UPPSALA
UNIVERSITET

# Non-coding constraint mutations impact the gene regulatory system in osteosarcoma

Raphaela Pensch

---

Degree project in bioinformatics, 2021
Examensarbete i bioinformatik 45 hp till masterexamen, 2021
Biology Education Centre and Dept. of Medical Biochemistry and Microbiology, Uppsala University
Supervisor: Professor Kerstin Lindblad-Toh

# Abstract

The non-coding space makes up around 98 % of the genome, but cancer-driving mutations have so far mostly been discovered in protein-coding regions. The majority of somatic non-coding mutations are neutral passenger mutations and identifying non-coding mutations with driving roles in cancer poses a challenge. In this work, evolutionary constraint was used to explore the non-coding space in human osteosarcoma to improve our understanding of how evolutionary constraint can be applied to identify non-coding driver mutations in cancer and describe the unknown role of non-coding mutations in osteosarcoma. Evolutionary constraint scores derived from an alignment of 33 mammals were used to extract non-coding mutations in functional elements from somatic variants of 38 osteosarcoma samples and genes with an enrichment of non-coding constraint mutations in their regulatory regions were identified. The investigation of those genes revealed that non-coding constraint mutations are likely involved in key osteosarcoma pathways. Furthermore, novel osteosarcoma genes and mechanisms were proposed based on the non-coding constraint mutation enrichment analysis. The regulatory potential of individual non-coding constraint mutations was evaluated based on regulatory annotations, functional evidence, transcription factor affinity predictions and electrophoretic mobility shift assays. We concluded that the analysis of non-coding constraint mutations is an efficient way to discover non-coding mutations with functional impact in osteosarcoma which likely play an important role in the disease.

# Investigating non-coding mutations in osteosarcoma

## Popular Science Summary

Raphaela Pensch

The cells in our body work together as a large community. To ensure that the complicated processes in our body run smoothly, all cells have to contribute and perform their tasks meticulously. They grow, process nutrients and synthesise proteins. The instructions for their tasks are noted in the DNA of which each cell receives the same copy. Included in the DNA are protein-coding genes, which are blueprints for the proteins that a cell builds itself. Other parts of the DNA that are called non-coding regions have less clearly defined roles, but describe for example at what point in time a cell needs to build how much of a particular protein. Cells grow and multiply by dividing and thereby forming new, identical cells. When cells divide, they first replicate their DNA and assign each daughter cell they then split into their own copy. This guarantees that every cell always knows exactly what to do.

The instruction manual, that is DNA, is long and complicated and mistakes in the replication process are unavoidable. Such mistakes are called mutations. When DNA in a cell is mutated, this particular cell does not know how to properly fulfil its tasks and acts up. This jeopardises the delicate balance in the cellular community and therefore needs to be avoided by all means. Luckily, the cells in our body are prepared: The DNA includes an emergency plan that tells them how to solve this situation and usually commands the affected cells to kill themselves. Serious problems arise, however, when mistakes are made in the replication of the emergency plan itself. Cells with an erroneous or incomplete emergency plan do not know how to deal with DNA mutations anymore. Because nothing is telling these cells to stop working, they continue to perform their tasks according to their instructions. When they divide, they pass on their faulty manual to their daughter cells which means the number of misbehaving cells increases steadily. Furthermore, once the emergency plan and its control mechanisms are disabled, the DNA accumulates additional mutations more easily. The affected cells start to form tumours and spread around the body. This is how cancer develops.

Osteosarcomas are particularly aggressive tumours and the most common form of bone cancer. To improve our understanding of the disease, we aimed to describe the previously unknown role of mutations in the non-coding regions of osteosarcoma. We have a fairly good idea of the consequences of mutations that occur in genes. However, genes only make up a small proportion of the DNA and most mutations happen outside of them. Determining which mutations in the non-coding regions of DNA are relevant is challenging, because they exist in great numbers and their effects are harder to pinpoint than those of mutations targeting genes. One approach to extract the important parts of non-coding regions in the DNA is to compare the DNA sequence of related species. As they have evolved from common ancestors, the DNA has differences and similarities among species. The idea is that regions of the DNA that are the same across multiple species remained unchanged over long evolutionary timescales because they have an important function. In this work, we used information about similarities in the DNA of 33 mammals to identify mutations in the non-coding regions that impact osteosarcoma.

Degree project in bioinformatics, 2021
Examensarbete i bioinformatik 45 hp till masterexamen, 2021
Biology Education Centre and Department of Medical Biochemistry and Microbiology, Uppsala University
Supervisor: Professor Kerstin Lindblad-Toh

# Table of Contents

# Abbreviations

| | |
|---|---|
| CGC | Cancer Gene Census |
| ChIP-seq | Chromatin immunoprecipitation sequencing |
| CNA | Copy number alteration |
| DNA | Deoxyribonucleic acid |
| EMSA | Electrophoretic mobility shift assay |
| eQTL | Expression quantitative train locus |
| FMG | Frequently mutated gene |
| GERP | Genomic Evolutionary Rate Profiling |
| GJ | Gap junction |
| GO | Gene Ontology |
| ICGC | International Cancer Genome Consortium |
| LincRNA | Long intergenic non-coding RNA |
| LncRNA | Long non-coding RNA |
| NCCM | Non-coding constraint mutation |
| OBT | Other bone tumour |
| PCAWG | Pan-cancer analysis of whole genomes |
| PhyloP | Phylogenetic p-values |
| RMG | Recurrently mutated gene |
| RNA | Ribonucleic acid |
| RNA-seq | RNA sequencing |
| SIM | Somatic indel mutation |
| SMG | Significantly mutated gene |
| SPIMs | Somatic point and indel mutations |
| SPM | Somatic point mutation |
| sTRAP | Sequence Transcription Factor Affinity Prediction |
| TAD | Topologically associating domain |
| TCGA | The Cancer Genome Atlas |
| UTR | Untranslated region |
| WGS | Whole-genome sequencing |

# 1 Introduction

In 2020, more than 19 million people worldwide were newly diagnosed with cancer and almost 10 million patients died of the disease (Sung *et al.* 2021). Our ability to treat cancer has greatly improved in the past decade and strong research efforts are put into continuously improving our understanding of the disease. Despite this, cancer is still difficult or, in many cases, impossible to cure.

Osteosarcoma is a very aggressive malignancy and the most common primary bone tumour (Mirabello *et al.* 2009). It is a predominantly paediatric disease with most cases below the age of 24, but shows a second incidence peak in cases above the age of 60 (Mirabello *et al.* 2009). Tumours most commonly affect the metaphysis of the lower long bones and there seems to be a correlation between paediatric onset of the disease and phases of rapid bone growth as they occur during puberty or *in utero* (Mirabello *et al.* 2009). Osteosarcoma is also frequently associated with certain cancer predisposition syndromes, such as Li-Fraumeni Syndrome, hereditary bilateral retinoblastoma and Bloom syndrome as well as Paget's disease (Chauveinc *et al.* 2001, Varley 2003, Hansen *et al.* 2006, Kansara & Thomas 2007, Mirabello *et al.* 2011). Tumour stage at the time of diagnosis and factors such as the distribution of metastases strongly influence the prognosis for osteosarcoma patients, but overall mortality is high. With the implementation of surgery and chemotherapy as standard treatment, survival rates have improved to 60 – 70 %, but progress has since stagnated (Mirabello *et al.* 2009, Rickel *et al.* 2017, Siegel *et al.* 2021). Five-year survival rates for patients who are initially diagnosed with metastatic osteosarcoma that has spread to distant tissues are as low as 20 - 27 % (Mialou *et al.* 2005, Howlader *et al.* 2019). Understanding the genetic mechanisms that drive osteosarcoma is a main objective in the endeavour to develop better treatment and cure more patients.

Cancer is driven by the accumulation of somatic mutations (although genetic predisposing mutations also exist) which eventually enable cancer cells to divide uncontrollably and invade healthy tissue. Somatic mutations that provide an advantage to the cancer are subject to natural selection and allow cancer cells to bypass normal control mechanisms and proliferate without adhering to cellular constraints (Stratton *et al.* 2009). Somatic mutations that support cancer growth and are causally involved in tumorigenesis are called driver mutations (Vogelstein *et al.* 2013). Driver mutations are very rare compared to neutral passenger mutations, but their discovery is crucial to gain a more thorough understanding of the mechanisms that underlie cancer growth and eventually develop better treatment (Bailey *et al.* 2018). Therefore, great efforts are put into understanding driver mutations and the driver genes they act on as well as developing better methods to identify them (Lawrence *et al.* 2013, Dietlein *et al.* 2020).

Most cancer-driving mutations have so far been discovered in the protein coding regions of the genome. Tumour suppressor genes have regulatory roles in cell division, replication and apoptosis and are frequently inactivated by driver mutations. Oncogenes, on the other hand, promote cancer proliferation and can be activated by gain-of-function mutations. The identification of driver mutations in protein coding regions undoubtedly leads to great progress in cancer research, however, setting the focus mainly on the coding space leaves the largest part of the genome unexplored (Hornshøj *et al.* 2018). Non-coding mutations can alter cellular functions and gene regulation when they are found in functional elements involved in transcriptional or post-transcriptional regulation (Hornshøj *et al.* 2018). The most well-known example of a non-coding driver element is the *TERT* promoter. Telomerase reverse

transcriptase (*TERT*) is responsible for maintaining telomere length which is pivotal for tumour persistence in many cancers. Somatic mutations in this gene's promoter region can enhance its expression, and associated telomere extension, making them important drivers of tumorigenesis (Fredriksson *et al.* 2014). The discovery of the driving role of mutations in the *TERT* promoter has heightened awareness of the importance of non-coding driver elements and shifted focus to the non-coding space. Higher accessibility of whole-genome sequencing (WGS), large sample sizes and new computational methods have led to additional findings (Mularoni *et al.* 2016, Cuykendall *et al.* 2017). However, the number of drivers that have been identified in the coding space continues to surpass the number of non-coding drivers by far. Investigating non-coding regions in the search for driver mutations has the potential to lead to exciting new discoveries (Cuykendall *et al.* 2017).

Identifying driver mutations among the vast amount of non-coding mutations poses a challenge. Restricting the search to annotated functional elements is not straightforward, because the workings of the gene regulatory system are complicated and its annotation in the human reference genome incomplete. Evolutionary constraint is one measure that can be applied to estimate the position of functional elements (Lindblad-Toh *et al.* 2011). Between 6 and 13 % of the human genome is estimated to be under evolutionary constraint which means that it has undergone purifying selection and is therefore likely functional (Davydov *et al.* 2010, Meader *et al.* 2010, Lindblad-Toh *et al.* 2011, Rands *et al.* 2014). Using comparative genomics, the level of evolutionary constraint can be determined for most sites in the genome.

Sakthikumar *et al.* developed a novel approach that uses evolutionary constraint scores to identify non-coding mutations in functional elements with regulatory potential and successfully applied it to investigate non-coding mutations in glioblastoma (Sakthikumar *et al.* 2020). They generated WGS data of matched tumour tissue and blood samples to perform variant calling and functional annotation. Genes with known roles in glioblastoma were selected and their associated non-coding regions examined for mutations occurring in constrained sites. They found that non-coding constraint mutations were enriched in the neighbourhood of key glioblastoma genes and proposed novel genes with putative roles in glioblastoma based on high numbers of non-coding constraint mutations in their regulatory regions. Furthermore, Sakthikumar *et al.* were able to show that a non-coding constraint mutation in the promoter region of semaphorin 3C (*SEMA3C*) disrupted a FOXA1 transcription factor binding site leading to decreased binding affinity of transcription factors to that region and potentially altering *SEMA3C* gene expression levels. The authors concluded that non-coding constraint mutations likely play an essential role in glioblastoma.

In the course of this project, we aimed to explore the landscape of non-coding constraint mutations in osteosarcoma to improve our understanding of how evolutionary constraint can be utilized to identify non-coding driver mutations and describe the previously unknown role of non-coding mutations in osteosarcoma. We mapped evolutionary constraint scores to somatic variant data generated by WGS of 38 osteosarcoma samples and extracted non-coding constraint mutations in potential regulatory regions. Established osteosarcoma genes with an enrichment of non-coding constraint mutations were used to investigate the involvement of non-coding constraint mutations in classical osteosarcoma pathways. Non-coding constraint mutation enrichment analysis also allowed us to propose novel osteosarcoma genes. We evaluated the regulatory potential of individual mutations using regulatory annotations, functional evidence, transcription factor affinity predictions and electrophoretic mobility shift assays and concluded that non-coding constraint mutation analysis is an efficient approach to identify non-coding mutation with functional impact in osteosarcoma.

# 2 Materials and Methods

## 2.1 ICGC bone tumour dataset

We obtained a dataset of somatic bone tumour variants from the International Cancer Genome Consortium (ICGC) Data Portal (https://dcc.icgc.org) to investigate non-coding constraint mutations in osteosarcoma. The data has previously been published as part of the ICGC/TCGA Pan-Cancer Analysis of Whole Genomes (PCAWG) Consortium's effort to collect WGS data and analyse genomic features across different tumour types (Campbell *et al.* 2020). The downloaded dataset comprised somatic variants from 64 patients and was generated by WGS of matched tumour and normal (blood) samples using Illumina HiSeq paired-end sequencing.

The consortium applied three different pipelines to call somatic variants against a version of the human reference build hg19 (hs37d5 from the 1000 Genomes Project )(Auton *et al.* 2015). A high-confidence consensus set of variants was derived by selecting single-nucleotide variants that have been called by at least two pipelines. Indel calls were integrated based on stacked logistic regression. Additionally, multiple optimized processing and quality control steps ensured that a set of high-quality somatic variants was provided (Campbell *et al.* 2020).

Somatic point mutation (SPM) and somatic indel mutation (SIM) data from seven bone tumour types was included in the dataset. These were osteosarcoma ($n$ = 38), chondroblastoma ($n$ = 7), osteoblastoma ($n$ = 6), adamantinoma ($n$ = 5), chordoma ($n$ = 5), chondromyxoid fibroma ($n$ = 2) and ameloblastoma ($n$ = 1). Across the entire cohort, 30 patients were female and 34 were male. Among them, there were 20 female and 18 male osteosarcoma patients. The median age of the entire bone tumour cohort was 22 and 19.5 for osteosarcoma patients (Appendix Figure 1). The age of four osteosarcoma patients was unknown and no additional clinical data about survival time, treatment, etc. was available for the entire cohort at the ICGC Data Portal or within the PCAWG publication.

Copy number alteration (CNA) data for the bone tumour cohort, which has been generated using the DKFZ somatic variant pipeline, was downloaded from the ICGC Data Portal as well (Campbell *et al.* 2020).

## 2.2 Somatic variant pre-processing pipeline

Three additional steps of filtering and quality control were applied to the dataset:

First, somatic point and indel mutations (SPIMs) that did not pass the quality criteria applied by the PCAWG consortium during variant calling were removed.

Then, the data was scanned for SPIMs that were called more than once per sample in the same position. The SPIM that was supported by more variant callers was chosen and the redundant variant removed. In case of a tie, the longer SIM was selected.

On top of this, SPIMs were compared to a dataset of known germline variants to determine the number of potential germline artefacts among them. This dataset consisted of germline variants from dbSNP (version 151) and SweGen (Sherry *et al.* 1999, Ameur *et al.* 2017). Confirmed somatic variants from the COSMIC database (version 91) were whitelisted (Tate *et al.* 2019). The number of SPIMs per sample that coincide with the dataset of know germline variants was small (on average 3 % per sample). Such low numbers of potential germline artefacts unlikely interfere with later analyses and were therefore kept in the dataset.

The remaining SPIMs were annotated with the functional annotation tool Funcotator (GATK 4.1.4.1) using pre-packaged data sources (version 1.6.20190124s) (McKenna *et al.* 2010). *--force-b37-to-hg19-reference-contig-conversion* was applied to translate to the correct reference genome.

CNAs were filtered according to the quality measures applied by the PCAWG consortium and annotated with gene information by intersecting CNAs with gene positions using BEDTools (Quinlan & Hall 2010). Only CNAs that had been identified as an amplification, deletion or loss of heterozygosity and overlapped with ≥ 90 % of a gene were considered for the following analyses.

After this step, the osteosarcoma cohort was disjoined from the bone tumour dataset and analysed separately.

## 2.3 Coding mutations

Because the ICGC bone tumour cohort has been published as part of a pan-cancer study and has not previously been analysed individually, we first examined osteosarcoma SPIMs in protein-coding sequences and compared the results to available literature on osteosarcoma. This analysis was performed to validate that this dataset could be used to reproduce established results, before investigating the unexplored non-coding space in osteosarcoma.

### 2.3.1 Recurrently mutated genes

SPIMs in coding regions were processed using different approaches.

First, significantly mutated genes (SMGs) in osteosarcoma were identified using MutSigCV (version 1.41) (Lawrence *et al.* 2013).

Then, SPIMs were filtered for coding, non-silent mutations and those counts used to assign genes to the sets of frequently mutated genes (FMGs) and recurrently mutated genes (RMGs). We defined FMGs as genes with mutations in the protein-coding sequence in at least 10 % of samples and RMGs as genes with coding mutations in at least two samples.

The absolute number of SPIMs found in a particular gene is not necessarily the most appropriate measure of its importance in cancer. To allow a more meaningful comparison of RMGs among each other, the number of SPIMs per Kbp protein-coding sequence was calculated as an additional metric for each gene (Ensembl Genes 103 for GRCh37).

All RMGs including SMGs and FMGs were intersected with the Cancer Gene Census (CGC) dataset based on the COSMIC database (v92) to identify known cancer genes among them (https://cancer.sanger.ac.uk/census). CGC is a curated collection of genes that drive cancer. They are divided into two groups: Tier 1 genes have documented roles in cancer supported by experimental evidence, while Tier 2 genes show strong indications of being cancer genes but are supported by less extensive evidence (Sondka *et al.* 2018).

Coding mutations and CNAs of relevant RMGs were visualised with brick plots that were generated using Oncoprinter from cBioPortal (Cerami *et al.* 2012, Gao *et al.* 2013).

### 2.3.2 Pathway enrichment

Pathway enrichment tools identify pathways that are significantly overrepresented in a given set of genes. In the context of cancer, pathway enrichment analysis of mutated genes provides insight into which pathways have been disrupted or altered. We performed pathway enrichment

analysis to reveal patterns and connections in the RMG set and relate these to previous findings in osteosarcoma. GSEA and MSigDB (version v7.2) were used to compute the overlap between RMGs and the MSigDB canonical pathways gene set which included pathway gene sets from BioCarta, KEGG, PID, Reactome and WikiPathways (Nishimura 2001, Daly *et al.* 2003, Subramanian *et al.* 2005, Schaefer *et al.* 2009, Liberzon *et al.* 2011, Fabregat *et al.* 2016, Kanehisa *et al.* 2017, Slenter *et al.* 2018)

## 2.4 Non-coding constraint mutations

To differentiate between neutral passenger mutations and non-coding mutations in functional elements, evolutionary constraint scores were mapped to the set of SPIMs. Because NCCMs in regulatory regions have the potential to alter gene transcription, we then extracted non-coding constraint mutations from the non-coding regions surrounding genes where we considered them most likely to be able to impact the gene regulatory system.

### 2.4.1 Annotation with evolutionary constraint scores

We downloaded precomputed evolutionary constraint scores for the human reference assembly hg19 from the UCSC Genome Browser database (Haeussler *et al.* 2019).

The scores have been generated with the statistical framework GERP (Genomic Evolutionary Rate Profiling) and were derived from a multiple sequence alignment of 33 mammals. Genomic positions which have been subjected to purifying selection were identified as regions with fewer substitutions than expected and the level of constraint quantified by measuring the substitution deficit as "rejected substitutions" (Cooper *et al.* 2005, Davydov *et al.* 2010).

GERP scores were mapped to SPIMs by intersecting both datasets using BEDOPS (version 2.4.39) and bigWigToBedGraph (Kent *et al.* 2010, Neph *et al.* 2012). Deletions that spanned across two or more sites were annotated by collecting GERP scores for all affected positions and reporting the highest score for the variant. We applied a threshold of GERP $\geq 2.0$ to retrieve SPIMs in constrained elements and termed SPIMs in non-coding regions under evolutionary constraint non-coding constraint mutations (NCCMs).

### 2.4.2 Non-coding constraint mutation enrichment analysis

To identify genes that are regulated by NCCMs in osteosarcoma, we scanned the genome across the osteosarcoma cohort for genes with an enrichment of NCCMs in their associated non-coding regions. We deemed introns, UTRs and the 100 Kbp flanking regions of genes as regions where NCCMs most likely have a functional impact on the regulation of a particular gene. Therefore, NCCMs that had a GERP score of $\geq 2$ and were positioned in these non-coding regions were retrieved from the dataset.

Subsequently, we calculated a rate for each gene that described the number of samples with NCCMs in the non-coding regions associated with it (NCCMs/ 100 Kbp). Genes that were assigned $\geq 2$ NCCMs/100 Kbp were considered to have an enrichment of NCCMs.

NCCM enrichment analysis was performed on the remaining bone tumour SPIM dataset excluding osteosarcoma samples as a negative control.

### 2.4.3 Characterisation of NCCM-enriched genes

We performed pathway enrichment and gene expression analyses to describe the roles of genes with an enrichment of NCCMs in osteosarcoma.

Using GSEA and MSigDB, we computed the overlap of the set of NCCM-enriched genes with the Gene Ontology (GO) gene set to find out which biological processes these genes were involved in (The Gene Ontology Consortium *et al.* 2000, Carbon *et al.* 2021).

Important osteosarcoma genes were expected to show expression differences when compared to normal osteoblasts. To investigate how many NCCM-enriched genes this applied to, we obtained a gene expression dataset which has been generated using Affymetrix Gene 1.0 ST expression arrays on 12 paediatric osteosarcoma and two normal human osteoblast samples and is available under the GEO accession number gse12865 (Sadikovic *et al.* 2009). Significant differences in gene expression between osteosarcomas and osteoblasts were determined with Student's t-tests ($p < 0.05$) using *ttest_ind()* from the Statistical functions (scipy.stats) Python module.

### 2.4.4 NCCM regulatory annotation

To evaluate the regulatory potential of individual mutations, NCCMs were annotated with regulatory information which was provided by UCSC Genome Browser, ENCODE Project, GENCODE (version v16) and others (downloaded http://larva.gersteinlab.org/) and visualized in the UCSC Genome Browser (Dunham *et al.* 2012, Davis *et al.* 2018, Frankish *et al.* 2019). The regulation data included annotations for sites of open chromatin, histone markers, transcription factor binding sites identified by ChIP-seq as well as enhancer and promoter regions.

### 2.4.5 Topologically associating domains

On a sub-chromosomal level, the genome organises into topologically associating domains (TADs). Genomic regions inside these TADs tend to interact with each other rather than with regions outside TAD boundaries. Hence, genes and the regions involved in their regulation are typically part of the same TAD, although TADs can vary across tissues. To confirm that NCCMs and the regulatory regions they are located in can interact with genes of interest, we visualized chromatin interactions of NCCM sites and TADs with 3DIV (3D-genome Interaction Viewer and database), focusing on an interaction range of 100 Kbp (Yang *et al.* 2018). Hi-C data from H1-derived Mesenchymal Stem Cell samples was used due to a lack of data from osteosarcoma cell lines and the TopDom algorithm was selected to identify TADs (Shin *et al.* 2015, Dali & Blanchette 2017).

### 2.4.6 NCCM impact prediction on transcription factor binding affinity

NCCMs in regulatory regions could impact the gene regulatory system by altering transcription factor binding sites. This can in turn change the affinity with which a transcription factor binds. sTRAP is a module of the Transcription factor Affinity Prediction (TRAP) Web Tools (http://trap.molgen.mpg.de/cgi-bin/home.cgi) and was used to detect transcription factors that could bind at a specific sequence. Furthermore, it predicted whether NCCMs could cause an increase or decrease in binding affinity for those transcription factors at the particular site (Manke *et al.* 2010, Thomas-Chollier *et al.* 2011).

Wild-type and variant alleles were used as input for the web tool with their 20 bp flanking regions which contain the length of most transcription factor binding sites. NCCMs that had regulatory annotations indicating a CTCF binding site at that position were additionally analysed with their 50 bp flanking regions since the binding site for this protein is long. Predictions were based on JASPAR vertebrate matrices using the human promoter background

model and Benjamini-Hochberg multiple test correction. Matrices of significant results were downloaded from http://jaspar.genereg.net/.

### 2.4.7 Electrophoretic mobility shift assay

To experimentally confirm that NCCMs could alter the affinity with which a transcription factor binds a binding site, we performed electrophoretic mobility shift assays (EMSA) which are assays used to investigate DNA-protein interactions. We annealed 5' biotin-labelled and unlabelled forward strand DNA oligos of 40 bp including the transcription factor binding site wild-type sequence and the sequence containing the NCCM variant with their unlabelled reverse complementary strands. Nuclear protein was extracted from the Saos-2 human osteosarcoma cell line. The experiment was conducted with the LightShift™ Chemiluminescent EMSA Kit (Thermo Scientific) according to the manufacturer's protocol.

# 3 Results

## 3.1 Somatic variant pre-processing

The ICGC bone tumour dataset was filtered in multiple steps and annotated with functional and gene information. The pre-processing pipeline removed on average 50.2 % of SPMs and 81.5 % of SIMs per sample. After this, 210,852 SPIMs and an average of 3,130 SPMs and 164 SIMs per sample remained (Figure 1). The 38 osteosarcoma samples alone contained 174,375 SPIMs, which means that osteosarcoma SPIMs make up 83 % of the bone tumour dataset.

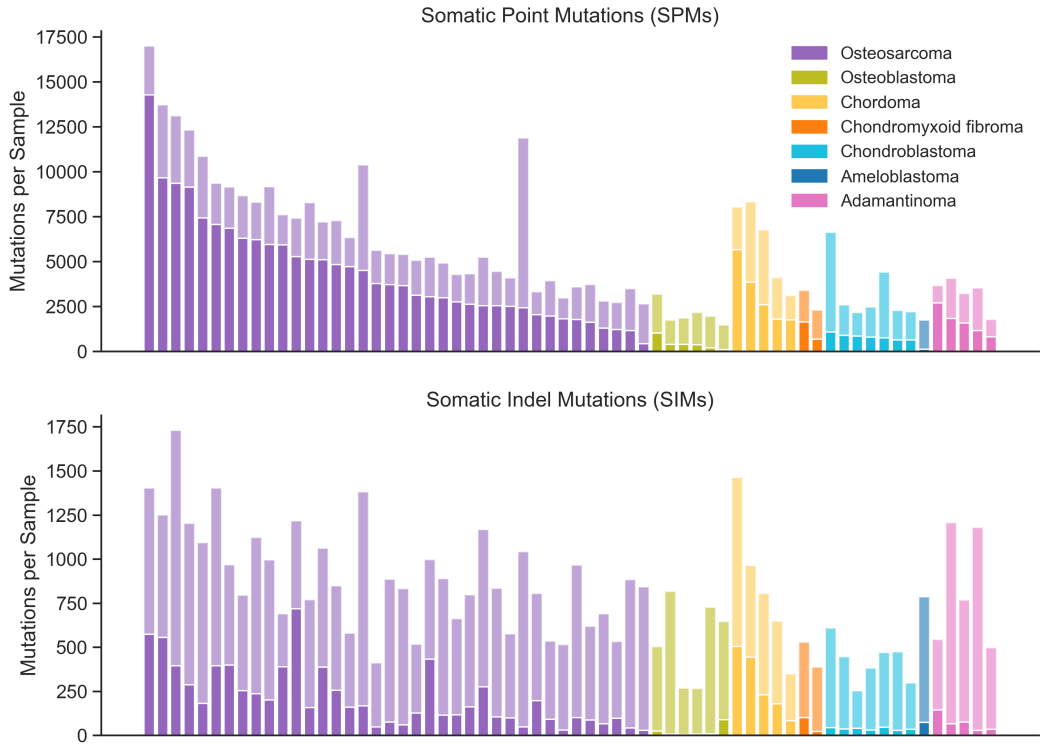



**Figure 1 | Somatic point and indel mutation (SPIM) distribution per sample.**
**The ICGC bone tumour somatic variant dataset consisted of samples from seven bone tumour types: osteosarcoma, osteoblastoma, chordoma, chondromyxoid fibroma, chondroblastoma, ameloblastoma and adamantinoma. Multiple steps of quality filtering were applied to the raw SPIM data (pale shade) to obtain the final set of SPIMs (dark shade). The majority of SPIMs in the dataset belonged to the osteosarcoma cohort.**

## 3.2 Coding mutations affect known osteosarcoma pathways

Analysing coding mutations and the genes with mutations in the coding sequence is important because it can point us towards genes that play a key role in osteosarcoma. Comparison with known osteosarcoma genes validates our pipeline and provides an indication whether our patient cohort is representative of the cancer.

*TP53* encodes the transcription factor p53 and was found to be significantly mutated in the osteosarcoma cohort ($q < 0.00001$). The p53 pathway is regarded as the main pathway involved in osteosarcoma development and the tumour suppressor gene *TP53* has previously been established as an SMG in osteosarcoma (Chen X *et al.* 2014, Perry *et al.* 2014). In this dataset, SPIMs in *TP53* were found in 32 % of osteosarcoma samples (Figure 2). *TP53* was also affected

by CNAs in 32 % of osteosarcoma samples. By extracting coding, non-silent mutations and counting their occurrence, we found two FMGs in addition to *TP53*. These were *TTN* and *RB1*. *TTN* (mutated in 21 % of samples) encodes one of the largest proteins in the human genome which means it is also prone to accumulate a larger absolute number of somatic mutations during cancer development. Its importance in cancer is dubious, despite its relatively large number of SPIMs (Lawrence *et al.* 2013). The transcription factor gene *RB1*, on the other hand, is another tumour suppressor gene with a well-established role in osteosarcoma (Chen X *et al.* 2014, Perry *et al.* 2014). Our osteosarcoma cohort showed *RB1* mutations in 11 % of samples and CNAs in 45 %.

A total number of 67 additional genes were found to be recurrently mutated. The total set of 70 genes including SMGs, FMGs and RMGs will be termed RMGs below (Appendix Figure 2). Several known cancer genes were found among the RMGs, eight of which were CGC genes: *TP53*, *RB1*, *LRP1B*, *ATRX*, *PIK3CA* and *PTPN13* are CGC Tier 1 and *MUC16* and *CSDM3* are Tier 2 genes (Sondka *et al.* 2018). All RMGs except *ATRX*, *CXorf30* and *TRIM50* showed CNAs in at least one sample.

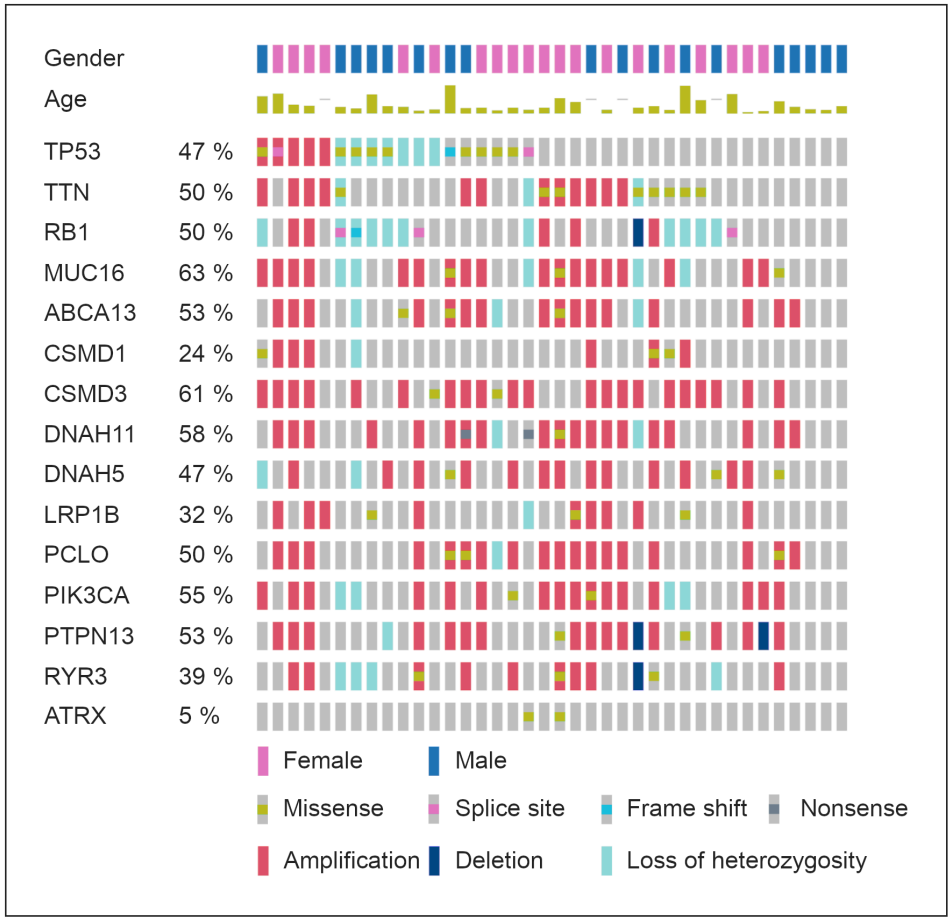



**Figure 2 | Recurrently mutated genes (RMGs) in osteosarcoma.**
**Genes that are recurrently mutated and show mutations in two or more samples of the osteosarcoma cohort were reported and copy number alterations (CNAs) in those genes analysed. The minimum age of the osteosarcoma cohort is 4 and the maximum age 85. The age of four samples was unknown (-). Several established osteosarcoma and cancer genes were discovered in the set.**

In the RMG set, 43 pathways were significantly enriched. Several pathways involved in cancer (head and neck squamous cell carcinoma, endometrial cancer and non-small cell lung cancer) and cancer-related processes were among the ten most significantly enriched pathways (Table 1). Most pathways centred around the osteosarcoma genes *TP53* and *RB1*. *TP53* was found in seven, *RB1* in five of these pathways. The ARF pathway was the most significantly enriched pathway in the set. RMGs found in the overlap with the BioCarta ARF pathway gene set were *TP53*, *RB1* and *PIK3CA*. It is an important cancer pathway which is involved in regulating p53 and osteosarcoma development as well as bone remodelling (Palmero *et al.* 1998, Rauch *et al.* 2010). The WNT signalling pathway was significantly enriched as well and plays a crucial role in bone development and osteosarcoma (Hill *et al.* 2005, Matsuoka *et al.* 2020). It was represented by the RMGs *TP53*, *ROCK2*, *APC2* and *DKK2*. The transcription regulator BTG2 suppresses osteosarcoma growth and has previously been linked to *TP53* (Rouault *et al.* 1996, Li *et al.* 2015); the BTG2 pathway was represented by *TP53* and *RB1*. *CACNA1A* and *CACNA1E* were found in two significantly enriched pathways involved in calcium regulation, calcium regulation in the cardiac cell and, together with *RYR2* and *RYR3*, presynaptic depolarization and calcium channel opening.

**Table 1 | Pathway enrichment analysis of recurrently mutated genes (RMGs).**
**Pathway enrichment analysis was performed with 70 recurrently mutated genes of the osteosarcoma cohort and the ten most significantly enriched pathways were examined. Several pathways related to cancer were enriched in the RMG set. Shown are the number of genes in the RMG-pathway gene set overlap, *q*-value for pathway enrichment and genes found in the overlap.**

| Pathway | Genes in overlap | q-value | Genes |
|---|---|---|---|
| BioCarta ARF pathway | 3 | 0.0098 | *TP53, RB1, PIK3CA* |
| WikiPathways head and neck squamous cell carcinoma | 4 | 0.0143 | *TP53, RB1, PIK3CA, CSMD3* |
| WikiPathways calcium regulation in the cardiac cell | 4 | 0.0334 | *CACNA1A, CACNA1E, RYR3, RYR2* |
| KEGG WNT signalling pathway | 4 | 0.0334 | *TP53, ROCK2, APC2, DKK2* |
| WikiPathways spinal cord injury | 4 | 0.0334 | *TP53, RB1, ROCK2, MAG* |
| KEGG non-small cell lung cancer | 3 | 0.0334 | *TP53, RB1, PIK3CA* |
| KEGG endometrial cancer | 3 | 0.0334 | *TP53, PIK3CA, APC2* |
| KEGG type II diabetes mellitus | 3 | 0.0334 | *PIK3CA, CACNA1A, CACNA1E* |
| Reactome presynaptic depolarisation and calcium channel opening | 2 | 0.0334 | *CACNA1A, CACNA1E* |
| BioCarta BTG2 pathway | 2 | 0.0334 | *TP53, RB1* |

## 3.3 9 % of non-coding mutations are found in constrained elements

In the osteosarcoma cohort, 99 % of SPIMs were found in non-coding regions. We annotated them with evolutionary constraint scores generated with GERP to identify SPIMs in functional elements. The distribution of GERP scores for non-coding SPIMs in osteosarcoma showed a clear peak around a score of zero (Figure 3): 78 % of non-coding SPIMs had a GERP score

between ± 2.0, 58 % of non-coding SPIMs had a score between ± 1.0 and 42 % had a score between ± 0.5. Considering a GERP score threshold of ≥ 2.0, 9 % of non-coding SPIMs were found in sites under evolutionary constraint and were therefore termed non-coding constraint mutations (NCCMs). Of these NCCMs, 44 % resided in intergenic regions, 41 % in introns and 7 % in lincRNAs. The remaining NCCMs were found in 3'UTRs (2 %), 5'UTRs (0.7 %), 5' flanking regions (2 %), and other non-coding RNAs (4 %). 13 NCCMs could not be assigned to either of these categories.

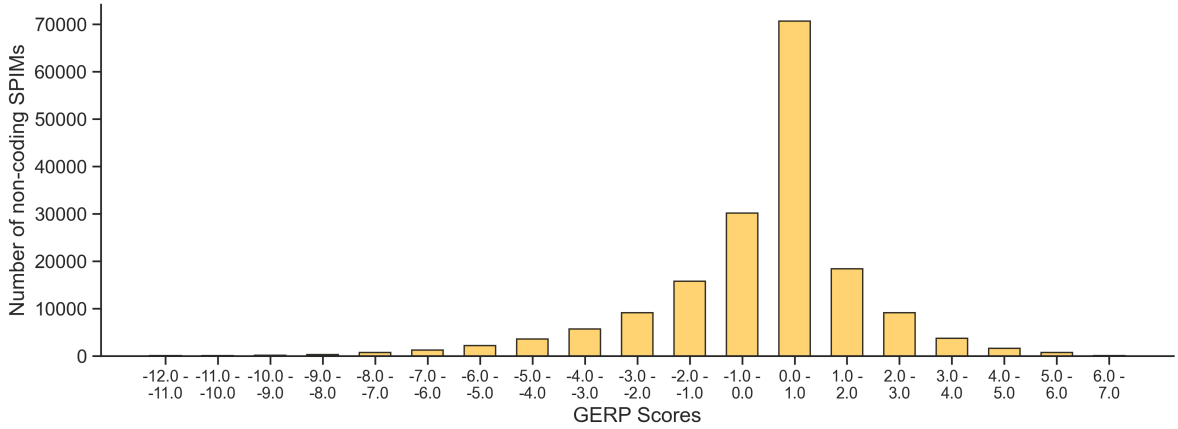



**Figure 3 | GERP score distribution of osteosarcoma non-coding mutations.**
**Somatic osteosarcoma mutations were annotated with GERP scores to identify sites under evolutionary constraint. The GERP distribution of non-coding mutations shows a peak around the score zero. 9% of positions of non-coding mutations have a GERP score ≥ 2 and are therefore considered to be constrained.**

## 3.4 NCCM enrichment around cancer genes

The next step was to identify genes with an enrichment of NCCMs in surrounding non-coding regions (Figure 4 a), as a high number of NCCMs in the regulatory regions of a gene could indicate that they are involved in the regulation of that gene in osteosarcoma.

The majority of genes (79.3 %) were associated with less than 0.5 NCCMs/100 Kbp and 99.7 % of genes had less than 2.0 NCCMs/100 Kbp (Figure 4 b). This left around 0.3 % of genes ($n$ = 64) with ≥ 2.0 NCCMs/100 Kbp at the tail end of the distribution. These genes were considered to have an enrichment of NCCMs and were investigated in more detail (Appendix Table 1). Because of overlapping flanking regions, some NCCMs were assigned to multiple genes. Of the 64 genes with an enrichment of NCCMs, 25 genes shared some or all their associated NCCMs with one or more genes.

As a control, the same analysis was carried out using all other bone tumour (OBT) samples from the ICGC bone tumour cohort excluding osteosarcoma. Bone tumours in this set were either benign or considered to be less aggressive than osteosarcomas. The GERP score distribution of non-coding OBT SPIMs was comparable to the distribution of non-coding osteosarcoma SPIMs (Appendix Figure 3). However, all genes fell below the threshold of ≥ 2 NCCMs/100 Kbp (Figure 4 b). 97.9 % of genes were associated with less than 0.5 NCCMs/100 Kbp, 1.9 % had 0.5 or more but less than 1.0 NCCMs/100 Kbp and 0.1 % ($n$ = 28) had 1.0 or more and less than 1.5 NCCMs/100 Kbp.

The 64 NCCM-enriched genes included genes related to cancer, osteosarcoma and bone development including a large number of transcription factors (Appendix Table 1). The top five

genes with the highest NCCM rate were *SOX2*, *NKX2-8*, *IZUMO3*, *NR4A2* and *COL1A2*. Four genes were CGC Tier 1 genes (*SOX2*, *NKX2-1*, *BCL11A* and *PREX2*).

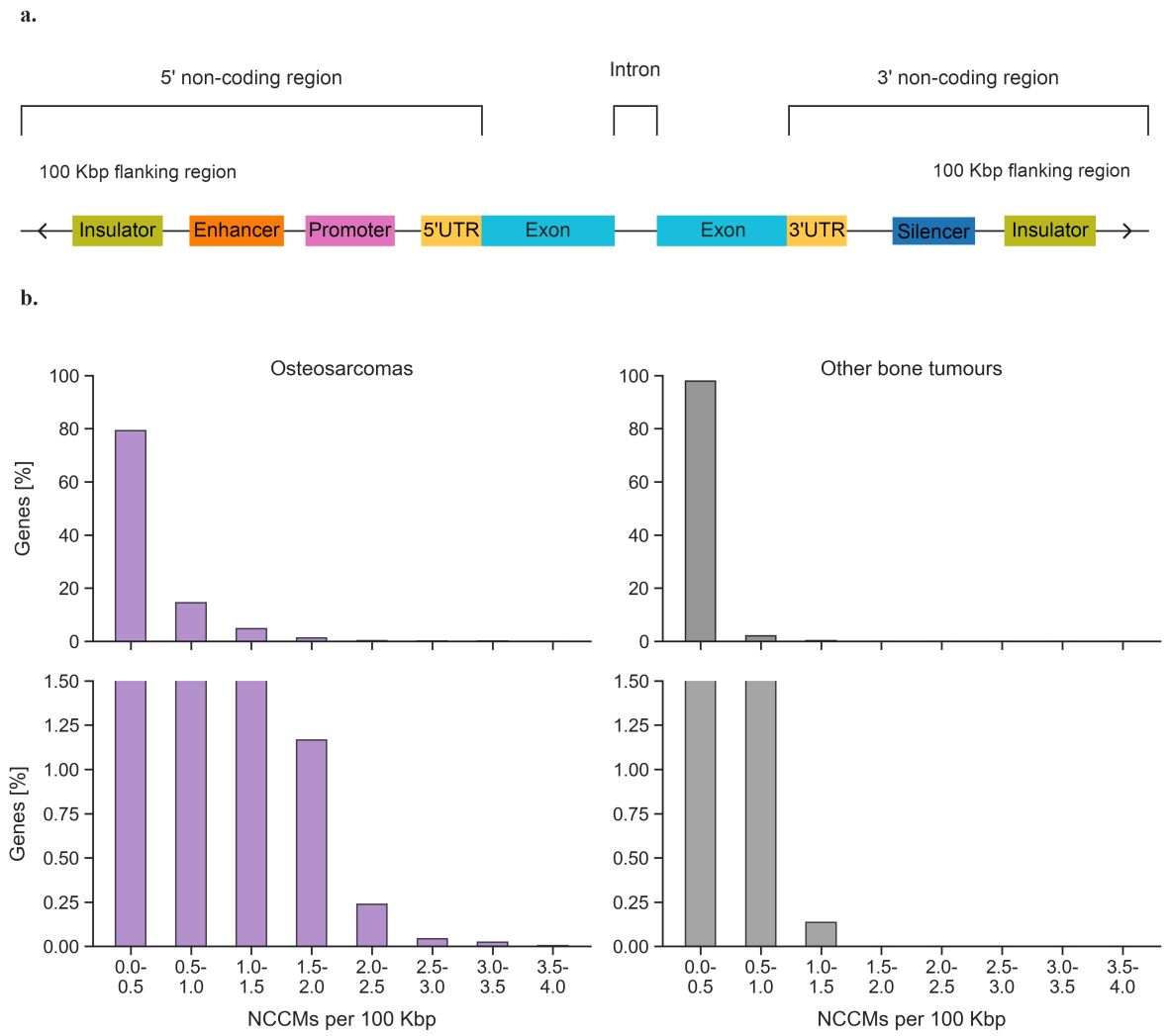



**Figure 4 | Enrichment of non-coding constraint mutations.**
**a. Enrichment of NCCMs in the non-coding space associated with genes was analysed. The non-coding regions that were considered are introns, 5' and 3' UTRs and the 100 Kbp flanking regions of each gene.**
**b. Distribution of the number of non-coding mutations associated with genes. Up: Distribution of the percentage of genes of the human genome and number of associated NCCMs for osteosarcoma and other bone tumours. Down: Magnified view of the tail ends of the distributions.**

The ten most significantly enriched GOs in the NCCM-enriched gene set were dominated by GOs involved in transcriptional regulation (Table 2). Fourteen genes were connected to each transcription regulator activities and sequence-specific DNA binding, 13 genes were associated with DNA binding transcription factor activities and 12 were related to chromatin. Another group of GOs among the ten most significantly enriched pathways were related to four gap junction genes (*GJA4*, *GJB4*, *GJB5*, *GJB3*). Gap junction genes are tumour suppressors that facilitate intracellular communication and are associated with bone development (Batra *et al.* 2012, Talbot *et al.* 2015). GOs that were solely represented by these four GJ genes were connexin complex, gap junction channel activity, wide pore channel activity and gap junction. Together with nine other genes they were found in the overlap of the membrane protein complex GO and with four other genes in passive transmembrane transporter activity.

**Table 2 | Gene Ontology (GO) enrichment analysis of NCCM genes.**
**GO enrichment analysis was performed with 64 genes that show an NCCM enrichment in the osteosarcoma cohort and the GO cellular component (GOCC), GO molecular function (GOMF) and GO biological process (GOBP) gene sets. The ten most significantly enriched GOs were examined. GOs related to transcription regulation and gap junctions are significantly enriched. Shown are the number of genes in the NCCM gene-GO overlap, *q*-value for GO enrichment and genes found in the overlap.**

| Gene Ontology | Genes in overlap | q-value | Genes |
|---|---|---|---|
| GOCC membrane protein complex | 13 | 0.0001 | *GJA4, GJB4, GJB5, GJB3, GRIN3A, KCNS2, TRPC3, GABRA4, CD40, ITGB4, CDH7, CPLX2, LAMTOR5* |
| GOMF gap junction channel activity | 4 | 0.0001 | *GJA4, GJB4, GJB5, GJB3* |
| GOCC connexin complex | 4 | 0.0001 | *GJA4, GJB4, GJB5, GJB3* |
| GOMF sequence specific DNA binding | 14 | 0.0001 | *NKX2-1, SOX2, NR4A2, POU3F3, TFAP2B, ALX1, NKX2-8, TFAP2D, FOXG1, SP8, SIM1, NR2F2, BCL11A, ORC5* |
| GOMF DNA binding transcription factor activity | 13 | 0.0001 | *NKX2-1, SOX2, NR4A2, POU3F3, TFAP2B, ALX1, NKX2-8, TFAP2D, FOXG1, SP8, SIM1, NR2F2, BCL11A* |
| GOCC chromatin | 12 | 0.0001 | *NKX2-1, SOX2, NR4A2, POU3F3, TFAP2B, ALX1, NKX2-8, TFAP2D, FOXG1, SP8, SIM1, ORC5* |
| GOMF wide pore channel activity | 4 | 0.0001 | *GJA4, GJB4, GJB5, GJB3* |
| GOCC gap junction | 4 | 0.0001 | *GJA4, GJB4, GJB5, GJB3* |
| GOMF transcription regulator activity | 14 | 0.0002 | *NKX2-1, SOX2, NR4A2, POU3F3, TFAP2B, ALX1, NKX2-8, TFAP2D, FOXG1, SP8, SIM1, NR2F2, BCL11A, VGLL1* |
| GOMF passive transmembrane transporter activity | 8 | 0.0002 | *GJA4, GJB4, GJB5, GJB3, GRIN3A, KCNS2, TRPC3, GABRA4* |

Gene expression differences between osteosarcomas and osteoblasts were analysed for 51 NCCM-enriched genes (no expression data was available for lncRNAs and *IZUMO3*) to further investigate their importance in cancer (Appendix Figure 4). Twelve genes showed significantly different expression between osteosarcomas and normal human osteoblasts (Student's t-test, $p \leq 0.05$). *CD40*, *CD84*, *GJA4*, *ITGB4*, *PREX2*, *SOX2* and *TFAP2D* expression was significantly higher while *SMIM12* (*C1ORF212*), *DPH5*, *HBXIP*, *MRPL19* and *SLC16A4* expression was significantly lower in osteosarcoma than in osteoblasts.

### 3.4.1 *SOX2* has the highest enrichment of NCCMs

*SOX2* was the gene with the highest NCCM rate (4.0 NCCMs/100 Kbp) and its expression was significantly higher in osteosarcomas than in osteoblasts (Appendix Table 1). It is a stem cell transcription factor that is involved in several important processes such as for example embryonic development and plays a crucial role in osteoblast self-renewal and proliferation

(Yuan *et al.* 1995, Basu-Roy *et al.* 2010). *SOX2* is also required for osteosarcoma development and proliferation (Maurizi *et al.* 2018).

In total, there were nine *SOX2* NCCMs found in eight different samples (Figure 5). Seven of these NCCMs were located in intergenic regions, one in the 5' flanking region and one in a noncoding RNA. To evaluate the regulatory potential of sites hit by NCCMs in detail, we intersected them with regulatory annotation information compiled from several sources. We found that one NCCM was located in the *SOX2* promoter region and three overlapped with DNase I hypersensitive sites which mark accessible chromatin and therefore potentially active regulatory regions (Frankish *et al.* 2019). Histone modifications regulate chromatin structure and thus influence the activation or inactivation of genes. Two NCCMs showed H3K4Me3 marks which are often found near promoters, six NCCMs were associated with H3K4Me1 marks which are found near regulatory elements and two NCCMs had an H3K27Ac mark which is a signature of a transcriptionally active regulatory region. Additionally, four NCCMs had transcription factor binding site annotations derived from ChIP-seq. In total, seven of the nine *SOX2* NCCMs had an additional regulatory annotation.

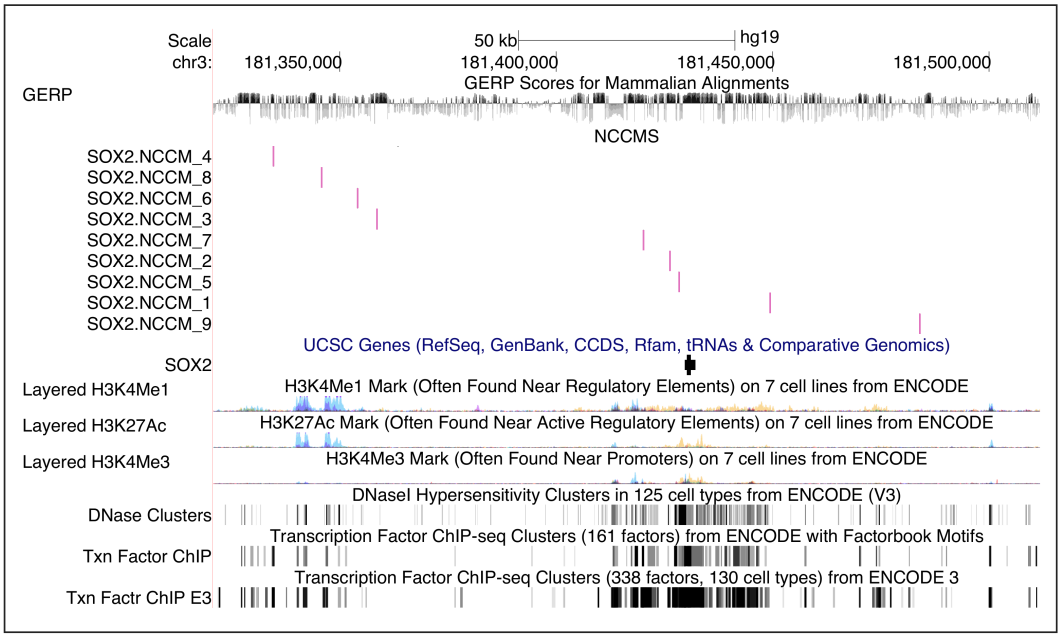



**Figure 5 | UCSC Genome Browser view of *SOX2* NCCMs.**
***SOX2* had the highest NCCM rate in the osteosarcoma cohort with nine NCCMs in associated non-coding regions.**

At the position of *SOX2*_NCCM_7, a transcription factor ChIP-seq cluster identified a CTCF binding site. CTCF is a zinc finger protein involved in genome regulation that often acts as an insulator, spatially separating genes from each other or from nearby enhancers (Phillips & Corces 2009). *SOX2*_NCCM_7 had a high GERP score of 4.9 and a variant allele fraction of 0.35. To test in silico whether *SOX2*_NCCM_7 could affect CTCF binding, the sTRAP module from the Transcription factor Affinity Prediction (TRAP) Web Tools, which predicts differences in transcription factor binding affinity between two sequences, was applied (Thomas-Chollier *et al.* 2011). Indeed, the wild-type sequence was predicted to have significant affinity for CTCF binding at this position, while the NCCM in the mutant sequences caused a strong decrease in binding affinity.

### 3.4.2 *BCL11A* NCCMs have strong regulatory potential

*BCL11A* is a transcription factor which is most commonly known for its role in lymphoid tumours. For example, high *BCL11A* expression in natural killer/T-cell lymphomas promotes tumour development and is correlated with poor clinical outcomes (Satterwhite *et al.* 2001, Shi *et al.* 2020). It is a CGC Tier 1 gene, which can interact with *SOX2* in cancer development (Lazarus *et al.* 2018). *BCL11A* expression was not significantly different in our small gene expression cohort, but a trend towards overexpression of *BCL11A* in osteosarcoma was observed (Figure 6 a).

We found 2.7 NCCMs/100 Kbp in the non-coding regions associated with *BCL11A*. This corresponds to an absolute number of 11 NCCMs in eight different samples (Appendix Table 1). Nine of these NCCMs were found in introns, while one was located in an intergenic region and one in a lincRNA. Moreover, there was an abundance of regulatory annotations in NCCM sites (Figure 6 b). Seven NCCMs were located in DNase I hypersensitive sites and four were in ChIP-seq transcription factor binding sites. All 11 NCCMs were in regions with H3K4Me1 marks which indicates that they were in or in close proximity to regulatory elements. Eight NCCMs were in sites that show H3K27Ac marks and four in sites with H3K4Me3 marks.

We performed sTRAP analysis for all NCCMs associated with *BCL11A*. HNF1B and PRRX2, which are transcription factors in the WNT signalling pathway, had significant binding affinity predictions for the wildtype sequence at and around the sites of *BCL11A*_NCCM_3 and *BCL11A*_NCCM_9, respectively. This was interesting, because *BCL11A* activates WNT signalling in breast cancer (Zhu *et al.* 2019). In both cases, the NCCM lead to a predicted decrease in binding affinity. *BCL11A*_NCCM_9 is located in a highly conserved position of the PRRX2 transcription factor binding site (Figure 6 c). We performed an EMSA which showed that the wild-type transcription factor binding sequence bound a protein, presumably PRRX2, while the mutant sequence with the *BCL11A*_NCCM_9 variant likely did not (Figure 6 d). However, the results of this experiment were slightly ambiguous and will therefore be confirmed via replication in the future.

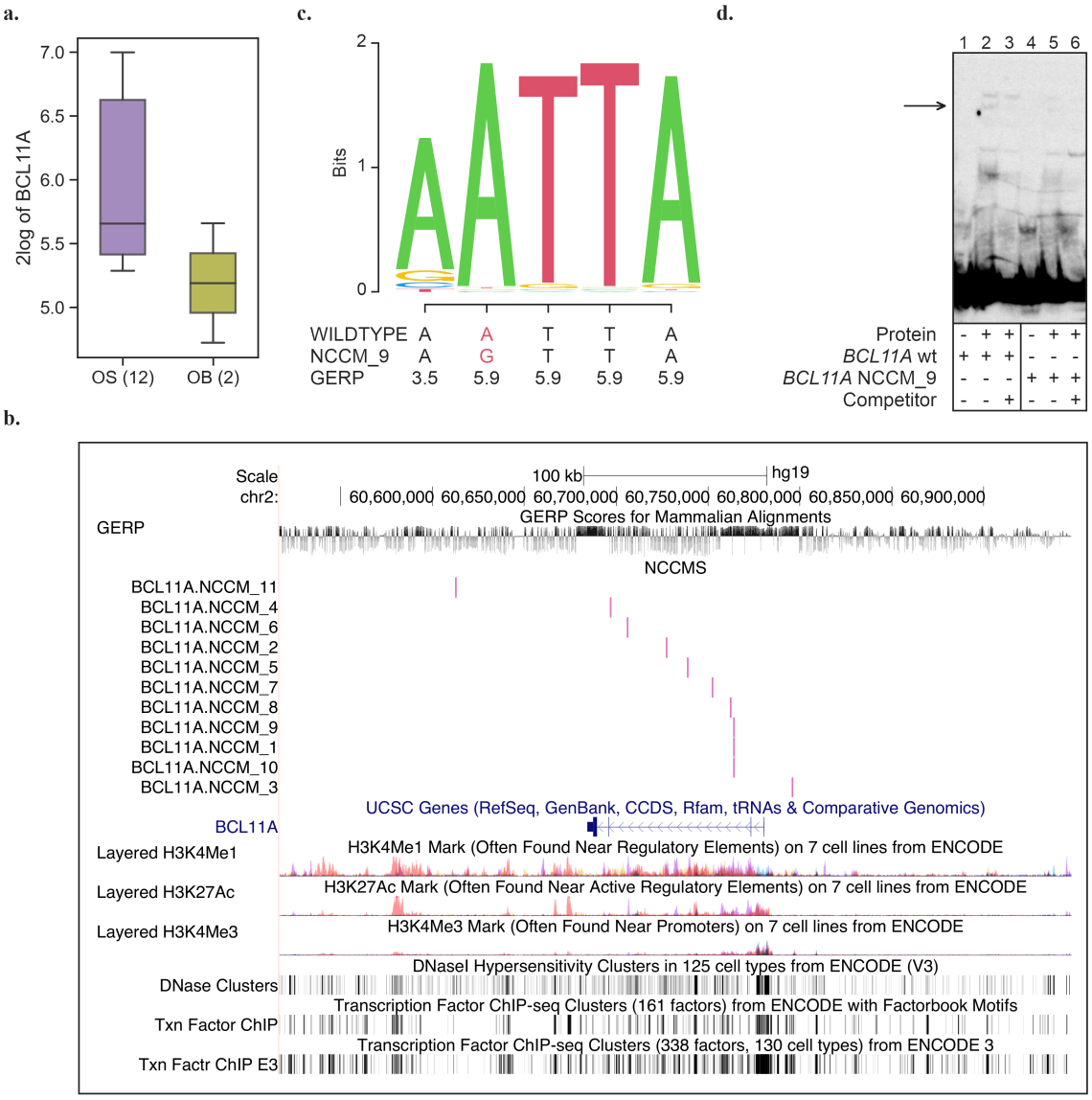



**Figure 6 | Non-coding constraint mutations analysis of *BCL11A*.**
**a. Normalised gene expression of *BCL11A* in osteosarcomas (OS) and osteoblasts (OB).**
**b. UCSC Genome Browser view of *BCL11A* NCCMs. *BCL11A* NCCMs were frequently found in regulatory elements.**
**c. PRRX2 Jaspar matrix. *BCL11A*_NCCM_9 affects a conserved site of the PRRX2 binding site.**
**d. Electrophoretic mobility shift assay (EMSA) comparing DNA-protein interaction of *BCL11A* wild-type (wt) and NCCM_9 mutant sequence. DNA binding of nuclear protein from Saos-2 osteosarcoma cell line to the predicted PRRX2 binding site in the *BCL11A* wild-type sequence (lanes 2-3) and the NCCM_9 mutant sequence (lanes 5-6) was tested. In lane 2, a shift was present for the wild-type sequence that was competed out in lane 3. This shift was weaker for the NCCM_9 mutant sequence in lane 5.**

### 3.4.3 *NR4A2* and *NKX2-1* NCCMs are mutually exclusive with *TP53* alterations

*NR4A2* and *NKX2-1* were among the top ten genes with the highest NCCMs/100 Kbp rate and are both known to regulate p53 (Zhang T *et al.* 2009, Chen PM *et al.* 2015).

*NR4A2* had a rate of 3.2 NCCMs/100 Kbp with a total number of seven NCCMs in seven different samples (Appendix Table 1). All seven NCCMs were located in intergenic regions,

two were in DNase I hypersensitive sites and two overlapped with ChIP-seq transcription factor binding sites. The H3K4Me1 histone mark that indicates proximity to regulatory elements was found in five NCCM sites.

*NKX2-1* had a rate of 3.0 NCCMs/100 Kbp and six NCCMs in seven different samples. All seven NCCMs were shared with *NKX2-8* which lies in close proximity to *NKX2-1* and therefore has an overlapping flanking region. Six NCCMs were also shared with *SFTA3*. Here we focused on *NKX2-1*, but NCCMs could be involved in the regulation of all three genes. Four NCCMs associated with *NKX2-1* lay in intergenic and three in intronic regions. Five NCCMs were in DNase I hypersensitive sites, five were annotated as ChIP-seq derived transcription factor binding sites and four had H3K4Me1 marks.

Because *NR4A2* and *NKX2-1* can regulate p53, we were interested in the distribution of NCCMs in those genes across samples compared to *TP53* mutations. *TP53* had protein-coding mutations in 12 samples and CNAs in 12 samples, but showed neither one in 20 samples. Five *NR4A2* NCCMs were found in samples with no *TP53* coding mutation and four *NR4A2* NCCMs in samples with neither a *TP53* coding mutation nor CNA (Figure 7). Three *NKX2-1* NCCMs were in samples that have neither a *TP53* coding mutation nor CNA. Together, seven *NR4A2* or *NKX2-1* NCCMs were in samples with no *TP53* coding mutation, six NCCMs were in samples with no *TP53* coding mutation or CNA. Considering CNA and NCCMs in *NR4A2* and *NKX2-1* as well as coding mutations and CNA in *TP53*, 84 % of osteosarcoma samples have mutations that could alter the p53 pathway. We tested mutual exclusivity of all *NR4A2* and *NKX2-1* alterations with *TP53* alterations as well as specifically of NCCMs in *NR4A2* and *NKX2-1* with *TP53* alterations and both results were significant ($q$ = 0.043 and $q$ = 0.003, respectively).

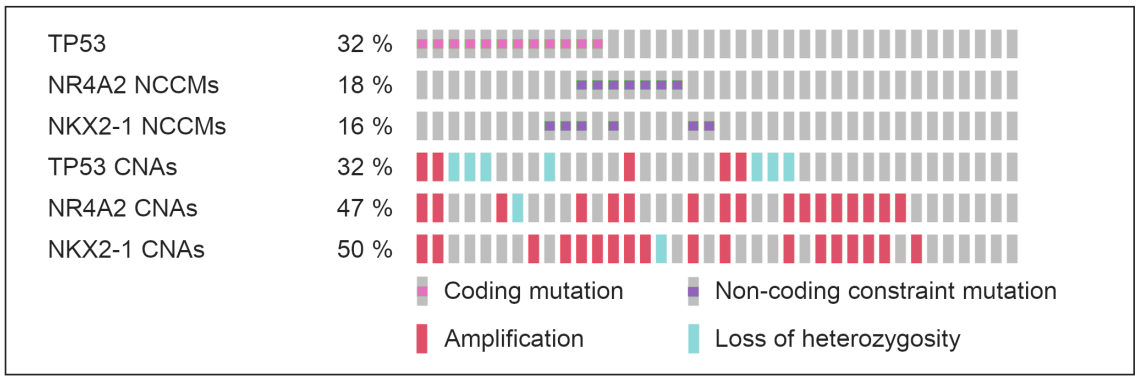



**Figure 7 | Comparison of *NR4A2*, *NKX2-1*, and *TP53* alterations per sample.**
***TP53* alterations are mutually exclusive with both *NR4A2* and *NKX2-1* alterations and *NR4A2* and *NKX2-1* NCCMs only.**

### 3.4.4 Gap junction genes share most NCCMs

Gap junction (GJ) proteins are responsible for intercellular communication and are often described as tumour suppressor genes (Talbot *et al.* 2015). They are also associated with bone development (Batra *et al.* 2012).

Looking at the gene expression of the four GJ genes of our NCCM-enriched gene set in osteosarcoma and osteoblasts revealed differences between them (Figure 8 a). *GJA4* was significantly overexpressed in osteosarcoma compared to osteoblasts. *GJB4*, *GJB3* and *GJB5*

expression was not significantly different in osteosarcoma than in osteoblasts, but a trend towards lower expression in osteosarcoma than in osteoblasts was observed.

In our dataset, four genes encoding for gap junction proteins had ≥ 2 NCCMs/100 Kbp (*GJA4*, *GJB3*, *GJB5* and *GJB4;* Appendix Table 1). These are rather short genes that are situated closely together. This means that their 100 Kbp flanking regions overlapped and most NCCMs found in their surroundings were shared between them. Some NCCMs were additionally shared with *SMIM12*. *GJA4* and *GJB3* had six NCCMs in six different samples each, while *GJB5* and *GJB4* had five NCCMs in five samples each. Collectively, these NCCMs were made up of seven unique NCCMs which will be referred to as *GJ*_NCCM_1 to *GJ*_NCCM_7 (Figure 8 b). Four NCCMs were found in the intergenic regions upstream of this *GJ* locus, and three downstream with one NCCM in the intergenic region and two in *DLGAP3* introns. Three NCCMs were in DNase I hypersensitive sites and one in a ChIP-seq transcription factor binding site. Moreover, three NCCMs had H3K4Me1 marks and one had an H3K27Ac mark.

Among other results, sTRAP predicted a decrease in binding affinity for ZNF354C caused by *GJ*_NCCM_6 (Figure 8 c). ZNF354C is a transcription factor with known roles in bone development. *GJ*_NCCM_6 resided in the flanking regions of all four GJ genes and could therefore have an effect on each of them. TAD domain analysis was performed with 3DIV to find out which of the four gap junction genes might be regulated by this transcription factor binding site and the NCCM. The analysis showed that the upstream NCCMs (*GJ*_NCCM_3, *GJ*_NCCM_4, *GJ*_NCCM_6 and *GJ*_NCCM7) were in the same TAD as *GJB4* and *GJB5*, and the downstream NCCMs (*GJ*_NCCM_1, *GJ*_NCCM_2 and *GJ*_NCCM_5) shared TAD with *GJA4* and *GJB3* (Figure 8 d). Since *GJ*_NCCM_6 as well as the ZNF354C transcription factor binding site shared TAD with only *GJB5* and *GJB4*, these were the likely targets of this regulatory region.

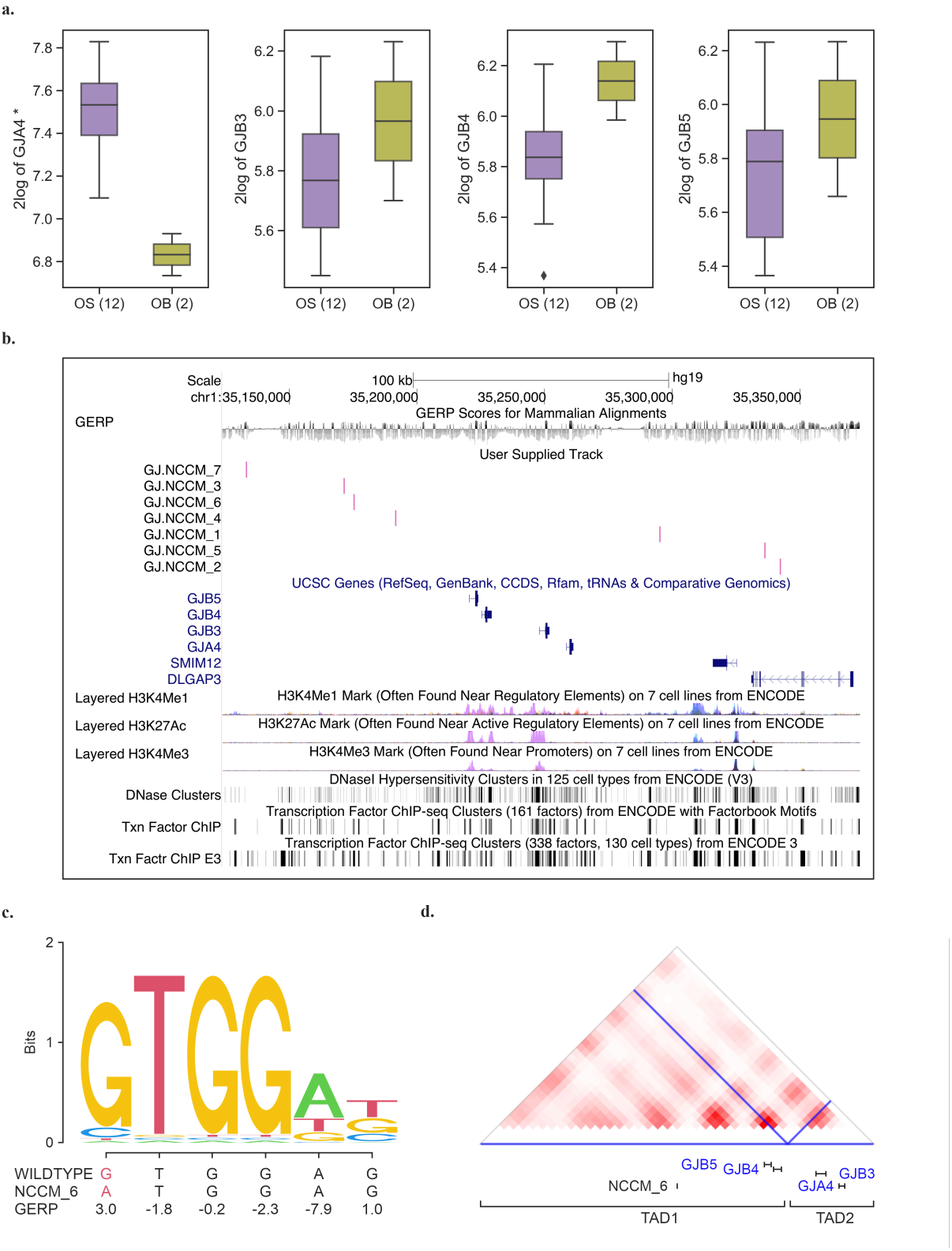



**Figure 8 | Gap junction (GJ) NCCM analysis.**
**a. Normalized expression comparison of *GJ* genes between osteosarcomas (OS) and osteoblasts (OB) showed different patterns. Gene expression was significantly different (*) in osteosarcomas and osteoblasts for *GJA4*.**
**b. UCSC Genome Browser view of *GJ* NCCMs. *GJ* NCCMs lie in the flanking regions of four *GJ* genes.**
**c. ZNF354C Jaspar matrix. *GJ*_NCCM_6 affects a conserved site of the ZNF354C binding site.**
**d. TAD analysis of the *GJ* locus. *GJB5* and *GJB4* share TAD1 with *GJ*_NCCM_6.**

# 4 Discussion

In our attempt to shine a light on the role of non-coding constraint mutations in osteosarcoma, we first analysed SPIMs in protein-coding regions and compared the results with findings from other studies.

*TP53* was significantly mutated in our osteosarcoma cohort. Mutations in the tumour suppressor gene play a crucial role in osteosarcoma tumorigenesis and *TP53* inactivation has been suggested to be involved in the structural instability of the genome that is a predominant characteristic of osteosarcoma (Perry *et al.* 2014). *TP53* has previously been identified as an osteosarcoma SMG by several studies (Chen X *et al.* 2014, Perry *et al.* 2014). The tumour suppressor gene *RB1* was frequently mutated in our osteosarcoma cohort. Studies have found that more than 50 % of osteosarcoma patients have mutations in both *TP53* and *RB1* and shown that mutations in *TP53* and *RB1* can synergistically accelerate tumorigenesis and thereby drive osteosarcoma development (Walkley *et al.* 2008, Perry *et al.* 2014). *ATRX* is a transcriptional regulator that has previously been reported as recurrently mutated in osteosarcoma and showed mutations in two samples of our cohort (Chen X *et al.* 2014). In addition to *TP53*, *RB1* and *ATRX*, several cancer and osteosarcoma genes were recurrently mutated. For example, *MUC16* is a tumour antigen, which is involved in tumour growth and metastasis in various cancers (Thériault *et al.* 2011, Chen S *et al.* 2012, Lakshmanan *et al.* 2012). *APC2* can inhibit osteosarcoma progression by acting as a negative regulator of the WNT signalling pathway (Wu *et al.* 2018). ABC transporters are involved in cancer in various ways and *ABCA13* has been linked to poor survival in metastatic ovarian serous carcinoma (Nymoen *et al.* 2015).

Pathway enrichment analysis revealed an enrichment of cancer pathways in the RMG set. Most centred once again around *TP53* and *RB1,* which are included in seven and five pathways, respectively. This emphasized the importance of these two genes in osteosarcoma. However, *TP53* and *RB1* are among the most well studied genes in cancer biology and are therefore more likely to have documented roles in a larger number of pathways. Pathway enrichment analyses are based on defined gene sets and databases, which can introduce bias towards well known pathways. WNT signalling was a highly relevant result, because it is supported by *TP53* as well as three other genes (*APC2*, *ROCK2* and *DKK2*). This pathway is not only involved in bone development and formation by regulating osteoblast lineage determination and differentiation (Glass *et al.* 2005, Hill *et al.* 2005). Aberrant activation of WNT signalling also promotes osteosarcoma proliferation and has been correlated with aggressive behaviour and decreased patient survival (Chen C *et al.* 2015, Matsuoka *et al.* 2020).

Overall, results from the analysis of recurrent protein-coding mutations in our osteosarcoma cohort and the pathways they act on coincide with previous findings, which establishes the osteosarcoma samples from our cohort as representative for the disease.

## 4.1 NCCM analysis yields results relevant for osteosarcoma

Non-coding mutations can be responsible for alterations of cellular function and gene regulation (Fredriksson *et al.* 2014). To identify somatic mutations that target functional elements with regulatory potential, we mapped evolutionary constraint scores to the set of somatic variants and extracted NCCMs in putative regulatory regions.

Between 6 and 13 % of the human genome are estimated to be under evolutionary constraint. Based on a moderate estimate of around 8 %, we used a threshold of GERP ≥ 2.0 to identify somatic mutations under evolutionary constraint in the osteosarcoma cohort (Rands *et al.* 2014).

Most variants that have an effect on the gene regulatory system are found in cis-acting eQTLs which are located in the 1 Mbp flanking regions of their target gene (Wittkopp & Kalay 2012, Battle *et al.* 2014, Albert & Kruglyak 2015, Sakthikumar *et al.* 2020). We decided to be more stringent and extracted NCCMs in introns, UTRs and the 100 Kbp flanking regions of all human genes to identify those with an enrichment of mutations with a potential functional impact.

The osteosarcoma NCCM enrichment analysis revealed that 0.3 % of genes had an enrichment of NCCMs in their associated regulatory regions. No genes were enriched with NCCMs in the OBT cohort where the majority of genes had less than 0.5 NCCMs/100 Kbp. This result is at least partly linked to a lower mutation rate in these benign or less aggressive bone tumours in general. It does, however, also indicate that genes with an NCCM enrichment in osteosarcoma are relevant to the cancer and suggests that NCCMs are an efficient way to identify mutations and genes with roles in osteosarcoma.

## 4.2 NCCMs are involved in important osteosarcoma pathways

Characterisation of NCCM-enriched genes revealed their involvement in several cancer and osteosarcoma processes.

A large number of genes with an enrichment of NCCMs were transcription factors. Transcription factors are strongly involved in tumorigenesis and make up 19 % of all known oncogenes (Lambert *et al.* 2018). Methods targeting transcription factors for cancer treatment, especially drivers of immune invasion, epithelial-to-mesenchymal transition, resistance development, replicative immortality, differentiation and cell death as well as transcription factors supporting stem cell properties, are being widely explored and show promising results (Lambert *et al.* 2018, Bushweller 2019). Transcription factors like p53 and SOX2 play an important role in osteosarcoma and our results suggest that non-coding mutations are a crucial factor in the regulation of their transcription.

Among the top ten significantly enriched GOs in NCCM genes were several GOs related to gap junctions. For four of these, the overlap with the NCCM gene set solely consists of the four gap junction genes *GJA4*, *GJB3*, *GJB4* and *GJB5*. Gap junctions play a role in cancer and bone development, but the fact that they dominate the results of this enrichment analysis leads to a potential overestimation of their importance. *GJA4*, *GJB3*, *GJB4* and *GJB5* share seven unique NCCMs due to their close proximity to each other and overlapping flanking regions, which raises the question of how such cases should be evaluated. On the one hand, the example of this GO enrichment analysis shows that genes sharing the same NCCMs could lead to an artificial enrichment and overemphasis of a certain gene family. On the other hand, regulatory elements, and therefore the NCCMs that potentially influence them, can act on multiple genes in a large genomic region. Until functional investigations of NCCMs and all associated genes can be conducted to assess the role of individual genes in osteosarcoma, it should be assumed that they are equally important.

Gap junctions consist of transmembrane channels that facilitate intercellular communication. By coordinating signal transmission between bone cells, gap junctions have been implicated in the regulation of bone development, differentiation, bone modelling and remodelling (Batra *et al.* 2012, Talbot *et al.* 2015). Moreover, gap junction genes are downregulated and gap junctions and their way of intracellular communication suppressed in several cancers (Naus *et al.* 1991, Huang *et al.* 1999, Laird *et al.* 1999, Saito T *et al.* 2001). Loss of gap junctional communication in cancer is suggested to be beneficial to cancer proliferation because it disrupts control mechanisms that surrounding healthy cells exert on cancerous cells (Loewenstein 1979, Mehta

*et al.* 1999). In osteosarcoma cells, overexpression of the gap junction gene *GJA1* (connexin-43) suppresses proliferation (Zhang YW *et al.* 2001). We therefore suggest that NCCMs might play their part in the downregulation of gap junction genes in osteosarcoma to prohibit their tumour suppressive activities.

*SOX2* is highly expressed in osteosarcoma. Aberrant *SOX2* expression inhibits osteoblast differentiation and WNT signalling (Mansukhani *et al.* 2005). In cancer, *SOX2* is mainly known for its role in tumour initiating or cancer stem cells. Cancer stem cells are a subpopulation of cells that maintain a tumour, can self-renew indefinitely and are able to initiate new tumour growth when transplanted into an animal host. Depletion of *SOX2* in murine osteosarcoma cells inhibits differentiation into osteoblasts, reduces the cells' potential to form new tumours and impairs tumour invasion and metastasis. This suggests that a subpopulation of tumour-initiating cancer stem cells exists in osteosarcoma in which *SOX2* is required for self-renewal and tumorigenicity (Basu-Roy *et al.* 2012). Subsequently, it has been shown that *SOX2* is not only necessary in cancer stem cells but required for osteosarcoma development and progression. *SOX2* has therefore been suggested as a potentially very effective therapeutic target (Maurizi *et al.* 2018). *SOX2* showed a strong enrichment of NCCMs which could play a crucial role in regulating *SOX2* expression in osteosarcoma.

## 4.3 Novel osteosarcoma mechanisms are proposed based on NCCM analysis

*BCL11A* is a CGC Tier 1 gene with oncogenic effects in several cancers. In natural killer/T-cell lymphoma, *BCL11A* promotes tumour development and has been linked to poor clinical outcomes (Satterwhite *et al.* 2001, Shi *et al.* 2020). In lung squamous cell carcinoma, *BCL11A* expression is involved in tumour development and maintenance and interacts with *SOX2* to control epigenetic regulators such as *SETD8* (Lazarus *et al.* 2018). In breast cancer, *BCL11A* promotes tumour formation and progression. Moreover, *BCL11A* activates the WNT signalling pathway and by doing so is suspected to initiate breast cancer stem cell renewal and metastasis (Zhu *et al.* 2019). To our knowledge, *BCL11A* has not directly been associated with osteosarcoma, but its connection to *SOX2* and cancer stemness, its ability to regulate WNT signalling as well as a high number of NCCMs and elevated expression levels in osteosarcoma point towards a potential role in the disease.

The nuclear receptor gene *NR4A2* is highly expressed in cancer and supports cancer progression by mediating cell proliferation, differentiation and apoptosis (Ke *et al.* 2004). In bladder cancer, *NR4A2* expression has been correlated with increasing tumour stage, enhanced invasiveness and adverse outcomes and in squamous cell carcinoma, colorectal carcinoma and gastric *NR4A2* has been linked to chemo-resistance (Inamoto *et al.* 2010, Shigeishi *et al.* 2011, Han *et al.* 2013b, Han *et al.* 2013a). Furthermore, *NR4A2* is an important factor in osteoblast differentiation and interaction with the WNT signalling pathway (Lee *et al.* 2006, Rajalin & Aarnisalo 2011). In osteosarcoma, *NR4A2* might be targeted by *PTHR1* and increased *PTHR1* expression levels mediated by WNT signalling have been correlated with tumour aggressiveness and poor outcome (Guan & Tian 2017). The *NKX* gene family is involved in oncogenesis of several cancers and *NKX2-1* serves for example as a prognostic marker in early-stage non-small cell lung cancer and might act as a tumour suppressor in lung adenocarcinoma (Saito RA *et al.* 2009, Moisés *et al.* 2017).

Both NCCM genes *NR4A2* and *NKX2-1* had an enrichment of NCCMs and their documented roles in bone development and cancer make them interesting candidate genes for osteosarcoma.

What they have in common is their ability to regulate p53. *NKX2-1* directly regulates p53 transcription and *NR4A2* interferes with p53 self-assembly and suppresses p53-mediated transcriptional activity (Zhang T *et al.* 2009, Chen PM *et al.* 2015). Based on our results, we suggest that *NR4A2* and *NKX2-1* could influence osteosarcoma development. Further, NCCMs and the gene regulatory alterations they likely cause in *NR4A2* and *NKX2-1* could support the characteristic osteosarcoma phenotype in *TP53* wild-type patients.

## 4.4 NCCMs have strong potential to alter gene regulation

The basic premise behind this effort to investigate non-coding constraint mutations was that evolutionary constraint scores can be utilized to identify non-coding mutations that drive osteosarcoma. Our results suggest that non-coding constraint mutations in regulatory regions can alter transcription levels and thereby promote tumorigenesis. Many NCCMs were found in regulatory regions and had an abundance of regulatory annotations. Of the sTRAP analyses that were performed on interesting candidates, more than 65 % predicted alterations of transcription factor binding affinities caused by NCCMs.

*SOX2* NCCMs were frequently annotated with regulatory information which was supported by functional evidence. ChIP-seq had identified a CTCF binding site at the position of *SOX2*_NCCM_7 and sTRAP predicted a decrease in binding affinity for CTCF caused by the mutation. CTCF is often involved in insulation, preventing enhancers from acting on promoters, while clumping together regions that interact within DNA loops (Phillips & Corces 2009). Disruption of CTCF binding causes changes in chromatin interaction and gene expression and may act as a mechanism of tumorigenesis (Fang *et al.* 2020). It has been shown that CTCF deficiency can lead to *SOX2* upregulation, while CTCF overexpression can repress *SOX2* which is strong evidence that *SOX2* is regulated by CTCF (Wang *et al.* 2020). We therefore hypothesize that *SOX2*_NCCM_7 facilitates *SOX2* overexpression by disrupting CTCF binding and preventing insulation.

Four *GJ* genes were enriched with NCCMs in the osteosarcoma cohort. At the position of *GJ*_NCCM_6, sTRAP predicted significant binding affinity for ZNF354C in the wild-type sequence. The NCCM was predicted to decrease binding affinity for ZNF354C. Considering that gap junction proteins and ZNF354C are important agents of bone development, we consider it likely that the transcription factor regulates GJ gene transcription (Jheon *et al.* 2001). TAD analysis suggested that *GJB5* and *GJB4* are potential targets of this regulatory element. We speculate that the predicted disruption of a ZNF354C binding site caused by *GJ*_NCCM_6 might be involved in tumorigenesis by facilitating a downregulation of gap junction genes.

Regulatory annotation and functional evidence suggested that *BCL11A* NCCMs were found in active regulatory regions. sTRAP predicted significant binding affinity for the WNT signalling transcription factors HNF1B and PRRX2 in the wild-type sequence at the positions of *BCL11A*_NCCM_3 and *BCL11A*_NCCM_9, respectively (Lv *et al.* 2017, Chai *et al.* 2019, Chan *et al.* 2019). Both mutations were predicted to cause decreased binding affinity compared to the wild-type. This indicates that *BCL11A*_NCCM_3 and *BCL11A*_NCCM_9 weaken or hinder the regulatory effects HNF1B and PRRX2 have on gene transcription.

Using EMSA, we aimed to experimentally confirm that *BCL11A*_NCCM_9 could disrupt a PRRX2 binding site. The preliminary results of this EMSA suggested that the wild-type sequence around this position indeed bound a protein, while the *BCL11A*_NCCM_9 mutant sequence did not. However, they were not definitive. The experiment will be replicated with a higher amount of nuclear protein extract to produce a clearer signal. Additional shifts visible

on the gel that were probably caused by unspecific binding could be prevented by using shorter DNA probes. To confirm that the bound protein is truly PRRX2, the EMSA could be repeated using recombinant PRRX2 instead of nuclear protein extract. Another option would be to add a selective antibody to cause a supershift. Until then, the results of this EMSA did indicate that *BCL11A*_NCCM_9 disrupts a PRRX2 transcription factor binding site and that, by doing so, it can interfere with transcription factor binding at that position.

This showed that NCCMs can manipulate functions of the gene regulatory system and suggests that they are likely able to cause changes in transcription levels.

## 4.5 Limitations

We used GERP scores based on an alignment of 33 mammals to identify NCCMs. During the time-frame of this project, a novel set of PhyloP scores based on an alignment of 241 mammals became available. PhyloP uses a multiple sequence alignment to compute p-values for evolutionary conservation and acceleration (Pollard *et al.* 2010). Because these scores are based on the alignment of more species that are part of the same class as humans, they are considered to be more precise in estimating the level of evolutionary constraint than the set of GERP scores. For this reason, the analysis was repeated using PhyloP scores which resulted in 43 genes with an enrichment of NCCMs. Due to time limits for this project, we continued using mainly GERP results, but focused on the 13 genes that showed an enrichment of NCCMs in both GERP and PhyloP analyses (*SOX2*, *IZUMO3*, *AC016251.1*, *GJA4*, *GJB3*, *BCL11A*, *C14orf23*, *TAS2R1*, *GJB5*, *GJB4*, *PRDM13*, *CD40* and *RP11-298I3.5*).

The utilization of evolutionary constraint is a powerful method to identify functional elements and allows us to estimate the functional importance of single nucleotides in the genome. Evolutionary constraint scores provide a simple, universally applicable measure that does not depend on experimental factors such as the analysed tissue type or choice of experiment and conditions. One multiple sequence alignment enables us to estimate the position of functional elements independent of tissue and cell types across all species that are included in it.

At the same time, the simplicity of evolutionary constraint scores can be considered a disadvantage as they do not provide insight on the kind of functional element that is identified and which tissue or cell type that element is active in. The Encyclopedia of DNA Elements (ENCODE) Consortium has launched a multi-phase effort to characterize functional elements in the human genome by applying a variety of methods, such as DNA hypersensitivity assays, DNA methylation assays, ChIP-seq for histones and transcription factors and many more on a wide range of tissue and cell types (Birney *et al.* 2007). Large-scale empirical projects such as ENCODE provide the detailed information on functional elements that evolutionary constraint scores are lacking which makes them invaluable resources to understand the role of functional elements.

The comparison of evolutionary constraint and the functional annotations generated by ENCODE revealed that a fraction of the functional elements ENCODE has identified in the human genome was not constrained across mammalian species (Borneman *et al.* 2007, Odom *et al.* 2007, Schmidt *et al.* 2010, Dunham *et al.* 2012). This discrepancy seems to have two main causes. Some functional elements have evolved under lineage-specific selection and are therefore not conserved across a wide range of species. Other elements are likely neutral which means that they do have a function but their presence does not affect the fitness of a species (Dunham *et al.* 2012). This shows that evolutionary constraint cannot identify all functional elements. On the other hand, ENCODE has so far not managed to characterize all sites in the

human genome that are under evolutionary constraint as functional elements. Despite great efforts and more than 9,000 experiments, the ENCODE data collection is still incomplete because assays have not been performed on all cell types or for example investigating all transcription factors (Abascal *et al.* 2020).

In this work, we showed that evolutionary constraint is an appropriate and effective measure to identify somatic mutations in functional elements. However, it cannot stand alone and data from additional resources such as ENCODE is necessary to further characterise regions of interest.

## 4.6 Conclusion

In conclusion, we found that non-coding mutations likely play an important role in osteosarcoma. We showed that mapping evolutionary constraint scores to somatic variants is an efficient method to extract non-coding mutations with functional impact in osteosarcoma. Furthermore, we applied NCCM analysis to improve our understanding of known osteosarcoma mechanisms as well as to discover new candidate genes and pathways that drive osteosarcoma.

# 5 Acknowledgement

In the course of this project, I had the pleasure of getting to know members from both Kerstin Lindblad-Toh and Karin Forsberg Nilsson's groups from all of whom I have received invaluable support and learned a great deal.

First of all, I would like to thank my supervisor Kerstin Lindblad-Toh for taking me in and giving me the amazing opportunity that was this project. I am immensely grateful for her guidance and input during the project as well as her feedback on my report.

I would also like to express my deep gratitude to Karin Forsberg Nilsson, for welcoming me into her group, for teaching me so much of what I know now about cancer and her comments on the report.

During this project, Sharadha Sakthikumar has mentored me and patiently walked me through the NCCM workflow both of which I am very grateful for. Very special thanks go to Sharadha and Suvi Mäkeläinen for all their amazing support in the past months and our weekly bioinformatics meetings that were always fun and instructive.

I am grateful to Elisabeth Sundström and Åsa Karlsson for carrying out the EMSAs for this project in the lab and to Elisabeth for helping me understand the experiments.

I want to thank Aristidis Moustakas for agreeing to be my subject reader and giving helpful comments on my report. Thanks to Yiwen Chen for being my student opponent.

Thank you Ananya Roy and Maja Louise Arendt for asking questions and providing input on the project during the pan-cancer meetings and Jessika Nordin for helping me get started on the dataset.

Finally, many thanks to Felix Teufel for constructive feedback on my report and to Felicitas Pensch and Bernadette Pensch for reading my popular science summary. I'm grateful to all three of them, my family and my friends for their continual support.

# References

Abascal F, Acosta R, Addleman NJ, Adrian J, Afzal V, Aken B, Akiyama JA, Jammal O Al, Amrhein H, Anderson SM, Andrews GR, Antoshechkin I, Ardlie KG, Armstrong J, Astley M, Banerjee B, Barkal AA, Barnes IHA, Barozzi I, Barrell D, Barson G, Bates D, Baymuradov UK, Bazile C, Beer MA, Beik S, Bender MA, Bennett R, Bouvrette LPB, Bernstein BE, Berry A, Bhaskar A, Bignell A, Blue SM, Bodine DM, Boix C, Boley N, Borrman T, Borsari B, Boyle AP, Brandsmeier LA, Breschi A, Bresnick EH, Brooks JA, Buckley M, Burge CB, Byron R, Cahill E, Cai L, Cao L, Carty M, Castanon RG, Castillo A, Chaib H, Chan ET, Chee DR, Chee S, Chen H, Chen H, Chen JY, Chen S, Cherry JM, Chhetri SB, Choudhary JS, Chrast J, Chung D, Clarke D, Cody NAL, Coppola CJ, Coursen J, D'Ippolito AM, Dalton S, Danyko C, Davidson C, Davila-Velderrain J, Davis CA, Dekker J, Deran A, DeSalvo G, Despacio-Reyes G, Dewey CN, Dickel DE, Diegel M, Diekhans M, Dileep V, Ding B, Djebali S, Dobin A, Dominguez D, Donaldson S, Drenkow J, Dreszer TR, Drier Y, Duff MO, Dunn D, Eastman C, Ecker JR, Edwards MD, El-Ali N, Elhajjajy SI, Elkins K, Emili A, Epstein CB, Evans RC, Ezkurdia I, Fan K, Farnham PJ, Farrell N, Feingold EA, Ferreira AM, Fisher-Aylor K, Fitzgerald S, Flicek P, Foo CS, Fortier K, Frankish A, Freese P, Fu S, Fu XD, Fu Y, Fukuda-Yuzawa Y, Fulciniti M, Funnell APW, Gabdank I, Galeev T, Gao M, Giron CG, Garvin TH, Gelboin-Burkhart CA, Georgolopoulos G, Gerstein MB, Giardine BM, Gifford DK, Gilbert DM, Gilchrist DA, Gillespie S, Gingeras TR, Gong P, Gonzalez A, Gonzalez JM, Good P, Goren A, Gorkin DU, Graveley BR, Gray M, Greenblatt JF, Griffiths E, Groudine MT, Grubert F, Gu M, Guigó R, Guo H, Guo Y, Guo Y, Gursoy G, Gutierrez-Arcelus M, Halow J, Hardison RC, Hardy M, Hariharan M, Harmanci A, Harrington A, Harrow JL, Hashimoto TB, Hasz RD, Hatan M, Haugen E, Hayes JE, He P, He Y, Heidari N, Hendrickson D, Heuston EF, Hilton JA, Hitz BC, Hochman A, Holgren C, Hou L, Hou S, Hsiao YHE, Hsu S, Huang H, Hubbard TJ, Huey J, Hughes TR, Hunt T, Ibarrientos S, Issner R, Iwata M, Izuogu O, Jaakkola T, Jameel N, Jansen C, Jiang L, Jiang P, Johnson A, Johnson R, Jungreis I, Kadaba M, Kasowski M, Kasparian M, Kato M, Kaul R, Kawli T, Kay M, Keen JC, Keles S, Keller CA, Kelley D, Kellis M, Kheradpour P, Kim DS, Kirilusha A, Klein RJ, Knoechel B, Kuan S, Kulik MJ, Kumar S, Kundaje A, Kutyavin T, Lagarde J, Lajoie BR, Lambert NJ, Lazar J, Lee AY, Lee D, Lee E, Lee JW, Lee K, Leslie CS, Levy S, Li B, Li H, Li N, Li X, Li YI, Li Y, Li Y, Li Y, Lian J, Libbrecht MW, Lin S, Lin Y, Liu D, Liu J, Liu P, Liu T, Liu XS, Liu Y, Liu Y, Long M, Lou S, Loveland J, Lu A, Lu Y, Lécuyer E, Ma L, Mackiewicz M, Mannion BJ, Mannstadt M, Manthravadi D, Marinov GK, Martin FJ, Mattei E, McCue K, McEown M, McVicker G, Meadows SK, Meissner A, Mendenhall EM, Messer CL, Meuleman W, Meyer C, Miller S, Milton MG, Mishra T, Moore DE, Moore HM, Moore JE, Moore SH, Moran J, Mortazavi A, Mudge JM, Munshi N, Murad R, Myers RM, Nandakumar V, Nandi P, Narasimha AM, Narayanan AK, Naughton H, Navarro FCP, Navas P, Nazarovs J, Nelson J, Neph S, Neri FJ, Nery JR, Nesmith AR, Newberry JS, Newberry KM, Ngo V, Nguyen R, Nguyen TB, Nguyen T, Nishida A, Noble WS, Novak CS, Novoa EM, Nuñez B, O'Donnell CW, Olson S, Onate KC, Otterman E, Ozadam H, Pagan M, Palden T, Pan X, Park Y, Partridge EC, Paten B, Pauli-Behn F, Pazin MJ, Pei B, Pennacchio LA, Perez AR, Perry EH, Pervouchine DD, Phalke NN, Pham Q, Phanstiel DH, Plajzer-Frick I, Pratt GA, Pratt HE, Preissl S, Pritchard JK, Pritykin Y, Purcaro MJ, Qin Q, Quinones-Valdez G, Rabano I, Radovani E, Raj A, Rajagopal N, Ram O, Ramirez L, Ramirez RN, Rausch D, Raychaudhuri S, Raymond J, Razavi R, Reddy TE, Reimonn TM, Ren B, Reymond A, Reynolds A, Rhie SK, Rinn J, Rivera M, Rivera-Mulia JC, Roberts B, Rodriguez JM, Rozowsky J, Ryan R, Rynes E,

Salins DN, Sandstrom R, Sasaki T, Sathe S, Savic D, Scavelli A, Scheiman J, Schlaffner C, Schloss JA, Schmitges FW, See LH, Sethi A, Setty M, Shafer A, Shan S, Sharon E, Shen Q, Shen Y, Sherwood RI, Shi M, Shin S, Shoresh N, Siebenthall K, Sisu C, Slifer T, Sloan CA, Smith A, Snetkova V, Snyder MP, Spacek D V., Srinivasan S, Srivas R, Stamatoyannopoulos G, Stamatoyannopoulos JA, Stanton R, Steffan D, Stehling-Sun S, Strattan JS, Su A, Sundararaman B, Suner MM, Syed T, Szynkarek M, Tanaka FY, Tenen D, Teng M, Thomas JA, Toffey D, Tress ML, Trout DE, Trynka G, Tsuji J, Upchurch SA, Ursu O, Uszczynska-Ratajczak B, Uziel MC, Valencia A, Biber B Van, van der Velde AG, Van Nostrand EL, Vaydylevich Y, Vazquez J, Victorsen A, Vielmetter J, Vierstra J, Visel A, Vlasova A, Vockley CM, Volpi S, Vong S, Wang H, Wang M, Wang Q, Wang R, Wang T, Wang W, Wang X, Wang Y, Watson NK, Wei X, Wei Z, Weisser H, Weissman SM, Welch R, Welikson RE, Weng Z, Westra HJ, Whitaker JW, White C, White KP, Wildberg A, Williams BA, Wine D, Witt HN, Wold B, Wolf M, Wright J, Xiao R, Xiao X, Xu J, Xu J, Yan KK, Yan Y, Yang H, Yang X, Yang YW, Yardımcı GG, Yee BA, Yeo GW, Young T, Yu T, Yue F, Zaleski C, Zang C, Zeng H, Zeng W, Zerbino DR, Zhai J, Zhan L, Zhan Y, Zhang B, Zhang J, Zhang J, Zhang K, Zhang L, Zhang P, Zhang Q, Zhang XO, Zhang Y, Zhang Z, Zhao Y, Zheng Y, Zhong G, Zhou XQ, Zhu Y, Zimmerman J, Snyder MP, Gingeras TR, Moore JE, Weng Z, Gerstein MB, Ren B, Hardison RC, Stamatoyannopoulos JA, Graveley BR, Feingold EA, Pazin MJ, Pagan M, Gilchrist DA, Hitz BC, Cherry JM, Bernstein BE, Mendenhall EM, Zerbino DR, Frankish A, Flicek P, Myers RM. 2020. Perspectives on ENCODE. Nature 583: 693–698.

Albert FW, Kruglyak L. 2015. The role of regulatory variation in complex traits and disease. Nature Reviews Genetics 16: 197–212.

Ameur A, Dahlberg J, Olason P, Vezzi F, Karlsson R, Martin M, Viklund J, Kähäri AK, Lundin P, Che H, Thutkawkorapin J, Eisfeldt J, Lampa S, Dahlberg M, Hagberg J, Jareborg N, Liljedahl U, Jonasson I, Johansson Å, Feuk L, Lundeberg J, Syvänen AC, Lundin S, Nilsson D, Nystedt B, Magnusson PKE, Gyllensten U. 2017. SweGen: A whole-genome data resource of genetic variability in a cross-section of the Swedish population. European Journal of Human Genetics 25: 1253–1260.

Auton A, Abecasis GR, Altshuler DM, Durbin RM, Bentley DR, Chakravarti A, Clark AG, Donnelly P, Eichler EE, Flicek P, Gabriel SB, Gibbs RA, Green ED, Hurles ME, Knoppers BM, Korbel JO, Lander ES, Lee C, Lehrach H, Mardis ER, Marth GT, McVean GA, Nickerson DA, Schmidt JP, Sherry ST, Wang J, Wilson RK, Boerwinkle E, Doddapaneni H, Han Y, Korchina V, Kovar C, Lee S, Muzny D, Reid JG, Zhu Y, Chang Y, Feng Q, Fang X, Guo X, Jian M, Jiang H, Jin X, Lan T, Li G, Li J, Li Y, Liu S, Liu X, Lu Y, Ma X, Tang M, Wang B, Wang G, Wu H, Wu R, Xu X, Yin Y, Zhang D, Zhang W, Zhao J, Zhao M, Zheng X, Gupta N, Gharani N, Toji LH, Gerry NP, Resch AM, Barker J, Clarke L, Gil L, Hunt SE, Kelman G, Kulesha E, Leinonen R, McLaren WM, Radhakrishnan R, Roa A, Smirnov D, Smith RE, Streeter I, Thormann A, Toneva I, Vaughan B, Zheng-Bradley X, Grocock R, Humphray S, James T, Kingsbury Z, Sudbrak R, Albrecht MW, Amstislavskiy VS, Borodina TA, Lienhard M, Mertes F, Sultan M, Timmermann B, Yaspo ML, Fulton L, Ananiev V, Belaia Z, Beloslyudtsev D, Bouk N, Chen C, Church D, Cohen R, Cook C, Garner J, Hefferon T, Kimelman M, Liu C, Lopez J, Meric P, O'Sullivan C, Ostapchuk Y, Phan L, Ponomarov S, Schneider V, Shekhtman E, Sirotkin K, Slotta D, Zhang H, Balasubramaniam S, Burton J, Danecek P, Keane TM, Kolb-Kokocinski A, McCarthy S, Stalker J, Quail M, Davies CJ, Gollub J, Webster T, Wong B, Zhan Y,

Campbell CL, Kong Y, Marcketta A, Yu F, Antunes L, Bainbridge M, Sabo A, Huang Z, Coin LJM, Fang L, Li Q, Li Z, Lin H, Liu B, Luo R, Shao H, Xie Y, Ye C, Yu C, Zhang F, Zheng H, Zhu H, Alkan C, Dal E, Kahveci F, Garrison EP, Kural D, Lee WP, Leong WF, Stromberg M, Ward AN, Wu J, Zhang M, Daly MJ, DePristo MA, Handsaker RE, Banks E, Bhatia G, Del Angel G, Genovese G, Li H, Kashin S, McCarroll SA, Nemesh JC, Poplin RE, Yoon SC, Lihm J, Makarov V, Gottipati S, Keinan A, Rodriguez-Flores JL, Rausch T, Fritz MH, Stütz AM, Beal K, Datta A, Herrero J, Ritchie GRS, Zerbino D, Sabeti PC, Shlyakhter I, Schaffner SF, Vitti J, Cooper DN, Ball E V., Stenson PD, Barnes B, Bauer M, Cheetham RK, Cox A, Eberle M, Kahn S, Murray L, Peden J, Shaw R, Kenny EE, Batzer MA, Konkel MK, Walker JA, MacArthur DG, Lek M, Herwig R, Ding L, Koboldt DC, Larson D, Ye K, Gravel S, Swaroop A, Chew E, Lappalainen T, Erlich Y, Gymrek M, Willems TF, Simpson JT, Shriver MD, Rosenfeld JA, Bustamante CD, Montgomery SB, De La Vega FM, Byrnes JK, Carroll AW, DeGorter MK, Lacroute P, Maples BK, Martin AR, Moreno-Estrada A, Shringarpure SS, Zakharia F, Halperin E, Baran Y, Cerveira E, Hwang J, Malhotra A, Plewczynski D, Radew K, Romanovitch M, Zhang C, Hyland FCL, Craig DW, Christoforides A, Homer N, Izatt T, Kurdoglu AA, Sinari SA, Squire K, Xiao C, Sebat J, Antaki D, Gujral M, Noor A, Ye K, Burchard EG, Hernandez RD, Gignoux CR, Haussler D, Katzman SJ, Kent WJ, Howie B, Ruiz-Linares A, Dermitzakis ET, Devine SE, Kang HM, Kidd JM, Blackwell T, Caron S, Chen W, Emery S, Fritsche L, Fuchsberger C, Jun G, Li B, Lyons R, Scheller C, Sidore C, Song S, Sliwerska E, Taliun D, Tan A, Welch R, Wing MK, Zhan X, Awadalla P, Hodgkinson A, Li Y, Shi X, Quitadamo A, Lunter G, Marchini JL, Myers S, Churchhouse C, Delaneau O, Gupta-Hinch A, Kretzschmar W, Iqbal Z, Mathieson I, Menelaou A, Rimmer A, Xifara DK, Oleksyk TK, Fu Y, Liu X, Xiong M, Jorde L, Witherspoon D, Xing J, Browning BL, Browning SR, Hormozdiari F, Sudmant PH, Khurana E, Tyler-Smith C, Albers CA, Ayub Q, Chen Y, Colonna V, Jostins L, Walter K, Xue Y, Gerstein MB, Abyzov A, Balasubramanian S, Chen J, Clarke D, Fu Y, Harmanci AO, Jin M, Lee D, Liu J, Mu XJ, Zhang J, Zhang Y, Hartl C, Shakir K, Degenhardt J, Meiers S, Raeder B, Casale FP, Stegle O, Lameijer EW, Hall I, Bafna V, Michaelson J, Gardner EJ, Mills RE, Dayama G, Chen K, Fan X, Chong Z, Chen T, Chaisson MJ, Huddleston J, Malig M, Nelson BJ, Parrish NF, Blackburne B, Lindsay SJ, Ning Z, Zhang Y, Lam H, Sisu C, Challis D, Evani US, Lu J, Nagaswamy U, Yu J, Li W, Habegger L, Yu H, Cunningham F, Dunham I, Lage K, Jespersen JB, Horn H, Kim D, Desalle R, Narechania A, Sayres MAW, Mendez FL, Poznik GD, Underhill PA, Mittelman D, Banerjee R, Cerezo M, Fitzgerald TW, Louzada S, Massaia A, Yang F, Kalra D, Hale W, Dan X, Barnes KC, Beiswanger C, Cai H, Cao H, Henn B, Jones D, Kaye JS, Kent A, Kerasidou A, Mathias R, Ossorio PN, Parker M, Rotimi CN, Royal CD, Sandoval K, Su Y, Tian Z, Tishkoff S, Via M, Wang Y, Yang H, Yang L, Zhu J, Bodmer W, Bedoya G, Cai Z, Gao Y, Chu J, Peltonen L, Garcia-Montero A, Orfao A, Dutil J, Martinez-Cruzado JC, Mathias RA, Hennis A, Watson H, McKenzie C, Qadri F, LaRocque R, Deng X, Asogun D, Folarin O, Happi C, Omoniwa O, Stremlau M, Tariyal R, Jallow M, Joof FS, Corrah T, Rockett K, Kwiatkowski D, Kooner J, Hien TT, Dunstan SJ, ThuyHang N, Fonnie R, Garry R, Kanneh L, Moses L, Schieffelin J, Grant DS, Gallo C, Poletti G, Saleheen D, Rasheed A, Brooks LD, Felsenfeld AL, McEwen JE, Vaydylevich Y, Duncanson A, Dunn M, Schloss JA. 2015. A global reference for human genetic variation. Nature 526: 68–74.

Bailey MH, Tokheim C, Porta-Pardo E, Sengupta S, Bertrand D, Weerasinghe A, Colaprico A, Wendl MC, Kim J, Reardon B, Ng PKS, Jeong KJ, Cao S, Wang Z, Gao J, Gao Q, Wang F, Liu

EM, Mularoni L, Rubio-Perez C, Nagarajan N, Cortés-Ciriano I, Zhou DC, Liang WW, Hess JM, Yellapantula VD, Tamborero D, Gonzalez-Perez A, Suphavilai C, Ko JY, Khurana E, Park PJ, Van Allen EM, Liang H, Caesar-Johnson SJ, Demchok JA, Felau I, Kasapi M, Ferguson ML, Hutter CM, Sofia HJ, Tarnuzzer R, Yang L, Zenklusen JC, Zhang J (Julia), Chudamani S, Liu J, Lolla L, Naresh R, Pihl T, Sun Q, Wan Y, Wu Y, Cho J, DeFreitas T, Frazer S, Gehlenborg N, Getz G, Heiman DI, Lawrence MS, Lin P, Meier S, Noble MS, Saksena G, Voet D, Zhang H, Bernard B, Chambwe N, Dhankani V, Knijnenburg T, Kramer R, Leinonen K, Liu Y, Miller M, Reynolds S, Shmulevich I, Thorsson V, Zhang W, Akbani R, Broom BM, Hegde AM, Ju Z, Kanchi RS, Korkut A, Li J, Ling S, Liu W, Lu Y, Mills GB, Ng KS, Rao A, Ryan M, Wang J, Weinstein JN, Zhang J, Abeshouse A, Armenia J, Chakravarty D, Chatila WK, de Bruijn I, Gross BE, Heins ZJ, Kundra R, La K, Ladanyi M, Luna A, Nissan MG, Ochoa A, Phillips SM, Reznik E, Sanchez-Vega F, Sander C, Schultz N, Sheridan R, Sumer SO, Sun Y, Taylor BS, Anur P, Peto M, Spellman P, Benz C, Stuart JM, Wong CK, Yau C, Hayes DN, Parker JS, Wilkerson MD, Ally A, Balasundaram M, Bowlby R, Brooks D, Carlsen R, Chuah E, Dhalla N, Holt R, Jones SJM, Kasaian K, Lee D, Ma Y, Marra MA, Mayo M, Moore RA, Mungall AJ, Mungall K, Robertson AG, Sadeghi S, Schein JE, Sipahimalani P, Tam A, Thiessen N, Tse K, Wong T, Berger AC, Beroukhim R, Cherniack AD, Cibulskis C, Gabriel SB, Gao GF, Ha G, Meyerson M, Schumacher SE, Shih J, Kucherlapati MH, Kucherlapati RS, Baylin S, Cope L, Danilova L, Bootwalla MS, Lai PH, Maglinte DT, Van Den Berg DJ, Weisenberger DJ, Auman JT, Balu S, Bodenheimer T, Fan C, Hoadley KA, Hoyle AP, Jefferys SR, Jones CD, Meng S, Mieczkowski PA, Mose LE, Perou AH, Perou CM, Roach J, Shi Y, Simons J V., Skelly T, Soloway MG, Tan D, Veluvolu U, Fan H, Hinoue T, Laird PW, Shen H, Zhou W, Bellair M, Chang K, Covington K, Creighton CJ, Dinh H, Doddapaneni HV, Donehower LA, Drummond J, Gibbs RA, Glenn R, Hale W, Han Y, Hu J, Korchina V, Lee S, Lewis L, Li W, Liu X, Morgan M, Morton D, Muzny D, Santibanez J, Sheth M, Shinbrot E, Wang L, Wang M, Wheeler DA, Xi L, Zhao F, Hess J, Appelbaum EL, Bailey M, Cordes MG, Ding L, Fronick CC, Fulton LA, Fulton RS, Kandoth C, Mardis ER, McLellan MD, Miller CA, Schmidt HK, Wilson RK, Crain D, Curley E, Gardner J, Lau K, Mallery D, Morris S, Paulauskis J, Penny R, Shelton C, Shelton T, Sherman M, Thompson E, Yena P, Bowen J, Gastier-Foster JM, Gerken M, Leraas KM, Lichtenberg TM, Ramirez NC, Wise L, Zmuda E, Corcoran N, Costello T, Hovens C, Carvalho AL, de Carvalho AC, Fregnani JH, Longatto-Filho A, Reis RM, Scapulatempo-Neto C, Silveira HCS, Vidal DO, Burnette A, Eschbacher J, Hermes B, Noss A, Singh R, Anderson ML, Castro PD, Ittmann M, Huntsman D, Kohl B, Le X, Thorp R, Andry C, Duffy ER, Lyadov V, Paklina O, Setdikova G, Shabunin A, Tavobilov M, McPherson C, Warnick R, Berkowitz R, Cramer D, Feltmate C, Horowitz N, Kibel A, Muto M, Raut CP, Malykh A, Barnholtz-Sloan JS, Barrett W, Devine K, Fulop J, Ostrom QT, Shimmel K, Wolinsky Y, Sloan AE, De Rose A, Giuliante F, Goodman M, Karlan BY, Hagedorn CH, Eckman J, Harr J, Myers J, Tucker K, Zach LA, Deyarmin B, Hu H, Kvecher L, Larson C, Mural RJ, Somiari S, Vicha A, Zelinka T, Bennett J, Iacocca M, Rabeno B, Swanson P, Latour M, Lacombe L, Têtu B, Bergeron A, McGraw M, Staugaitis SM, Chabot J, Hibshoosh H, Sepulveda A, Su T, Wang T, Potapova O, Voronina O, Desjardins L, Mariani O, Roman-Roman S, Sastre X, Stern MH, Cheng F, Signoretti S, Berchuck A, Bigner D, Lipp E, Marks J, McCall S, McLendon R, Secord A, Sharp A, Behera M, Brat DJ, Chen A, Delman K, Force S, Khuri F, Magliocca K, Maithel S, Olson JJ, Owonikoko T, Pickens A, Ramalingam S, Shin DM, Sica G, Van Meir EG, Eijckenboom W, Gillis A, Korpershoek E, Looijenga L, Oosterhuis W, Stoop H, van Kessel

KE, Zwarthoff EC, Calatozzolo C, Cuppini L, Cuzzubbo S, DiMeco F, Finocchiaro G, Mattei L, Perin A, Pollo B, Chen C, Houck J, Lohavanichbutr P, Hartmann A, Stoehr C, Stoehr R, Taubert H, Wach S, Wullich B, Kycler W, Murawa D, Wiznerowicz M, Chung K, Edenfield WJ, Martin J, Baudin E, Bubley G, Bueno R, De Rienzo A, Richards WG, Kalkanis S, Mikkelsen T, Noushmehr H, Scarpace L, Girard N, Aymerich M, Campo E, Giné E, Guillermo AL, Van Bang N, Hanh PT, Phu BD, Tang Y, Colman H, Evason K, Dottino PR, Martignetti JA, Gabra H, Juhl H, Akeredolu T, Stepa S, Hoon D, Ahn K, Kang KJ, Beuschlein F, Breggia A, Birrer M, Bell D, Borad M, Bryce AH, Castle E, Chandan V, Cheville J, Copland JA, Farnell M, Flotte T, Giama N, Ho T, Kendrick M, Kocher JP, Kopp K, Moser C, Nagorney D, O'Brien D, O'Neill BP, Patel T, Petersen G, Que F, Rivera M, Roberts L, Smallridge R, Smyrk T, Stanton M, Thompson RH, Torbenson M, Yang JD, Zhang L, Brimo F, Ajani JA, Gonzalez AMA, Behrens C, Bondaruk J, Broaddus R, Czerniak B, Esmaeli B, Fujimoto J, Gershenwald J, Guo C, Lazar AJ, Logothetis C, Meric-Bernstam F, Moran C, Ramondetta L, Rice D, Sood A, Tamboli P, Thompson T, Troncoso P, Tsao A, Wistuba I, Carter C, Haydu L, Hersey P, Jakrot V, Kakavand H, Kefford R, Lee K, Long G, Mann G, Quinn M, Saw R, Scolyer R, Shannon K, Spillane A, Stretch J, Synott M, Thompson J, Wilmott J, Al-Ahmadie H, Chan TA, Ghossein R, Gopalan A, Levine DA, Reuter V, Singer S, Singh B, Tien NV, Broudy T, Mirsaidi C, Nair P, Drwiega P, Miller J, Smith J, Zaren H, Park JW, Hung NP, Kebebew E, Linehan WM, Metwalli AR, Pacak K, Pinto PA, Schiffman M, Schmidt LS, Vocke CD, Wentzensen N, Worrell R, Yang H, Moncrieff M, Goparaju C, Melamed J, Pass H, Botnariuc N, Caraman I, Cernat M, Chemencedji I, Clipca A, Doruc S, Gorincioi G, Mura S, Pirtac M, Stancul I, Tcaciuc D, Albert M, Alexopoulou I, Arnaout A, Bartlett J, Engel J, Gilbert S, Parfitt J, Sekhon H, Thomas G, Rassl DM, Rintoul RC, Bifulco C, Tamakawa R, Urba W, Hayward N, Timmers H, Antenucci A, Facciolo F, Grazi G, Marino M, Merola R, de Krijger R, Gimenez-Roqueplo AP, Piché A, Chevalier S, McKercher G, Birsoy K, Barnett G, Brewer C, Farver C, Naska T, Pennell NA, Raymond D, Schilero C, Smolenski K, Williams F, Morrison C, Borgia JA, Liptay MJ, Pool M, Seder CW, Junker K, Omberg L, Dinkin M, Manikhas G, Alvaro D, Bragazzi MC, Cardinale V, Carpino G, Gaudio E, Chesla D, Cottingham S, Dubina M, Moiseenko F, Dhanasekaran R, Becker KF, Janssen KP, Slotta-Huspenina J, Abdel-Rahman MH, Aziz D, Bell S, Cebulla CM, Davis A, Duell R, Elder JB, Hilty J, Kumar B, Lang J, Lehman NL, Mandt R, Nguyen P, Pilarski R, Rai K, Schoenfield L, Senecal K, Wakely P, Hansen P, Lechan R, Powers J, Tischler A, Grizzle WE, Sexton KC, Kastl A, Henderson J, Porten S, Waldmann J, Fassnacht M, Asa SL, Schadendorf D, Couce M, Graefen M, Huland H, Sauter G, Schlomm T, Simon R, Tennstedt P, Olabode O, Nelson M, Bathe O, Carroll PR, Chan JM, Disaia P, Glenn P, Kelley RK, Landen CN, Phillips J, Prados M, Simko J, Smith-McCune K, VandenBerg S, Roggin K, Fehrenbach A, Kendler A, Sifri S, Steele R, Jimeno A, Carey F, Forgie I, Mannelli M, Carney M, Hernandez B, Campos B, Herold-Mende C, Jungk C, Unterberg A, von Deimling A, Bossler A, Galbraith J, Jacobus L, Knudson M, Knutson T, Ma D, Milhem M, Sigmund R, Godwin AK, Madan R, Rosenthal HG, Adebamowo C, Adebamowo SN, Boussioutas A, Beer D, Giordano T, Mes-Masson AM, Saad F, Bocklage T, Landrum L, Mannel R, Moore K, Moxley K, Postier R, Walker J, Zuna R, Feldman M, Valdivieso F, Dhir R, Luketich J, Pinero EMM, Quintero-Aguilo M, Carlotti CG, Dos Santos JS, Kemp R, Sankarankuty A, Tirapelli D, Catto J, Agnew K, Swisher E, Creaney J, Robinson B, Shelley CS, Godwin EM, Kendall S, Shipman C, Bradford C, Carey T, Haddad A, Moyer J, Peterson L, Prince M, Rozek L, Wolf G, Bowman R, Fong KM, Yang

I, Korst R, Rathmell WK, Fantacone-Campbell JL, Hooke JA, Kovatich AJ, Shriver CD, DiPersio J, Drake B, Govindan R, Heath S, Ley T, Van Tine B, Westervelt P, Rubin MA, Lee J Il, Aredes ND, Mariamidze A, Godzik A, Lopez-Bigas N, Stuart J, Wheeler D, Chen K, Karchin R. 2018. Comprehensive Characterization of Cancer Driver Genes and Mutations. Cell 173: 371-385.e18.

Basu-Roy U, Ambrosetti D, Favaro R, Nicolis SK, Mansukhani A, Basilico C. 2010. The transcription factor Sox2 is required for osteoblast self-renewal. Cell Death and Differentiation 17: 1345–1353.

Basu-Roy U, Seo E, Ramanathapuram L, Rapp TB, Perry JA, Orkin SH, Mansukhani A, Basilico C. 2012. Sox2 maintains self renewal of tumor-initiating cells in osteosarcomas. Oncogene 31: 2270–2282.

Batra N, Kar R, Jiang JX. 2012. Gap junctions and hemichannels in signal transmission, function and development of bone. Biochimica et Biophysica Acta - Biomembranes 1818: 1909–1918.

Battle A, Mostafavi S, Zhu X, Potash JB, Weissman MM, McCormick C, Haudenschild CD, Beckman KB, Shi J, Mei R, Urban AE, Montgomery SB, Levinson DF, Koller D. 2014. Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. Genome Research 24: 14–24.

Birney E, Stamatoyannopoulos JA, Dutta A, Guigó R, Gingeras TR, Margulies EH, Weng Z, Snyder M, Dermitzakis ET, Thurman RE, Kuehn MS, Taylor CM, Neph S, Koch CM, Asthana S, Malhotra A, Adzhubei I, Greenbaum JA, Andrews RM, Flicek P, Boyle PJ, Cao H, Carter NP, Clelland GK, Davis S, Day N, Dhami P, Dillon SC, Dorschner MO, Fiegler H, Giresi PG, Goldy J, Hawrylycz M, Haydock A, Humbert R, James KD, Johnson BE, Johnson EM, Frum TT, Rosenzweig ER, Karnani N, Lee K, Lefebvre GC, Navas PA, Neri F, Parker SCJ, Sabo PJ, Sandstrom R, Shafer A, Vetrie D, Weaver M, Wilcox S, Yu M, Collins FS, Dekker J, Lieb JD, Tullius TD, Crawford GE, Sunyaev S, Noble WS, Dunham I, Denoeud F, Reymond A, Kapranov P, Rozowsky J, Zheng D, Castelo R, Frankish A, Harrow J, Ghosh S, Sandelin A, Hofacker IL, Baertsch R, Keefe D, Dike S, Cheng J, Hirsch HA, Sekinger EA, Lagarde J, Abril JF, Shahab A, Flamm C, Fried C, Hackermüller J, Hertel J, Lindemeyer M, Missal K, Tanzer A, Washietl S, Korbel J, Emanuelsson O, Pedersen JS, Holroyd N, Taylor R, Swarbreck D, Matthews N, Dickson MC, Thomas DJ, Weirauch MT, Gilbert J, Drenkow J, Bell I, Zhao X, Srinivasan KG, Sung WK, Ooi HS, Chiu KP, Foissac S, Alioto T, Brent M, Pachter L, Tress ML, Valencia A, Choo SW, Choo CY, Ucla C, Manzano C, Wyss C, Cheung E, Clark TG, Brown JB, Ganesh M, Patel S, Tammana H, Chrast J, Henrichsen CN, Kai C, Kawai J, Nagalakshmi U, Wu J, Lian Z, Lian J, Newburger P, Zhang X, Bickel P, Mattick JS, Carninci P, Hayashizaki Y, Weissman S, Hubbard T, Myers RM, Rogers J, Stadler PF, Lowe TM, Wei CL, Ruan Y, Struhl K, Gerstein M, Antonarakis SE, Fu Y, Green ED, Karaöz U, Siepel A, Taylor J, Liefer LA, Wetterstrand KA, Good PJ, Feingold EA, Guyer MS, Cooper GM, Asimenos G, Dewey CN, Hou M, Nikolaev S, Montoya-Burgos JI, Löytynoja A, Whelan S, Pardi F, Massingham T, Huang H, Zhang NR, Holmes I, Mullikin JC, Ureta-Vidal A, Paten B, Seringhaus M, Church D, Rosenbloom K, Kent WJ, Stone EA, Batzoglou S, Goldman N, Hardison RC, Haussler D, Miller W, Sidow A, Trinklein ND, Zhang ZD, Barrera L, Stuart R, King DC, Ameur A, Enroth S, Bieda MC, Kim J, Bhinge AA, Jiang N, Liu J, Yao F, Vega VB, Lee CWH, Ng P, Yang A,

Moqtaderi Z, Zhu Z, Xu X, Squazzo S, Oberley MJ, Inman D, Singer MA, Richmond TA, Munn KJ, Rada-Iglesias A, Wallerman O, Komorowski J, Fowler JC, Couttet P, Bruce AW, Dovey OM, Ellis PD, Langford CF, Nix DA, Euskirchen G, Hartman S, Urban AE, Kraus P, Van Calcar S, Heintzman N, Hoon Kim T, Wang K, Qu C, Hon G, Luna R, Glass CK, Rosenfeld MG, Aldred SF, Cooper SJ, Halees A, Lin JM, Shulha HP, Zhang X, Xu M, Haidar JNS, Yu Y, Iyer VR, Green RD, Wadelius C, Farnham PJ, Ren B, Harte RA, Hinrichs AS, Trumbower H, Clawson H, Hillman-Jackson J, Zweig AS, Smith K, Thakkapallayil A, Barber G, Kuhn RM, Karolchik D, Armengol L, Bird CP, De Bakker PIW, Kern AD, Lopez-Bigas N, Martin JD, Stranger BE, Woodroffe A, Davydov E, Dimas A, Eyras E, Hallgrímsdóttir IB, Huppert J, Zody MC, Abecasis GR, Estivill X, Bouffard GG, Guan X, Hansen NF, Idol JR, Maduro VVB, Maskeri B, McDowell JC, Park M, Thomas PJ, Young AC, Blakesley RW, Muzny DM, Sodergren E, Wheeler DA, Worley KC, Jiang H, Weinstock GM, Gibbs RA, Graves T, Fulton R, Mardis ER, Wilson RK, Clamp M, Cuff J, Gnerre S, Jaffe DB, Chang JL, Lindblad-Toh K, Lander ES, Koriabine M, Nefedov M, Osoegawa K, Yoshinaga Y, Zhu B, De Jong PJ. 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 447: 799–816.

Borneman AR, Gianoulis TA, Zhang ZD, Yu H, Rozowsky J, Seringhaus MR, Wang LY, Gerstein M, Snyder M. 2007. Divergence of Transcription Factor Binding Sites Across Related Yeast Species. Science 317: 815–819.

Bushweller JH. 2019. Targeting transcription factors in cancer — from undruggable to reality. Nature Reviews Cancer 19: 611–624.

Campbell PJ, Getz G, Korbel JO, Stuart JM, Jennings JL, Stein LD, Perry MD, Nahal-Bose HK, Ouellette BFF, Li CH, Rheinbay E, Nielsen GP, Sgroi DC, Wu CL, Faquin WC, Deshpande V, Boutros PC, Lazar AJ, Hoadley KA, Louis DN, Dursi LJ, Yung CK, Bailey MH, Saksena G, Raine KM, Buchhalter I, Kleinheinz K, Schlesner M, Zhang J, Wang W, Wheeler DA, Ding L, Simpson JT, O'Connor BD, Yakneen S, Ellrott K, Miyoshi N, Butler AP, Royo R, Shorser SI, Vazquez M, Rausch T, Tiao G, Waszak SM, Rodriguez-Martin B, Shringarpure S, Wu DY, Demidov GM, Delaneau O, Hayashi S, Imoto S, Habermann N, Segre A V., Garrison E, Cafferkey A, Alvarez EG, Heredia-Genestar JM, Muyas F, Drechsel O, Bruzos AL, Temes J, Zamora J, Baez-Ortega A, Kim HL, Mashl RJ, Ye K, DiBiase A, Huang K lin, Letunic I, McLellan MD, Newhouse SJ, Shmaya T, Kumar S, Wedge DC, Wright MH, Yellapantula VD, Gerstein M, Khurana E, Marques-Bonet T, Navarro A, Bustamante CD, Siebert R, Nakagawa H, Easton DF, Ossowski S, Tubio JMC, De La Vega FM, Estivill X, Yuen D, Mihaiescu GL, Omberg L, Ferretti V, Sabarinathan R, Pich O, Gonzalez-Perez A, Taylor-Weiner A, Fittall MW, Demeulemeester J, Tarabichi M, Roberts ND, Van Loo P, Cortés-Ciriano I, Urban L, Park P, Zhu B, Pitkänen E, Li Y, Saini N, Klimczak LJ, Weischenfeldt J, Sidiropoulos N, Alexandrov LB, Rabionet R, Escaramis G, Bosio M, Holik AZ, Susak H, Prasad A, Erkek S, Calabrese C, Raeder B, Harrington E, Mayes S, Turner D, Juul S, Roberts SA, Song L, Koster R, Mirabello L, Hua X, Tanskanen TJ, Tojo M, Chen J, Aaltonen LA, Rätsch G, Schwarz RF, Butte AJ, Brazma A, Chanock SJ, Chatterjee N, Stegle O, Harismendy O, Bova GS, Gordenin DA, Haan D, Sieverling L, Feuerbach L, Chalmers D, Joly Y, Knoppers B, Molnár-Gábor F, Phillips M, Thorogood A, Townend D, Goldman M, Fonseca NA, Xiang Q, Craft B, Piñeiro-Yáñez E, Muñoz A, Petryszak R, Füllgrabe A, Al-Shahrour F, Keays M, Haussler D, Weinstein J, Huber W, Valencia A, Papatheodorou I, Zhu J, Fan Y, Torrents D,

Bieg M, Chen K, Chong Z, Cibulskis K, Eils R, Fulton RS, Gelpi JL, Gonzalez S, Gut IG, Hach F, Heinold M, Hu T, Huang V, Hutter B, Jäger N, Jung J, Kumar Y, Lalansingh C, Leshchiner I, Livitz D, Ma EZ, Maruvka YE, Milovanovic A, Nielsen MM, Paramasivam N, Pedersen JS, Puiggròs M, Sahinalp SC, Sarrafi I, Stewart C, Stobbe MD, Wala JA, Wang J, Wendl M, Werner J, Wu Z, Xue H, Yamaguchi TN, Yellapantula V, Davis-Dusenbery BN, Grossman RL, Kim Y, Heinold MC, Hinton J, Jones DR, Menzies A, Stebbings L, Hess JM, Rosenberg M, Dunford AJ, Gupta M, Imielinski M, Meyerson M, Beroukhim R, Reimand J, Dhingra P, Favero F, Dentro S, Wintersinger J, Rudneva V, Park JW, Hong EP, Heo SG, Kahles A, Lehmann K Van, Soulette CM, Shiraishi Y, Liu F, He Y, Demircioğlu D, Davidson NR, Greger L, Li S, Liu D, Stark SG, Zhang F, Amin SB, Bailey P, Chateigner A, Frenkel-Morgenstern M, Hou Y, Huska MR, Kilpinen H, Lamaze FC, Li C, Li X, Li X, Liu X, Marin MG, Markowski J, Nandi T, Ojesina AI, Pan-Hammarström Q, Park PJ, Pedamallu CS, Su H, Tan P, Teh BT, Wang J, Xiong H, Ye C, Yung C, Zhang X, Zheng L, Zhu S, Awadalla P, Creighton CJ, Wu K, Yang H, Göke J, Zhang Z, Brooks AN, Fittall MW, Martincorena I, Rubio-Perez C, Juul M, Schumacher S, Shapira O, Tamborero D, Mularoni L, Hornshøj H, Deu-Pons J, Muiños F, Bertl J, Guo Q, Gonzalez-Perez A, Xiang Q, Bazant W, Barrera E, Al-Sedairy ST, Aretz A, Bell C, Betancourt M, Buchholz C, Calvo F, Chomienne C, Dunn M, Edmonds S, Green E, Gupta S, Hutter CM, Jegalian K, Jones N, Lu Y, Nakagama H, Nettekoven G, Planko L, Scott D, Shibata T, Shimizu K, Stratton MR, Yugawa T, Tortora G, VijayRaghavan K, Zenklusen JC, Townend D, Knoppers BM, Aminou B, Bartolome J, Boroevich KA, Boyce R, Buchanan A, Byrne NJ, Chen Z, Cho S, Choi W, Clapham P, Dow MT, Dursi LJ, Eils J, Farcas C, Fayzullaev N, Flicek P, Heath AP, Hofmann O, Hong JH, Hudson TJ, Hübschmann D, Ivkovic S, Jeon SH, Jiao W, Kabbe R, Kahles A, Kerssemakers JNA, Kim H, Kim J, Koscher M, Koures A, Kovacevic M, Lawerenz C, Liu J, Mijalkovic S, Mijalkovic-Lazic AM, Miyano S, Nastic M, Nicholson J, Ocana D, Ohi K, Ohno-Machado L, Pihl TD, Prinz M, Radovic P, Short C, Sofia HJ, Spring J, Struck AJ, Tijanic N, Vicente D, Wang Z, Williams A, Woo Y, Wright AJ, Yang L, Hamilton MP, Johnson TA, Kahraman A, Kellis M, Polak P, Sallari R, Sinnott-Armstrong N, von Mering C, Beltran S, Gerhard DS, Gut M, Trotta JR, Whalley JP, Niu B, Espiritu SMG, Gao S, Huang Y, Lalansingh CM, Teague JW, Wendl MC, Abascal F, Bader GD, Bandopadhayay P, Barenboim J, Brunak S, Carlevaro-Fita J, Chakravarty D, Chan CWY, Choi JK, Diamanti K, Fink JL, Frigola J, Gambacorti-Passerini C, Garsed DW, Haradhvala NJ, Harmanci AO, Helmy M, Herrmann C, Hobolth A, Hodzic E, Hong C, Isaev K, Izarzugaza JMG, Johnson R, Juul RI, Kim J, Kim JK, Jan Komorowski, Lanzós A, Larsson E, Lee D, Li S, Li X, Lin Z, Liu EM, Lochovsky L, Lou S, Madsen T, Marchal K, Martinez-Fundichely A, McGillivray PD, Meyerson W, Paczkowska M, Park K, Park K, Pons T, Pulido-Tamayo S, Reyes-Salazar I, Reyna MA, Rubin MA, Salichos L, Sander C, Schumacher SE, Shackleton M, Shen C, Shrestha R, Shuai S, Tsunoda T, Umer HM, Uusküla-Reimand L, Verbeke LPC, Wadelius C, Wadi L, Warrell J, Wu G, Yu J, Zhang J, Zhang X, Zhang Y, Zhao Z, Zou L, Lawrence MS, Raphael BJ, Bailey PJ, Craft D, Goldman MJ, Aburatani H, Binder H, Dinh HQ, Heath SC, Hoffmann S, Imbusch CD, Kretzmer H, Laird PW, Martin-Subero JI, Nagae G, Shen H, Wang Q, Weichenhan D, Zhou W, Berman BP, Brors B, Plass C, Akdemir KC, Bowtell DDL, Burns KH, Busanovich J, Chan K, Dueso-Barroso A, Edwards PA, Etemadmoghadam D, Haber JE, Jones DTW, Ju YS, Kazanov MD, Koh Y, Kumar K, Lee EA, Lee JJK, Lynch AG, Macintyre G, Markowetz F, Navarro FCP, Pearson J V., Rippe K, Scully R, Villasante I, Waddell N, Yang L, Yao X, Yoon SS, Zhang CZ, Bergstrom EN, Boot A,

Covington K, Fujimoto A, Huang MN, Islam SMA, McPherson JR, Morganella S, Mustonen V, Ng AWT, Prokopec SD, Vázquez-García I, Wu Y, Yousif F, Yu W, Rozen SG, Rudneva VA, Shringarpure SS, Turner DJ, Xia T, Atwal G, Chang DK, Cooke SL, Faltas BM, Haider S, Kaiser VB, Karlić R, Kato M, Kübler K, Margolin A, Martin S, Nik-Zainal S, P'ng C, Semple CA, Smith J, Sun RX, Thai K, Wright DW, Yuan K, Biankin A V., Garraway L, Grimmond SM, Adams DJ, Anur P, Cao S, Christie EL, Cmero M, Cun Y, Dawson KJ, Dentro SC, Deshwar AG, Donmez N, Drews RM, Gerstung M, Ha G, Haase K, Jerman L, Ji Y, Jolly C, Lee J, Lee-Six H, Malikic S, Mitchell TJ, Morris QD, Oesper L, Peifer M, Peto M, Rosebrock D, Rubanova Y, Salcedo A, Sengupta S, Shi R, Shin SJ, Spiro O, Vembu S, Wintersinger JA, Yang TP, Yu K, Zhu H, Spellman PT, Weinstein JN, Chen Y, Fujita M, Han L, Hasegawa T, Komura M, Li J, Mizuno S, Shimizu E, Wang Y, Xu Y, Yamaguchi R, Yang F, Yang Y, Yoon CJ, Yuan Y, Liang H, Alawi M, Borozan I, Brewer DS, Cooper CS, Desai N, Grundhoff A, Iskar M, Su X, Zapatka M, Lichter P, Alsop K, Bruxner TJC, Christ AN, Cordner SM, Cowin PA, Drapkin R, Fereday S, George J, Hamilton A, Holmes O, Hung JA, Kassahn KS, Kazakoff SH, Kennedy CJ, Leonard CR, Mileshkin L, Miller DK, Arnau GM, Mitchell C, Newell F, Nones K, Patch AM, Quinn MC, Taylor DF, Thorne H, Traficante N, Vedururu R, Waddell NM, Waring PM, Wood S, Xu Q, deFazio A, Anderson MJ, Antonello D, Barbour AP, Bassi C, Bersani S, Cataldo I, Chantrill LA, Chiew YE, Chou A, Cingarlini S, Cloonan N, Corbo V, Davi MV, Duthie FR, Gill AJ, Graham JS, Harliwong I, Jamieson NB, Johns AL, Kench JG, Landoni L, Lawlor RT, Mafficini A, Merrett ND, Miotto M, Musgrove EA, Nagrial AM, Oien KA, Pajic M, Pinese M, Robertson AJ, Rooman I, Rusev BC, Samra JS, Scardoni M, Scarlett CJ, Scarpa A, Sereni E, Sikora KO, Simbolo M, Taschuk ML, Toon CW, Vicentini C, Wu J, Zeps N, Behren A, Burke H, Cebon J, Dagg RA, De Paoli-Iseppi R, Dutton-Regester K, Field MA, Fitzgerald A, Hersey P, Jakrot V, Johansson PA, Kakavand H, Kefford RF, Lau LMS, Long G V., Pickett HA, Pritchard AL, Pupo GM, Saw RPM, Schramm SJ, Shang CA, Shang P, Spillane AJ, Stretch JR, Tembe V, Thompson JF, Vilain RE, Wilmott JS, Yang JY, Hayward NK, Mann GJ, Scolyer RA, Bartlett J, Bavi P, Chadwick DE, Chan-Seng-Yue M, Cleary S, Connor AA, Czajka K, Denroche RE, Dhani NC, Eagles J, Gallinger S, Grant RC, Hedley D, Hollingsworth MA, Jang GH, Johns J, Kalimuthu S, Liang S Ben, Lungu I, Luo X, Mbabaali F, McPherson TA, Miller JK, Moore MJ, Notta F, Pasternack D, Petersen GM, Roehrl MHA, Sam M, Selander I, Serra S, Shahabi S, Thayer SP, Timms LE, Wilson GW, Wilson JM, Wouters BG, McPherson JD, Beck TA, Bhandari V, Collins CC, Fleshner NE, Fox NS, Fraser M, Heisler LE, Lalonde E, Livingstone J, Meng A, Sabelnykova VY, Shiah YJ, Van der Kwast T, Bristow RG, Ding S, Fan D, Li L, Nie Y, Xiao X, Xing R, Yang S, Yu Y, Zhou Y, Banks RE, Bourque G, Brennan P, Letourneau L, Riazalhosseini Y, Scelo G, Vasudev N, Viksna J, Lathrop M, Tost J, Ahn SM, Aparicio S, Arnould L, Aure MR, Bhosle SG, Birney E, Borg A, Boyault S, Brinkman AB, Brock JE, Broeks A, Børresen-Dale AL, Caldas C, Chin SF, Davies H, Desmedt C, Dirix L, Dronov S, Ehinger A, Eyfjord JE, Fatima A, Foekens JA, Futreal PA, Garred Ø, Giri DD, Glodzik D, Grabau D, Hilmarsdottir H, Hooijer GK, Jacquemier J, Jang SJ, Jonasson JG, Jonkers J, Kim HY, King TA, Knappskog S, Kong G, Krishnamurthy S, Lakhani SR, Langerød A, Larsimont D, Lee HJ, Lee JY, Lee MTM, Lingjærde OC, MacGrogan G, Martens JWM, O'Meara S, Pauporté I, Pinder S, Pivot X, Provenzano E, Purdie CA, Ramakrishna M, Ramakrishnan K, Reis-Filho J, Richardson AL, Ringnér M, Rodriguez JB, Rodríguez-González FG, Romieu G, Salgado R, Sauer T, Shepherd R, Sieuwerts AM, Simpson PT, Smid M, Sotiriou C, Span PN, Stefánsson ÓA, Stenhouse A, Stunnenberg HG, Sweep F, Tan BKT,

Thomas G, Thompson AM, Tommasi S, Treilleux I, Tutt A, Ueno NT, Van Laere S, Van den Eynden GG, Vermeulen P, Viari A, Vincent-Salomon A, Wong BH, Yates L, Zou X, van Deurzen CHM, van de Vijver MJ, van't Veer L, Ammerpohl O, Aukema S, Bergmann AK, Bernhart SH, Borkhardt A, Borst C, Burkhardt B, Claviez A, Goebler ME, Haake A, Haas S, Hansmann M, Hoell JI, Hummel M, Karsch D, Klapper W, Kneba M, Kreuz M, Kube D, Küppers R, Lenze D, Loeffler M, López C, Mantovani-Löffler L, Möller P, Ott G, Radlwimmer B, Richter J, Rohde M, Rosenstiel PC, Rosenwald A, Schilhabel MB, Schreiber S, Stadler PF, Staib P, Stilgenbauer S, Sungalee S, Szczepanowski M, Toprak UH, Trümper LHP, Wagener R, Zenz T, Hovestadt V, von Kalle C, Kool M, Korshunov A, Landgraf P, Lehrach H, Northcott PA, Pfister SM, Reifenberger G, Warnatz HJ, Wolf S, Yaspo ML, Assenov Y, Gerhauser C, Minner S, Schlomm T, Simon R, Sauter G, Sültmann H, Biswas NK, Maitra A, Majumder PP, Sarin R, Barbi S, Bonizzato G, Cantù C, Dei Tos AP, Fassan M, Grimaldi S, Luchini C, Malleo G, Marchegiani G, Milella M, Paiella S, Pea A, Pederzoli P, Ruzzenente A, Salvia R, Sperandio N, Arai Y, Hama N, Hiraoka N, Hosoda F, Nakamura H, Ojima H, Okusaka T, Totoki Y, Urushidate T, Fukayama M, Ishikawa S, Katai H, Katoh H, Komura D, Rokutan H, Saito-Adachi M, Suzuki A, Taniguchi H, Tatsuno K, Ushiku T, Yachida S, Yamamoto S, Aikata H, Arihiro K, Ariizumi S ichi, Chayama K, Furuta M, Gotoh K, Hayami S, Hirano S, Kawakami Y, Maejima K, Nakamura T, Nakano K, Ohdan H, Sasaki-Oku A, Tanaka H, Ueno M, Yamamoto M, Yamaue H, Choo SP, Cutcutache I, Khuntikeo N, Ong CK, Pairojkul C, Popescu I, Ahn KS, Aymerich M, Lopez-Guillermo A, López-Otín C, Puente XS, Campo E, Amary F, Baumhoer D, Behjati S, Bjerkehagen B, Futreal PA, Myklebost O, Pillay N, Tarpey P, Tirabosco R, Zaikova O, Flanagan AM, Boultwood J, Bowen DT, Cazzola M, Green AR, Hellstrom-Lindberg E, Malcovati L, Nangalia J, Papaemmanuil E, Vyas P, Ang Y, Barr H, Beardsmore D, Eldridge M, Gossage J, Grehan N, Hanna GB, Hayes SJ, Hupp TR, Khoo D, Lagergren J, Lovat LB, MacRae S, O'Donovan M, O'Neill JR, Parsons SL, Preston SR, Puig S, Roques T, Sanders G, Sothi S, Tavaré S, Tucker O, Turkington R, Underwood TJ, Welch I, Fitzgerald RC, Berney DM, De Bono JS, Cahill D, Camacho N, Dennis NM, Dudderidge T, Edwards SE, Fisher C, Foster CS, Ghori M, Gill P, Gnanapragasam VJ, Gundem G, Hamdy FC, Hawkins S, Hazell S, Howat W, Isaacs WB, Karaszi K, Kay JD, Khoo V, Kote-Jarai Z, Kremeyer B, Kumar P, Lambert A, Leongamornlert DA, Livni N, Lu YJ, Luxton HJ, Marsden L, Massie CE, Matthews L, Mayer E, McDermott U, Merson S, Neal DE, Ng A, Nicol D, Ogden C, Rowe EW, Shah NC, Thomas S, Thompson A, Verrill C, Visakorpi T, Warren AY, Whitaker HC, Zhang H, van As N, Eeles RA, Abeshouse A, Agrawal N, Akbani R, Al-Ahmadie H, Albert M, Aldape K, Ally A, Appelbaum EL, Armenia J, Asa S, Auman JT, Balasundaram M, Balu S, Barnholtz-Sloan J, Bathe OF, Baylin SB, Benz C, Berchuck A, Berrios M, Bigner D, Birrer M, Bodenheimer T, Boice L, Bootwalla MS, Bosenberg M, Bowlby R, Boyd J, Broaddus RR, Brock M, Brooks D, Bullman S, Caesar-Johnson SJ, Carey TE, Carlsen R, Cerfolio R, Chandan VS, Chen HW, Cherniack AD, Chien J, Cho J, Chuah E, Cibulskis C, Cope L, Cordes MG, Curley E, Czerniak B, Danilova L, Davis IJ, Defreitas T, Demchok JA, Dhalla N, Dhir R, Doddapaneni HV, El-Naggar A, Felau I, Ferguson ML, Finocchiaro G, Fong KM, Frazer S, Friedman W, Fronick CC, Fulton LA, Gabriel SB, Gao J, Gehlenborg N, Gershenwald JE, Ghossein R, Giama NH, Gibbs RA, Gomez C, Govindan R, Hayes DN, Hegde AM, Heiman DI, Heins Z, Hepperla AJ, Holbrook A, Holt RA, Hoyle AP, Hruban RH, Hu J, Huang M, Huntsman D, Huse J, Iacobuzio-Donahue CA, Ittmann M, Jayaseelan JC, Jefferys SR, Jones CD, Jones SJM, Juhl H, Kang KJ, Karlan B,

Kasaian K, Kebebew E, Kim HK, Korchina V, Kundra R, Lai PH, Lander E, Le X, Lee D, Levine DA, Lewis L, Ley T, Li HI, Lin P, Linehan WM, Liu FF, Lu Y, Lype L, Ma Y, Maglinte DT, Mardis ER, Marks J, Marra MA, Matthew TJ, Mayo M, McCune K, Meier SR, Meng S, Mieczkowski PA, Mikkelsen T, Miller CA, Mills GB, Moore RA, Morrison C, Mose LE, Moser CD, Mungall AJ, Mungall K, Mutch D, Muzny DM, Myers J, Newton Y, Noble MS, O'Donnell P, O'Neill BP, Ochoa A, Park JW, Parker JS, Pass H, Pastore A, Pennell NA, Perou CM, Petrelli N, Potapova O, Rader JS, Ramalingam S, Rathmell WK, Reuter V, Reynolds SM, Ringel M, Roach J, Roberts LR, Robertson AG, Sadeghi S, Saller C, Sanchez-Vega F, Schadendorf D, Schein JE, Schmidt HK, Schultz N, Seethala R, Senbabaoglu Y, Shelton T, Shi Y, Shih J, Shmulevich I, Shriver C, Signoretti S, Simons J V., Singer S, Sipahimalani P, Skelly TJ, Smith-McCune K, Socci ND, Soloway MG, Sood AK, Tam A, Tan D, Tarnuzzer R, Thiessen N, Thompson RH, Thorne LB, Tsao M, Umbricht C, Van Den Berg DJ, Van Meir EG, Veluvolu U, Voet D, Wang L, Weinberger P, Weisenberger DJ, Wigle D, Wilkerson MD, Wilson RK, Winterhoff B, Wiznerowicz M, Wong T, Wong W, Xi L, Yau C, Zhang H, Zhang H, Zhang J. 2020. Pan-cancer analysis of whole genomes. Nature 578: 82–93.

Carbon S, Douglass E, Good BM, Unni DR, Harris NL, Mungall CJ, Basu S, Chisholm RL, Dodson RJ, Hartline E, Fey P, Thomas PD, Albou LP, Ebert D, Kesling MJ, Mi H, Muruganujan A, Huang X, Mushayahama T, LaBonte SA, Siegele DA, Antonazzo G, Attrill H, Brown NH, Garapati P, Marygold SJ, Trovisco V, dos Santos G, Falls K, Tabone C, Zhou P, Goodman JL, Strelets VB, Thurmond J, Garmiri P, Ishtiaq R, Rodríguez-López M, Acencio ML, Kuiper M, Lægreid A, Logie C, Lovering RC, Kramarz B, Saverimuttu SCC, Pinheiro SM, Gunn H, Su R, Thurlow KE, Chibucos M, Giglio M, Nadendla S, Munro J, Jackson R, Duesbury MJ, Del-Toro N, Meldal BHM, Paneerselvam K, Perfetto L, Porras P, Orchard S, Shrivastava A, Chang HY, Finn RD, Mitchell AL, Rawlings ND, Richardson L, Sangrador-Vegas A, Blake JA, Christie KR, Dolan ME, Drabkin HJ, Hill DP, Ni L, Sitnikov DM, Harris MA, Oliver SG, Rutherford K, Wood V, Hayles J, Bähler J, Bolton ER, de Pons JL, Dwinell MR, Hayman GT, Kaldunski ML, Kwitek AE, Laulederkind SJF, Plasterer C, Tutaj MA, Vedi M, Wang SJ, D'Eustachio P, Matthews L, Balhoff JP, Aleksander SA, Alexander MJ, Cherry JM, Engel SR, Gondwe F, Karra K, Miyasato SR, Nash RS, Simison M, Skrzypek MS, Weng S, Wong ED, Feuermann M, Gaudet P, Morgat A, Bakker E, Berardini TZ, Reiser L, Subramaniam S, Huala E, Arighi CN, Auchincloss A, Axelsen K, Argoud-Puy G, Bateman A, Blatter MC, Boutet E, Bowler E, Breuza L, Bridge A, Britto R, Bye-A-Jee H, Casas CC, Coudert E, Denny P, Es-Treicher A, Famiglietti ML, Georghiou G, Gos AN, Gruaz-Gumowski N, Hatton-Ellis E, Hulo C, Ignatchenko A, Jungo F, Laiho K, Le Mercier P, Lieberherr D, Lock A, Lussi Y, MacDougall A, Ma-Grane M, Martin MJ, Masson P, Natale DA, Hyka-Nouspikel N, Orchard S, Pedruzzi I, Pourcel L, Poux S, Pundir S, Rivoire C, Speretta E, Sundaram S, Tyagi N, Warner K, Zaru R, Wu CH, Diehl AD, Chan JN, Grove C, Lee RYN, Muller HM, Raciti D, van Auken K, Sternberg PW, Berriman M, Paulini M, Howe K, Gao S, Wright A, Stein L, Howe DG, Toro S, Westerfield M, Jaiswal P, Cooper L, Elser J. 2021. The Gene Ontology resource: Enriching a GOld mine. Nucleic Acids Research 49: D325–D334.

Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, Jacobsen A, Byrne CJ, Heuer ML, Larsson E, Antipin Y, Reva B, Goldberg AP, Sander C, Schultz N. 2012. The cBio Cancer Genomics Portal: An open platform for exploring multidimensional cancer genomics data. Cancer Discovery 2: 401–404.

Chai WX, Sun LG, Dai FH, Shao HS, Zheng NG, Cai HY. 2019. Inhibition of PRRX2 suppressed colon cancer liver metastasis via inactivation of Wnt/β-catenin signaling pathway. Pathology Research and Practice 215: 152593.

Chan SC, Zhang Y, Pontoglio M, Igarashi P. 2019. Hepatocyte nuclear factor-1β regulates Wnt signaling through genome-wide competition with β-catenin/ lymphoid enhancer binding factor. Proceedings of the National Academy of Sciences of the United States of America 116: 24133–24142.

Chauveinc L, Mosseri V, Quintana E, Desjardins L, Schlienger P, Doz F, Dutrillaux B. 2001. Osteosarcoma following retinoblastoma: Age at onset and latency period. Ophthalmic Genetics 22: 77–88.

Chen C, Zhao M, Tian A, Zhang X, Yao Z, Ma X. 2015. Aberrant activation of Wnt/β-catenin signaling drives proliferation of bone sarcoma cells. Oncotarget 6: 17570–17583.

Chen PM, Wu TC, Cheng YW, Chen CY, Lee H. 2015. NKX2-1-mediated p53 expression modulates lung adenocarcinoma progression via modulating IKKβ/NF-κB activation. Oncotarget 6: 14274–14289.

Chen S, Dallas MR, Balzer EM, Konstantopoulos K. 2012. Mucin 16 is a functional selectin ligand on pancreatic cancer cells. The FASEB Journal 26: 1349–1359.

Chen X, Bahrami A, Pappo A, Easton J, Dalton J, Hedlund E, Ellison D, Shurtleff S, Wu G, Wei L, Parker M, Rusch M, Nagahawatte P, Wu J, Mao S, Boggs K, Mulder H, Yergeau D, Lu C, Ding L, Edmonson M, Qu C, Wang J, Li Y, Navid F, Daw NC, Mardis ER, Wilson RK, Downing JR, Zhang J, Dyer MA. 2014. Recurrent somatic structural variations contribute to tumorigenesis in pediatric osteosarcoma. Cell Reports 7: 104–112.

Cooper GM, Stone EA, Asimenos G, Green ED, Batzoglou S, Sidow A. 2005. Distribution and intensity of constraint in mammalian genomic sequence. Genome Research 15: 901–913.

Cuykendall TN, Rubin MA, Khurana E. 2017. Non-coding genetic variation in cancer. Current Opinion in Systems Biology 1: 9–15.

Dali R, Blanchette M. 2017. A critical assessment of topologically associating domain prediction tools. Nucleic Acids Research 45: 2994–3005.

Daly MJ, Patterson N, Mesirov JP, Golub TR, Tamayo P, Spiegelman B. 2003. PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nature Genetics 34: 267–273.

Davis CA, Hitz BC, Sloan CA, Chan ET, Davidson JM, Gabdank I, Hilton JA, Jain K, Baymuradov UK, Narayanan AK, Onate KC, Graham K, Miyasato SR, Dreszer TR, Strattan JS, Jolanki O, Tanaka FY, Cherry JM. 2018. The Encyclopedia of DNA elements (ENCODE): Data portal update. Nucleic Acids Research 46: D794–D801.

Davydov E V., Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. 2010. Identifying a high fraction of the human genome to be under selective constraint using GERP++. PLoS Computational Biology 6: e1001025.

Dietlein F, Weghorn D, Taylor-Weiner A, Richters A, Reardon B, Liu D, Lander ES, Van Allen EM, Sunyaev SR. 2020. Identification of cancer driver genes based on nucleotide context.

Nature Genetics 52: 208–218.

Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, Doyle F, Epstein CB, Frietze S, Harrow J, Kaul R, Khatun J, Lajoie BR, Landt SG, Lee BK, Pauli F, Rosenbloom KR, Sabo P, Safi A, Sanyal A, Shoresh N, Simon JM, Song L, Trinklein ND, Altshuler RC, Birney E, Brown JB, Cheng C, Djebali S, Dong X, Ernst J, Furey TS, Gerstein M, Giardine B, Greven M, Hardison RC, Harris RS, Herrero J, Hoffman MM, Iyer S, Kellis M, Kheradpour P, Lassmann T, Li Q, Lin X, Marinov GK, Merkel A, Mortazavi A, Parker SCJ, Reddy TE, Rozowsky J, Schlesinger F, Thurman RE, Wang J, Ward LD, Whitfield TW, Wilder SP, Wu W, Xi HS, Yip KY, Zhuang J, Bernstein BE, Green ED, Gunter C, Snyder M, Pazin MJ, Lowdon RF, Dillon LAL, Adams LB, Kelly CJ, Zhang J, Wexler JR, Good PJ, Feingold EA, Crawford GE, Dekker J, Elnitski L, Farnham PJ, Giddings MC, Gingeras TR, Guigó R, Hubbard TJ, Kent WJ, Lieb JD, Margulies EH, Myers RM, Stamatoyannopoulos JA, Tenenbaum SA, Weng Z, White KP, Wold B, Yu Y, Wrobel J, Risk BA, Gunawardena HP, Kuiper HC, Maier CW, Xie L, Chen X, Mikkelsen TS, Gillespie S, Goren A, Ram O, Zhang X, Wang L, Issner R, Coyne MJ, Durham T, Ku M, Truong T, Eaton ML, Dobin A, Tanzer A, Lagarde J, Lin W, Xue C, Williams BA, Zaleski C, Röder M, Kokocinski F, Abdelhamid RF, Alioto T, Antoshechkin I, Baer MT, Batut P, Bell I, Bell K, Chakrabortty S, Chrast J, Curado J, Derrien T, Drenkow J, Dumais E, Dumais J, Duttagupta R, Fastuca M, Fejes-Toth K, Ferreira P, Foissac S, Fullwood MJ, Gao H, Gonzalez D, Gordon A, Howald C, Jha S, Johnson R, Kapranov P, King B, Kingswood C, Li G, Luo OJ, Park E, Preall JB, Presaud K, Ribeca P, Robyr D, Ruan X, Sammeth M, Sandhu KS, Schaeffer L, See LH, Shahab A, Skancke J, Suzuki AM, Takahashi H, Tilgner H, Trout D, Walters N, Wang H, Hayashizaki Y, Reymond A, Antonarakis SE, Hannon GJ, Ruan Y, Carninci P, Sloan CA, Learned K, Malladi VS, Wong MC, Barber GP, Cline MS, Dreszer TR, Heitner SG, Karolchik D, Kirkup VM, Meyer LR, Long JC, Maddren M, Raney BJ, Grasfeder LL, Giresi PG, Battenhouse A, Sheffield NC, Showers KA, London D, Bhinge AA, Shestak C, Schaner MR, Kim SK, Zhang ZZ, Mieczkowski PA, Mieczkowska JO, Liu Z, McDaniell RM, Ni Y, Rashid NU, Kim MJ, Adar S, Zhang Z, Wang T, Winter D, Keefe D, Iyer VR, Zheng M, Wang P, Gertz J, Vielmetter J, Partridge EC, Varley KE, Gasper C, Bansal A, Pepke S, Jain P, Amrhein H, Bowling KM, Anaya M, Cross MK, Muratet MA, Newberry KM, McCue K, Nesmith AS, Fisher-Aylor KI, Pusey B, DeSalvo G, Parker SL, Balasubramanian S, Davis NS, Meadows SK, Eggleston T, Newberry JS, Levy SE, Absher DM, Wong WH, Blow MJ, Visel A, Pennachio LA, Petrykowska HM, Abyzov A, Aken B, Barrell D, Barson G, Berry A, Bignell A, Boychenko V, Bussotti G, Davidson C, Despacio-Reyes G, Diekhans M, Ezkurdia I, Frankish A, Gilbert J, Gonzalez JM, Griffiths E, Harte R, Hendrix DA, Hunt T, Jungreis I, Kay M, Khurana E, Leng J, Lin MF, Loveland J, Lu Z, Manthravadi D, Mariotti M, Mudge J, Mukherjee G, Notredame C, Pei B, Rodriguez JM, Saunders G, Sboner A, Searle S, Sisu C, Snow C, Steward C, Tapanari E, Tress ML, Van Baren MJ, Washietl S, Wilming L, Zadissa A, Zhang Z, Brent M, Haussler D, Valencia A, Addleman N, Alexander RP, Auerbach RK, Balasubramanian S, Bettinger K, Bhardwaj N, Boyle AP, Cao AR, Cayting P, Charos A, Cheng Y, Eastman C, Euskirchen G, Fleming JD, Grubert F, Habegger L, Hariharan M, Harmanci A, Iyengar S, Jin VX, Karczewski KJ, Kasowski M, Lacroute P, Lam H, Lamarre-Vincent N, Lian J, Lindahl-Allen M, Min R, Miotto B, Monahan H, Moqtaderi Z, Mu XJ, O'Geen H, Ouyang Z, Patacsil D, Raha D, Ramirez L, Reed B, Shi M, Slifer T, Witt H, Wu L, Xu X, Yan KK, Yang X, Struhl K, Weissman SM, Penalva LO, Karmakar S, Bhanvadia RR, Choudhury A, Domanus M, Ma L, Moran J, Victorsen A, Auer T, Centanin L, Eichenlaub M,

Gruhl F, Heermann S, Hoeckendorf B, Inoue D, Kellner T, Kirchmaier S, Mueller C, Reinhardt R, Schertel L, Schneider S, Sinn R, Wittbrodt B, Wittbrodt J, Jain G, Balasundaram G, Bates DL, Byron R, Canfield TK, Diegel MJ, Dunn D, Ebersol AK, Frum T, Garg K, Gist E, Hansen RS, Boatman L, Haugen E, Humbert R, Johnson AK, Johnson EM, Kutyavin T V., Lee K, Lotakis D, Maurano MT, Neph SJ, Neri F V., Nguyen ED, Qu H, Reynolds AP, Roach V, Rynes E, Sanchez ME, Sandstrom RS, Shafer AO, Stergachis AB, Thomas S, Vernot B, Vierstra J, Vong S, Wang H, Weaver MA, Yan Y, Zhang M, Akey JM, Bender M, Dorschner MO, Groudine M, MacCoss MJ, Navas P, Stamatoyannopoulos G, Beal K, Brazma A, Flicek P, Johnson N, Lukk M, Luscombe NM, Sobral D, Vaquerizas JM, Batzoglou S, Sidow A, Hussami N, Kyriazopoulou-Panagiotopoulou S, Libbrecht MW, Schaub MA, Miller W, Bickel PJ, Banfai B, Boley NP, Huang H, Li JJ, Noble WS, Bilmes JA, Buske OJ, Sahu AD, Kharchenko P V., Park PJ, Baker D, Taylor J, Lochovsky L. 2012. An integrated encyclopedia of DNA elements in the human genome. Nature 489: 57–74.

Fabregat A, Sidiropoulos K, Garapati P, Gillespie M, Hausmann K, Haw R, Jassal B, Jupe S, Korninger F, McKay S, Matthews L, May B, Milacic M, Rothfels K, Shamovsky V, Webber M, Weiser J, Williams M, Wu G, Stein L, Hermjakob H, D'Eustachio P. 2016. The reactome pathway knowledgebase. Nucleic Acids Research 44: D481–D487.

Fang C, Wang Z, Han C, Safgren SL, Helmin KA, Adelman ER, Eagen KP, Gaspar-Maia A, Figueroa ME, Singer BD, Ratan A, Ntziachristos P, Zang C. 2020. Cancer-specific CTCF binding facilitates oncogenic transcriptional dysregulation. Genome Biology 21: 1–30.

Frankish A, Diekhans M, Ferreira AM, Johnson R, Jungreis I, Loveland J, Mudge JM, Sisu C, Wright J, Armstrong J, Barnes I, Berry A, Bignell A, Carbonell Sala S, Chrast J, Cunningham F, Di Domenico T, Donaldson S, Fiddes IT, García Girón C, Gonzalez JM, Grego T, Hardy M, Hourlier T, Hunt T, Izuogu OG, Lagarde J, Martin FJ, Martínez L, Mohanan S, Muir P, Navarro FCP, Parker A, Pei B, Pozo F, Ruffier M, Schmitt BM, Stapleton E, Suner MM, Sycheva I, Uszczynska-Ratajczak B, Xu J, Yates A, Zerbino D, Zhang Y, Aken B, Choudhary JS, Gerstein M, Guigó R, Hubbard TJP, Kellis M, Paten B, Reymond A, Tress ML, Flicek P. 2019. GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Research 47: D766–D773.

Fredriksson NJ, Ny L, Nilsson JA, Larsson E. 2014. Systematic analysis of noncoding somatic mutations and gene expression alterations across 14 tumor types. Nature Genetics 46: 1258–1264.

Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, Sun Y, Jacobsen A, Sinha R, Larsson E, Cerami E, Sander C, Schultz N. 2013. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Science Signaling 6: 1–20.

Glass DA, Bialek P, Ahn JD, Starbuck M, Patel MS, Clevers H, Taketo MM, Long F, McMahon AP, Lang RA, Karsenty G. 2005. Canonical Wnt signaling in differentiated osteoblasts controls osteoclast differentiation. Developmental Cell 8: 751–764.

Guan D, Tian H. 2017. Integrated network analysis to explore the key genes regulated by parathyroid hormone receptor 1 in osteosarcoma. World Journal of Surgical Oncology 15: 1–10.

Haeussler M, Zweig AS, Tyner C, Speir ML, Rosenbloom KR, Raney BJ, Lee CM, Lee BT, Hinrichs

AS, Gonzalez JN, Gibson D, Diekhans M, Clawson H, Casper J, Barber GP, Haussler D, Kuhn RM, Kent WJ. 2019. The UCSC Genome Browser database: 2019 update. Nucleic Acids Research, doi 10.1093/nar/gky1095.

Han Y, Cai H, Ma L, Ding Y, Tan X, Chang W, Guan W, Liu Y, Shen Q, Yu Y, Zhang H, Cao G. 2013a. Expression of orphan nuclear receptor NR4A2 in gastric cancer cells confers chemoresistance and predicts an unfavorable postoperative survival of gastric cancer patients with chemotherapy. Cancer 119: 3436–3445.

Han Y, Cai H, Ma L, Ding Y, Tan X, Liu Y, Su T, Yu Y, Chang W, Zhang H, Fu C, Cao G. 2013b. Nuclear orphan receptor NR4A2 confers chemoresistance and predicts unfavorable prognosis of colorectal carcinoma patients who received postoperative chemotherapy. European Journal of Cancer 49: 3420–3430.

Hansen MF, Seton M, Merchant A. 2006. Osteosarcoma in Paget's disease of bone. Journal of Bone and Mineral Research 21: P58–P63.

Hill TP, Später D, Taketo MM, Birchmeier W, Hartmann C. 2005. Canonical Wnt/β-catenin signaling prevents osteoblasts from differentiating into chondrocytes. Developmental Cell 8: 727–738.

Hornshøj H, Nielsen MM, Sinnott-Armstrong NA, Świtnicki MP, Juul M, Madsen T, Sallari R, Kellis M, Ørntoft T, Hobolth A, Pedersen JS. 2018. Pan-cancer screen for mutations in non-coding elements with conservation and cancer specificity reveals correlations with expression and survival. npj Genomic Medicine 3: 1–14.

Huang RP, Hossain MZ, Sehgal A, Boynton AL. 1999. Reduced connexin43 expression in high-grade human brain glioma cells. Journal of Surgical Oncology 70: 21–24.

Inamoto T, Czerniak BA, Dinney CP, Kamat AM. 2010. Cytoplasmic mislocalization of the orphan nuclear receptor Nurr1 is a prognostic factor in bladder cancer. Cancer 116: 340–346.

Jheon AH, Ganss B, Cheifetz S, Sodek J. 2001. Characterization of a Novel KRAB/C2H2 Zinc Finger Transcription Factor Involved in Bone Development. Journal of Biological Chemistry 276: 18282–18289.

Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. 2017. KEGG: New perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Research 45: D353–D361.

Kansara M, Thomas DM. 2007. Molecular pathogenesis of osteosarcoma. DNA and Cell Biology 26: 1–18.

Ke N, Claassen G, Yu DH, Albers A, Fan W, Tan P, Grifman M, Hu X, DeFife K, Nguy V, Meyhack B, Brachat A, Wong-Staal F, Li QX. 2004. Nuclear hormone receptor NR4A2 is involved in cell transformation and apoptosis. Cancer Research 64: 8208–8212.

Kent WJ, Zweig AS, Barber G, Hinrichs AS, Karolchik D. 2010. BigWig and BigBed: Enabling browsing of large distributed datasets. Bioinformatics 26: 2204–2207.

Laird DW, Fistouris P, Batist G, Alpert L, Huynh HT, Carystinos GD, Alaoui-Jamali MA. 1999. Deficiency of connexin43 gap junctions is an independent marker for breast tumors. Cancer Research 59: 4104–4110.

Lakshmanan I, Ponnusamy MP, Das S, Chakraborty S, Haridas D, Mukhopadhyay P, Lele SM, Batra SK. 2012. MUC16 induced rapid G2/M transition via interactions with JAK2 for increased proliferation and anti-apoptosis in breast cancer cells. Oncogene 31: 805–817.

Lambert M, Jambon S, Depauw S, David-Cordonnier MH. 2018. Targeting transcription factors for cancer treatment. Molecules, doi 10.3390/molecules23061479.

Lawrence MS, Stojanov P, Polak P, Kryukov G V., Cibulskis K, Sivachenko A, Carter SL, Stewart C, Mermel CH, Roberts SA, Kiezun A, Hammerman PS, McKenna A, Drier Y, Zou L, Ramos AH, Pugh TJ, Stransky N, Helman E, Kim J, Sougnez C, Ambrogio L, Nickerson E, Shefler E, Cortés ML, Auclair D, Saksena G, Voet D, Noble M, Dicara D, Lin P, Lichtenstein L, Heiman DI, Fennell T, Imielinski M, Hernandez B, Hodis E, Baca S, Dulak AM, Lohr J, Landau DA, Wu CJ, Melendez-Zajgla J, Hidalgo-Miranda A, Koren A, McCarroll SA, Mora J, Lee RS, Crompton B, Onofrio R, Parkin M, Winckler W, Ardlie K, Gabriel SB, Roberts CWM, Biegel JA, Stegmaier K, Bass AJ, Garraway LA, Meyerson M, Golub TR, Gordenin DA, Sunyaev S, Lander ES, Getz G. 2013. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature 499: 214–218.

Lazarus KA, Hadi F, Zambon E, Bach K, Santolla MF, Watson JK, Correia LL, Das M, Ugur R, Pensa S, Becker L, Campos LS, Ladds G, Liu P, Evan GI, McCaughan FM, Le Quesne J, Lee JH, Calado D, Khaled WT. 2018. BCL11A interacts with SOX2 to control the expression of epigenetic regulators in lung squamous carcinoma. Nature Communications 9: 1–11.

Lee MK, Choi H, Gil M, Nikodem VM. 2006. Regulation of osteoblast differentiation by Nurr1 in MC3T3-E1 cell line and mouse calvarial osteoblasts. Journal of Cellular Biochemistry 99: 986–994.

Li YJ, Dong BK, Fan M, Jiang WX. 2015. BTG2 inhibits the proliferation and metastasis of osteosarcoma cells by suppressing the PI3K/AKT pathway. International Journal of Clinical and Experimental Pathology 8: 12410–12418.

Liberzon A, Subramanian A, Pinchback R, Thorvaldsdóttir H, Tamayo P, Mesirov JP. 2011. Molecular signatures database (MSigDB) 3.0. Bioinformatics, doi 10.1093/bioinformatics/btr260.

Lindblad-Toh K, Garber M, Zuk O, Lin MF, Parker BJ, Washietl S, Kheradpour P, Ernst J, Jordan G, Mauceli E, Ward LD, Lowe CB, Holloway AK, Clamp M, Gnerre S, Alföldi J, Beal K, Chang J, Clawson H, Cuff J, Di Palma F, Fitzgerald S, Flicek P, Guttman M, Hubisz MJ, Jaffe DB, Jungreis I, Kent WJ, Kostka D, Lara M, Martins AL, Massingham T, Moltke I, Raney BJ, Rasmussen MD, Robinson J, Stark A, Vilella AJ, Wen J, Xie X, Zody MC, Worley KC, Kovar CL, Muzny DM, Gibbs RA, Warren WC, Mardis ER, Weinstock GM, Wilson RK, Birney E, Margulies EH, Herrero J, Green ED, Haussler D, Siepel A, Goldman N, Pollard KS, Pedersen JS, Lander ES, Kellis M, Baldwin J, Bloom T, Chin CW, Heiman D, Nicol R, Nusbaum C, Young S, Wilkinson J, Cree A, Dihn HH, Fowler G, Jhangiani S, Joshi V, Lee S, Lewis LR, Nazareth L V., Okwuonu G, Santibanez J, Delehaunty K, Dooling D, Fronik C, Fulton L, Fulton B, Graves T, Minx P, Sodergren E. 2011. A high-resolution map of human evolutionary constraint using 29 mammals. Nature 478: 476–482.

Loewenstein WR. 1979. Junctional intercellular communication and the control of growth. BBA - Reviews on Cancer 560: 1–65.

Lv ZD, Wang HB, Liu XP, Jin LY, Shen RW, Wang XG, Kong B, Qu HL, Li FN, Yang QF. 2017. Silencing of Prrx2 Inhibits the Invasion and Metastasis of Breast Cancer both in Vitro and in Vivo by Reversing Epithelial-Mesenchymal Transition. Cellular Physiology and Biochemistry 42: 1847–1856.

Manke T, Heinig M, Vingron M. 2010. Quantifying the effect of sequence variation on regulatory interactions. Human Mutation 31: 477–483.

Mansukhani A, Ambrosetti D, Holmes G, Cornivelli L, Basilico C. 2005. Sox2 induction by FGF and FGFR2 activating mutations inhibits Wnt signaling and osteoblast differentiation. Journal of Cell Biology 168: 1065–1076.

Matsuoka K, Bakiri L, Wolff LI, Linder M, Mikels-Vigdal A, Patiño-García A, Lecanda F, Hartmann C, Sibilia M, Wagner EF. 2020. Wnt signaling and Loxl2 promote aggressive osteosarcoma. Cell Research 30: 885–901.

Maurizi G, Verma N, Gadi A, Mansukhani A, Basilico C. 2018. Sox2 is required for tumor development and cancer cell proliferation in osteosarcoma. Oncogene 37: 4626–4632.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. 2010. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research 20: 1297–1303.

Meader S, Ponting CP, Lunter G. 2010. Massive turnover of functional sequence in human and other mammalian genomes. Genome Research 20: 1335–1343.

Mehta PP, Perez-Stable C, Nadji M, Mian M, Asotra K, Roos BA. 1999. Suppression of human prostate cancer cell growth by forced expression of connexin genes. Developmental Genetics 24: 91–110.

Mialou V, Philip T, Kalifa C, Perol D, Gentet JC, Marec-Berard P, Pacquement H, Chastagner P, Defaschelles AS, Hartmann O. 2005. Metastatic osteosarcoma at diagnosis: Prognostic factors and long-term outcome - The French pediatric experience. Cancer 104: 1100–1109.

Mirabello L, Troisi RJ, Savage SA. 2009. Osteosarcoma incidence and survival rates from 1973 to 2004: Data from the surveillance, epidemiology, and end results program. Cancer 115: 1531–1543.

Mirabello L, Yu K, Berndt SI, Burdett L, Wang Z, Chowdhury S, Teshome K, Uzoka A, Hutchinson A, Grotmol T, Douglass C, Hayes RB, Hoover RN, Savage SA. 2011. A comprehensive candidate gene approach identifies genetic variation associated with osteosarcoma. BMC Cancer, doi 10.1186/1471-2407-11-209.

Moisés J, Navarro A, Santasusagna S, Viñolas N, Molins L, Ramirez J, Osorio J, Saco A, Castellano JJ, Muñoz C, Morales S, Monzó M, Marrades RM. 2017. NKX2-1 expression as a prognostic marker in early-stage non-small-cell lung cancer. BMC Pulmonary Medicine 17: 1–9.

Mularoni L, Sabarinathan R, Deu-Pons J, Gonzalez-Perez A, López-Bigas N. 2016. OncodriveFML: A general framework to identify coding and non-coding regions with

cancer driver mutations. Genome Biology 17: 1–13.

Naus CCG, Bechberger JF, Caveney S, Wilson JX. 1991. Expression of gap junction genes in astrocytes and C6 glioma cells. Neuroscience Letters 126: 33–36.

Neph S, Kuehn MS, Reynolds AP, Haugen E, Thurman RE, Johnson AK, Rynes E, Maurano MT, Vierstra J, Thomas S, Sandstrom R, Humbert R, Stamatoyannopoulos JA. 2012. BEDOPS: High-performance genomic feature operations. Bioinformatics 28: 1919–1920.

Nishimura D. 2001. A View from the Web: BioCarta. Biotech Software & Internet Report 2: 117–120.

Nymoen DA, Holth A, Falkentha TEH, Tropé CG, Davidson B. 2015. CIAPIN1 and ABCA13 are markers of poor survival in metastatic ovarian serous carcinoma. Molecular Cancer 14: 1–9.

Odom DT, Dowell RD, Jacobsen ES, Gordon W, Danford TW, MacIsaac KD, Rolfe PA, Conboy CM, Gifford DK, Fraenkel E. 2007. Tissue-specific transcriptional regulation has diverged significantly between human and mouse. Nature Genetics 39: 730–732.

Palmero I, Pantoja C, Serrano M. 1998. p19ARF links the tumour suppressor p53 to Ras. Nature 395: 125–126.

Perry JA, Kiezun A, Tonzi P, Van Allen EM, Carter SL, Baca SC, Cowley GS, Bhatt AS, Rheinbay E, Pedamallu CS, Helman E, Taylor-Weiner A, McKenna A, DeLuca DS, Lawrence MS, Ambrogio L, Sougnez C, Sivachenko A, Walensky LD, Wagle N, Mora J, De Torres C, Lavarino C, Dos Santos Aguiar S, Yunes JA, Brandalise SR, Mercado-Celis GE, Melendez-Zajgla J, Cárdenas-Cardós R, Velasco-Hidalgo L, Roberts CWM, Garraway LA, Rodriguez-Galindo C, Gabriel SB, Lander ES, Golub TR, Orkin SH, Getz G, Janeway KA. 2014. Complementary genomic approaches highlight the PI3K/mTOR pathway as a common vulnerability in osteosarcoma. Proceedings of the National Academy of Sciences of the United States of America 111: E5564–E5573.

Phillips JE, Corces VG. 2009. CTCF: Master Weaver of the Genome. Cell 137: 1194–1211.

Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. 2010. Detection of nonneutral substitution rates on mammalian phylogenies. Genome Research 20: 110–121.

Quinlan AR, Hall IM. 2010. BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics 26: 841–842.

Rajalin AM, Aarnisalo P. 2011. Cross-talk between NR4A orphan nuclear receptors and β-catenin signaling pathway in osteoblasts. Archives of Biochemistry and Biophysics 509: 44–51.

Rands CM, Meader S, Ponting CP, Lunter G. 2014. 8.2% of the Human Genome Is Constrained: Variation in Rates of Turnover across Functional Element Classes in the Human Lineage. PLoS Genetics, doi 10.1371/journal.pgen.1004525.

Rauch DA, Hurchla MA, Harding JC, Deng H, Shea LK, Eagleton MC, Stefan NS, Lairmore MD, Piwnica-Worms D, Rosol TJ, Weber JD, Ratner L, Weilbaecher KN. 2010. The ARF tumor suppressor regulates bone remodeling and osteosarcoma development in mice. PLoS ONE, doi 10.1371/journal.pone.0015755.

Rickel K, Fang F, Tao J. 2017. Molecular genetics of osteosarcoma. Bone 102: 69–79.

Rouault JP, Falette N, Guéhenneux F, Guillot C, Rimokh R, Wang Q, Berthet C, Moyret-Lalle C, Savatier P, Pain B, Shaw P, Berger R, Samarut J, Magaud JP, Ozturk M, Samarut C, Puisieux A. 1996. Identification of BTG2, an antiproliferative p53-dependent component of the DNA damage cellular response pathway. Nature Genetics 14: 482–486.

Sadikovic B, Yoshimoto M, Chilton-MacNeill S, Thorner P, Squire JA, Zielenska M. 2009. Identification of interactive networks of gene expression associated with osteosarcoma oncogenesis by integrated molecular profiling. Human Molecular Genetics 18: 1962–1975.

Saito RA, Watabe T, Horiguchi K, Kohyama T, Saitoh M, Nagase T, Miyazono K. 2009. Thyroid transcription factor-1 inhibits transforming growth factor-ß-mediated epithel ial-to-mesenchymal transition in lung adenocarcinoma cells. Cancer Research 69: 2783–2791.

Saito T, Nishimura M, Kudo R, Yamasaki H. 2001. Suppressed gap junctional intercellular communication in carcinogenesis of endometrium. International Journal of Cancer 93: 317–323.

Sakthikumar S, Sakthikumar S, Roy A, Haseeb L, Pettersson ME, Sundström E, Marinescu VD, Lindblad-Toh K, Lindblad-Toh K, Forsberg-Nilsson K. 2020. Whole-genome sequencing of glioblastoma reveals enrichment of non-coding constraint mutations in known and novel genes. Genome Biology 21: 1–22.

Satterwhite E, Sonoki T, Willis TG, Harder L, Nowak R, Arriola EL, Liu H, Price HP, Gesk S, Steinemann D, Schlegelberger B, Oscier DG, Siebert R, Tucker PW, Dyer MJS. 2001. The BCL11 gene family: Involvement of BCL11A in lymphoid malignancies. Blood 98: 3413–3420.

Schaefer CF, Anthony K, Krupa S, Buchoff J, Day M, Hannay T, Buetow KH. 2009. PID: The pathway interaction database. Nucleic Acids Research 37: 674–679.

Schmidt D, Wilson MD, Ballester B, Schwalie PC, Brown GD, Marshall A, Kutter C, Watt S, Martinez-jimenez CP, Mackay S, Talianidis I, Flicek P, Odom DT. 2010. Five-Vertebrate ChIP-seq Reveals the Evolutionary Dynamics of Transcription Factor Binding. Science 328: 1036–1040.

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

Shi H, Li C, Feng W, Yue J, Song J, Peng A, Wang H. 2020. BCL11A Is Oncogenic and Predicts Poor Outcomes in Natural Killer/T-Cell Lymphoma. Frontiers in Pharmacology 11: 1–10.

Shigeishi H, Higashikawa K, Hatano H, Okui G, Tanaka F, Tran TT, Rizqiawan A, Ono S, Tobiume K, Kamata N. 2011. PGE2 targets squamous cell carcinoma cell with the activated epidermal growth factor receptor family for survival against 5-fluorouracil through NR4A2 induction. Cancer Letters 307: 227–236.

Shin H, Shi Y, Dai C, Tjong H, Gong K, Alber F, Zhou XJ. 2015. TopDom: An efficient and deterministic method for identifying topological domains in genomes. Nucleic Acids Research 44: 1–13.

Siegel RL, Miller KD, Fuchs HE, Jemal A. 2021. Cancer Statistics, 2021. CA: A Cancer Journal for Clinicians 71: 7–33.

Slenter DN, Kutmon M, Hanspers K, Riutta A, Windsor J, Nunes N, Mélius J, Cirillo E, Coort SL, DIgles D, Ehrhart F, Giesbertz P, Kalafati M, Martens M, Miller R, Nishida K, Rieswijk L, Waagmeester A, Eijssen LMT, Evelo CT, Pico AR, Willighagen EL. 2018. WikiPathways: A multifaceted pathway database bridging metabolomics to other omics research. Nucleic Acids Research 46: D661–D667.

Sondka Z, Bamford S, Cole CG, Ward SA, Dunham I, Forbes SA. 2018. The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. Nature Reviews Cancer 18: 696–705.

Stratton MR, Campbell PJ, Futreal PA. 2009. The cancer genome. Nature 458: 719–724.

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. 2005. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences of the United States of America, doi 10.1073/pnas.0506580102.

Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, Bray F. 2021. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA: A Cancer Journal for Clinicians 71: 209–249.

Talbot J, Lamora A, Stresing V, Verrecchia F. 2015. Gap junction in bone remodeling and in primary bone tumors: Osteosarcoma and Ewing sarcoma, Second Edi. Bone Cancer: Primary Bone Cancers and Bone Metastases: Second Edition, doi 10.1016/B978-0-12-416721-6.00008-X.

Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, Boutselakis H, Cole CG, Creatore C, Dawson E, Fish P, Harsha B, Hathaway C, Jupe SC, Kok CY, Noble K, Ponting L, Ramshaw CC, Rye CE, Speedy HE, Stefancsik R, Thompson SL, Wang S, Ward S, Campbell PJ, Forbes SA. 2019. COSMIC: The Catalogue Of Somatic Mutations In Cancer. Nucleic Acids Research, doi 10.1093/nar/gky1015.

The Gene Ontology Consortium, Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, MartinRingwald, Rubin GM, Sherlock G. 2000. Gene Ontology: tool for the unification of biology. Nature Genetics 25: 25–29.

Thériault C, Pinard M, Comamala M, Migneault M, Beaudin J, Matte I, Boivin M, Piché A, Rancourt C. 2011. MUC16 (CA125) regulates epithelial ovarian cancer cell growth, tumorigenesis and metastasis. Gynecologic Oncology 121: 434–443.

Thomas-Chollier M, Hufton A, Heinig M, O'Keeffe S, Masri N El, Roider HG, Manke T, Vingron M. 2011. Transcription factor binding predictions using TRAP for the analysis of ChIP-seq data and regulatory SNPs. Nature Protocols 6: 1860–1869.

Varley JM. 2003. Germline TP53 mutations and Li-Fraumeni syndrome. Human Mutation, doi 10.1002/humu.10185.

Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA, Kinzler KW. 2013. Cancer genome landscapes. Science 340: 1546–1558.

Walkley CR, Qudsi R, Sankaran VG, Perry JA, Gostissa M, Roth SI, Rodda SJ, Snay E, Dunning P, Fahey FH, Alt FW, McMahon AP, Orkin SH. 2008. Conditional mouse osteosarcoma, dependent on p53 loss and potentiated by loss of Rb, mimics the human disease. Genes and Development 22: 1662–1676.

Wang J, Wang J, Yang L, Zhao C, Wu LN, Xu L, Zhang F, Weng Q, Wegner M, Lu QR. 2020. CTCF-mediated chromatin looping in EGR2 regulation and SUZ12 recruitment critical for peripheral myelination and repair. Nature Communications, doi 10.1038/s41467-020-17955-2.

Wittkopp PJ, Kalay G. 2012. Cis-regulatory elements: Molecular mechanisms and evolutionary processes underlying divergence. Nature Reviews Genetics 13: 59–69.

Wu Z, Shi W, Jiang C. 2018. Overexpressing circular RNA hsa_circ_0002052 impairs osteosarcoma progression via inhibiting Wnt/β-catenin pathway by regulating miR-1205/APC2 axis. Biochemical and Biophysical Research Communications 502: 465–471.

Yang D, Jang I, Choi J, Kim MS, Lee AJ, Kim H, Eom J, Kim D, Jung I, Lee B. 2018. 3DIV: A 3D-genome Interaction Viewer and database. Nucleic Acids Research 46: D52–D57.

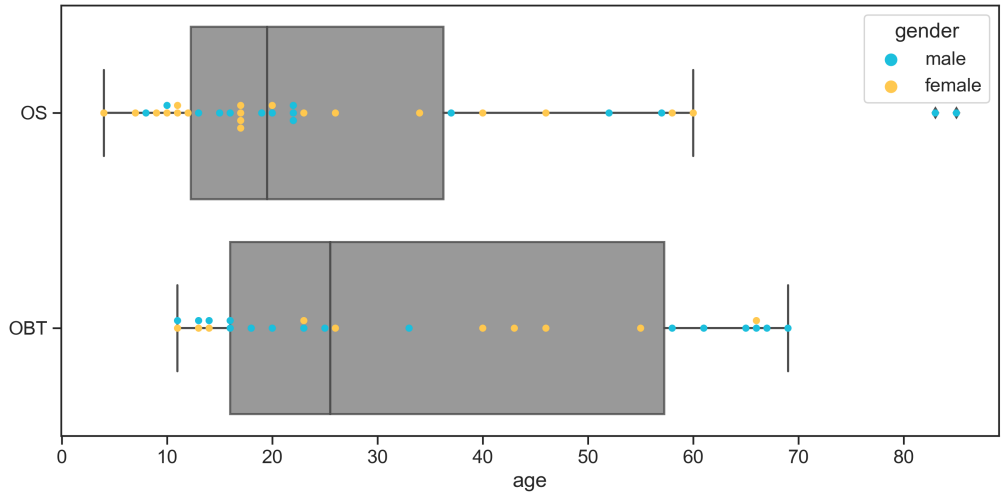
Yuan H, Corbi N, Basilico C, Dailey L. 1995. Developmental-specific activity of the FGF-4 enhancer requires the synergistic action of Sox2 and Oct-3. Genes and Development 9: 2635–2645.

Zhang T, Wang P, Ren H, Fan J, Wang G. 2009. NGFI-B Nuclear Orphan Receptor Nurr1 Interacts with p53 and Suppresses Its Transcriptional Activity. 7: 1408–1416.

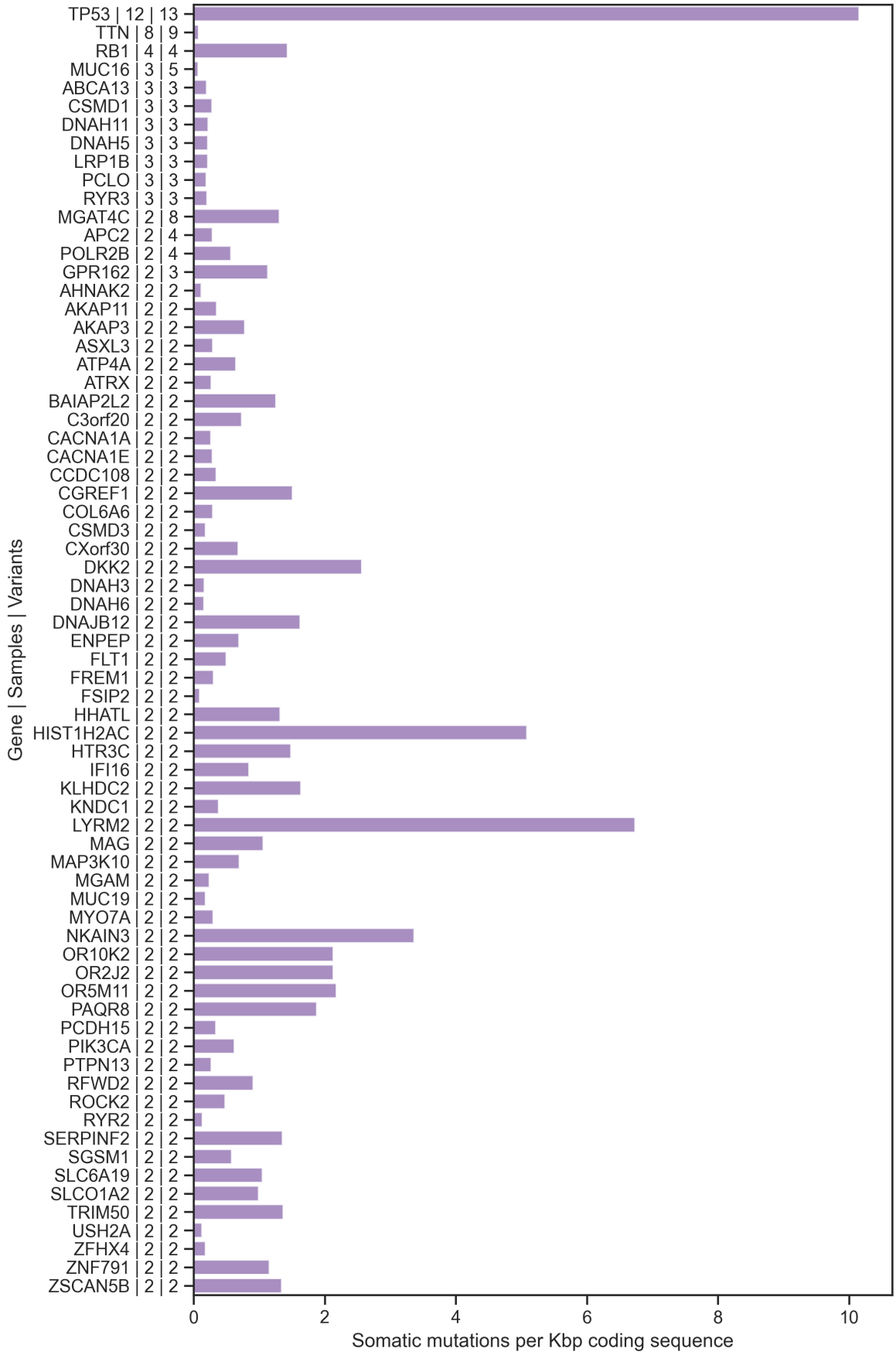
Zhang YW, Morita I, Ikeda M, Ma KW, Murota S. 2001. Connexin43 suppresses proliferation of osteosarcoma U2OS cells through post-transcriptional regulation of p27. Oncogene 20: 4138–4149.

Zhu L, Pan R, Zhou D, Ye G, Tan W. 2019. BCL11A enhances stemness and promotes progression by activating Wnt/β-catenin signaling in breast cancer. Cancer Management and Research 11: 2997–3007.

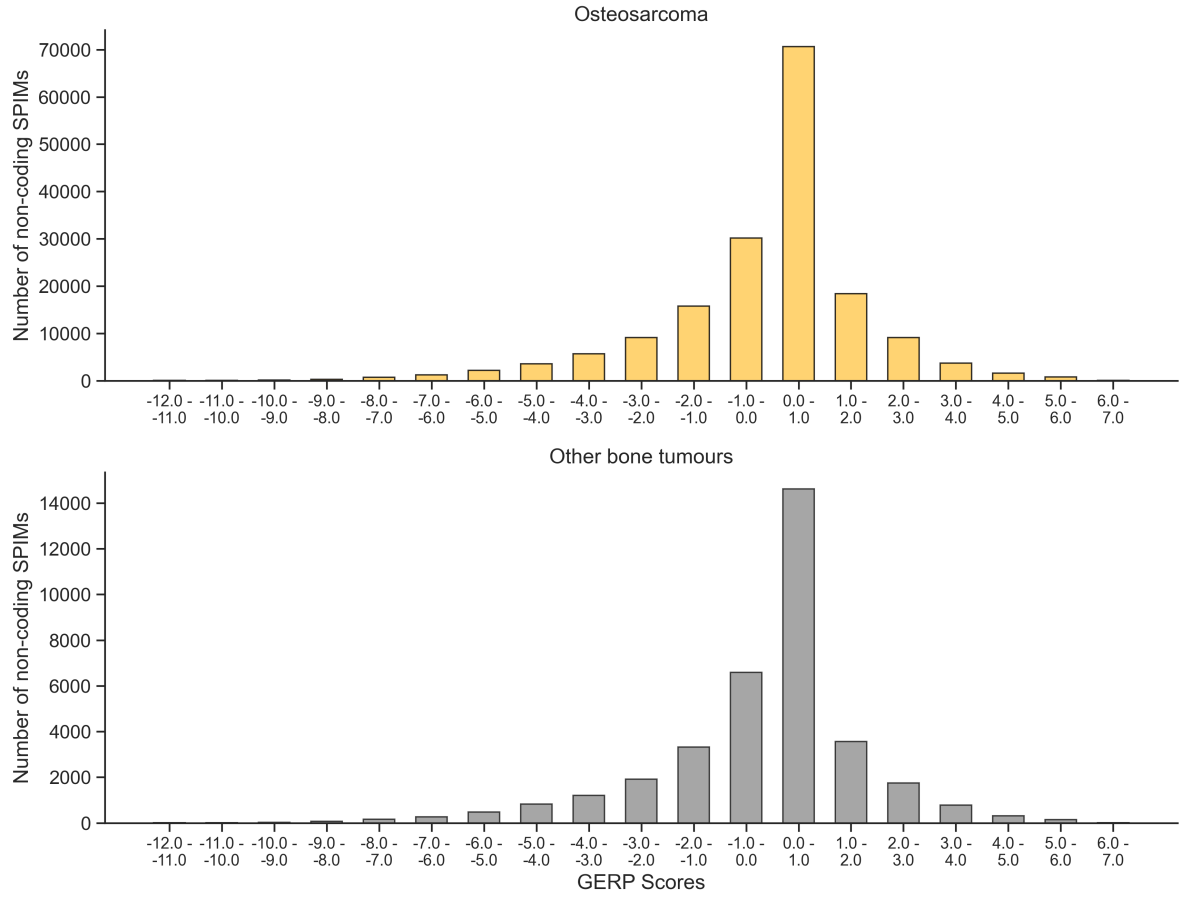
# Appendix




**Appendix Figure 1 | Age and gender distribution of ICGC bone tumour patient cohort for osteosarcoma (OS) and other bone tumours (OBTs).**


TP53 | 12 | 13
TTN | 8 | 9
RB1 | 4 | 4
MUC16 | 3 | 5
ABCA13 | 3 | 3
CSMD1 | 3 | 3
DNAH11 | 3 | 3
DNAH5 | 3 | 3
LRP1B | 3 | 3
PCLO | 3 | 3
RYR3 | 3 | 3
MGAT4C | 2 | 8
APC2 | 2 | 4
POLR2B | 2 | 4
GPR162 | 2 | 3
AHNAK2 | 2 | 2
AKAP11 | 2 | 2
AKAP3 | 2 | 2
ASXL3 | 2 | 2
ATP4A | 2 | 2
ATRX | 2 | 2
BAIAP2L2 | 2 | 2
C3orf20 | 2 | 2
CACNA1A | 2 | 2
CACNA1E | 2 | 2
CCDC108 | 2 | 2
CGREF1 | 2 | 2
COL6A6 | 2 | 2
CSMD3 | 2 | 2
CXorf30 | 2 | 2
DKK2 | 2 | 2
DNAH3 | 2 | 2
DNAH6 | 2 | 2
DNAJB12 | 2 | 2
ENPEP | 2 | 2
FLT1 | 2 | 2
FREM1 | 2 | 2
FSIP2 | 2 | 2
HHATL | 2 | 2
HIST1H2AC | 2 | 2
HTR3C | 2 | 2
IFI16 | 2 | 2
KLHDC2 | 2 | 2
KNDC1 | 2 | 2
LYRM2 | 2 | 2
MAG | 2 | 2
MAP3K10 | 2 | 2
MGAM | 2 | 2
MUC19 | 2 | 2
MYO7A | 2 | 2
NKAIN3 | 2 | 2
OR10K2 | 2 | 2
OR2J2 | 2 | 2
OR5M11 | 2 | 2
PAQR8 | 2 | 2
PCDH15 | 2 | 2
PIK3CA | 2 | 2
PTPN13 | 2 | 2
RFWD2 | 2 | 2
ROCK2 | 2 | 2
RYR2 | 2 | 2
SERPINF2 | 2 | 2
SGSM1 | 2 | 2
SLC6A19 | 2 | 2
SLCO1A2 | 2 | 2
TRIM50 | 2 | 2
USH2A | 2 | 2
ZFHX4 | 2 | 2
ZNF791 | 2 | 2
ZSCAN5B | 2 | 2
Gene | Samples | Variants
0
2
4
6
8
10
Somatic mutations per Kbp coding sequence

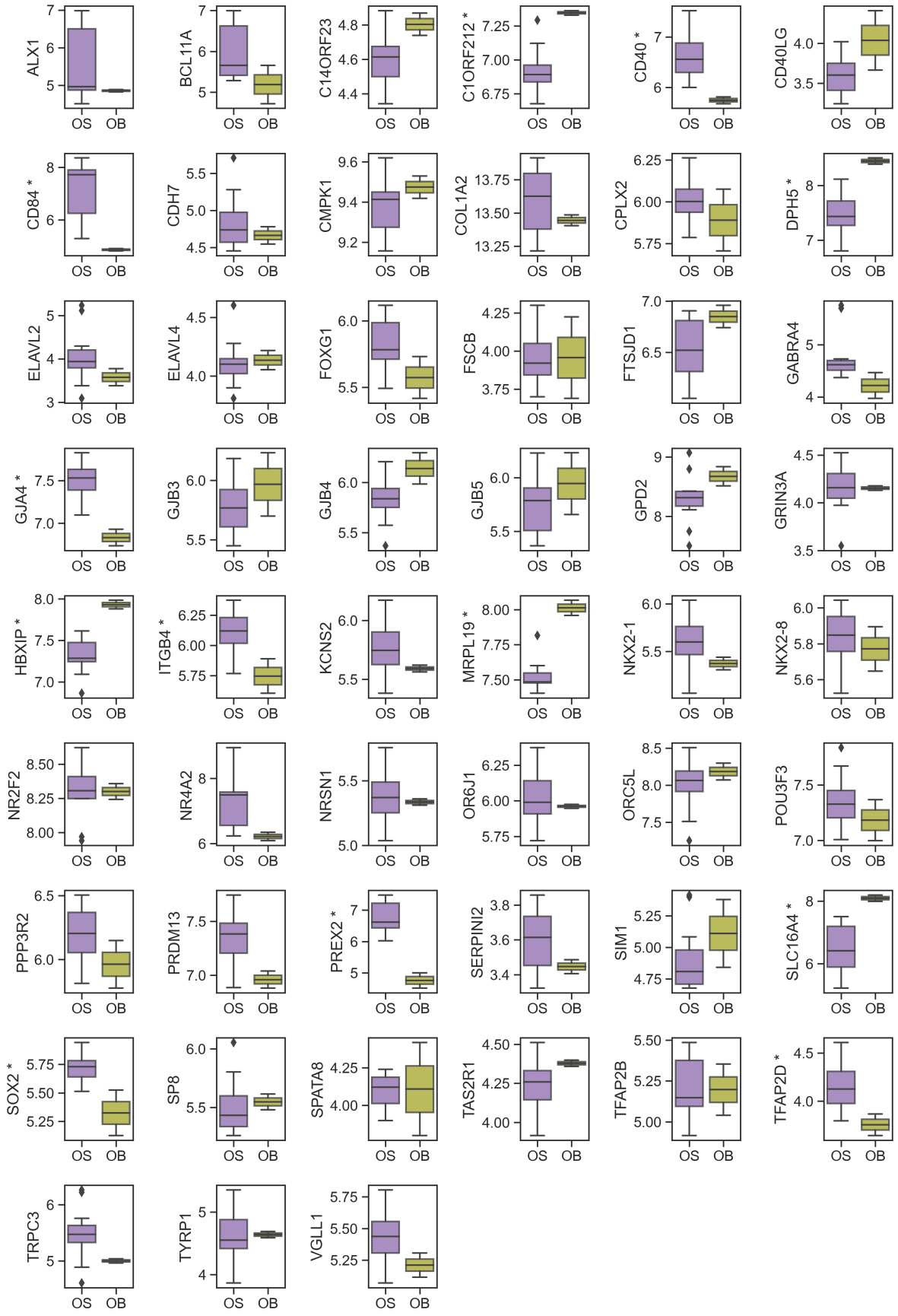
**Appendix Figure 2 | Recurrently mutated genes.**
**Recurrently mutated genes have ≥ 2 protein-coding mutations and are shown with the number of unique samples with mutations in the gene, the total number of mutations in that gene in the osteosarcoma cohort and the total number of mutations normalized with the gene's coding-sequence length.**




**Appendix Figure 3 | GERP score distributions of osteosarcoma and other bone tumours.**
**The distribution of GERP scores across non-coding mutations is comparable between osteosarcoma and other bone tumour samples.**

**Appendix Table 1 | NCCM enrichment analysis.**
**Genes with an enrichment of NCCMs are shown with the total number of NCCMs in the associated non-coding regions, the number of unique samples with NCCMs and the NCCM rate (NCCMs/100 Kbp).**

| Gene | NCCMs | Samples | NCCMs/ 100 Kbp | Gene | NCCMs | Samples | NCCMs/ 100 Kbp |
|---|---|---|---|---|---|---|---|
| *SOX2* | 9 | 8 | 4.0 | *RP11-386G21.1* | 5 | 5 | 2.4 |
| *NKX2-8* | 8 | 7 | 3.5 | *GPD2* | 11 | 9 | 2.4 |
| *IZUMO3* | 8 | 7 | 3.5 | *ELAVL2* | 8 | 8 | 2.4 |
| *NR4A2* | 7 | 7 | 3.2 | *CD40* | 6 | 5 | 2.4 |
| *COL1A2* | 7 | 7 | 3.0 | *AL139147.1* | 5 | 5 | 2.4 |
| *AC016251.1* | 6 | 6 | 3.0 | *CD40LG* | 5 | 5 | 2.4 |
| *GJA4* | 6 | 6 | 3.0 | *NR2F2* | 5 | 5 | 2.3 |
| *NKX2-1* | 7 | 6 | 3.0 | *ALX1* | 5 | 5 | 2.3 |
| *GJB3* | 6 | 6 | 2.9 | *TYRP1* | 5 | 5 | 2.2 |
| *TFAP2D* | 8 | 7 | 2.7 | *VGLL1* | 5 | 5 | 2.2 |
| *BCL11A* | 11 | 8 | 2.7 | *TRPC3* | 6 | 6 | 2.2 |
| *GABRA4* | 7 | 7 | 2.6 | *SLC16A4* | 5 | 5 | 2.2 |
| *ELAVL4* | 12 | 9 | 2.5 | *TFAP2B* | 6 | 5 | 2.2 |
| *SIM1* | 9 | 7 | 2.5 | *NRSN1* | 5 | 5 | 2.2 |
| *AL109659.1* | 5 | 5 | 2.5 | *GRIN3A* | 8 | 8 | 2.2 |
| *C14orf23* | 6 | 6 | 2.5 | *ITGB4* | 10 | 5 | 2.2 |
| *FSCB* | 5 | 5 | 2.5 | *ORC5L* | 8 | 6 | 2.1 |
| *SP8* | 5 | 5 | 2.5 | *DPH5* | 5 | 5 | 2.1 |
| *SPATA8* | 5 | 5 | 2.5 | *CDH7* | 8 | 7 | 2.1 |
| *TAS2R1* | 7 | 7 | 2.5 | *RP11-298I3.5* | 5 | 5 | 2.1 |
| *FOXG1* | 5 | 5 | 2.5 | *SERPINI2* | 5 | 5 | 2.1 |
| *GJB5* | 5 | 5 | 2.5 | *CD84* | 5 | 5 | 2.1 |
| *GJB4* | 5 | 5 | 2.5 | *SFTA3* | 6 | 5 | 2.1 |
| *PPP3R2* | 5 | 5 | 2.5 | *PREX2* | 12 | 10 | 2.1 |
| *RP11-386G21.2* | 5 | 5 | 2.5 | *MRPL19* | 5 | 5 | 2.1 |
| *POU3F3* | 5 | 5 | 2.5 | *CMPK1* | 5 | 5 | 2.0 |
| *KCNS2* | 6 | 5 | 2.5 | *SMIM12* | 7 | 7 | 2.0 |
| *CPLX2* | 8 | 7 | 2.4 | *AC011308.1* | 5 | 4 | 2.0 |
| *HBXIP* | 5 | 5 | 2.4 | *AC087477.1* | 4 | 4 | 2.0 |
| *FTSJD1* | 5 | 5 | 2.4 | *FLJ00388* | 4 | 4 | 2.0 |
| *PRDM13* | 5 | 5 | 2.4 | *OR6J1* | 6 | 4 | 2.0 |
| *LINC00923* | 9 | 8 | 2.4 | *TSRM* | 5 | 4 | 2.0 |


ALX1
BCL11A
C14ORF23
C1ORF212 *
CD40 *
CD40LG
CD84 *
CDH7
CMPK1
COL1A2
CPLX2
DPH5 *
ELAVL2
ELAVL4
FOXG1
FSCB
FTSJD1
GABRA4
GJA4 *
GJB3
GJB4
GJB5
GPD2
GRIN3A
HBXIP *
ITGB4 *
KCNS2
MRPL19 *
NKX2-1
NKX2-8
NR2F2
NR4A2
NRSN1
OR6J1
ORC5L
POU3F3
PPP3R2
PRDM13
PREX2 *
SERPINI2
SIM1
SLC16A4 *
SOX2 *
SP8
SPATA8
TAS2R1
TFAP2B
TFAP2D *
TRPC3
TYRP1
VGLL1
OS
OB

**Appendix Figure 4 | Expression of genes with an enrichment of NCCMs.**
**Gene expression of genes with an NCCM enrichment was compared between osteosarcoma (OS) and osteoblasts (OB). Significant differences were determined with Student's t-test (*).**