNAs are functional RNA molecules that doesn't rely on the translation process to exert its function. In the last decade, several systematic screening revealed a surprising number of novel ncRNAs capable of acting in a variety of cellular functions such as post-transcriptional regulation of gene expression and guiding RNA and DNA modifications[19,21,24–26]. The term ncRNA is rather vague since they are present in a huge variety of sizes from 21nt long referred as microRNAs (miRNAs) to huge non-coding genes such as the 32,103nt long X-inactive specific transcript (*XIST*) gene[27]. To differentiate short from large ncRNAs, a ncRNA larger than 200nt is often referred as long non-coding RNAs (lncRNAs). While shorter ncRNAs usually presents a strict function and way of acting, lncRNAs are far more diverse in their acting mechanisms and function. These characteristics has put them into spotlight in the recent year, with novel genes, role and mechanisms being uncovered at a quick pace[28–31].

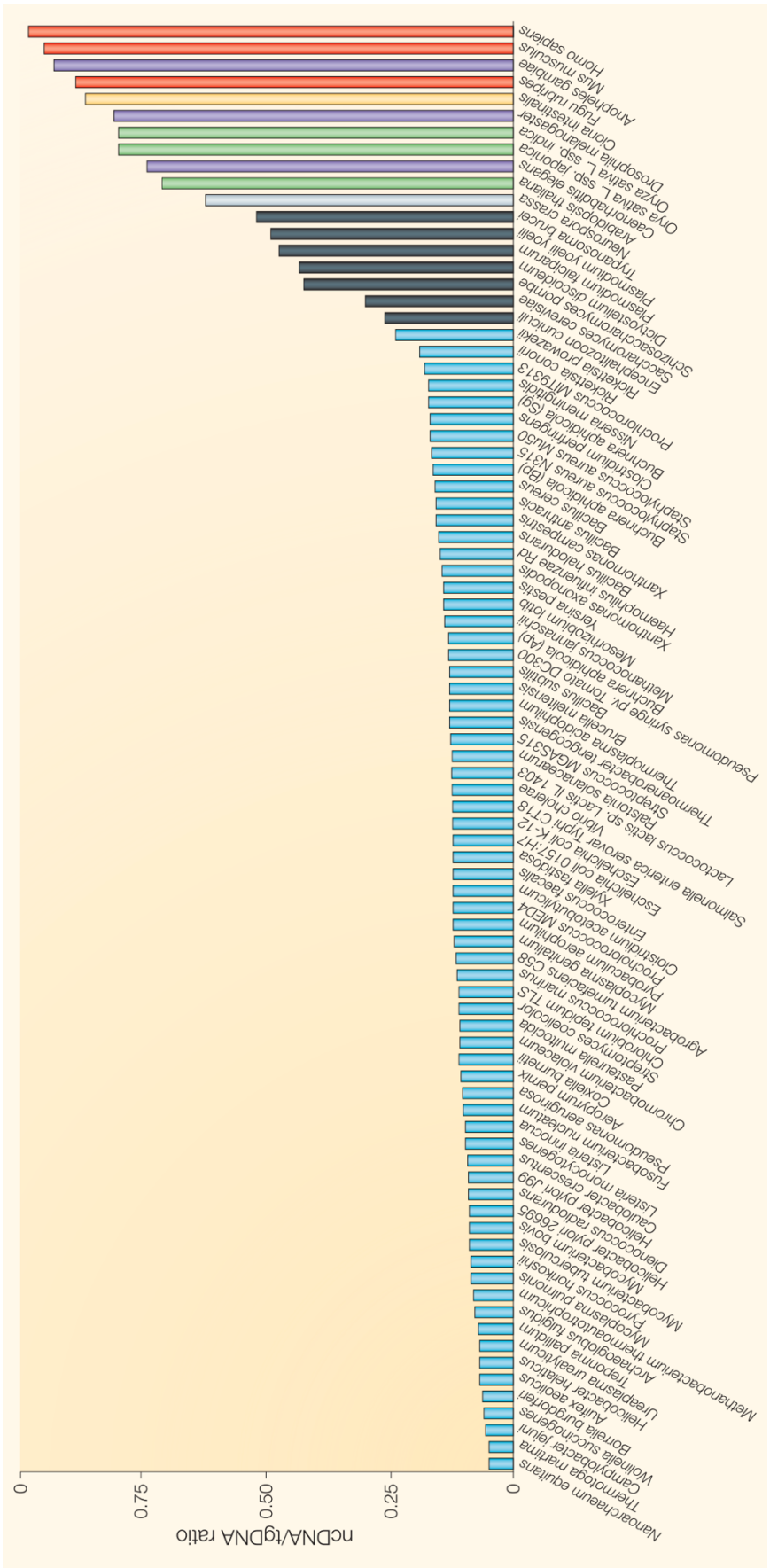



**Figure 1 - The ratio of non-coding to protein-coding DNA rises as a function of developmental complexity.** Figure shows the increase of non-coding regions that tracks together with developmental complexity. Prokaryotes the least amount of non-coding DNA in their genome with up to 25% whereas in highly complex organisms such as humans the non-coding DNA content can rise up to 98.5%. The distinct colours represent prokaryotes (bacteria and archaea) (blue), simple eukaryotes (black), Neurospora crassa (grey), plants (green), non-chordate invertebrates (nematodes, insects) (purple), Ciona intestinalis (urochordate) (yellow) and vertebrates (red). ncDNA, non-coding DNA; tgDNA, total genomic DNA. Original figure from Mattick (2001) [19].

So far, they have been implicated in a variety of roles (e.g. regulation of allelic expression, control of pluripotency, lineage specification, epigenetic control)[28] and diseases (e.g. cancer, obesity, diabetes)[31,32]. Although many lncRNAs have been confidentially associated with several phenotypes, elucidating all the mechanisms by which they act are still a work in progress. Nevertheless, some mechanisms have already been studied among which are transcription factor localization, increase in mRNA stability, disruption of translation, transcription coactivation, and others[33]. A comprehensive review can be found in Kung et al.[33].

**Annotations**

Assembling the genome was only the first step in understanding the complexity of our genome. The next challenge would be making sense of all uncovered sequences revealed by the HGP. This process of uncovering genes and their function is referred as annotation, which is the process of annotating regions of a genome with discovered genes. The process of annotating a complex genome is very laborious and needs constant work as novel genes or new roles for known genes are uncovered every year.

To tackle this issue, several consortiums were launched to annotate the human genome. Two of the biggest consortiums devoted to annotating the human genome are the GENCODE and the FANTOM (Functional Annotation of the Mammalian Genome). Both consortiums have the same goal, but the approaches used in each are different.

The GENCODE project was launched in 2003 to carry out a project whose objective was to identify all functional elements in the human genome sequence. The GENCODE leverage computational and experimental methods to identify new genes and their isoforms with manual curation of regions requiring expert investigation[34]. In its latest version to date GENCODE 29

contains annotations for 58,721 genes with 19,940 and 38,781 coding and non-coding (including pseudogenes) genes, respectively.

The FANTOM consortium was established in 1995 to assign functional annotations to the mouse genome. Overtime the project expanded to encompass the human genome as well. The objective of the consortium is to obtain a systemic overview of the transcriptional regulatory network of the human organism[23]. Over its iterations the FANTOM consortium aimed to increase our understanding of the regulatory landscape of the human transcriptome using CAGE-seq. CAGE-seq technique measures expression from the 5' end of capped molecules and provides very accurate mapping of the transcription start site (TSS) and enables the identification of promoters and other lncRNAs as it is very sensitive due to sequencing only the 5' end of the transcript[35]. In 2017, the FANTOM consortium released the FANTOM Cage Associated Transcriptome (FANTOM-CAT) which integrated accurate mappings of TSS by CAGE-seq, RNA-seq and epigenomic data from the Readmap DNase I hypersensitive sites (DHS), which revealed over 19,000 (in its most stringent set) novel lncRNAs not included in GENCODE annotations. In its most permissive set FANTOM-CAT harbors 124,245 genes[29] (Figure 2).

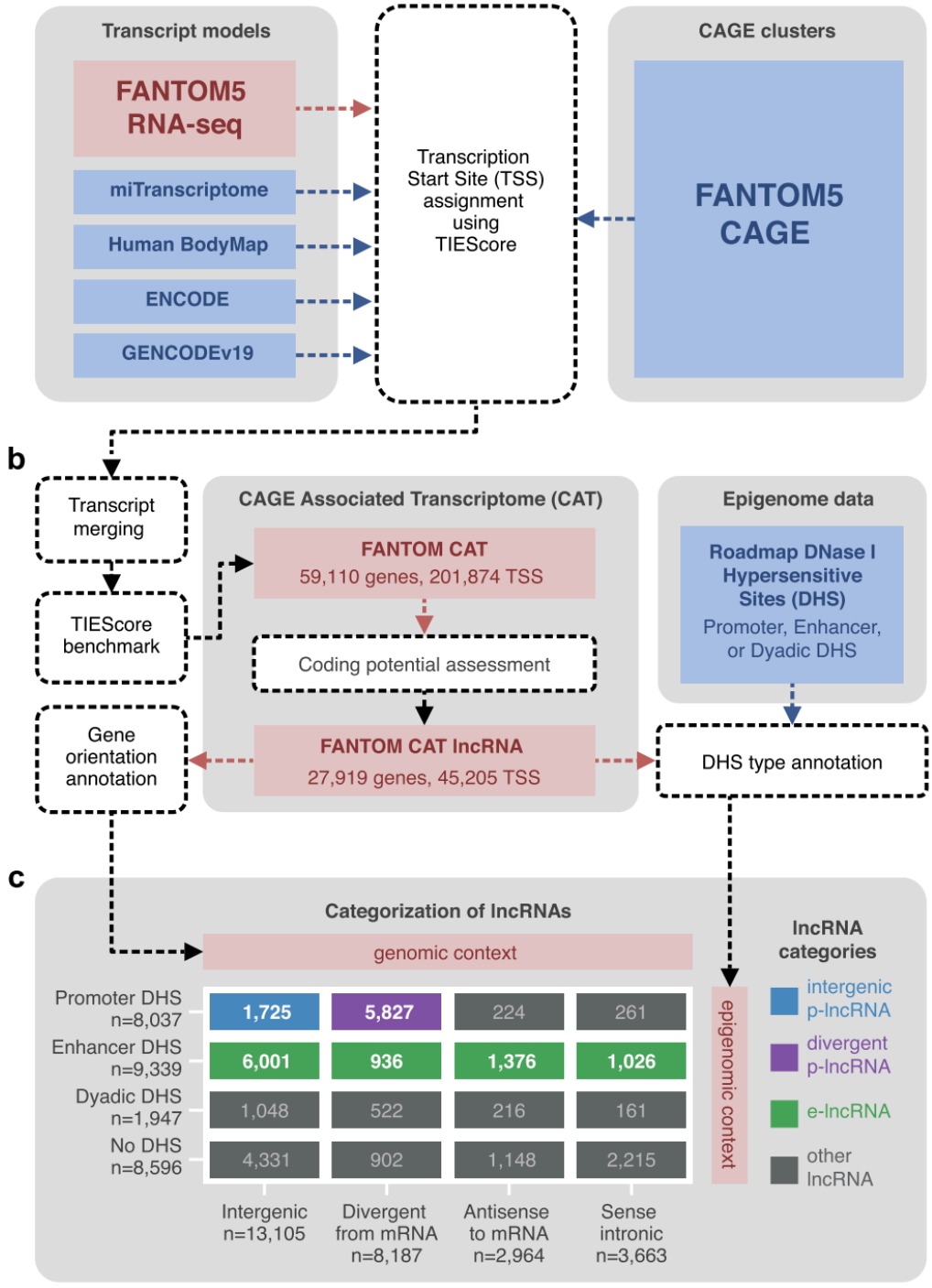



**Figure 2 - Overview of the FANTOM-CAT meta-assembly.**The FANTOM-CAT meta-assembly encompasses annotation from a variety of sources that were refined through CAGE-seq data derived from the FANTOM5 project. Epigenomics data from DNase I hypersensitivity arrays were also used to define the chromatin state of the genes and assign functional lncRNAs categories. Original figure from Hon et al. (2017) [29].

## Expression databases

The modernization of NGS platforms over the decades had led to decreased costs and a huge increase of data throughput. This enabled larger studies which, by consequence, generated substantial amounts of data that quickly put biological data under the big data spectrum.

To couple with the huge amount of data generated every day, databases play a crucial role in making these data safely stored and accessible to other scientists ensuring that science is still reproducible at this scale. Although thousands of databases are available through efforts of research groups from all around the world, most of them are devoted to specific subjects. In a more general scope, the largest databases available are the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA)[36] and Gene Expression Omnibus (GEO)[37]. The first, SRA, stores raw sequencing data and alignment information from high-throughput sequencing platforms of all sources (genomic or transcriptomic) and make it available to the research community to enhance reproducibility and promote new discoveries by comparing data sets. The second, GEO, harbors only expression (transcriptomic) data and encompass both microarrays and NGS data. GEO however, unlike SRA, allows researchers to not only obtain raw data, but also processed data deposited by its submitter. Processing of biological data is usually very resource intensive and giving the opportunity to skip this process ensure that researchers without access to high-end machines can still use the data, or at least avoid double efforts.

However, one huge drawback of providing processed data is that expression data can be processed in several ways that may better adjust to the particular aims of the original study and even in cases where the aim is the same there are still a plethora of pipelines that can be used to achieve the same goal and not all of them yield the same results, as parameters are often adjusted on a need basis. Obtaining and comparing processed data from distinct groups can be prove

difficult, not only due batch effects but also by the different methods used to process the data. In an attempt to tackle this issue, Collado-Torres et al. made an effort to make all Illumina-based human RNA-seq data from the SRA processed through a standardized pipeline which makes comparing data from different studies less prone to the variability introduced during the processing step. The database *recount2* harbor processed expression data from over 70,000 RNA-seq samples from 2,041 studies making it the largest collection of processed expression data to date[38].

**Precision medicine**

As technologies advances, the cost and time to generate massive amounts of biological data starts to decline. Over the last decades we have been observing an increasingly amount of biomedical data made available in databases. These data have helped researchers to gain valuable insights in the molecular mechanisms that drives several diseases, which could be leveraged to improve and guide current methods of treatments[39,40]. These insights have led researchers to uncover a layer of heterogeneity within diseases that until then were treated in the same manner.

Translating insights from genomics to medicine, gathering information at an individual level, enabled more precise calls on treatment of several diseases[41]. The use of genomics-driven medical decision is often referred as *precision medicine* were the treatment is not guided by the disease *per se*, but by unique features of the molecular profile nested within a disease at an individual or at least subgroup level[41] (Table 2).

**Table 2 - Examples of precision medicine** [41].

| Condition | Gene | Action |
|---|---|---|
| ***Mendelian disease*** | | |
| Cystic fibrosis | *CFTR* | Specific therapies such as ivacaftor and a combination of lumacaftor and ivacaftor |
| Long QT syndrome | *KCNQ1*, *KCNH2* and *SCN5A* | Specific therapy for patients with *SCN5A* mutations |
| Duchenne muscular dystrophy | *DMD* | Ongoing phase III clinical trials of exon-skipping therapies |
| Malignant hyperthermia susceptibility | *RYR1* | Avoid volatile anaesthetic agents; avoid extremes of heat |
| Familial hypercholesterolaemia (FH) | *PCSK9*, *APOB* and *LDLR* | • Heterozygous FH (HeFH): eligible for PCSK9 inhibitor drugs<br>• Homozygous FH (HoFH): eligible for PCSK9 inhibitor drugs in addition to lomitapide and mipomersen |
| Dopa-responsive dystonia | *SPR* | Therapy with dopamine precursor L-dopa and the serotonin precursor 5-hydroxytryptophan |
| Thoracic aortic aneurysm | *SMAD3*, *ACTA2*, *TGFBR1*, *TGFBR2* and *FBN1* | Customization of surgical thresholds based on patient genotype |
| Left ventricular hypertrophy | *MYH7*, *MYBPC3*, *GLA* and *TTR* | Sarcomeric cardiomyopathy, Fabry disease and transthyretin cardiac amyloid disease have specific therapies |
| ***Precision oncology*** | | |
| Lung adenocarcinoma | *EGFR* and *ALK* | Targeted kinase inhibitors, such as gefitinib and crizotinib |
| Breast cancer | *HER2* | HER2 (also known as ERBB2)-targeted treatment, such as trastuzumab and pertuzumab |
| Gastrointestinal stromal tumour | *KIT* | Targeted KIT kinase activity inhibitors, such as imatinib |
| Melanoma | *BRAF* | BRAF inhibitors, such as vemurafenib and dabrafenib |
| ***Pharmacogenomics*** | | |
| Warfarin sensitivity | *CYP2C9* and *VKORC1* | Adjust dosage of warfarin or consider alternative anticoagulant |
| Clopidogrel sensitivity, post-stent procedure | *CYP2C19* | Consider alternative antiplatelet therapy (for example, prasugrel or ticagrelor) |
| Thiopurine sensitivity | *TPMT* | Reduce thiopurine dosage or consider alternative agent |
| Codeine sensitivity | *CYP2D6* | Avoid use of codeine; consider alternatives such as morphine and non-opioid analgesics |
| Simvastatin sensitivity | *SLCO1B1* | Reduce dose of simvastatin or consider an alternative statin; consider routine creatine kinase surveillance |

Original table from Asheley (2016)

In oncology for example, the classification of solid tumors was traditionally focused in their tissue of origin. However, since the success of a more directed and personalized treatment based on the molecular subtype, oncology has moved towards a molecular classification of tumors. Taking lung cancer as an example, non-small cell lung adenocarcinoma expressing epidermal growth factor receptor (EGFR) is currently treated with a different chemotherapy from that of non-

EGFR-driven adenocarcinomas[42]. In the precision medicine era since patients with different biomarkers present different risks of developing a disease and, therefore, have different prognoses and response to treatments, these biomarkers are expected to be treated as a standard phenotypic feature (e.g. symptoms, histology and medical history), leading to a revised definition of a disease to include a new subtype[40] (Figure 3).

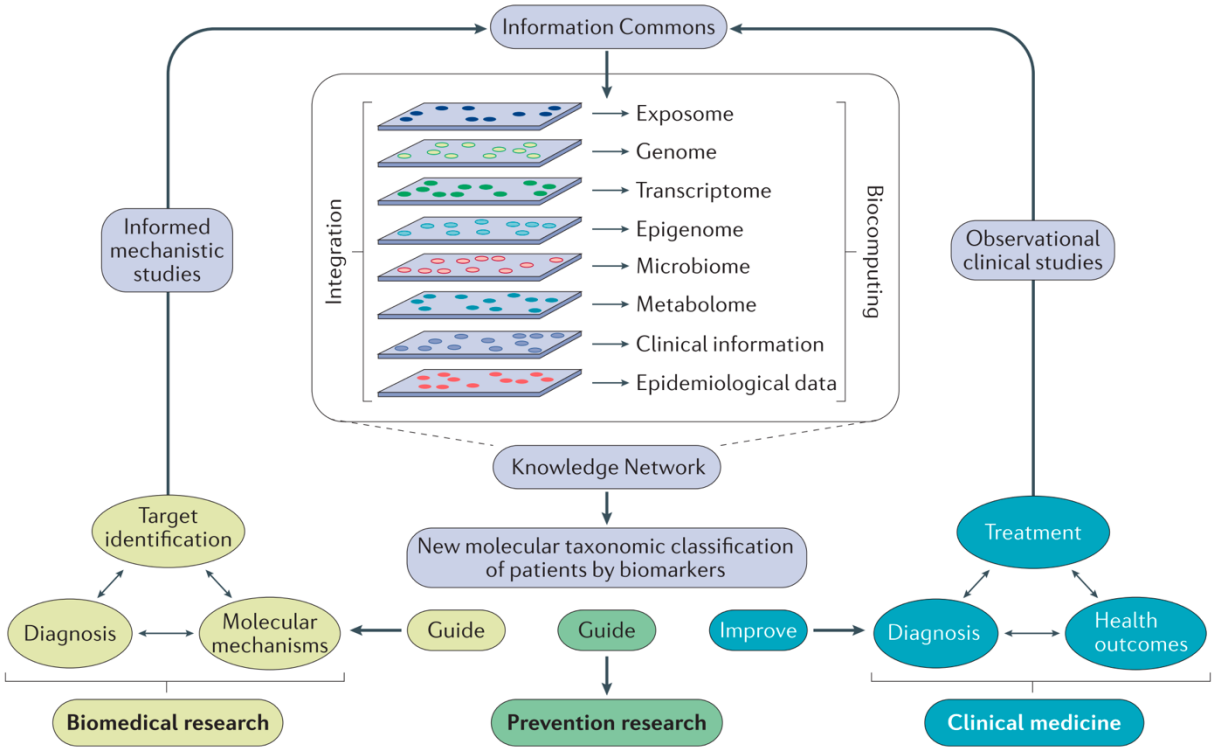



**Figure 3 - A precision medicine research strategy.**Omics studies can be integrated in a knowledge network that can be used to define new molecular classifications for diseases. This improved classification and stratification can be used to refine mechanistic processes of a disease and improve the way patients are diagnosticated and treated. Original figure from Vargas & Harris (2016)[40].

The precision by which new subtypes can be assigned to a disease, rests heavily on the success of the research framework, as outlined in the American Institute of Medicine's report[43]. Naturally, biomarkers are the foundation of improving diagnostic precision. Biomarkers can be

associated with a specific disease, or they can present a different mechanism of action related to a disease. They can be used to infer risk, diagnosis, prognosis and therapeutic response as a single feature or as group, which is often the case of complex diseases such as cancer were a single biomarkers is not enough to capture the heterogeneity of tumors[40].

When biomarkers are paired with a companion therapeutic agent, they are termed companion diagnostics[44]. Naturally, as any other subject implicated in human health, the development and use of companion diagnostics have been recently regulated. In this regard, the US Food and Drug Administration is the leader in the regulation of precision medicine because it was the first to set regulatory guidelines for this field[45]. The European Medicines Agency and Japan's Pharmaceutical and Medical Device Agency have also their own guide to companion diagnostics and there are ongoing efforts to harmonize and improve the regulatory pathways within and between countries[40].

The ultimate goal of precision medicine and its regulation is to move molecular findings through validation and then to patients in need of an improved diagnostic precision. Although the Holy Grail of precision medicine would be a large national cohort study, such data is still not available to date. Currently, the precision medicine research approach makes use of existing cohorts with a relatively small number of individuals (usually hundreds for a single disease) such as the The Cancer Genome Atlas (TCGA). Nevertheless, these cohorts harbor large amounts of data from a variety of sources (e.g. DNA, RNA, protein, histology) that can be analyzed in order to obtain predictors of disease risk, prognosis and treatment response that can be further validated and used in the clinics[40].

**Differential gene expression analysis**

RNA-seq and expression microarrays provide scientists with a comprehensive overview of the transcriptional landscape of sample. The experimental design for these experiments follows the same set of 'rules' from traditional benchwork experiments where performing replicates is crucial for giving statistical confidence that the results observed are not purely by chance. While obtaining the transcriptional landscape of a given phenotype can lead to insights by itself, pairing it with another contrasting phenotype of interest (e.g. control, a subtype of the phenotype) enable scientists to background noise by focusing only in the transcriptional set that are specific to the phenotypes of interest. This is usually done with what is called differential gene expression analysis (DGE). DGE analysis is a powerful tool that allow researchers to screen dozens of thousands genomic features of a sample and compare it with another one of interest, ultimately yielding a list of features that are differentially activated or repressed in the conditions analyzed. This list often provides insightful information of key genes or pathways that leads to the phenotype in question. Performing a DGE analysis however is not straightforward, dozens of tools have been developed to perform this task and each one of them takes a different approach in doing so. Methods can vary according platforms (e.g. microarrays, RNA-seq, etc.) in this section only RNA-seq approaches will be discussed since it is currently the state-of-the-art for classic transcriptomics studies (Figure 4).

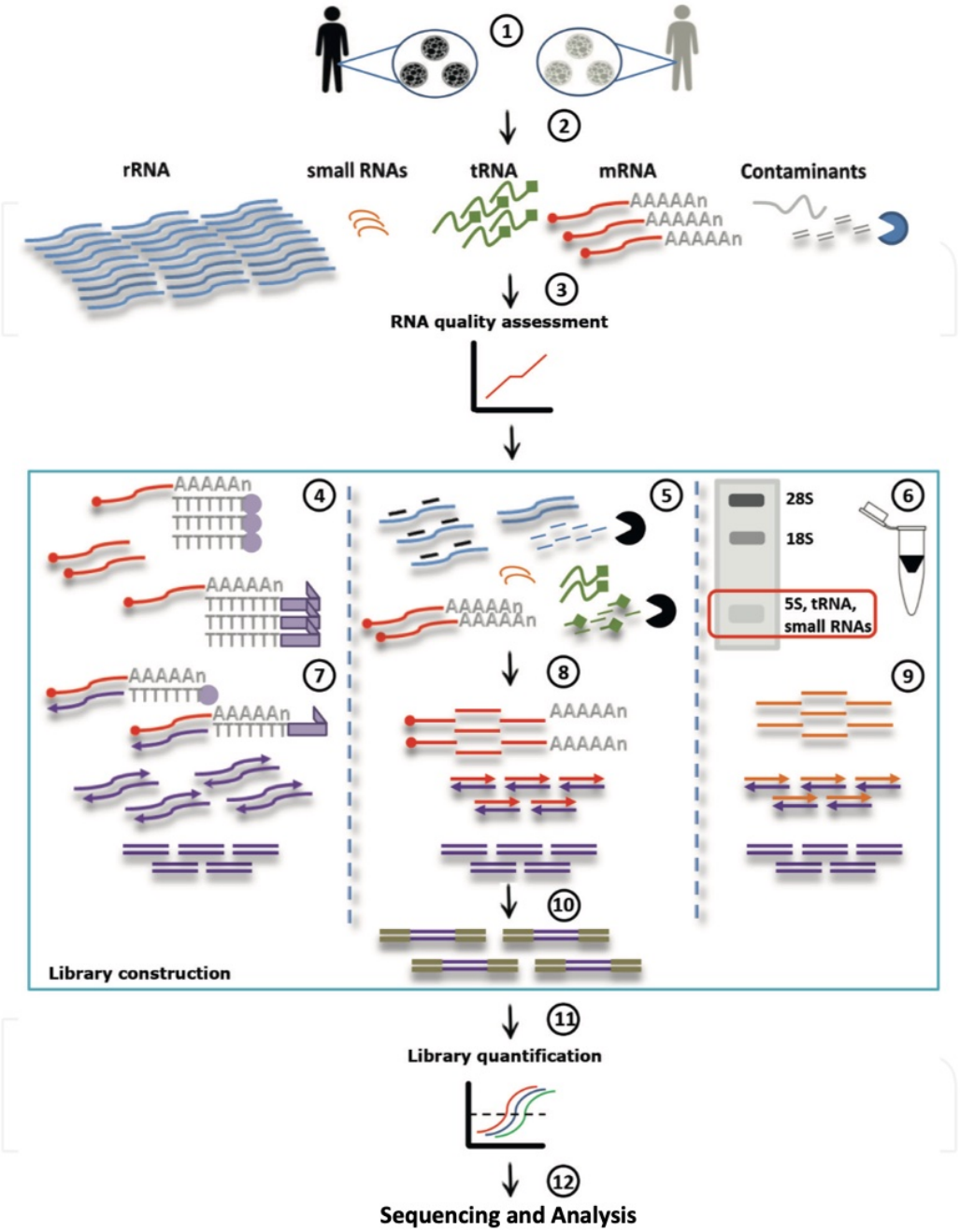



**Figure 4 - A typical RNA-seq workflow.**(1) Experimental design definition of qualitative and quantitative goals. Differential gene expression among different conditions is exemplified; (2) Sample selection, RNA extraction and elimination of contaminants such as genomic DNA; (3) Assessment of RNA integrity; (4-6) RNA enrichment. (4) mRNA enrichment using magnetic or cellulose beads coated with oligo(dT) molecules or oligo(dT) priming; (5) mRNA enrichment through rRNA depletion with conserved probes or Selective Depletion of abundant RNA (SDRNA); (6) Small RNA size-selection through electrophoresis or based on solid phase extraction; (7-9) cDNA single/double strand synthesis. (7) cDNA synthesis followed by fragmentation; (8) mRNA fragmentation followed by cDNA synthesis; (9) cDNA synthesis for small RNA without fragmentation; (10) Adapters ligation; (11) Library quantification and (12) Library sequencing and data analysis. Original figure from Pereira et al. (2017)[11]

264

The RNA-seq approach ultimately yields counts of sequenced molecules for each gene which is a discrete type of data. Since sequencing can be performed at different depths for each sample, the first step in analyzing RNA-seq data is to normalize it. Several approaches to normalization have been proposed[46–48], each with their own advantages and disadvantages with the choice being a somewhat an arbitrary since most of them perform well in most conditions. There are a few distributions that can model this type of data. Most tools designed to model counts data relies on what is called a negative binomial (NB) distribution. While a Poisson distribution is often used to model counts type of data, NB distributions has several properties that better fit the modelling of gene expression than the Poisson distribution. The Poisson distribution has a limiting factor in which it assumes that the data modelled had a single mean/variance which will tend to underestimate the variance since usually expression data is derived from different subjects which will introduce their own variability for each gene tested. Because of this property, Poisson based model tends to rely on many samples to yield precise mean/variance estimates, which is not currently feasible due to cost/time factors. In a scenario where $n \to \infty$, the variability introduced by each individual would shrink, leading to similar results of a NB model. The NB distribution allows for different parameters that can better estimate the real variance of a gene.

One drawback of NB distribution methods developed for RNA-seq counts (or any probabilistic distribution for what matters) is that they rely on approximations of various kinds. They treat the estimated dispersions as if they are known parameters, without allowing for the uncertainty of estimation. While some methods account for this uncertainty in later steps, they still rely on other kind of approximation[49].

One popular approach to deal with this issues is implemented in the *limma* model were it seeks to robustly estimate the mean-variance relationship at the observation level of the counts in

a non-parametrically way from the data and incorporating it as weights to effectively eliminate it allowing it to be analyzed as log-normal distributions[49].

Finally, once counts are modeled, they can be tested for differences across the phenotypes in several ways (e.g. quasi-likelihood, generalized linear models (GLMs), linear models) yielding the final list of differentially expressed (DE) genes to analyze.

# Objectives

- Leverage publicly available RNA-Seq data to create a comprehensive gene expression atlas comprising thousands of novel lncRNAs recently uncovered by the FNATOM consortium, allowing researchers to easily study these lncRNAs.
- Validate the resource by reproducing recent key findings about lncRNAs (i.e. expression patterns, prognostic potential, etc.)

# Methods

## Data and preprocessing

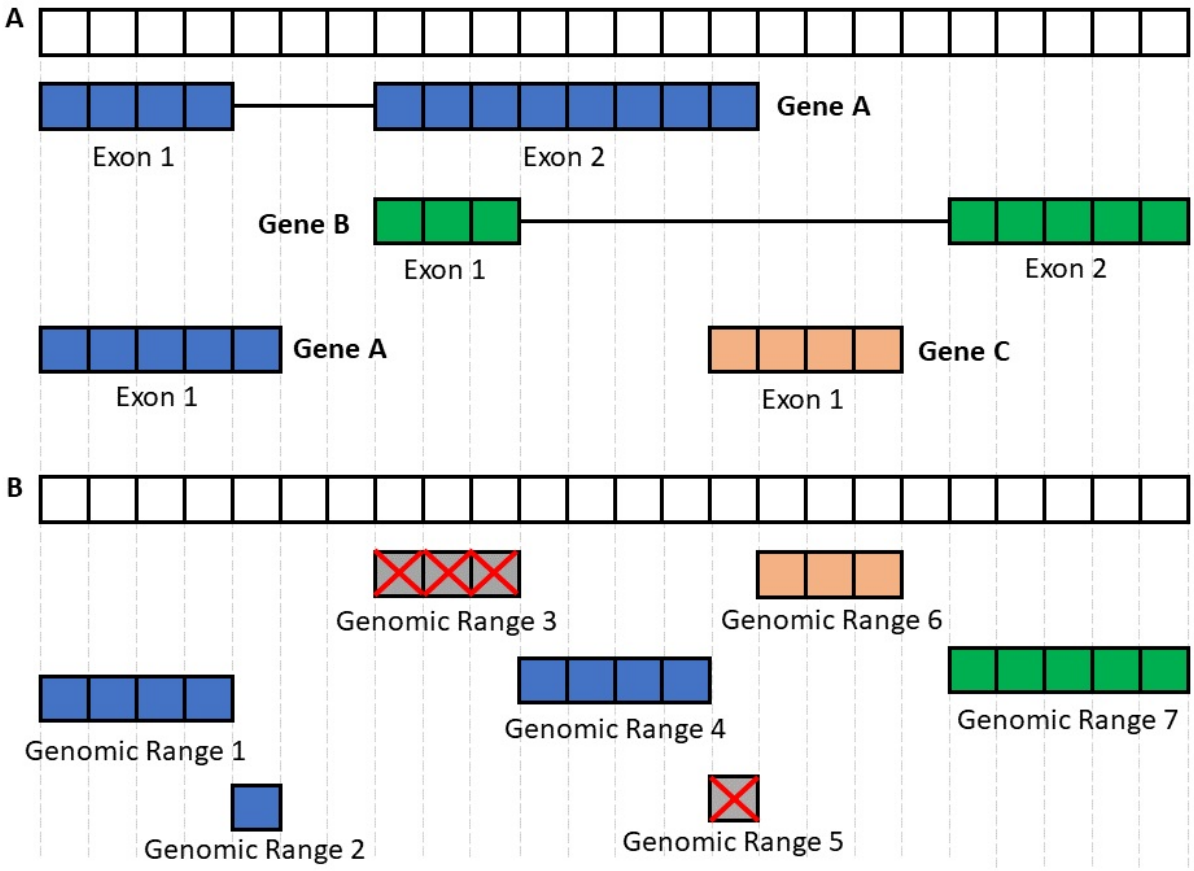



**Figure 5 - Representation of the disjoining and exon disambiguation processes.**(A) Representation of a genome segment and its annotation containing 3 genes with gene A having two isoforms, and genes B and C with one isoform each. Each box can be interpreted as one nucleotide with boxes colored blue or orange to represent exons on opposite strands. (B) Representation of disjoined exon ranges from example A. Each feature is reduced to a set of non-overlapping genomic ranges, then genomic ranges mapping back to two or more genes are removed (crossed boxes). After removal of ambiguous ranges, the remaining ranges are summarized at gene level. Grey boxes represent segments with ambiguous strand.

To create the FC-R2 expression atlas we obtained an updated version of FANTOM-CAT permissive catalog from the FANTOM consortium containing data from the ongoing FANTOM6

project and therefore still not publicly available. This catalog was initially comprised of 124,245 genes defined by Cap Analysis of Gene Expression (CAGE) peaks published by Hon et al[29].

Given the unstranded nature of recount2 framework we opted to remove overlapping gene regions from the quantification process to avoid multiple measures from these ambiguous regions. In order to remove ambiguity we imported into an R session the coordinates for each gene/exon from a BED file and processed with the GenomicRanges package[50] by disjoining the exon coordinates. To avoid losing strand information from the annotations we processed it in a two-step approach by first disjoining overlapping segments on the same strand and latter across strands (Figure 5). Genomic ranges (disjoint exons segments) that were assigned to more than one gene were removed from the expression atlas.

We submitted these ranges to the Recount2 framework[38,51] obtaining expression information for coding mRNAs, enhancers and promoters (divergent and intergenic) for 9,662 samples from the Genotype-Tissue Expression (GTEx) project, 11,350 samples from TCGA consortium and over 50,000 human samples from SRA.

**Correlation with other studies**

In order to verify if our pre-processing step had any major impact on the expression quantification, we compared our counts estimates to the published GTEx counts from recount2 which was based on GENCODE annotations. The version 2 of the gene counts for the GTEx samples was downloaded from the recount website (https://jhubiostatistics.shinyapps.io/recount/). Next, we compared distribution of tissue specific genes across tissues and computed the Pearson correlation for each protein coding gene in common across the original recount2 gene counts estimates and our version.

366

367 **Expression specificity of tissue facets**

368

369 We used our FC-R2 GTEx data to evaluate the expression specificity of lncRNAs categories. First,
370 we normalized the expression levels accounting for the library depth and gene length, with the
371 gene length being the sum of non-overlapping disjoint ranges belonging to a gene. Next, we
372 grouped GTEx samples by tissue type, totalizing 54 tissues groups and evaluated the expression
373 level and specificity of each gene. The expression level for each gene was represented by the
374 maximum transcripts per million (TPM) of all samples within a tissue type. The expression
375 specificity was calculated as the empirical entropy of the median expression values of each tissue
376 type. We then computed the 99.99 percent confidence intervals for the expression of each lncRNA
377 category by tissue type based on TPM values. During this analysis only genes with a TPM greater
378 than 0.01 were considered expressed and set to 0 otherwise.

379

380 **Global enhancer activation**

381

382 In order to compare enhancer activation in tumor vs normal tissues we selected all tumor types
383 available from TCGA with at least 10 tumor/normal paired samples. For each of the samples, we
384 measured its global enhancer expression by summing the TPM values of all enhancers in our atlas
385 stratified by tissue type. We then proceed to compute the global enhancer activation for matched
386 tumor and normal samples with the activation score being computed as the ratio of mean global
387 enhancer expression between tumor and normal samples minus 1 given a cancer type.

388

389 $$\frac{A_{tumor}}{A_{normal}} - 1;\ A_{ij} \leftarrow \sum \mu_{ij}$$

390

Where *i* is the sample type (tumor or normal) and *j* is a given cancer type. Significance for differential global enhancer activation was computed using a paired t-test.

**Prognostic enhancers analysis**

A univariate Cox proportional regression was performed using each of 17,404 enhancer lncRNAs (e-lncRNA) as predictors on each of the 13 TCGA cancer types with available survival follow-up. Enhancers with FDR equal or less than 0.05, after multiple hypothesis correction with Benjamini-Hochberg[52] method within cancer type were selected as significant prognostic factors.

In order to compare our results, we obtained supplementary data from Chen et al.[53] containing all enhancers position used in the study and prognostic potential were also obtained from the original publication. A liftover from hg19 enhancer positions evaluated by Chen et al.to hg38 genome assembly was performed to match FC-R2 coordinates.

**Identification of differentially expressed genes**

Differential gene expression analysis was performed in FC-R2 TCGA gene expression summaries across 13 cancer types with at least 10 normal samples. The original dataset for each cancer type was split by RNA class (coding mRNA, intergenic-lncRNA promoter (i-lncRNA), divergent-lncRNA (d-lncRNA) promoter and e-lncRNA) and treated independently to avoid compromising the model with artifactually low variance from lowly expressed lncRNAs categories. We applied a generalized linear model approach coupled with empirical Bayes standard errors[54] to estimate differential expression between tumor and normal samples within each tumor type and RNA class. The model was also adjusted for the three most variable coefficients for data heterogeneity as estimated by surrogate variable analysis (SVA)[55]. P-values were corrected for multiple testing

using the Benjamini-Hochberg method[52] and genes with an adjusted p-value equal or less than 0.01 were considered as differentially expressed.

# Results

## Building the FC-R2 resource

We built the FC-R2 expression atlas by extracting expression levels from *recount2* coverage tracks that were contained within unambiguous exon coordinates for the permissive set of *FC-R2* transcripts, as shown in Figure 5 (see Methods). Due to lack of genomic strand specificity/information in *recount2*, we removed ambiguous exonic segments from overlapping genes to more precisely measure expression levels of the individual transcripts (which we assessed in detail below). After removing these ambiguous genomic ranges, we ended up with 1,066,515 exonic segments mapping back to 109,873 genes in *FC-R2*. The resulting resource in total includes expression information for 109,873 genes including 22,110 previously annotated coding and presumably 87,693 non-coding genes, such as enhancers, promoters, and others lncRNAs. Specifically, it encompasses expression data for 109,873 genes across 2,041 studies with over 70,000 RNA-seq samples.

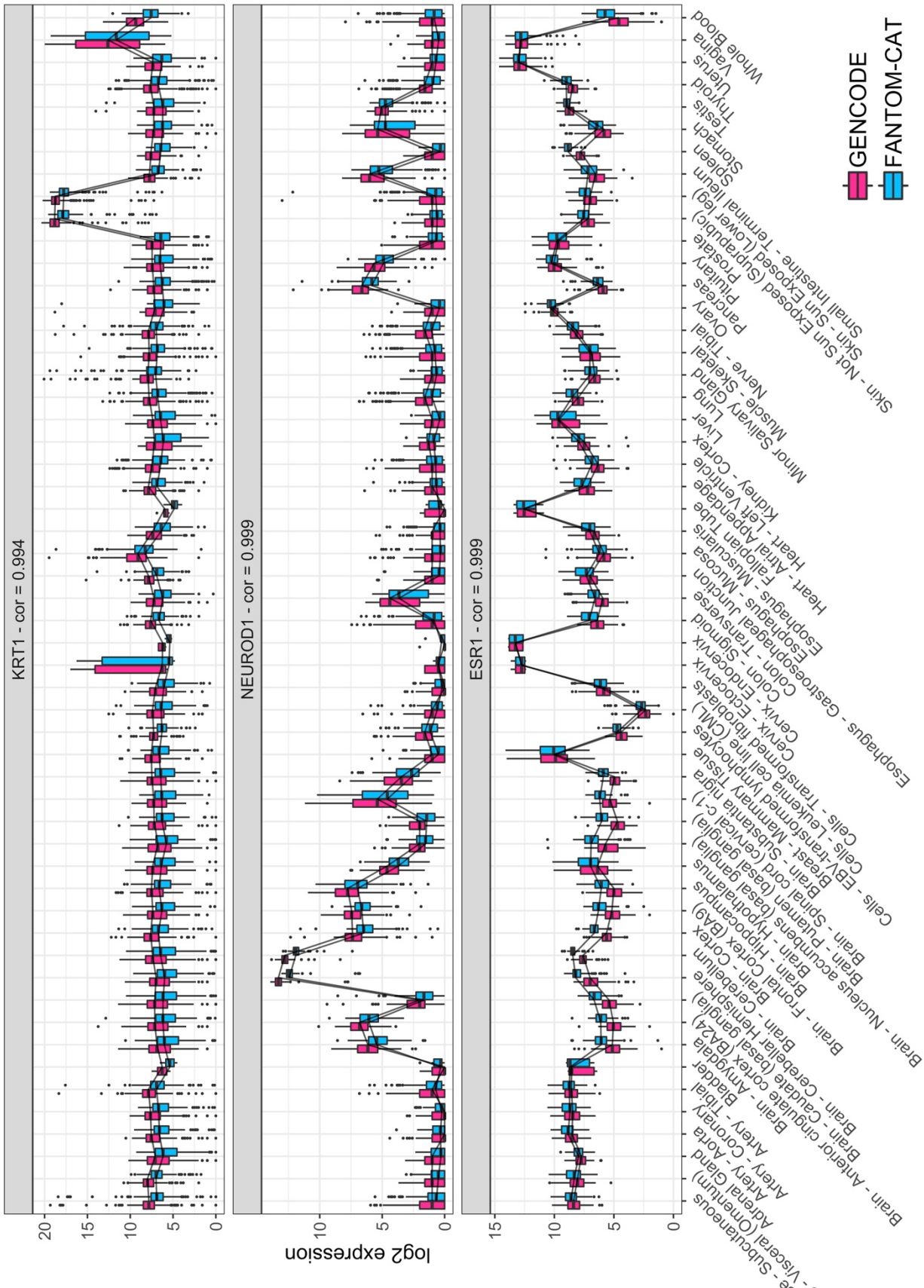



**Figure 6 – Tissue specific expression in GTEx.** Log2 expression for three tissue specific genes (*KRT1*, *NEUROD1*, and *ESR1*) in GTEx data stratify by tissue type using FC-R2 and *recount2*/GENCODE based quantification. Expression profiles are highly correlated and expressed consistently in the expected tissue types (e.g., *KRT1* is most expressed in skin, *NEUROD1* in brain, and *ESR1* in estrogen sensitive tissue types like uterus, Fallopian tubes, and breast

## Validating the FC-R2 resource

We first performed quality assessment of the expression data in FC-R2 compared to previous efforts. We computed the correlation of the overall expression levels between the FC-R2 atlas and gene counts quantifications from GENCODE-based *recount2*.

In particular, we observed a high correlation between gene expression values based on these two related gene models, with a median correlation of 0.986 (p-value ≤ 0.0001) for the 32,922 genes in common. This result supports the notion that our pre-processing steps to disambiguate overlapping exons between strands did not significantly alter the gene expression quantification. Next, we analyzed the GTEx consortium dataset, which accounts for 9,662 samples from 551 individuals and 54 tissues types, to confirm tissue specificity for a selection of coding genes. We picked genes with known tissue-specific expression patterns, such as Keratin 1 (*KRT1*) Figure 6, Estrogen Receptor 1 (*ESR1*), and Neuronal Differentiation 1 (*NEUROD1*). These genes (as well as other tissue markers, data not shown) were confirmed to mostly expressed in skin, uterus and brain tissue samples, respectively, as expected (see Figure 6). Overall, all these genes presented very similar expression distribution across GTEx tissue samples when compared to the GENCODE-based *recount2* gene expression levels.

## Tissue-specific expression of lncRNAs

We used GTEx data to assess expression and specificity profiles across samples from each of the 54 tissues, stratified into four categories: coding mRNA, intergenic-lncRNA (i-lncRNA), divergent-lncRNA (d-lncRNA), and enhancers-lncRNA (e-lncRNA).

To this end, we were able to reproduce key findings from a recent FANTOM5 study where coding and long non-coding RNAs expression levels and specificity profiles among cell lines were evaluated[29]. Overall, the FC-R2 data showed similar profiles as Hon et. al, but with more variability likely due to the cellular complexity of tissue versus cell line data. Overall, coding mRNAs (log median expression = 6.6) were shown to be, more expressed than lncRNAs (log median expression = 4.14, 3.83 and 3.14, for i-lncRNA, d-lncRNA and e-lncRNAs respectively), although no clear difference in expression was observed among lncRNAs. For specificity levels, however, differences were observed among lncRNAS. Enhancers and intergenic promoters expression were notably more tissue-specific (median = 0.41 and 0.3) than divergent promoters and coding mRNAs (median = 0.13 and 0.09) (Figure 7A). With specificity being calculated based on the entropy of the median expression of each 54 tissues, normalized by log2 of the number of tissues:

$$S = empirical.entropy(\frac{\tilde{x}_1, \tilde{x}_2, \tilde{x}_3, \dots, \tilde{x}_n}{log_2 n})$$

When analyzing the percentage of genes expressed by category we can observe that coding mRNAs are ubiquitously expressed across all tissues types (mean = 88.42%), while lncRNA expression was more coordinated and specific, with enhancers showing the lowest percentage of expression by tissue (mean = 41.98%) (Figure 7B). These results are on par with common knowledge about lncRNAs expression, such as enhancer transcription being more tissue specific[56].

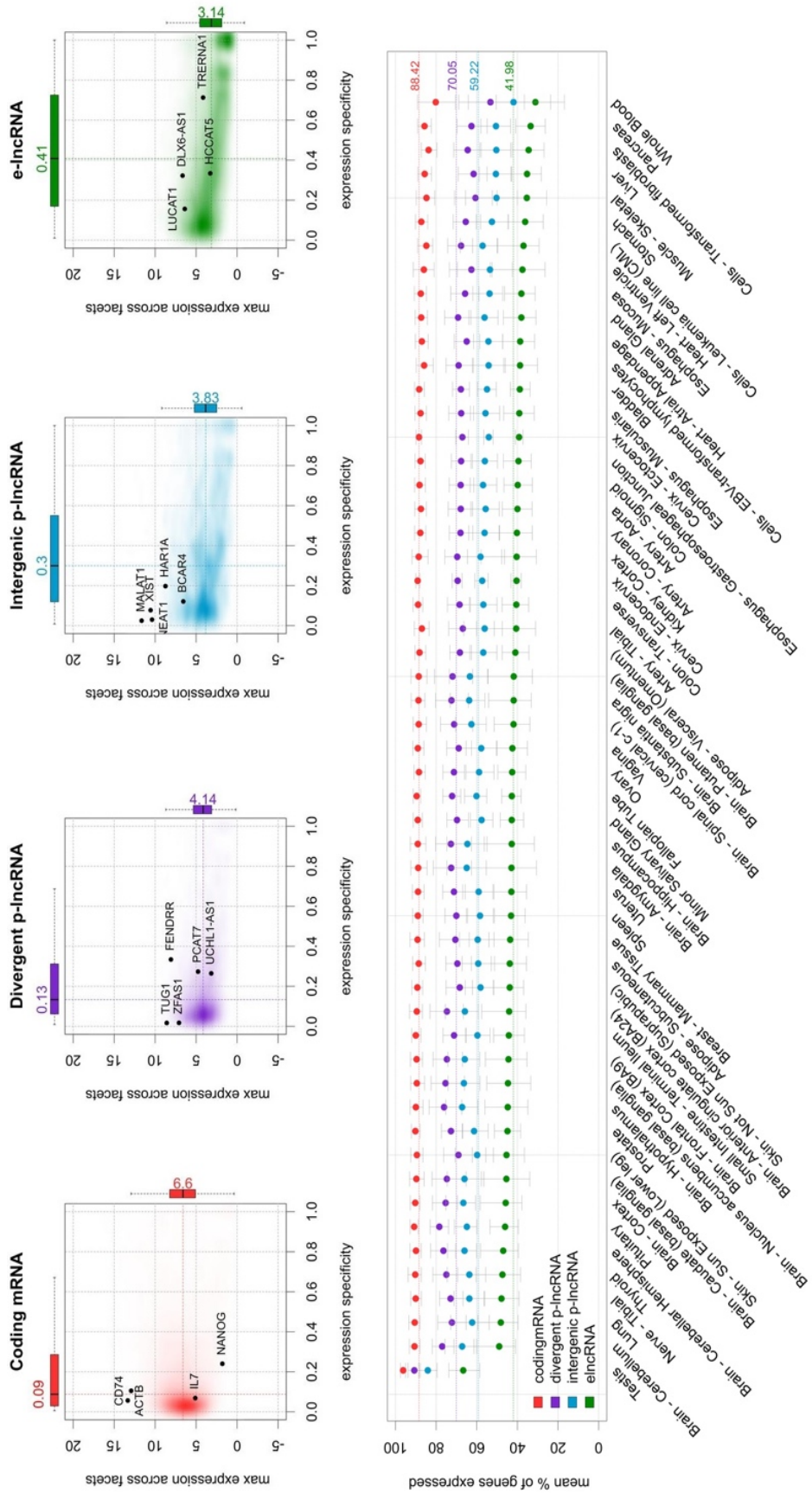



**Figure 7 - Expression profiles across GTEx tissues.**(A) Expression level and specificity of each RNA category. Y-axis shows the maximum expression level for each feature represented in transcripts per million (TPM) in log scale. X-axis shows specificity (entropy of genes among tissue facets) for median expression values summarized by tissue type. (B) Percentage of genes within categories expressed within each tissue facet. The dots represent the mean among samples within a facet and the error bars represent 99.99% confidence intervals. Dashed lines represent the means among all samples.

**Differentially expressed lncRNAs in cancer across the TCGA**

We performed differential gene expression analysis comparing tumor versus normal samples across 13 tumor types separately using re-quantified data from the TCGA consortium. Our results yielded a list table of DE genes that contained both overall and cancer-specific expression signatures that have previously been reported by several studies over the years. Overall, we identified 476 coding mRNAs and 48 lncRNAs genes differentially expressed across all the 13 tumor types analyzed at False Discovery Rate (FDR) ≤ 0.01 (Table 3).

Downregulated d-lncRNAs, were mostly those associated with immune cells (e.g. natural killer cells, -ß T cell, and mature ß-cells). Three genes, *RP11-276H19*, *RPL34-AS1*, and *RAP2C-AS1* were reported to be implicated in cancer. The first allowed for epithelial-mesenchymal transition, the second is related with increase in tumor size, and the latter was reported upregulated in urothelial cancer developed in patients after renal transplantation[57–59]. Among d-lncRNAs that were upregulated, *SNHG1* was implicated in cellular proliferation, migration, invasion of cancers and particularly described as upregulated in gastric cancer[60].

Similarly, the gene *RP11-572O17* was best implicated and associated with urinary bladder neoplasms. The remaining 13 genes were associated with a wide-variety of tissues or cells, including embryonic stem cell, pituitary gland, myeloid progenitor cells, lymphoid cells, among others. Among these, *AC068831* was recently reported as upregulated in myocardial infarction compared to unstable angina using patients whole blood samples mRNA after the event[61].

In downregulated e-lncRNAs, although there was no cancer related information in this category, some genes were associated with other human diseases e.g. *RP5-965F6* was reported to be upregulated in late-onset Alzheimer disease[62].

**Table 3 - Number of significantly differentially expressed genes (FDR ≤ 0.01) comparing tumor vs normal samples across tumor types. Values in parenthesis represent the number of FANTOM-CAT exclusive genes.**

| | | d-lncRNA | | e-lncRNA | | i-lncRNA | | mRNA | |
|---|---|---|---|---|---|---|---|---|---|
| Cancer type | Total | Up | Down | Up | Down | Up | Down | Up | Down |
| Bile | 7010 | 200 (60) | 313 (90) | 186 (89) | 203 (99) | 47 (12) | 84 (17) | 2658 (106) | 3319 (97) |
| Bladder | 7680 | 344 (125) | 319 (87) | 140 (68) | 149 (67) | 65 (19) | 82 (7) | 3112 (201) | 3469 (61) |
| Breast | 15290 | 753 (291) | 721 (202) | 656 (377) | 583 (305) | 207 (50) | 178 (32) | 6109 (296) | 6083 (244) |
| Colorectal | 13685 | 490 (164) | 592 (168) | 381 (203) | 400 (196) | 130 (32) | 160 (28) | 5538 (371) | 5994 (132) |
| Esophagus | 4883 | 87 (21) | 193 (50) | 90 (38) | 184 (103) | 40 (11) | 48 (2) | 1921 (83) | 2320 (77) |
| Head and Neck | 10517 | 442 (138) | 401 (96) | 267 (139) | 251 (112) | 100 (23) | 109 (18) | 4329 (256) | 4618 (53) |
| Kidney | 15697 | 734 (238) | 820 (281) | 535 (299) | 486 (209) | 203 (45) | 200 (48) | 6349 (525) | 6370 (114) |
| Liver | 10554 | 346 (94) | 395 (106) | 230 (102) | 248 (123) | 90 (16) | 112 (19) | 4164 (174) | 4969 (95) |
| Lung | 17143 | 864 (338) | 835 (304) | 893 (512) | 729 (396) | 242 (76) | 213 (39) | 7523 (532) | 5844 (212) |
| Prostate | 13183 | 686 (287) | 654 (218) | 418 (254) | 452 (214) | 175 (55) | 167 (30) | 5153 (489) | 5478 (128) |
| Stomach | 11309 | 528 (213) | 518 (164) | 462 (291) | 436 (240) | 144 (51) | 129 (22) | 4509 (558) | 4583 (89) |
| Thyroid | 14264 | 752 (284) | 804 (318) | 527 (295) | 594 (332) | 161 (39) | 174 (47) | 5403 (189) | 5849 (308) |
| Uterus | 12906 | 641 (285) | 713 (235) | 454 (263) | 612 (341) | 210 (79) | 225 (54) | 5135 (335) | 4916 (181) |

The one gene found to be ubiquitously downregulated in i-lncRNAs, *LINC00478*, has been reported in a wide variety of neoplasms from leukemia, breast, vulvar, prostate and bladder cancer[63–67]. In vulvar squamous cell carcinoma, there is a statistical relationship between *LINC00478* and *MIR31HG* expression and tumor differentiation[64]. Additionally, *LINC00478* was found to be significantly downregulated in patients with Estrogen Receptor positive ($ER^+$) breast tumors. The loss was associated with tumor progression, recurrence, and metastasis[65]. In contrast, an upregulated i-lncRNA, *SNHG17*, was associated with long term survival and its overexpression was correlated with tumor size, TNM (Tumor, Node, Metastasis) stage, lymph node metastasis in colorectal patients [68,69]. i-lncRNA *AC004463* was found upregulated in liver cancer and metastatic prostate cancer[70].

For the downregulated mRNA's, most of the genes that appeared on the list were associated with metabolism/oxidative stress. Among common categories that genes were associated with were transcriptional activator/repressors, and cell migration/adhesion. There were very few genes associated with DNA damage repair, and apoptosis. Conversely, in the upregulated mRNA's, most of the genes had functions that dealt with the cell cycle and replication, as well DNA damage/repair, chromosome/chromatid segregation and spindle checkpoints.

When looking at specific signatures such as in prostate cancer for example, several coding genes such as *ERG*, *FOXA1*, *RNASEL*, *ARVCF* and *SLC43A1* already reported[71,72] to be involved in prostate cancer progression and mortality were significantly DE (FDR ≤ 0.01) and ranked (by absolute FC) high in our table of DE genes.

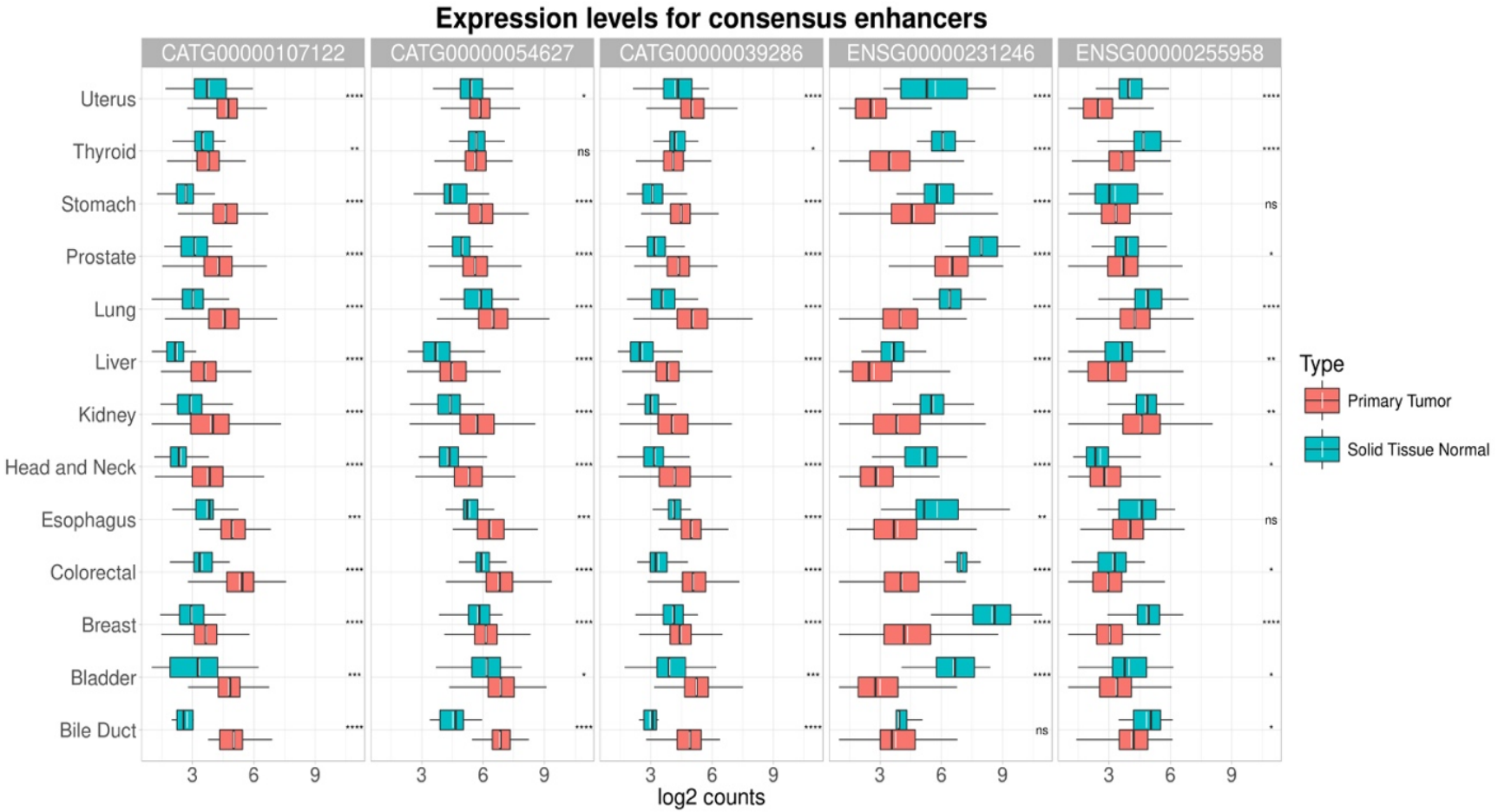



**Figure 8 - Enhancer RNA (e-lncRNA) differential expression across cancer types in TCGA.**The figure show enhancers ubiquitously differentially expressed across all cancer types in FC-R2 TCGA data. Three of these are annotated only in the FANTOM-CAT transcriptome meta-assembly.

Most notably, our approach was also able to identify several classes of non-coding RNAs involved in prostate cancer such as divergent and intergenic lncRNAs, as well as host genes of miRNAs.

Among differentially expressed non-coding genes were: *PCA3*, the first clinically approved lncRNA marker for prostate cancer and highly expressed in high grade prostate cancer tumors [73,74];*PCAT1*, a prostate-specific lncRNA reported to be involved in disease progression in high grade prostate cancer[75]; *MALAT1*, a lncRNA previously associated with several types of cancers and linked to poor prognosis in prostate cancer[76]; *ANRIL*, an anti-sense lncRNA that blocks activity of tumor suppressor genes and has shown elevated levels of expression in prostate cancer[77,78], among several others lncRNAs already reported to be DE in prostate cancer.

We have also been able to detect DE for small ncRNAs such as mir-375, with signal likely coming from the host gene, reported as biomarker for castration-resistant prostate cancer[79]. In addition to observing known coding and non-coding genes, our resource also highlighted several non-coding genes restricted only to FANTOM-CAT annotations that were not been associated with tumorigenesis before. This includes potential oncogenes such as novel enhancers found differentially expressed across almost all tumor types (Figure 8) showing the potential of the resource in prospective analysis.

**Enhancer expression levels associated with increased cancer survival**

Since TCGA analysis highlighted several potential lncRNAs markers across tumor types, specially enhancers due its high specificity, we evaluated the prognostic potential of enhancers.

In a recent work Chen and collaborators surveyed enhancers expression in nearly 9000 patients[53]. By using enhancer coordinates from Anderson et al.[80] they were able to detect enhancers

549 with prognostic potential across TCGA samples. Out of 65,433 enhancers analyzed in the
550 mentioned study, 4,803 were found to have prognostic potential. We compared these findings with
551 our resource. However, since FC-R2 is based on the meta-assembly by Hon et al.[29], some
552 differences were observed, as we could not map some of the enhancers detected by Anderson et
553 al. (2014) to Hon et al. (2017) enhancers since some enhancers were later classified as another
554 RNA class or removed in the latest atlas.

555 When comparing normal with primary tumor tissues from the same patients, we could
556 observe that most cancer types showed global enhancer activation (paired t-test, p-value ≤ 0.05,
557 with at least 10 matched tumor-normal pairs). The results we obtained were 100% on par with
558 enhancer activation patterns reported by Chen et al.[53] for cancer types with significant p-value
559 (Figure 9).

**Figure 9 - Global enhancer activation in cancer.**Figure contain enhancer expression changes across tumors with at least 10 matched normal and primary tumor samples. The y-axis shows changes in global enhancer expression (TPMtumor/TPMnormal - 1); statistics were performed with paired t-test corrected by BH method. On the left is data presented by Chen et al., on the right is data based on FC-R2 (COAD includes colon and rectal cancers.)

Moreover, using Cox proportional models for survival analysis, we evaluated the predictive power of e-lncRNAs among patient samples available within TCGA cohorts (Supplementary Table 1). Out of the 13 TCGA types analyzed, 11 showed significant predictive enhancers expression levels ranging from 3 in head and neck neoplasms to 3,850 in kidney cancers (average: 561). A total of 5,382 e-lncRNAs were identified with predictive potential (FDR ≤ 0.05), and no single e-lncRNA presented predictive power across all cancer types. When paired evaluations were performed, enhancers from kidney neoplasms showed important overlapping with neoplasms arisen from different tissues, the most striking from the uterus (n = 261) and the stomach (n = 141). Four of the five enhancers differentially expressed across all tumor types (Figure 8) were identified among the predictive enhancers. The majority were identified among kidney tumors: *CATG00000107122*, *CATG00000054627*, *CATG00000039286*; one in stomach tumors: *ENSG00000255958*; and, one in uterine tumors: *CATG00000039286*. Despite differences across annotations, we could detect 3054 enhancers out of 4803 evidenced by Chen et al. with prognostic potential. This includes "enhancer 22"/*ENSG00000272666* (chr22:50980817 - 50981280), which was highlighted as a promising marker of poor survival prognosis for some tumors in the mentioned study. We identified this e-lncRNA and obtained a similar survival curve on kidney cancer depicting poor prognosis for patients in the higher expression group (Figure 10).

# Discussion

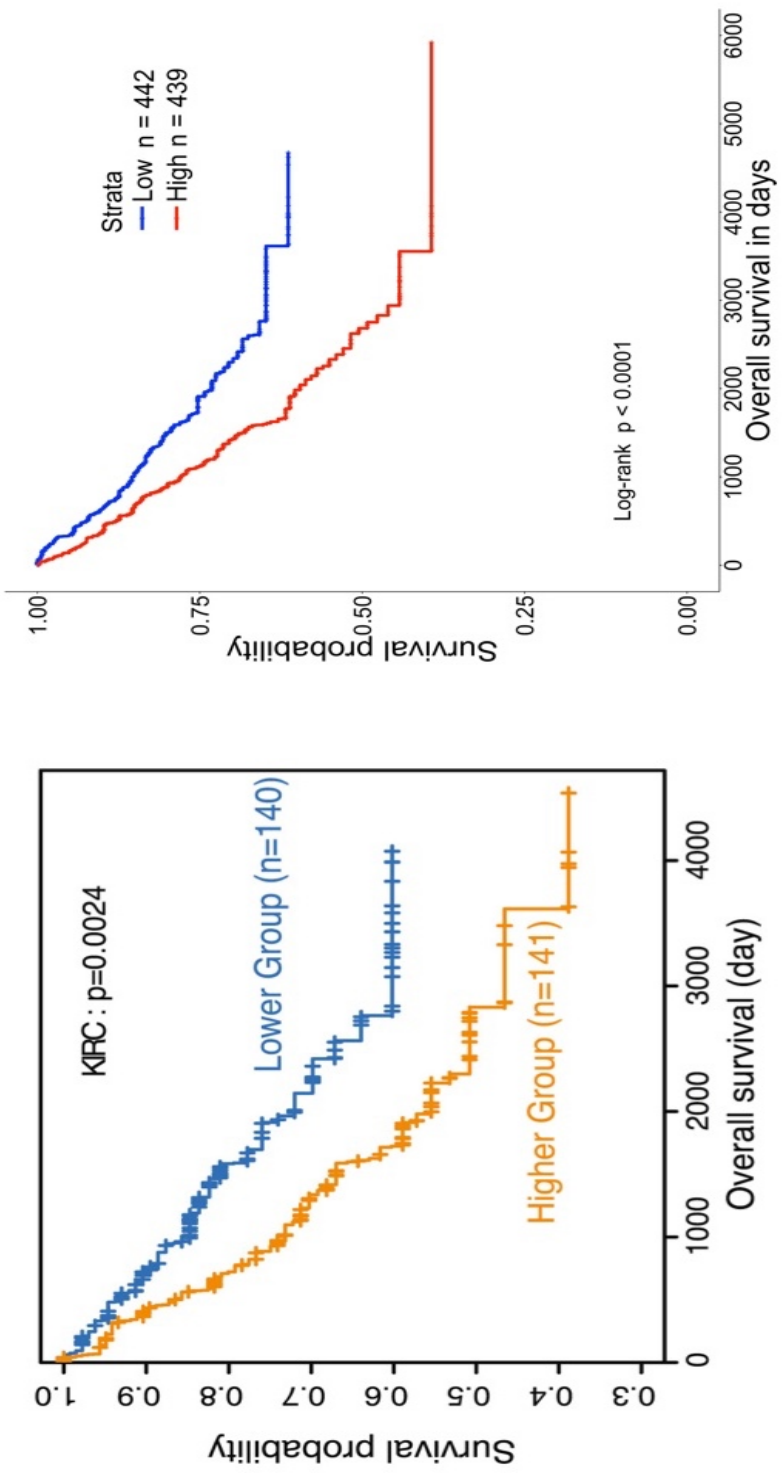



**Figure 10 - Enhancer 22 prognostic potential. Left:** Kaplan-Meier curve depicting predictive Enhancer 22 expression associated with worse cancer prognostic in Kidney RCC. **Right:** Same enhancer in FC-R2 showing similar worse behavior for all kidney cancer subtypes in TCGA. Cut-off made by median.

583 The role of lncRNAs in human diseases, such as cancer, has been increasingly appreciated
584 over the past few years. Recently, several classes of lncRNAs have been demonstrated to play
585 crucial roles in cell regulation and homeostasis[28]. Enhancers are a major category of regulatory
586 elements in cell regulation, which are also critical players in oncogenic process[81]. Despite its
587 importance, large-scale studies of enhancers are rare, in part due to the technical difficulty of
588 applying high-throughput techniques such as Chromatin Immunopreciptation (ChIP)-seq and Hi-
589 C over large cohorts to study enhancer activity. Chen et al.[53] have recently presented the detection
590 and characterization of many expressed enhancers in a genome-wide analysis using RNA-seq data
591 from the TCGA. As pointed in the study, despite relatively low depth of RNA-seq data from large
592 cohorts such as TCGA, which might increase variation, the expression data provides a valuable
593 dimension of information that is complementary to other technologies and can help in the study of
594 these elements.
595 By integrating recent findings from the FANTOM consortium and providing a resource by
596 the means of recount2, using a standardized pipeline, we enable the possibility of an analysis
597 integrating multiple datasets from thousands of studies. The FANTOM consortium recently
598 refined transcription start sites (TSS) for annotated genes and unveiled thousands of new lncRNAs
599 supported by CAGE-seq data[29]. Our resource also provides complementary expression evidence
600 from thousands of samples for several transcripts discovered by the FANTOM consortium. These
601 characteristics makes the recount2/FANTOM-CAT a unique resource to study, particularly when
602 focusing on lncRNAs. By providing a comprehensive atlas of expression for several lncRNAs,
603 such as enhancers and promoters it enables scientists to investigate the role of uncharacterized
604 lncRNAs, and associate these to phenotypes of interest such as human diseases.

We first applied GLMs to detect differential expressed genes across tumor and normal samples to test if our resource was able to capture the current landscape of genes that are already known to be involved in those tumor types. The analysis across 13 cancer types (summary in table 3) yielded a comprehensive list of mRNAs, promoters, enhancers and other ncRNAs significantly differentially expressed. In these lists we were able to confirm several biomarkers and genes that were already reported to be shared across cancer types and genes whose expression is restrict to specific tissue of origin. Interestingly, we also observed many lncRNAs whose function is unknown or that have never been reported to be involved in the development of several cancer types shedding a light on potential new oncogenes and tumor suppressors.

Strikingly, we observed how much information would not be capture if one would not rely on FANTOM-CAT annotations. We found that several genes differentially expressed across tumor types were exclusive tied to the FANTOM-CAT annotations. Across all models we applied, a total of 28,207 differentially expressed genes were contained only in FANTOM-CAT, suggesting that by relying solely on other resources such as GENCODE v25 quantifications, one would have missed on average 1,087 up-regulated and 982 down-regulated genes for each of the phenotypes analyzed in this study (Table 3), with most of them being lncRNAs.

In addition to GLMs, we also applied univariate cox proportional-hazard models looking particularly for enhancers showing predictive survival potential. We have been able to uncover several enhancers that when stratified by expression levels (split by the median) led to better or worse clinical outcomes. The potential of enhancers as prognostic features has been recently explored by Chen et al.[53] which led to the detection of promising enhancers with high prognostic potential. Using our data, we have been able to recover a high number of predictive enhancers that were described in the mentioned study, including key ones such as “enhancer 22”. Additionally,

since our resource is based on more recent data, we have been able to uncover additional prognostic enhancers or curating the ones that were no longer classified as enhancers. Moreover, we also computed the proportion of global enhancer activation between tumor and normal samples and compared it to the profiles reported in the same study. Although the levels of enhancer activation were different, which was expected given the different approaches in the resources used, the direction and significance in each statistically significant cancer type were 100% on par (Figure 9).

By confirming findings reported by other studies, we are providing extra evidence that these genes, despite not yet fully understood, might have a biological function and can potentially be leverage as prognostic biomarkers. Also, we demonstrate how our resource can be of use in providing quick and reliable expression information for several lncRNAs classes, such as enhancers and promoters which is not readily available in any database to date. This resource is opening the doors for further research on the mechanisms implicated in the development and behavior of cancer and other diseases. The results establish our gene expression atlas as a reliable resource to perform large scale transcriptomics studies and with over 70,000 samples ready to analyze it provides a suitable environment for the study of the role ncRNAs play in cancer development, as well in other diseases which in turn can reveal important cues to understand their biological function. All code used in these analysis are available in: https://eddieimada.github.io/fcr2/ and data can be downloaded from: https://jhubiostatistics.shinyapps.io/recount/ or through the Bioconductor package *recount.*

# Case studies

## Case study 1: Transcriptional landscape of PTEN loss in PCa

### Background

Prostate cancer (PCa) is the second most prevalent form of cancer in men (after skin cancer), with an estimated worldwide number of 1,600,000 cases and 366,000 deaths annually[82]. PCa in younger men (age < 40) is rare, but the chance of developing PCa rises rapidly after age 50. It is estimated that 6 in 10 cases of PCa are found in men older than 65[83]. Despite recent progress in treatment and detection, PCa remains a significant medical problem. Due to its complexity, overtreatment of inherently benign tumor and inadequate therapy choice for metastatic PCa is often observed. PCa follows a multistep process of development. It initiates as prostatic intraepithelial neoplasia followed by localized PCa and prostate adenocarcinoma (PRAD) with local invasion which ultimately culminates in metastatic PCa.

667 Being such complex disease, PCa evaluation is guided by the Gleason grading system.
668 Originally defined by Donald Gleason[84], the score is based on the cell morphology assessed by a
669 pathologist. It is composed of 2 scores, the primary and secondary grade, where the former is based

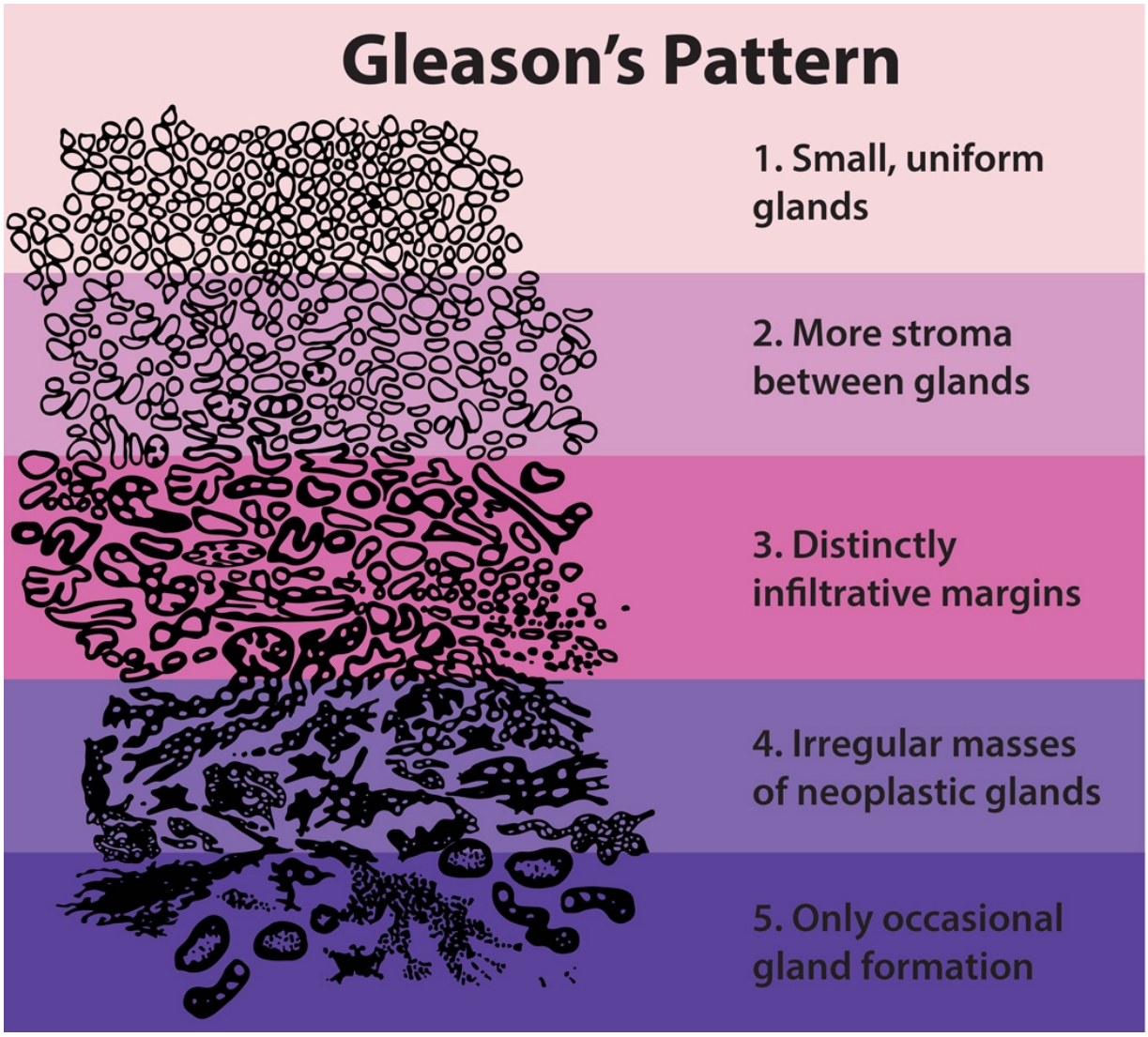



**Figure 11 – Gleason Scoring System.** The Gleason scoring system is based on cell morphology. Higher scores are linked to worst prognosis and vice-versa.

on the dominant cell morphology and the latter based on the non-dominant. Each grade ranges from 1 to 5, with the final Gleason score being the sum of both grades and therefore ranges from 2 to 10 with higher numbers indicating greater risks and mortality.

Although Gleason score is still widely used, the scoring system has several problematic aspects. Although the score starts at 2, the score of the majority of PCa cases starts at 6. While this is not a problem with people familiar with the grading system and the disease biology (e.g. pathologists, researchers, and medical staff), patients who are not familiar with the subject may interpret the 6 out 10 score as a more aggressive cancer, causing greater anxiety. Another major problem is that the classification system fails to distinguish between 3+4 and 4+3 scores, with the latter having a worse prognosis. Therefore, in 2014 a 5-point Gleason grade was proposed, validated and latter accepted by the World Health organization the new method to be used in conjunction with the 2005 Gleason system[85,86].

Genomic alterations are very common in cancer diseases. In primary PCa the most common genomic alterations involve androgen regulated promoters and the ETS family of transcription factors (e.g. *ERG* and *ETV* genes)[87]. *TMPRSS2* to *ERG* fusion (*TMPRSS2-ERG*) is the most prevalent form of alteration with a prevalence of around 50% in localized PCa[88]. Similar alterations involving *TMPRSS2* fusion are often found between *ETV1*, *ETV4* and *ETV5*. *TMPRSS2* expression is regulated by androgenic hormones which are often increased in PC tumors, leading to the overexpression of fused genes. Overexpression of *ETS* family genes is associated with *PTEN* inactivation by deletion. *PTEN* is a key tumor suppressor gene, often deleted in PCa, which control many aspects of cellular proliferation by regulating the PI3K-Akt-mTOR pathway[89,90].

An increase in *MYC* transcription factor copy number is often present in PC tumors even at early stages of development[91]. *MYC* promotes expression of many proliferative genes by binding

693 at enhancer regions and histone acetylation via recruitment of histone acetyltransferases and it is
694 a key gene in cell cycle regulation. Overexpression of *MYC* is directly associated with increased
695 cellular proliferation.

696 Another commonly mutated gene is *SPOP*, which encodes an E3 ubiquitin ligase
697 component, and the mutated protein causes stabilization of oncogenic substrates such as *MAPK8*
698 (*JNK*), *NCOA3*, and *DEK*[92,93]. Recently, a *SPOP* mutant *GEMM* pointed *SPOP* as a driver of
699 prostate tumorigenesis through activation of both PI3K/mTOR and Androgen Receptor (AR)
700 signaling, and effective uncoupling of the normal negative feedback between these two
701 pathways[93]. Some studies have shown that *SPOP* mutants displays loss of the chromatin
702 remodeling factor *CHD1*[92,93], but these observations are in contrast to recent work demonstrating
703 that *CHD1* represents an essential effector of *PTEN* deficiency in prostate cancer[94].

704 Epigenetic control plays a vital role in cellular homeostasis therefore it is not surprising to
705 find that genes controlling epigenetic process (e.g., methylation, histone modification and
706 nucleosome remodeling) are often found deregulated in many cancer types[95], including PCa.

707 DNA methylation leads to suppressed gene expression when occurring in its promoter
708 region. DNA can be methylated by canonical DNA methyltransferase (*DNMT*) which is often
709 causing methylation of cytosines in CpG islands. The methylated cytosine can then be converted
710 into 5-hydroxymethylcytosine (5hmC) by proteins of the TET family (i.e. *TET1*, *TET2* and *TET3*).
711 While DNA methylation is normal in cellular homeostasis and it is required to ensure proper
712 regulation of gene expression, aberrant DNA methylation. Hypermethylation of promoter regions
713 and global hypomethylation – can lead to genome instability and tumor development through
714 silencing of tumor suppressor genes[96].

*DNMT1* has been shown to have a dual role in PCa, acting as a tumor suppressor in early stage and as oncogene in late stage[97]. Expression of DNMT has been shown to be regulated by TGF-beta in PCa, with their expression levels associated with aggressiveness and recurrence[94]. Both *TET1* and *TET2* genes were shown to be tumor suppressors in PCa capable of regulating cell proliferation, migration, and invasion[94,98]. Many mutated epigenetic regulators and chromatin remodelers have been identified by genomic profiling in up to 20% of primary PCa. Among these regulators are the *ASXL1*, *KMT2C*, *KMT2D*, *KMTD2A*, *KDM6A*, *SETDB2* and *SETDB1* and among chromatin remodelers are *ARID1A*, *ARID4A*, *ARID2*, *SMARCA1* and some members of the SWI/SNF nucleosome remodeling complex. These mutations are significantly enriched in PCa were *ETS* fusions or driver mutations such as *IDH1*, *SPOP*, *CUL3* or *FOXA1* are present. These mutations are also associated with higher Gleason score in primary tumors[99]. Interestingly, the long non-coding *SChLAP1* has been shown to antagonize the function of the SWI/SNF complex which contributes to its oncogenic funsction[100]. Some members of the Polycomb group protein complexes, which suppresses transcriptional programs by methylation also contribute to PCa development. *EZH2*, a methyltransferase of Polycomb-repressive complex 2 (*PRC2*), is often overexpressed in cancers and has been demonstrated to promote PCa progression[100].

In this section, we present two case studies using the FC-R2 atlas to show how this resource can be used to uncover novel lncRNAs associated with diverse phenotypes.

## Methods

### *Data collection*

All expression data used in this work were gathered from public domain databases. In this work we made use of three cohorts: FC-R2 TCGA, Natural History (NH) and Health Professionals Follow-up Study (HPFS). Information about each cohort is summarize on Table 4.

Information about *PTEN* status by immunohistochemistry for the HPFS cohort was readily available and therefore obtained from public domain. For NH cohort samples IHC staining for PTEN was performed in partnership with Dr. Tamara Lotan. Last, for TCGA we evaluated a classification approach using expectation maximization (EM) algorithm and Copy-Number-Variation (CNV) called by the GISTIC algorithm to define *PTEN* status.

**Table 4 - Cohorts summary.**

| Cohort | PTEN(-) | PTEN(+) | N |
|---|---|---|---|
| TCGA | 95 | 321 | 416 |
| HPFS | 91 | 299 | 390 |
| Natural History | 56 | 151 | 207 |
| Total | 242 | 771 | 1103 |

***FC-R2 TCGA status call***

Since PTEN status was not available for the TCGA cohort we evaluated a classification approach using the EM algorithm to call PTEN status based on its expression level. We applied this method on all cohorts with available Imunohistochemistry (IHC) data for *PTEN* (HPFS and NH). For this analysis IHC status was used as a gold-standard. Sensitivity and Specificity values were used to assess the approach performance. Alternatively, we defined *PTEN* status by copy-number-variation called by the GISTIC algorithm. To reduce data heterogeneity, we kept only samples with a GISTIC score of -2 (PTEN-null) or 0 (PTEN-normal).

***Differential expression analysis***

Differential expression analysis was performed in each cohort by applying a GLM approach coupled with empirical Bayes moderation of standard errors[54]. Surrogate variables were detected and estimated with SVA package and inputted in the model. Adjusted p-values controlling for multiple hypothesis testing was performed using Benjamini-Hochberg method[52] and genes with FDR ≤ 0.1 were reported.

***Meta-analysis of microarray-based cohorts***

We applied a meta-analysis approach using a Bayesian hierarchical multi-level model for cross-study detection of differential gene expression implemented in the Bioconductor package XDE[101] on microarray-based cohorts in order to obtain a PTEN-null signature from PTEN IHC validated samples. The model was fitted using the $\Delta_{gp}$ model with empirical starting values and 1000 bootstraps was performed. All remaining parameters were set to default values.

***Gene set enrichment analysis (GSEA)***

The lists of differentially expressed genes were tested for enrichment with Gene Set Enrichment Analysis (GSEA). GSEA was performed using fast geneset enrichment analysis implemented in the fgsea[102] package from Bioconductor with 10000 permutations. A collection of genesets were obtained from the Broad Institute MSigDB database[103]. Genesets with less than 15 and more than 500 genes were removed from the analysis. The lists of differentially expressed genes were ranked by t-statistics and were used as input together with the genesets. The analysis was performed in two ranked lists: 1- TCGA differential expression analysis ranked by t-statistics and 2- Meta-analysis weighted size effect computed as:

$$(Concordance\ probability - Discordance\ probability) * Average\ Size\ Effect$$

## Results

### *GLMs models fail to capture homogeneous signal across cohorts*

We have analyzed all available gene expression datasets to identify genes and pathways differentially expressed upon *PTEN* and *ERG* loss, as well stratified contrasts e.g. *PTEN* loss in different *ERG* loss background. To this end, we applied a generalized linear model (GLM) approach coupled with empirical Bayes moderation of standard errors and adjusted p-values controlling for multiple hypothesis testing using Benjamini-Hochberg method and reported genes with FDR ≤ 0.01. This approach resulted in differential gene expression lists for each cohort, which revealed important insights about our datasets.

First, we ranked genes by t-statistics and computed the correspondence-at-the-top. Overall, this analysis revealed that models used for *ERG* status presented a good agreement across all cohorts (Figure 12). However, *PTEN* status models showed subpar agreement despite still being higher than expected by chance. Finally, agreement for *PTEN-ERG* interaction models showed no agreement across the analyzed cohorts. By looking at the number of significantly differentially

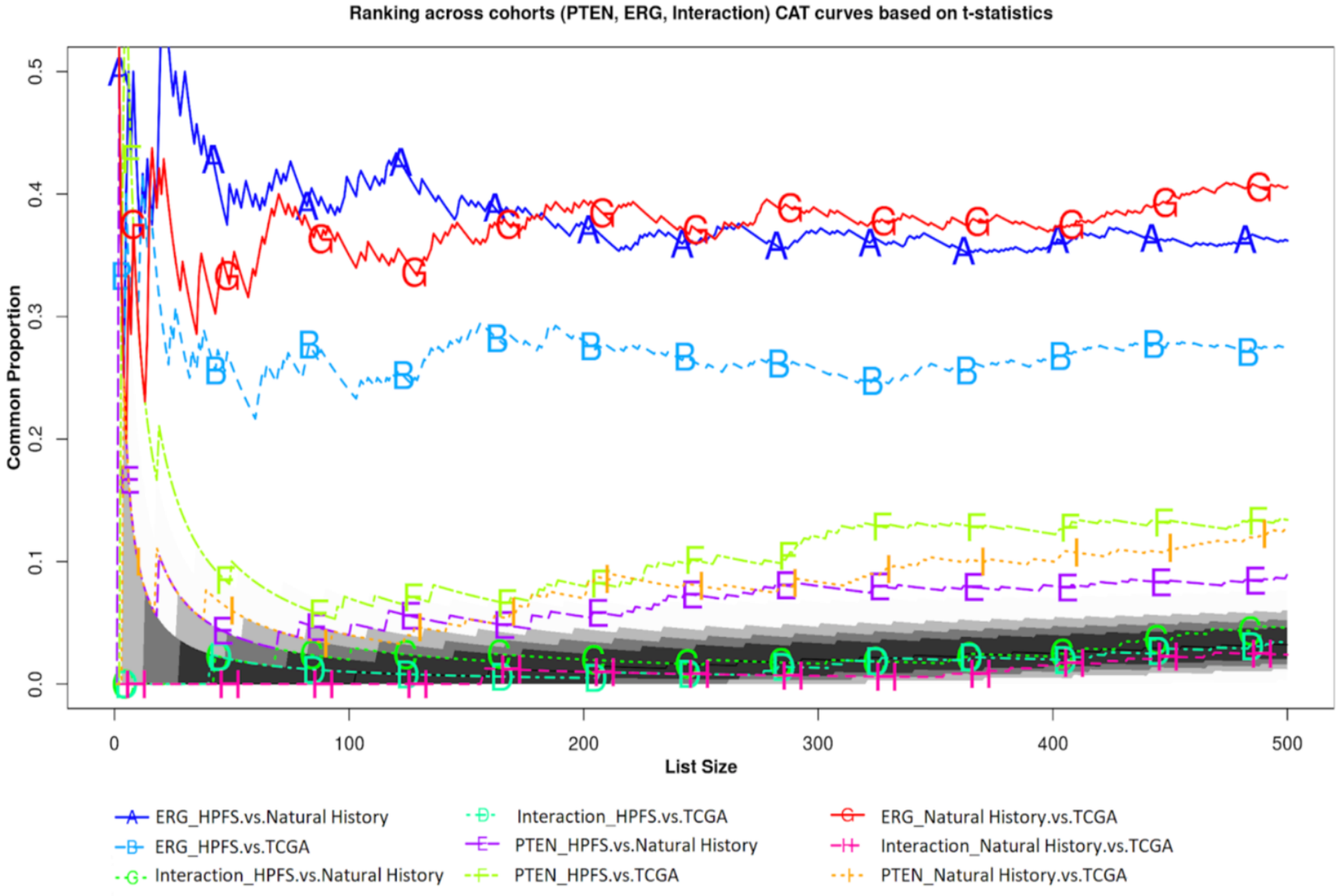



**Figure 12 - Correspondence-At-the-Top (CAT) plot.** Agreement of genes ranked by t-statistics as obtained from differential gene expression analysis. Lines represent agreement between tested cohorts for ERG, PTEN and PTEN-ERG Interaction. Black-to-light grey shades represent the decreasing probability of agreeing by chance based on the hypergeometric distribution, with intervals ranging from 0.999999 (light grey) to 0.95 (dark grey). Lines outside this range represent agreement in different cohorts with a higher agreement than expected by chance.

803 the interaction term models were detected in the remaining cohorts (Figure 12).

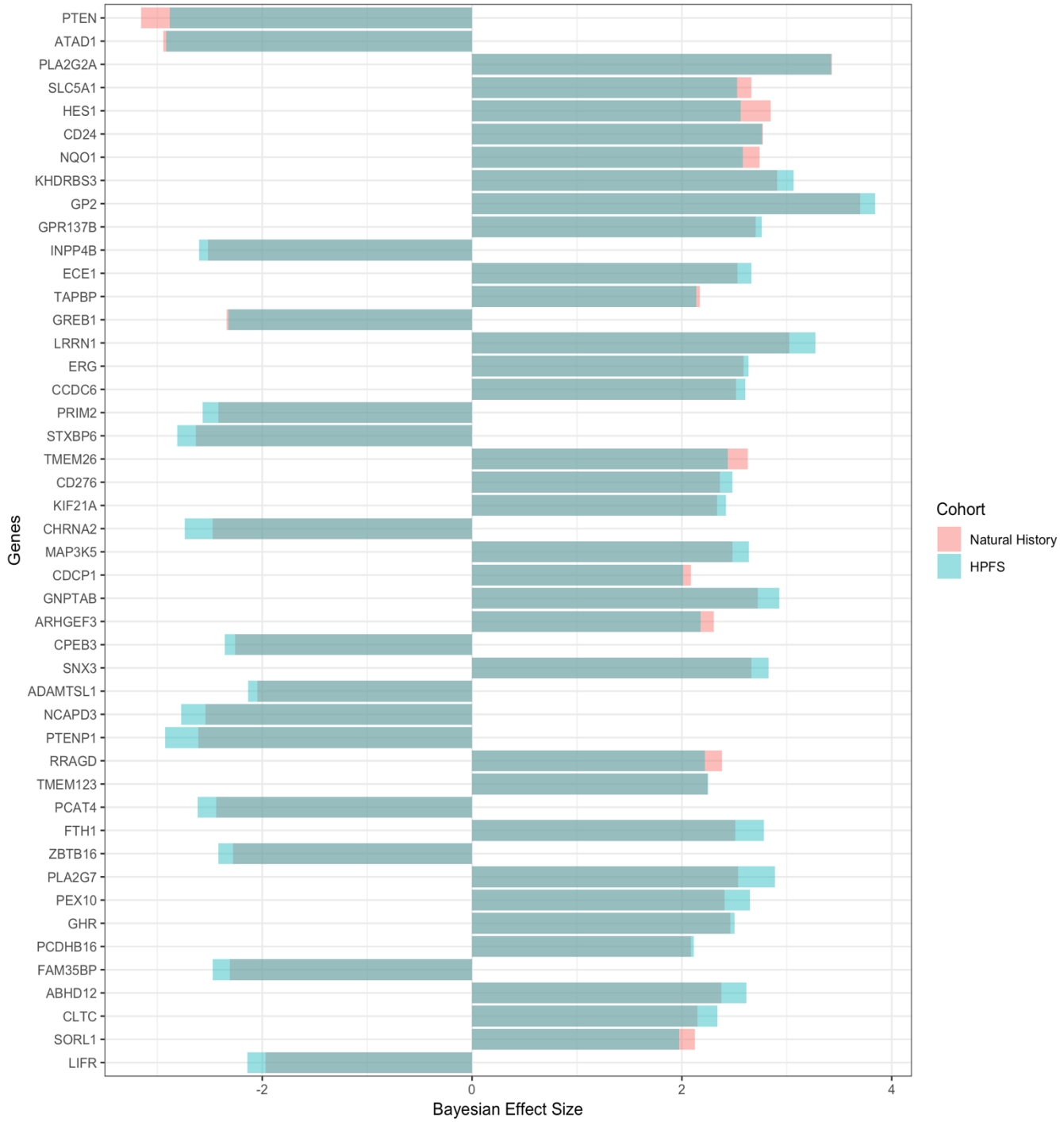



**Figure 13 - PTEN signature from meta-analysis.** PTEN signature obtained by multi-level model for cross-study detection of differential gene expression based on IHC calls on Natural History and HPFS cohorts. Figure shows the posterior probability of concordant differential expression across tested cohorts (left); and the effect size of each cohort.

804

805

806 ***Meta-analysis increases statistical power leading to higher rank agreement***
807

808 In order to tackle the issue, describe in the previous section, we switched to a meta-analysis
809 approach using multi-level model for cross-study detection of differential gene expression. Fitting
810 a Bayesian hierarchical model for analysis of differential expression across multiple studies
811 allowed us to aggregate information from microarray-based cohorts with IHC calls (our gold
812 standard), leading to a high statistical power to calculate the effect size and the posterior
813 probabilities of concordant/discordant DGE. We relied on this approach to generate a set of
814 signatures for both *PTEN* and *ERG* status (interaction models are not possible in this framework).
815 This approach yielded signatures with high concordance across two independent cohorts and
816 allowed us to obtain a reliable ranking leading to results with a much higher agreement with the
817 TCGA cohort for *PTEN* status (Figure 14) a result we could not obtain with the previous approach.

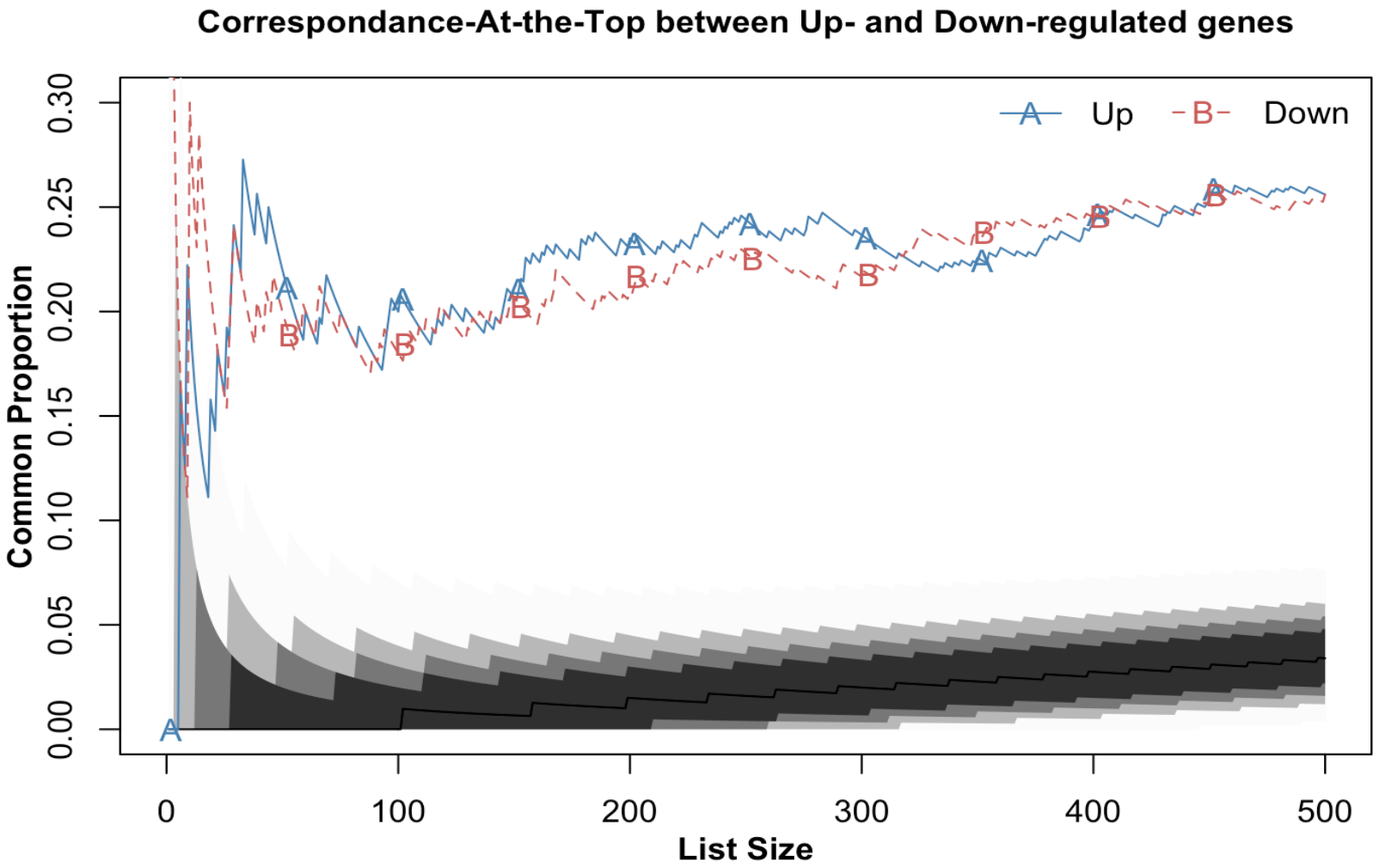



**Figure 14 - Correspondence-At-the-Top (CAT) plot.** Agreement of genes ranked by t-statistics in TCGA and weighted effect size ((Concordance probability - Discordant probability) * Average Size Effect) in Meta-analysis as obtained from each approach. Line represents agreement between PTEN model. Black-to-light grey shades represent the decreasing probability of agreeing by chance based on the hypergeometric distribution, with intervals ranging from 0.999999 (light grey) to 0.95 (dark grey). Lines outside this range represent agreement in different approaches with a higher agreement than expected by chance.

### *GSEA analysis reveals pathways leading to aggressive phenotype*

GSEA analysis were performed to obtain an overview of general processes and pathways affected by PTEN loss. This analysis revealed several pathways enriched in concordance across both cohorts (Table 5). Most interestingly, genesets associated with cellular proliferation (MYC1-targets, PI3K/AKT/mTOR pathway, mTORC1) and cellular mobility (CDC42-Rac and Actin-Y pathways) were found positively enriched across two independent cohorts. These processes are on par with the phenotype described in the literature which reports a more aggressive phenotype in PCa upon PTEN loss[104].

**Table 5 - Geneset enrichment analysis.**

| Gene Set Name | NES XDE | adj.P.Val | NES TCGA | adj.P.Val |
|---|---|---|---|---|
| HALLMARK_PROTEIN_SECRETION | 3.45 | 7.9e-04 | 2.12 | 9.5e-04 |
| HALLMARK_OXIDATIVE_PHOSPHORYLATION | 3.37 | 7.9e-04 | NS | NS |
| HALLMARK_INTERFERON_ALPHA_RESPONSE | 3.36 | 7.9e-04 | NS | NS |
| HALLMARK_MTORC1_SIGNALING | 2.88 | 7.9e-04 | 2.14 | 9.5e-04 |
| HALLMARK_MYC_TARGETS_V1 | 2.87 | 7.9e-04 | 2.10 | 9.5e-04 |
| HALLMARK_TGF_BETA_SIGNALING | 2.57 | 7.9e-04 | 1.81 | 9.5e-04 |
| HALLMARK_PI3K_AKT_MTOR_SIGNALING | 2.50 | 7.9e-04 | 1.67 | 3.6e-03 |
| HALLMARK_G2M_CHECKPOINT | NS | NS | 2.33 | 9.5e-04 |
| HALLMARK_E2F_TARGETS | NS | NS | 2.32 | 9.5e-04 |
| HALLMARK_ALLOGRAPH_REJECTION | 2.28 | 7.9e-04 | 1.42 | 3.5e-02 |
| HALLMARK_UNFOLDED_PROTEIN_RESPONSE | 2.08 | 4.3e-03 | 1.70 | 3.6e-03 |
| HALLMARK_ANGIOGENESIS | 1.86 | 2.8e-02 | 1.66 | 3.2e-02 |
| HALLMARK_MITOTIC_SPINDLE | 1.71 | 2.3e-02 | 2.03 | 9.5e-04 |
| HALLMARK_GLYCOLYSIS | 1.63 | 3.5e-02 | 1.83 | 9.5e-04 |
| HALLMARK_MYOGENISIS | -2.60 | 1.4e-03 | -2.22 | 1.4e-03 |
| BIOCARTA_CDC42RAC_PATHWAY | 2.35 | 6.4e-03 | 2.13 | 3.8e-03 |
| BIOCARTA_ACTINY_PATHWAY | 2.39 | 7.8e-03 | 2.02 | 1e-03 |
| REACTOME_SIGNALING_BY_WNT | 2.40 | 2.23e-04 | 2.08 | 1.5e04 |
| GO_REGULATION_OF_ESTABLISHMENT_OF_PLANAR_POLARITY | 3.09 | 5.5e-04 | 2.27 | 2.89e-04 |

## Discussion

*PTEN* is the most frequently mutated tumor suppressor (TS) gene in PRAD and other human cancer. With an estimated prevalence of up to 50%, *PTEN* loss is recognized as one of the major driving events in PRAD[89]. *PTEN* antagonizes PI3K-AKT/PKB and is a key modulator of the AKT-mTOR signaling pathways which are important in regulating the cell cycle. Therefore, *PTEN* loss

836 is consistently associated with more aggressive disease features and poor outcomes. Saal and
837 collaborators previously generated a transcriptomic signature of *PTEN* loss in breast cancer[105].
838 While this signature was correlated with worse patient outcome in breast and others independent
839 cancer datasets, including PRAD, the signature unsurprisingly, fails to capture key characteristics
840 of PCa such as *ERG*-rearrangement[105,106]. Moreover, despite the effects of *PTEN* loss being
841 extensively associated with a more aggressive phenotype and many genes already being associated
842 to it, a genomic signature, derived specifically from PRAD, reflecting the landscape of *PTEN* loss
843 in PRAD has not been described to date.

844 During our first approach using GLMs models to detect differentially expressed genes upon
845 *PTEN* loss and *ERG* rearrangement we discovered that this approach was problematic in assessing
846 DE across microarray-based cohorts (i.e. HPFS and NH) as pointed by Figure 12 these cohorts
847 presented variable results across different contrasts, more specifically in the *PTEN*-null vs *PTEN*-
848 normal. These results raised two possible hypothesis: 1) the technology used to measure expression
849 levels of the genes yields different results (TCGA cohort is based on Illumina sequencing, while
850 NH and HPFS/HPS are microarray-based); or 2) a larger sample size is necessary to reliably detect
851 signal for some of the models used in this approach. Based on CAT-plots shown in Figure 12, we
852 could expect that if the first hypothesis was right, we would observe low agreement across all
853 models used in this analysis. We rejected the former hypothesis due to *ERG* model presenting high
854 concordance levels. Thus, the latter hypothesis is more likely to account for the divergence in the
855 results.

856 To test this hypothesis and tackle this issue we applied a meta-analysis approach using a
857 Bayesian hierarchical multi-level model across microarray-based cohorts which allowed us to
858 increase the statistical power of the analysis by combining both cohorts in a single analysis. This

approach yielded results in a much higher agreement with the TCGA cohort for *PTEN* status (Figure 14), a result we could not obtain with the previous approach, thus confirming our hypothesis. The meta-analysis approach yielded a high number of highly concordant differently expressed genes (n=1745 with posterior probability higher than 70%) involved in a variety of processes known to be affected upon *PTEN* loss (e.g. cell cycle, proliferation, immune system). Similar results for differentially expressed genes were found in the TCGA cohort analyzed with GLM models (n=3940 with FDR ≤ 0.01).

GSEA analysis for *PTEN* rank showed several functional gene sets enriched in common with both approaches. For example, for *PTEN* loss functional gene sets related to cell cycle progression (i.e. *MYC* targets, *MTORC1* signaling, PI3K/AKT/MTOR signaling) and cellular motility (i.e. *CDC42-Rac* and Actin-Y pathways) were found positively enriched upon *PTEN* loss in both approaches (Table 5**)**. *MYC1* is a transcription factor belonging to the Myc family which increases the expression of several other genes, most of which are involved in cell proliferation. PI3K/AKT/mTOR is a signaling pathway important in cell cycle regulation which is often over-reactive in the absence of *PTEN* which leads to reduced cell apoptosis and increased cell proliferation. *mTORC1* signaling pathway is responsible for the control of protein synthesis. Together, increased activation of these pathways leads to an increased cell proliferation which are often observed in cancer. In addition, *CDC42* and *RAC1* genes are proteins that regulate several cellular processes, among which is cell migration. *CDC42* gene activates *PAK* genes (*PAK1*, *PAK2*, *PAK3*) which primary role is to initiate actin reorganization and regulate cell adhesion, migration, and invasion. These pathways are positively enriched upon *PTEN* loss, meaning that *PTEN*-null tumors present increased invasiveness due to increased mobility and proliferation. This invasive profile has been extensively reported across literature[90,104].

Collectively, the findings obtained at the gene level and those obtained at the gene set level confirmed that there is substantial agreement between our XDE meta-analysis and the analysis performed on TCGA (Figure 14). Hence, we can expect that results from TCGA cohort based on CNV calls are on par with other cohorts based on IHC calls. This confirmation is important because TCGA cohort harbor crucial information that is not available in other cohorts, such as expression levels for lncRNAs and SNPs associated with traits.

By leveraging our FC-R2 resource we have been able to detect a variety of lncRNAs that have already been described in PCa development and progression such as *PCA3*, *PCGEM1* and *KRTAP5-AS1*. *PCA3* is a lncRNA prostate-specific which is overexpressed in PCa tissue. *PCA3* acts by a variety of mechanisms such as downregulation of the oncogene *PRUNE2* and upregulation of *PRKD3* gene by acting as a miRNA sponge for mir-1261leading to increase proliferation and migration[107,108]. Conversely, knockdown of *PCA3* can lead to partial reversion of epithelial-mesenchymal transition (EMT)[109] which can lead to increased cell invasion, motility and survival[110]. Similarly, lncRNA *PCGEM1* expression is increased and highly specific in PCa it promotes cell growth and it has been associated with high-risk PCa patients[111,112]. On the other hand, *KRTAP5-AS1* expression has not been directly associated with PCa. However, it has recently shown that *KRTAP5-AS1* can act as a miRNA sponge for miRNAs, such as mir-596, targeting the oncogene *CLDN4* which enhances the invasion capacity of cancer cells and promote EMT[110,113], thereby overexpression of *KRTAP5-AS1* can lead increased levels of *CLDN4*[114]. *Mir-596* has also been shown to be overexpressed in response to androgen signaling and associated with antiandrogen therapy resistance[115]. In our analysis we observed that both *PCA3* and *PCGEM1* were downregulated upon loss which tracks together with the increased invasive profile observed

in *PTEN*-null tumors. Similarly, *KRATAP5-AS1* was found upregulated in *PTEN*-null tumors reinforcing a potential role for EMT leading to more aggressive phenotype.

We also observed several lncRNAs exclusively annotated in the FANTOM-CAT associated with *PTEN*-loss. Since these genes are novel genes without elucidated function, we analyzed potential roles for these genes by looking at other genes in same loci. Among the most downregulated FANTOM-CAT exclusive genes were *CATG00000038715*, *CATG00000079217* and *CATG00000000330*. *CATG00000038715* is in proximity of cytochrome P450 enzymes, more specifically *CYP4F2* and *CYP4F11*. Both enzymes are involved in the process of inactivating and degrading leukotriene B4 (*LTB4*). *LTB4* is a key gene in inflammatory response which are produced in leukocytes in response to inflammatory mediators and is able to induce the adhesion and activation of leukocytes on the endothelium[116]. As recently demonstrated by Wculek & Malanchi[117], leukotrienes can provide a selective proliferative advantage to cancer cells with intrinsically higher tumorigenicity, therefore downregulation of the intergenic promoter *CATG00000038715* together with *CYP4F2* and *CYP4F11* can lead to increased leukotrienes levels upon *PTEN* loss resulting in the selection of PCa cells with higher tumorigenicity. *CATG00000079217* is closely located with the coding gene *FBXL7* which has been shown to regulate Survivin stability which overexpression is known to lead to poor prognosis in several cancers[118]. *FBXL7* is also known to regulate mitotic arrest and mediate Class I MHC antigen processing and presentation[119]. *CATG00000000330* is located in the same loci of *PTEN* which has already been extensively document as a key oncogene in PCa[120].

Among upregulated lncRNA FANTOM-CAT genes, *CATG00000117664* was among the most upregulated lncRNA. Located in close proximity with the androgen regulated gene *GPR158*

which is reported to stimulate cell proliferation in prostate cancer cell lines, and it is linked to neuroendocrine differentiation[121].

Altogether, we have shown that these novel lncRNAs harmoniously track together with several coding mRNAs and lncRNAs already reported to be involved in PCa development and progression. This analysis reveals a plethora of lncRNAs, known or novel, that have never been associated with PCa and therefore empower further studies on the mechanisms leading to the development of PCa as well its more aggressive subtypes and aids in the future development of potential biomarkers and drug targets.

## Case Study 2: Landscape of CDK12-mutant primary tumors in PCa

### Methods

#### *Subtype annotation*

We obtained Mutation Annotation Files (MAF) for primary solid tumors in PCa from Genomic Data Commons (GDC) using TCGAbiolink package[122] and CNV level 4 data from firehose portal using RTCGAtoolbox package[123]. The MAF files were parsed to assign each sample to subtypes based on their mutation/loss status for common cancer drivers (i.e. *PTEN*, *ERG*, *SPOP* and *CDK12*) in PCa, only non-silent mutations were considered during subtype assignment. CNV data from firehose containing discrete indicators of gene copy number status, ranging from -2 (deep deletion) to 2 (amplification) were also used to assign subtypes. Samples with loss of driver gene were considered mutated and therefore assigned to its respective subtype. Finally, with exception of *PTEN* + *ERG* concomitant occurrence which were considered a subtype, only samples with exclusive mutation/loss for a driver were assigned to its subtype. Samples containing co-occurring

mutations where removed and samples without any of the mentioned drivers mutated/lost were considered "wild" primary tumor.

***Identification of recurrent lncRNAs CNV in PRAD***

Recurrent CNV in each subtype were identified using CNV SNP6 level 3 data for PCa primary solid tumor samples in the legacy database, this data contains pre-processed segmented data from Affymetrix Genome-Wide Human SNP Array 6.0 from GDC. Briefly, segment mean values ($log2\,{}^{Copy\ Number}/_{2}$) representing the extent of copy number changes were used to identify regions of amplifications/deletions. Segmented regions with segment mean ≤ -0.3 were considered deletions, where segments with segment mean ≥ 0.3 were considered amplifications, this data was used to generate a matrix including all needed information about the observed aberrant regions. This matrix was used as input to GAIA[124], a conservative permutation test allowing the estimation of the probability distribution of the contemporary mutations expected for non-driver markers. Genomic regions identified as significantly altered in copy number (corrected p-value ≤ 0.0001) were then annotated using FANTOM-CAT annotations from the FANTOM consortium to report amplified and deleted genes potentially related with *CDK12* mutation.

***Transcriptomic analysis***

Expression data were obtained from FC-R2 gene expression atlas for PCa primary tumor samples. Next, we pre-processed the data by filtering genes with low expression and normalized it with TMM method. Using generalized linear models approach coupled with empirical Bayes moderation of standard errors[54] and voom precision weigths[49], we performed differential

975 expression analysis of each subtype in contrast with wild samples. Genes with FDR ≤ 0.01 and
976 logFC ≥ 1 were considered differentially expressed.
977
978 ***Gene set enrichment analysis***
979
980 The lists of differentially expressed genes were tested for enrichment with GSEA. GSEA was
981 performed using fast geneset enrichment analysis implemented in the fgsea[102] package from
982 Bioconductor with 100000 permutations. We performed the enrichment test on the hallmarks
983 collection from Broad Institute MSigDB database[103]. Gene sets with less than 15 and more than
984 500 genes were removed from the analysis. The lists of differentially expressed genes were ranked
985 by t-statistics and were used as input together with the gene sets. Gene sets with FDR ≤ 0.05 were
986 considered significant.
987
988 ***Differential methylation analysis***
989
990 To detect differentially methylated regions (DMRs) we obtained level 3 methylation data
991 presented in the form of beta-values that uses a scale ranging from 0.0 (probes completely
992 unmethylated) up to 1.0 (probes completely methylated) from GDC using TCGAbiolinks. Next,
993 we tested for differential methylation between the groups using Wilcoxon test and adjusting by the
994 Benjamin-Hochberg method. A minimum mean difference of 0.2 and adjusted p-value of less than
995 0.01 was required to be considered significant. Significant DMRs were annotated by assigning
996 then to the closest transcription start site (TSS) using FANTOM-CAT annotations.
997
998 **Results**
999

***Annotation***

Samples in TCGA-PRAD cohort were assigned to subtypes based on their mutation status for the selected driver genes. After parsing the MAF files and the CNV data we were able to assign 471 out of 505 samples to a unique subtype. Considering only uniquely mutated samples, *CDK12* mutation/loss (herein referred just as mutation) showed a relatively low prevalence (2.4%) with 12 samples while the remaining driver genes showed prevalence between 9% (*SPOP*, 46 samples) to 15% (*PTEN* and *PTEN+ERG*, 73 samples each). Most of the samples showed no mutation in the selected driver genes, with 212 samples considered WILD primary tumors.

***CDK12 subtype shows large aberrations***

Recurrent amplifications and deletions were identified for all defined subtypes (Table 6). The number of deleted regions ranged from 3 (CDK12) to 78 (WILD) and from 0 (SPOP) to 123 (WILD) amplified regions. The amount of significant aberrations was directly related to the sample size, with CDK12 having the least number of samples (12) we opted to adopt a relaxed adjusted p-value of 0.01 for this subtype. Interestingly, the power of the analysis did not directly translate to the number of genes affected by the aberrations. Although having fewer significant aberrations, CDK12 showed significantly larger aberrations (Figure 15) and the greatest average number of amplified (61) and deleted (484) genes per aberration among the subtypes. ETS subtype followed CDK12 subtype in aberration size and average number of deleted genes (156). Overall PTEN, PTEN+ERG and SPOP subtypes showed similar numbers of aberrations and gene affected.

**Table 6 - Summary of recurrent aberrations.**

| Gene | CDK12 | ETS | PTEN | PTEN+ERG | SPOP | WILD |
|---|---|---|---|---|---|---|
| Deleted regions | 3 | 14 | 49 | 33 | 35 | 78 |
| Amplified Regions | 1 | 0 | 2 | 2 | 0 | 123 |
| Number of deleted genes | 1451 | 2184 | 4467 | 4029 | 4149 | 6074 |
| Number of amplified genes | 61 | 0 | 26 | 29 | 0 | 1988 |
| Chromosomes with deletion | 8 | 12, 21, 17, 8, 3, 16 | 10, 17, 13, 12, 16, 6, 8, 3, 22, 15 | 10, 17, 13, 12, 16, 21, 8, 19, 5, 11, 3, 1 | 13, 2, 6, 8, 5 | 13, 16, 17, 2, 6, 8, 3, 12, 19, 18 |
| Chromosomes with amplification | 23 | | 14, 4 | 1 | | 8, 11 |

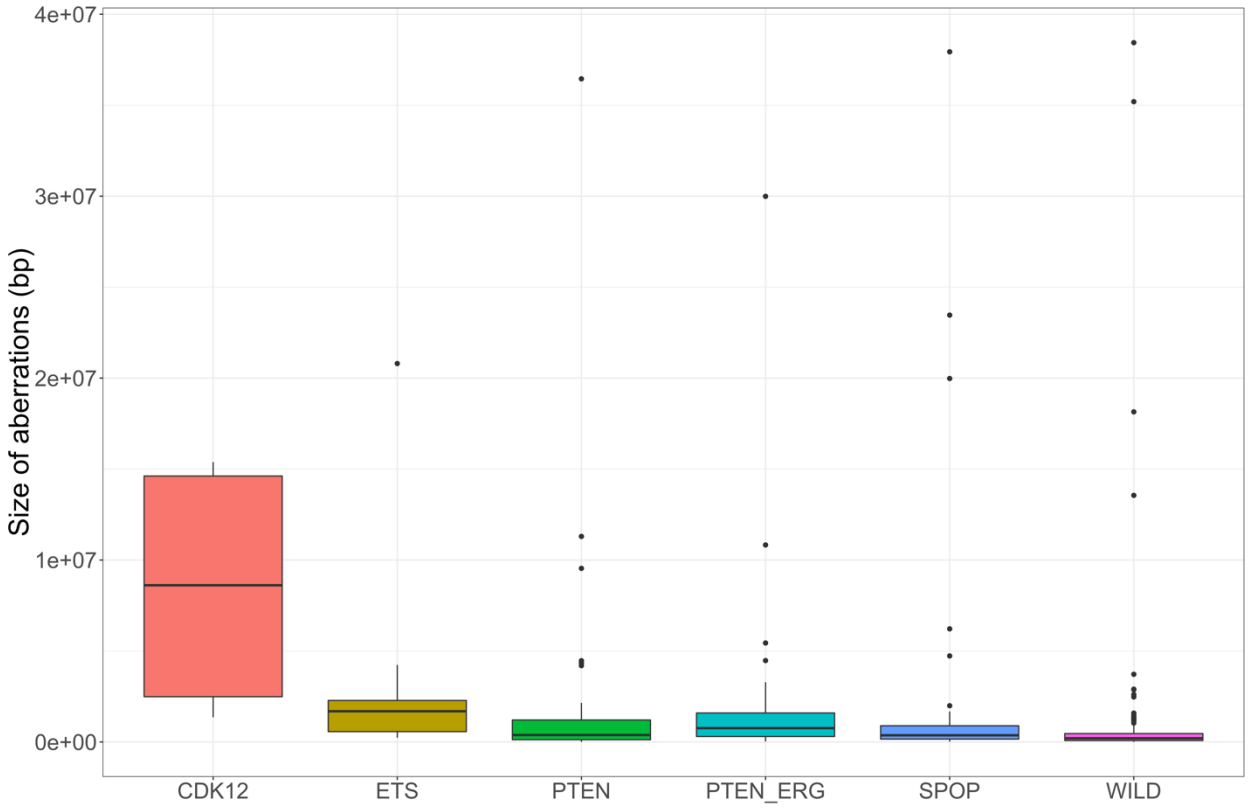



**Figure 15 – Aberration sizes distributions.** Boxplots shows size distributions of the recurrent aberrations found. Despite all subtypes presenting a few large aberration, CDK12 aberrations are significantly larger.

***CDK12 mutation mostly up-regulate genes***

In order to profile changes in genes expression for each subtype we obtained expression matrix from the FC-R2 atlas for PRAD samples. This matrix initially contained expression levels for 109,873 genes, which were reduce to 42,024 after filtering for lowly expressed genes (counts < 5). After filtering, counts were adjusted for RNA composition with TMM normalization and standard DE analysis were carried with limma/voom using the WILD subtype as control. We were able to detect DE genes across all subtypes, ranging from 183 to 5003 up-regulated genes and 22 to 4884 down-regulated genes (Table 7). Interestingly, when observing the ratio of up- and down-regulated genes for each subtype, CDK12 was the only subtype which presented a ratio larger than 1. In fact, CDK12 subtype showed only few genes down- regulated (22) in contrast with up-regulated genes (183) with a ratio of up-/down-regulated genes of 8.3, while the ratios for the other subtypes remained below 2 (Table 7).

**Table 7 - Summary of DGE analysis.**

| Gene | Up-Regulated Genes | Down-Regulated Genes | CAT genes Up-regulated | CAT genes Dn-regulated | Up/Down Ratio |
|---|---|---|---|---|---|
| CDK12 | 183 | 22 | 77 | 13 | 8.3 |
| ETS | 1221 | 768 | 188 | 358 | 1.5 |
| SPOP | 2279 | 1180 | 332 | 734 | 1.93 |
| PTEN | 2521 | 2260 | 131 | 1218 | 1.12 |
| PTEN+ERG | 5003 | 4884 | 484 | 2439 | 1.02 |

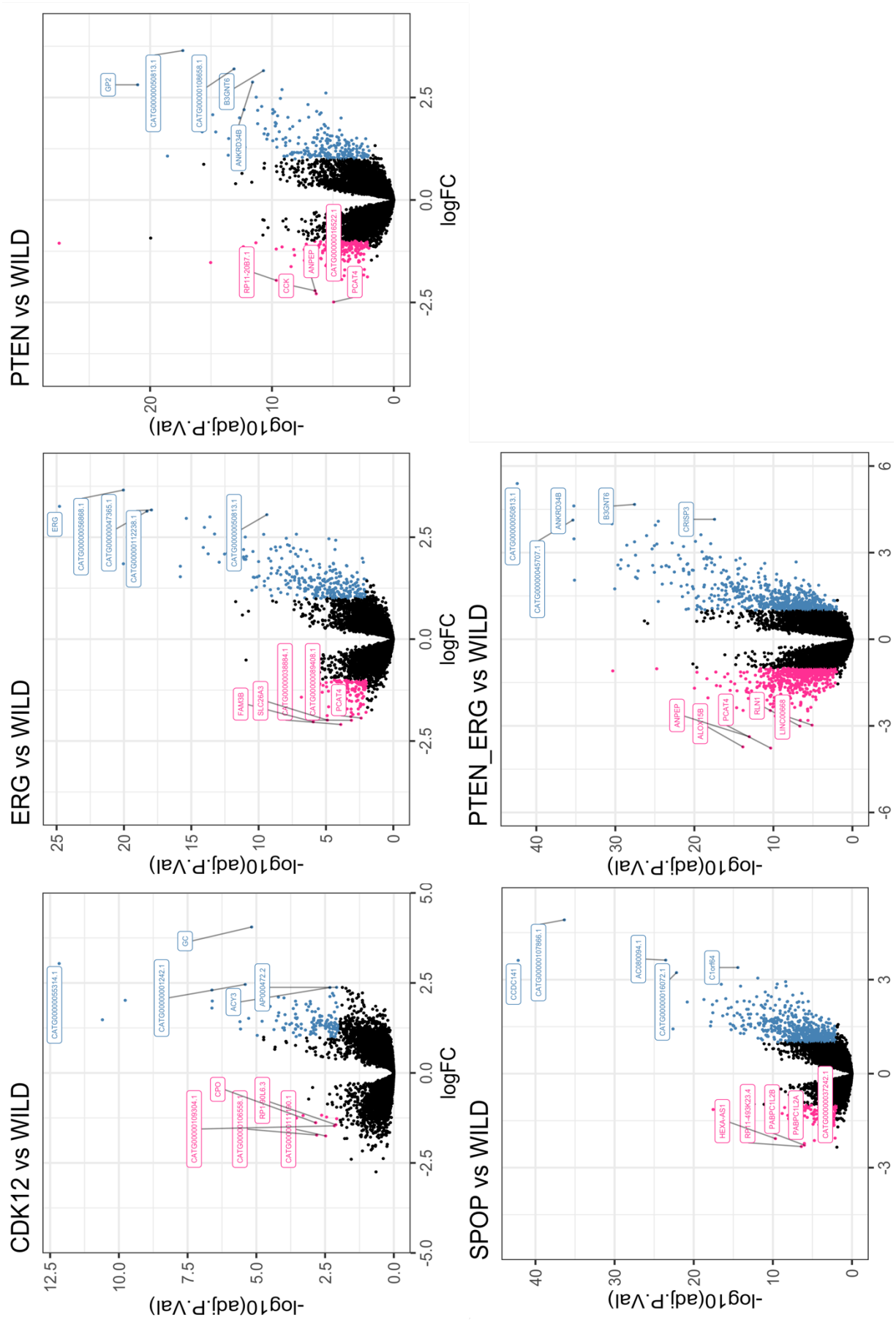



**Figure 16 – Differentially expressed genes across subtypes in contrast with wild subtype.** Volcano plots shows the impact of each mutation at transcriptional level. Dots in pink represents genes down-regulated with logFC ≥ 1 and adjusted p-value ≤ 0.01, blue dots represents genes up-regulated with logFC ≥ 1 and adjusted p-value ≤ 0.01. Labeled dots shows the top 5 up- and down-regulated genes in each contrast.

### *CDK12 mutation presents increased xenobiotic metabolism*

Enrichment analysis of hallmarks of cancer gene sets presented only a few gene sets enriched (Figure 17). Nevertheless, it showed that CDK12-mutated samples present increased activity of cell cycle and DNA metabolism. Interestingly, CDK12 was the only subtype to present increased xenobiotic metabolism among all other subtypes. Among decreased activity, the only gene set enriched was IL6-JAK-STAT3 signaling pathway.

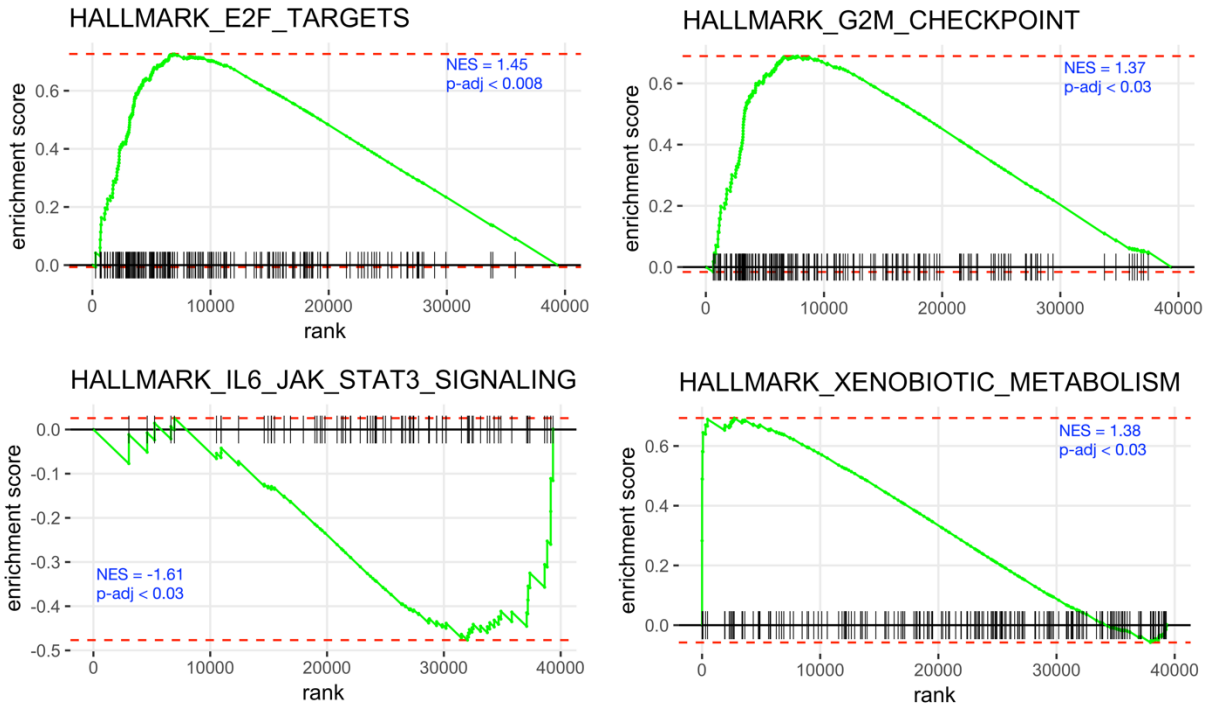



**Figure 17 – Gene sets enriched upon CDK12 mutation.** CDK12 mutated samples shows increased activity of cell cycle and DNA metabolism genes, as well increased xenobiotic metabolism which might suggest the susceptibility of this subtypes to certain drugs.

### *Methylation levels in CDK12 is unchanged*

We also investigate methylation profiles of each subtype. During our analysis we did not found DMRs in CDK12 subtype when compared to WILD subtype, while other subtypes showed several

DMRs in contrast with WILD subtype. We hypothesize that either CDK12 mutation does not impact methylation levels in PRAD, or that we did not reach enough statistical power due to low number of samples for this subtype. In order to assess which hypothesis is true, we repeated the analysis for the PTEN subtype randomly sampling the same number of samples (12) available for the CDK12 subtype. Since we were able to detect DMRs in PTEN subtype with reduced number of samples, we therefore conclude that CDK12 mutation does not impact methylation levels at a significant level (Figure 19). Also, by computing mean methylation levels of the probes, we observed that there is no global methylation changes across subtypes (Figure 18) despite subtypes presenting DMRs, this suggests that PRAD presents a fine control of methylated regions.

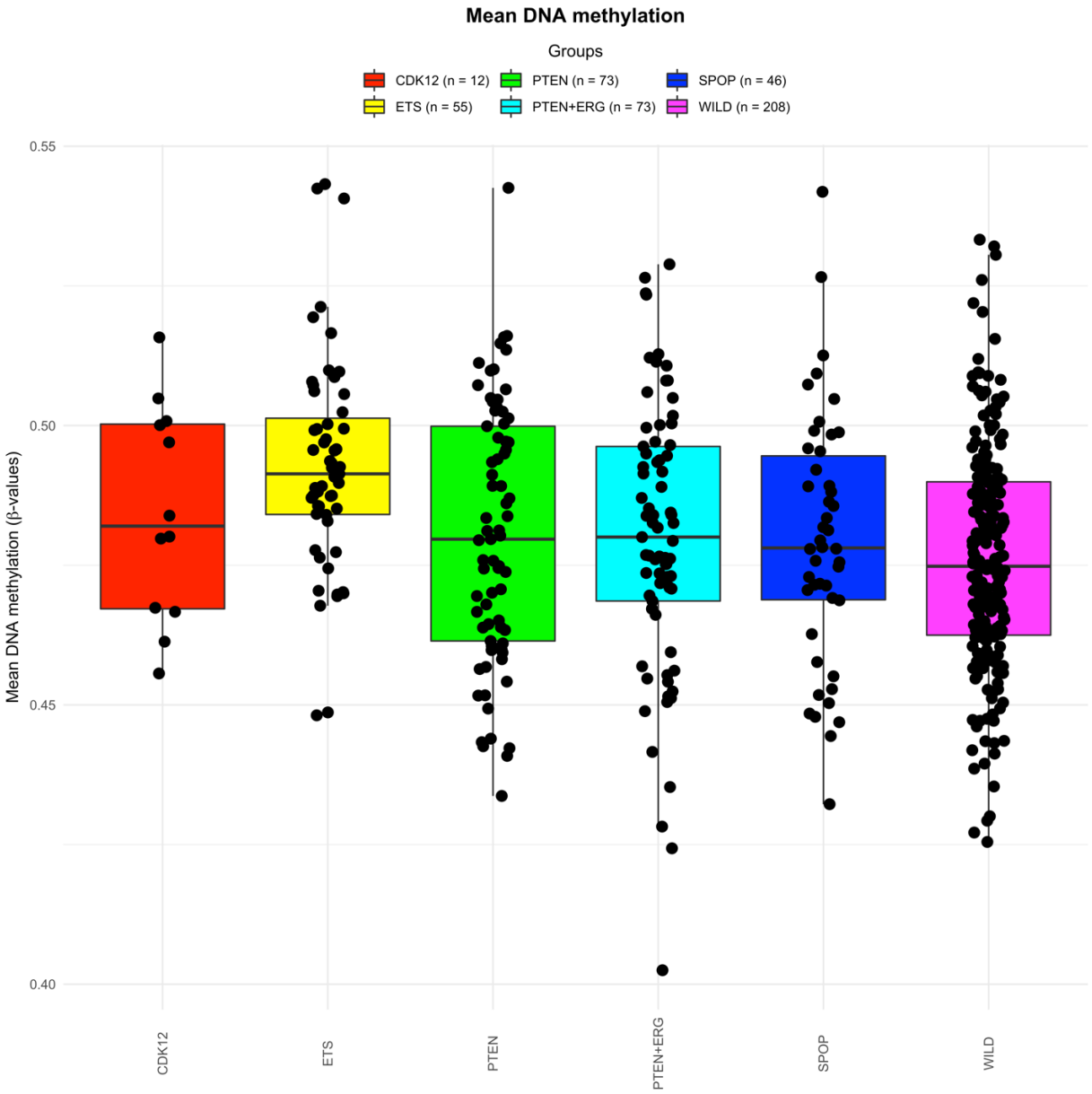



**Figure 18 – Global methylation levels across subtypes.** Boxplots shows mean methylation level for each subtype. As shown, there are no global shift in methylation levels for any subtype suggesting that methylation is controlled at a fine level.

1069

1070

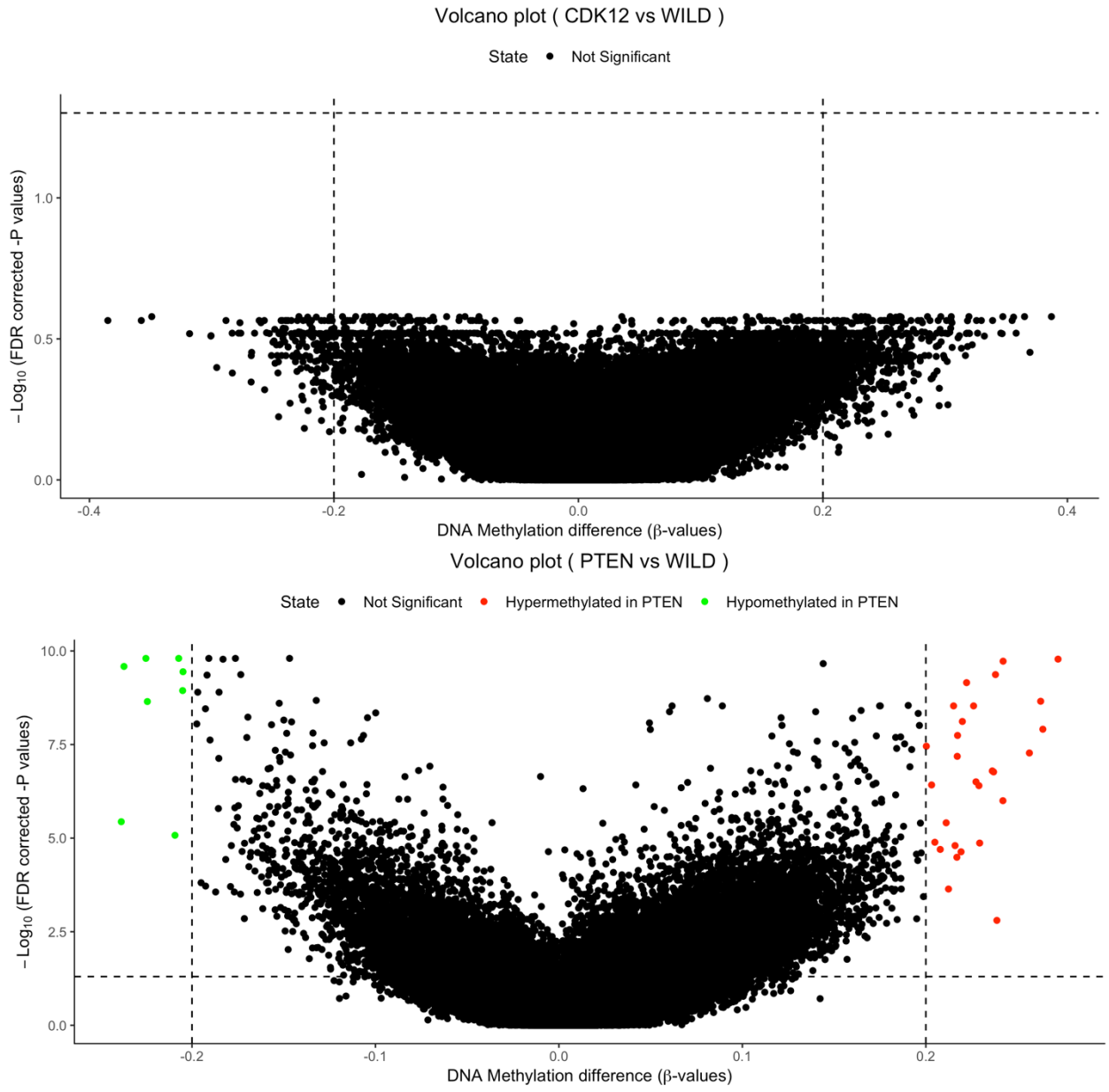



**Figure 19 – Differential methylation in mutants.** On top: CDK12 subtype shows no significant methylated regions when compared to WILD subtype. On bottom: PTEN subtype shows hypomethylated and hypermethylated regions.

## Discussion

The advent of immune checkpoint blockade therapies that use programmed death 1 (PD-1) or programmed death ligand 1 (PD-L1) inhibitors for the treatment of multiple cancer types represent

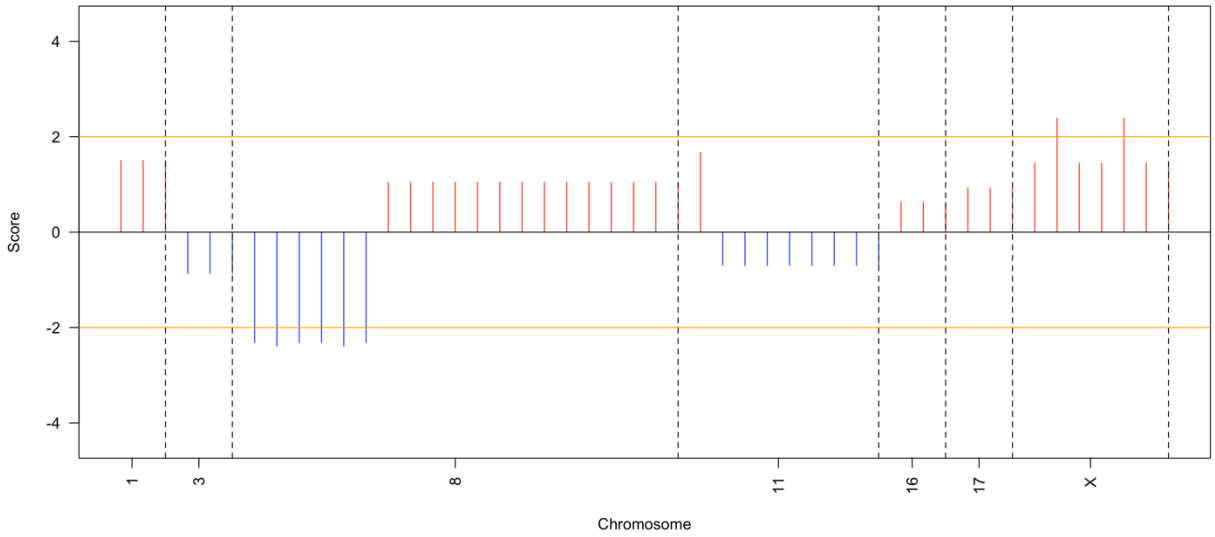
a major step in the care of patients with cancer. Recently, it was discovered that certain genetic subtypes of cancers may be remarkably sensitive to PD-1 inhibitor therapies. This discovery is a clear example of precision oncology, where the tumor genomic status can be used to guide interventions[125].

Following the development of immune checkpoint blockade therapies, it was discovered that cancers with mutations in genes belonging to the homologous recombination DNA damage repair pathway, especially cancers of the breast and ovary, have greater chance to respond well to treatment with poly-adenosine diphosphate ribose polymerase (PARP) inhibitors or platinum chemotherapies[125].

Wu and collaborators[126] have recently characterized a new subtype of metastatic castration-resistant prostate cancer (mCRPC) whose main characteristic is the biallelic inactivation of the tumor suppressor CDK12. This gene is thought to maintain DNA repair through the regulation of DNA damage-response genes (*BRCA1*, *FANCD2*, and *ATR*), and it had been suggested that the enzyme was associated with PARP inhibitor sensitivity when it was genetically inactivated[127]. Here we present the genomic, transcriptomic and epigenetic landscape of *CDK12*-mutant primary tumors in PRAD.

Copy number variations have a critical role in cancer development and progression. A chromosomal segment can be deleted or amplified as a result of genomic rearrangements, such as deletions, duplications, insertions and translocations. Using public data, we analyzed genomic data to discover recurrent aberration in the genome of the CDK12 subtype. Most interestingly, our results showed that despite having the least amount recurrent aberrations, CDK12 subtype presented the highest number of amplified genes among which where cell cycle and DNA replication related genes (e.g. *CTPS2*, *TXLNG*, *RBBP7*, *CA5B*, *GRPR*, etc.). As reported by Wu

and collaborators[126], one of the main characteristics of mCRPC CDK12 subtype is recurrent gains at loci involved in cell cycle and DNA replication. It is also noteworthy, that these genes were mainly located on chromosome X and it was the only subtype to present aberration in this chromosome (




Supplementary Figure 1). Apart from chromosome X, CDK12 subtype exhibited large deletions across chromosome 8, contrasting with results from Wu and collaborators were gain of 8q arm was observed in mCRPC[126].

Interestingly, the recurrent gains often observed CDK12 subtype were also observed at transcriptomic level. With a ratio of up- and down-regulated genes of 8.3, CDK12 subtype was by far the highest ratio (Table 7). In fact, it was the only subtype to present more than twice up-regulated genes than down-regulated ones. It is also noteworthy that CDK12 subtype presented few down-regulated genes, indicating that CDK12 profile at transcriptomic level might reflect events at genomic level, such as duplications in chromosome. Curiously, none of DE genes were present in recurrently amplified regions detected by GAIA at FDR of $1x10^{-4}$. However, if a relaxed FDR of 0.1 is adopted we start observing DE genes in aberration regions, suggesting that the small

number of samples in CDK12 subtype might be limiting the detection of recurrent aberrant regions. Furthermore, we observed that several lncRNAs recently annotated in the FANTOM-CAT meta-assembly were among the top differentially expressed genes in the CDK12 subtype (Figure 16 – Differentially expressed genes across subtypes in contrast with wild subtype. Supplementary Table 2), as well in others subtypes analyzed (Figure 16).

In our enrichment analysis we observed enriched gene sets that are on par with the literature. In hallmarks of cancer collection, we observed for gene sets significantly enriched: E2F targets, G2-M DNA damage checkpoint, IL6-JAK-STAT3 signaling and xenobiotic metabolism. E2F are a group of genes that encode a group of transcription factor in higher eukaryotes. All of them are involved in the cell cycle regulation and DNA synthesis in mammalian cells. Similarly, G2-M DNA damage checkpoint is an important cell cycle checkpoint in eukaryotic organisms. Both gensets were found positively enriched in CDK12 vs WILD subtype contrast. Wu and collaborators recently characterized the CDK12 subtype in mCRPC. Their findings showed that mCRPC CDK12 subtype exhibit focal tandem duplication (FTD) in regions containing cell cycle and DNA repair related genes and that these FTDs induces expression in a dosage-dependent and independent manner[126]. Strikingly, CDK12 subtype was the only subtype to present enrichment of genes involved xenobiotioc metabolism. This might explain a key clinical characteristic of CDK12 subtype which are susceptible to treatment with PARP inhibitors or platinum chemotherapies. Moreover, IL6-JAK-STAT3 signaling pathway was the only negatively enriched geneset in CDK12 subtype. This pathway communicates information from chemical signals outside of a cell to the cell nucleus and is involved in processes such as immunity, cell division, cell death and tumor formation. Also, this pathway can regulate other pathways such as PI3K/AKT/mTOR pathway. PI3K/AKT/mTOR pathway and MYC activation is often found in most types of cancer,

including PRAD. These pathways however, were not found enriched in CDK12 and SPOP subtypes, which are the only subtypes presenting negative enrichment of IL6-JAK-STAT3 signaling pathway, but are highly enriched in PTEN, ETS and PTEN + ETS subtypes which might indicate that increased proliferation of these subtypes are independent of this pathway.

At methylation level, we did not observe global methylation changes in the subtypes in contrast with the WILD subtype (Figure 18), suggesting that methylation levels in PCa are regulated at finer level. Interestingly, we did not observe significant methylation differences in CDK12 subtype, while all remaining subtypes present DMRs. This might suggest that genomic alterations are the main driver of CDK12 phenotype.

Finally, in this case study we have shown the complete landscape of CDK12 mutant subtype and all genomics events and transcriptional processes that are impacted in this subtype. We have shown results that are on par with the literature as well novel insights that can help further characterize this subtype. Given the concordance of our results from known genes, we can believe that the remaining genes whose function is yet unknown, are important players in this subtype. At both genomic and transcriptomic level, we showed that several lncRNAs uniquely annotated in FANTOM-CAT are shown to be differentially expressed and susceptible to genomic alterations. Given the susceptibility of CDK12 subtype to current treatments it is of paramount importance to fully characterize what makes this treatment susceptible in the first place. By uncovering players that so far have been overlook, we hope these results can advance our understanding of this subtype.

# Conclusion

In this work we present the FC-R2 expression atlas encompassing thousands of lncRNAs recently uncovered by the FANTOM consortium. We demonstrate that data from FC-R2 can robustly replicate several recent findings about lncRNA expression in humans, while uncovering hundreds of potential lncRNAs players in a variety of scenarios that would otherwise not be capture by other resources. Moreover, we presented two distinct case studies showing how our resource can be used to uncover lncRNAs that might play important roles in several phenotypes.

Finally, all results and data from the FANTOM-CAT/recount2 atlas are available as a public tool. With uniformly processed expression data for over 70,000 samples and 109,873 genes ready to analyze, we want to encourage researchers to dive deeper into the study of ncRNAs, their interaction with coding and non-coding genes and the influences on the phenotypes in normal and altered tissues. While uncovering the exact mechanisms and roles of these lncRNAs are beyond the scope of this work, the ongoing FANTOM 6 project is aiming in characterize these genes. We hope this new tool can help paving the way to develop new hypothesis that can be followed to unwind the biological role of the RNAs as a whole.

# References


1. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
2. Human Genome Sequencing Consortium, I. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931–945 (2004).
3. Venter, J. C. *et al.* The Sequence of the Human Genome. *Science (80-. ).* **291**, 1304–1351 (2001).
4. Willyard, C. New human gene tally reignites debate. *Nature* **558**, 354–355 (2018).
5. Sim, G. K. *et al.* Use of a cDNA library for studies on evolution and developmental expression of the chorion multigene families. *Cell* **18**, 1303–16 (1979).

6. Putney, S. D., Herlihy, W. C. & Schimmel, P. A new troponin T and cDNA clones for 13 different muscle proteins, found by shotgun sequencing. *Nature* **302**, 718–721 (1983).
7. Adams, M. D. *et al.* Complementary DNA sequencing: expressed sequence tags and human genome project. *Science* **252**, 1651–6 (1991).
8. Schena, M., Shalon, D., Davis, R. W. & Brown, P. O. Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray. *Science (80-. ).* **270**, 467–470 (1995).
9. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10**, 57–63 (2009).
10. Martin, J. a & Wang, Z. Next-generation transcriptome assembly. *Nat. Rev. Genet.* **12**, 671–682 (2011).
11. Pereira, M. A., Imada, E. L. & Guedes, R. L. M. RNA-seq: Applications and Best Practices. in *Applications of RNA-Seq and Omics Strategies - From Microorganisms to Human Health* **7**, 43936 (InTech, 2017).
12. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12**, 357–360 (2015).
13. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
14. Nellore, A. *et al.* Rail-RNA: scalable analysis of RNA-seq splicing and coverage. *Bioinformatics* btw575 (2016). doi:10.1093/bioinformatics/btw575
15. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–7 (2016).
16. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast

and bias-aware quantification of transcript expression. *Nat. Methods* **14**, 417–419 (2017).

17. Pennisi, E. ENCODE Project Writes Eulogy for Junk DNA. *Science (80-. ).* **337**, 1159–1161 (2012).

18. Mattick, J. S. Challenging the dogma: the hidden layer of non-protein-coding RNAs in complex organisms. *Bioessays* **25**, 930–9 (2003).

19. Mattick, J. S. RNA regulation: a new genetics? *Nat. Rev. Genet.* **5**, 316–323 (2004).

20. Mattick, J. S. Non-coding RNAs: the architects of eukaryotic complexity. *EMBO Rep.* **2**, 986–91 (2001).

21. Eddy, S. R. Non-Coding Rna Genes and the Modern Rna World. *Nat. Rev. Genet.* **2**, 919–929 (2001).

22. Shiraki, T. *et al.* Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc. Natl. Acad. Sci.* **100**, 15776–15781 (2003).

23. de Hoon, M. *et al.* Paradigm shifts in genomics through the FANTOM projects. *Mamm. Genome* **26**, 391–402 (2015).

24. Karapetyan, A. R., Buiting, C., Kuiper, R. a. & Coolen, M. W. Regulatory roles for long ncRNA and mRNA. *Cancers (Basel).* **5**, 462–490 (2013).

25. Freyhult, E. K., Bollback, J. P. & Gardner, P. P. Exploring genomic dark matter: A critical assessment of the performance of homology search methods on noncoding RNA. *Genome Res.* **17**, 117–125 (2006).

26. Anastasiadou, E., Jacob, L. S. & Slack, F. J. Non-coding RNA networks in cancer. *Nat. Rev. Cancer* **18**, 5–18 (2017).

27. Brown, C. J. *et al.* The human XIST gene: Analysis of a 17 kb inactive X-specific RNA

that contains conserved repeats and is highly localized within the nucleus. *Cell* **71**, 527–542 (1992).

28. Quinn, J. J. & Chang, H. Y. Unique features of long non-coding RNA biogenesis and function. *Nat. Rev. Genet.* **17**, 47–62 (2016).

29. Hon, C.-C. *et al.* An atlas of human long non-coding RNAs with accurate 5′ ends. *Nature* **543**, 199–204 (2017).

30. Iyer, M. K. *et al.* The landscape of long noncoding RNAs in the human transcriptome. *Nat. Genet.* **47**, 199–208 (2015).

31. Ling, H. *et al.* Junk DNA and the long non-coding RNA twist in cancer genetics. *Oncogene* **34**, 5003–5011 (2015).

32. Qi, M. *et al.* Analysis of Long Non-Coding RNA Expression of Lymphatic Endothelial Cells in Response to Type 2 Diabetes. *Cell. Physiol. Biochem.* **41**, 466–474 (2017).

33. Kung, J. T. Y., Colognori, D. & Lee, J. T. Long Noncoding RNAs: Past, Present, and Future. *Genetics* **193**, 651–669 (2013).

34. Frankish, A. *et al.* GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.* **47**, D766–D773 (2019).

35. Kawaji, H. *et al.* Comparison of CAGE and RNA-seq transcriptome profiling using clonally amplified and single-molecule next-generation sequencing. *Genome Res.* **24**, 708–717 (2014).

36. Kodama, Y., Shumway, M. & Leinonen, R. The sequence read archive: explosive growth of sequencing data. *Nucleic Acids Res.* **40**, D54–D56 (2012).

37. Edgar, R., Domrachev, M. & Lash, A. E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* **30**, 207–210

(2002).

38. Collado-Torres, L. *et al.* Reproducible RNA-seq analysis using recount2. *Nat. Biotechnol.* **35**, 319–321 (2017).

39. McCrea, E. M., Lee, D. K., Sissung, T. M. & Figg, W. D. Precision medicine applications in prostate cancer. *Ther. Adv. Med. Oncol.* **10**, 1–13 (2018).

40. Vargas, A. J. & Harris, C. C. Biomarker development in the precision medicine era: Lung cancer as a case study. *Nat. Rev. Cancer* **16**, 525–537 (2016).

41. Ashley, E. A. Towards precision medicine. *Nat. Rev. Genet.* **17**, 507–522 (2016).

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

43. Council, N. R. *Toward precision medicine: building a knowledge network for biomedical research and a new taxonomy of disease*. (2011).

44. Mansfield, E. A. FDA Perspective on Companion Diagnostics: An Evolving Paradigm. *Clin. Cancer Res.* **20**, 1453–1457 (2014).

45. Administration, U. F. and D. In vitro companion diagnostic devices: Guidance for industry and Food and Drug Administration staff. *Cent. Drug Eval. Res.*

46. Risso, D., Ngai, J., Speed, T. P. & Dudoit, S. Normalization of RNA-seq data using factor analysis of control genes or samples. *Nat. Biotechnol.* **32**, 896–902 (2014).

47. Hansen, K. D., Irizarry, R. A. & WU, Z. Removing technical variability in RNA-seq data using conditional quantile normalization. *Biostatistics* **13**, 204–216 (2012).

48. Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* **11**, R25 (2010).

49. Law, C. W., Chen, Y., Shi, W. & Smyth, G. K. voom: precision weights unlock linear

model analysis tools for RNA-seq read counts. *Genome Biol.* **15**, R29 (2014).

50. Lawrence, M. *et al.* Software for computing and annotating genomic ranges. *PLoS Comput. Biol.* **9**, e1003118 (2013).

51. Collado-Torres, L., Nellore, A. & Jaffe, A. E. recount workflow: Accessing over 70,000 human RNA-seq samples with Bioconductor. *F1000Research* **6**, 1558 (2017).

52. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc.* **Series B,** 289–300 (1995).

53. Chen, H. *et al.* A Pan-Cancer Analysis of Enhancer Expression in Nearly 9000 Patient Samples. *Cell* **173**, 386–399.e12 (2018).

54. Smyth, G. K. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.* **3**, Article3 (2004).

55. Leek, J. T. & Storey, J. D. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet.* **3**, 1724–1735 (2007).

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

57. Zhou, M. *et al.* LncRNA-Hh strengthen cancer stem cells generation in twist-positive breast cancer via activation of hedgehog signaling pathway. *Stem Cells* **34**, 55–66 (2016).

58. Zhao, J. *et al.* Long non-coding RNA Linc00152 is involved in cell cycle arrest, apoptosis, epithelial to mesenchymal transition, cell migration and invasion in gastric cancer. *Cell cycle* **14**, 3112–3123 (2015).

59. Shang, D., Zheng, T., Zhang, J., Tian, Y. & Liu, Y. Profiling of mRNA and long non-coding RNA of urothelial cancer in recipients after renal transplantation. *Tumor Biol.* **37**, 12673–12684 (2016).

60. Cao, W.-J., Wu, H.-L., He, B.-S., Zhang, Y.-S. & Zhang, Z.-Y. Analysis of long non-coding RNA expression profiles in gastric cancer. *World J. Gastroenterol. WJG* **19**, 3658 (2013).

61. Lu, Y., Meng, X., Wang, L. & Wang, X. Analysis of long non-coding RNA expression profiles identifies functional lncRNAs associated with the progression of acute coronary syndromes. *Exp. Ther. Med.* **15**, 1376–1384 (2018).

62. Humphries, C. E. *et al.* Integrated whole transcriptome and DNA methylation analysis identifies gene networks specific to late-onset Alzheimer's disease. *J. Alzheimer's Dis.* **44**, 977–987 (2015).

63. Emmrich, S. *et al.* LincRNAs MONC and MIR100HG act as oncogenes in acute megakaryoblastic leukemia. *Mol. Cancer* **13**, 1 (2014).

64. Ni, S., Zhao, X. & Ouyang, L. Long non-coding RNA expression profile in vulvar squamous cell carcinoma and its clinical significance. *Oncol. Rep.* **36**, 2571–2578 (2016).

65. Gökmen-Polar, Y. *et al.* Abstract P2-06-05: LINC00478: A novel tumor suppressor in breast cancer. (2016).

66. Sun, D. *et al.* Regulation of several androgen-induced genes through the repression of the miR-99a/let-7c/miR-125b-2 miRNA cluster in prostate cancer cells. *Oncogene* **33**, 1448 (2014).

67. Li, S. *et al.* Exploring functions of long noncoding RNAs across multiple cancers through co-expression network. *Sci. Rep.* **7**, 754 (2017).

68. Zhao, W., Luo, J. & Jiao, S. Comprehensive characterization of cancer subtype associated long non-coding RNAs and their clinical implications. *Sci. Rep.* **4**, 6591 (2014).

69. Ma, Z. *et al.* Long non-coding RNA SNHG17 is an unfavourable prognostic factor and

promotes cell proliferation by epigenetically silencing P57 in colorectal cancer. *Mol. Biosyst.* **13**, 2350–2361 (2017).

70. Zhu, S. *et al.* Genome-scale deletion screening of human long non-coding RNAs using a paired-guide RNA CRISPR--Cas9 library. *Nat. Biotechnol.* **34**, 1279 (2016).

71. Lin, D. W. *et al.* Genetic variants in the LEPR, CRY1, RNASEL, IL4, and ARVCF genes are prognostic markers of prostate cancer-specific mortality. *Cancer Epidemiol Biomarkers Prev* **20**, 1928–1936 (2011).

72. Yu, J. J. *et al.* An integrated network of androgen receptor, polycomb, and TMPRSS2-ERG gene fusions in prostate cancer progression. *Cancer Cell* **17**, 443–454 (2010).

73. Bussemakers, M. J. *et al.* DD3: a new prostate-specific gene, highly overexpressed in prostate cancer. *Cancer Res.* **59**, 5975–5979 (1999).

74. de Kok, J. B. *et al.* DD3(PCA3), a very sensitive and specific marker to detect prostate tumors. *Cancer Res* **62**, 2695–2698 (2002).

75. Prensner, J. R. *et al.* Transcriptome sequencing across a prostate cancer cohort identifies PCAT-1, an unannotated lincRNA implicated in disease progression. *Nat. Biotechnol.* **29**, 742–749 (2011).

76. Ren, S. *et al.* Long noncoding RNA MALAT-1 is a new potential therapeutic target for castration resistant prostate cancer. *J. Urol.* **190**, 2278–2287 (2013).

77. Gutschner, T. & Diederichs, S. The hallmarks of cancer: a long non-coding RNA point of view. *RNA Biol* **9**, 703–719 (2012).

78. Kotake, Y. *et al.* Long non-coding RNA ANRIL is required for the PRC2 recruitment to and silencing of p15(INK4B) tumor suppressor gene. *Oncogene* **30**, 1956–1962 (2011).

79. Huang, X. *et al.* Exosomal miR-1290 and miR-375 as prognostic markers in castration-

resistant prostate cancer. *Eur. Urol.* **67**, 33–41 (2015).

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

81. Sur, I. & Taipale, J. The role of enhancers in cancer. *Nat. Rev. Cancer* **16**, 483 (2016).

82. Wang, G., Zhao, D., Spring, D. J. & Depinho, R. A. Genetics and biology of prostate cancer. *Genes Dev.* **32**, 1105–1140 (2018).

83. American Cancer Society. What is prostate cancer ? *Am. Cancer Soc.* **192**, 1,2 (2016).

84. Gleason, D. F. & Mellinger, G. T. Prediction of prognosis for prostatic adenocarcinoma by combined histological grading and clinical staging. *J. Urol.* **167**, 953–958 (2002).

85. Montironi, R. *et al.* Original Gleason System Versus 2005 ISUP Modified Gleason System: The Importance of Indicating Which System Is Used in the Patient's Pathology and Clinical Reports. *Eur. Urol.* **58**, 369–373 (2010).

86. Epstein, J. I. *et al.* The 2014 International Society of Urological Pathology (ISUP) Consensus Conference on Gleason Grading of Prostatic Carcinoma. *Am. J. Surg. Pathol.* 1 (2015). doi:10.1097/PAS.0000000000000530

87. Sizemore, G. M., Pitarresi, J. R., Balakrishnan, S. & Ostrowski, M. C. The ETS family of oncogenic transcription factors in solid tumours. *Nat. Rev. Cancer* **17**, 337–351 (2017).

88. Tomlins, S. A. *et al.* ETS Gene Fusions in Prostate Cancer: From Discovery to Daily Clinical Practice. *European Urology* (2009). doi:10.1016/j.eururo.2009.04.036

89. Wise, H. M., Hermida, M. A. & Leslie, N. R. Prostate cancer, PI3K, PTEN and prognosis. *Clin. Sci.* **131**, 197–210 (2017).

90. Squire, J. A. TMPRSS2-ERG and PTEN loss in prostate cancer. *Nat. Genet.* **41**, 509–510 (2009).

91. Boutros, P. C. *et al.* Spatial genomic heterogeneity within localized, multifocal prostate cancer. *Nat. Genet.* **47**, 736–745 (2015).

92. Geng, C. *et al.* Prostate cancer-associated mutations in speckle-type POZ protein (SPOP) regulate steroid receptor coactivator 3 protein turnover. *Proc. Natl. Acad. Sci.* **110**, 6997–7002 (2013).

93. Blattner, M. *et al.* SPOP Mutation Drives Prostate Tumorigenesis In Vivo through Coordinate Regulation of PI3K/mTOR and AR Signaling. *Cancer Cell* **31**, 436–451 (2017).

94. Zhang, P. *et al.* Intrinsic BET inhibitor resistance in SPOP-mutated prostate cancer is mediated by BET protein stabilization and AKT–mTORC1 activation. *Nat. Med.* **23**, 1055–1062 (2017).

95. Baylin, S. B. DNA methylation and gene silencing in cancer. *Nat. Clin. Pract. Oncol.* **2**, S4–S11 (2005).

96. Schübeler, D. Function and information content of DNA methylation. *Nature* **517**, 321–326 (2015).

97. Morey Kinney, S. R. *et al.* Opposing Roles of Dnmt1 in Early- and Late-Stage Murine Prostate Cancer. *Mol. Cell. Biol.* **30**, 4159–4174 (2010).

98. Hsu, C.-H. *et al.* TET1 Suppresses Cancer Invasion by Activating the Tissue Inhibitors of Metalloproteinases. *Cell Rep.* **2**, 568–579 (2012).

99. Armenia, J. *et al.* The long tail of oncogenic drivers in prostate cancer. *Nat. Genet.* **50**, 645–651 (2018).

100. Prensner, J. R. *et al.* The long noncoding RNA SChLAP1 promotes aggressive prostate cancer and antagonizes the SWI/SNF complex. *Nat. Genet.* **45**, 1392–1403 (2013).

101. Scharpf, R. B., Tjelmeland, H., Parmigiani, G. & Nobel, A. B. A Bayesian model for cross-study differential gene expression. *J. Am. Stat. Assoc.* **104**, 1295–1310 (2009).

102. Sergushichev, A. A. An algorithm for fast preranked gene set enrichment analysis using cumulative statistic calculation. *bioRxiv* 060012 (2016). doi:10.1101/060012

103. Liberzon, A. *et al.* Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**, 1739–1740 (2011).

104. Leinonen, K. A. *et al.* Loss of PTEN Is Associated with Aggressive Behavior in ERG-Positive Prostate Cancer. *Cancer Epidemiol. Biomarkers Prev.* **22**, 2333–2344 (2013).

105. Saal, L. H. *et al.* Poor prognosis in carcinoma is associated with a gene expression signature of aberrant PTEN tumor suppressor pathway activity. *Proc. Natl. Acad. Sci.* **104**, 7564–7569 (2007).

106. Han, B. *et al.* Fluorescence in situ hybridization study shows association of PTEN deletion with ERG rearrangement during prostate cancer progression. *Mod. Pathol.* **22**, 1083–1093 (2009).

107. Salameh, A. *et al.* PRUNE2 is a human prostate cancer suppressor regulated by the intronic long noncoding RNA *PCA3*. *Proc. Natl. Acad. Sci.* **112**, 8403–8408 (2015).

108. He, J. H. *et al.* Snail-activated long non-coding RNA PCA3 up-regulates PRKD3 expression by miR-1261 sponging, thereby promotes invasion and migration of prostate cancer cells. *Tumor Biol.* **37**, 16163–16176 (2016).

109. Lemos, A. E. G. *et al.* PCA3 long noncoding RNA modulates the expression of key cancer-related genes in LNCaP prostate cancer cells. *Tumor Biol.* **37**, 11339–11348 (2016).

110. Agarwal, R., D'Souza, T. & Morin, P. J. Claudin-3 and claudin-4 expression in ovarian

epithelial cells enhances invasion and is associated with increased matrix metalloproteinase-2 activity. *Cancer Res.* **65**, 7378–7385 (2005).

111. Srikantan, V. *et al.* PCGEM1, a prostate-specific gene, is overexpressed in prostate cancer. *Proc. Natl. Acad. Sci.* **97**, 12216–12221 (2000).

112. Petrovics, G. *et al.* Elevated expression of PCGEM1, a prostate-specific gene with cell growth-promoting function, is associated with high-risk prostate cancer patients. *Oncogene* **23**, 605–611 (2004).

113. Lin, X., Shang, X., Manorek, G. & Howell, S. B. Regulation of the Epithelial-Mesenchymal Transition by Claudin-3 and Claudin-4. *PLoS One* **8**, (2013).

114. Song, Y. X. *et al.* Non-coding RNAs participate in the regulatory network of CLDN4 via ceRNA mediated miRNA evasion. *Nat. Commun.* **8**, 1–16 (2017).

115. Ottman, R., Nguyen, C., Lorch, R. & Chakrabarti, R. MicroRNA expressions associated with progression of prostate cancer cells to antiandrogen therapy resistance. *Mol. Cancer* **13**, 1–21 (2014).

116. Hardwick, J. P. Cytochrome P450 omega hydroxylase (CYP4) function in fatty acid metabolism and metabolic diseases. *Biochem. Pharmacol.* **75**, 2263–2275 (2008).

117. Wculek, S. K. & Malanchi, I. Neutrophils support lung colonization of metastasis-initiating breast cancer cells. *Nature* **528**, 413–417 (2015).

118. Kamran, M. *et al.* Aurora kinase A regulates Survivin stability through targeting FBXL7 in gastric cancer drug resistance and prognosis. *Oncogenesis* **6**, (2017).

119. Coon, T. A., Glasser, J. R., Mallampalli, R. K. & Chen, B. B. Novel E3 ligase component FBXL7 ubiquitinates and degrades Aurora A, causing mitotic arrest. *Cell Cycle* **11**, 721–729 (2012).

120. Jamaspishvili, T. *et al.* Clinical implications of PTEN loss in prostate cancer. *Nat. Rev. Urol.* **15**, 222–234 (2018).

121. Patel, N. *et al.* Expression and functional role of orphan receptor GPR158 in prostate cancer growth and progression. *PLoS One* **10**, 1–30 (2015).

122. Colaprico, A. *et al.* TCGAbiolinks : an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res.* **44**, e71–e71 (2016).

123. Samur, M. K. RTCGAToolbox: A New Tool for Exporting TCGA Firehose Data. *PLoS One* **9**, e106397 (2014).

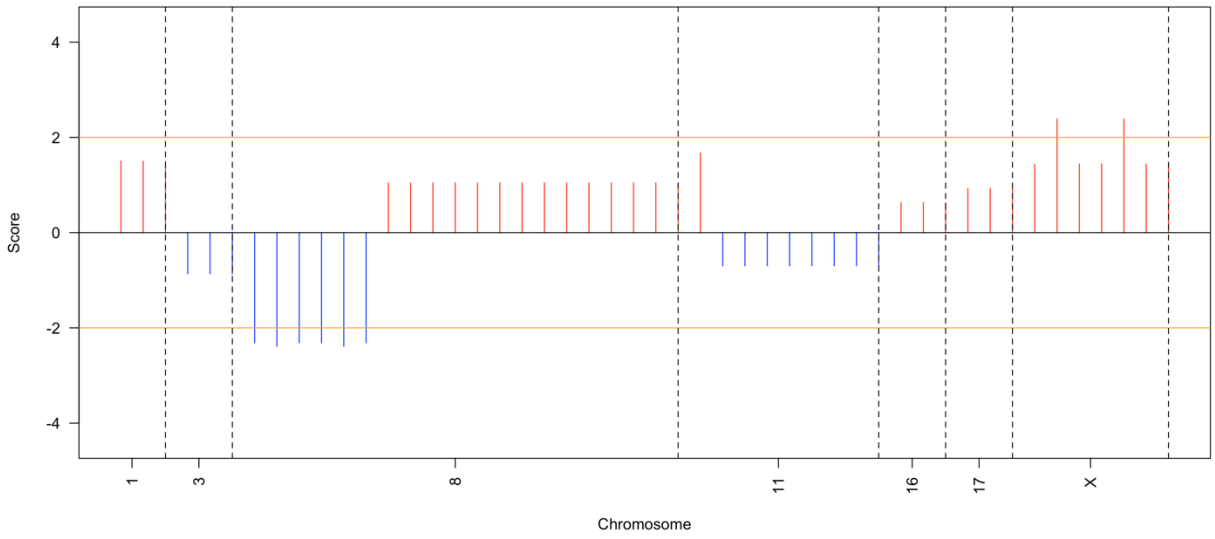
124. Morganella, S. GAIA: An R package for genomic analysis of significant chromosomal aberrations. *R Packag. version 2.26.0* (2018).

125. Antonarakis, E. S. Cyclin-Dependent Kinase 12, Immunity, and Prostate Cancer. *N. Engl. J. Med.* **379**, 1087–1089 (2018).

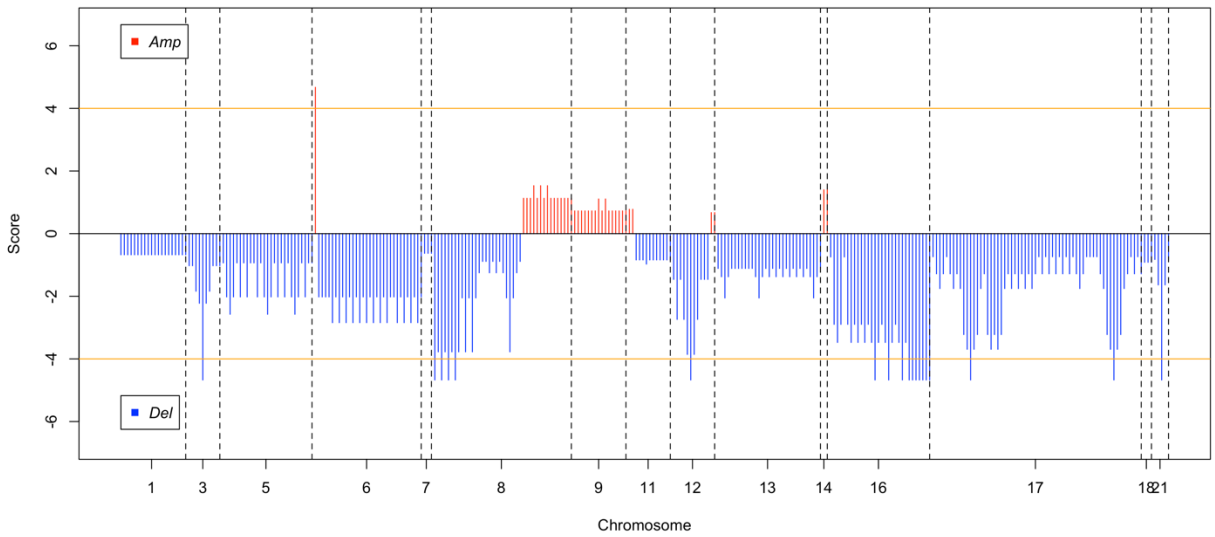
126. Wu, Y. M. *et al.* Inactivation of CDK12 Delineates a Distinct Immunogenic Class of Advanced Prostate Cancer. *Cell* **173**, 1770–1782.e14 (2018).

127. Bajrami, I. *et al.* Genome-wide Profiling of Genetic Synthetic Lethality Identifies CDK12 as a Novel Determinant of PARP1/2 Inhibitor Sensitivity. *Cancer Res.* **74**, 287–297 (2014).

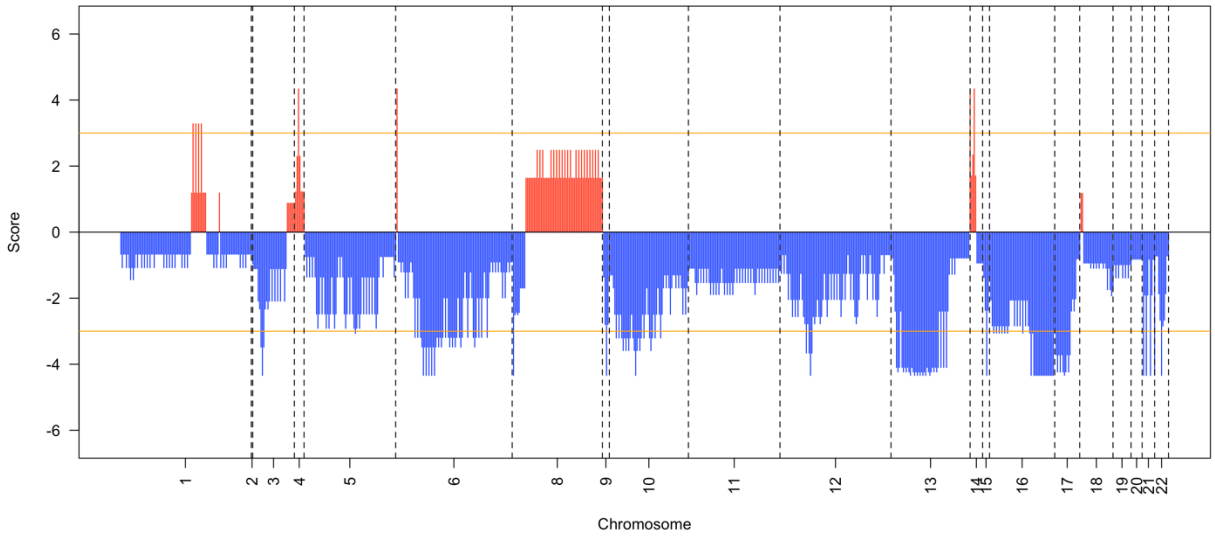
# ATTACHMENTS



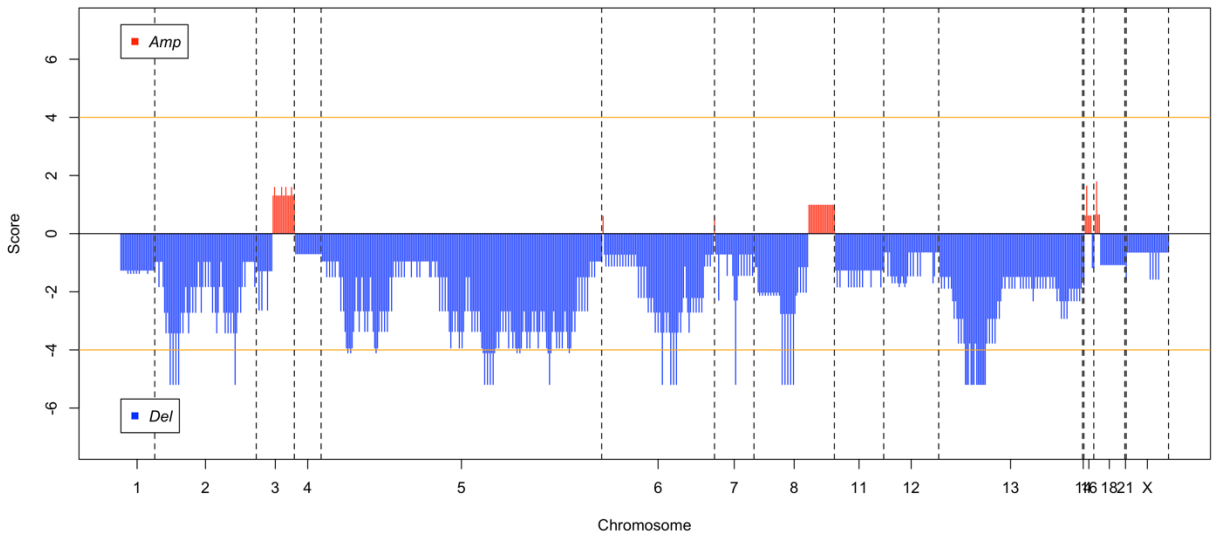

**Supplementary Figure 1 - GAIA plot for CDK12 subtype.** Figure shows regions subject to aberrations. Blue lines represent loss events while red lines gains. Orange line represents log10 of significance threshold.



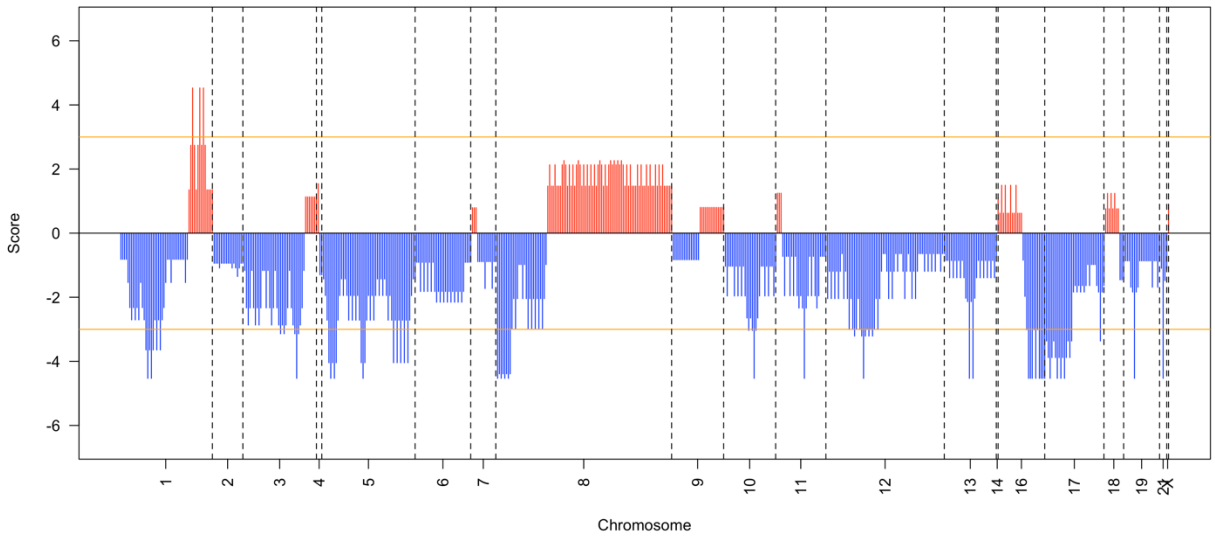

**Supplementary Figure 2 - GAIA plot for ETS subtype.** Figure shows regions subject to aberrations. Blue lines represent loss events while red lines gains. Orange line represents log10 of significance threshold.



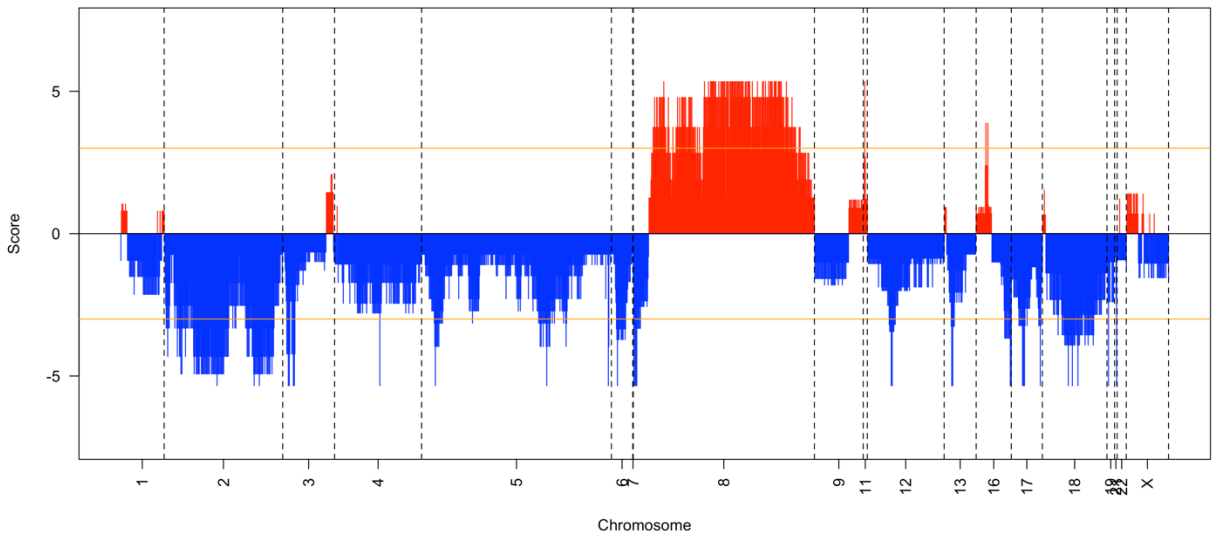

**Supplementary Figure 3 - GAIA plot for PTEN subtype.** Figure shows regions subject to aberrations. Blue lines represent loss events while red lines gains. Orange line represents log10 of significance threshold.




**Supplementary Figure 4 - GAIA plot for SPOP subtype.** Figure shows regions subject to aberrations. Blue lines represent loss events while red lines gains. Orange line represents log10 of significance threshold.




**Supplementary Figure 5 - GAIA plot for PTEN+ERG subtype.** Figure shows regions subject to aberrations. Blue lines represent loss events while red lines gains. Orange line represents log10 of significance threshold.




**Supplementary Figure 6 - GAIA plot for WILD subtype.** Figure shows regions subject to aberrations. Blue lines represent loss events while red lines gains. Orange line represents log10 of significance threshold.

**Supplementary Table 1. Survival analysis using Cox proportional regression showing the number of enhancers with prognostic power.**

| Tumor Tissue | Non-significant | FDR < 0.05 | Cases | Events | Median time |
|---|---|---|---|---|---|
| Kidney | 13,554 | 3,850 | 881 | 227 | NA |
| Uterus | 16,563 | 831 | 596 | 125 | 3,365 |
| Stomach | 16,850 | 554 | 392 | 158 | 940 |
| Liver | 16,970 | 369 | 365 | 130 | 1,694 |
| Bladder | 17,234 | 153 | 407 | 178 | 1,008 |
| Thyroid | 17,277 | 111 | 504 | 16 | NA |
| Breast | 17,305 | 96 | 1,080 | 151 | 3.941 |
| Colorectal | 17,292 | 87 | 602 | 128 | 2,532 |
| Lung | 17,335 | 69 | 998 | 395 | 1,531 |
| Prostate | 17,332 | 53 | 496 | 10 | NA |
| HeadNeck | 17,398 | 3 | 501 | 217 | 1,671 |
| Bile | 15,725 | 0 | 36 | 18 | 1,220 |
| Esophagus | 17,404 | 0 | 184 | 77 | 784 |

**Supplementary Table 2. List of differentially expressed genes in CDK12 vs WILD subtypes.**

| Gene ID | Gene Name | logFC | t | adj. p-value |
|---|---|---|---|---|
| ENSG00000145321 | GC | 4.05 | 6.10 | 6.74E-06 |
| CATG00000055314 | CATG00000055314.1 | 3.04 | 8.84 | 6.74E-13 |

| | | | | |
|---|---|---|---|---|
| CATG00000001242 | CATG00000001242.1 | 2.46 | 6.22 | 3.95E-06 |
| ENSG00000233215 | AP000472.2 | 2.38 | 4.16 | 8.32E-03 |
| ENSG00000132744 | ACY3 | 2.37 | 4.36 | 4.74E-03 |
| CATG00000096426 | CATG00000096426.1 | 2.30 | 6.80 | 2.43E-07 |
| ENSG00000121351 | IAPP | 2.21 | 4.28 | 5.90E-03 |
| ENSG00000241388 | HNF1A-AS1 | 2.15 | 4.25 | 6.54E-03 |
| ENSG00000185332 | TMEM105 | 2.12 | 4.36 | 4.74E-03 |
| ENSG00000255050 | RP11-661A12.9 | 2.11 | 5.58 | 6.34E-05 |
| CATG00000002737 | CATG00000002737.1 | 2.08 | 5.47 | 1.01E-04 |
| ENSG00000255236 | CTD-2655K5.1 | 2.05 | 6.43 | 1.74E-06 |
| ENSG00000260954 | LA16c-425C2.1 | 2.02 | 7.95 | 1.70E-10 |
| CATG00000004618 | CATG00000004618.1 | 2.02 | 5.98 | 1.17E-05 |
| CATG00000094347 | CATG00000094347.1 | 2.00 | 6.77 | 2.50E-07 |
| ENSG00000262877 | RP11-1055B8.4 | 1.99 | 5.21 | 2.66E-04 |
| ENSG00000144485 | HES6 | 1.99 | 4.87 | 9.10E-04 |
| ENSG00000176593 | CTD-2368P22.1 | 1.96 | 6.38 | 2.07E-06 |
| CATG00000094793 | CATG00000094793.1 | 1.95 | 4.18 | 7.82E-03 |
| ENSG00000131142 | CCL25 | 1.92 | 4.86 | 9.34E-04 |
| CATG00000071594 | CATG00000071594.1 | 1.91 | 4.97 | 6.40E-04 |
| CATG00000110168 | CATG00000110168.1 | 1.90 | 5.17 | 2.87E-04 |
| CATG00000061772 | CATG00000061772.1 | 1.88 | 5.80 | 2.66E-05 |
| ENSG00000216621 | RP11-244K5.6 | 1.85 | 5.75 | 3.18E-05 |
| CATG00000028402 | CATG00000028402.1 | 1.84 | 4.91 | 7.99E-04 |
| ENSG00000181656 | GPR88 | 1.83 | 4.14 | 8.76E-03 |
| ENSG00000225431 | AP001626.1 | 1.80 | 6.83 | 2.43E-07 |
| ENSG00000016490 | CLCA1 | 1.79 | 5.27 | 2.21E-04 |
| CATG00000083355 | CATG00000083355.1 | 1.78 | 4.65 | 1.80E-03 |
| ENSG00000172568 | FNDC9 | 1.74 | 4.94 | 7.11E-04 |
| ENSG00000111981 | ULBP1 | 1.71 | 4.82 | 1.07E-03 |
| ENSG00000162761 | LMX1A | 1.70 | 4.73 | 1.39E-03 |
| ENSG00000237194 | SNAI1P1 | 1.69 | 4.23 | 6.88E-03 |
| CATG00000038061 | CATG00000038061.1 | 1.67 | 5.57 | 6.54E-05 |
| ENSG00000205143 | ARID3C | 1.64 | 5.20 | 2.66E-04 |
| CATG00000102439 | CATG00000102439.1 | 1.62 | 5.44 | 1.10E-04 |
| ENSG00000254338 | RP11-909N17.3 | 1.60 | 4.39 | 4.37E-03 |
| ENSG00000175329 | ISX | 1.59 | 4.61 | 2.05E-03 |
| ENSG00000255346 | NOX5 | 1.59 | 4.34 | 5.06E-03 |
| ENSG00000118156 | ZNF541 | 1.58 | 5.59 | 6.22E-05 |
| CATG00000005468 | CATG00000005468.1 | 1.58 | 4.56 | 2.39E-03 |
| ENSG00000269915 | AP006621.9 | 1.57 | 4.61 | 2.05E-03 |

| | | | | |
|---|---|---|---|---|
| CATG00000073078 | CATG00000073078.1 | 1.54 | 4.31 | 5.44E-03 |
| CATG00000106348 | CATG00000106348.1 | 1.53 | 5.15 | 3.05E-04 |
| CATG00000026950 | CATG00000026950.1 | 1.53 | 5.20 | 2.66E-04 |
| CATG00000038150 | CATG00000038150.1 | 1.52 | 4.76 | 1.24E-03 |
| CATG00000041483 | CATG00000041483.1 | 1.52 | 6.16 | 5.26E-06 |
| ENSG00000223503 | RP11-29H23.6 | 1.51 | 5.16 | 2.93E-04 |
| ENSG00000145536 | ADAMTS16 | 1.51 | 4.18 | 7.80E-03 |
| CATG00000063540 | CATG00000063540.1 | 1.50 | 4.78 | 1.21E-03 |
| CATG00000049673 | CATG00000049673.1 | 1.48 | 8.27 | 2.52E-11 |
| CATG00000001331 | CATG00000001331.1 | 1.47 | 4.43 | 3.87E-03 |
| ENSG00000221944 | TIGD1 | 1.46 | 4.60 | 2.12E-03 |
| CATG00000083638 | CATG00000083638.1 | 1.46 | 4.77 | 1.21E-03 |
| CATG00000042831 | CATG00000042831.1 | 1.45 | 4.56 | 2.38E-03 |
| CATG00000064652 | CATG00000064652.1 | 1.45 | 5.63 | 5.36E-05 |
| CATG00000083353 | CATG00000083353.1 | 1.45 | 4.65 | 1.80E-03 |
| ENSG00000270487 | RP11-230C9.3 | 1.44 | 5.91 | 1.62E-05 |
| ENSG00000113249 | HAVCR1 | 1.43 | 4.46 | 3.57E-03 |
| CATG00000013670 | CATG00000013670.1 | 1.42 | 6.33 | 2.53E-06 |
| CATG00000113639 | CATG00000113639.1 | 1.41 | 4.83 | 1.04E-03 |
| CATG00000079832 | CATG00000079832.1 | 1.40 | 4.21 | 7.14E-03 |
| CATG00000099987 | CATG00000099987.1 | 1.40 | 4.32 | 5.26E-03 |
| CATG00000098414 | CATG00000098414.1 | 1.40 | 4.33 | 5.18E-03 |
| CATG00000039185 | CATG00000039185.1 | 1.39 | 4.68 | 1.59E-03 |
| CATG00000104566 | CATG00000104566.1 | 1.39 | 4.23 | 6.87E-03 |
| ENSG00000254982 | HMGB1P24 | 1.38 | 4.61 | 2.05E-03 |
| CATG00000038892 | CATG00000038892.1 | 1.38 | 5.30 | 1.92E-04 |
| CATG00000010722 | CATG00000010722.1 | 1.37 | 4.48 | 3.32E-03 |
| CATG00000049360 | CATG00000049360.1 | 1.37 | 5.41 | 1.19E-04 |
| CATG00000040000 | CATG00000040000.1 | 1.36 | 4.39 | 4.37E-03 |
| CATG00000078852 | CATG00000078852.1 | 1.36 | 5.06 | 4.30E-04 |
| CATG00000018127 | CATG00000018127.1 | 1.35 | 4.17 | 8.01E-03 |
| ENSG00000212864 | RNF208 | 1.35 | 4.37 | 4.63E-03 |
| CATG00000005873 | CATG00000005873.1 | 1.35 | 4.77 | 1.23E-03 |
| CATG00000036922 | CATG00000036922.1 | 1.34 | 5.34 | 1.63E-04 |
| CATG00000116159 | CATG00000116159.1 | 1.34 | 5.17 | 2.87E-04 |
| ENSG00000174527 | MYO1H | 1.34 | 4.32 | 5.26E-03 |
| ENSG00000254463 | RP11-484D2.3 | 1.33 | 4.37 | 4.63E-03 |
| CATG00000005872 | CATG00000005872.1 | 1.33 | 4.57 | 2.35E-03 |
| CATG00000022162 | CATG00000022162.1 | 1.32 | 4.10 | 9.70E-03 |
| ENSG00000007001 | UPP2 | 1.32 | 4.61 | 2.05E-03 |

| | | | | |
|---|---|---|---|---|
| CATG00000036880 | CATG00000036880.1 | 1.31 | 4.50 | 3.05E-03 |
| CATG00000001295 | CATG00000001295.1 | 1.30 | 4.71 | 1.44E-03 |
| ENSG00000235296 | AC137723.5 | 1.30 | 4.42 | 3.95E-03 |
| CATG00000113638 | CATG00000113638.1 | 1.30 | 4.23 | 6.88E-03 |
| CATG00000038960 | CATG00000038960.1 | 1.30 | 4.26 | 6.26E-03 |
| ENSG00000226174 | TEX22 | 1.29 | 5.13 | 3.29E-04 |
| CATG00000046244 | CATG00000046244.1 | 1.29 | 4.44 | 3.75E-03 |
| ENSG00000204839 | MROH6 | 1.27 | 4.76 | 1.24E-03 |
| CATG00000045991 | CATG00000045991.1 | 1.27 | 4.39 | 4.37E-03 |
| ENSG00000110375 | UPK2 | 1.26 | 4.80 | 1.11E-03 |
| CATG00000096385 | CATG00000096385.1 | 1.26 | 4.11 | 9.64E-03 |
| ENSG00000204248 | COL11A2 | 1.26 | 4.12 | 9.35E-03 |
| ENSG00000253731 | PCDHGA6 | 1.26 | 4.15 | 8.45E-03 |
| ENSG00000059915 | PSD | 1.25 | 5.46 | 1.02E-04 |
| ENSG00000180998 | GPR137C | 1.25 | 5.50 | 9.20E-05 |
| ENSG00000239247 | RN7SL589P | 1.24 | 4.24 | 6.74E-03 |
| ENSG00000172382 | PRSS27 | 1.24 | 5.06 | 4.30E-04 |
| CATG00000105296 | CATG00000105296.1 | 1.24 | 4.48 | 3.30E-03 |
| CATG00000099573 | CATG00000099573.1 | 1.23 | 4.36 | 4.74E-03 |
| CATG00000055615 | CATG00000055615.1 | 1.23 | 5.22 | 2.66E-04 |
| CATG00000094807 | CATG00000094807.1 | 1.22 | 4.27 | 5.98E-03 |
| ENSG00000260493 | RP11-219B4.7 | 1.22 | 4.27 | 6.02E-03 |
| ENSG00000258982 | RP11-638I2.4 | 1.22 | 6.30 | 2.66E-06 |
| ENSG00000224647 | AC026954.6 | 1.22 | 4.13 | 8.84E-03 |
| CATG00000066066 | CATG00000066066.1 | 1.21 | 4.20 | 7.37E-03 |
| CATG00000109082 | CATG00000109082.1 | 1.21 | 5.22 | 2.66E-04 |
| ENSG00000252821 | RNU6-388P | 1.20 | 5.69 | 4.37E-05 |
| ENSG00000271993 | RP11-285J16.1 | 1.20 | 4.71 | 1.44E-03 |
| ENSG00000205704 | LINC00634 | 1.20 | 4.14 | 8.69E-03 |
| CATG00000036982 | CATG00000036982.1 | 1.19 | 5.45 | 1.05E-04 |
| ENSG00000154035 | C17orf103 | 1.19 | 4.98 | 6.10E-04 |
| CATG00000095984 | CATG00000095984.1 | 1.19 | 4.77 | 1.22E-03 |
| ENSG00000232082 | RPS6KA2-IT1 | 1.18 | 4.11 | 9.38E-03 |
| CATG00000034566 | CATG00000034566.1 | 1.16 | 4.13 | 8.98E-03 |
| CATG00000109148 | CATG00000109148.1 | 1.15 | 4.39 | 4.40E-03 |
| CATG00000107360 | CATG00000107360.1 | 1.15 | 4.85 | 9.65E-04 |
| ENSG00000238098 | ABCA17P | 1.14 | 4.85 | 9.60E-04 |
| ENSG00000239911 | PRKAG2-AS1 | 1.13 | 4.36 | 4.74E-03 |
| ENSG00000230841 | RP5-915N17.3 | 1.13 | 4.28 | 5.90E-03 |
| CATG00000052130 | CATG00000052130.1 | 1.12 | 5.26 | 2.23E-04 |

| | | | | |
|---|---|---|---|---|
| ENSG00000229869 | RP11-363N22.2 | 1.11 | 4.76 | 1.24E-03 |
| CATG00000028303 | CATG00000028303.1 | 1.10 | 4.82 | 1.04E-03 |
| CATG00000076772 | CATG00000076772.1 | 1.09 | 4.19 | 7.71E-03 |
| ENSG00000176268 | CYCSP34 | 1.09 | 4.19 | 7.71E-03 |
| CATG00000004547 | CATG00000004547.1 | 1.08 | 4.30 | 5.50E-03 |
| CATG00000096427 | CATG00000096427.1 | 1.07 | 4.31 | 5.44E-03 |
| ENSG00000224939 | LINC00184 | 1.07 | 4.70 | 1.45E-03 |
| CATG00000058452 | CATG00000058452.1 | 1.07 | 4.87 | 9.21E-04 |
| ENSG00000230002 | ALMS1-IT1 | 1.06 | 4.33 | 5.18E-03 |
| ENSG00000233654 | AC093388.3 | 1.05 | 5.10 | 3.68E-04 |
| CATG00000103257 | CATG00000103257.1 | 1.05 | 5.09 | 3.92E-04 |
| ENSG00000101850 | GPR143 | 1.04 | 6.01 | 1.08E-05 |
| CATG00000018529 | CATG00000018529.1 | 1.01 | 5.18 | 2.80E-04 |
| CATG00000067650 | CATG00000067650.1 | 1.01 | 4.21 | 7.30E-03 |
| CATG00000043927 | CATG00000043927.1 | 1.01 | 4.25 | 6.54E-03 |
| ENSG00000245322 | RP11-15B17.1 | 1.01 | 4.33 | 5.18E-03 |
| ENSG00000230797 | YY2 | 0.99 | 5.87 | 1.85E-05 |
| CATG00000058711 | CATG00000058711.1 | 0.98 | 4.62 | 2.04E-03 |
| CATG00000059716 | CATG00000059716.1 | 0.97 | 4.96 | 6.68E-04 |
| ENSG00000168350 | DEGS2 | 0.95 | 4.34 | 5.06E-03 |
| CATG00000018128 | CATG00000018128.1 | 0.95 | 4.29 | 5.70E-03 |
| CATG00000108312 | CATG00000108312.1 | 0.93 | 4.41 | 4.09E-03 |
| CATG00000039249 | CATG00000039249.1 | 0.93 | 4.25 | 6.51E-03 |
| ENSG00000271966 | RP11-7F18.2 | 0.92 | 4.16 | 8.32E-03 |
| ENSG00000152926 | ZNF117 | 0.92 | 5.87 | 1.85E-05 |
| ENSG00000251136 | RP11-37B2.1 | 0.92 | 4.33 | 5.18E-03 |
| ENSG00000226900 | RP11-432J24.5 | 0.92 | 4.17 | 8.22E-03 |
| ENSG00000011021 | CLCN6 | 0.89 | 4.94 | 7.10E-04 |
| CATG00000050335 | CATG00000050335.1 | 0.88 | 4.23 | 6.88E-03 |
| ENSG00000227394 | AC007386.3 | 0.88 | 5.41 | 1.19E-04 |
| ENSG00000229689 | AC009237.8 | 0.86 | 4.32 | 5.23E-03 |
| ENSG00000261662 | RP5-1042I8.7 | 0.84 | 5.40 | 1.25E-04 |
| CATG00000039106 | CATG00000039106.1 | 0.84 | 4.25 | 6.54E-03 |
| ENSG00000146263 | MMS22L | 0.84 | 4.47 | 3.36E-03 |
| ENSG00000163666 | HESX1 | 0.83 | 4.21 | 7.14E-03 |
| CATG00000014111 | CATG00000014111.1 | 0.81 | 4.36 | 4.74E-03 |
| ENSG00000160229 | ZNF66 | 0.80 | 4.50 | 3.07E-03 |
| ENSG00000174483 | BBS1 | 0.77 | 5.21 | 2.66E-04 |
| ENSG00000178665 | ZNF713 | 0.76 | 4.16 | 8.33E-03 |
| CATG00000041526 | CATG00000041526.1 | 0.74 | 4.32 | 5.27E-03 |

| | | | | |
|---|---|---|---|---|
| ENSG00000172977 | KAT5 | 0.72 | 5.65 | 4.83E-05 |
| ENSG00000040275 | SPDL1 | 0.71 | 4.89 | 8.67E-04 |
| CATG00000047951 | CATG00000047951.1 | 0.71 | 4.31 | 5.44E-03 |
| ENSG00000159259 | CHAF1B | 0.71 | 4.72 | 1.44E-03 |
| ENSG00000204947 | ZNF425 | 0.69 | 4.78 | 1.21E-03 |
| ENSG00000173715 | C11orf80 | 0.68 | 4.86 | 9.34E-04 |
| ENSG00000238058 | RP11-432J22.2 | 0.68 | 4.79 | 1.17E-03 |
| ENSG00000186312 | CA5BP1 | 0.66 | 5.67 | 4.68E-05 |
| ENSG00000116017 | ARID3A | 0.65 | 4.12 | 9.35E-03 |
| ENSG00000119772 | DNMT3A | 0.60 | 4.95 | 6.72E-04 |
| ENSG00000176809 | LRRC37A3 | 0.59 | 4.73 | 1.39E-03 |
| ENSG00000172613 | RAD9A | 0.58 | 4.12 | 9.35E-03 |
| ENSG00000242338 | BMS1P4 | 0.57 | 4.40 | 4.37E-03 |
| ENSG00000132740 | IGHMBP2 | 0.52 | 4.43 | 3.87E-03 |
| ENSG00000188690 | UROS | 0.50 | 4.71 | 1.44E-03 |
| CATG00000084188 | CATG00000084188.1 | 0.49 | 4.20 | 7.37E-03 |
| CATG00000096591 | CATG00000096591.1 | 0.46 | 4.57 | 2.36E-03 |
| ENSG00000125450 | NUP85 | 0.46 | 4.58 | 2.27E-03 |
| ENSG00000120784 | ZFP30 | 0.45 | 4.54 | 2.59E-03 |
| ENSG00000047230 | CTPS2 | 0.42 | 4.40 | 4.28E-03 |
| ENSG00000173120 | KDM2A | 0.39 | 4.19 | 7.64E-03 |
| ENSG00000163312 | HELQ | -0.38 | -4.19 | 7.63E-03 |
| ENSG00000108264 | TADA2A | -0.40 | -4.80 | 1.11E-03 |
| ENSG00000158793 | NIT1 | -0.42 | -4.16 | 8.33E-03 |
| ENSG00000166262 | FAM227B | -0.59 | -4.37 | 4.63E-03 |
| ENSG00000120860 | CCDC53 | -0.60 | -4.14 | 8.76E-03 |
| ENSG00000221817 | RP11-137L10.6 | -0.73 | -4.91 | 8.00E-04 |
| CATG00000039122 | CATG00000039122.1 | -0.82 | -4.29 | 5.70E-03 |
| ENSG00000185267 | CDNF | -0.86 | -4.43 | 3.87E-03 |
| ENSG00000224424 | PRKAR2A-AS1 | -0.88 | -4.28 | 5.90E-03 |
| ENSG00000186976 | EFCAB6 | -0.95 | -5.34 | 1.60E-04 |
| ENSG00000064199 | SPA17 | -0.95 | -4.16 | 8.32E-03 |
| ENSG00000071082 | RPL31 | -1.17 | -4.57 | 2.37E-03 |
| ENSG00000272983 | RP11-508N22.12 | -1.18 | -5.03 | 5.00E-04 |
| ENSG00000186567 | CEACAM19 | -1.23 | -4.46 | 3.58E-03 |
| ENSG00000242220 | TCP10L | -1.25 | -5.17 | 2.87E-04 |
| CATG00000000073 | CATG00000000073.1 | -1.27 | -4.15 | 8.36E-03 |
| ENSG00000169085 | C8orf46 | -1.38 | -4.71 | 1.44E-03 |
| ENSG00000144410 | CPO | -1.38 | -4.72 | 1.44E-03 |
| ENSG00000235271 | RP1-90L6.3 | -1.44 | -4.18 | 7.80E-03 |

| | | | | |
|---|---|---|---|---|
| CATG00000109304 | CATG00000109304.1 | -1.47 | -4.22 | 6.91E-03 |
| CATG00000106558 | CATG00000106558.1 | -1.73 | -4.68 | 1.57E-03 |
| CATG00000111190 | CATG00000111190.1 | -1.75 | -4.48 | 3.32E-03 |

1548