The role of brachyury in the pathogenesis of chordoma

A thesis submitted to University College London (UCL) for the degree of Doctor of Philosophy

Nischalan Pillay

2013

UCL Cancer Institute, London, UK

I, Nischalan Pillay, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Abstract

Chordoma is a rare malignant tumour of bone, the molecular marker of which is the expression of the transcription factor *T* (also referred to as brachyury). Silencing of T induces growth arrest in chordoma cell lines; however its downstream genomic targets are unknown. In this thesis I have identified these targets with validation in human chordoma samples. This was achieved by using an integrated functional genomics approach involving shRNA-mediated brachyury knockdown, gene expression microarray, ChIP-seq experiments and bioinformatics analyses. The results show that T regulates a downstream network that involves key cell cycle related genes amongst other potential oncogenic programmes. This is the first documentation of genomic T targets in humans and provides a molecular insight into the pathogenesis of chordoma.

In a separate but related piece of work, a genetic association study was conducted to determine the genetic susceptibility determinants in patients with sporadic chordomas. Whole-exome and Sanger sequencing of *T* exons revealed a strong risk association with the common non-synonymous SNP rs2305089 in chordoma. The degree of risk imparted by this variant is large and is an exceptional finding in cancer genetics.

Overall the work presented in this thesis contributes to the evolving understanding of T's role in the pathogenesis of this rare bone cancer.

Acknowledgments

I owe an enormous debt of gratitude to my supervisor, Prof Adrienne Flanagan for granting me the opportunity to pursue a doctoral research degree when conventional wisdom may have suggested otherwise. I continue to benefit immensely from her mentorship and tutelage.

I am grateful to all members of Adrienne's group, past and present in the UCL Cancer Institute and the Royal National Orthopaedic Hospital. There are too many to mention by name but there is no doubt that the questions that I sought to address in this thesis would not have been possible without the seminal and erudite discoveries made by both previous and current group members. Moreover the easy sharing of the varied expertise within the group has allowed me to broaden my knowledge- base on areas of molecular biology and pathology that would not have been possible otherwise. I am also grateful to my secondary supervisor Prof Claudio Stern whose all-round passion for designing "clean" experiments to help unravel fundamental biological mechanisms through the lens of a gene or a cell is infectious. Although he may not have known it, it was through valuable discussions with him at the infancy of my thesis that I was "bitten by the research bug".

I am also appreciative of the collaborative research environment provided at the UCL Cancer Institute by Prof Chris Boshoff. I have found sharing the laboratory space with his group both stimulating and beneficial. In particular I am grateful to Dr Leonid Nikitenko. Leonid is an oracle, disarmingly in multiple facets of science and life, through whom I have avoided many a rookie mistake by talking through my experiments for the day. He is meticulous and his scepticism,

fastidiousness and perfectionism provided the perfect foil for my inexperience. Although I hope to have paid this back in part through providing some histopathology insights through collaborative projects, I fear the balance of favour weighs heavily on his side. I am grateful that our paths have crossed. He has become a close friend.

It is said that “no man is an island” and I have learned that is no truer than in science. In that vein I am indebted to all the collaborators I have worked with on the various projects that make up this thesis. They are all named co-authors in the published papers. In particular, I would like to thank Drs Fiona Wardle and Andrew Nelson for their invaluable assistance and expertise in helping generate and analysing the data for the ChIP-seq experiment. This project was initiated when ChIP-seq was just about becoming a mainstream technique and much of the analytical and bio-informatic approaches were rapidly evolving over this period. That it took us more than 18 months from the time of generation of data to finalising the analysis and publishing bears testament to this.

I am eternally grateful to my maternal aunt Sundree and my parents for their encouragement and all their sacrifices on my behalf. I am grateful also to my siblings for tolerating me pursuing my dream away from home and entrusting many domestic responsibilities upon them.

Lastly, I would not have been able to undertake this PhD without the moral and financial support of my wife, Ula. She has provided me with the space and time to fully commit myself to satisfying this rather self-indulgent exercise. Moreover,

her unwavering love and faith carried me through my many moments of self-doubt. She is the greatest person that I know.

"Fate gives us all three teachers, three friends, three enemies and three great loves in our lives. But these twelve are always disguised and we can never know which one is which until we've loved them, left them or fought them." Gregory David Roberts, Shantaram, 2007

Publications

My thesis is partly based on the following journal articles, which are appropriately referenced in the main body of the thesis and reference list. Permission for use of the figures and content from these articles has been obtained from the publishers where necessary.

1. Presneau N, Shalaby A, Ye H, **Pillay N**, Halai D, Idowu B, Tirabosco R, Whitwell D, Jacques TS, Kindblom LG, Bruderlein S, Moller P, Leithner A, Liegl B, Amary FM, Athanasou NN, Hogendoorn PC, Mertens F, Szuhai K, Flanagan AM: **Role of the transcription factor T (brachyury) in the pathogenesis of sporadic chordoma: a genetic and functional-based study**, J Pathol 2011, 223:327-335

2. **Pillay N**, Amary FM, Berisha F, Tirabosco R, Flanagan AM: **P63 does not regulate brachyury expression in human chordomas and osteosarcomas**, Histopathology 2011, 59:1025-1027

3. Nelson AC and **Pillay N***, Henderson S, Presneau N, Tirabosco R, Halai D, Berisha F, Flicek P, Stemple DL, Stern C, Wardle FC, Flanagan AM: **An integrated functional genomics approach identifies the regulatory network directed by brachyury (T) in chordoma**. J Pathol. 2012 Nov;228(3):274-85.

- *Joint first authors

4. **Pillay N**, Plagnol V, Tarpey PS, Lobo SB, Presneau N, Szuhai K, Halai D, Berisha F, Cannon SR, Mead S, Kasperaviciute D, Palmen J, Talmud P, Kindblom LG, Amary FM, Tirabosco R, Flanagan AM: **A common single nucleotide variant in *T* is strongly associated with chordoma**. Nat Genet. 2012 Nov;44(11):1185-7

Contents

List of figures

List of tables

List of abbreviations

3D	Three dimensional
ACAN	Aggrecan
ADAMTS3	ADAM metallopeptidase with thrombospondin type 1 motif, 3
AKR1B10	Aldo-keto reductase family 1, member B10
AKT2	v-akt murine thymoma viral oncogene homolog 2
APC	Adenomatous polyposis coli
ARID1A	AT rich interactive domain 1A (SWI-like)
BMP 4/6	Bone morphogenetic protein 4/6
BNCT	Benign notochordal cell tumour
BRAF	v-raf murine sarcoma viral oncogene homolog B1
BSA	Bovine serum albumin
BUB1	Budding uninhibited by benzimidazoles 1 homolog
BUB1B	Budding uninhibited by benzimidazoles 1 homolog beta
cDNA	Complementary DNA
C-KIT	v-kit Hardy-Zuckerman 4 feline sarcoma viral oncogene homolog
C-MET	met proto-oncogene (hepatocyte growth factor receptor)
CCNA2	Cyclin A2
CD24	CD24 antigen (small cell lung carcinoma cluster 4 antigen)
CDKN2A/2B	Cyclin-dependent kinase inhibitor 2A/2B
CHST4	Carbohydrate (N-acetylglucosamine 6-O) sulfotransferase 4
CK18	Keratin 18
CNTN1	Contactin 1
COL6A3	Collagen, type VI, alpha 3
CTCF	CCCTC-binding factor (zinc finger protein)
CTGF	Connective tissue growth factor
CTNNB1	Catenin (cadherin-associated protein), beta 1
CNTNAP2	Contactin associated protein-like 2
CX3CR1	Chemokine (C-X3-C motif) receptor 1
DNA	Deoxyribonucleic acid
DDR1	Discoidin domain receptor tyrosine kinase 1
ECM	Extracellular matrix
EGF /R	Epidermal growth factor / receptor
EMT	Epthelial-mesenchymal transition
EPYC	Epiphycan
ETS	ETS transcription factor
ETV 1/4/5	Ets variant 1/4/5
FBXW7	F-box and WD repeat domain containing 7
FDR	False discovery rate

FGF1	Fibroblast growth factor 1
FGFR1/2/3/4	Fibroblast growth factor receptor 1/2/3/4
FISH	Fluorescence in situ hybridisation
FLT3	fms-related tyrosine kinase 3
FN1	Fibronectin 1
FOXA	Forkhead box A
FOXD3	Forkhead box D3
FZD4	Frizzled family receptor 4
GAPDH	Glyceraldehyde-3-phosphate dehydrogenase
GATA	GATA binding protein
GFP	Green fluorescent protein
GNR	Giant notochordal rest
GRN	Gene regulatory network
HAPLN1	Hyaluronan and proteoglycan link protein 1
HEK293T	Human embryonic kidney cell line
HeLa	Cervical carcinoma cell line
HIF1A	Hypoxia inducible factor 1, alpha subunit
HOPX	HOP homeobox
HT1080	Fibrosarcoma cell line
H&E	Haematoxylin and eosin
ICGC	International Cancer Genome Consortium
IGF1-R	Insulin-like growth factor 1/ receptor
IL-8	Interleukin 8
IMDM	Iscove's Modified Dulbecco's medium
INI1	SWI/SNF related, actin dependent regulator of chromatin
ITGA3	Integrin, alpha 3
JAK2	Janus kinase 2
KD	Knockdown
KIF11/15	Kinesin family member 11/15
KLF4	Kruppel-like factor 4
KRAS	v-Ki-ras2 Kirsten rat sarcoma viral oncogene homolog
KRT 15/18/19	Keratin 15/18/19
LEF1	Lymphoid enhancer-binding factor 1
LIM	Lim homeodomain transcription factor
MAPK	Mitogen activated kinase-like protein
MMP16	Matrix metallopeptidase 16
MRI	Magnetic resonance imaging
MSGN1	Mesogenin 1
MYOD1	Myogenic differentiation 1
NANOG	Nanog homeobox
NEK2	NIMA (never in mitosis gene a)-related kinase 2
NF-kβ	Nuclear factor of kappa light polypeptide gene enhancer in B-cells
NHS	National Health Service

NKX2-5	NK2 homeobox 5
NRAS	Neuroblastoma RAS viral (v-ras) oncogene homolog
NUSAP1	Nucleolar and spindle associated protein 1
OCT4	POU class 5 homeobox 1
OLFM4	Olfactomedin 4
PBS	Phosphate buffered saline
PCR	Polymerase chain reaction
PDGFRB	Platelet-derived growth factor receptor, beta
PEA3	ETS-domain transcription factor pea3
PI3K	Phosphatidylinositol-4,5-bisphosphate 3-kinase
PI3KCA	hosphatidylinositol-4,5-bisphosphate 3-kinase, Subunit alpha
PKP2	Plakophilin 2
PTEN	Phosphatase and tensin homolog
PTN	Pleiotrophin
P53	Tumour protein p53
qRT-PCR	Real time quantitative reverse transcriptase PCR
RNA	Ribonucleic acid
RPMI	Roswell Park Memorial Institute
S100	S100 calcium binding protein
SD	Standard deviation
shRNA	Short hairpin RNA
SHH	Sonic Hedgehog
siRNA	Short interfering RNA
SMAD1	SMAD family member 1
SNP	Single nucleotide pleomorphism
T*	T, brachyury homolog (mouse)
TBX	T-box
TCF	Transcription factor
TGFA	Transforming growth factor alpha
TF	Transcription factor
TSPAN7	Tetraspanin 7
UCL	University College London
U-CH1 /2	Chordoma cell lines
VNTR	Variable number tandem repeat
WNT 3/8A/11	Wingless-type MMTV integration site family, member 3/8a/11
WRN	Werner syndrome, RecQ helicase-like
Xbra	Xenopus brachyury

***In order to distinguish mouse and human brachyury which share the same gene symbol T, the species designation for mouse has been added as a prefix to the gene symbol in parentheses in the main text i.e. (MOUSE) T as per Mouse Genome Informatic guidelines**.

Chapter 1

General Introduction

Chordomas, like most cancers are *"Borg"*[*] like. They are insidious, possessing interminable growth, co-opting the normal physiological functions of the cell to their own end - survival, expansion, and colonisation. The chief regulator in this relentless process in chordomas is a gene that is crucial for the development of all vertebrate life, the transcription factor brachyury.

See footnote [2]

[2]The Borg are a fictional pseudo-species from the science fiction series Star Trek. They are cybernetic organisms that assimilate other species into their collective, incorporating their genetic makeup to enhance their own fitness with the aim of universal colonisation and the attainment of perfection.

1. GENERAL INTRODUCTION

1.1 Background

Primary malignant tumours of bone are rare and account for approximately 1% of all cancers [1]. Chordomas constitute between 3-6% of these malignant bone tumours but are the most common of the vertebral column [2, 3]. Apart from this location they can occur in the clivus (base of skull) and rarely in extra-axial locations [4]: an exceptional example of a chordoma arising in the background of an ovarian teratoma has been reported [5]. They are characterised by their expression of the transcription factor T, brachyury homolog (mouse) (commonly referred to as *T* or brachyury).

1.1.1 Epidemiology and natural history

Initiating environmental risk factors for the development of chordoma are unknown.

Due to the rarity of the disease, information regarding chordoma incidence and survival patterns has largely been assessed using small case series or considered in combination with other rare bone neoplasms in epidemiological studies. Population-based surveys specifically assessing chordoma as distinct pathological entities are few in number. The largest and the most comprehensive epidemiological study on chordoma emanates from the USA in which data from the Surveillance, Epidemiology, and End Results (SEER) program of the National Cancer Institute was examined to calculate incidence and survival patterns of 400 cases of histologically confirmed chordoma between the period of 1973-1995. This revealed an incidence rate of

0.08/100,000 population [6]. In this cohort of patients it was found that chordoma was more frequent in males than in females (Incidence rate 0.10 versus 0.06) and was distinctly uncommon in African Americans. The incidence rate in whites/Caucasians was four times that in blacks. Historically it was presumed that chordoma was more prevalent in the bones of the pelvis than in the skull base, however evidence from the SEER data suggest an almost equal distribution in the skull base (32%), mobile spine (32.8%) and sacrum (29.2%) [6].

Median survival is approximately 6yrs with 5 year, 10 year and 20 year survival rates dropping to 67.6%, 39.9% and 13.1% respectively [6]. Similar survival rates has been documented from data on chordoma patients in UK tumour registries [3].

Histologically chordomas have been considered low grade neoplasms; however they are locally destructive and progressive. Due to the relatively inaccessible locations of these tumours, particularly base of skull lesions, they are generally not amenable to complete surgical resection, and therefore commonly recur. Metastases occur late in the disease process and can be seen in up to 40% of patients with chordoma [7].

The disease occurs sporadically in most cases, however rarely it can be inherited as an autosomal dominant trait. There are a few reported families with chordoma occurring in two or more close relatives. There have also been case reports in which chordomas have been identified in patients with the tuberous sclerosis complex [8, 9]. In two studies it was reported that a tumour suppressor locus on chromosome 1p36 segregated with the single chordoma family

investigated using a combination of loss of heterozygosity and segregation analysis [10, 11]. This was refuted in two larger studies from the National Cancer Institute (USA) where microsatellite and SNP linkage analysis in three of four families narrowed down the susceptibility locus to chromosome 7q33 [12, 13]. However the putative “chordoma gene” was not identified using conventional capillary sequencing analysis. In view of a new member of one of the families developing chordoma without a 7q33 association and the lack of this association in the fourth family, these same investigators performed a further search using a more refined SNP linkage analysis [14]. This revealed strong linkage with the 6q25-27 locus which harbours the *T* gene. Sequencing did not reveal mutations. However using an array comparative genomic hybridisation approach (aCGH), it was found that four of eight families demonstrated a tandem duplication of *T*. These familial inheritance cases of chordoma are rare, tend to affect younger patients (<40yrs) and occur more commonly in the clival region [12, 15, 16]. This contrasts with sporadic chordomas in which the median age of presentation is ~60yrs and are generally sacral based tumours [6, 17].

1.1.2 Therapeutic options

The location of chordomas along critical bony and neural structures makes the clinical management of these patients difficult. The treatment options are primarily surgical [18, 19] with *en bloc* resection being feasible in up to 50% of sacral chordomas. More aggressive surgery with wide margins can control the advent of local recurrence and has improved the recurrence rates for chordomas of the sacrum [20]. However this is achieved with considerable morbidity (including loss of normal bladder and bowel control due to sacrificing

the sacral nerve [21, 22]), notwithstanding the extensive reconstructive efforts needed with wide excision.

Although subtotal resection (debulking) is often the goal of surgery in the base of skull chordomas, the optimum treatment remains total resection with neurological preservation followed by radiation therapy. As quality of life and preservation of neurological function take priority in this group of patients, base of skull tumours tend to have lower rates of complete excision and recurrences are therefore more common in this location.

Most surgery for chordomas is followed by adjuvant radiotherapy. Conventional stand-alone radiotherapy is ineffective as a palliative approach [23]. Photon-beam therapy can be used as an adjunct to surgery with local control rates between 10-40% reported [23-25]. This control rate is more commonly seen in sacral based chordomas because the sacro-coccygeal region can withstand higher doses of photon radiation (45-80 Gy) than the cervical spine and base of skull [26]. Recent reports of hadron based radiotherapy have been slightly more effective with 5 year local control rates of around 60% [23].

The tumour is totally resistant to conventional forms of chemotherapy apart from a few documented cases of response in the rare de-differentiated subtype [27]. The incidence of morbidity is high as treatment of recurrent tumours often requires re-operation or re-radiation, invariably with higher doses.

Recent case reports and small clinical trials using receptor tyrosine kinase inhibitors, for example inhibitors of both *c-KIT* and *PDGFR*β have been reported [28-30]. These have shown variable success. There are now trials using tyrosine kinase inhibitors, such as Imatinib, in combination with the histone

deacetylase inhibitor LBH589 and Nilotinib combined with high dose radiotherapy which are currently recruiting patients (http://clinicaltrials.gov/ct2/home). EGFR, PI3K, and c-MET inhibitors have also been trialled but early data suggest that they are only partially effective [31-33]. In a study of 70 chordomas, activation of phosphorylated STAT3 protein was identified. *In vitro* experiments on chordoma cell lines have shown that STAT3 inhibitors are effective in controlling cell growth and proliferation suggesting that this may represent a new therapeutic avenue [34, 35]. It is also worth mentioning that to date, the response outcome to some of these "gene targeting" small molecule inhibitors has not been correlated with the specific genetic abnormalities in the tumours. There is therefore an urgent need for identifying new therapeutic interventions that are associated with the underlying genetic aberrations that are present in chordomas.

1.1.3 Pathology

There are three histological variants of chordoma that is, conventional, chondroid and dedifferentiated.

1.1.3.1 Macroscopy

Grossly conventional chordomas have a lobular architecture and are infiltrative destroying the bone and often invading into the surrounding soft tissue (**Figure 1.1**). The tumour cells secrete copious amounts of extracellular matrix which impart a myxoid/jelly-like consistency to the tumour. Haemorrhage and necrosis are common especially in the larger tumours. Chondroid chordoma is less gelatinous and is more firm resembling cartilage. Dedifferentiation in a chordoma implies the transformation of the more conventional component to a

higher grade tumour. This may take on the form of a fibrosarcoma, osteosarcoma or an undifferentiated pleomorphic sarcoma. Dedifferentiation usually occurs following radiotherapy but may occur *de novo* and usually portends a worse prognosis [36]. The dedifferentiated component of a chordoma tends to be fleshy and does not show the gelatinous or chondral appearance [37, 38].

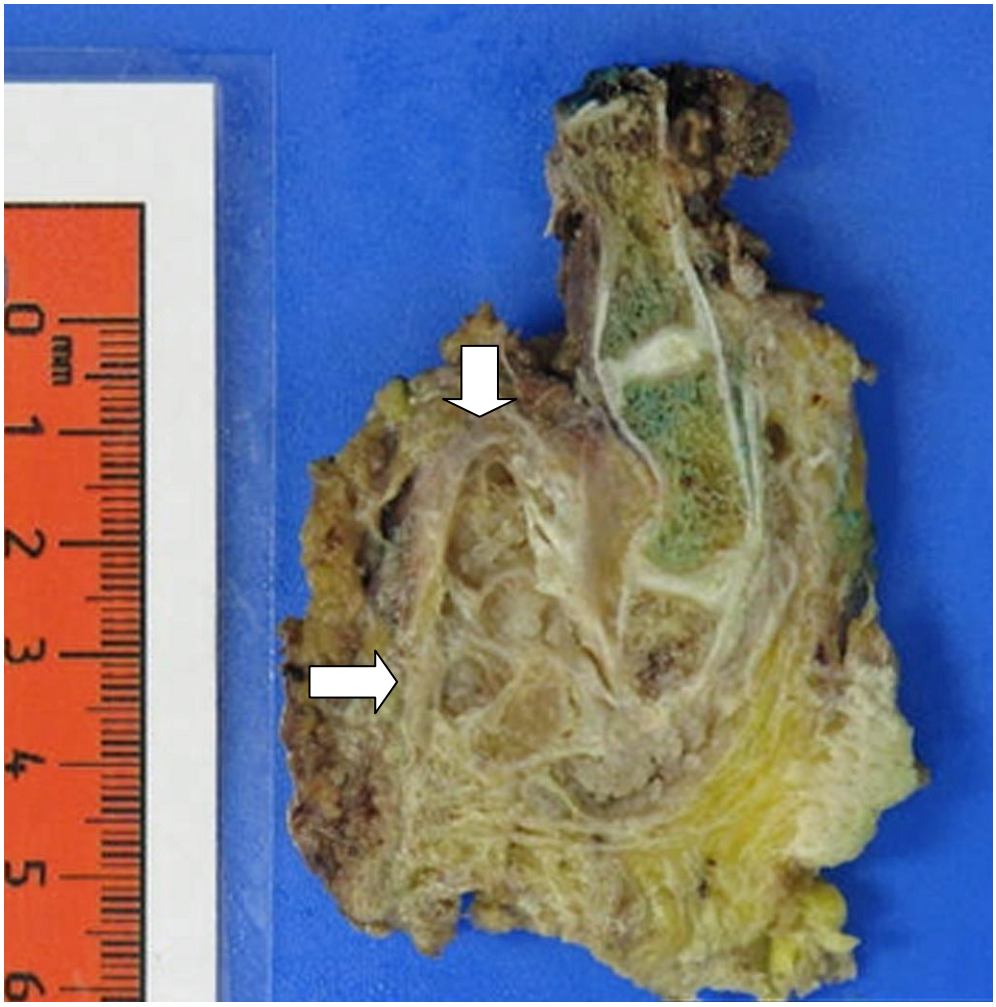


Figure 1.1 Macroscopy: A typical example of a coccygeal chordoma (arrows) demonstrating destruction of the cortical bone with a lobulated expansion into the surrounding soft tissue.

1.1.3.2 Histology

Chordoma cells are epithelioid and have been designated with the appellation – physaliphorous (Greek for “sail like”). Their characteristic feature is their abundant vacuolated cytoplasm. In comparison to other bone and soft tissue sarcomas, conventional chordomas do not demonstrate the usual hallmarks of aggressive malignancies; the tumour is generally slow growing. Chordomas demonstrate variable histology with three main histological types: conventional, chondroid and de-differentiated. As mentioned above the chondroid variant bears a striking similarity to cartilage and diagnostically can be easily confused with a primary cartilaginous tumour such as chondrosarcoma. The conventional chordoma is composed of “bubbly” cells arranged in cords and strands (from which chordoma gets its name) with hyperchromatic nuclei and abundant surrounding myxoid matrix. Areas of tumour demonstrating more than the usual cytological atypia, mitosis and necrosis are sometimes evident in otherwise conventional chordomas and it as yet uncertain if these are of prognostic importance. De-differentiation of typical chordoma-like areas to a more spindled undifferentiated sarcoma-like appearance tends to be more aggressive and has been associated with loss of INI1 and T protein expression [39].

1.1.3.3 Immunohistochemistry

Diagnostic applications

The combined immunohistochemical expression of nuclear T and CK 19 protein is regarded as diagnostic for chordoma [40] and can be used for differentiating it from cartilage-forming tumours. Haemangioblastomas which may also occur in the base of the skull also express T protein but are phenotypically distinct from chordomas [4]. Chordoma cells express S100 protein to a variable degree and CD24, CK 8 and 18 and epithelial membrane antigen. The co-existence of chordomas with renal cell carcinoma in some patients prompted the investigation of the immunohistochemical expression of the renal cancer markers CD10 and RCC antigen in chordomas [41]. These markers were found not to be expressed in a set of 31 chordomas and are useful to distinguish chordomas from metastatic renal cell carcinoma. It is interesting to note however that a subset of renal cancers express the chemo-resistance gene *AKR1B10* [42] which is also expressed at an mRNA and protein level in chordomas [43].

Prognostic utility

Immunohistochemical expression of c-MET has been associated with a more favourable prognosis compared to those without it in younger patients with skull-based chordomas [44]. This association does not appear to be maintained in older patients with spinal chordomas [45]. The expression of fascin protein has been noted in recurrent skull-based chordomas with dural erosion suggesting that it is a marker of poor outcome [46], and on the basis of strong HIF1α [47, 48], and IGF1-R [49] protein expression, these have been posited

as potential therapeutic targets. Chordomas also demonstrate strong protein expression of the PDGFβ receptor but there is less expression of the PDGFα and *c-KIT* receptors although these are phosphorylated [50]. This has provided a rationale for the use of the PDGFR/KIT small molecule inhibitor Imatinib mesylate in patients with these tumours [30]. None of these markers is currently used in routine clinical practice for predicting outcome for patients.

1.1.4 Genetics of chordoma

A wide variety of genetic studies have been performed on chordomas. These include gene expression studies, chromosome analysis, metaphase comparative genomic hybridisation (mCGH), array comparative hybridisation (aCGH), fluorescent *in-situ* hybridisation (FISH), DNA micro-satellite analysis, SNP array analysis, loss of heterozygosity (LOH), and clonality studies [51-58].

1.1.4.1 Gene expression studies

Gene expression microarray studies show that chordomas have a specific gene signature compared to other soft tissue and bone sarcomas. The first genome wide expression studies to investigate this showed that apart from this signature, chordoma shared many similarities in mRNA expression with chondrosarcomas, particularly with the expression of genes known to play a role in cartilage development such as *COL2A1, SOX9* and *ACAN* [56, 58, 59]. This is not entirely surprising as both chondrosarcoma and chordoma bear a striking morphological similarity and there is a relationship between the developing notochord and embryonic cartilage. However these studies demonstrated that there were important differences too, namely in the mRNA expression of the transcription factor *T* and a set of other genes including *KRT*

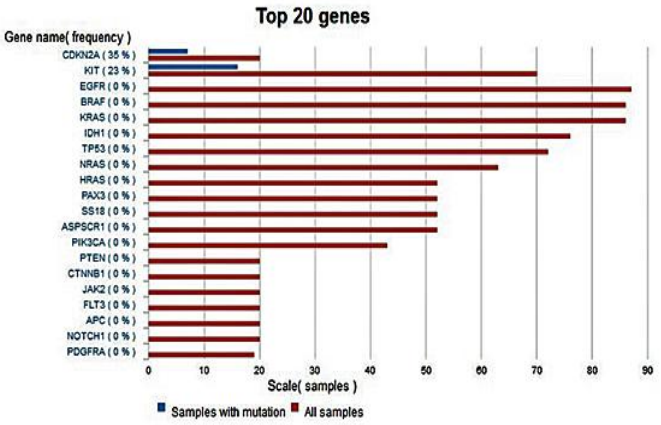
19 and *AKR1B10* which provided the specific gene signature which differentiated chordoma from chondrosarcoma. This was the first molecular link between chordoma and the notochord as both are characterised by the expression of *T* which is now regarded as the diagnostic hallmark of chordomas [58, 59]. More recently it has been shown that *T* mRNA can also be expressed in a subset of colon and lung carcinomas [60, 61], and in some testicular germ cells [40]. It has shown to be expressed in degenerative intervertebral discs [62]. Apart from these tissues and haemangioblastoma [4], a tumour that is more common in the paediatric population, *T* is not expressed in humans outside of the embryonic period. In chordomas, the basis of its expression is largely unexplained. In familial chordomas in which duplication of the *T* gene is a major susceptibility determinant [14], it has not been described that the duplication influences expression levels of *T* mRNA. In sporadic chordomas which are the more frequent subtype the relatively high expression levels of *T* mRNA can be accounted for by amplification of the *T* gene locus in only 7% of cases [17]. Other factors which may influence *T* mRNA expression levels namely overexpression of upstream transcriptional regulators and/or coding somatic mutations have not been identified. Minor allelic gain of *T*, present in ~5% of chordomas [17, 63] is of uncertain significance.

1.1.4.2 Somatic mutations in chordoma

To date, recurrent somatic oncogenic driver mutations have not been identified in chordomas. A study of somatic point mutations in common cancer genes (*APC, BRAF, CTNNB1, FLT3, JAK2, KIT, KRAS, NOTCH1, NRAS, PIK3CA*, *PTEN* and *P53*) using a single base extension genotyping assay in 21 sporadic chordoma samples revealed no recurrent mutations [64]. In view of activation of

MAPK signalling pathways in chordomas, direct sequencing of common mutational hotspots *for EGFR, KRAS, NRAS, HRAS, BRAF, FGFR1, FGFR2, FGFR3,* and *FGFR4* have been performed and have not shown recurrent mutations [17, 63]. Somatic mutations have not been shown in the coding exons and putative promoter of *T* [15, 17]. Data from the Catalogue of Somatic Mutations in Cancer (COSMIC) database [65] show that the most common recurrent mutations in chordoma occur in *CDKN2A* (35%) and *KIT* (23%) (**Figure 1.2**). The recurrent *KIT* mutation identified is the exon 10 M541L point mutation which is known to be non-activating [66].

A



B

Color	Mutation Type	Mutant samples	Percentage
	Substitution nonsense	0	0.00
	Substitution missense	11	44.00
	Substitution synonymous	7	28.00
	Insertion inframe	0	0.00
	Insertion frameshift	0	0.00
	Deletion inframe	0	0.00
	Deletion frameshift	0	0.00
	Complex	0	0.00
	Other	7	28.00
	Total	23	100%

Figure 1.2

Somatic mutations in chordoma (Cosmic database)

A- The most frequent recurrent mutations identified in the COSMIC database are in *CDKN2A* and *KIT* with no mutations identified in a host of other common cancer genes tested.

B- 11/18(61%) *KIT* mutations are protein coding but not activating, whilst the *CDKN2A* mutation events (7/7-100%) are deletion events.

There is a collaborative effort supported by the Chordoma Foundation, and International Cancer Genome Consortium to determine the genome-wide mutational profile of sporadic chordomas using whole exome sequencing (www.chordomafoundation.org/research/chordoma-genome-project/). Although the results are preliminary, there are no recurrent cancer-causing mutations that have been identified in 25 chordomas (personal communication Prof Adrienne Flanagan and Prof Peter Campbell, Chordoma Genome Project).

1.1.4.3 Somatic chromosomal copy number alterations in chordomas

Copy number losses are more frequent than copy number gains in chordoma [53, 54, 64]. The genomic loss of the tumour suppressors *CDKN2A, CDKNA2B* (Chromosome 9p) and *PTEN* (Chromosome 10q) have been shown to occur in up to 70- 80% of chordomas using arrayCGH and qPCR assays [54, 64] **(Table 1.1)**. These are considered late events and not involved in initiation of the tumour. Chromosomal losses of *CDKN2A* and *PTEN* have been corroborated using fluorescence *in-situ* hybridisation and immunohistochemistry analysis on chordoma tissue sections [54, 64]. Other copy number alterations include gain of the 7q33-34 region [53] which has also been identified as a putative susceptibility locus in familial chordomas [13]. A few studies have demonstrated chromosome 7 polysomy which is a common event in primary and recurrent chordomas and correlates with EGFR and c-MET protein expression [67, 68]. According to the Mitelman Database of Chromosome Aberrations in Cancer, 48 cytogenetically abnormal chordomas have been described with the majority exhibiting complex karyotypes [69]. Losses and rearrangements of chromosome 1p have been noted but no chordoma specific translocations have been identified. The recurring aberrations involving the loss of 1p (up to 85% in

skull based chordomas) suggests that an association between the loss of 1p and tumour progression may exist [70]. Interestingly, there appears to be fewer genomic copy number alterations in skull based chordomas, which are seen in a younger patient population, compared to the sacral chordomas that are found in an older patient population. Skull base chordomas with abnormal karyotypes tend to have higher recurrence rates (45% vs 3%) compared to those with relatively normal karyotypes at an average of 48 months follow-up and are associated with shorter patient survival [70].

The 6q27 locus which harbours the *T* gene has been shown by arrayCGH studies to be gained in 29% of sporadic chordomas [54]. Another study performed by our group showed that up to 39% of chordomas were polysomic for chromosome 6, 4.5% showed chromosomal gains of the *T* gene and that up to 7% of chordomas had amplification of the *T* locus using aCGH, FISH and qPCR techniques [17]. The copy number gain of the *T* gene was seen in a similar percentage in sacral, mobile spine and base of skull chordomas.

Table 1.1 Summary of recurrent somatic chromosomal abnormalities in chordoma.

Cytogenetic locus (Chromosome)	Gain/Loss	Frequency	Candidate gene
1p36.2-p11.1	Loss	0.85 [a] 0.57 [b] 0.44 [c]	*RUNX3*
9p24.3-q34.3	Loss	0.55 [a] 0.76 [b]	*CDKN2A/B* [*]
10p15.3-q26.3	Loss	0.80 [a] 0.81 [b]	*PTEN* [**]
17p13.3-p11.1	Loss	0.35 [a] 0.45 [b]	*P53/NF1*
6q25.3-q27	Gain	0.29 [a]	*T*
7q	Gain	0.52 [b] 0.69 [c]	*SHH, AKR1B10, PTN*

The table represents comparative genomic hybridisation data regarding the more common chromosomal abnormalities in sporadic chordoma derived from the literature: [a-] 21 cases [64], [b-] 21 cases [54], [c-] 16 cases [53]. The differences in frequency may reflect the different array platforms used and their resolutions. *- 30% homozygous deletion, **-80% hemizygous deletion.

A recent study using SNP arrays to interrogate somatic copy number changes showed that up to 21% of chordomas, were aneuploid for chromosome 3 [55]. A gain or loss encompassing a ~200-Mb region mapping to 3p26.3-3q26.1 was seen. This finding is similar to a previous study in which FISH analysis of skull based chordomas showed deletion or amplification of 3p in 2 of 6 primary and 2 of 10 recurrent tumours [71]. One of the candidate genes in this region, *FHIT* (fragile histidine triad, a tumour suppressor gene) demonstrated reduced or no protein expression in 98% of sacral chordomas and 67% of skull base

chordomas which suggests that chromosome 3 aneuploidy and/or epigenetic regulation of *FHIT* contribute to loss of the *FHIT* tumour suppressor in chordoma.

1.1.4.4 Chromothripsis and chordoma

The long held Vogelstein paradigm that the development of malignancy is characterised by the relentless, stepwise acquisition of somatic mutations and rearrangements has recently been complemented by a newly described mechanism [72]. Whole genome massive parallel sequencing has unravelled a novel cancer-initiating event, termed chromothripsis [73]. This new concept posits that up to 25% of bone cancers and 2-3% of all cancers develop clustered somatically acquired rearrangements that occur simultaneously rather than by stepwise accumulation. Two of 11 chordomas studied in this seminal paper demonstrated this phenomenon. In one case, there were 147 rearrangements involving localised regions of chromosomes 3q, 4q, 7q, 8p and 9p. Interestingly, the *CDKN2A* locus was deleted in one arm of 9p whilst it was lost, presumably by chromothripsis on the other arm. Other cancer genes that showed rearrangements included *FBXW7, WRN* and *ARID1A.* There were no rearrangements involving the *T* locus. The initiating event for this genomic crisis is unknown but yields important insights into the temporal development of these tumours.

1.1.5 Mouse models of chordoma

Mice have been used for decades to investigate various aspects of mammalian biology. They serve as an invaluable resource to develop preclinical models to identify suitable drugs for clinical trials in humans. Progress in this area in chordoma is hampered by a lack of transgenic mouse models of chordoma. To this end, Dr Brian Harfe's group in Florida has been working on generating a mouse model of chordoma over the last few years [74-77]. In the course of fate mapping experiments of notochordal cells, Dr Harfes team discovered that apart from forming the nucleus pulposis, small collections of notochordal cells were "left behind" in the vertebral column similar to the findings of Yamaguchi *et al* in human cadavers [78]. These fate-mapping experiments were the first documentation of these notochordal remnants in mice. Interestingly these notochordal remnants do not express SHH and its cognate receptor PTCH1 protein which are both co-expressed in the developing notochord and chordomas [77]. Reasoning that activating this pathway, together with engineering other chordoma related genetic changes such as deletion of *Tsc1* and *Pten* in these remnants may induce chordoma formation in these mouse models, this group has been attempting to generate murine chordomas which has thus far been unsuccessful (http://www.chordomafoundation.org/research/model-development/).

1.2 T and T box genes

1.2.1 Background

T, encoding the protein brachyury (Greek~ brakhus = short, oura =tail), is located on chromosome 6q27. It is the founding member of the T-box family of transcription factors, all of which share a conserved DNA binding domain. The expression of (MOUSE) T is restricted to the notochord and paraxial mesoderm during development of mice [79, 80]. During mouse gastrulation the three primary germ layers (ectoderm, mesoderm and endoderm) become arranged such that the inner layer of cells (epiblast) in the embryo ingresses and moves between the remaining epiblast layer (future ectoderm) and the outer primitive endoderm layer. This gradual movement of the presumptive mesoderm and endoderm cells, begins in the posterior aspect of the embryo and creates a morphologically visible structure known as the primitive streak. This is a fundamental process in the establishment of the anterio-posterior axis in all mammals and the T protein is essential for this process. As the embryo develops, the mesodermal cells exit the streak, and take on a lateral and anterior migratory path to occupy the lateral and dorsal positions along the anterio-posterior axis, and subsequently form the axial and paraxial mesoderm structures such as the notochord and somites. T is critical to this process and is responsible for notochord differentiation, in the establishment of the anterior-posterior axis and in mesoderm specification [81]. (MOUSE) $T^{-/-}$ homozygote mice are embryonic lethal with profound defects in the posterior half of the trunk related to aberrations in the notochord and primitive streak [82]. Mice with heritable short tails were first identified in X-ray mutagenesis screens in 1927 and these were subsequently ascribed to (MOUSE) $T^{+/-}$ heterozygous mutations

[83]. There is a dosage phenomenon in axial development identified by the graded degree of severity of defects dependant on the mutant (MOUSE) *T* alleles carried by mice. These range from shortening of the tail with sacral malformations in heterozygotes to a disruption of trunk development and embryonic death in homozygotes [84]. The disruption of the primitive streak, premature cessation of mesoderm induction and absence of the notochord subsequently leads to abnormalities of the neural tube and somites. Null mutants of the zebrafish T ortholog (Ntla) are named “no tail”, because they do not develop the posterior somites and notochord [85]. These observations provide strong evidence that *T* is a critical gene that is necessary for normal development. In embryonic development of Xenopus it has been shown that a promoter element 500bp 5’ of (MOUSE) *T* is necessary and sufficient for induction of (MOUSE) *T* mRNA expression in the primitive streak and tailbud and that this can be activated by FGF signalling but this is separable from the element controlling (MOUSE) *T* expression in the notochord [86, 87]. This suggests that different or additional signals are required for the activation of (MOUSE) *T* in the notochord than those involved in (MOUSE) *T* activation in the primitive streak in mice. In zebrafish stable expression of the T ortholog Ntla in the notochord however does require continued FGF signalling. This results in an autoregulatory feedback loop with FGF, where Ntla regulates its own expression either directly or indirectly [86, 88, 89]. In humans (outside of embryonic development), *T* is “switched on” again in the context of cancer rather than in normal differentiated tissue apart from isolated germ cells in the testis and some instances of degenerative interverterbral discs [4, 62]. Given its important role in development and emerging links with cancer, *T* therefore joins

the company of other essential embryonic genes that have been ignominiously associated with the development of malignancy [90, 91].

1.2.2 Structure of *T*

The (MOUSE) *T* gene was first cloned by Herrmann [92] where it is located near the centromeric region of chromosome 17 contrasting with its telomeric location on chromosome 6 in humans. The human *T* gene is ~10kb in length with the longest expressed transcript consisting of nine exons (Transcript ID: ENST00000296946), the first exon of which is non-coding (**Figure 1.1**). This transcript encodes a protein that consists of 435 amino acids and is localised to the nucleus. Cloning of T homologues in mouse, frog (Xbra), zebrafish (Ntla) and chick (ch-T) has demonstrated that conservation is maintained in the N-terminal portion of the gene whilst the C-terminus is variable across these vertebrate species [81]. The human homolog of (MOUSE) *T* was first cloned by Edwards et al. at UCL from foetal intervertebral discs [93]. This demonstrated a high level of protein sequence conservation with other vertebrate *T* homologs particularly in the DNA binding domain (T-box) where there is a 100% degree of conservation between mouse and human *T* between amino acids 1 and 223 but less conservation between *T* and other T box genes.

1.2.3 T-box domain

The highly conserved T-box binding domain is located in the N-terminal portion of the gene with the C-terminus responsible for transcriptional modulation (trans-activation and repressor domains) [93]. (MOUSE) T and Xbra (Xenopus)

recognise and bind two palindromic 12bp elements one of which is 5'-AGGTGTGAAATT-3' while the other can substantially differ to it [94, 95]. The two palindromic DNA binding motifs represent an unusually long binding region, and it has been shown that the (MOUSE) T protein can bind it either as a dimer or monomer. This ability of T-box genes to bind as homodimers is similar to other transcription factors like helix-loop-helix proteins and zinc finger proteins. *In vitro* experiments show that two half sites are required for binding and their spacing is required for transactivation which is potentiated when palindromic T binding sites are present [96]. The binding of the T-box element to DNA occurs with contacts of the major groove which have been resolved by X-ray crystallography (**Figure 1.3**) [95]. There is evidence that a common human polymorphism Gly-177-Asp affects T dimer stability *in vitro* [94]. Using electrophoretic mobility shift assays, researchers have demonstrated that the polymorphism in this highly conserved region of DNA in humans reduces the stability of T dimer formation decreasing its affinity to the target DNA and possibly allowing for heterodimerisation with other transcription factors. This is important as heterodimerisation with other transcription factor proteins offers scope for diversity of gene regulation. It would seem that the ability of T-box proteins to form heterodimers has precedence as it has been proven in *Xenopus* sp. that the functional specificity of Xbra binding is enhanced by interaction with other transcription factors, for example, Smad1, Gata and Lim proteins [97].

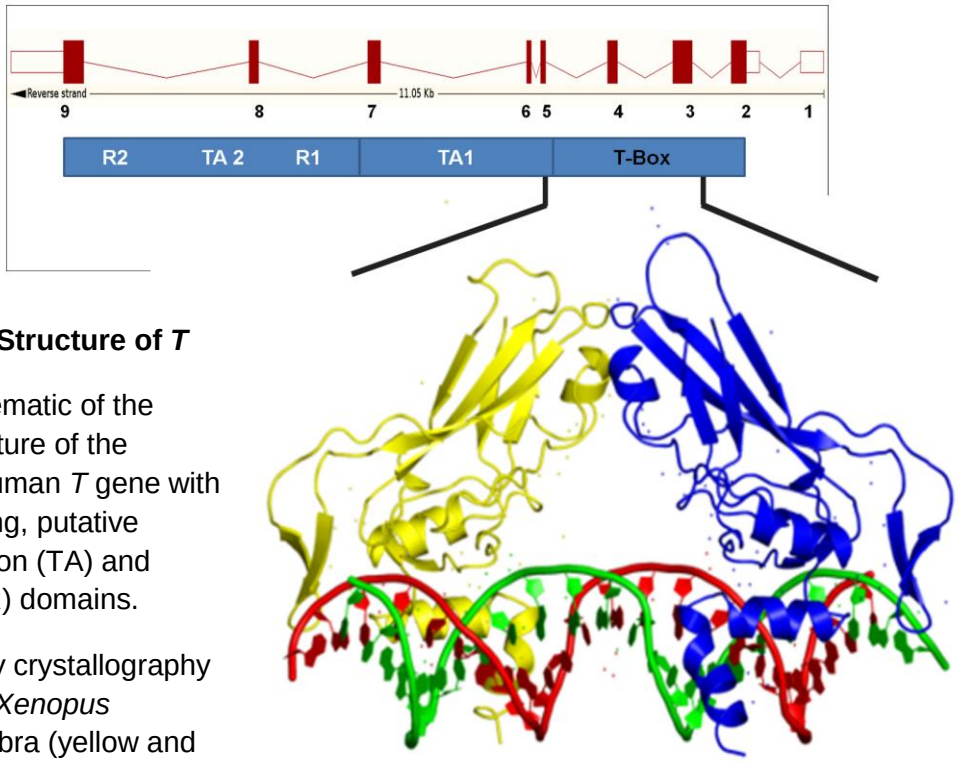


Figure 1.3: Structure of *T*

Upper: Schematic of the exonic structure of the reference human *T* gene with T-Box binding, putative transactivation (TA) and repressor (R) domains.

Lower: X-ray crystallography structure of *Xenopus* brachyury-Xbra (yellow and blue ribbons denote T-box domain) bound to DNA duplex as a dimer **(adapted from reference (90))**.

1.2.4 Genes bound by the T protein during development.

In view of *T*'s essential role during embryonic development, multiple efforts have been made to identify its transcriptional targets [98-100]. More commonly this has been performed in developing mesoderm in *Xenopus sp*, Zebrafish and murine models compared to the notochord. Specific targets have however been identified in the notochord albeit in the invertebrate chordate, *Ciona intestinalis* [101]. In zebrafish, Ntla directs a transcriptional program in the mesoderm that involves the activation of muscle fate determinants (MyoD1, Msgn1, FoxD3),

gastrulation movements (Wnt11, SnailA, Tbx16) and posterior fate-components of the Fgf and Wnt signalling pathway [99]. In particular, it regulates the balance between Wnt and Fgf signalling during posterior body development. These findings are similar to that found in *Xenopus sp*. implying a degree of functional conservation.

1.2.5 T-Box genes and disease

Many T-box genes have multiple overlapping roles during development and organogenesis. For instance, murine TBX1 and TBX5 are active in mesoderm patterning outside of gastrulation and their expression specifies transcriptional programs necessary for normal limb and cardiac development [81]. In humans, heritable germline mutations in *TBX5* manifest in syndromes related to its roles during development as in the Holt-Oram syndrome in which atrial septal defects of the heart and absent radial bones in arms of affected individuals are identified [102]. Furthermore, deregulation of T-box gene expression is increasingly being implicated in a variety of malignancies (**Table 1.2**). *TBX2* mRNA overexpression for instance is responsible for maintaining proliferation in melanomas [103] and downregulation of *TBX5* expression is associated with colon cancer [104].

Table 1.2 T-box genes and disease

Gene	Germline mutations that manifest in human Syndromes	Mutations** or alterations in gene expression+ implicated in human cancers and other disorders
T	**Not known**	**Chordoma**[**,+,14, 17, 105] **, Lung carcinoma**[+,61] **, Colon carcinoma**[+,106]
TBX1	**DiGeorge**[107]	**Non-syndromic Tetralogy of Fallot**[*,107]
TBX2	**Not known**	**Melanoma**[+,103] **, Breast carcinoma**[+,108]
TBX3	**Ulnar mammary**[109]	**Melanoma**[+,110] **, Breast carcinoma**[+,111]
TBX5	**Holt-Oram**[102]	**Colon carcinoma**[**,+,104]
TBX21	**Not known**	**Graves disease**[**,112] **, Aplastic anaemia**[**,113]
TBX22	**X-linked cleft palate and ankyloglossia**[114]	**Colon carcinoma**[*,115]

1.3 The relationship between *T* and chordomas

1.3.1 The origin of chordoma

Chordomas are tumours that show differentiation towards the notochord and there is evidence that they may arise from notochordal remnants [78]. The notochord is a midline embryonic structure that is essential for the normal development of chordates. It is a transient structure that serves two main

functions in the embryos of higher vertebrates. Firstly it provides structural support to developing embryos. Indeed, it has been referred to as the primitive axial skeleton [116]. In zebrafish, without a fully differentiated notochord, embryos fail to elongate resulting in an inability to swim making them susceptible to predation [117]. It represents a primitive form of cartilage that eventually ossifies to form the vertebral bodies and contributes to the nucleus pulposis of the intervertebral discs [118]. Its second role is in patterning surrounding tissues. It is responsible for induction of the neural tube and floor plate and patterning of the somites [119, 120]. This is chiefly achieved by secretion of Hedgehog proteins. It is also involved in the establishment of Left-Right asymmetry and normal specification of the cardiac field [121]. The notochord expresses various orthologues of the transcription factors, T and FOXA in different model organisms which are important in its developmental, function and differentiation [122] with mutations in (MOUSE) *T* resulting in disruption of the notochord and embryonic death [84].

Vertebral and skull-based chordomas have been postulated to arise from the notochordal cells because of their shared morphology, immunophenotype (T, S100 and cytokeratin profile) and anatomical location (**Figure 1.4**) [123, 124]. Virchow coined the term **ec*chondroses*** physaliphora to describe "cartilaginous" outgrowths found in the base of the skull. The proposition that these were in fact notochordal remnants was proposed by Müller and the term **ec*chordoses*** physaliphora was then adopted [125]. The concept of the existence of notochordal remnants has been supported by anatomical and 3D reconstruction studies showing that there is inconsistent regression of the human notochord particularly in the skull base and sacral region of foetuses

where the notochord tends to have a more convoluted course [126]. More recently, it has been shown by cell tracking that isolated notochordal cells survive as remnants in murine vertebral columns [127]. However, this link between the notochord and chordoma was firmly clinched at a molecular level when gene expression microarray studies demonstrated the consistent, specific mRNA expression of *T* in chordomas [59]. T protein expression is therefore regarded as the diagnostic hallmark of chordomas [4, 128].

There is now evidence th*at T* is implicated in the pathogenesis of chordoma. This is on the basis that the protein is expressed in chordoma, it is amplified in ~7% of sporadic chordomas, and silencing of the gene *in vitro* induces growth arrest of chordoma cells. Furthermore, germline duplication of *T* is associated with an increased susceptibility to developing chordoma in the familial form of the disease.

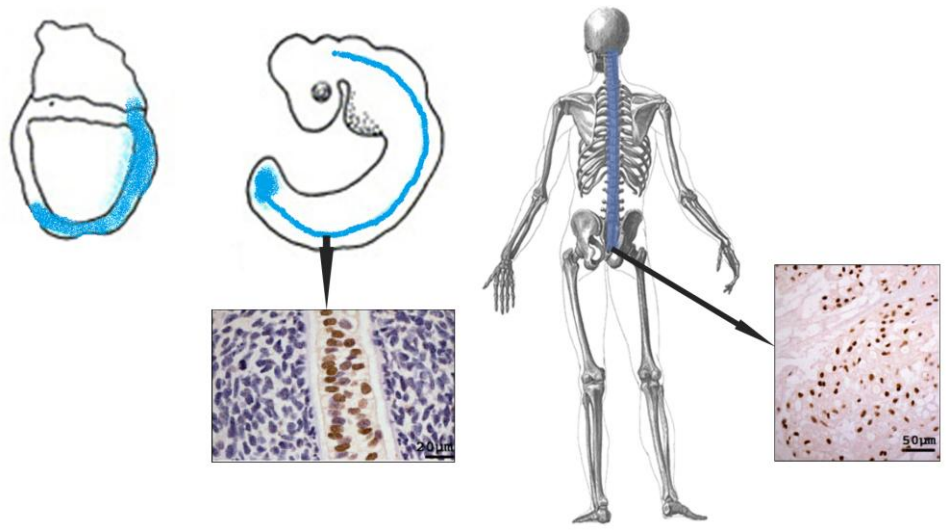


Figure 1.4: T expression and chordomas

In the schematic diagram of an early developing mouse embryo (left), (MOUSE) T expression is present in the notochordal plate and primitive streak. Later during development (centre), this expression is restricted to the posterior tail-bud and notochord. In the post embryonic stage in humans, T is usually only expressed in chordomas which may occur anywhere along the vertebral column or base of skull

1.3.2 Benign notochordal cell tumours or Giant notochordal rests?

The origins of chordoma are not without controversy. There is a division of opinion with regards to terminology of the precursor lesion which develop from vertebral notochordal vestiges or remnants [124]. Yamaguchi et al. coined the term benign notochordal cell tumour (BNCT) to differentiate what they deem to be a neoplastic collection of cells with a similar immunophenotype as the notochord in the vertebrae but differ from embryonic vestiges of notochordal cells found in intervertebral discs [78, 123]. BNCTs share a similar immunophenotype to chordomas co-expressing S100, CK19 and T protein. However, they differ morphologically in that BNCTs comprise compact nests of adipocyte-like cells with little intervening extracellular matrix. They may elicit a sclerotic reaction in the surrounding bone and can grow to such large sizes that they replace much of the vertebral body. They are frequent (up to 20% of the population) and are increasingly being linked to the development of chordoma on the basis of their immunoexpression and co-existence in vertebrae resected for chordoma (**Figure1.5 a- h**).

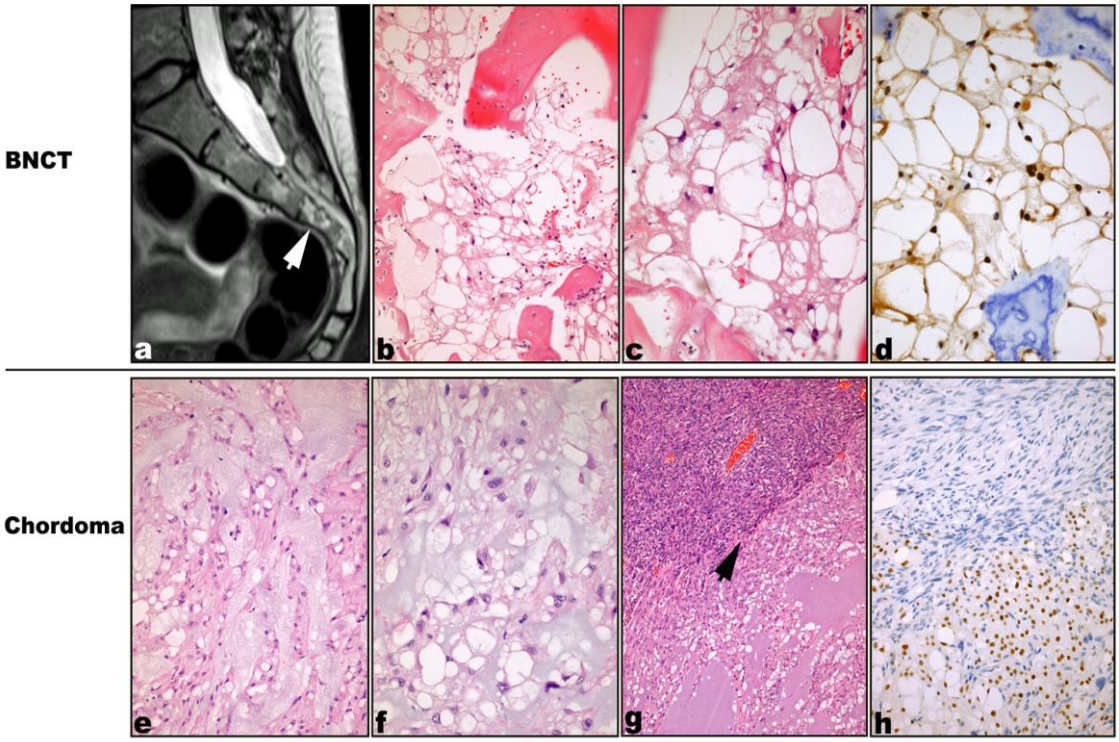


Figure 1.5: Radiology and histology of benign notochordal tumours and chordomas

a) T2-weighted MRI image with hyperintense large coccygeal BNCT (white arrow).

b-d) Photomicrographs depict intraosseus location of BNCT (b, 4x magnification). Uni and multivacuolar notochordal cells of BNCT lacking nuclear atypia and demonstrating nuclear immunoreactivity for T (c-d, 20x magnification).

e-f) Typical chordoma composed of cords of infiltrative physaliphorous cells embedded in abundant eosinophilic extracellular matrix (e, 10x magnification). Higher magnification showing nuclear atypia seen in chordomas and not in BNCT (f, 20x magnification).

g-h) Classical chordoma transitioning to dedifferentiation (arrow) (g, 4x magnification). T is expressed in the classical component but not in dedifferentiated foci (h, 10x magnification).

Kyriakos and co-authors question the terminology employed by Yamaguchi *et al*. and prefer to call these lesions, giant notochordal rests (GNRs). The GNRs are larger than the BNCTs originally identified in the Yamaguchi *et al.* autopsy study and have a different vertebral distribution, being more common in the mobile spine than the sacro-coccygeal region. In any event, the histomorphology of ecchordoses physaliphora, BNCT and GNR are identical and all have been found in association with chordoma re-affirming earlier observations. However, to date there have been no cases reported of chordomas arising in the intervertebral disk.

1.4 *T* and other malignancies

The T protein is involved in mesoderm specification and differentiation during early development. During gastrulation it plays a pivotal role in regulating the EMT switch. There have been recent reports of *T* expression (mRNA) in a number of colon and lung cancer cell lines as well a variety of human-derived cancers including oesophageal, stomach, small intestine, kidney, bladder, uterine, ovary and testis [60, 61]. This is intriguing, as a number of studies have unequivocally demonstrated no protein expression of T in a multitude of malignancies [4, 59, 128]. This discrepancy is unresolved but may be explained by the different antibodies utilised in these studies. It has also been demonstrated *in vitro* that plasmid mediated overexpression of *T* mRNA in a pancreatic cancer cell line results in induction of the EMT with a subsequent increase in proliferation, invasion and metastases of this cancer in mouse xenografts. The complementary experiment in which the *T* mRNA expression levels are reduced results in the opposite effect [129]. In view of *T*'s putative role in cancer cells of epithelial origin and its function in regulating the EMT, the

prediction of response to EGFR inhibition was investigated in T protein expressing human lung carcinomas. It was shown that up to 41% of non-small cell lung carcinomas express T protein and the expression level corresponded to resistance of the tumour cells to EGFR kinase inhibition [130]. Further evidence for *T*'s role in carcinoma has come from investigations of the *T* mRNA expressing human colon cancer cell line, SW480. In this scenario, *T* mRNA expression was shown to be a confluency-dependent phenomenon with lower confluence associated with higher expression and *vice versa* [106]. In the context of colon carcinoma, it appears that a subset of cells, those with a stem cell phenotype which is typically located at the invasive front of the tumour express T protein. These cells also demonstrate active WNT signalling [131]. The localisation of T protein in these tumours appears to be cytoplasmic rather than nuclear, the significance of which is yet to be determined [132]. In some types of non-small cell lung carcinoma, T protein is also expressed in the cytoplasm [61]. Given this accumulating evidence (see above) it appears that *T* may be expressed in other malignancies, although the expression may be restricted to a specialised cellular niche in carcinomas undergoing mesenchymal transition. It is not expressed in a wide range of sarcomas apart from chordoma [59, 128].

1.5 *T*, chordoma and stem cells

Stem cells are progenitor cells that have the ability to self-renew and are totipotent [133]. There is currently a great interest in them because of the potential therapeutic benefits of being able to generate a limitless supply of tissue to replace and restore pathological tissue, for example, in degenerative diseases [134].

1.5.1 Cancer stem cells

There is a hypothesis that within many cancers a small population of cells exist that have the capacity for limitless self-renewal, have the ability to differentiate, and are resistant to chemotherapy. These are tumour-initiating cells often referred to as cancer stem cells [135]. The ability to characterise these cells is challenging as there are as yet no consensus markers that unequivocally identifies this population: furthermore the cancer stem cell phenotype may vary in different tumour types. The bulk of evidence for their existence is primarily based on xenograft models in severely immunodeficient mice [135], where it has been shown that only a small fraction of the main tumour bulk (0.0001-0.1% of cells) is capable of generating a tumour *in vivo,* and that these cells share characteristics with normal stem cells of the same tissue. Also in a study of leukaemia, in which the major discoveries in stem cell biology have been made, it has been shown that whilst chemotherapy reduces much of the tumour burden, there is a rare population of cells that are resistant which accounts for recurrent disease [136]. The proposed mechanism for this is the ability of the cancer stem cell to activate efflux mechanisms to eject the chemotherapeutic agents that have penetrated the cell membrane [137].

1.5.2 *T* and stem cells

There exist in developing embryos, stem cell niches (an environment that sustains the stem cell population) responsible for the differentiation of various tissues. During mouse vertebrate development the tailbud has a population of stem cells that is responsible for somitogenesis and the differentiation of various tissues in this region [88]. T functions in an autoregulatory feedback loop with

the canonical WNT signalling pathway that is responsible for maintaining this stem cell niche by activating β-catenin. Also the mesodermal differentiation of embryonic stem cells is chiefly directed by T *in vitro* and spermatogonial stem cell renewal is mediated by T in mice [138]. Differentiation assays of mouse and human embryonic stem cells show high levels of T expression [100]. T is therefore used as a marker together with OCT4, NANOG and KLF4 to characterise this population of cells.

1.5.3 *T* and cancer stem cells

The role of the WNT-T feedback loop [139] in stem cell maintenance has been explored in colon cancers [106, 131]. *APC* mutations result in constitutive WNT signalling and β-catenin activation in the majority of colon cancers [140]. There is heterogeneity of WNT signalling in these tumours with the invasive front showing higher levels of β-catenin expression [131]. The evidence suggests that this is responsible for maintaining a stem cell population that has the ability to undergo the epithelial to mesenchymal transition and respond to invasive and metastatic cues [61, 106, 129]. This population of cells has been shown to express T in response to β-catenin activation.

The human chordoma cell line, U-CH1 expresses T protein, and has a population of cells that express known cancer stem cell markers [141]. It also has the ability to show osteogenic differentiation and form xenografts in mice [17, 141]. These findings coupled with their inherent resistance to chemotherapy provide evidence that the U-CH1 cell line possesses cells with stem cell characteristics. The therapeutic applicability of this finding would be the ability to differentiate these cells into a more “mature” lineage *in vivo*.

1.6 Overall scope of my thesis

Chordoma, like some other cancers, can occur in families. The risk mechanism in members of these families is the inheritance of a second copy of *T*. In sporadic chordomas, a minority show amplification of the *T* locus and abrogation of T expression in chordoma cells is associated with decreased proliferation.

To this end it has been the long term research interest in our group to determine how this transcription factor mediates its function in these tumours.

The first aim of this project was to unravel the downstream targets of T in chordoma with a view to understanding the transcriptional network it controls. A secondary benefit from the identification of these targets would be the detection of genes that could potentially be targeted therapeutically.

The second aim was to determine if patients with sporadic chordomas harboured any genetic susceptibility determinants in *T* or in other genes.

The third aim was to investigate the role of P63, an upstream regulator of (MOUSE) *T* in murine embryos and murine cancer models.

Chapter 2

MATERIALS AND METHODS

"First, have a definite, clear practical ideal; a goal, an objective. Second, have the necessary means to achieve your ends; wisdom, money, materials, and methods. Third, adjust all your means to that end" Aristotle

2.1 Clinical samples

a) Chordoma with matching germline and control samples were retrieved from Royal National Orthopaedic Hospital Musculoskeletal Biobank, UK (approved by the Cambridgeshire Research Ethics committee, Cambs., UK; Reference Number: 09/H0304/78).

b) Chordoma and matching germline samples for replication analysis were obtained from Royal Orthopaedic Hospital Biobank, Birmingham, UK, and other external referral centres with appropriate ethics approval.

c) Clinical information regarding the patients were obtained by direct interview and/or retrieved from histological reports and clinical notes. All cases which were diagnosed as chordoma were confirmed to be immunoreactive for T.

d) Control DNA samples for the SNP association study were obtained from the UCL Department of Cardiovascular Genetics. These samples were obtained with appropriate consent from healthy Caucasians aged 50 to 61 years, registered with 9 UK general medical practices [142].

e) Control samples for exome sequencing were participants in studies of early-onset dementias none of which have any associations with chordoma. Patients were recruited by the NHS National Prion Clinic and the Dementia Research Centre or others at University College London Hospitals NHS Trust and approved under the auspices of the National Hospital for Neurology & Neurosurgery Research Ethics Committee.

2.2 Cell culture

2.2.1 Chordoma cell lines:

a) **U-CH1**: (Kindly provided by Dr David Alcorta, Duke University, Durham, NC, USA through the Chordoma Foundation).
b) **U-CH2**: (Kindly provided by Dr Silke Bruderlein, University of Ulm, Germany).
c) **Mug-Chor1**: (Kindly provided by Dr Beate Rinner, Austria).

2.2.2 Media and supplements:

a) IMDM (GIBCO®, Invitrogen, Paisley, UK).
b) RPMI-1640 (Invitrogen).
c) Foetal calf serum (Invitrogen).
d) Antibiotics: 100U/ml penicillin G, and 100µg/ml streptomycin (Invitrogen) and Puromycin 1mg/ml (Sigma-Aldrich, Ayrshire, UK).

Cells were cultured as a monolayer in a 4:1 ratio of IMDM: RPMI media supplemented with 10% FCS and 5ml Pen-Strep. These were maintained at low passage in 75cm^2 tissue culture flasks coated with 0.1%gelatin (Sigma-Aldrich) flasks in 37^{o}C, humidified incubators maintained at 5% CO_2.

2.2.3 Other cell lines:

HEK293T (obtained from UCL Scientific Support Services), **HeLa** cells (obtained from American Type Culture Collection, Middlesex, UK) and GFP expressing **HT1080** cells (kindly provided by Dr Leonid Nikitenko-UCL Cancer Institute) were used for various experiments detailed below. These cells were

cultured in 75cm^2 tissue culture flasks in RPMI media supplemented with 10% FCS.

2.3 Stable shRNA knockdown of T in U-CH1 and U-CH2

2.3.1 Virus production

Approximately 2million HEK293T cells per 10cm culture dish were seeded 24hours before transfection. Transfection was performed using 1μg p8.91 (gag-pol expressor), 1μg pMDG.2 (VSV-G expressor) and 1.5μg pGIPz DNA with Fugene transfection reagent (Roche Diagnostics Ltd., Burgess Hill, West Sussex, UK, using an in-house protocol. The supernatant containing virus was harvested and filtered on days 3, 4, 5 post transfection and titrated using a biological flow cytometry assay.

2.3.2 Stable lentiviral transduction

The establishment of a stable T knockdown cell line was performed using lentiviral mediated shRNA transduction. This model for knockdown of T in chordoma cells was established and validated in our laboratory using two different hairpin constructs (V2LHS 153729, and V2LHS 153725) (pGIPZTM, Thermoscientific Open Biosystems, Chesterville, Al, USA) which induce >95% permanent reduction of *T* mRNA and protein [17].

The U-CH1 and U-CH2 cells were harvested the day before transduction and plated at a density of 1x10^6 cells/10cm tissue culture dish (Corning Life Sciences, Corning, NY, USA). The cells were infected in triplicate with the V2LHS 153729 shRNA construct, which targets the 3'UTR of human *T* mRNA, at a multiplicity of infection of 1 in the presence of polybrene (10μg/ml) (Sigma-

Aldrich, St Louis, MO. USA). GFP fluorescence was monitored daily by microscopy to determine efficiency of transduction. Optimal fluorescence was obtained 96hours post transduction. These cells were then placed on antibiotic selection using puromycin (4μg/ml) for 48 hours. Results were compiled from three biologically independent transduction experiments.

2.4 Cell proliferation and cell cycle assay

Cell proliferation was assessed using the Click-IT®Edu Alexa Fluor 647 system (Invitrogen). Briefly, cells were grown as a monolayer, and labelled with 10μM (final concentration) Edu overnight. Cells were permeabilised with the saponin-based solution, and Edu detection was performed using Alexa fluor 647 azide dye chemistry. Cell cycle labelling was performed using propidium iodide (Sigma-Aldrich). Non Edu- and non-propidium iodide-containing cells were labelled with the Alexa Fluor 647 dye and served as the background control. The percentage of actively cycling cells and the cell cycle distribution phase was determined by flow cytometry using the CyAN-ADP™ flow instrument (Beckman Coulter Inc., UK) using appropriate parameters and controls.

2.5 Migration assay

In order to investigate the migratory ability of T knockdown cells towards a chemotactic stimulus, a modified Boyden chamber assay was performed. The V2LHS 153729 shRNA plasmid has a GFP expressing cassette. To take advantage of this, a transwell system with a fluorescence blocking coating was used. This coating is specifically designed to absorb fluorescence in the 490-700nm range and can therefore be used to detect GFP fluorescent cells that have migrated to the underside of the membrane without detecting the non-

migrating cells above. Control and T knockdown cells were serum starved overnight and were then seeded onto BD Falcon Fluoroblock 24-Multiwell 8μm pore inserts (Becton Dickinson,Franklin Lakes, NJ, USA) at a density of 5000 cells/well (300ul in each insert) and the insert was placed in BD falcon 24 well multiwell insert plates. The lower chamber was loaded with 1000μl of serum containing media as a chemotactic stimulis. The number of cells which had migrated across the membrane was determined by GFP expression measured on a microplate reader at 24 and 48 hour time-points and confirmed using fluorescent microscopy. From previous optimisation experiments, 48hours is the optimal time point for U-CH1 cells which are slow migrators compared to stable GFP expressing HT1080 cells which were used as the positive control.

2.6 Adhesion assay

Adhesion assays were performed for T KD U-CH1 cells as previously described [143]. Briefly, cells from three biologically independent knockdown experiments were seeded into wells of a 96 well tissue culture plate (TPP, Trasadingen Switzerland). The wells were coated in triplicate with 100ul of human placental laminin(Sigma Aldrich), human plasma fibronectin(BD), rat tail collagen 1 (First Link, Wolverhampton,UK) and growth factor reduced extracellular matrix gel (Sigma) at 20μg/ml and 0.1% bovine gelatin (Sigma-Aldrich) for 2 hours at room temperature. The substrates were then removed and the wells were blocked with 1% bovine serum albumin (BSA) + phosphate buffered saline (PBS) for 1 hour at room temperature. The BSA was removed and wells washed with PBS. Serum starved control and T KD cells were then seeded in the appropriate wells at 5000 cells/well in 100ul serum free media and incubated in a tissue culture

incubator at 37°C, 5%CO2 for 1 hour. The wells were then gently washed with PBS to remove non and loosely adherent cells. Quantification of adherent cells was then performed on a microplate reader measuring GFP expression.

2.7 Molecular biology: RNA

2.7.1 RNA extraction

Total RNA was extracted from cell lines and frozen tissue samples according to the manufacturer's instructions using the Qiagen mirnEasy kit (GmBH, Hilden, Germany) and quantified using Nanodrop® spectrophotometry (Thermoscientifc, Wilmington, DE, USA).

2.7.2 cDNA synthesis

a) Downstream gene target and P63 study: Reverse transcription to cDNA was performed according to the manufacturer's instructions using the Applied Biosystems High Capacity RNA to cDNA reverse transcription mastermix kit (Applied Biosystems, CA, USA) using 100ng of total RNA.

b) SNP association study: The RNA from 13 samples for allelic and differential *T* mRNA expression analysis were converted to single strand complementary DNA (GoScript reverse transcription system, Promega, Madison, WI, USA) utilizing 100ng of total RNA.

2.7.3 qRT-PCR

a) Downstream gene target study: qRT-PCR was performed using Applied Biosystems Sybr® green chemistry with primers custom designed for *T* (Eurofins MWG-Operon, Germany) *GAPDH, TGFA, AKR1B10, FGF1, EGF, HOPX, ETV1* and *IL-8* (Invitrogen, Paisley, Scotland, UK) (**Table 2.1**). Detection was performed using the Mastercycler® Realplex, Eppendorf system (Cambridge, Cambs., UK). The amplification step was performed in a Microamp optical 96-well plate Applied Biosystems, Birchwood, Warrington, UK). Each well contained a total reaction volume of 25µl diluted in RNAse free water: 1µl of cDNA, 0.75µl of forward and 0.75µl of reverse primer at 10pmol, 12.5µl of SYBR® green PCR Master Mix. Each reaction was performed in duplicate and the PCR experiment was performed a minimum of two times with most experiments performed in triplicate. The amplification was performed with a first step at 95°C for 10min and then 40 cycles with 95°C for 15s, 60°C for 1min with a fluorescent reading at the end of this step (with a 520nM filter, SYBR dye).
The forward and reverse primer concentration was optimised for amplification efficiency and primer dimer formation monitored by melting curve analysis.
The mRNA expression level of *T* served as validation of efficacy of the knockdown of this gene in the U-CH1 cell line. The relative gene expression level was determined using the $2^{-\Delta\Delta}$CT comparative method normalized to the housekeeping gene *GAPDH* [144].

For confirmation of *T* target expression in human chordoma samples, the $2^{-\Delta}$CT method was employed, again normalized to *GAPDH* expression [144]. These values were $\log_2$ transformed and used to generate the heatmap. Each datum point of the heatmap represents the mean expression derived from a minimum of 2 qRT-PCR runs performed in duplicate. Both the heatmap and unsupervised hierarchical clustering were generated with GENE-E using Euclidian distance as a parameter [145].

b) SNP association study

- ***T* mRNA expression**: qRT-PCR was performed using a specific *T* fluorescent primer-probe assay [17].

- **Allele specific expression analysis.**

Allelic specific gene expression analysis was performed as previously described [146]. Using a SNP genotyping Taqman assay, a linear regression curve for the $\log_2$ fluorescent intensity vs $\log_2$ allele ratio from the serial dilution of AA and GG genotyped genomic DNA from control samples was performed in the following ratios (AA:GG 8:1, 4:1, 2:1, 1:1, 1:2, 1:4, 1:8). Allele specific expression of cDNA for disomic samples heterozygote for SNP rs2305089 were then obtained by extrapolating these values on the regression line.

2.7.4 RT-PCR

- *P63* study

Primers to detect the various isoforms were designed based on a previous publication [147] (**Table 2.1**). PCR was carried out with AmpliTaq polymerase (Applied Biosystems) on an Eppendorf Mastercycler using a touchdown PCR protocol. cDNA from squamous cell carcinoma (PanP63, ΔNP63, and TAP63) and giant cell tumour (PanP63, TAp63) served as the positive controls.

2.7.5 Primers

Desalted primers, were designed using the Primer3 design tool [148] (**Table 2.1 and 2.2**).

Table 2.1 Primer list.

Gene	Primer sequence (5'-3')		Product length (bp)
AKR1B10	GGCCTGGGCACTTGGAAGTCTCC	FWD	131
	TGGATGGCTTCCCCCACTTCATGT	REV	
EGF	TGTCTTGACTCTACTCCACCCCCTC	FWD	104
	AGGCAGTACCCATCGTGGGACA	REV	
ETV1	GCCAGATGTCTGAGCCAAA	FWD	154
	GCAAAATCTCTGGGTTCCTG	REV	
FGF1	GCTCTTTAGTCTTGAAAGCGCCACA	FWD	150
	TGGCCCCCGTTGCTACAGTA	REV	
GAPDH	GGAGTCAACGGATTTGGTCGTA	FWD	78
	GGCAACAATATCCACTTTACCAGAGT	REV	
HOPX	CGTGCTATCAGCAGCCTGCGT	FWD	113
	CATTAGTGTGGAAGAGGCAAAGGCA	REV	
IL-8	TTCTGCAGCTCTGTGTGAAGGTGC	FWD	148
	TGTGTTGGCGCAGTGTGGTC	REV	
PGK	GAAGAAGGAGCTGAACTACTTTGC	FWD	104
	TGATGAGCTGGATCTTGTCTGC	REV	
T	CCCGTCTCCTTCAGCAAAGTC	FWD	89
	TGGATTCGAGGCTCATACTTATGC	REV	
TGFA	GCTGCAGCAGTGGTGTCCCA	FWD	115
	AATGGCAGACACATGCTGGCTTGTC	REV	
WNT8A	TCAGTGCCTCTGCCTGGTCAGT	FWD	92
	CTGGGCACCCAAGGCCACAC	REV	
PAN P63	TCCTCAGGGAGCTGTTATCC	FWD	101
	ATTCACGGCTCAGCTCATGG	REV	
DNP63	GAAAACAATGCCCAGACTCAA	FWD	184
	AGAGAGCATCGAAGGTGGAG	REV	
TAP63	TGTTCAGTTCAGCCCATTGA	FWD	211
	CTGTGTTATAGGGACTGGTGGAC	REV	

Table 2.2 Probe list

Gene	FLUORESCENT PROBE SEQUENCE
T	FAM-CTCACCAACAAGCTCAACGGAGGG-TAMRA
PGK	VIC-CCTTGGAGAGCCCAGAGCGACCCT-TAMRA

2.7.6 Gene expression microarray

a) Downstream target gene study: Total RNA was quantified using Nanodrop 1000 Spectrophotometer and the Agilent Bioanalyser 2100 instrument. RNA samples with RNA Integrity Number (RIN) > 7 were chosen for further analysis by the in-house microarray facility – UCL Genomics. Total starting RNA material of 175ng was prepared using the Applause WT-Amp ST System (NuGEN, San Carlos, CA, USA), and fragmentation and labelling of the cDNA was carried out using the NuGEN Encore Biotin Module. The fragmented and labelled cDNA (5ug) was then hybridised to Affymetrix GeneChip® Human Exon 1.0 ST arrays (Santa Clara, CA, USA) as per NuGEN instructions for 16 hours at 45^{o}C. The arrays were washed and stained using the GeneChip Fluidics Station 450, and scanned using the Affymetrix GeneChip® Scanner. Expression Console 1.1 (Affymetrix) was used to assess quality metrics.

b) SNP association study: Total RNA was extracted from frozen chordoma samples from 20 patients which met the SNP association study criteria. RNA was extracted as described above. Only cases that were disomic for chromosome 6 with no *T* amplification (as

determined previously by FISH and qPCR) [17] were selected for further analysis (n=13). Five (4 AA genotype, 1 GA genotype) of these cases with a RIN>7 were hybridized to Affymetrix GeneChip® Human Gene 1.0 ST arrays (Santa Clara, CA, USA) as described above. Gene expression data from 2 additional cases (GA genotype) which met the above criteria were available from a previous gene expression study [56]. The intensity file data from these were batched with the results from the Gene ST1.0 array for normalization and batch correction as previously described [57].

2.8 Molecular biology: DNA

2.8.1 Genomic DNA isolation

a) Extraction from frozen tumour and non-neoplastic tissue samples:

DNA extraction was performed using the Qiagen DNA extraction kit (Qiagen, Crawley, West Sussex, UK) according to the manufacturer's instructions. The tumour samples comprised at least 80% tumour.

b) Extraction from blood:

DNA from blood was extracted using an in-house protocol detailed below:

Table 2.3 Lysis buffer preparation:

Reagent	Final concentration	Amount
DTAB (Sigma-Adlrich D8638)	8%	20g
NaCl (Sigma-Aldrich)	1.5mM	22g
Tris Cl pH 8.6 (Sigma Aldrich)	100mM	25ml of 1M Tris
EDTA (Sigma-Aldric)	50mM	25ml of 0.5M EDTA
dH20		Make up to 250ml

2. Chloroform

3. Ethanol (100% and 75%)

Method for 10ml of blood: Frozen blood was thawed by rotating the tube in a water bath at 37^{o}C. This was then transferred to a 50ml falcon tube to which 40ml of PBS was added. This was spun at 2,000 rpm in a centrifuge for 15 mins to form a pellet. The supernatant was removed and the pellet was resuspended using 2ml PBS. 4ml of lysis buffer was then added and mixed well, placed in a 68 ^{0}C water bath for 5 minutes, the cooled down to room temperature. 6ml chloroform was added, mixed and centrifuged at 4000rpm for 15 minutes. The

upper layer containing DNA in solution above a protein plug was transferred to a clean falcon tube containing 6ml of absolute ethanol. This was then centrifuged at 2,000 rpm for 5 minutes. The ethanol was removed and the DNA washed with 2ml of 75% ethanol. The ethanol was removed and any excess ethanol was briefly allowed to evaporate for a few minutes, The DNA was then dissolved in 1000μl DNAse molecular grade water and allowed to resuspend by being placed on a rotator overnight at 4^0C. DNA was then quantified using spectrophotomery and stored at -20^{o}C.

2.8.2 PCR amplification

Table 2.4 Primer list

Gene	Primer sequence (5'-3')		Product length (bp)
T (Exon 4)	TTCCCTCAACAGCAGAGACA	FWD	368
	CTTCCTCTCAGTGCGGGTTA	REV	
T(VNTR-Intron 6)	TTGTGAGGCAGACTGTGTCC	FWD	445/361
	CAGGGTTGGGTACCTGTCAC	REV	

a) PCR was performed for exon 4 using the following conditions:

For each reaction, 10μl of 10X Hot start buffer (Qiagen), 2.5μl of a 10pM dilution of each forward and reverse primers (**Table 2.4**), 1μl of DNA (50ng), 1μl of a mix of 10nM dNTP and 0.25μl of 1U Hotstart polymerase enzyme (Qiagen) were mixed in a total volume of 50μl. PCR was performed using a touchdown PCR protocol with annealing temperature reduced by 1°C per cycle from 65°C to 56°C, followed by 35 further cycles at 56°C. Each cycle was performed as follows: 94°C for 30sec, annealing temperature for 45sec, 72°C for 45sec. The polymerase was activated at by an initial step at 95°C for 15min before cycling and a final elongation step at 72°C was performed for 10min.

b) PCR for the VNTR was performed using a modified version of the above reaction conditions. For these reactions, 10µl of Q solution (Qiagen) was added to enhance the specificity of the amplification reaction. The PCR annealing reactions were conducted at 5°C higher than described above.

2.8.3 Taqman genotyping

Validatory genotyping on cases and controls was performed with the Applied Biosystems SNP Genotyping Taqman assay (ID C__11223433_10, Paisley, UK) for rs2305089. Allele assignment was carried out using the Applied Biosystems Sequence Detection System Software.

2.8.4 Whole exome sequencing [Performed by Dr Patrick Tarpey and members of Cancer Genome Project- WTSI]

a) Exome design

Exome enrichment was performed using the Agilent 50Mb Human Exome kit (Agilent Technologies Inc, Santa Clara, CA, USA) [149].

b) Genomic library preparation

Genomic DNA (5µg) was fragmented by Adaptive Focused Acoustics on a Covaris E120 (Covaris Inc, Woburn, MA, USA), purified using a Qiaquick PCR purification column (Qiagen, 28104), and quantified on a Bioanalyser using the Agilent DNA 1000 kit (Agilent, 5067-1504). The resulting DNA ranged from ~100-400bp, with a modal fragment size of ~250bp. Genomic libraries were prepared using the Illumina Paired-End Sample Prep Kit (Illumina, San Diego CA, USA). Adapter-ligated DNA was purified using AMPure beads (Agencourt

BioSciences Corporation, Beverly, MA, USA). The prepared library was used directly in the subsequent enrichment procedure without prior size-selection or PCR amplification.

c) Exome enrichment

The genomic library (500ng) was mixed with 7.5μg human Cot1 DNA, lyophilized in a speedvac and rehydrated in 3.4μl of nuclease-free water. Enrichment of the genomic DNA was performed using the Agilent SureSelect kit (Human 50Mb) with minor modifications to the manufacturer's protocol. Briefly, the genomic DNA library (3.4μl) was combined with 2.5μl of Block reagent 1, 2.5μl of Block reagent 2 and 0.6μl of Block reagent 3 and transferred to a well of a microtitre plate. The sample was denatured by incubating the plate on a thermocycler at 95°C for 5 min then snap-cooled on ice. A hybridization mix was prepared and a 13μl aliquot of this mastermix was added to the denatured DNA. The sample was incubated at 95°C for 5 min, then 65°C for 5 min. In a separate microtitre plate, the baits were prepared by combining 5μl of SureSelect capture library with 1μl of nuclease free water and 1μl of RNAse block, and then incubated at 65°C for 3 min. The pre-warmed DNA (22μl) was transferred to the pre-warmed bait mix and the solution incubated for 24hr at 65°C. Following hybridisation, the captured DNA was isolated using streptavidin-coated magnetic Dynabeads, (Invitrogen) and washed following the standard Agilent SureSelect protocol. The isolated DNA was purified using a Qiagen MinElute purification column, eluted and PCR-amplified for 14 cycles as previously described [150].

d) Sequencing

Each exome was sequenced using a 76bp paired-end protocol, on the Illumina GAII Analyser, to produce approximately 10 GB of sequence per exome. Sequencing reads were aligned to the human genome (NCBI build 37) using the BWA algorithm on default settings [151]. Unmapped reads, PCR-derived duplicates, or outside the targeted region of the genome were excluded from the analysis. The remaining uniquely mapping reads (~50-60%) provided 60-80% coverage over the targeted exons at a minimum depth of 30X.

2.8.5 ArrayCGH (design of chip in collaboration with Dr Karoly Szuhai, Leiden)

A custom high definition (HD) 8*15K Agilent array (Santa Clara, CA, USA) was designed specifically to interrogate the 6q27 *T* locus in addition to the centromeric and telomeric regions surrounding the locus to correspond to the flanking regions of the familial cases as described [15]. A replicate matching the *T* locus and a triplicate set corresponding to *T* gene exons only was also included in the design. The chip also contains whole genome coverage with lower density enabling interrogation of other known regions of copy number variation [54]. At the duplication locus and flanking sequence (Feb 2009 assemble GRCh37/hg19), probes were placed at high density at 1 probe per 100 bp.

The array hybridisations were performed according to the manufacturer's recommended protocols. 500 ng of genomic DNA was digested with restriction enzymes AluI and RsaI and fluorescently labelled using the BioPrime Total genomic labelling System kit (Invitrogen). Test samples were labelled with

Alexa Fluor® 3-dUTP and the reference sample with Alexa Fluor® 5-dUTP. Arrays were scanned using an Agilent scanner, and image analysis performed using default CGH settings of Feature Extraction Software 9.1.1.1 (Agilent Technologies). Data analysis was performed using Genomic workbench® (Agilent) using the AMD-2 algorithm with a conservative threshold of 9.8.

2.8.6 Chromatin immunoprecipitation and next generation sequencing (ChIP-seq). [Performed in collaboration with Dr Fiona Wardle's laboratory]

Three biologically independent ChIP experiments were performed. For each replicate, adherent U-CH1 cells (~$2x10^6$) were fixed in situ in the tissue culture flasks with 1% formaldehyde solution, which was quenched after 10 min by the addition of glycine (final concentration 0.125M). Cells were then washed twice with 1X Dulbecco's phosphate buffered saline (Sigma-Aldrich) and harvested using a cell scraper (Sarstedt, NC, USA), pelleted by centrifugation and flash frozen in liquid nitrogen.

ChIP was carried out as previously described [152] using 5µg of ChIP-grade anti-brachyury (N19) antibody (Santa Cruz; sc17743, CA, USA) which has been shown to recognise T in U-CH1 cells [57]. A paired-end library suitable for sequencing was generated from each immunoprecipitated DNA sample, and a whole cell extract input sample (30ng). The Illumina Paired-End DNA Sample Prep Kit (Illumina, CA, USA) was used for end repair A-tailing and adaptor ligation according to the manufacturer's instructions except that the adaptor ligation reaction constituted a 50µl reaction including 4µl of a 1-in-100 dilution of paired-end adaptor oligonucelotide mix. The ligated products were then

amplified with Phusion HF polymerase enzyme (Finnzymes, Vantaa, Finland) using PCR primers PE 1.0 and 2.0 (Illumina) before size selection and sequencing. 54 bp paired-end sequencing was carried out on the Illumina Genome Analyser II platform.

2.9 Molecular biology: Protein

2.9.1 Western blotting

A tumour lysate was produced from snap-frozen tumour sections using a lysis buffer mix consisting of Radio-Immunoprecipitation Assay (RIPA) lysis buffer (150 mM NaCl, 1%Igepal CA-630, 0.5% sodium deoxycholate, 0.1% sodium dodecyl sulfate, 50 mM Tris [pH 8.0]),), 5µl of protease inhibitor, 5µl of phosphatase inhibitor (Sigma) and 5µl of 1% SDS (Sigma). The lysates were incubated on ice for 15 min and then centrifuged at 13,000rpm at 4°C for 10 min to remove debris. Proteins were quantified with BCA protein assay kit (Fisher Scientific UK Ltd, Loughborough, Leicestershire, UK).

30µg of protein lysate were resolved by SDS-8% polyacrylamide gel electrophoresis (SDS-8% PAGE) and transferred to a polyvinylidenefluoride (PVDF) Immobilion-P transfer membrane (Millipore Corporation, Bedford, MA, USA) by standard semi-dry electro-transfer methods. The membrane was blocked with TBS, 0.1% Tween 20, 5% dry skim-milk for a minimum of 30 min and probed with the appropriate primary antibody overnight at 4°C on a shaker. Blots were washed three times (15 min each) in 1X TBS with 0.1% Tween 20 (TBS-T) and incubated for 1 hr at room temperature with the appropriate secondary horseradish peroxidise (HRP)-conjugated antibody, followed by

further washing and enhanced using chemiluminescence (ECL detection (GEHealthcare Ltd, Amersham, Buckinghamshire, UK). The details of antibodies used for Western blotting are shown in **Table 2.5**.

Table 2.5 List of antibodies

Antibody	Source	Primary/Secondary	Dilution	Incubation
P63, clone 4A4, mouse monoclonal	Santa Cruz, sc-8431	Primary	1:5,000	4^{o}C overnight
GAPDH, clone 6C5, mouse monoclonal	Advanced immunochemical Inc, Long Beach, CA, USA	Primary	1:10,000	Room temp, 1hr
Anti mouse-HRP, goat polyclonal	Dako, P0447	Secondary	1:5,000	Room temp, 1hr

2.9.2 Immunohistochemistry

a) P63

Three µm thick paraffin-embedded whole tissue sections and tissue microarrays (chordoma and osteosarcoma) were de-waxed by successive baths of xylene, ethanol and water, pre-treated and incubated with the p63 primary antibody (Santa Cruz, 4A4 clone). The reaction was performed on pre-coated slides using the Ventana NexES Autostainer (Ventana Medical Systems), following the manufacturer's instructions. Diaminobenzidine (DAB) was used as a chromogen in all reactions. The slides were counterstained with haematoxylin. A positive, IgG isotype and a negative control (e.g. no primary antibody added) were performed on tissue controls.

b) β-catenin

Immunohistochemistry for β-catenin (DAKO, mouse monoclonal, 17C2) was performed on chordoma tissue microarrays by UCL Advanced Diagnostics using in-house protocols. Appropriate positive and negative controls were performed.

2.10 Bioinformatics

2.10.1 Downstream target study:

a) Differential gene expression analysis [collaboration with Dr Stephen Henderson]

Expression values were calculated from Affymetrix GeneChip® Human Exon 1.0 ST data using the aroma.affymetrix package for the R statistical programming language (http://www.aroma-project.org/). Expression summaries for each gene were calculated using a custom chip description file (CDF) that collects individual probes into probe sets for ENSEMBL genes (V13;http://brainarray.mbni.med.umich.edu/Brainarray/Database/CustomCDF /13.0.0/ensg.asp). Significance of differential expression was estimated using the limma package from Bioconductor [153].

b) ChIP-seq data analysis [collaboration with Dr Andy Nelson]

All reads were mapped to the human genome (GRCh37) using Bowtie [154] implemented in Galaxy [155] using the following parameters: pairtries 1000, -k 2, –m 2, --best, -y. Peak finding was performed using CCAT using default parameters [156]. Regions exhibiting >5-fold enrichment over local background in all 3 replicate ChIP experiments was identified. The overlapping peak

regions present in all 3 replicates were used for further analysis. These are referred to as “ChIP-seq peaks” throughout. ChIP-seq peaks were categorised as promoter (±2kb from transcription start site of Ensembl genes), enhancer (-50kb from TSS to +5kb from transcription end site of Ensembl genes, excluding promoter regions but including the remaining gene) and intergenic [157, 158].

To determine if the ChIP-seq peaks were significantly enriched in promoter and enhancer regions these were compared to a simulated random distribution of the same peaks. Randomly selected regions of the genome of the same size as each of the ChIP-seq peaks were categorised as promoter, enhancer or intergenic as above. This was repeated 100 times. The number of peaks in each category from the ChIP-seq experiment was compared to the simulated distribution by one sample t-test.

c) De novo motif discovery:

Motif discovery within the genomic sequences of all ChIP-seq peaks was performed using default parameters in Weeder, except that both strands of the input sequences were searched [159]. Visualisation of the resulting position frequency matrices was performed using STAMP [160].

d) Gene set enrichment analysis (GSEA)

GSEA is a statistical method to determine whether a given gene set is enriched at one end of a rank ordered gene list, that is, whether a given gene set showed significant concordant differences between two biological states [161] . For the analysis of the microarray experiment GSEA v2.07 was used to generate a ranked list of differential expression based on signal to noise ratio of the processed microarray data. This was then interrogated with the GSEA C2

(Pathway module) and C5 (Gene Ontology module) gene sets. Gene sets which were enriched at a false discovery rate of <0.05 were deemed to be significant. GSEA was also used to test whether ChIP-seq peaks were associated with genes up- or down-regulated on knockdown of brachyury. Gene sets were defined as genes with promoter peaks only, genes with enhancer peaks only, and genes with both promoter and enhancer peaks. Each set was compared to genes rank ordered on t statistic obtained from microarray data analysis as defined above.

e) ARACNE algorithm

The ARACNE algorithm is a robust computational tool for determining transcriptional interactions and has been used successfully to reverse engineer cellular networks in silico [162]. Using the published protocol [163] and gene expression profiles of chordoma and chordoma cell lines made available by Josh Sommer (Chordoma Foundation, http://www.chordomafoundation.org/), a pairwise adjacency interaction matrix using mutual information was performed with ARACNE. The data from the intensity files were derived from four HG-U133A and six HG-U133 Plus2.0 Affymetrix chips. The data from the respective chip types were background corrected, quantile normalised and log2 transformed as separate datasets using GenePattern [164]. After normalisation, the probe sets with corresponding expression values from all the samples were exported into a single spreadsheet. These were then corrected for batch effect using the COMBAT algorithm [165] within GenePattern. The common probe sets were then collapsed into gene names, merged (average centred) and filtered. A stringent threshold of $P < 1x10\text{-}7$ was used for a significance threshold for the ARACNE analysis.

2.10.2 SNP association study:

a) Whole exome sequencing analysis [in collaboration with Dr Vincent Plagnol, UCL Genetics Institute and Dr Patrick Tarpey –WTSI]

The 21 germ line chordoma cases and the 220 controls (50M Agilent capture design) were re-aligned using the same algorithm (novoalign, www.novocraft.com). All samples (cases and controls) were called jointly using samtools mpileup (version 0.17). The resulting VCF file was annotated using the ANNOVAR package (http://www.openbioinformatics.org/annovar/).

To detect common variant associations an 800kb region around T (400 kb on each side) was used. The VCF file was filtered as follows: non-missing calls for more than 90% of cases and 90% of controls, SNP quality greater than 20 (Phred scaled), Hardy-Weinberg equilibrium less significant than 10^{-5}. This analysis identified 33 SNPs in this region, with 27 of them in dbSNP version 135. A SNP call was considered if the differential between the highest and the second highest genotype likelihood was greater than 5 (Phred scaled), otherwise the SNP was labelled as missing. The resulting genotypes were used to generate an association P-value for each SNP using a 2x2 Fisher exact test (R software).

In the second stage, potential rare variants were sought in chordoma cases. SNPs and indels were filtered on the basis of 1,000 Genomes frequency (< 0.5%) and the NHLBI Exome Sequencing project frequencies (< 0.5%). To filter for false positive SNPs which might be specific to the choice of tools for alignment and genotyping, an additional set of 150 control samples (non-cancer

related, heterogeneous set of Mendelian disorders) and filtered variants present in three or more control samples. SNP/indels were further filtered based on function (non-synonymous and stop-gain/loss SNPs, non-frameshift and frameshift indels). These filters defined a set of candidate calls. To investigate whether candidate calls accumulated in specific genes, the collapsed candidate variant frequencies in each gene with the genome-wide values between the 21 cases and the 220 controls (Fisher exact test) were compared. A small number of genes showed significant case control differences. A manual inspection of sequencing reads (Integrative Genomics Viewer, http://www.broadinstitute.org/igv) showed that these were driven by single low-quality SNPs, generally in DNA repeat regions, which were subsequently ruled out as genotyping errors in cases. After applying these filters, no gene showed convincing evidence of excess of rare variants in cases. Additionally, the subset of genes identified as targets of *T* were examined in depth which did not identify potential candidate genes in this set.

2.11 **Statistical analysis**

All experiments were performed as independent replicates. Error bars correspond to the standard deviation from the calculated mean. Microsoft Excel and SPSS (Version 19.0, Chicago, IL) were used for statistical analyses.Mean value and standard deviation was calculated using descriptive statistics.

Chi-square tests, Student T test, Mann Witney U test or Fisher exact tests were used, where applicable, to analyze differences between independent variants. Levels of significance are indicated in the main text and/or figure legends.

Chapter 3

The gene expression profile of chordoma cells depleted of T

In this chapter I introduce current details of the validated chordoma cell lines and some biological concepts regarding RNA interference. This is followed by experiments detailing the gene expression profiling of two chordoma cell lines which have been depleted of T and the functional significance thereof.

3.1 Introduction

3.1.1 Cell culture

Tumour cell lines are an invaluable resource for exploration of functional biological mechanisms and can be used to test hypotheses. In the recent past, research into chordoma has been hampered by the scarcity of chordoma cell lines. Although there has been an effort to encourage the generation of chordoma cell lines (www.chordomafoundation.org/ research/cell-line-prize/), *bona fide* examples are still not widely available. Recently the phenotypic and genotypic criteria for the definitive characterisation of a cell line as originating from a chordoma have been established [57]. These include:

1. the cells having a physaliphorous morphology
2. a common gene expression signature which includes the expression of *T* and *CD24* mRNA
3. a characteristic cytokeratin profile and
4. other reliable features: copy number alterations such as loss of *CDKN2A/2B* and *PTEN.*

Two well established cell lines that meet these criteria are U-CH1 and U-CH2 generated at the University of Ulm, Germany (**Figure 3.1**).

3.1.1.1 U-CH1

U-CH1 was the first human chordoma cell line generated and was established from a recurrence of a sacro-coccygeal chordoma from a 46 year old male [53]. Microscopically these cells resemble chordoma. They are physaliphorous with abundant cytoplasm, cytoplasmic vacuoles and round nuclei (**Figure 3.1**). Like chordomas, they co- express T protein, S100 protein, vimentin, EMA and cytokeratin. Cytogenetic studies of the U-CH1 cell line at low passage (pass 4) compared to its parental tumour have shown nearly identical profiles with gains in chromosomes 7,8p,9q34,12q24,15 and 20q and losses in 1p21-p34,10,11,14,18,22 [53]. The only difference between the cell line and the parental tumour was loss of 4q13-q28 in the cell line. An array CGH study has demonstrated homozygous loss of the *CDKN2A/B* loci on chromosome 9p21 with no regions of high level amplification in U-CH1 cells [57]. Gains in chromosomes 1,7,9,14,15,17,18 and 19 with losses on chromosomes 1,3,4,9,10 (*PTEN* locus), 11,13 (*Rb* locus), 18 and 22 were observed. These cells express both *T* mRNA and protein [57]. Hierachical clustering analysis of gene expression microarray data of chordomas show that U-CH1 clusters closely with human chordoma samples [57]. The U-CH1 cells show activation of the EGF receptor without evidence of mutation or amplification [68]. Treatment of the cells with the selective EGFR inhibitor tyrphostin (AG-1478) results in decreased proliferation, but active phosphorylated AKT protein is maintained. This suggests that other ligand receptor interactions such as through FGFR protein may maintain activated AKT protein in these cells and single agent therapy may result in a transient anti-proliferative effect with other “escape” pathways available to the tumour [63, 68]. The U-CH1 cells have also shown

sensitivity to STAT3 and mTOR inhibitors [34, 35, 166]. The first chordoma xenograft model was generated using the U-CH1 cell line in NOD/ SCID/interleukin 2 receptor [IL2r]γ^{null} mice. These tumours morphologically and immunohistochemically recapitulated primary human chordomas [17].

3.1.1.2 U-CH2

U-CH2 is a more recently characterised human chordoma cell line. It was established from a recurrent sacral tumour from a 76 year old female patient and has a similar cytogenetic profile to U-CH1 [57]. U-CH2 also demonstrates homozygous loss of *CDKN2A/B* but no losses of the *Rb* and *PTEN* loci. U-CH2 has losses on chromosomes 1p,2,3,4,8,10,14,16,17,19,20, X and gains on 1,6,7,12,15,16,19 and X. U-CH2 also expresses *T* mRNA but 630 fold lower than U-CH1[57]. The preserved *PTEN* locus in U-CH2 is responsible for the subsequent difference in AKT protein activation levels between the cell lines. Both cell lines are slow growing, adhere better to gelatine-coated flasks than uncoated flasks and have a doubling time of roughly one week.

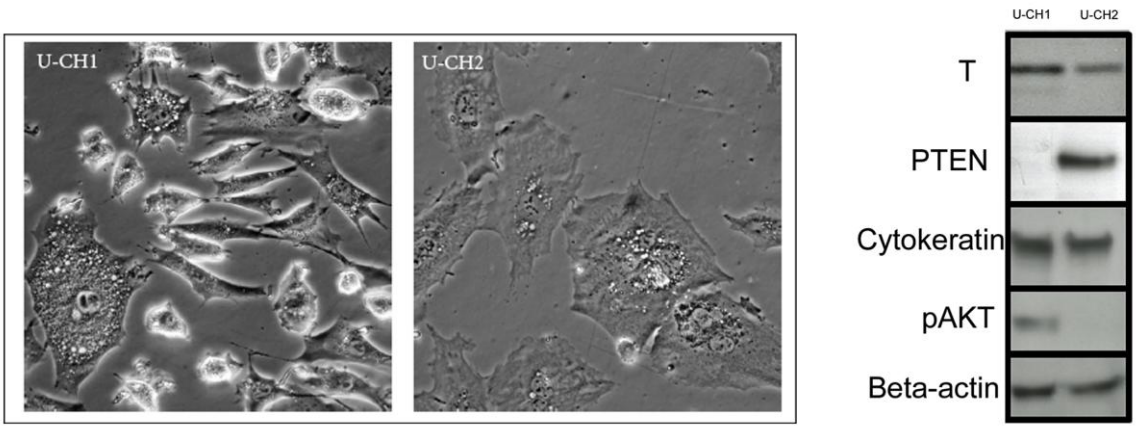


Figure 3.1 (adapted from reference[57]): Characteristics of chordoma cell lines.

Photomicrographs: Both U-CH1 and U-CH2 grow at a low density and are composed of a heterogeneous cell population. The physaliphorous cell component varies and is fewer in number in the U-CH2 cell line. Stellate and smaller rounded cells are also present in culture.

Western blot: The U-CH1 and U-CH2 cell lines have differing levels of T protein and U-CH1 shows loss of PTEN protein expression.

3.1.1.3 Other chordoma cell lines

Data regarding three new chordoma cell lines have recently been published which meet the validation criteria viz:

1. CH22 [167] established from a sacral chordoma of a 56-year old female patient with recurrent chordoma and liver metastasis. Xenografts have been established using this cell line which recapitulate the primary tumour and demonstrates immunoreactivity for T, cytokeratin 19, and S100 proteins.

2. Mug-Chor1 [168] established from a sacral chordoma of a 58 year old female. It expresses T protein and harbours chordoma typical cytogenetic aberrations such as chromosomal gains at 6q27 (*T* locus),

losses at 9p (*CDKN2A/CDKN2B* locus) and 10p (*PTEN* locus). It has failed to form xenografts in immunosuppressed Nod scid gamma mice (personal communication from Prof. A. Flanagan and Dr Beate Rinner).

3. JHC7 [169] established from a sacral tumour of a 60year old female. It displays many of the characteristics of chordoma mentioned above and has been successfully xenografted.

4. EACH -1 [170] is a cell line established from an extra-axial chordoma. It has a rapid doubling time and has been successfully xenografted.

3.1.2 RNA interference

The demonstration that short non-coding stretches of double-stranded RNA that were widespread in the eukaryotic genome and were responsible for post-transcriptional gene silencing was discovered by Andrew Fire and Craig Mello [171] for which they were awarded the Nobel prize in 2006.

3.1.2.1 Mechanism of RNA interference

Endogenous microRNAs or exogenous siRNAs repress gene expression by using specific sense-antisense binding to mRNA targets. This decreases protein expression by either destabilising the mRNA resulting in decreased mRNA levels through decay (the major mechanism) or through translational repression predominantly at the initiation phase (**Figure 3.2**). The application of this mechanism to probe gene function has since proved invaluable to researchers. The ability to deplete the expression levels of a gene provides insight into its possible functions within a given cellular context. Applications of this

technology vary from using synthetic molecules for example siRNAs, and morpholino oligonucleotides which can be delivered to the cell using liposomal-based transfection methods or electroporation. Stable knockdown can also be achieved by using viral transduction to integrate and constitutively express the short interfering hairpin constructs in the cell.

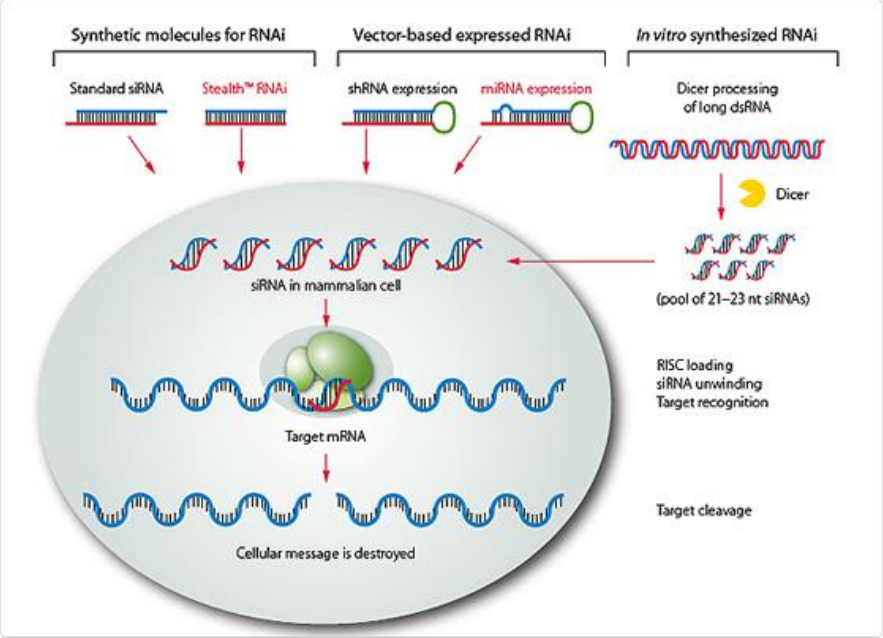


Figure 3.2 Schematic diagram of RNA interference mechanism.

RNA interference can be achieved by a variety of modalities. The mechanism involves, the double stranded oligonucleotides being processed in the cytoplasm by the enzyme, Dicer. This cleavage results in the antisense strand being incorporated into the RNA-induced silencing complex (RISC) which then targets the mRNA transcript for degradation or represses translation thereby reducing gene expression levels.

The choice of RNA interference approach is dependent on the experimental strategy and the protein dynamics of the gene in question. Some proteins are quite stable, and have a long half life. Therefore, prolonged suppression of mRNA is required to ensure that the residual protein in the cell has sufficient time to degrade, whilst no new protein is produced. This strategy involves the generation of a stable “knockdown” cell line using integrated shRNA constructs.

3.1.2.2 shRNA knockdown of T in chordoma cell lines

By using a shRNA gene knockdown approach we have previously shown that decreased expression levels of *T* mRNA and protein result in growth arrest [17]. This experiment has since been replicated by an independent group using different sets of shRNA constructs on the JHC7 human chordoma cell line [169]. The complementary experiment in which *T* mRNA was up-regulated in pancreatic cancer cells (which do not normally express *T*) resulted in increased cellular proliferation, motility and invasiveness *in vitro* [129]. These data implicate *T* as a determinant of cell proliferation in the context of chordoma and potentially in other cancers as well.

3.1.3 Proliferation assays

To date, the diminished proliferation of chordoma cells with reduced *T* expression has been assessed using indirect measurements of proliferation such as cellular confluence kinetics (Incucyte), and metabolic enzyme assays [17, 169]. Assays which incorporate nucleoside analogues are regarded as more sensitive readouts of proliferation with the added advantage of determining S phase cell cycle dynamics [172].

3.1.4 Migration assays

Chemotactic migration responses are required for local growth of cancer cells. Additionally, this ability is a pre-requisite for cells to invade, intravasate, extravasate and grow in distant sites and is chiefly mediated by a host of growth factors, integrins and chemokines [173].

There are various modalities that can be used to study different modes of migration in cancer, the choice of which is cost and experiment dependent viz:

1. *In vitro*:

- Cell culture systems using transwell assays (eg. Boyden chamber)
- 2D chemotaxis assays (eg. Dunn chamber or Zigmoid chamber)
- 3D cultures with cells embedded in reconstituted extracellular matrix gels

2. *In vivo*:

- Intravital multiphoton microscopy of live animals

3.2 AIM

To determine downstream transcriptional response and functional role of reduced T expression in chordoma.

3.3 OBJECTIVES

1. The knockdown of T expression has previously been performed in our laboratory by Dr Asem Shalaby [17, 174] on a single chordoma cell line. The objective is to optimise the protocol for a larger scale experiment and to perform it on two chordoma cell lines
2. Examine the proliferative and migratory ability of the knockdown cells
3. Perform gene expression microarray analysis to determine global transcriptional response to T knockdown
4. Determine if T-responsive genes in the cell line have relevance to human chordoma samples

3.4 RESULTS

3.4.1 Silencing of T in chordoma cells

The specific silencing of T in U-CH1 using a lentiviral infection with the shRNA plasmid V2HS_153729 which is directed at the 3'UTR of the transcript has previously been demonstrated [17]. This experiment was repeated using a larger number of cells for the purposes of performing the functional assays undertaken below. Silencing of T using the same vectors was repeated in the U-CH2 cell line.

3.4.2 *T* mRNA expression is significantly reduced by vector-mediated RNA interference

qRT-PCR assays revealed significant reduction of *T* mRNA expression in both U-CH1 and U-CH2 cells (**Figure 3.3**). This result is reproducible and has been demonstrated independently by others in another study using a different cell line [169].

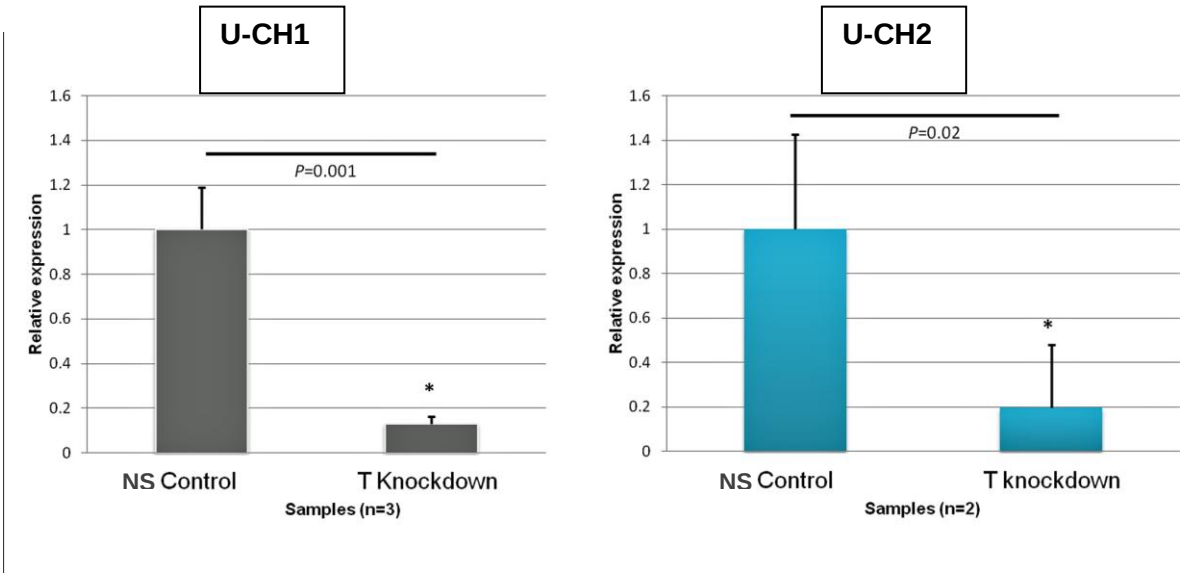


Figure 3.3: Validation of T knockdown in U-CH1 and U-CH2 cell lines.

Bar chart of qRT-PCR validation of T knockdown in chordoma demonstrating significant reduction of *T* mRNA expression in T knockdown cells ($P<0.05$) compared to the non-silencing control (NS). The expression of *T* mRNA was normalised to the housekeeping gene *GAPDH*. Bars represent the average of biologically independent experiments with +/- standard deviation. The qRT-PCR assays were performed as 3 technical replicates in duplicate.

3.4.3 T knockdown U-CH1 cells have diminished proliferative capacity *in vitro*

A flow cytometry-based nucleoside analogue assay was performed on U-CH1 and U-CH2 cells to determine their proliferative status upon T knockdown. This confirmed a significant reduction in proliferation capacity in the U-CH1 knockdown cells (**Figure 3.4**).

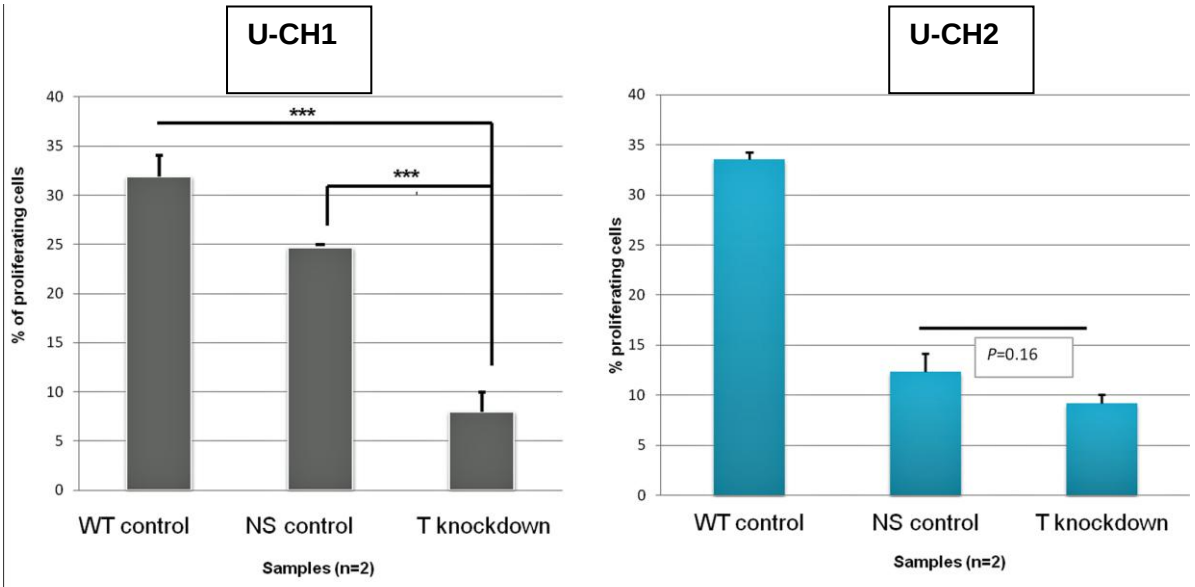


Figure 3.4: Proliferation assay of T knockdown cells.

The bars represent percentage of U-CH1 and U-CH2 cells incorporated with Edu nucleoside analogue. Left: There is a significant reduction in proliferative capacity in the U-CH1 T knockdown cells (***$P < 0.01$, Students t-test). Right: there is a reduction in proliferation of NS control cells compared to the WT cells but no decrease in proliferation in U-CH2 T knockdown cells compared to the NS control. (NS=non-silencing, WT=wild type)

3.4.4 T knockdown U-CH1 cells are arrested in G1 phase of the cell cycle

Bivariate analysis assessing the propidium iodide incorporation into DNA versus Edu incorporation allowed determination of cell cycle phase of actively cycling cells. This revealed that the T knockdown U-CH1 cells were arrested in the G1 phase in keeping with the reduction in proliferation (**Figure 3.5**).

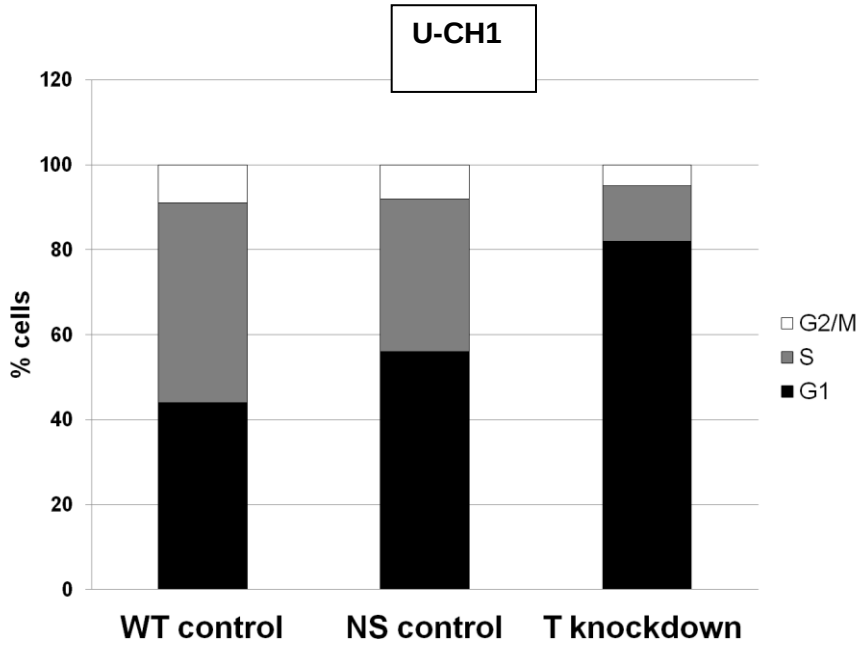


Figure 3.5 Cell cycle analysis of U-CH1 T knockdown cells.

Composite bar chart reflects that T knockdown blocks G1-S progression. G1 phase arrest is noted in T knockdown cells compared to NS and WT controls ($P<0.05$, Students t-test). Bars represent the mean percentages of biologically duplicate experiments. (NS=non-silencing, WT=wild type)

3.4.5 Serum-mediated migration responses are abrogated in T knockdown cells

Following serum starvation, the migratory capabilities of non-silencing control and T knockdown cells towards a serum-rich chemotactic stimulus was investigated. The directional migratory ability of the cells was assessed using their relative fluorescent intensity values which were determined using a microplate fluorescent plate reader. Visualisation of migrated cells was also confirmed by fluorescent microscopy. Both sets of T knockdown cells (U-CH1

and U-CH2) showed significantly diminished capacity to migrate across a membrane towards high serum conditions compared to control cells (*P*<0.05, T-test) (**Figure 3.6**).

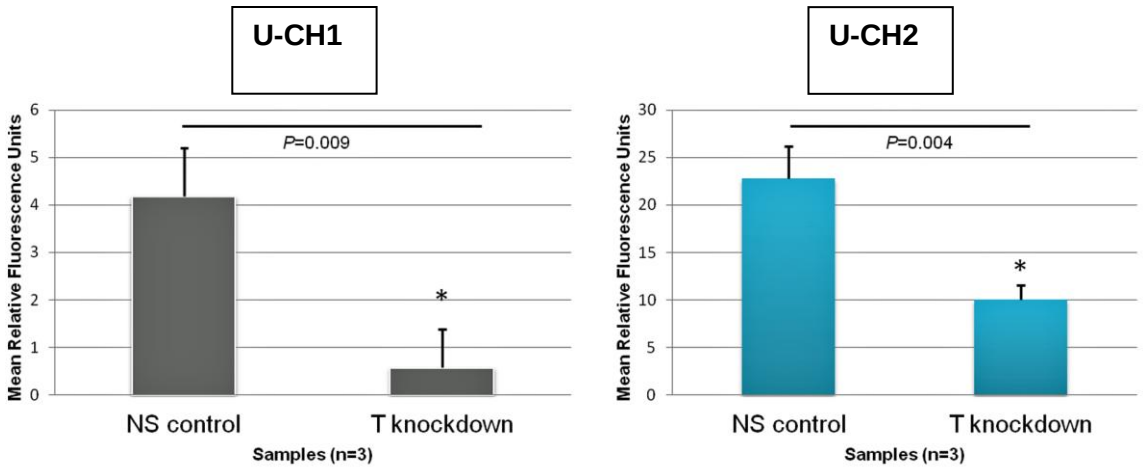


Figure 3.6 Migratory ability of T knockdown cells.

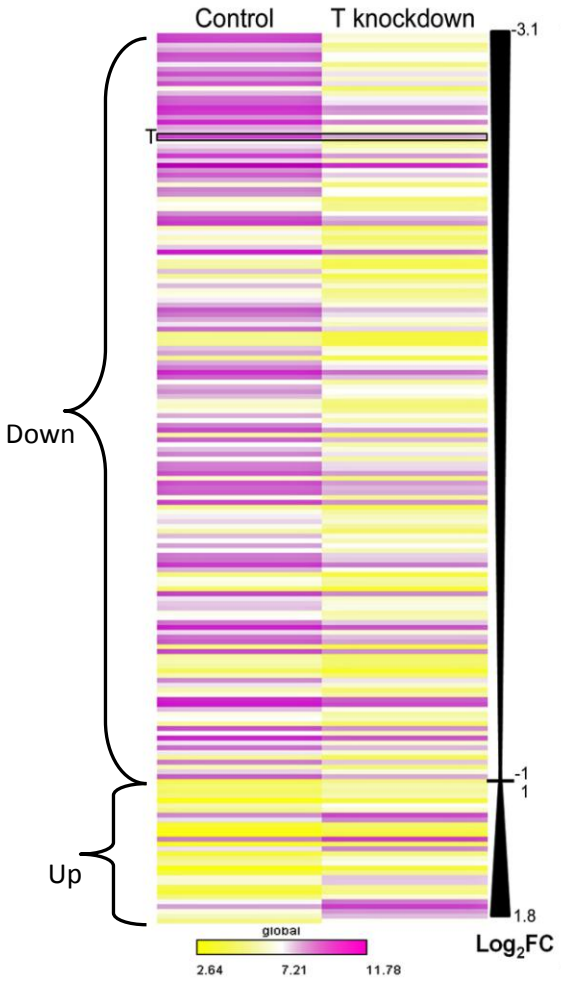
T knockdown significantly reduces migration of chordoma cell lines U-CH1 and U-CH2 on *in vitro* transwell migration assays. All cell lines stably express GFP. Bars represent mean relative fluorescent units of migrated cells +/-SD of 3 biologically independent experiments, (*$P < 0.05$, Students t-test). (NS=non silencing)

3.4.6 The transcriptional response of T knockdown U-CH1 cells accurately reflect their phenotype and behaviour

In order to determine the global transcriptional response to T knockdown, gene expression microarrays from triplicate knockdown experiments on the U-CH1 cells were performed. U-CH1 was chosen for this purpose as it is the most well characterised chordoma cell line.

Analysis at a genome-wide level revealed 1075 genes differentially expressed (at a threshold of *P*<0.01) between knockdown and control cells. In keeping with T's known role in development as a transcriptional activator there were a

greater number of genes down-regulated (176 genes with Log fold change ≤-1) than up-regulated (45 genes with Log fold change ≥1). As expected, this analysis confirmed that T was amongst the most down-regulated genes (**Figure 3.7**). To validate further the results from the microarray, qRT-PCR assays were then conducted using a selection of the down-regulated genes. The microarray results were confirmed with lower expression levels identified in the knockdown cells (**Figure 3.8**).



Downregulated					Upregulated
HGNC	AKR1B10	CDH19	ELK3	MKI67	PTCHD3
HPGDS	PTGS2	CHST4	ITM2A	SCRG1	EDNRB
EPYC	SCARA5	MPZL2	PTPRD	NA	TUFT1
WNT8A	SCEL	SPTLC3	KLRC3	DEFB118	MKNK2
KCND2	ENPP6	CPM	MMP16	CDH6	RDH10
GDA	ARHGAP11A	TXNDC16	SORCS3	USP12PX	NEFH
SLC15A5	PLA2R1	DLGAP5	KLRK1	ODZ3	LBP
KIAA1024L	DOCK11	SGCD	SLC6A12	NCAPG2	STARD5
MTMR7	NA	RAD51AP1	PLK4	NA	CXCL2
MYL1	HPCAL4	C8orf75	MMP1	ESYT2	IDO1
CX3CR1	CRB1	SOSTDC1	PRIM1	ADAMTS19	IL26
IL7	CD226	SHISA3	MAP2K6	ERCC6L	IL2RG
NA	CHST9	CKAP2L	RHOJ	MIR492	FBXO32
TRPM8	GLIPR1	BRIP1	TMPO	LIX1	C3
HVCN1	DSC1	ITGA10	HAS2	SLC7A2	AKR1C1
PLEKHH2	ALG10	IL8	MGAT5	ZNF385D	AKR1C2
RAB3B	KLRC2	PBK	PDE3B	CTGF	SPANXC
MRGPRX3	CNTN1	GALNT3	KIAA0101	ALPK2	THBS1
ACAN	NCRNA0011	GAP43	NA	ZNF385B	HMOX1
FGL2	FAM135B	SLCO5A1	CDH11	EXO1	MSC
RAPGEF5	ESCO2	GHR	TET1	CYP4X1	IFI27
T	GPC3	DUSP6	SPATA16	TMEM65	
GPR116	KIF15	ATAD2	FRY		
POSTN	C18orf54	TGFA	CDH10		
KIF11	TRIL	GPR64	GRPR		
HAPLN1	KRT19	SGOL1	ACSM3		
STC1	ADAMTS3	CD3G	NEIL3		
TSPAN7	CCDC141	SLC4A11	HIST1H3B		
KCNK2	SERTAD4	PRR11	PTCHD1		
HUNK	CENPI	HIST1H2BI	LAMA2		
FAP	ETV1	GABRA2	BUB1B		

Figure 3.7 Gene expression profile of T knockdown cells in U-CH1

A: Heatmap. The global representation of statistically significant differentially expressed genes ($Log_2FC \geq 1$ and $P<0.05$) in T knockdown U-CH1 cells reflects a predominance of down-regulated genes. Yellow=lower expression, violet=higher expression, Log2FC=log2 fold change.

B: Table. List of significantly differentially expressed genes ($Log_2FC \geq 1$, $P<0.01$)

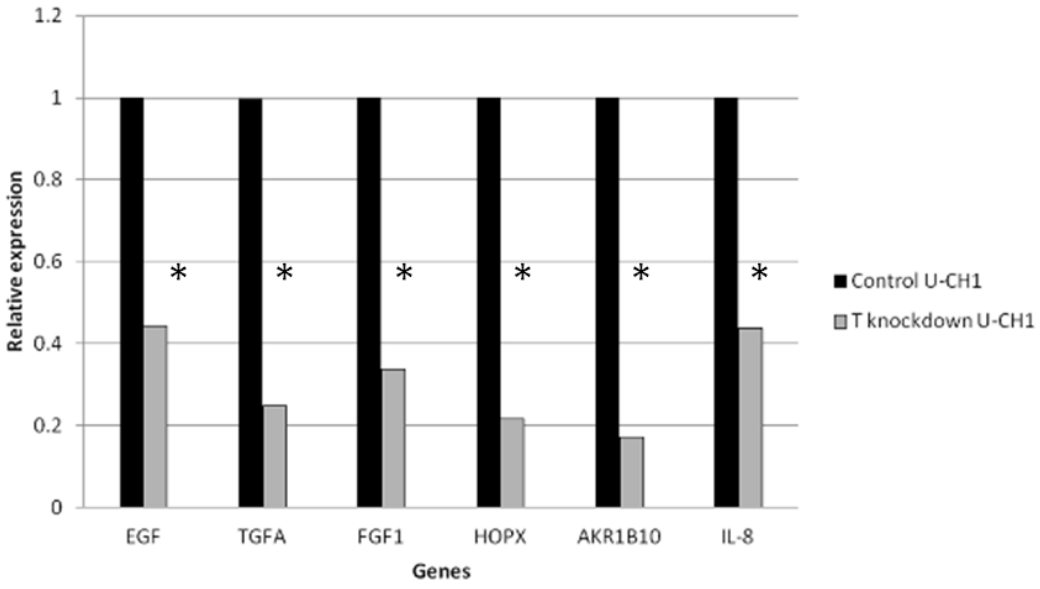


Figure 3.8 Validation of gene expression microarray data

qRT-PCR validation of Affymetrix microarray data of a selection of genes analysed using the comparative CT method. The fold change (relative expression) represents the average mRNA expression of 3 biological replicates. *P<0.01 (T-test). Control cells are cells stably expressing non-silencing control plasmid.

3.4.7 Functional annotation of differentially expressed genes

Modular and functional gene annotation enrichment of the significantly differentially expressed genes (P<0.01) was determined using Gene Set Enrichment Analysis (GSEA), and DAVID (Database for integrated visual discovery). This revealed enrichment for genes involved in regulation of the cell cycle (spindle, mitosis, sister chromatid segregation) (**Figure 3.9**), in the production of extracellular matrix and in glycosaminoglycan binding. There was also enrichment for the GO term locomotory behaviour, cell chemotaxis and for the Reactome-chemokine receptor-binding pathway (P=0.02).

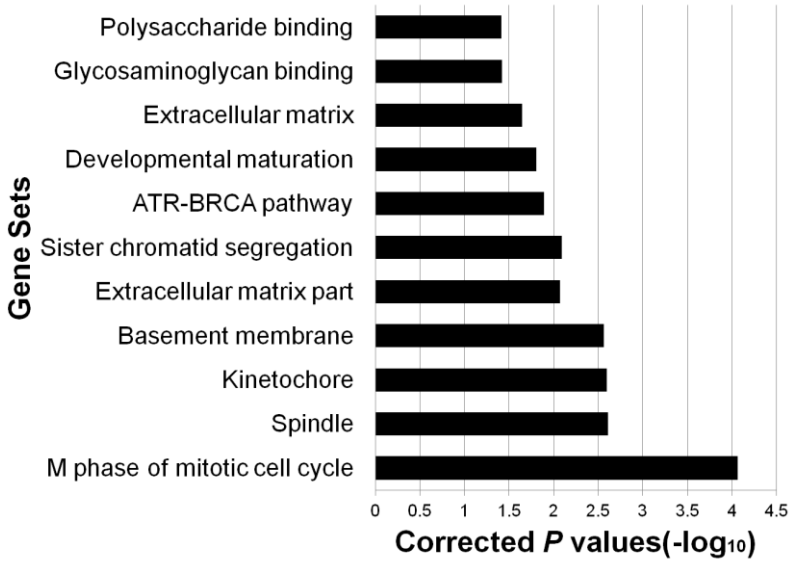


Figure 3.9: Gene ontology analysis of gene expression data in U-CH1.

Functional annotation analysis of the microarray data demonstrates significant enrichment ($P < 0.05$) of biological pathway and gene ontology gene sets.

3.4.8 GSEA leading edge analysis

Specific biological functions are dependent on the cellular context, environmental stimuli and tissue lineage. However, not all members (genes) of a gene set necessarily participate in their representative biological process/function. In order to determine which genes contribute to the cell cycle and extracellular matrix enrichment score in the gene set enrichment analysis in T knockdown U-CH1 cells, a leading edge analysis was performed (**Figures 3.10 and 3.11**).

Cell Cycle

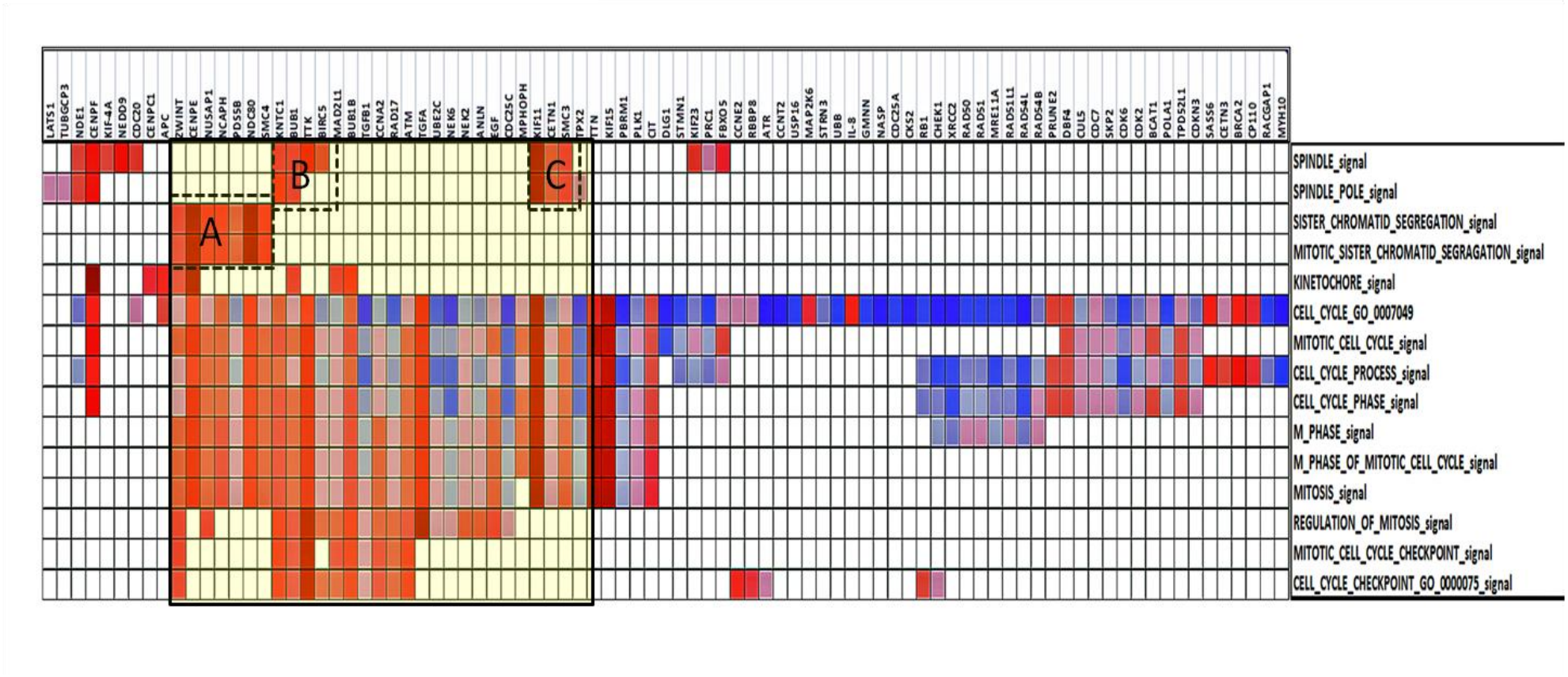


Figure 3.10: Leading edge analysis-cell cycle genes in U-CH1 T-knockdown cells

Heatmap of leading edge subset analysis for cell cycle related genes contributing to the core enrichment score. This reveals that a subset of genes involved in the chromatid segregation pathway (A and B) and spindle apparatus (C) separately contribute to the overall enrichment score for the cell cycle. Red=lower expression in T knockdown cells, Blue=higher expression in T knockdown cells

Extracellular matrix

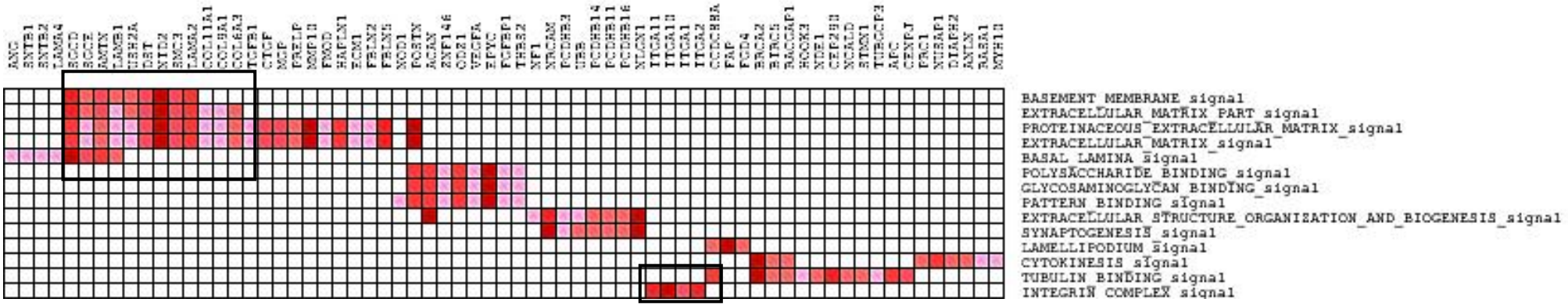


Figure 3.11: Leading edge analysis-extracellular matrix genes in U-CH1 T knockdown cells.

The core genes that contribute to the enrichment of the extracellular matrix and cell adhesion gene sets. Key components of the extracellular matrix such as Laminins (*LAMA2, LAMA4, LAMB1*) and collagens and of the integrin complex (*ITGA1, ITGA2, ITGA11*) (indicated in the rectangles) are contribute to the enrichment score for this gene set in the T knockdown cells.

3.4.9 Knockdown of T results in weaker adhesion to extracellular matrices *in vitro*

As a core set of extracellular matrix genes and basement membrane components for example laminins and integrins are significantly down-regulated by T knockdown it was reasoned that this may in part explain the reduced migratory ability of the knockdown cells. To investigate further the cells adhesive properties *in vitro*, tissue culture adhesion assays were conducted. This revealed that there was a diminished binding capacity of the U-CH1 T knockdown cells to laminin, collagen 1, and fibronectin. No major differences were noted in the binding ability to gelatine and growth factor reduced extracellular matrix gel (**Figure 3.12**).

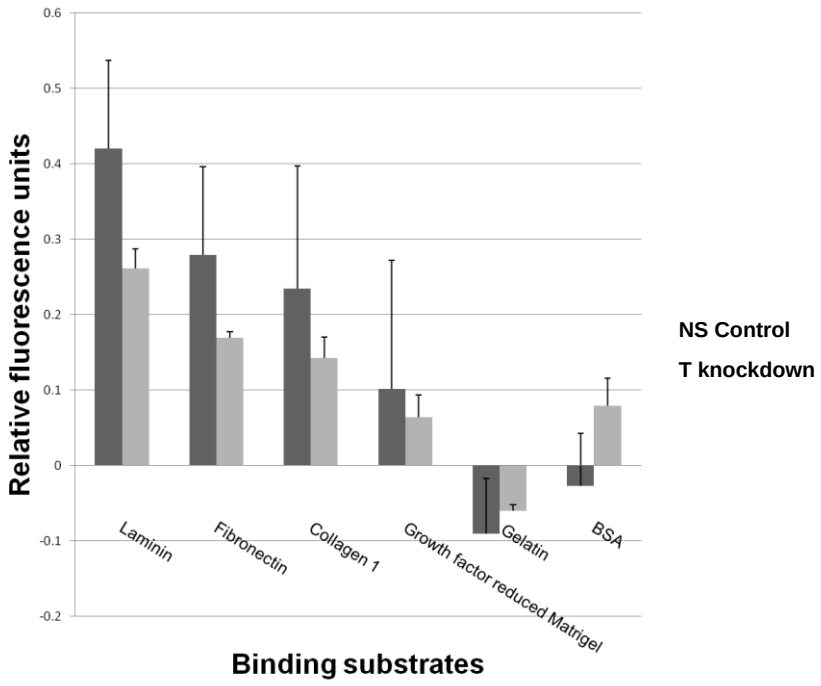


Figure 3.12 Adhesion of U-CH1 cells is reduced when *T* mRNA is knocked down. Bar charts show a general reduction of adhesion of T knockdown U-CH1 cells to a variety of matricellular substrates. Although a definite trend towards reduction of binding is noted in the T knockdown cells this did not reach statistical significance. (NS=non-silencing).

3.4.10 Differentially expressed genes in the U-CH1 cell line are relevant to human chordomas

If the differentially expressed genes on knockdown of T had functional relevance *in vivo* and were not cell line-specific, these T-responsive genes should be expressed in human chordoma samples at comparable levels to the cell line. We therefore analysed the global mRNA expression values from 8 human chordoma samples compared to three human chordoma cell lines (U-CH1, U-CH2, and K001) using gene expression microarray data provided by the Chordoma Foundation. A selection of genes responsive to T knockdown were then analysed in these datasets which revealed comparable levels of gene expression between the cell lines and human tumour samples (**Figure 3.13**). Global differential expression analysis between the primary tumours and cell lines revealed that there were only 12 significantly differentially expressed genes with a fold change ≥1.2 and threshold of $P<0.01$. Gene ontology analysis revealed that these genes were primarily involved in the extracellular matrix and plasma membrane.

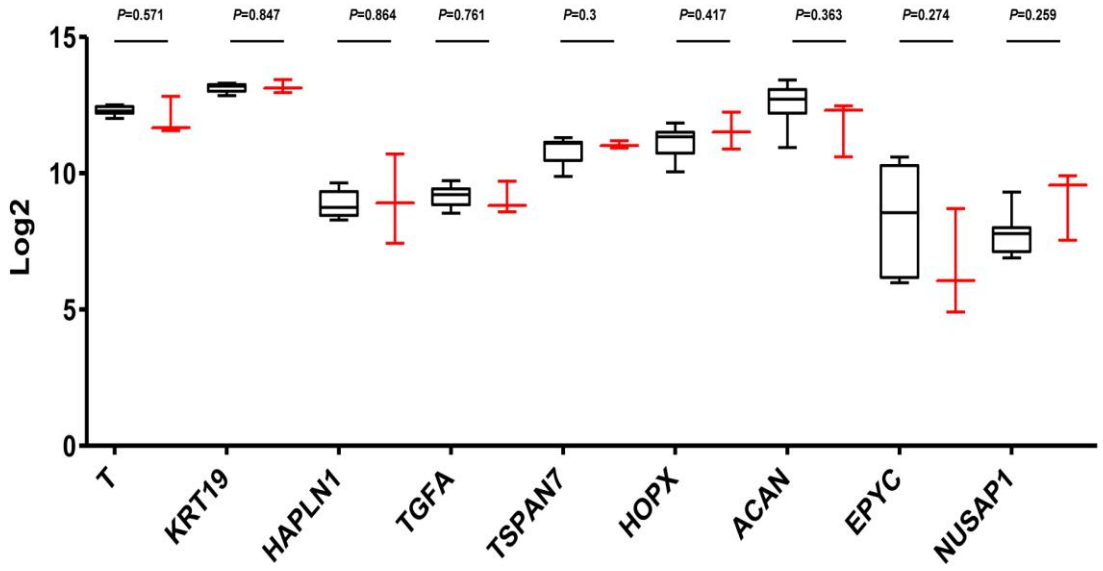


Figure 3.13 Comparison of mRNA expression levels between chordoma cell lines and primary tumours.

The pairwise comparison between the expression levels of relevant genes in human chordoma samples (black) and chordoma cell lines (red) reveals no significant differences between the two groups (t-test).

3.5 Discussion

Gene expression profiling studies of human chordomas and cells lines have demonstrated that chordomas have a set of specific highly expressed genes compared to other sarcomas [56-58]. Some of the relatively highly expressed genes in chordomas include *T, CD24, KRT 15, 18,19, DDR1, ITGA3, EGF, HOPX*. Chordomas also share high expression of extracellular matrix genes such as *ACAN, HAPLN1, CTGF and FN1* with chondrosarcoma [56, 58].

By functionally diminishing T expression in chordoma cells there is reduced expression of many of the core genes of this specific chordoma gene signature

providing evidence that they may be regulated by T in either a direct or indirect manner. The limitations of using one cell line for the knockdown and differential gene expression analysis is acknowledged. This is important as a single cell line would not be representative of the entire spectrum of genetic changes present in other chordoma cell lines and tissue samples. However the phenotypic and behavioural response of the U-CH1 cells upon T knockdown has been replicated by an independent research group on another chordoma cell line [169]. This provides some level of re-assurance that the effect of diminishing T expression may be similar in other chordomas. However there is no data regarding the effect on gene expression in this cell line to conduct a comparative analysis to U-CH1. The comparable expression levels of chordoma signature genes in the two cell lines, U-CH1 and U-CH2, compared to human chordoma samples also provides evidence that the *in vitro* experimental model is representative of the human tumours at least at the level of expression of core “chordoma genes”.

3.5.1 T and cell proliferation

Cancers are heterogeneous diseases, and chordoma is no exception. It is therefore necessary to have available as many cell lines as possible in an attempt to “capture” as much of that heterogeneity as possible in the *in vitro* environment. By examining the biology of multiple cell lines, experimental models can then be generated that reflect the variety/spectrum of *in vivo* disease processes as closely as possible. This point is illustrated from the results which show that knockdown of T in U-CH2 does not have the same anti-proliferative effect as in U-CH1 (**Figure 3.4**), and the JHC7 cell line [169]. It is important to note that in this experiment the effect of the non-silencing control

on the U-CH2 cell line results in a substantial reduction of proliferation compared to wild type cells and is similar to T knockdown cells. Although a reduction in proliferation is seen in NS control cells in U-CH1 the effect in U-CH2 contrasts markedly with the U-CH1 cells. This is unlikely to be a technical issue as the samples are biological replicates performed in technical duplicates and the virus has been biologically titrated. It is probable that this may represent an “on-target” effect of the NS control in this cell line either due to the nature of the scrambled sequence itself (i.e. a gene is inadvertently being targeted) or perhaps is related to an “off target” effect by the multiple random genomic integration of the plasmid with consequent over-production of the shRNAs within the cell thereby overwhelming the cells transcriptional machinery. Alternatively an aberration in cell culture for instance an undetected microbial contamination may be an explanation. This makes the effects seen on migration of U-CH2 cells upon T knockdown uninformative.

In U-CH1 cells, the propidium iodide cell cycle assay provided additional molecular evidence to the proliferation assay that these cells are growth arrested containing large amounts of G1phase DNA (**Figure 3.5**), a finding that is consistent with the growth arrest phenotype that has previously been reported [17].

Analysis of the downstream transcriptional response suggests that there is a deregulation of genes involved in assembling of the mitotic apparatus as determined by the leading edge analysis. The genes identified in this process include *BUB1, BUB1B, CCNA2* and *NEK2* (**Figure 3.10**). Although this may reflect the cells being in growth arrest, hence the lower expression of spindle checkpoint genes, there is evidence in the literature which suggests that cells

with defective spindle apparatus checkpoints can escape mitosis without chromosome segregation [175, 176]. The sequelae of this escape from mitosis may be either apoptosis, aneuploidy or G1 arrest. Whilst it is intriguing to speculate that the knockdown U-CH1 cells are in G1 arrest because they have escaped mitosis following downregulation of spindle apparatus checkpoint genes, there is no functional evidence for this.

The gene expression profile of T knockdown in U-CH1 cells corroborate the decreased proliferative capacity seen in these cells i.e. they are in growth arrest (**Figure 3.4**).

3.5.2 *T* and cell migration

In contrast to the differing proliferative capacities of T knockdown in U-CH1 and U-CH2, both cell lines show similar effects when assayed for their directional chemotactic migratory ability (**Figure 3.6**). This suggests that the responsible downstream genes and pathways that are responsible for this may be similar in both cell lines. Genes such as *CX3CR1*, *CNTN1* and *TGFA* which are down-regulated in the U-CH1 T knockdown cells have been implicated in cancer chemotactic pathways and regulation of adhesion associated integrins [173]. Also many of the growth factor ligands which are down-regulated in U-CH1 T knockdown cells have been shown to be crucial to invasive and metastatic processes [173]. This provides additional evidence that *T*, not only behaves as a proto-oncogene by regulating cell proliferation but is important for the motility of cells, a necessary determinant for malignancy.

3.5.3 *T* and cell adhesion

Cell adhesion is central to a variety of physiological and pathological processes [177]. It guides anchorage-dependent processes (for example migration, tissue remodelling) and signalling events (for example activation, survival, proliferation, differentiation) during embryogenesis, immune responses, wound repair and cancer.

The process of adhesion requires specific protein-to-protein or protein to carbohydrate bonds at the cell surface. Also, the interactions between adhesion receptors and cell membrane protein macromolecular complexes determine downstream signalling events by interacting with the cytoskeleton and signalling pathways. Cell adhesion can therefore occur between cell to cell and cell to extracellular matrix. Cell to cell adhesion is mediated by cadherins resulting in broad areas of the cell plasma membrane binding [178]. Specific binding is mediated by desmosomes and tight junction complexes which are required for migration. Integrins assist this process by regulating the binding of epithelial cells to the basement membrane for instance [179].

In cell-matrix interactions, plasma membrane adhesion receptors, such as integrins, bind to extracellular matrix substrates. Although there is some specificity of binding conferred by this process there are also large areas of overlap in the binding affinity mediated by various integrin dimer complexes on the cell surface [180] (**Table 3.1** – adapted from reference 15).

Table 3.1 Table of the common ligand-matrix interaction pairs in human cells.

Integrin	Ligand
α1β1	Collagen IV
	Laminin
α2β1	Collagen I
α3β1	Laminin
	Thrombospondin
α4β1	Fibronectin
	Opsonin
α5β1	Fibronectin
α6β1	Laminin
αIIBβ3	Fibrinogen
	Fibronectin
	Vitronectin
	vWF
αVβ3	Fibronectin
	Vitronectin
	Fibrinogen
	vWF
	Thrombospondin
	Opsonin
	Tenascin

The finding of significantly down-regulated extracellular matrix genes such as laminin (*LAMA2*) and ligands such as *ITGA3* prompted the investigation of whether this influenced matrix binding of the U-CH1 cells *in vitro*. These results demonstrate a decreased adherence of the U-CH1 cells to various extracellular matrix substrates tested. This result is bolstered from evidence from the developmental biology literature where it has been shown that the (MOUSE) $T^{-/-}$ mutation is embryonic lethal. Ultrastructural examination of these embryos show significant disorganisation of the extracellular matrix that comprises the notochordal sheath with disruption of the notochordal cells and adjacent somites and neural tube [82]. This provides additional support for implicating *T* in the regulation of the production of extracellular matrix components so critical for migration and invasion.

3.6 Conclusion

Overall the gene expression analysis and functional assays reveal important insight into key functions of *T* not previously described at a genetic level. Furthermore, it provides a list of downstream T-responsive genes. The comparative expression analysis between chordoma cell lines and human chordoma samples suggest that these potential target genes are not restricted to chordoma cell lines but appear to reflect what occurs in human disease, and have relevance to chordomas in patients.

Chapter 4

The transcriptional network regulated by T in chordoma.

In this section I introduce the utility of next generation sequencing to ascertain genome-wide transcription factor binding profiles (ChIP-seq). I then examine and discuss the genome-wide binding profile of T using ChIP-seq. These data are then integrated with gene expression data from the previous chapter to identify direct transcriptional targets and generate a gene regulatory network directed by T in chordomas.

4.1 Introduction

Before the advent of whole genome sequencing, determining transcription factor binding sites was generally a low throughput process. This was accomplished using chromatin immunoprecipitation (ChIP) followed by conventional PCR techniques [181]. The advancements in microarray technology (ChIP-chip), and next generation sequencing (ChIP-seq) has facilitated the genome-wide investigation of transcription factor binding.

4.1.1 Chromatin immunoprecipitation with next generation sequencing (ChIP-seq)

The genome-wide identification of protein-DNA interactions is a pre-requisite to understanding transcriptional regulation. This can be achieved by ChIP which involves the isolation of protein-DNA fragments bound to a specific antibody *in vitro*. Hybridisation of these fragments to promoter microarray chips (ChIP-chip) was the one of the logical early iterations of employing this technique in a broader, genome-wide context [182]. This allowed, for the first time, an unprecedented insight into large scale binding profiles of transcription factors.

However, with the advent of next generation sequencing (NGS) technology this has been superseded by chromatin immunoprecipitation followed by NGS (ChIP-seq) which is a less biased approach to genome-wide binding identification. In this process, the bound fragments are sequenced instead of being hybridised to a chip that is spotted with predetermined oligonucleotides. Computational mapping of these sequences back to a reference genome and determination of enrichment of the sequenced fragments enables precise identification of transcription factor binding sites, also referred to as peaks (**Figure 4.1**) at high resolution. The exploitation of this technique to investigate factors such as nucleosome positioning and histone modifications has yielded unheralded insight into the complex processes of transcription. An international collaborative effort using ChIP-seq and a variety of related modalities is currently underway to determine and document the regulatory dynamics of transcription factors and histone modifications in a variety of cell lines – the ENCODE project [183].

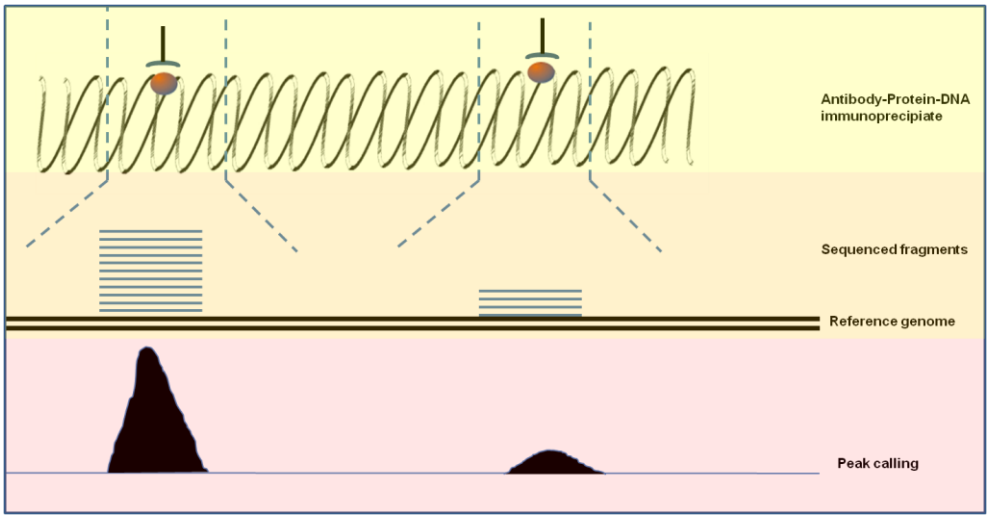


Figure 4.1: Schematic of the ChIP-seq method.

The protein of interest (eg transcription factor) is crosslinked to DNA using formaldehyde. The DNA is then sonicated to fragments of about 150-500bp. A specific antibody is then immunoprecipitated with the protein-DNA complex of interest. DNA libraries are then constructed, amplified and sequenced using next generation sequencing platforms. The short sequenced fragments (tags) are mapped to a reference genome and regions of enrichment (binding events) are identified as peaks using peak-calling software.

4.1.2 DNA binding motifs

The DNA binding domains of transcription factors enable them to bind to specific DNA sequences. These are short stretches of nucleotide sequences which in large part regulate the specificity of the binding, the other major determinants being the availability of co-factors and chromatin accessibility. These DNA binding sites are referred to as consensus sequence motifs and can be represented by positional weight matrices which specify a score for the frequency of each nucleotide in the binding site based on the number of known instances of the transcription factor binding to this sequence (**Figure 4.2**).

	1	2	3	4	5	6	7
A	1	4	1	2	0	17	13
C	28	5	5	0	3	3	2
G	0	0	4	0	25	1	7
T	2	22	21	29	4	10	9

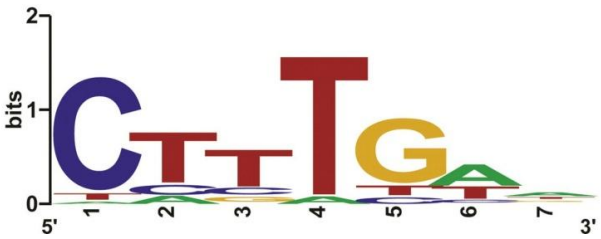


Figure 4.2 DNA transcription factor motifs.

An example of a positional weight matrix where the height of each letter is proportional to the score of the nucleotide position in the motif. Many gene specific transcription factors have specific motifs and this example represents the *SOX* trio (*SOX5*, *SOX6,SOX9*) recognition motif in chondrocyte differentiation.

These motifs are catalogued in various databases such as Jaspar and Transfac [184, 185]. In theory, with the availability of annotated genomes, and the knowledge of the specific binding motifs, it should be straightforward to identify all target genomic regions of transcription factors. In reality, the number of motifs far outweighs the number of direct functional targets. This is because not all motifs are bound and not all binding events occur at motifs. In fact the majority of the motifs in the genome are not bound due to availability of regulatory factors, chromatin dynamics, and competition by other transcription factors. Secondary motifs (sequences that are bound less frequently) can be bound with almost as high an affinity as for primary motifs. Also, motifs are short and degenerate leading to the high number of false positives calls *in silico* [186]. Other issues that relate to the identification of high quality peaks with low quality motifs are due to the influence of co-operative binding with a protein partner resulting in indirect protein DNA interactions [182]. Therefore, although ChIP-seq is an invaluable tool to identify the binding dynamics of proteins, in the context of identifying functional gene targets, it is not entirely specific.

4.1.3 Dynamics of transcription factor binding

A *cis* (Latin ~ “on the same side as”) element is a region of DNA that regulates the expression of genes located in the same vicinity of the DNA molecule. This is usually on the same chromosome and multiple genes can be co-ordinately transcribed by the binding of a transcription factor to a common *cis* regulatory sequence. These sequences are referred to as promoters, enhancers, silencers and insulators depending on how they regulate gene expression.

Core promoters, for instance the well known TATA box, are usually located just a few hundred base pairs upstream of the transcriptional start site of a gene. Although most core promoters are necessary for initiation of transcription, the complexes formed at these sites are unstable, weakly active and are insufficient for complete initiation of transcription. In fact data from the ENCODE project have shown that promoters such as the TATA box are the exception rather than the rule. Enhancers of gene transcription are required for this purpose. These can be located some distance away from the transcriptional start site but stabilise the core transcriptional complex at the site of initiation and potentiate transcription (**Figure 4.3**). A well studied example is the oestrogen receptor in breast cancer. For its direct target genes, a third of the enhancers lie within 50kb from their promoter sites and for its other regulated genes these enhancers can be considerably further away [187]. The stabilisation of these transcriptional complexes at these regulatory sites is maintained by core transcription factors such as CTCF, and proteins known as cohesions.

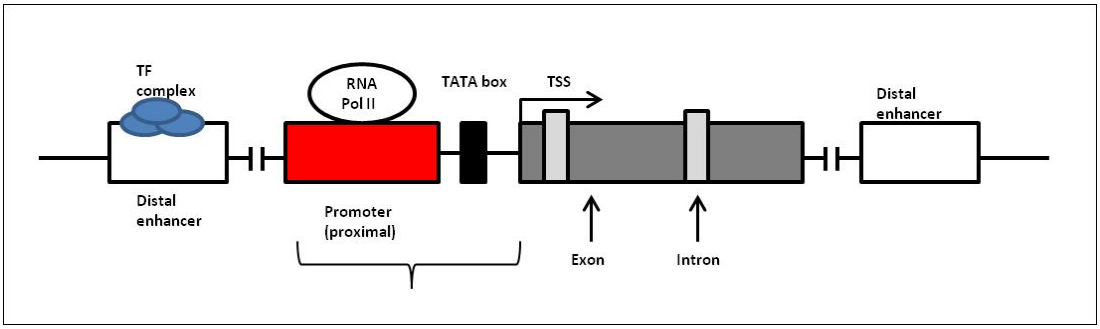


Figure 4.3 Schematic of the structure of genetic elements of transcription.

DNA sequence motifs at regulatory regions confer specificity of gene expression by recruitment of transcription factors. RNA Pol II and other TFs bind to core promoter regions such as TATA boxes thus providing a low level of transcriptional activity. Promoter activity can then be increased by the binding of other factors with *cis* elements (e.g. enhancers) which can stabilise the recruitment of the transcriptional machinery through a direct interaction between the site-specific factor and the general factors. The enhancers also recruit histone modifying enzymes which facilitate a chromatin environment conducive for transcription.

4.1.4 Chromatin accessibility

All $2x10^{13}$ metres of DNA is efficiently packaged in the nucleus by chromatin almost like a thread around a spool. This chromatin fibre is referred to as the nucleosome and consists of wound up packages of DNA around an octamer of histone cores. In order for transcription factor proteins to access the DNA it is important that chromatin makes these regions accessible. This process involves chromatin remodelling, nucleosome re-positioning and histone modifications. Researchers have been able to take advantage of these changes to identify regulatory regions of the genome. Sites with nucleosome disruption are indicative of open regions of DNA accessible to transcription factors and can by identified by their DNAse1 hypersensitivity.

Promoter regions are characterised by specific methylation marks on histones such as Histone 3 Lysine 4 Trimethylation (H3K4Me3) and enhancers are characterised by Histone 3 Lysine 4 Monomethylation (H3K4Me1). ChIP grade antibodies for these specific methylation marks are available and are increasingly being used to define these regulatory regions of the genome.

4.1.5 Variation in transcription factor binding

Are transcription factor binding sites constrained amongst divergent species?

Comparative genomics or phylogenetic footprinting have been mooted as complementary strategies to ChIP-seq as a means of identifying transcriptional targets in human tissue based on the hypothesis that conservation implies function. For example, the combination of conservation, gene expression and binding site mapping have been used to identify putative constrained enhancers in mice [188]. However, evolutionary constraint appears to be the exception rather than the rule. Large scale divergence in transcriptional factor binding was demonstrated in homologous mammalian tissues of mouse and human. The genome-wide binding profile of four highly conserved tissue specific transcription factors were examined in murine and human hepatocytes [189]. This revealed that the majority of binding events were species specific and that these binding events near orthologous genes did not align at a sequence level. The same researchers, on the same tact, used ChIP-seq to examine the binding of two transcription factors across five vertebrates: human, mouse, chicken, dog and opossum and showed that a large proportion of binding was species specific [190, 191]. They next investigated whether genetic sequences or epigenetic phenomena were responsible for divergent binding between species. This was performed using an elegant model of an aneuploid mouse, carrying an extra chromosome, that of human chromosome 21, and showed that the binding repertoire of the tissue-specific transcription factor in humans was conserved in the “hybrid mouse” compared to wild-type mouse [192]. This suggests that genetics (that is, sequence of DNA) plays a larger role with regards to transcription factor binding specificity compared to epigenetic phenomena. In view of findings such as these a strong element of caution is required when attempting to use conservation as a means to identify transcriptional regulatory targets.

Are transcription factor binding sites constrained <u>within</u> species?

Genetic variation is the proverbial “spice of life”. Single nucleotide polymorphisms occur at about 1-2SNPs per kb of the human genome and play an important role in our phenotypic diversity. In view of this, it is not surprising that variations in *cis* regulatory elements occur in humans. Using lymphoblastoid cell lines from candidates of European, Asian and African ancestry, it was shown that the general transcription factor binding site selection of RNA polymerase III varied by as much as 25% amongst individuals [193]. A smaller but no less significant element of variation was seen with NFκB. These differences were not only attributable to SNPs in the binding regions but also to larger structural variants and were correlated to differences in gene expression.

Variations in *trans* (Latin ~”on the other side”) also influence transcription factor binding. Functional polymorphisms within transcription factors themselves (particularly the DNA binding domain) can result in varied binding affinity and selection of target genes [194]. Also, to a large extent differences in binding within species are determined by the underlying tissue type. Many tissues have specific transcriptional programs and respond to environmental stimuli in different ways. The availability of co-factors can therefore critically determine cell type specific binding events. This has been demonstrated for many T-box genes for example Xbra (Xenopus brachyury), the functional specificity of which is determined by interactions with a partner protein Smad1 [97]. Likewise, TBX2, and TBX5 proteins associate with their partner protein NKX2-5 during cardiomyocyte differentiation [195, 196].

4.2 Aims

1. To determine the genome-wide binding profile of T in chordoma
2. To identify the direct transcriptional targets of T in chordoma

4.3 Objectives

1. Perform ChIP-seq experiments on the U-CH1 chordoma cell line using a specific antibody to determine T protein binding sites.
2. Conduct *de novo* motif finding from T protein bound sites.
3. Integration of ChIP-seq data with gene expression microarray data to determine direct targets of T.
4. Validation of targets in human tissue samples using *in silico* approaches including qRT-PCR and Immunohistochemistry.
5. Modular enrichment analysis of target genes to determine their putative functions, and network generation.

4.4 Results

4.4.1 Genome wide identification of T binding sites

Transcription factors carry out their activation and repressor functions by binding to gene regulatory elements. These include promoters which lie in close proximity to genes and enhancers and silencers whose genomic distribution can be variable (see above). In order to identify genome wide binding of T protein in chordoma, triplicate ChIP-seq experiments were performed in the U-CH1 cell line. Following library preparation, the paired end sequenced fragments were mapped to the GRCh37/hg19 version of the reference human genome. 6420 reproducibly bound regions were identified from the overlap of the three experiments. Using this overlap is a deliberately stringent approach with the intention of reducing the false positive call rate. Peak calling was performed using the bioinformatics package, CCAT whose algorithm is based on a signal to background noise ratio (**Figure 4.4**) [156].

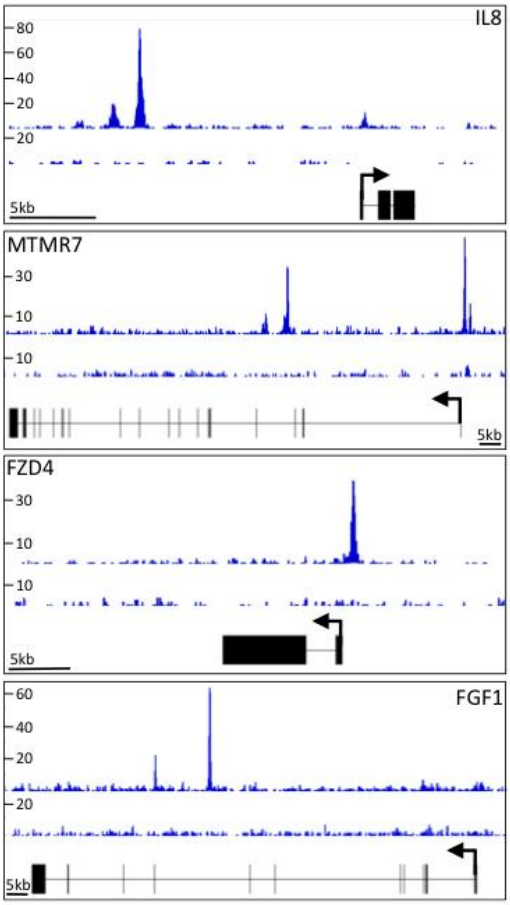


Figure 4.4 Representative ChIP-seq reads.

Sequence depth is indicated on the left and is representative of merged replicate ChIP-seq experiments.

This analysis revealed that 69% of the binding occurred in close proximity to the transcriptional start sites of genes. These binding sites were then categorised as promoters if they lay within +/-2kb of transcription start sites, as enhancers if they extended from -50kb of transcription start sites to +5kb of transcription end sites (excluding promoter regions) or anywhere else on the genome as intergenic binding [157]. The distances from genes were determined from those annotated in the Ensembl database [158]. Analysis of the binding events showed that there was a significant enrichment in both promoter and enhancer regions compared to intergenic regions (**Figure 4.5**) supporting prior evidence that being a transcriptional activator, T binds in close proximity to genes rather than to other regions of the genome.

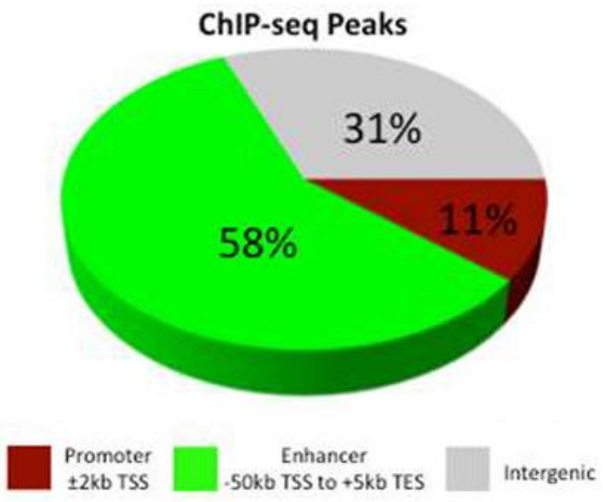


Figure 4.5: Distribution of ChIP-seq peaks in U-CH1

Pie chart indicating distribution of ChIP-seq peaks. ChIP-seq peaks are significantly enriched within promoter and enhancer regions as defined above (P=2.10x10^{-101} and 7.13x10^{-110}).

4.4.2 T binds the canonical T-box in chordoma

T binds the palindromic sequence T(G/C)ACACCTAGGTGTGAAATT as a dimer. Electrophoretic mobility shift assays show that It can also bind to the half palindrome sequence (T/C)TTCACACCT with high affinity [94, 95].

In order to determine if the binding events identified in U-CH1 were true T binding events *de novo* motif finding was performed on the ChIP-seq peaks. This revealed that the binding identified in the chordoma cell line was to a consensus T box motif and was virtually identical to those motifs identified in other model organisms (**Figure 4.6**) [99]. In view of the previous reports of the ability of T to sometimes bind to unusually long DNA sequences [95] the motif finding analysis was extended to search for possible longer binding motifs. Interestingly this analysis demonstrated a putative ETS family binding site adjacent to the canonical T-box suggestive of co-binding with a member of this the ETS transcription factor family.

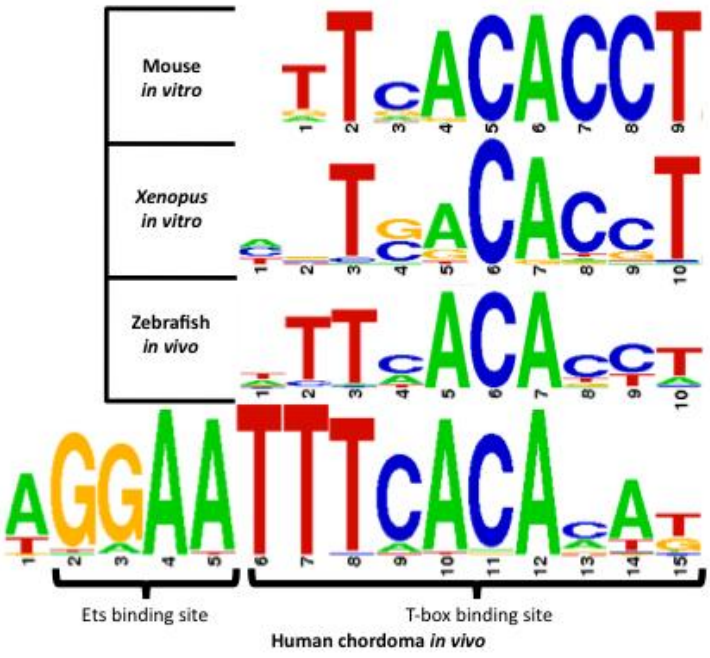


Figure 4.6 *De novo* motif finding:

The discovered binding motif for T protein in the U-CH1 chordoma cell line is similar to the published consensus T-binding site for Xenopus brachyury, mouse T and zebrafish Ntla. The discovered motif includes an Ets binding site adjacent to the canonical T-box.

4.4.3 Integration of ChIP-seq and gene expression data

Although the gene perturbation experiments in the U-CH1 cell line followed by gene expression microarrays provides insight into the downstream functional transcriptional programs active in chordoma, it does not specifically identify direct targets. Conversely, the identification of T protein binding sites in U-CH1 using ChIP-seq provides additional information regarding binding but does not necessarily imply function. Reasoning that the gene expression and ChIP-seq data would provide complementary information regarding functionality of the binding sites, this data was then integrated (**Figure 4.7**).

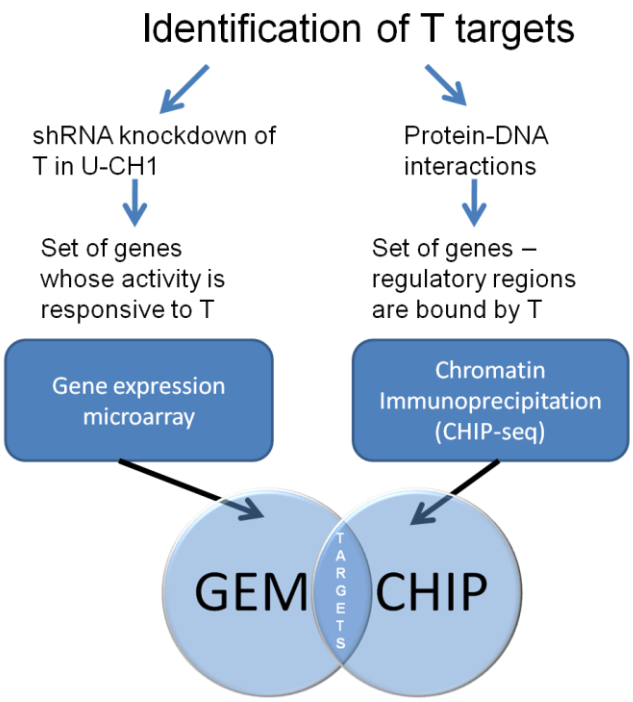


Figure 4.7 Experimental and analysis workflow: Schematic diagram explaining the rationale and experimental workflow for identifying T targets.

4.4.4 T positively regulates the transcription of genes identified by ChIP-seq

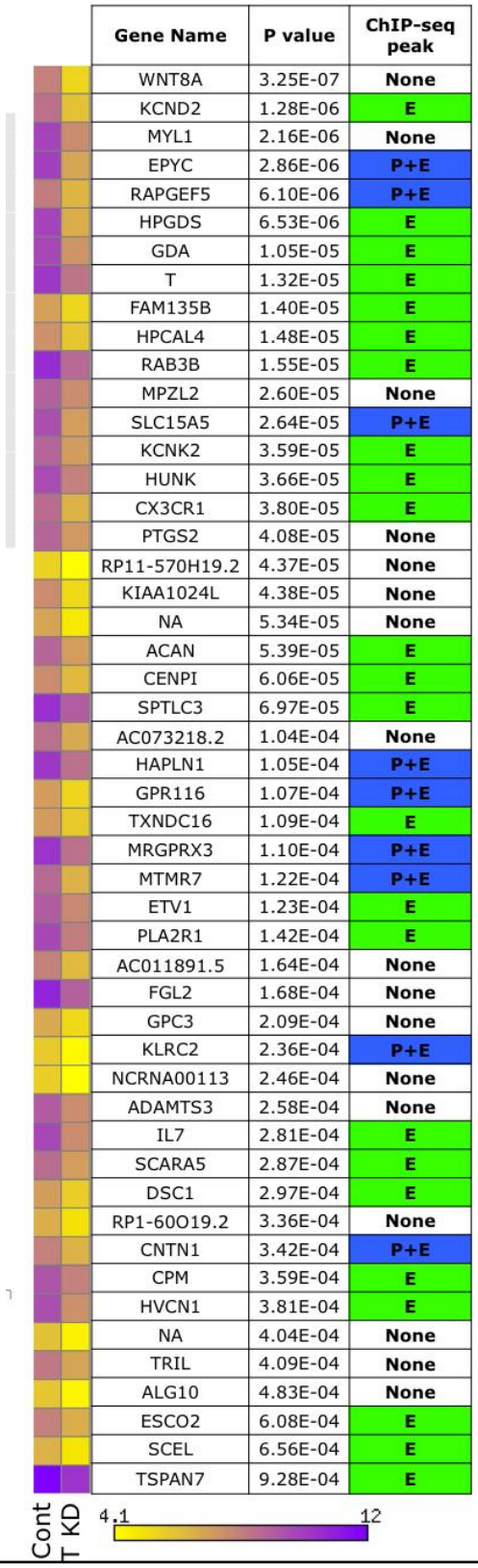
An integration of the ChIP-seq and GEM data was performed to determine whether the T protein regulates the genes it binds to (henceforth referred to as marked genes).

If the identified binding regions were functional, then these marked genes were likely to be transcriptionally active. Conversely, genes without a ChIP-seq peak may be less likely to be expressed. Based on this the differentially expressed genes were categorised (**Figure 4.8**) as:

1. genes with both promoter and enhancer peaks (P+E),
2. genes with enhancer peaks only (E)
3. genes with promoter peaks only (P)
4. genes without peaks.

The distribution of expression levels of these sets of genes (P,E, P+E) in the U-CH1 cell line were then tested to determine if they had a significantly different distribution of expression to those genes without a ChIP-seq peak. The results show that all three categories of marked genes have a significantly higher expression than genes without ChIP-seq peaks (**Figure 4.9A**).This indicates that ChIP-seq peaks are associated with transcriptionally active genes in U-CH1 cells.

If T binding is a key determinant of these expression levels (described above) then it would be expected that these genes are likely to be differentially expressed on T knockdown. The integrative analysis shows that the three categories of marked genes show significantly lower *P* values for differential expression than unmarked genes in the microarray experiment (**Figure 4.9B**), suggesting these are the likely transcriptional targets of T.



Gene Name	P value	ChIP-seq peak
WNT8A	3.25E-07	None
KCND2	1.28E-06	E
MYL1	2.16E-06	None
EPYC	2.86E-06	P+E
RAPGEF5	6.10E-06	P+E
HPGDS	6.53E-06	E
GDA	1.05E-05	E
T	1.32E-05	E
FAM135B	1.40E-05	E
HPCAL4	1.48E-05	E
RAB3B	1.55E-05	E
MPZL2	2.60E-05	None
SLC15A5	2.64E-05	P+E
KCNK2	3.59E-05	E
HUNK	3.66E-05	E
CX3CR1	3.80E-05	E
PTGS2	4.08E-05	None
RP11-570H19.2	4.37E-05	None
KIAA1024L	4.38E-05	None
NA	5.34E-05	None
ACAN	5.39E-05	E
CENPI	6.06E-05	E
SPTLC3	6.97E-05	E
AC073218.2	1.04E-04	None
HAPLN1	1.05E-04	P+E
GPR116	1.07E-04	P+E
TXNDC16	1.09E-04	E
MRGPRX3	1.10E-04	P+E
MTMR7	1.22E-04	P+E
ETV1	1.23E-04	E
PLA2R1	1.42E-04	E
AC011891.5	1.64E-04	None
FGL2	1.68E-04	None
GPC3	2.09E-04	None
KLRC2	2.36E-04	P+E
NCRNA00113	2.46E-04	None
ADAMTS3	2.58E-04	None
IL7	2.81E-04	E
SCARA5	2.87E-04	E
DSC1	2.97E-04	E
RP1-60O19.2	3.36E-04	None
CNTN1	3.42E-04	P+E
CPM	3.59E-04	E
HVCN1	3.81E-04	E
NA	4.04E-04	None
TRIL	4.09E-04	None
ALG10	4.83E-04	None
ESCO2	6.08E-04	E
SCEL	6.56E-04	E
TSPAN7	9.28E-04	E

Figure 4.8 Integration of ChIP-seq and GEM data:

Heatmap of expression for 50 genes downregulated on T knockdown ($P \leq 0.01$, ≥2.5-fold change in expression). For each gene name, representative heatmap, *P* value and ChIP-seq peaks are indicated. (Cont=wild type cells, KD=knockdown)

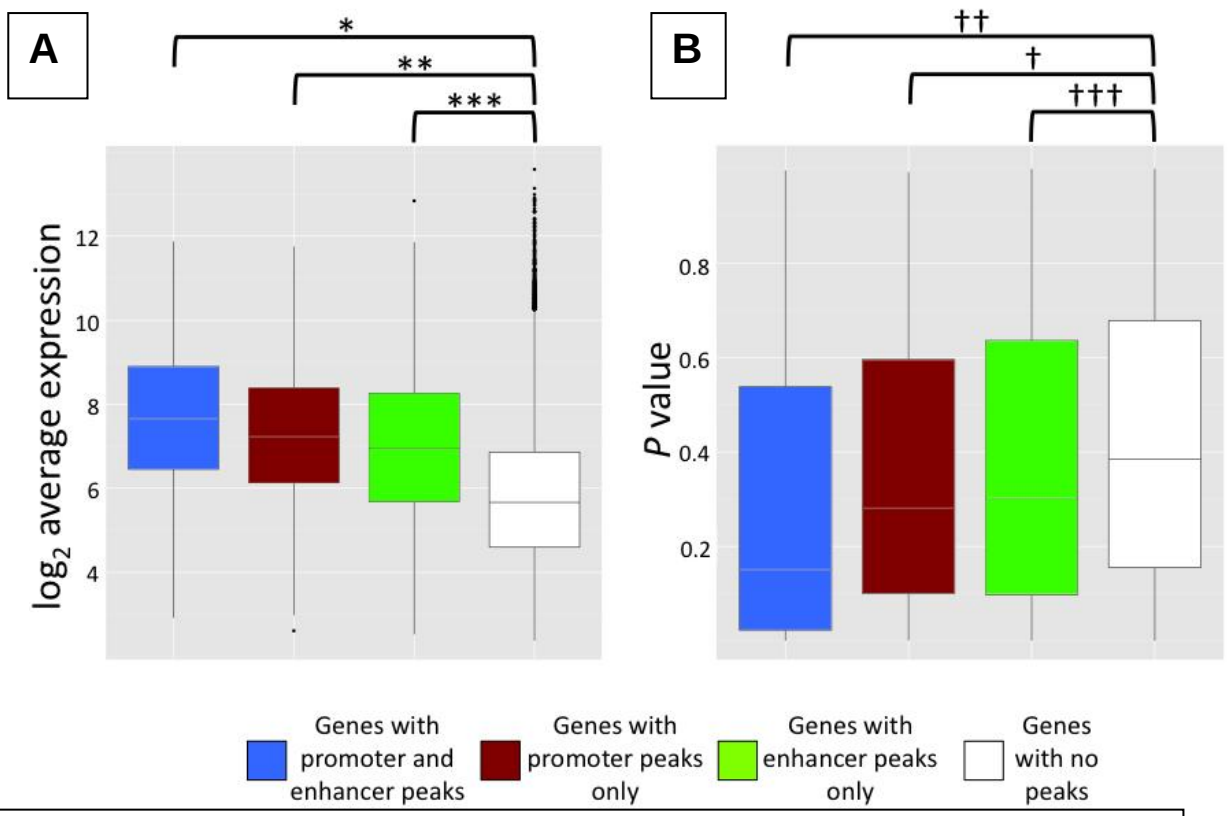


Figure 4.9 Dynamics of T binding (i):

Box plots of A) average expression in U-CH1 cells, and B) P value for differential expression on T KD. Represented is expression of genes with promoter and enhancer peaks, genes with only promoter peaks, genes with only enhancer peaks, and genes with no peaks respectively. * $P = 4.95 \times 10^{-37}$; ** $P = 1.23 \times 10^{-45}$; *** $P = 1.09 \times 10^{-210}$; † $P = 3.47 \times 10^{-5}$; †† $P = 3.18 \times 10^{-9}$; ††† $P = 9.42 \times 10^{-16}$.

The next question to be addressed was whether T positively or negatively regulates its target genes. Specifically, the question as to whether the genes that are significantly up-regulated or down-regulated on T knockdown ($P \leq 0.01$) are more likely to have a ChIP-seq peak than all other genes was investigated. It was found that the down-regulated genes were enriched in each category of marked genes, whereas up-regulated genes were not (**Figure 4.10A**). A complementary analysis using Gene Set Enrichment Analysis (GSEA) indicated a significant association of ChIP-seq peaks with genes down-regulated on T knockdown (**Figure 4.10 B**).

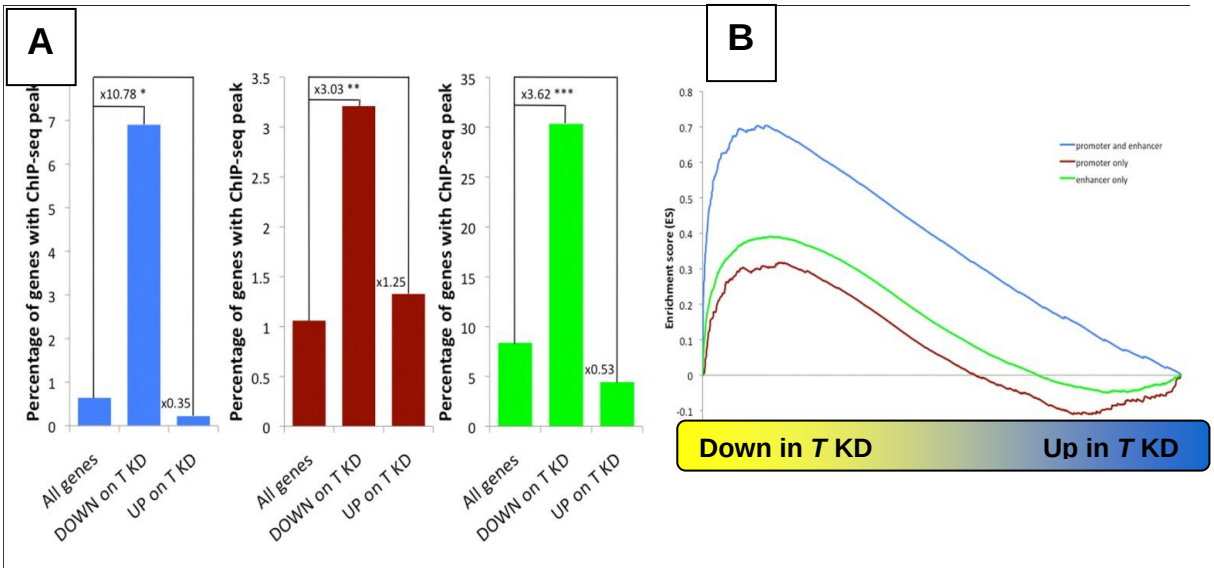


Figure 4.10 Dynamics of T binding (ii)

A. Plots of percentage of all genes and differentially expressed genes ($P \leq 0.01$) with promoter and enhancer peaks, genes with only promoter peaks and genes with only enhancer peaks. * $P = 8.46 \times 10^{-85}$, ** $P = 3.58 \times 10^{-7}$, *** $P = 5.18 \times 10^{-88}$.

B. GSEA enrichment plots for genes with promoter and enhancer peaks (blue), genes with only promoter peaks (red) and genes with only enhancer peaks (green) are shown. All show a significant bias towards genes down-regulated on T KD (FDR q-val < 0.001, FWER $P < 0.001$).

Overall this analysis revealed 257 significantly down-regulated genes ($P \leq 0.01$) with associated ChIP-seq peaks (Appendix). These genes are likely to be the direct transcriptional targets of T, whereas the other 366 differentially expressed genes, not associated with ChIP-seq peaks, represent genes whose expression is controlled indirectly by T. Based on these results, we constructed a gene regulatory network (GRN) for *T* in U-CH1 cells (**Figure 4.11**) using functional genomics data from the Molecular Signature Database tool [197].

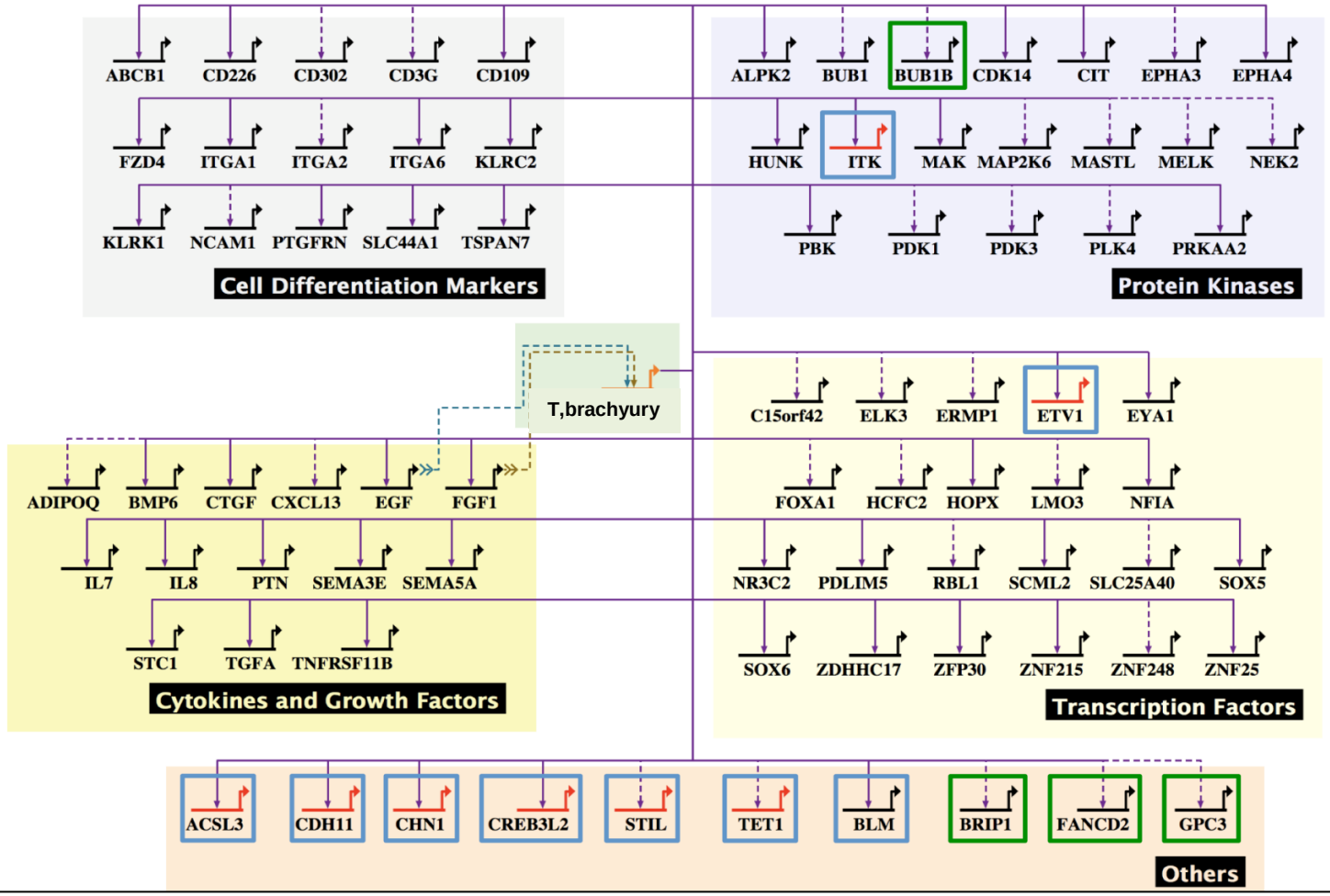


Figure 4.11 Gene regulatory network for T in U-CH1 cells.

T directs a network of transcription factors, protein kinases, cytokines and growth factors, and known cancer-associated genes. Genes are categorised as in the Broad Institute Molecular Signature Database (MSigDB) v3.0. Red gene = oncogene; blue box = translocated in cancer; green box = tumour suppressor; solid lines indicate T binding within promoter and/or enhancer region and significant ($P \leq 0.01$) down-regulation on T knockdown; dotted lines indicate significant ($P \leq 0.01$) down-regulation on T knockdown.

4.4.5 The *T* gene regulatory network in U-CH1 cells reflects transcriptional networks active in primary tumours

To determine if the proposed *T* gene regulatory network in U-CH1 cells was representative of that in other chordoma cell lines and human chordoma samples an *in silico* target prediction software was used. ARACNE is a bioinformatics package that is used to predict transcriptional networks from gene expression microarray datasets [163]. In theory, the integration of its output with the data above would allow stringent filtering for a set of true targets that may be relevant to human tumour samples and correct for cell line specific features (**Table 4.1**).

Table 4.1 Table of techniques utilised to identify set of targets regulated by *T*.

	Cell line (Direct binding)	Cell line(Functional-Transcriptional profile)	Network relationships-in-silico(human samples)
ChIP-seq	+	-	-
GEM profiling	-	+	
ARACNe algorithim	-	-	+ *

* The overlap of results from *in vitro* experiments with human chordoma data is likely to represent targets relevant *in vivo*.

The previously published expression profiles of 8 human primary chordomas, and 2 chordoma cell lines (U-CH2 and K001) was used to predict the T targets [57]. From this 2,660 putative T targets were identified (**Figure 4.12**). Of the direct targets and other changing genes identified from the U-CH1 cell line (see above), 99 direct targets (39%) and 64 other genes (17%) were common to the U-CH1 cell line, the primary tumours and other cell lines. The overlap of the direct targets is greater than would be expected by chance (χ^2, $P < 0.002$). This analysis therefore identified a conservative set of 99 "core" direct targets, that

is, genes with proximal T binding which are down-regulated in U-CH1 cells as a result of T knockdown, which are also identified in primary tumour samples by ARACNE.

To provide additional evidence that the “core” gene set represented *bona fide* direct targets in primary tumours, qRT-PCR assays on a new set of 18 chordomas, all of which are known to express *T* mRNA and protein, were performed. Targets of potential interest, selected on the basis of previous publications [12, 56, 63, 68, 157, 198], were found not only to be expressed in chordoma samples, but unsupervised hierarchical clustering and correlation analysis revealed that higher levels of expression of *T* mRNA were found to be associated with higher mRNA levels of target genes (**Figure 4.13 and 4.14**).

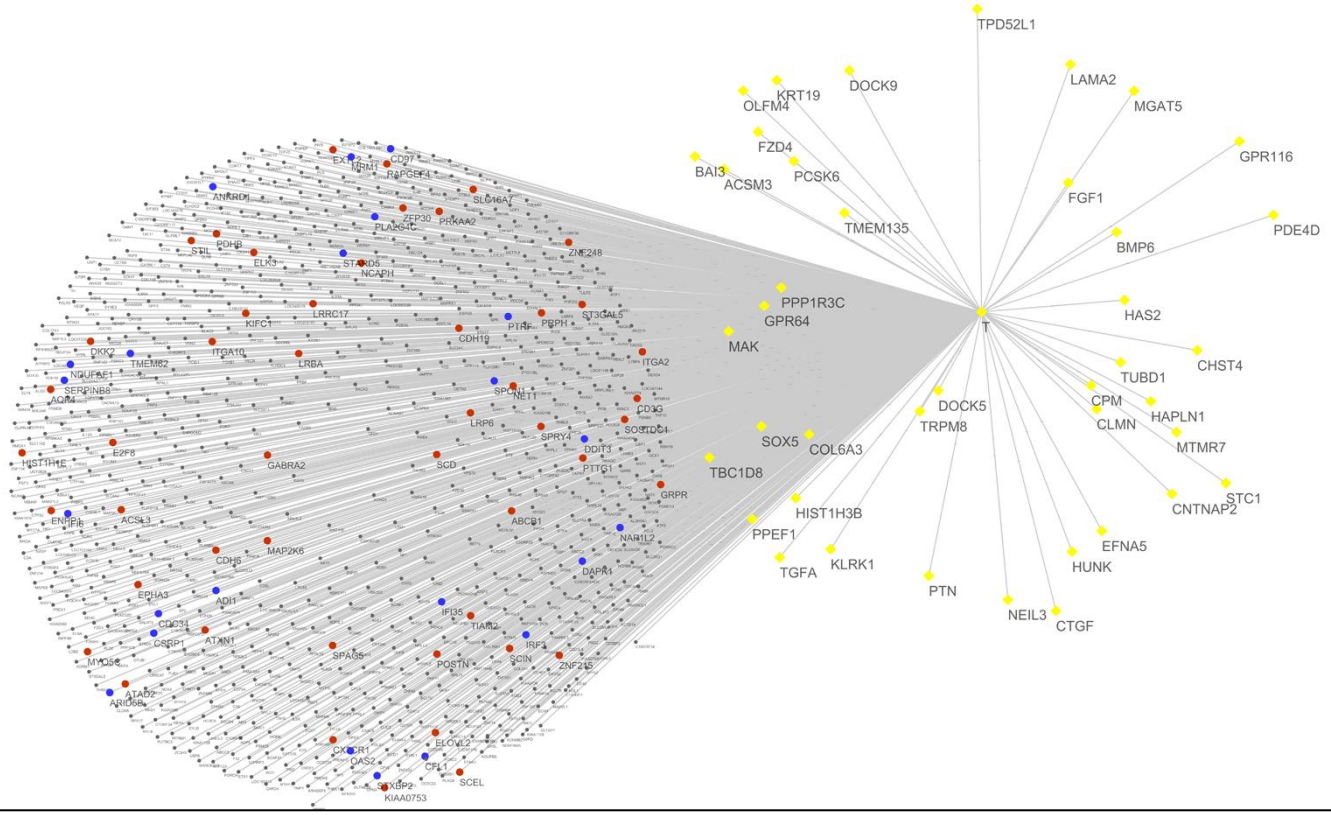
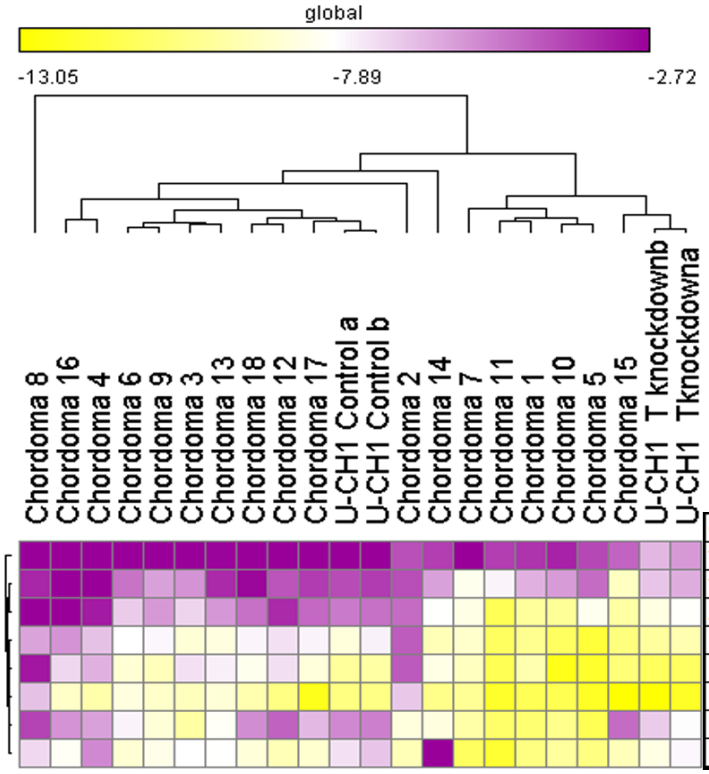


Figure 4.12 Visual representation of ARACNE output

An *in silico* T transcriptional subnetwork was identified from human chordoma samples and chordoma cell lines by using the ARACNE algorithm to identify genes that have $P<1.0E\text{-}^{7}$ based on pairwise mutual information with T. For visual purposes the only the top 1000 predicted target genes are plotted as an undirected network using Cytoscape with T represented as the master regulator. The small grey nodes represent probable false positive calls by the software. The yellow, blue and red nodes are calls from the software that overlap with the data from the *in vitro* experiments. Yellow nodes=direct targets Red nodes=Indirect down-regulated targets Blue nodes = Indirect up-regulated targets.



Gene	*P* Value	FDR(BH)
T	5.45E-04	7.27E-04
TGFA	3.56E-04	5.69E-04
AKR1B10	1.20E-06	9.58E-06
FGF1	8.96E-06	3.58E-05
EGF	5.03E-05	1.34E-04
HOPX	1.08E-04	2.17E-04
ETV1	3.11E-02	3.11E-02
IL8	3.70E-03	4.23E-03

Figure 4.13 qRT-PCR validation of targets in human samples.

Heatmap. q-RT-PCR mRNA expression data (selection of core targets) analysis of eighteen chordoma samples. Unsupervised hierarchical clustering demonstrates two distinct groups of chordomas (High and Low). *P* represents the significant difference in the expression between these two groups (t-test). Statistically higher expression levels of *T* mRNA are associated with higher levels of expression of downstream targets and vice versa. KD=Knockdown, FDR (BH) =False discovery rate (Benjamini-Hochberg method).

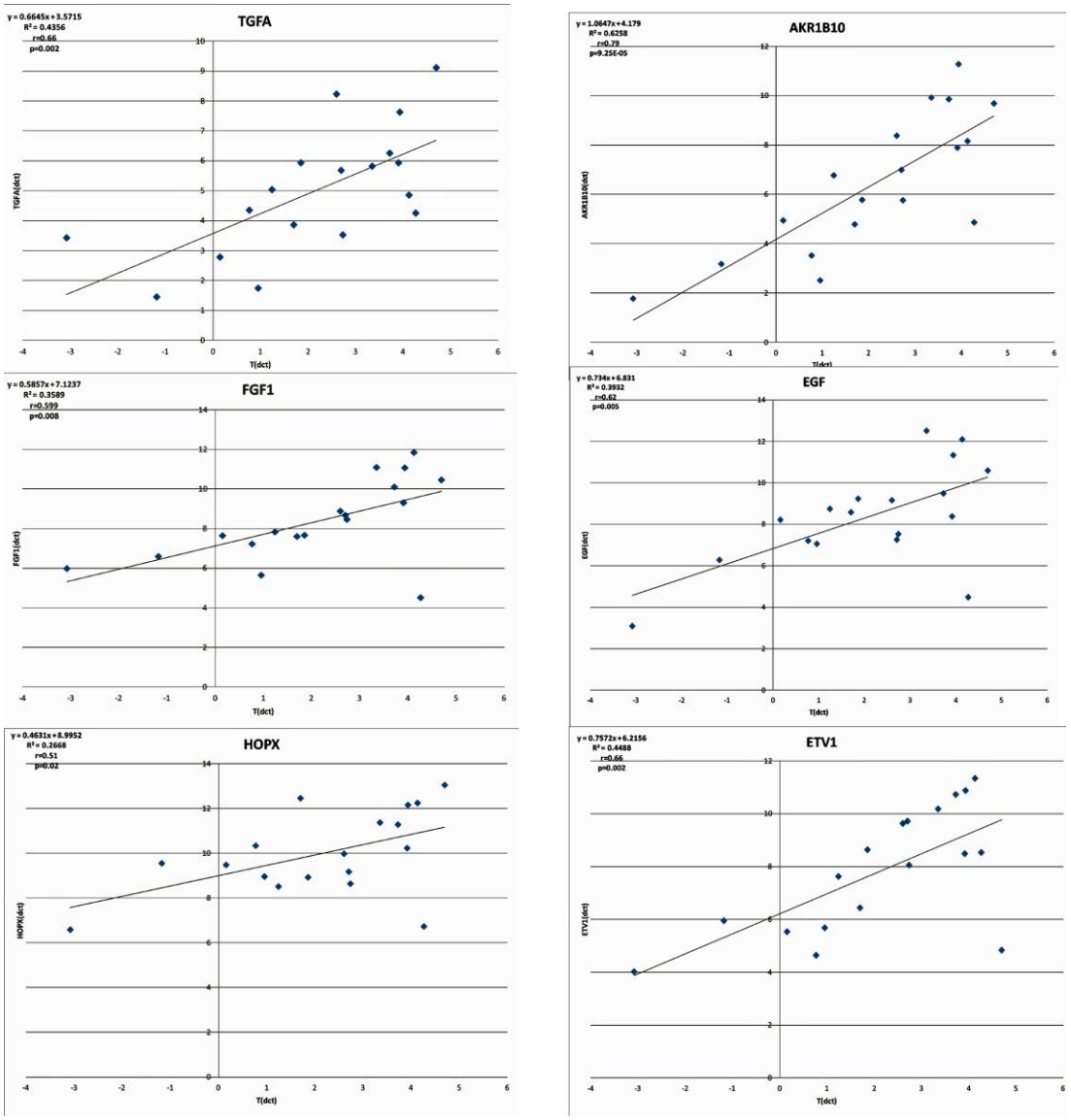


Figure 4.14 Correlation plots of qRT-PCR expression data: There is a positive correlation between *T* mRNA expression and the expression of selected target genes. The strongest correlation is seen with *AKR1B10* (r^2=0.63, *P*=9.25E-^{05}).

4.4.6 *AKR1B10* and chordomas

Aldo-keto reductase family 1 member B10 (*AKR1B10*) is identified as a target in the U-CH1 cell line and is located on a putative chordoma susceptibility region [12]. Also, it has recently been shown to be a key antioxidant response element factor in renal cancer [42]. In view of its documented roles in cancer development and chemo-resistance [199], in addition to its strong co-expression with *T* mRNA in chordomas [56, 57], made it a reasonable candidate to perform further validation in human tumour samples.

Immunohistochemical analysis was performed on a tissue microarray comprising 50 conventional chordomas. This demonstrated unequivocal cytoplasmic expression in all cases (**Figure 4.15**).

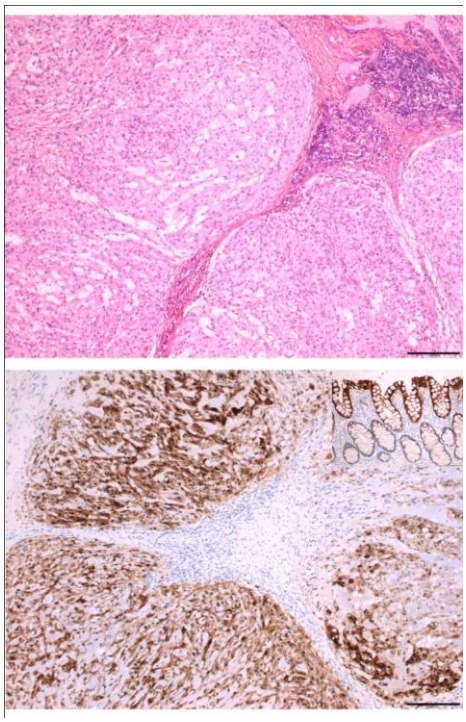


Figure 4.15 AKR1B10 protein expression in chordoma

Light photomicrograph of a representative case: Haematoxylin and eosin-stained section of a conventional chordoma (top panel) and a chordoma showing immunoreactivity for AKR1B10 (bottom panel). Scalebar = 200μm. Inset (positive control, normal colonic tissue).

4.4.7 The WNT pathway and chordomas

One of the genes significantly differentially expressed upon T knockdown is *WNT8A* (mRNA). However this was not identified as a direct target as there was no evidence of T binding after integration with ChIP-seq and *in silico* data. Nevertheless, in view of the known autoregulatory feedback loop present between (MOUSE) T and the WNT signalling pathway during development [139] and in order to ensure that this was not a false negative finding, the expression of WNT8A was investigated. This was accomplished by analysing WNT8A mRNA expression using qRT-PCR and the protein expression of its downstream effector β-catenin using immunohistochemistry in human chordoma samples.

qRT-PCR assays of 10 human chordoma samples demonstrated no expression of WNT8A. This was confirmed using appropriate positive and negative controls and melt curve analysis (**Figure 4.16**). Furthermore, there was no nuclear β-catenin protein expression using immunohistochemistry on a tissue microarray of 50 chordomas (**Figure 4.17**).

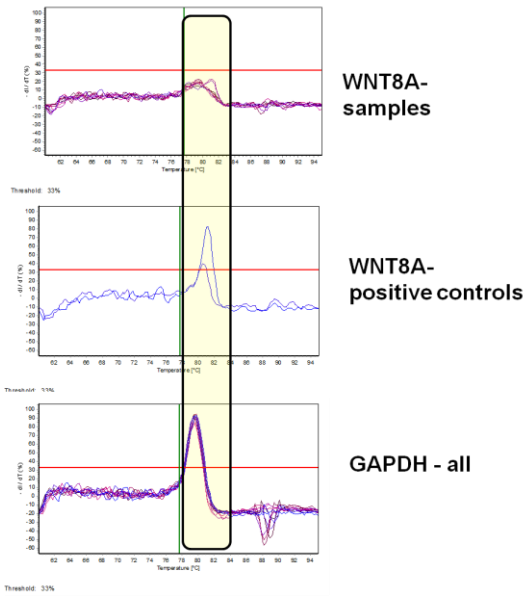


Figure 4.16 *WNT8A* mRNA expression in chordoma.

Melting curve analysis of RT-PCR *WNT8A* products compared to *GAPDH*. Top panel: All chordoma samples demonstrate no amplification of *WNT8A* product. Middle panel: Positive control shows specific products in the U-CH1 and SW480 cell lines. Bottom panel: All samples and controls show specific amplification of housekeeping gene *GAPDH*.

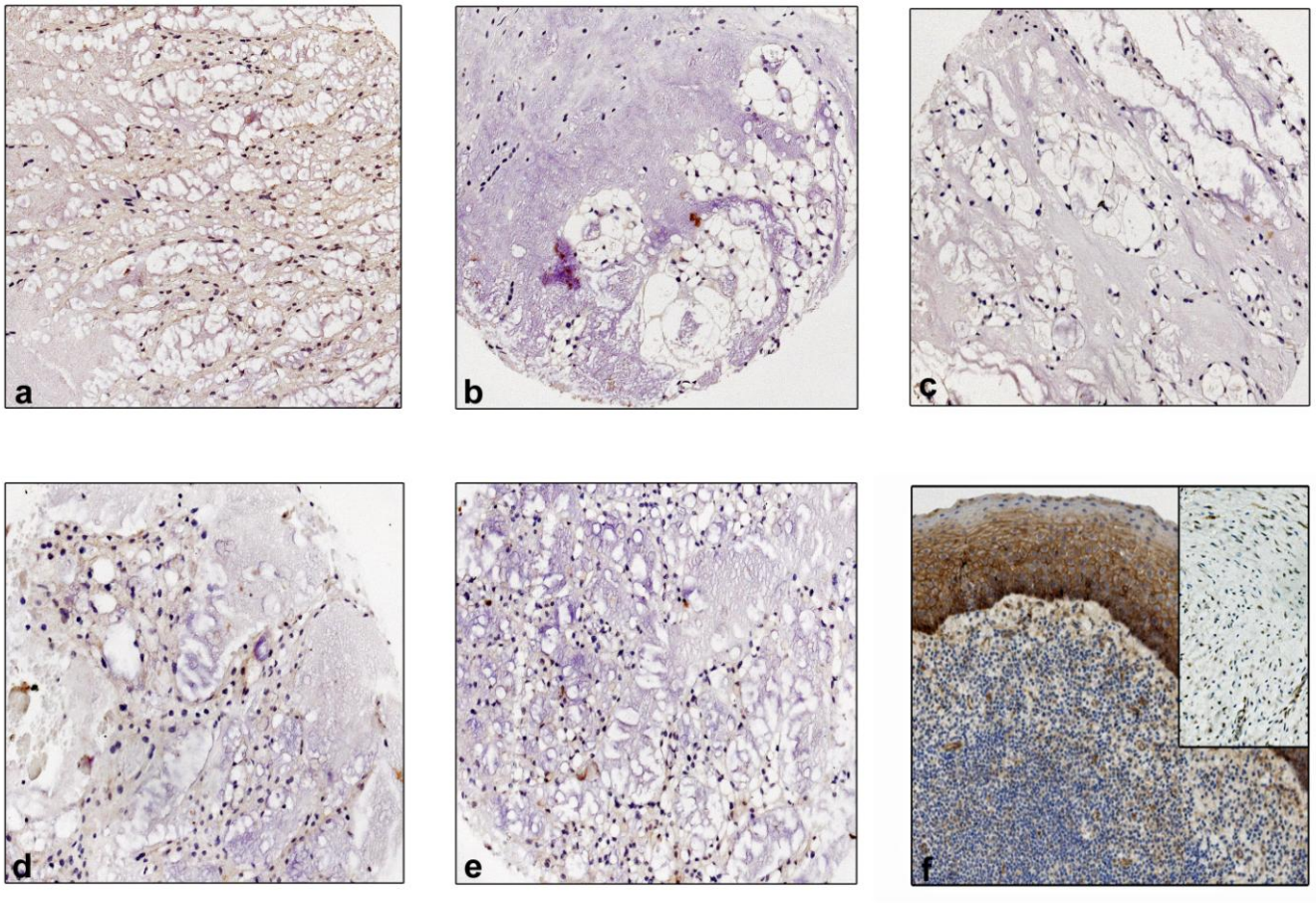


Figure 4.17 β-catenin expression in chordoma

a-e: Light photomicrographs of five representative chordomas which show no immunohistochemical expression of β-catenin. f: Positive control – squamous epithelium. Inset - nuclear expression in fibromatosis.

This core set of genes derived from the integration of data from cell lines and primary disease allowed the generation of a putative GRN of T target genes in chordoma samples (**Figure 4.17**). Gene ontology analysis of the core targets using the Broad's Molecular Signature database [197] revealed enrichment for gene sets involved in regulation of the cell cycle. Other enriched gene sets of pathophysiological interest included those involved in the production of extracellular matrix and growth factor activity. These results are similar to those obtained from the gene expression microarray experiment generated from the U-CH1 cell line in which T was silenced. Hence, the reproducibility of the data provides another level of confirmation that the function of the identified core set of target genes accurately reflect the global changes seen when T is knocked-down.

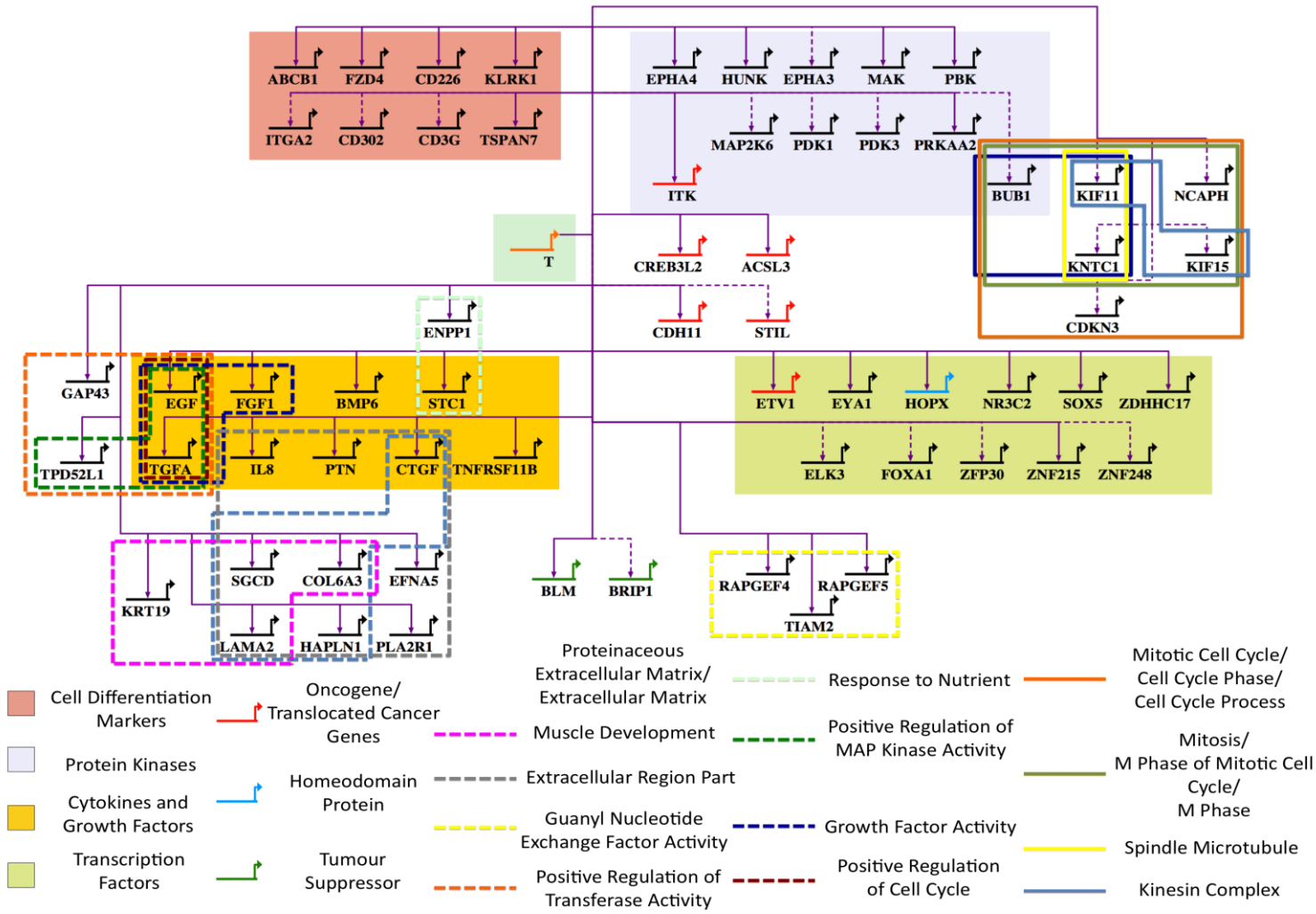


Figure 4.18 Gene regulatory network for T in chordoma

Genes identified by both U-CH1 study and ARACNE analysis of primary tumour data. Genes are categorised by terms enriched in GSEA (FDR < 0.05). Coloured background and coloured genes represent Broad Institute Molecular Signatures Database (MSigDB) v3.0 categories; genes in outlined boxes are categorised by Gene Ontology (GO) term. Solid lines connecting genes indicate T protein binding within promoter and/or enhancer regions and significant down-regulation on knockdown of T ($P \leq 0.01$) in U-CH1 cells; dotted lines indicate significant down-regulation on knockdown of T ($P \leq 0.01$). All genes were also identified in ARACNE analysis of primary tumour data.

4.5 Discussion

4.5.1 The binding repertoire of T

By using ChIP-seq, the binding dynamics of the T protein in humans has been identified for the first time. The finding that T binds in close proximity to genes and to promoter and enhancer regions confirms what is known about its transcriptional activation role during development of other organisms [98, 99]. Furthermore, the finding of a canonical T-box motif in the T bound DNA sequences in the human cell line, indicates that that these are *bona fide* targets.

4.5.2 *T* and the ETS family of transcription factors

It is well established that T-box genes regulate gene expression of targets co-operatively with other transcription factors [97, 195, 196]. It is therefore of interest that an ETS binding site is associated with the canonical T-box in U-CH1 suggesting that T may regulate expression with an ETS family co-factor in chordomas. This is not without precedent as ETS factors have been associated with T-box member families. In mice, a Pea3 binding site was identified within the (MOUSE) *T* promoter [200]. Pea3 belongs to the ETS subfamily that includes Etv1, Etv4 and Etv5. Although the ETS members demonstrate relative sequence specificity for binding, this profile is interchangeable amongst members of the family. ETV5 has been shown to directly regulate T expression in human spermatagonial stem cells [138] and Pea3 is co-expressed with Ntla during zebrafish mesoderm development [88]. Recently it also been shown that Etv1 associates with T-pit (Tbx19) in a combinatorial manner to regulate gene expression in the pituitary gland [201] of zebrafish.

From the target identification and gene expression correlation analysis described above (**Figure 4.14 and 4.18**), the gene expression of *ETV1* is regulated by T in chordomas. This leads to an intriguing hypothesis that the downstream transcriptional program regulated by T requires ETV1 expression. Furthermore, there exists a possibility at least on the basis of the literature cited above that a positive feedback loop might also exist between these two transcription factors. This is an avenue for future research into the co-operative transcriptional dynamics of T expression in chordomas.

4.5.3 *T* and cell proliferation

During development, T-box genes, including *T*, have recognised roles in cell fate decisions, tissue specification and differentiation [81]. Specifically, (MOUSE) *T* is indispensible for development as shown by the embryonic lethal murine $T^{-/-}$ mutant, and in ascidians it is involved in the regulation of cell division of notochordal cells [82, 84, 202].

In the context of cancer, melanomas and breast carcinoma demonstrate TBX2 and TBX3 overexpression, which act as transcriptional repressors of key senescent, and proliferation checkpoints, including CDKN2A [103, 203]. Conversely, *TBX5* is a tumour suppressor and loss of its function in colorectal cancer results in cell proliferation [104]. In complementary experiments to those described in Chapters 3 and 4, it has been shown in pancreatic cancer cells that induction of T expression enhances cell proliferation, migration and invasion [129]. There is therefore good evidence that T-box genes have a role in regulating cell proliferation in some malignancies. However, the downstream transcriptional program and target genes regulated by these T-Box genes, and the mechanism through which they mediate their effects are not known.

The shRNA experiments show that growth arrest is induced by silencing T. It is therefore of interest that a set of genes, including *EGF, TGFA* and *NUSAP1*, are identified as targets of T as these are known to regulate cell cycle progression (**Figure 4.17**). *NUSAP1* is of particular interest being a spindle checkpoint gene, which is exclusively expressed in proliferating cells and is a requirement for mitosis [204]. Little is known about its upstream regulation although c-myc, lin9 and nf-y have previously been identified as transcriptional regulators [205-207]. NUSAP1-knockout mice are embryonic lethal highlighting its critical role in embryogenesis [208]. In cancer, *NUSAP1* mRNA expression has been associated with aggressive clinical behaviour in melanomas, glioblastomas and breast cancers, and is associated with chemotherapy resistance [209-211]. It is also noteworthy that *NUSAP1* is a target of NKX2.1 protein, and a member of a cluster of mitotic spindle checkpoint genes, along with *BUB1, CDKN3, KIF15, KIF11*[212], which have been reported as representing a specific gene signature for a “proliferative” subtype of T-acute

lymphoblastic leukaemia, and that these same genes are also down-regulated on silencing of T in the U-CH1 cell line (**Figure 4.17**).

BUB1 another target gene regulated by T is a protein kinase and is overexpressed in many cancers, particularly melanoma [209, 213] although the mechanism by which it mediates its effect is not well described. *BUB1* is a critical co-regulator of the spindle assembly checkpoint (ref), and it is also a direct target of TBX2 protein, together with a cohort of other cell cycle genes [214]. Furthermore, it is associated with spontaneous tumourigenesis through defects in chromosomal segregation in murine models [215].

These data argue that T mediates its effect through regulating the mitotic spindle checkpoint genes.

4.5.4 Regulation of chemokines and growth factors and the dynamic interplay with extracellular matrix

It has previously been reported that a number of growth factor receptors including EGFR, FGFR are activated in chordomas, and this has not been adequately explained by the presence of mutations or structural variation [63, 68, 216]. It is therefore interesting to find that T appears to orchestrate the expression of multiple growth factors and cytokines including CTGF, PTN, TGFA, EGF, FGF1, and BMP6 (**Figure 4.17**). Hence, the findings argue that activation of their respective pathways is ligand-dependent.

ECM genes represent another major group of genes that are targeted by T in chordoma (**Figure 4.17**). This finding is consistent with the reduction of ECM production, and the associated disorganisation of the notochord, neural tube and somites, in (MOUSE) T-/- embryos, and Ntla's role in the development of structural components of the notochordal sheath matrix in zebrafish [82, 116]. It is recognised that critical interactions between the ECM and components of the cell plasma membrane (adhesion molecules and receptor tyrosine kinases) determine the downstream transcriptional response of various cytokines through their receptors. *EGFR* is of particular interest as its protein product forms a physical complex with ITGA2B1 [217], an integrin, which not only determines specificity of EGF signalling but also controls cell proliferation,

survival and migration. It is therefore noteworthy that T indirectly regulates expression of selected integrins (ITGA2, ITGA10), ECM, and adhesion proteins (LAMA2, HAPLN1, OLFM4, PKP2, COL6A3, CNTNAP2, CHST4, MMP16, ADAMTS3). This data is supported by the results of the adhesion assay in Chapter 3. Furthermore, connective tissue growth regulating factor (*CTGF*), which is required for ECM formation in development, and is known to play a pivotal role in regulating the functions of integrins, signalling and motility in the notochord has been identified as a direct T target in this study. This adds weight to the speculative relationship, posited by others, between the functions of T and CTGF in notochordal cells [216, 218]. Overall these results provide evidence that T is the master regulator of a complex signalling network involved in controlling the growth of chordoma through it downstream transcriptional programme.

4.5.5 The role of WNT8A signalling in chordomas

WNT8A and its cognate receptor *FZD4* were amongst the most down-regulated genes in the T knockdown U-CH1 cell line. This is interesting from an embryological perspective as *T* is a target gene of the WNT/β- catenin axis during mouse development [139, 219]. In developing zebrafish mesoderm, it has also been shown that Wnt8A and Ntla/Bra co-operate in a positive feedback loop during somite formation, while in the notochord this loop is regulated by Fgf signalling [88]. Although, the absence of β-catenin and *WNT8A* mRNA expression in chordoma samples argues against the role of active WNT/β-catenin signalling in these tumours, there is a possibility that a small subpopulation of chordoma cells, a stem cell niche, has not been identified due to tumour sampling, and are not present at detectable levels using bulk processing methods for RNA. This is raised because in the context of colorectal cancer in which there is constitutive WNT signalling, there is very focal and discretely localised β-catenin expression [131]. Moreover, there is a recent report of a niche population at the invasive tumour front of colorectal cancers which have stem cell like characteristics conferred by T and mediated through β-catenin [106].

4.5.6 Therapeutic implications of T targets in chordoma

The integrative functional approach described in this chapter has identified genes and pathways not previously associated with T. The genomic analyses presented here provide insight into the molecular events involved in the disease, and have important implications for stratified medicine. The finding that T targets many growth factor ligands is consistent with the known activation of multiple tyrosine kinase receptors in chordoma. It may also account for the limited value of single targeted therapies, for example antagonists to EGFR, and ckit/PDGFR, observed in treating patients with this disease [28, 32, 220]. Given this background, and that chordoma is a rare cancer with few therapeutic options [221], the targeting of T directly using RNA interference-mediated gene therapy, or other methodologies [60], has a biologically justifiable rationale.

Conclusion

The identification of functional transcription binding sites is a requirement for understanding the complex regulatory dynamics, and networks that govern the behaviour and phenotype of the cancer cell. In view of T's putative oncogenic role it is therefore critical to identify its downstream targets if its function is to be elucidated. By integrating the global transcriptome data (GEM), from a *bona fide* chordoma cell line, in which T was silenced, with genome-wide T protein-DNA interactions (ChIP-Seq) generated from the same cell line, the chief transcriptional targets of T have been identified in for the first time in humans.

Chapter 5

A common non-synonymous single nucleotide variant in *T* is strongly associated with chordoma

In this chapter I introduce topical concepts regarding the study of genetic variation in cancer. This is followed by the results of a genetic association study which shows that a coding SNP in *T* is strongly associated with the development of chordoma.

5.1 Introduction

5.1.1 Genetic variation and disease

Genetic variation is a major source of phenotypic and evolutionary diversity [222]. This diversity is achieved by the complex interactions between multiple genetic interacting partners (“epistasis”), epigenetics and the overall interaction of these genetic loci with the environment [223, 224]. Genetic variation can be inherited and in many cases plays a role in contributing to the risk for developing complex diseases such as cancer. Single nucleotide polymorphisms (SNPs) and copy number variations (CNVs) are two of the more common form of genetic variation that have been widely investigated (**Table 5.1**) and the extent of genetic variation in the human genome across multiple population groups is being catalogued through large international collaborations such as the HapMap and 1000Genomes projects [225, 226].

A single nucleotide polymorphism is a category of genetic variation that is defined by the allele frequency of a genetic mutation in the population under study. Variations that occur in ≥1% of the population are regarded as common polymorphisms whilst those variants that have a minor allele frequency of ≥0.1% but <1% are regarded as rare.

Genetic association studies are conducted to establish statistical associations between ≥1 genetic polymorphisms and phenotypes or disease traits. The ability to screen for common single nucleotide polymorphisms that influence the risk of developing a disease is possible without making any prior assumptions of the nature of the nucleotide variants involved. This can be performed on a

genome wide scale [Genome Wide Association Studies (GWAS)] using oligonucleotide arrays (SNP arrays) that contain a large number of probes for common SNPs that essentially provide for genome wide coverage of nucleotide variation. This approach takes advantage of the linkage disequilibrium structure of the human genome and the association signals generated by the marker SNPs in principle then allow for the interrogation of nearby genes for variants that may have an association with the disease under investigation. The challenges in these types of studies are many and require large numbers of samples with replication in order to obtain unequivocal statistical significance [227]. Moreover confounding factors such as population substructure and statistical issues such as multiple comparison testing have to be taken into account to obtain reliable results. In the large number of studies performed on a variety of common diseases, GWAS studies have demonstrated that the majority of common variants are associated with odds ratios of 1.2-1.6 [228].

The small effect sizes imparted by these signals and the proportion of heritability conferred by them is thought to be modest [229]. This suggests that common diseases are more strongly influenced by the combination of many common variants with small effects on risk. There is also some evidence to suggest that there may be a contribution of rare variants to this “missing heritability”. This rare variant hypothesis posits that a large proportion of inherited susceptibility to relatively common diseases may be due to the combined effects of low frequency dominantly and independently acting variant alleles of a variety of different genes [230, 231]. Unlike in common variants where population association, case control studies using GWAS is necessary for their detection, rare variant discovery strategies would involve DNA

sequencing. This can now be accomplished on a genome wide scale using second generation sequencing technologies such as whole genome and whole exome sequencing. Some of discovery strategies proposed to elucidate these "missing heritability" variants include the sequencing of affected individuals afflicted with common diseases in families and sequencing of individuals who demonstrate extreme phenotypic/disease traits [232].

Table 5.1 Categories of human genetic variation.

	Size of variant	Category of variation
Sequence variation	**Single base pair**	SNP
		Indels
Structural variation	**2bp-1kb**	Microsatellites
		Minisatellites
		Indels
		Inversions
	>1kb	CNV
		Duplication events
		Inversions
		Translocations
	Microscopic and submicroscopic	Segmental aneusomy
		Chromosomal deletions
		Chromosomal insertions
		Translocations
	Chromosome to whole genome	Interchromosomal translocations
		Ring chromosomes
		Aneuploidy
		Aneusomy

Common types of genetic variation can be categorised into two major groups- single base changes and those that alter more than one base (i.e. structural variants).

5.1.2 Single nucleotide polymorphisms

By far, SNPs are the most common form of variation with one occurring at least every 0.3-1kb of the genome, totalling 5-10 million SNPs across the genome. The regions in which these SNPs occur are inherited together as haplotype blocks. Therefore one SNP can serve as a proxy for another on the same haplotype block, a phenomenon known as linkage disequilibrium. The convenience, accessibility and relative economical cost of interrogating the entire genome using this technology has led to an escalation in research on inherited genetic variation of disease.

5.1.3 Copy number variation

CNVs are variable regions of the genome larger than 1kb in size that involve gains or losses of DNA which can be microscopic or sub-microscopic. They are relatively common and are a considerable source of variation that may be responsible for disease [233]. Duplications or deletions of various cancer genes have been associated with increased cancer risk [234]. In the context of chordoma, duplication of *T* has been shown to increase susceptibility to the familial form of the disease [14]. It has been shown that most common CNVs are in linkage disequilibrium with SNPs on the same haplotype block suggesting that their signal can be determined in GWAS studies [235, 236] using high density SNP arrays. The architecture of the CNVs can then be investigated using a variety of modalities for example, arrayCGH, qPCR, FISH and more recently algorithms have been developed to use whole exome and whole genome sequences to identify and characterise CNVs [237].

5.1.4 Indel polymorphisms

Indel polymorphisms are regions of small insertions or deletions of DNA (<1kb). They frequently occur at repetitive DNA hotspots, such as variable number tandem repeats (VNTRs) also known as micro or minisatellite regions and in transposable genomic elements. These are complex genomic regions that "expand and contract" as a result of insertion or deletion of the repetitive DNA sequences and are prone to slippage and consequently errors in repair [238]. Even with the current advances in sequencing technology, the repetitive nature of these regions means that they are not completely mapped and are largely responsible for the incomplete annotation of mammalian genomes. Because of this, variation in these regions is probably the least well catalogued.

5.1.5 The functional effects of germline DNA variation on disease.

Germline variants that occur within the coding region of a gene more readily explain their impact on gene function than other variants. These may be deleterious with alteration of protein coding and hence have an effect on function. A good example of this is the *PALB2* gene which is associated with the risk of developing pancreatic cancer. PALB2 is the localising partner of BRCA2. It is involved in the DNA repair pathway (double strand break repair) and has been associated with an increased risk of breast cancer and Fanconi anaemia [239-241]. In a whole exome sequencing study of familial pancreatic cancer it was discovered that germline protein truncating mutations of *PALB2* were responsible for the significant risk of developing this disease in these families [242].

However in common diseases, most strong association signals are not identified near genes making their impact on function and therefore on disease less clear. Their roles in contributing to disease are being elucidated by the integration of the genetic variation datasets with functional regulatory genomic outputs such as that seen from the ENCODE project. This is exemplified by a recent study on breast cancer in which the majority of risk associated SNPs were not found in genes. These researchers have shown that these SNPS are distributed in enhancer regions of the genome that influence the DNA binding affinity of a transcription factor FOXA1, a key regulator in oestrogen (ER) responsive breast cancer. This results in enhanced transcriptional activity of genes specific to the ER responsive pathway in these tumours [243]. Another mechanistic insight into the role of common susceptibility determinants in cancer was shown in renal carcinoma. Here it was demonstrated that an intergenic susceptibility SNP modulated the binding and function of the HIF transcription factor in an enhancer region of the *CCND1* gene [244]. *CCND1* is an oncogenic cell cycle regulator that is known to be up-regulated in these tumours [245]. These two studies provide functional insights into the role of intergenic common single nucleotide variations in the development of common cancers and it is likely that further studies of this kind will be necessary to unlock the association signals of common variants [246]. Copy number variations also contribute to disease risk [247], and the limited examples in which strong risk has been found involve inflammatory diseases [248, 249], which may be explained by the fact that copy number variable regions appear to be enriched in genes involved in immune response [250]. In cancer, rare CNV susceptibility alleles have been identified in well-known cancer-predisposing

genes, including the *BRCA1* and *BRCA2* genes, *VHL* and *APC* genes [251-253]. The mechanisms underlying their functional effects involve large deletions resulting in loss of function mutations. Another novel mechanism imparted by a germline CNV was noted in a family with mismatch repair gene defects. A micro deletion of the 3' end of *EPCAM* was identified. This disrupts the termination polyadenylation signal of *EPCAM* resulting in read-through into the *MSH2* gene which was hypermethylated both in normal and cancer tissues. This represented a novel type of germline copy number mutation that led to mono-allelic inactivation of the *MSH2* gene resulting in a Lynch phenotype. In chordoma, germline duplication of the *T* gene has been reported as a risk factor for the familial form of the disease [14]. Presumably the functional effect of the duplication involves an increase in dosage of the gene product; however this has not yet been reported.

5.1.6 Genetic association studies and cancer

The promise that genome-wide susceptibility studies into complex diseases believed to have a genetic basis may accurately predict risk is underpinned by variation in gene structure and single nucleotide polymorphisms. The evidence that this genetic variation may play a role in cancer comes from the familial clustering of some tumours. Moreover, the higher frequency of cancers seen in monozygotic twins compared to dizygotic twins or siblings argues that the aggregation in families is strongly influenced by genetic variation rather than environmental or lifestyle factors [254]. The well characterised cancer genetic associations such as mutations in the *APC* gene in colorectal cancer and *BRCA1/2* genes in breast cancer have a Mendelian pattern of inheritance [255, 256]. In general though, most common cancers follow a polygenic model of

inheritance where multiple risk alleles of low to moderate effect combine to influence disease risk. GWAS studies have revealed multiple susceptibility regions in some cancers whilst in others which have a strong environmental exposure component, much fewer genetic associations have been established. As expected, on the whole, the risk conferred by these common variants in cancer has been modest with allelic odds ratio <1.5 [257, 258]. Notable exceptions include testicular germ cell tumour and thyroid cancer [259, 260]. Nevertheless, even the highly penetrant susceptibility genes such as *BRCA1/2* only confer about 20% of the genetic risk for breast cancer in these families. There is therefore a line of reasoning that proposes that rare genetic variants may account for this missing heritability [261].

Because of these small but indisputable risks there has been a call in this post-GWAS era for a new paradigm for the functional characterisation of these genetic predisposition loci in cancer [246]. The association signals associated with these risk loci are strong but they still have not revealed variants that could easily explain the signal in most cases. Although the majority of these variants do not alter the amino-acid sequence they may influence gene function, and this is likely to be accounted for by their being situated in regulatory genomic regions as discussed above. In order to derive some biological understanding and clinical benefit from these studies there is a need to associate the variants with functional characteristics such as gene expression or pathway mapping. There is a rationale that the identification of true causal SNPs and epistatic interactions will require a more focussed approach possibly using prior biological knowledge, functional genomics and targeted re-sequencing of disease susceptibility loci with next generation sequencing technology [262].

5.1.7 Genetic variation and chordomas

Genome-wide linkage studies in familial chordoma patients have identified the 7q33 locus as a susceptibility determinant in these families [12, 13]. The putative “chordoma gene” had not been discovered by re-sequencing of this region using more conventional techniques as there were no mutations nor loss of heterozygosity events identified. The strongest candidate genes included growth factors (for example *PTN*), transmembrane receptors (*EFNA1, EFNAB6*, and *PLXNA4*), signalling factors (*BRAF*, and *ARHGEF5*), transcription factors (*TIF1, CNOT1, CREB3L2,* and *EZH2*) and apoptotic pathway genes (*CASP 2*). Less likely candidates were thought to be the family of aldo-ketoreductases (e.g. *AKR1B10*) and cell adhesion molecules (for example C-type lectin superfamily member 5). Subsequently, a member of one of these chordoma families presented with a clival chordoma without any evidence of linkage to 7q33. A more refined search was conducted using a new SNP linkage platform. This narrowed down the putative susceptibility region to chromosome 6q27, the *T* locus [14]. Conventional sequencing did not reveal any mutations; however high resolution arrayCGH detected a tandem duplication of the *T* gene in the germline of this patient and this was confirmed by qPCR. Wider screening of 7 chordoma families revealed that multiple members of 3 of these families harbour *T* duplication. The susceptibility region in the other 4 families is still unknown.

5.1.8 Copy number variation and chordoma

T is ~10kb in length. In the seminal familial chordoma *T* duplication study the region harbouring the duplicated copy of *T* ranged from 52kb to 489kb. Current "off the shelf" arrayCGH chips do not have sufficient coverage of this region to detect a tandem duplication thus requiring a high resolution arrayCGH approach.

Somatic copy number gains of *T* have been identified using various modalities including arrayCGH, qPCR and FISH. In two studies, using FISH, a minor allelic gain (ratio of >3:1 of *T*: centromeric probe) of *T* was identified in 3 of 39 (7.6%) and 8 of 81 (9.9%) cases respectively [17, 63]. A similar finding using qPCR showed that 2 of 16 (12.5%) sporadic cases harboured minor gains of *T*. Amplification of the *T* locus has been reported in ~7% of chordomas using a combination of qPCR and FISH [17]. Polysomy of chromosome 6 has also been identified in 39% of chordomas.

5.1.9 Variable number tandem repeats and *T*

VNTR polymorphisms play an important role in human genetic variation. They can influence gene expression by serving as transcription factor binding sites in promoter or enhancer regions [263] or by regulating the production of short non-coding RNAs [264]. When located within coding regions they alter the structure, production and function of the protein produced [265]. They increase chromosome fragility and are associated with pathological conditions such as the triplet repeat expansion neurodegenerative disorders [263].

There are two VNTRs present in *T,* in introns 1 and 6. The VNTR located in intron 7 is in a putative regulatory region (ENCODE) and lies upstream of a known variably spliced exon 7 (**Figure 5.1**). This region exhibits a length polymorphism and is composed of five near perfect 42bp repeat sequences. The annotated deletion polymorphism is composed of 2 of the 42bp repeats and constitutes either deletion of the first 3 repeats or the last 3 repeat sequences. Although these VNTR polymorphisms have been assigned identification numbers in dbSNP (rs67234608 and rs71658389) there is no frequency data regarding its prevalence in human population groups. As there is minimal degeneracy between the repeats which appear to be compensated for by SNPs, the sequences for rs67234608 and rs7165389 are identical.

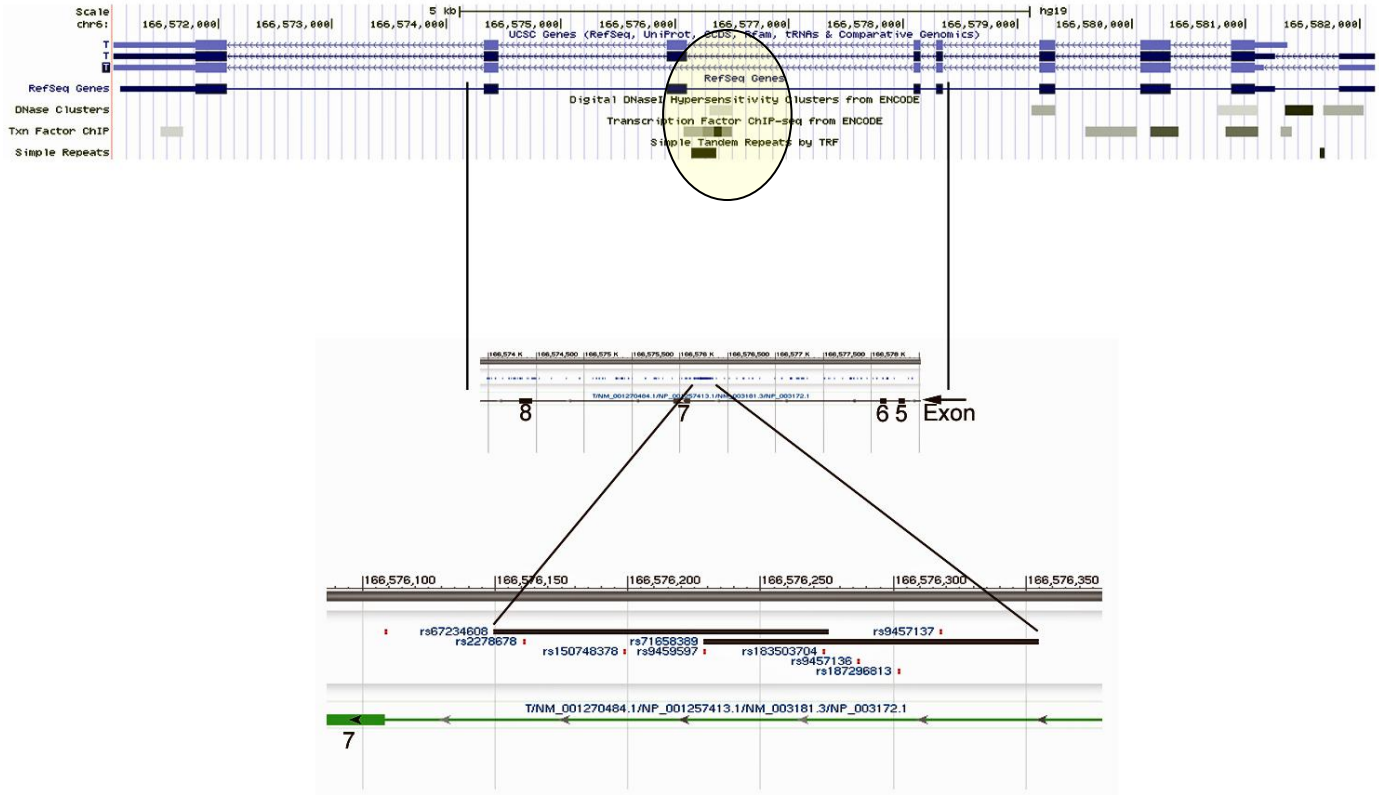


Figure 5.1 Genomic structure of the VNTR in intron 6. The repetitive region is in a known transcription factor binding site upstream of a variably spliced exon 7 (Denoted within the area of the ellipse). The reference mRNA transcript splice variants, functional domains, and repetitive regions are represented in the upper panel. The lower panel is an enlarged schematic diagram of the VNTR locus and its annotation from dbSNP 135.

Table 5.2 DNA sequence of the VNTR locus

Genomic Position	Sequence
166576532	tgaggccaggctcctgggtttactcaggtacacacaggattggaagccttgggttacctg
166576472	tggtagtgagtgtggggctgatggcctcagatgtcttgtgaggcagactgtgtcctaagt
166576412	atgagacactcagggtgcttcccccatcaggggcccctgaggggaaggtgggaag[tcact
166576352	ggcctgggctctgctctctggagtgatggagggacag][tccctggcctgggctctgctctc
166576292	tggagtgatggagggagag][tccctggcctgggctctgctctctggagtgatggagggaga
166576232	g][tccctggcctgggctctgctctctggagtgatggagggacag][tccctggcctgggctct
166576172	gctctctggagagatggagggag]cactaatgcaggttactcagcactgtcttttctcaca
166576112	gcag

(The bracketed highlighted sequence represents the VNTR region composed of 5x 42bp repeat sequences)

5.2 Aims

1. To determine the genetic susceptibility variants in patients with sporadic chordomas.

5.3 Objectives

5.3.1 Perform high resolution arrayCGH to determine if a duplication of *T* is a common event in sporadic chordomas.

5.3.2 Sequence *T*, and perform whole exome sequencing to identify possible coding genetic susceptibility variants.

5.3.3 Investigate the frequency of the VNTR in intron 6 in chordomas.

5.4 Results

5.4.1 Tandem duplications of *T* are uncommon in sporadic chordomas

In order to determine specifically if a duplication event of *T* was prevalent in sporadic chordomas, a custom “high resolution” arrayCGH chip with multiple probes covering the *T* region using the breakpoints previously described in the familial study was designed. Additionally, the chip was designed with lower density genome-wide probes to investigate the other common copy number gains and losses seen in chordomas.

Genomic DNA from 22 chordomas (16 disomic for chromosome 6, 5 polysomic for chromosome 6, 1 with amplification of *T*, 2 with minor allelic gain, 1 cell line with gains in chromosome 6q27 suggestive of duplication served as a putative positive control) were hybridised to the chip in comparison to reference disomic human genomic DNA. The rationale for utilising tumour DNA for this analysis was that if *T* duplication was not identified, it would be unlikely to be present in the germline DNA. Conversely the presence of duplication would provide a guide as to the DNA breakpoints to be interrogated in the matching germline DNA. The comparative genomic results (analysis described in materials and methods) show that there are no tandem duplications of *T* in any of the sporadic tumour samples (**Figure 5.2**).

<u>Other structural aberrations</u>

11 of 22 (50%) cases showed homozygous loss of the chromosome 9p region which harbours the *CDKN2A/B* locus and 5 of 22 (23%) had losses of chromosome 17p (*P53* locus).

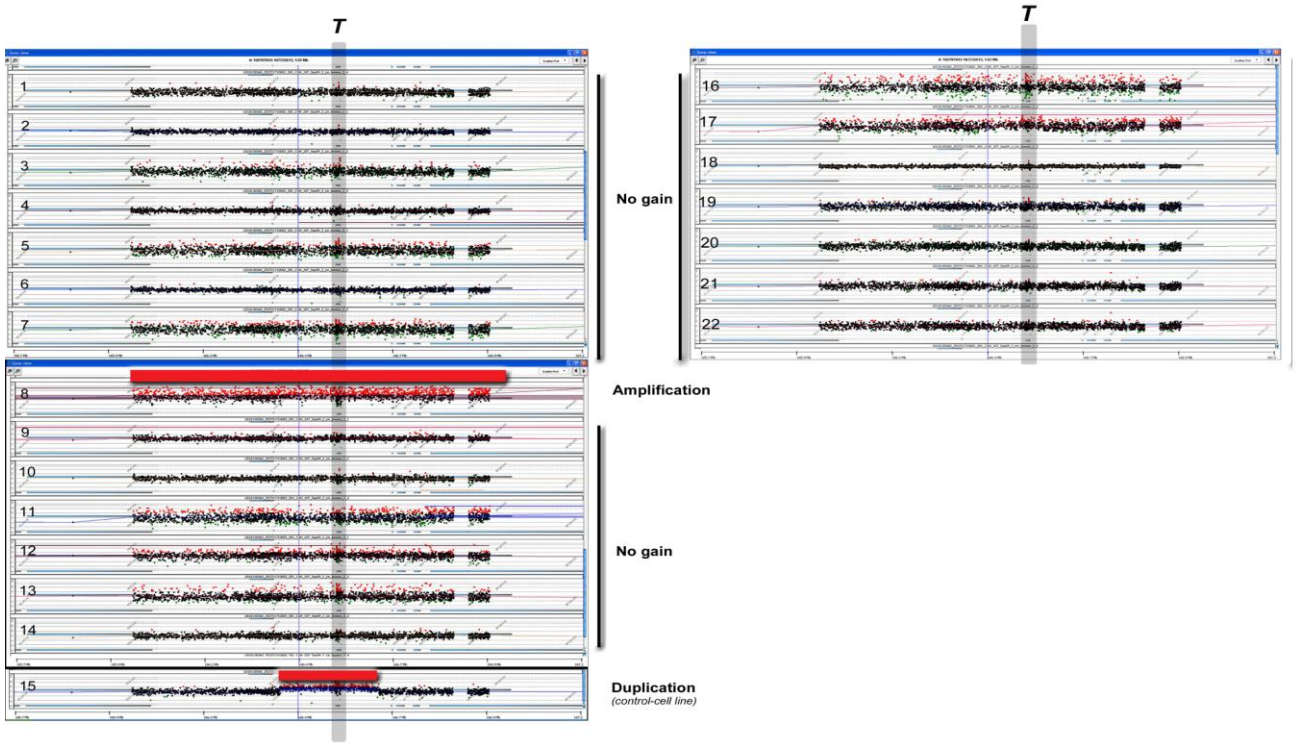


Figure 5.2 ArrayCGH results: Twenty of 22 cases of sporadic chordoma show no evidence of *T* duplication. A case with known amplification demonstrated previously using FISH is confirmed by arrayCGH (Case 8). Duplication of *T* was confirmed in the positive control (Case 15). The location of the *T* gene is indicated by the semi-transparent vertical bar.

5.4.2 A common single nucleotide variant in *T* is strongly associated with chordoma

As duplication of *T* is an uncommon event in sporadic chordomas and is therefore unlikely to be a susceptibility determinant in this cohort, a search for possible sequence variants in *T* (i.e. SNPs or INDELS) that may contribute to disease susceptibility was conducted.

5.4.2.1 Sequencing of *T* exons

A genotyping study of germline DNA from 45 patients with sporadic chordoma was performed (**Figure 5.3**). This was accomplished by combining Sanger sequencing exon data of *T* (n = 23 samples, Founder group-discovery cohort) with whole exome sequencing data from 20 of these patients (**Figure 5.3**). This showed that rs230589 (Gly177Asp, G>A) was the only nsSNP identified in the *T* exons, for which the non-reference allele was present in more than one case of the 23 cases, and it was identified in all samples (**Figure 5.4, Table 5.3**) suggesting a trend towards a disease association.

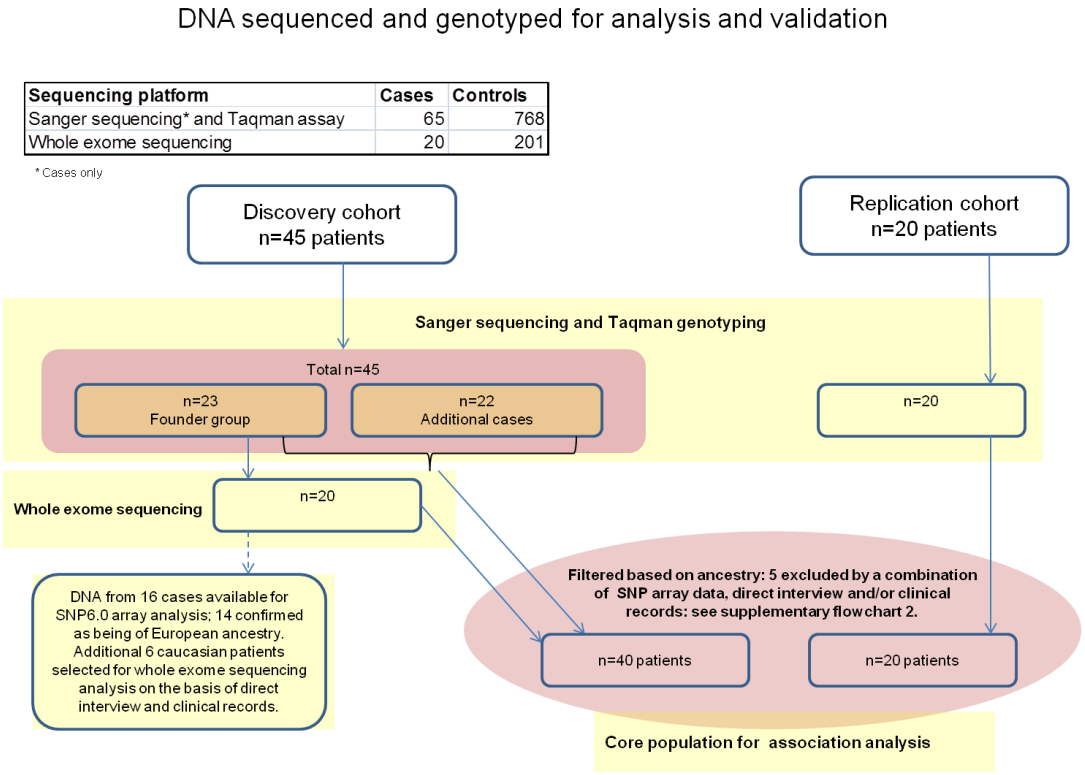


Figure 5.3 Flowchart of the sequencing analysis.

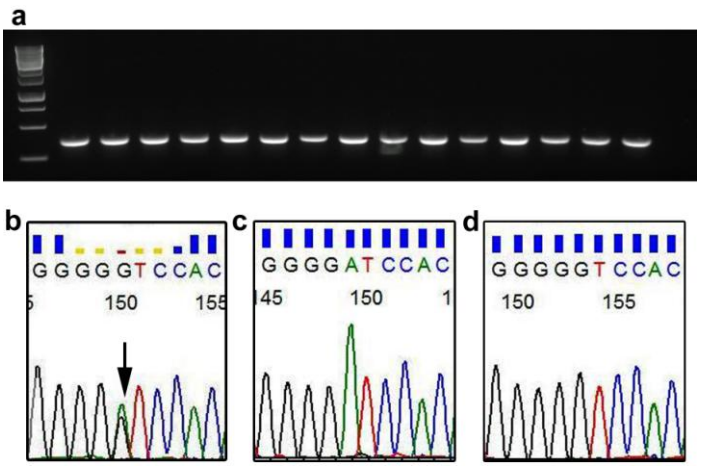


Figure 5.4 Identification of rs2305089 by sequencing of *T*

a) Representative agarose gel demonstrating PCR products of amplified exon 4.

b, c) Sequencing chromatograms showing samples heterozygous (G/A) and homozygous (A/A) for rs2305089.

d) A genomic control homozygous (G/G) for the reference allele.

Table 5.3 Exonic SNPs identified in *T* by whole exome sequencing.

Marker Name	Position	Gene	Exonic Function
rs3816300	**chr6:166572005**	***T***	**nonsynonymous SNV**
rs3127328	**chr6:166572045**	***T***	**nonsynonymous SNV**
rs117097130	**chr6:166574346**	***T***	**nonsynonymous SNV**
rs2305089*	**chr6:166579270**	***T***	**nonsynonymous SNV**
rs1056048	**chr6:166580188**	***T***	**synonymous SNV**

(*rs2305089 is the only recurrent nsSNP identified in all samples.)

5.4.2.2 SNP rs2305089 becomes the focus for further research

rs230589 is a common SNP (ancestral allele G, frequency ~47% (CEU) - HapMap) [266]. It is located in exon 4 of *T* and lies in the DNA binding domain of this transcription factor [94]. Using the linkage disequilibrium tools in SNAP [267] on 1000Genomes data it is evident that rs2305089 is in linkage disequilibrium (r^2>0.8) with 11 other SNPs across a 500kb haplotype block, but none of these is predicted to be either deleterious or to be in a putative regulatory region (**Figure 5.5**). These findings together with the missense nature of the Gly177Asp polymorphism and aberrant binding ability makes this a potentially strong candidate causal SNP.

a

SNP	Proxy	RSquared	Chromosome	Major	Minor	MAF
rs2305089	rs2305089	1	chr6	T	C	0.408
rs2305089	rs9459598	1	chr6	T	G	0.408
rs2305089	rs10806845	0.932	chr6	T	C	0.408
rs2305089	rs10806844	0.932	chr6	T	A	0.408
rs2305089	rs9355558	0.899	chr6	G	A	0.417
rs2305089	rs9348079	0.899	chr6	G	C	0.417
rs2305089	rs9347106	0.899	chr6	G	A	0.417
rs2305089	rs9347104	0.899	chr6	T	C	0.417
rs2305089	rs9365956	0.899	chr6	T	C	0.417
rs2305089	rs3127434	0.899	chr6	G	A	0.417
rs2305089	rs3127344	0.865	chr6	A	T	0.392
rs2305089	rs4510650	0.837	chr6	T	C	0.433

b

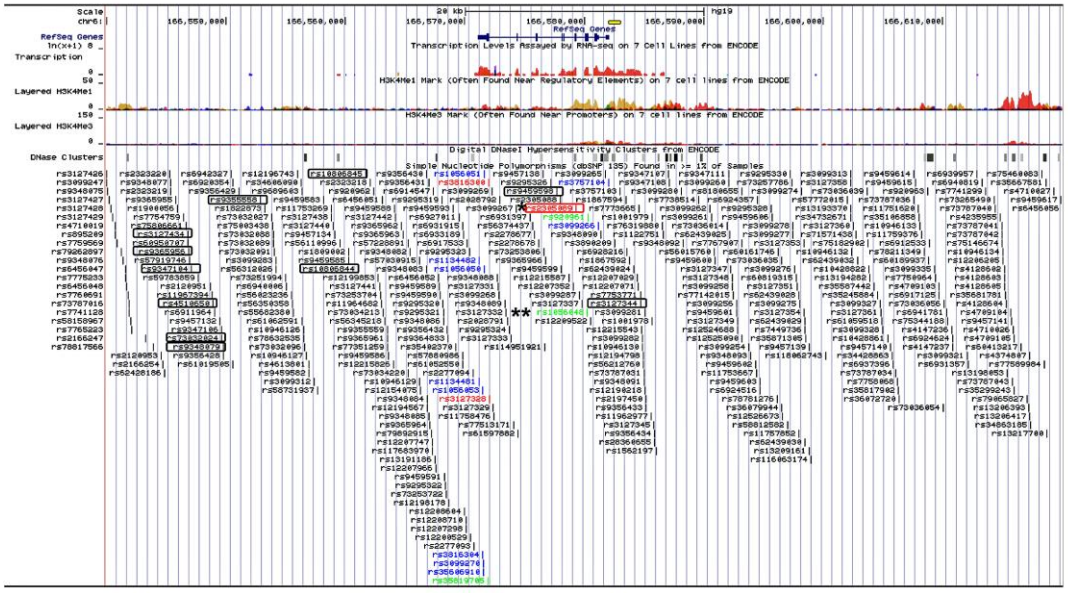


Figure 5.5 Architecture of genetic variation around the *T* locus

a. Table of SNPs in linkage with rs2305089 (r^2>8) in a 500kb around *T*. Data are derived from 1000genomes project using SNAP http://www.broadinstitute.org/mpg/snap/

b. UCSC view of SNPs in linkage disequilibrium (black boxes) with rs2305089 (red box) and the relationship of these SNPs with putative regulatory elements surrounding *T*. **rs1056048, a synonymous SNP, which was identified (*P*=0.0049) in chordomas from whole exome sequencing data, is not in strong linkage with *rs2305089 and is not located in a putative regulatory region.

5.4.2.3 Association study between rs2305089 and chordoma

A case-control analysis was then performed to assess the association between rs2305089 and chordoma risk. In order to increase the power of the study, Sanger sequencing of exon 4 was then performed in an additional 22 cases (**Figure 5.3**). This confirmed that the risk allele was present in all cases.

In view of chordomas being significantly more frequent in people of European ancestry, compared to other population groups, the analysis was restricted to that population cohort (**Figures 5.3 and 5.6**).

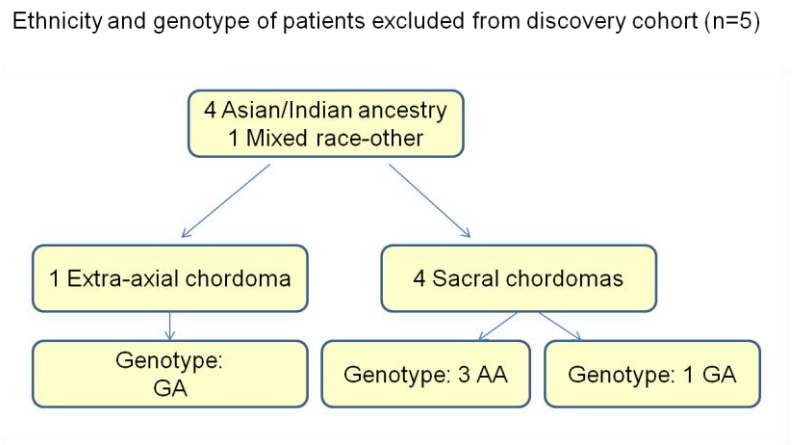


Figure 5.6 Flowchart of cases filtered after ancestry analysis.

Using TaqMan genotyping, the sequencing result was validated in all cases and additional genotyping of 358 matching controls was performed (**Figure 5.7**). The ethnicity of patients was determined by direct interview of patients and/or review of clinical records. In addition, genome-wide SNP array data analysis [73] for ethnicity outlier analysis was performed on a subset of the samples on which whole exome sequencing was available (**Figure 5.8**). Following this analysis and following interviews and/or reviews of the clinical records there were two cases that did not cluster with those of European ancestry and these

were excluded from the final analyses (**Figures 5.3 and 5.6**). An additional 3 patients from the discovery cohort were excluded on the basis of ethnicity determined by direct interview and/or clinical records.

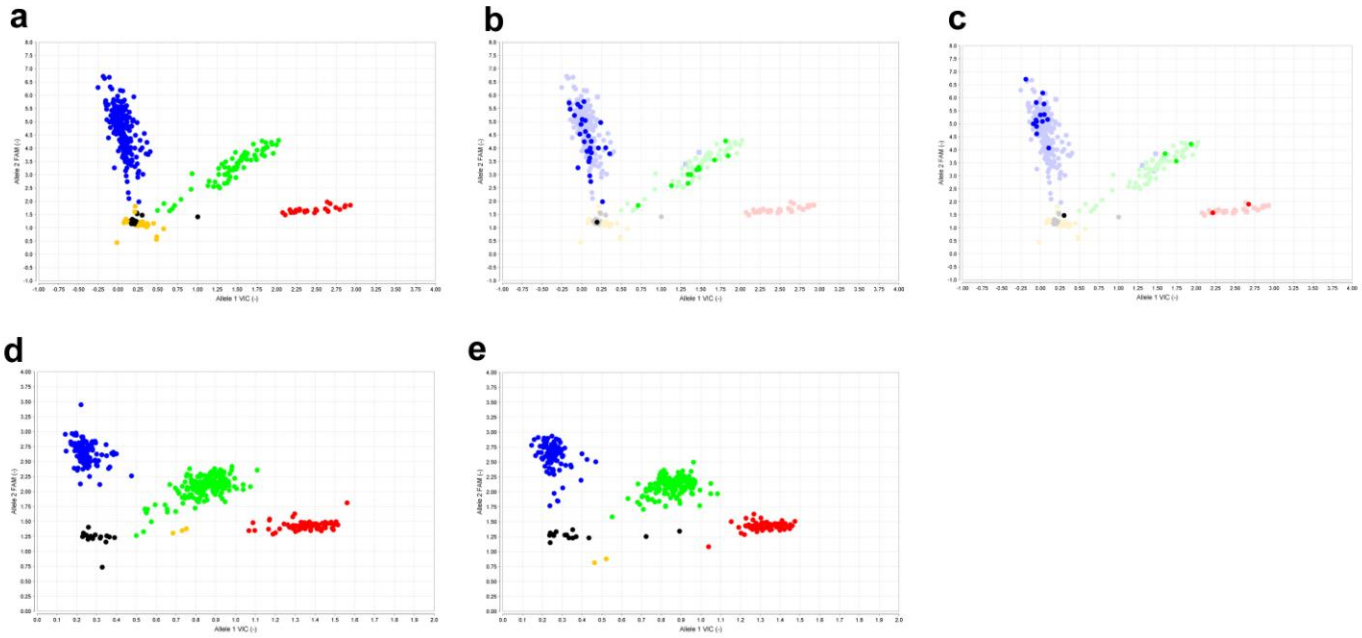


Figure 5.7 Cluster plots of Taqman assay results for rs2305089

a. Combined discovery and replication results for rs2305089 Taqman genotyping assay. Large number of data points reflects the assay: performed in quadruplicate.
b. Discovery cohort superimposed on combined results. One of four replicate experiments displayed in foreground.
c. Replication cohort superimposed on combined results. One of four replicate experiments displayed in foreground.

d, e. Genotyping plots of controls used for discovery and replicate groups. Data points: Blue=AA, Green=GA, Red=GG, Black=undetermined/No template controls, Semi-transparent = replicate data points. Squares=positive control

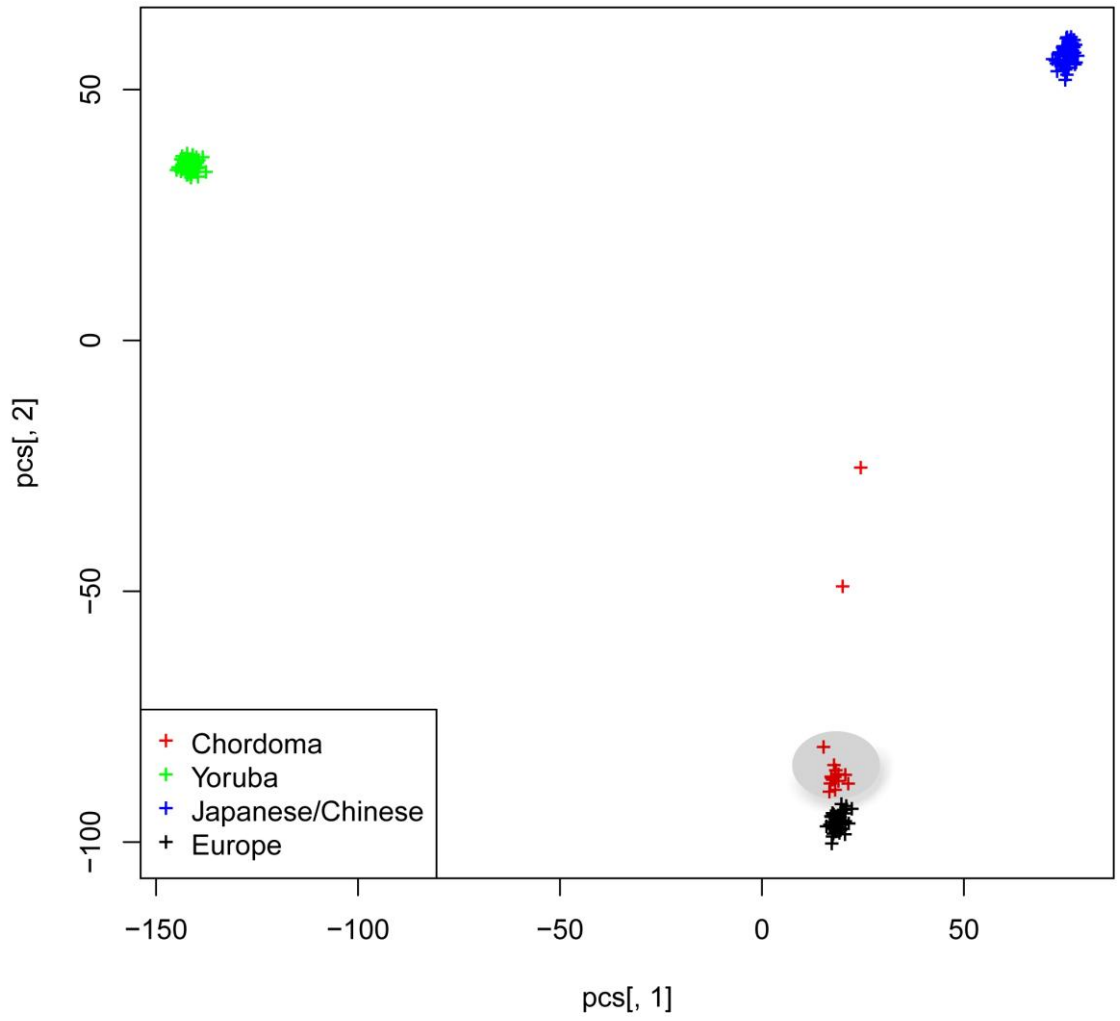


Figure 5.8 Ethnicity outlier analysis-Cases

Principal component analysis results of subset of patients' DNA on which whole exome sequencing was performed using HAPMAP and chordoma SNP6.0 data. The HAPMAP samples are labelled in blue (Japanese/Chinese), green (Yoruba) and black (European). Shaded area indicates tight clustering of chordoma samples (red) used for exome analysis. Two outliers identified were excluded from the study.

Overall in this filtered discovery set (40 cases, 358 controls) there was a strong association between the A risk allele on rs2305089 and chordoma (OR=6.1, $P=4.4x10^{-9}$, **Table 5.4**). To determine if the results could be replicated, a second case-control analysis on DNA from 20 additional patients from other referral

centres, and 363 controls was performed. All patients were of European ancestry, an assessment based on interview and/or clinical records. These were genotyped using the TaqMan assay and further validated using Sanger sequencing (**Figure 5.7**). This study also revealed a strong association between the A allele on rs2305089 with chordoma patients (OR= 4.1, P=2.8x10^{-4}, CI=1.8-9.5, **Table 5.4**). Combining discovery and replication sets we found the A allele strongly associated with the presence of chordoma (Combined OR=5.3, P combined=4.6x10^{-12}**, Table 5.4**).

Table 5.4 Association analysis of rs2305089 and chordoma

		Allelic model									
Ancestry matched cases	**RAF**	**Cases**			**RAF**	**Controls**					
		AA	**GA**	**GG**		**AA**	**GA**	**GG**	**P value**	**OR**	**95% CI**
Discovery	0.88	30	10	0	0.53	101	179	78	4.4x10^{-9}	6.1	3.1-12.1
Replication	0.825	15	3	2	0.53	99	188	76	2.8x10^{-4}	4.1	1.8-9.5
Combined									4.6x10^{-12}	5.3	3.1-8.9

5.4.2.4 Genome-wide search for other coding variants

Although the functional role of rs2305089 makes it a potential causal variant for developing chordoma, it is a common SNP. To determine if there were a potentially rare, coding variant on this common haplotype an interrogation of a 800 kb region across *T* using the chordoma whole exome sequencing data (n=20) and 201 Caucasian controls was conducted (**Figure 5.9-outlier analysis**). This confirmed that rs2305089 was associated with chordoma (OR = 18.4, P = 6.25x10^{-9}, 95%CI 4.64-159), and the analysis also identified a total of 23 other coding variants (all SNPs) (**Table 5.5 and Figure 5.10**) none of which reached a convincing significance level ($P < 0.001$). These results indicate that rs2305089 is the most likely causal variant in this chromosome region. A gene-

based collapsing analysis was then performed to test whether rare functional variants might be present at elevated frequencies in other genes (Chapter 2-**Methods**) but this showed no convincing associations, even amongst the target genes of T [43].

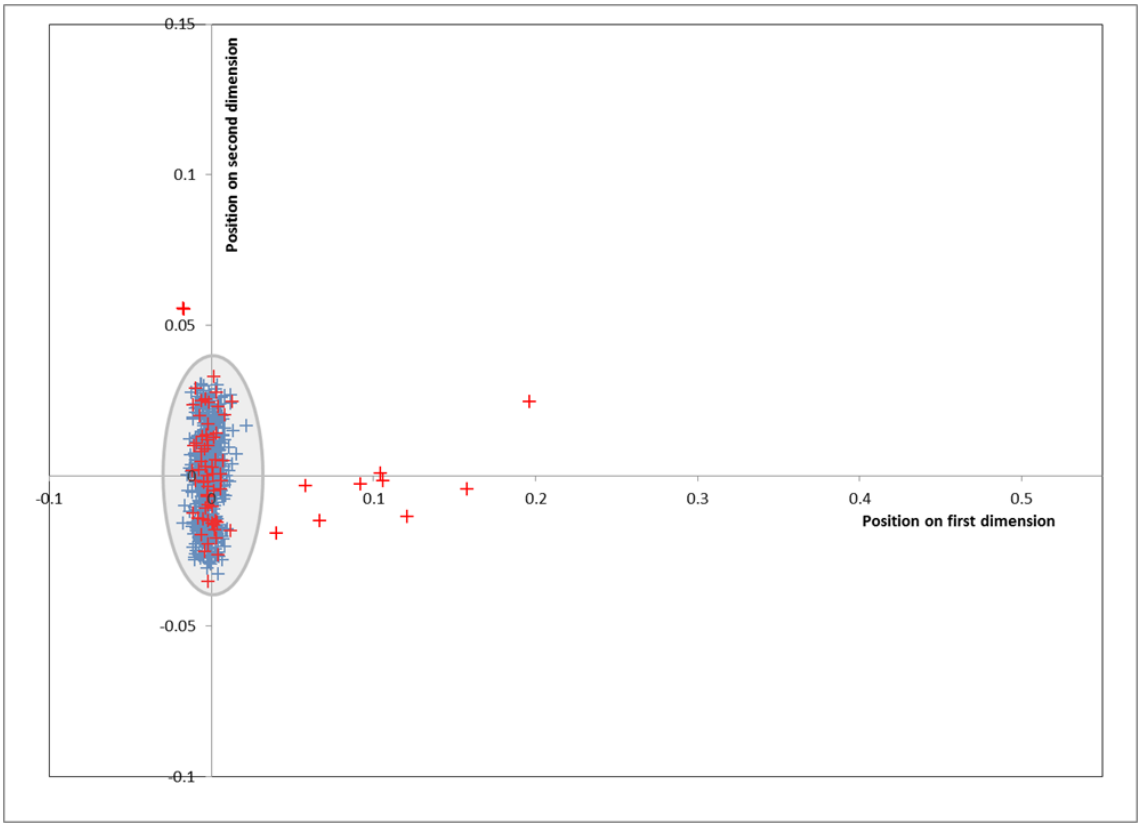


Figure 5.9 Ethnicity outlier analysis-Controls

Ethnic outliers were detected in the control exome series using multi-dimensional scaling analysis implemented with PLINK. 5200 UK (blue) sample data were provided by the Wellcome Trust Case Control Consortium, and Control Exome (red) samples were genotyped using several different genome-wide arrays. 18 outliers were identified empirically from analysis of the first four dimensions, and are illustrated outside the shaded area on a plot of the first two dimensions. 16/224 exome samples had no genome-wide array data available. In each of these cases clinical notes were reviewed and three further individuals were identified with non-white British ethnicity and excluded from the study.

Table 5.5 SNPs identified in 800kb region surrounding *T* derived from exome sequencing data of chordoma

MarkerName	Position	P.value	*Gene*	ExonicFunc	Ref	Alt	Ref AF.controls	Ref AF.cases
rs3816300	chr6:166572005	0.22484	*T*	nonsynonymous SNV	T	C	0.906896552	0.973684211
rs3127328	chr6:166572045	0.48126	*T*	nonsynonymous SNV	C	T	0.914893617	0.970588235
rs117097130	chr6:166574346	1	*T*	nonsynonymous SNV	G	A	0.990049751	1
rs2305089*	chr6:166579270	6.25E-09	*T*	nonsynonymous SNV	C	T	0.492537313	0.05
rs1056048**	chr6:166580188	0.00496	*T*	synonymous SNV	G	A	0.826633166	0.625
rs911203	chr6:166720806	0.06947	*PRR18*	synonymous SNV	G	C	0.3984375	0.166666667
rs13204594	chr6:166721211	1	*PRR18*	synonymous SNV	G	T	0.35046729	0.333333333
rs7757150	chr6:166721224	0.50555	*PRR18*	nonsynonymous SNV	C	G	0.291666667	0.214285714
rs7088	chr6:166736362	0.61662	*SFT2D1*	synonymous SNV	C	T	0.577114428	0.625
NA	chr6:166736363	1	*SFT2D1*	nonsynonymous SNV	G	A	0.997512438	1
rs11551053	chr6:166739646	1	*SFT2D1*	nonsynonymous SNV	T	C	0.965174129	0.975
NA	chr6:166743080	1	NA	NA	A	G	0.997512438	1
NA	chr6:166743666	1	*SFT2D1*	nonsynonymous SNV	C	T	0.9975	1
rs12205572	chr6:166779464	0.74501	*BRP44L*	synonymous SNV	G	A	0.932835821	0.925
rs11557064	chr6:166780349	1	*BRP44L*	nonsynonymous SNV	G	T	0.957711443	0.975
rs760670	chr6:166826304	1	*RPS6KA2*	synonymous SNV	T	C	0.245901639	0.25
rs147557165	chr6:166836792	1	*RPS6KA2*	synonymous SNV	G	A	0.98	1
rs12214821	chr6:166862233	0.1004	*RPS6KA2*	synonymous SNV	T	C	0.725888325	0.6
rs3736681	chr6:166873010	0.65434	*RPS6KA2*	synonymous SNV	C	T	0.830729167	0.875
rs35257815	chr6:166904212	1	*RPS6KA2*	synonymous SNV	C	T	0.973958333	0.975
rs2230732	chr6:166912062	1	*RPS6KA2*	synonymous SNV	C	T	0.910447761	0.925
rs150080589	chr6:166923787	1	*RPS6KA2*	synonymous SNV	G	A	0.995	1
rs1003857	chr6:166952264	0.53481	*RPS6KA2*	synonymous SNV	G	A	0.205	0.15
rs10946179	chr6:167040463	0.09729	*RPS6KA2*	synonymous SNV	A	G	0.858778626	0.970588235

Ref AF=reference allele frequency.

*Estimated Odds Ratio =18.4, 95% Confidence Interval: 4.64-159

**Estimated Odds Ratio=2.85, 95% Confidence Interval: 1.33-5.9

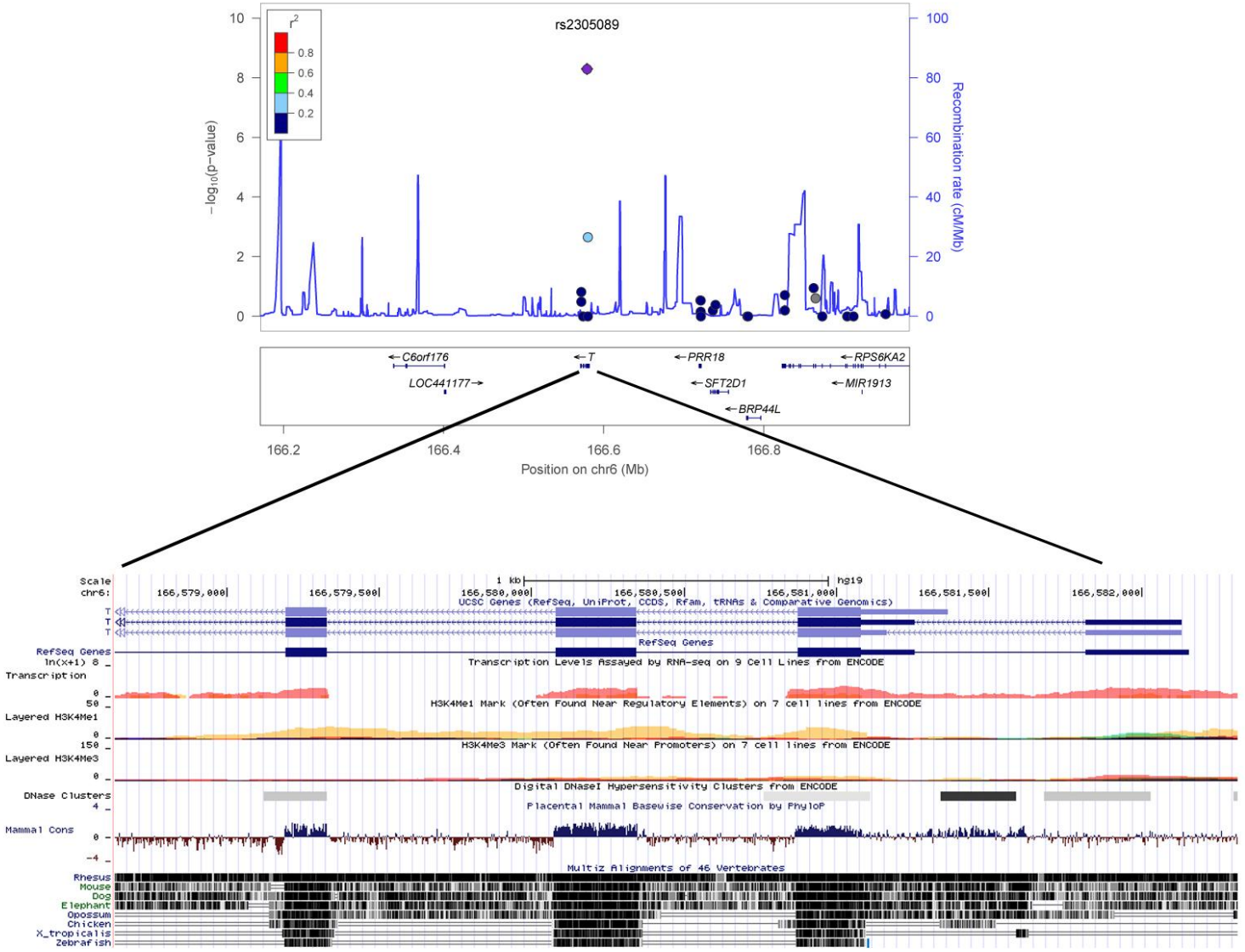


Figure 5.10 Association of rs2305089 with chordoma

Upper panel: Association plot of the 6q27 region from whole-exome sequencing data. Genes with direction of transcription are also indicated. The colour of filled circles representing SNPs depicts the r^2 between that SNP and rs2305089 (diamond).

Bottom panel: UCSC genome browser view (http://genome.ucsc.edu/cgi-bin/hgGateway) of *T* exon 4. RNA-seq expression data derived from human embryonic stem cells with regulatory tracks for H3K4Me1 and DNase1 hypersensitivity. Vertebrate Multiz alignment track indicates high degree of conservation for the ancestral allele.

5.4.2.5 Functional role of rs2305089

The rs2305089 polymorphism alters the binding ability of *T* [94]. Additional functional significance of rs2305089 can be inferred by its amino acid conservation and regional regulatory features identified in human embryonic stem cells (ENCODE) which express *T* (**Figure 5.10**). In view of this, different *T* mRNA expression levels may be reflected by the GA and AA genotypes. qRT-PCR assays on *T* copy number neutral cases were used to investigate this possibility. This analysis revealed that the relative expression ($2^{-\Delta CT}$) of *T* was higher in the AA genotype compared to the heterozygous GA genotype (**Figure 5.11** *P*=0.02, Mann-Witney U test). However, no significant allele specific imbalance was identified amongst the heterozygote cases (**Figure 5.12**). This may in part be accounted for by the expression analysis being performed on tumour tissue known to have multiple structural rearrangements [73], which can regulate allele specific expression [268].

The next question to be addressed was whether the AA genotype with its associated higher level of *T* mRNA correlates with expression levels of downstream targets [43]. Gene Set Enrichment Analysis (Chapter 2-**Methods**)**,** generated from gene expression microarray data, demonstrated that there was significant enrichment for targets in the AA genotype compared to the GA genotype in chordomas (*P* <0.001) (**Figure 5.11**). However, no significant clinical or histological associations were found to correlate with the different genotypes (**Table 5.6**).

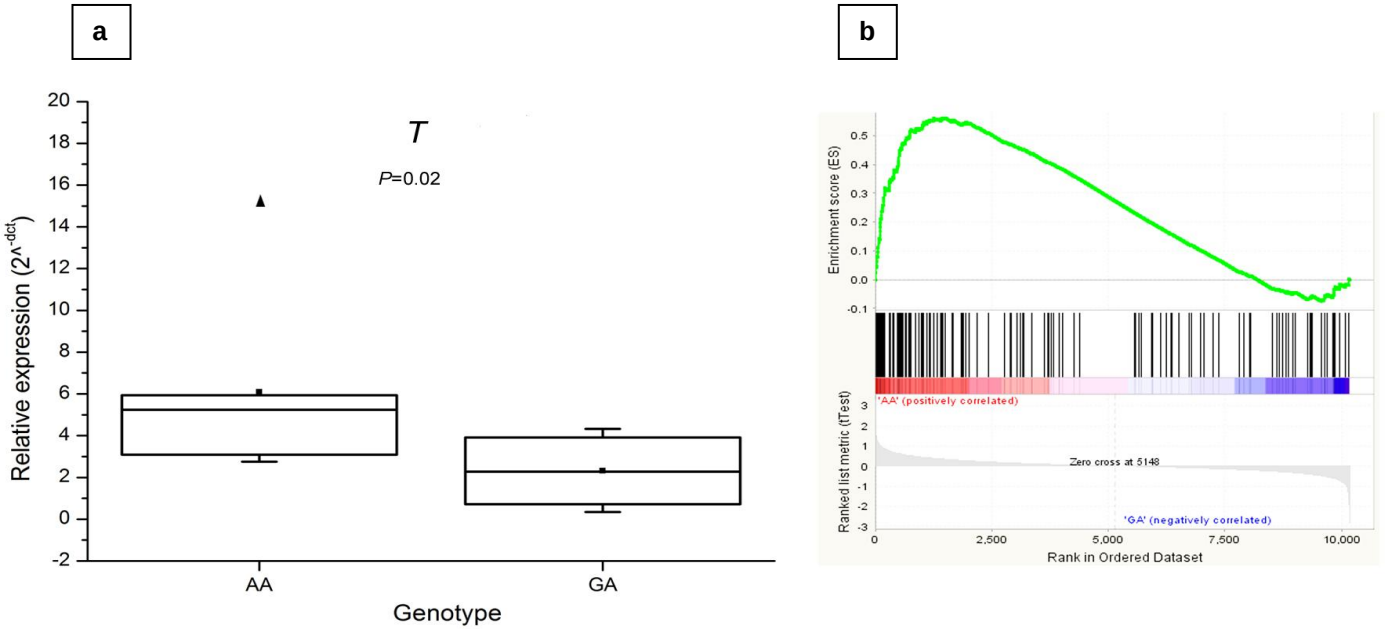
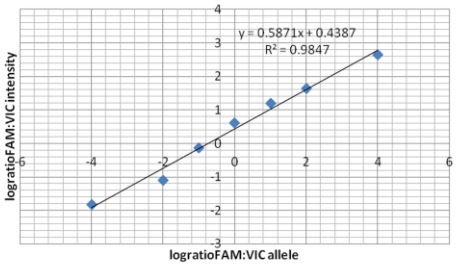


Figure 5.11 Association of rs2305089 and *T* expression

a) Box plots show that significant differences in expression levels of *T* mRNA in chordomas (n=13) depend on the matching genotype. The box represents the middle 50% of data points and interquartile ranges. Triangle represents outlier (>1 standard deviation from ends of box) and square represents mean expression.

b) GSEA plot of T target genes derived from ranked gene expression microarray data between AA and GA genotypes. Normalised enrichment score=2.15, False discovery rate q<0.001.

ratio FAM (A):VIC(G) allele (dilution series)	FAM fluorescence intensity	VIC fluorescence intensity	log ratio FAM:VIC intensity
8:1	202.67	32.33	2.65
4:1	558.67	179.67	1.64
2:1	867.67	376.33	1.21
1:1	649.00	421.00	0.62
1:2	631.00	691.33	-0.13
1:4	340.00	726.67	-1.10
1:8	211.67	745.33	-1.82



1. Fluorescent intensity data for generation of standard curve from serial dilution of homozygote genomic DNA (AA:GG) in various ratios.

Chordoma	FAM (A) fluorescent intensity	VIC (G) flourescent intensity	log ratio FAM:VIC intensity
Case 1	772.5	758	0.027337085
Case 2	1068	1422.5	-0.413517006
Case 3	1325.5	909	0.544184471
Case 4	1404	1143	0.296717532
Case 5	629.5	626.5	0.006891868
Case 6	1407.5	1197.5	0.233109266
Case 7	1908.5	661	1.52971701

2. Fluorescent intensity data for cases disomic for chromosome 6 and heterozygote for rs2305089

Figure 5.12 Allele specific gene expression analysis. The log2 ratio of FAM (A) intensity vs VIC (G) intensity for rs2305089 was generated from six dilutions of homozygous genomic DNA to generate a standard curve. The FAM (A): VIC (G) ratio was then determined on cDNA from seven cases heterozygote for rs2305089.

Table 5.6 Association of rs2305089 genotype and clinicopathological characteristics in discovery and replication groups.

		Genotype			Frequency		Allele count		
	Total	AA	GA	GG	A	G	A	G	$\chi 2$ P value
GENDER									
Male	30	25	4	1	0.9	0.1	54	6	0.075524349
Female	22	13	8	1	0.77	0.23	34	10	
AGE AT PRESENTATION *									
≤58yrs	21	16	4	1	0.86	0.14	36	6	0.798225519
≥59yrs	31	22	8	1	0.84	0.16	52	10	
OUTCOME									
Primary tumour	38	30	7	1	0.88	0.12	67	9	0.099017876
Recurrence and/or metastasis	14	8	5	1	0.75	0.25	21	7	
HISTOLOGY									
Conventional	49	37	10	2	0.86	0.14	84	14	0.209372966
De-differentiated	3	1	2	0	0.67	0.33	4	2	

Association analysis done on cases only, by dividing the patients into two groups based on the parameters in the various categories. Analysis was performed on only those cases in which all data were available (n=52). *mean age of diagnosis used as parameter.

5.4.3 The VNTR in *T* is not in linkage with rs2305089

PCR amplification of the known VNTR region (see Introduction) followed by direct sequencing revealed a heterozygous deletion polymorphism in intron 6 of *T* in the U-CH1 cell line (**Figure 5.13**).

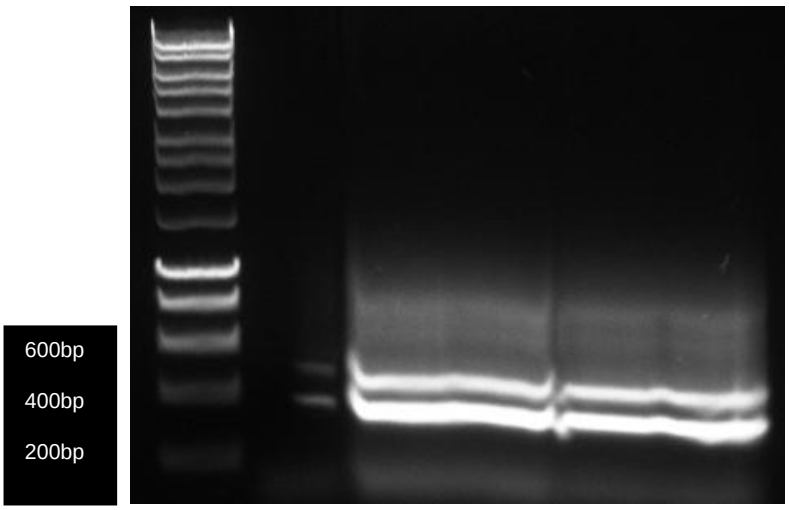


Figure 5.13 VNTR in U-CH1 cell line: The top PCR band consisted of five near perfect 42bp repeats and the bottom band of 2 repeats corresponding to rs67234608/ rs7165389.

To determine if the VNTR polymorphism represented a second, possibly independent susceptibility locus in chordomas, a second case-control study was initiated. As there is no population frequency data regarding this deletion polymorphism, DNA from 30 healthy controls of European ancestry were assembled for this purpose. The VNTR was examined by PCR (**Figure 5.14**) followed by direct sequencing and was categorised as follows:

1. Homozygous for reference sequence (5 repeats) **AA** (one upper band).
2. Heterozygous (5 and 2 repeats) **AB** (two bands)
3. Homozygous (2 repeats) **BB** (one lower band)

The frequency data for this cohort is represented in **Table 5.7** and shows that the homozygous deletion (BB) is the least frequent allele.

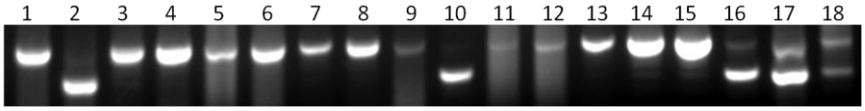


Figure 5.14 PCR of VNTR in controls: Identification of deletion polymorphisms in a representative sample consisting of 18 of the healthy controls. Deletions confirmed by PCR. Individuals 2, 10 are homozygous for the deletion compared to individuals 16, 17, 18 who are heterozygous.

Table 5.7 Frequency of VNTR in control population group.

	AA	**AB**	**BB**
VNTR status	19 (63%)	7 (23%)	4 (14%)

Next, the frequency of the VNTR polymorphism was examined in chordoma patients. PCR followed by sequencing of the germline DNA of 29 chordoma patients revealed a similar distribution of the VNTR to that of the control population. In order to determine if a potential pathogenic somatic expansion of the VNTR existed, the matching chordoma tumour DNA of these patients was also examined. PCR followed by direct sequencing showed 100% concordance between the germline and tumour DNA with no deviation from the germline characteristics in all 29 cases. In view of this and to establish a larger cohort for

the case-control analysis, the VNTR locus was then investigated in a further 27 chordomas (**Table 5.8**).

Table 5.8 Frequency of VNTR in chordoma patients

	AA	AB	BB
Matching germline and tumour DNA	13 (45%)	10 (34%)	6 (21%)
Additional tumour DNA	11 (41%)	14 (52%)	2 (7%)
Total	24 (43%)	24 (43%)	8 (14%)

A case-control association analysis was then performed which did not reveal a significant association between the deletion polymorphism and chordomas (**Table 5.9**).

Table 5.9 Association analysis of VNTR polymorphism in chordoma patients.

RAF	Case			RAF	Controls					
	AA	AB	BB		AA	AB	BB	P value	OR	95% CI
0.36	24	24	8	0.25	19	7	4	0.151	0.6	0.3-1.2

This data was supported by multiple logistic regression analysis which revealed that the VNTR did not add to the rs2305089 model of chordoma risk prediction (*P*=0.519). Overall, this indicates that the VNTR polymorphism is not a significant susceptibility determinant in sporadic chordomas.

5.5 Discussion

Many cancer GWAS studies have identified genetic risk loci for common cancers, however their effect sizes have been modest. Common variations that may influence risk in rare cancers are not well documented possibly because the statistical power of GWAS cannot be leveraged in these patients owing to low sample numbers. The investigation of polymorphisms in these cases therefore requires a more directed approach and the use of prior biological knowledge and targeted sequencing of potential disease susceptibility loci plays a key role in this regard.

T plays a pivotal role in the pathogenesis of chordoma (Chapter 4). It is specifically expressed in chordomas compared to other sarcomas, amplification or minor allelic gain of the gene occurs in a subset of patients with the disease, gene knockdown results in growth arrest and duplication of *T* is a susceptibility determinant in familial chordoma. This provides evidence that *T* is a plausible candidate gene involved in the pathogenesis of chordoma. Using this prior knowledge, strengthened by whole exome sequencing analysis, shows a strong association between the genetic variant in *T* and chordomas. Moreover, the search for potential epistatic interactions in the downstream transcriptional pathway regulated by T shows no significant susceptibility variants. This makes the nsSNP rs2305089 the most likely causal variant. The functional effect imparted by rs2305089 alters the T proteins homodimerisation abilities [94]. This finding is supported by an X-ray crystallography study of DNA bound by the T protein which has revealed surprisingly few direct base pair specific contacts between the DNA and T-box binding domain [95]. This suggests that the target specificity of T-box genes is be achieved through heterodimerisation

with non-T box binding factors as has been seen in Xbra with Smad1 in Xenopus and between TBX2, TBX5 with NKX2-5 protein [97, 195, 196]. It is a reasonable postulate in view of this evidence that heterodimerisation of T protein is also likely to be achieved in chordoma by the presence of the A allele. The altered expression of the target genes between the AA and GA genotypes provides further support for this concept. Additionally, there is evidence for an ETS member family member potentially heterodimerising with T from the *de novo* motif finding results (Chapter 4) in the U-CH1 chordoma cell line.

As rs2305089 is a common SNP and deletion polymorphisms are known to be in linkage with common SNPs in the human genome [269], a VNTR polymorphism in *T* was investigated. This revealed no significant associations with chordoma either independently or additively in a rs2305089 model of risk.

5.6 Conclusion

Overall, the heritability risk conferred by common genetic variants (MAF≥1%) in cancers is generally modest (~1-2 fold) [258]. The finding of a per allele OR>5 in chordomas is exceptional amongst those cancers for which there is a non-Mendelian mode of inheritance. In view of this, and the dearth of functional variants, other than rs2305089, there is a strong case that this SNP not only contributes significantly to the development of chordoma but also represents a novel example in a rare tumour of a strong risk allele that is prevalent in the general population.

Chapter 6

P63 is an upstream regulator of (MOUSE) *T* in murine models but not in human chordomas.

There is good evidence that P63 regulates the expression of (MOUSE) *T* in embryos and murine osteosarcomas. In chordomas, the upstream transcriptional regulators of *T* have not been identified. In this chapter I demonstrate that P63 does not play a role in regulating *T* in human chordomas.

6.1 Introduction

6.1.1 Upstream regulators of *T*

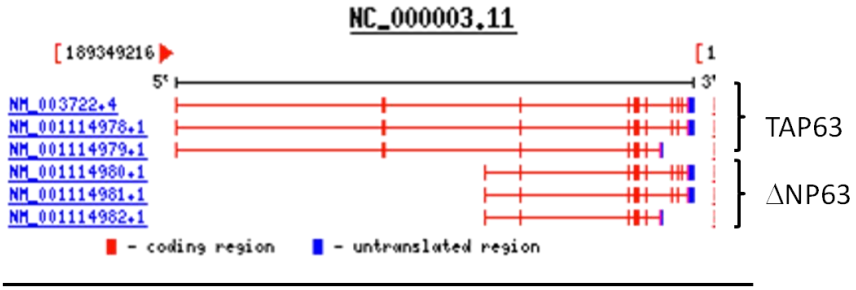
(MOUSE) *T* is a target of the Wnt/β-catenin signalling pathway during murine development [139]. Active WNT signalling can drive reporter expression of a Tcf/Lef promoter 500bp upstream of T in murine embryonic stem cells. T is also a direct target of the active canonical Wnt pathway induced by WNT8/WNT3A during murine paraxial mesoderm specification [88, 219]. Embryos lacking WNT3, TCF or LEF1 have a similar somite phenotype to the T (mouse) mutant. By contrast, unlike in the mesoderm, there is no active canonical WNT signalling in the notochord [88]. Similarly, in chordomas we have demonstrated through our integrative functional genetics analysis that the WNT pathway is unlikely to be active in chordomas (see Chapter 3). Upstream regulation of (MOUSE) T expression during mesoderm development and in the notochord is also known to be brought about by increasing gradient concentrations of morphogens [270]. In Xenopus, Xbra expression shows early activation after induction by increasing gradients of Fgf, Activin and Bmp-4 in the presumptive mesoderm. Both (MOUSE) T and FGF have been shown to function in an autoregulatory positive feedback loop maintaining T expression in the developing notochord in mice. FGF mediates this effect through signal transduction of the MAPK pathway [271]. In chordomas, approximately 94% of cases show expression of the FGF receptor and evidence of active MAPK signalling through activation of this receptor family. However, no mutations of the receptor or of key downstream signalling components or amplification of the common *FGF* signalling effector *ETS2* have been detected in chordomas [63].

The direct upstream regulators of *T* in chordoma have therefore not been established. However, there is strong evidence in murine embryos and in murine cell lines that P63 protein may be a direct upstream regulator of T during embryogenesis and cancer development [272].

6.1.2 P63 isoforms and disease

P63 is a transcription factor essential for the formation of the epidermis and other stratified epithelia. During embryonic development it has a role in limb bud formation and (MOUSE)P63$^{-/-}$ mutants are embryonic lethal [273]. In humans, inherited mutations in *P63* result in ectodermal dysplasias [274].

The role of *P63* in the context of cancer is dependent on the isoforms expressed. There are three variants of the short isoforms (*ΔNP63*) and three of the longer isoform (*TAP63*). The *ΔNP63* variants are transcribed from an alternative promoter in intron 3 and have been shown to act as dominant negative inhibitors of the *P53* family thus acting as proto-oncogenes. The isoform of *P63* that is expressed determines whether it mediates an oncogenic or tumour suppressor effect (**Figure 6.1**) [147, 275, 276].



Tissue/organ site	Variant mRNA expressed	Comments
Breast cancer	ΔNp63	Protein expression
Bladder cancer	TAp63 and ΔNp63	Protein expression
Urothelial cancer	TAp63 and ΔNp63	Protein expression
Cervical cancer	Not determined	Protein expression
Uterine cancer	Not determined	Protein expression
Head and neck cancer	TAp63 and ΔNp63	Protein expression
Lung cancer	TAp63 and ΔNp63	Protein expression
Prostate cancer	ΔNp63	Protein expression
Esophageal cancer	ΔNp63	Squamous cell carcinoma, protein expression
Gastric carcinoma	TAp63 and ΔNp63	Protein expression, mRNA expression
Pancreatic cancer	Not determined	Protein expression

Figure 6.1: Alternative splicing patterns of *P63* mRNA

Upper panel: Variant splicing of *P63* results in multiple isoforms. The TA variants retain a transactivation domain in the 5' region of the mRNA. Alternative promoter usage is also responsible for the differences in transcriptional start sites of the TA and ΔN variants of the gene.

Lower panel: Table of the various malignancies that have been associated with *P63* isoform expression.

In keeping with the disparate roles between these isoforms there are also different functions associated with the same isoform depending on the cellular context. For instance, the *ΔNP63* variant is exclusively expressed in invasive breast carcinoma compared to benign breast disease [275]. This contrasts with urothelial carcinomas in which the more aggressive subtypes are associated with impaired *ΔNP63* expression [277, 278].

6.1.3 P63 is an upstream regulator of (MOUSE) *T* in murine models

During murine development, peak ΔNP63 levels are seen during limb bud development (embryonic day 9.5-12) in the apical ectodermal ridge [279]. T is expressed in the mesoderm around the same time as limb development and T mutant mice have disrupted apical ectodermal ridge development. Furthermore (MOUSE) *T* has previously been identified as a putative transcriptional target of P63 using microarray analysis and P63 mutant embryos show significantly less T expression than their wild type counterparts [280]. In a series of complementary experiments on murine embryonic fibroblasts it has been shown that T is a direct transcriptional target of ΔNP63 and not the TAP63 isoform [272]. Chromatin immunoprecipitation and luciferase reporter transactivation assays support this finding. Also, murine osteosarcoma cell lines co-express both ΔNP63 and (MOUSE) T. In this model when the ΔNP63 message was knocked down, the levels of (MOUSE) T were also reduced, and conversely when expression of ΔNP63alpha and ΔNP63beta were induced in a cell line exhibiting low levels of these transcripts, the levels of (MOUSE) T increased. The higher expression level of (MOUSE) T was therefore implicated in the pathogenesis of osteosarcoma on the basis that reduced expression of ΔNP63 and (MOUSE) T occurred as a consequence of exposing the cells to siRNA directed against ΔNP63 and this was associated with reduced proliferation and migration [272].

6.2 Aim

To investigate whether *T* is transcriptionally regulated by ΔNP63 in human chordomas, and osteosarcomas.

6.3 Objective

1. Determine, by using RT-PCR and protein expression studies, if human chordomas and osteosarcomas express P63 and if so which of the isoforms are expressed.

6.4 Results

6.4.1 mRNA expression of *P63* in chordomas

Twenty six chordomas and the U-CH1 human chordoma cell line were examined for *P63* mRNA expression by RT-PCR, initially using a pan *P63* primer set and thereafter algorithmically using sets of transcript-specific primers for the *TAP63* and *ΔNP63* mRNA variants. All 27 samples showed mRNA expression of pan *P63* but were negative for the *ΔNP63* transcript (**Figure 6.2**). Expression of the alternative transcript *TAP63* was demonstrated in all samples.

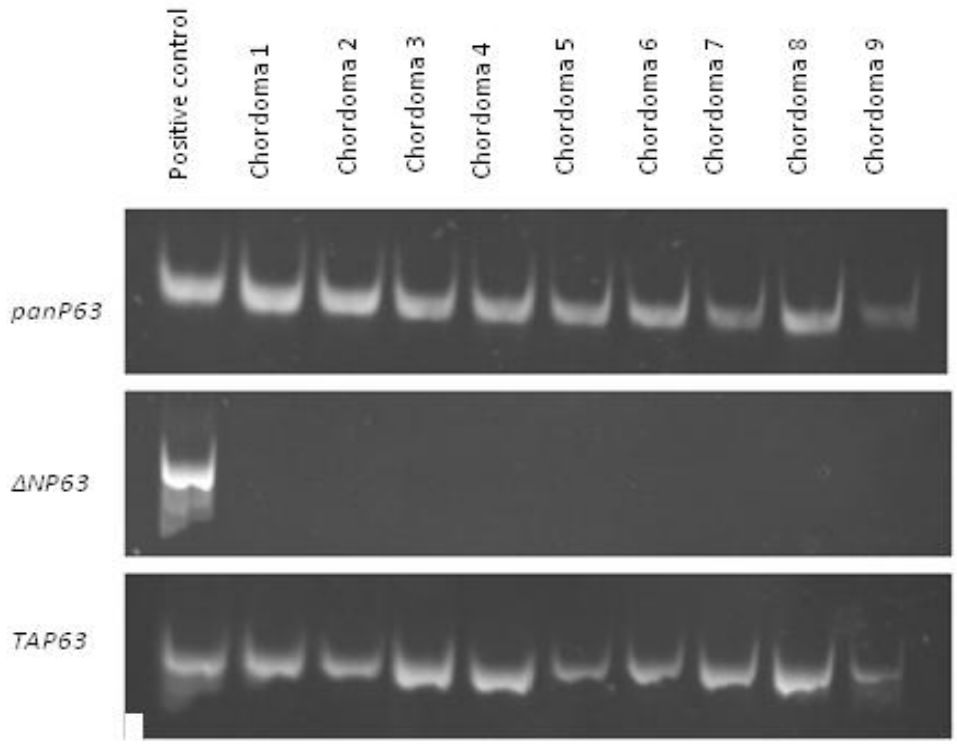


Figure 6.2 RT-PCR expression of *P63* in chordomas.

Polyacrylamide gel showing RT-PCR products of nine representative chordoma samples demonstrating lack of *ΔNP63* mRNA expression.

6.4.2 Protein expression of P63 in chordomas

Protein expression, as assessed by immunohistochemistry, using the pan-p63 antibody (4A4 clone) was performed on an in-house tissue microarray containing duplicate cores of 50 chordomas including the 26 cases that had been analysed by RT-PCR. None of the cases showed immunoreactivity (**Figure 6.3**). The positive and negative controls provided the appropriate results. The 4A4 clone recognises all 6 isoforms of P63, however there have been reports that occasionally the absence of immunohistochemical expression may be related to the inability of conventional antigen retrieval methods to expose the P63 protein epitope. Epitope unmasking can be achieved by protein denaturation techniques as utilised in Western blot. Therefore in order to determine if this was a reason for the absence of immunohistochemical expression, western blotting using the 4A4 antibody was performed on 4 chordomas, all of which expressed T at both mRNA and protein level (**Figure 6.4**).

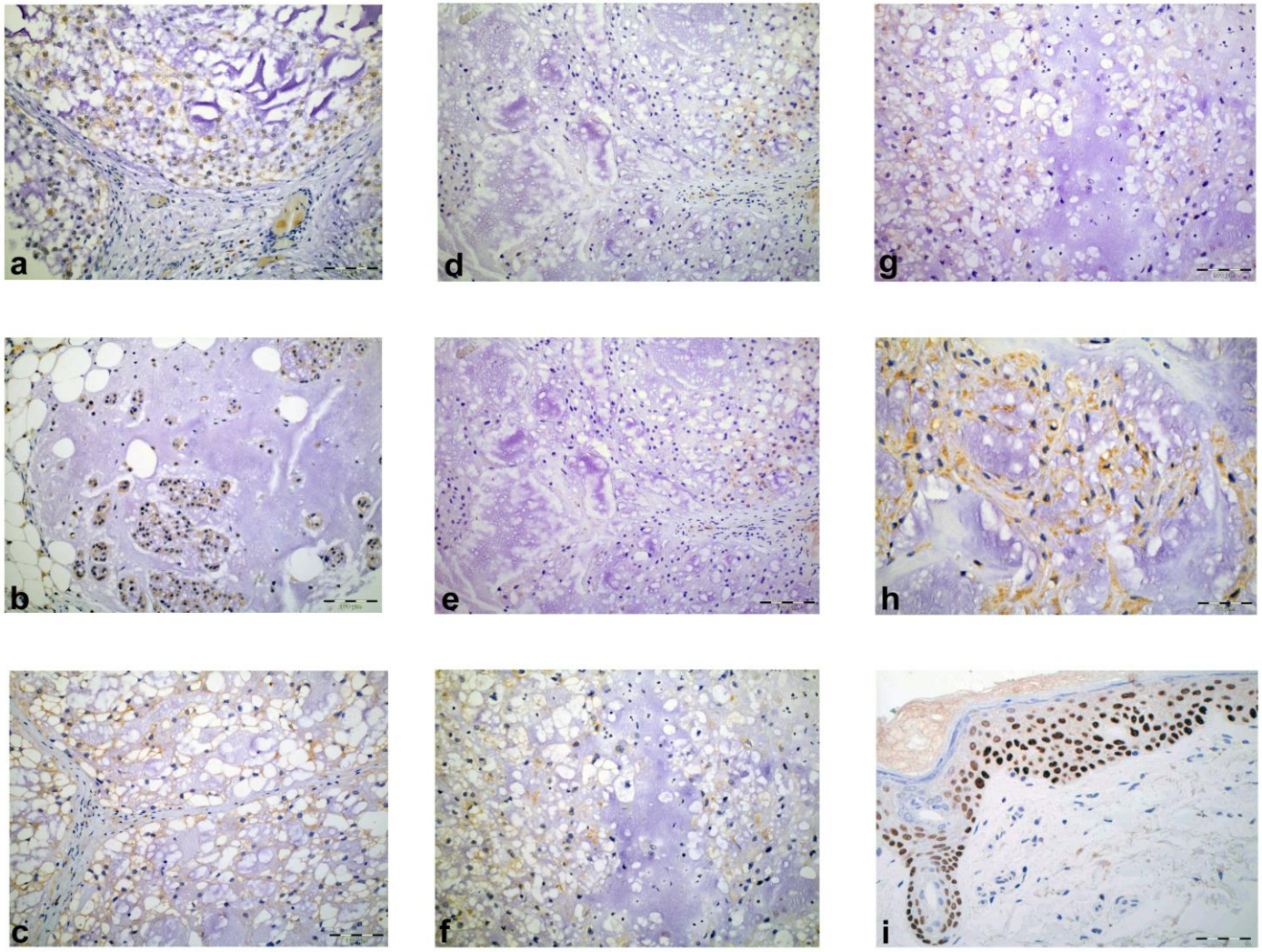


Figure 6.3 Immunohistochemistry results of P63 in human chordomas.

Representative photomicrographs of chordomas counterstained with haematoxylin showing no evidence of nuclear P63 immunoreactivity: i- Keratinocytes of epidermis served as a positive control with strongest expression in the basal layer.

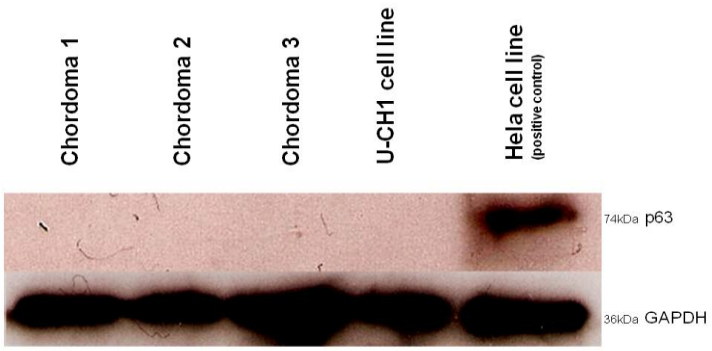


Figure 6.4 Western blot of P63 in chordoma. Three chordomas and U-CH1 cell line showing no protein expression of P63. Protein from the cervical carcinoma Hela cell line served as the positive control.

6.5 Discussion

The P63 protein is a marker of basal cells of stratified epithelia and there is evidence that it plays a role in stem cell renewal in this compartment [281]. Its role in various cancers is diverse but interestingly mutations of *P63* are not common [273]. The question of whether *P63* is a tumour suppressor or oncogene is controversial as these functions can be carried out by variable expression of its isoforms and is context dependent. Evidence for its oncogenic function includes amplification of the *P63* locus in head and neck squamous carcinomas and high expression levels in other squamous malignancies [282, 283].

Tumours with myoepithelial cell differentiation show strong and consistent P63 protein expression. Tumours of this type in the soft tissue are sometimes controversially designated as parachordomas because of their superficial resemblance to chordomas [284]. These could be diagnostically challenging if they are occur in anatomical locations more common for chordoma or when the diagnosis of an extra-axial chordoma is entertained. A case report of a

myoepithelioma/parachordoma arising in the iliac bones was differentiated from chordomas on the basis of its expression of P63 and absence of T protein [285] and it is recommended that T is included in an immunohistochemical panel to differentiate between these two entities [4].

There is little evidence to suggest a broad role for *P63* in tumours of mesenchymal origin. In an immunohistochemical study examining 40 subtypes (650 samples) of sarcomas and there was little or no expression in the majority of tumours [286], the exception being giant cell tumour of bone which demonstrates expression of the TAP63 protein isoform. It has previously been demonstrated using tissue microarrays containing more than 200 human osteosarcomas and gene expression microarrays that there is no expression of *T* in this group of tumours (17, 18). P63 immunoexpression was not identified in any of the tumours in this same tissue microarray (personal communication, Dr Fernanda Amary-Royal National Orthopaedic Hospital). Furthermore, there is evidence in the literature that P63 immunoexpression is rarely identified in osteosarcomas [286, 287]. In view of this it is unlikely that T regulates *P63* in human osteosarcomas. The recent finding of P63 expression in murine osteosarcomas is therefore quite topical [272]. Moreover the strong evidence that the ΔNP63 isoform regulates (MOUSE) *T* in these tumours is exciting as potentially this would be a new avenue of investigation for chordomas.

Through the experiments described above, there is evidence that ΔNP63 protein is rarely if ever a transcriptional regulator of *T* in human chordomas as it is not expressed in these tumours. Although the *TAP63* mRNA transcript is expressed in chordoma it is not expressed at a protein level suggestive of RNA

decay. Nevertheless, it is interesting to note that the gene promoter regulating the expression of (MOUSE) *T* mRNA in murine-derived embryonal carcinoma P19 is similar to that in the murine primitive streak and tailbud but differs from that of the mouse notochord, the structure which chordoma recapitulates [87]. Hence, the discrepant findings with respect to murine and human regulation of *T* by P63 indicate that the regulation of *T* in presumptive mesoderm may differ from that of notochord.

6.6 Conclusion

While there is good experimental evidence that P63 is involved in the transcriptional regulation of (MOUSE) *T* in murine embryos, and that it is implicated in the pathogenesis of murine osteosarcoma, these findings cannot be translated into human disease, at least not in the context of chordoma and osteosarcoma.

Chapter 7

Conclusion and future work

"..... we eventually will develop, in principle, a theory which is capable of explaining everything so well that there will be nothing left to explain. This does not mean, of course, that our explanation necessarily will reflect the way that things actually are. We still will not be able to open the watch, as Einstein put it, but every occurrence in the real world

Gene expression microarray studies and immunohistochemical analysis of a wide variety of sarcomas and other tumours has identified T as being specifically and almost exclusively expressed in chordomas.

At the initiation of this thesis, the genetic events involved in the development and progression of chordomas were largely unknown but the recognition that *T* was expressed in this tumour opened up a previously unrecognized avenue of investigation into the pathogenesis of the disease. The expression of T in chordomas provided an opportunity to identify for the first time, key human T transcriptional targets and to investigate potential regulatory upstream molecules, which may provide therapeutic targets.

The main findings of my work are:

1. Identification of target genes regulated by T in chordoma (Chapters 3 and 4).

    a) Characterisation of the gene expression signature associated with *T* mRNA *expression* in two chordoma cell lines with validation in human tumour samples.
    b) In collaboration with Dr Fiona Wardle's group determining the genome wide binding profile of T in chordoma.
    c) Developing a gene regulatory network that provides molecular evidence for the functional effects noted upon diminishing the expression of T.

(inside the watch) will be accounted for by a corresponding element of our final supertheory." Gary Zukav, Dancing Wu Li Masters,1991

2. Determining that the *T* gene harbours a nsSNP which is a significant genetic susceptibility determinant in patients with sporadic chordomas (Chapter 5).

    a) Although familial and sporadic chordomas both express T, most sporadic chordomas do not harbour a duplication of the *T* gene and this is therefore not a risk determinant in this cohort of patients.
    b) In collaboration with Dr Vincent Plagnol (UGI) we show that the majority of patients of European ancestry with sporadic chordoma inherit an amino acid changing SNP in the DNA binding domain of T. It has previously been shown that this SNP alters the binding ability of the T protein and I show that this influences the selection of downstream target genes.
    c) The risk imparted by this SNP is unusual as the risk variant A (rs2305089) has a frequency of 53% in the general population and 88% in the chordoma cohort, which translates into an estimated odds ratio of 5. This is a rare combination of a large allele frequency with a large odds ratio. The fact that the risk variant is also found in a substantial fraction of the general population makes the finding intriguing.
    d) Through whole exome sequencing analysis (in collaboration with the UCL Genetics Institute and Wellcome Trust Sanger Institute) it is also shown that there isn't another exonic or splice site SNP that is as strongly associated with chordoma as rs2305089 is.

3. In Chapter 6, I discuss the study which shows that in murine embryos and murine osteosarcomas, the ΔNP63 isoform is an upstream regulator of T expression. In investigating a large number of chordomas using immunohistochemistry and q-RTPCR I show that this P63 isoform is not expressed in human chordomas and is therefore unlikely to be a regulator of T expression in human tumours. Moreover there is compelling evidence that both T and P63 are not expressed in human osteosarcomas.

Future work

The search for the downstream target genes and subsequent transcriptional network directed by T has revealed that by targeting the *T* mRNA by an RNA interference approach itself may be a rational therapeutic option. Work on this area has already begun in collaboration with gene therapy experts in UCL.

However it is clear from the *in vitro* work on the cell lines and subsequent research on the rs2305089 locus that a gene therapy approach may not be effective in every chordoma. Thus stratification of patients for this type of therapy could potentially be performed based on either the SNP status of the patient or the gene expression profile of the tumour or both.

Another avenue of research worth pursuing is the role of *ETV1* in chordomas and whether this may operate in a positive feedback loop as other ETS proteins do with T during development.

Chordomas are inherently chemo-resistant and the finding that a well known resistance gene *AKR1B10* is a target of T opens up new possibilities for treatment. AKR1B10 inhibitors are available and pre-clinical research would need to be developed to determine if they could be effective in an adjuvant setting.

References

1. Mazanet R, Antman KH: **Sarcomas of soft tissue and bone**. *Cancer* 1991, **68**(3):463-473.

2. Kelley SP, Ashford RU, Rao AS, Dickson RA: **Primary bone tumours of the spine: a 42-year survey from the Leeds Regional Bone Tumour Registry**. *European spine journal : official publication of the European Spine Society, the European Spinal Deformity Society, and the European Section of the Cervical Spine Research Society* 2007, **16**(3):405-409.

3. Whelan J, McTiernan A, Cooper N, Wong YK, Francis M, Vernon S, Strauss SJ: **Incidence and survival of malignant bone sarcomas in England 1979-2007**. *Int J Cancer* 2011.

4. Tirabosco R, Mangham DC, Rosenberg AE, Vujovic S, Bousdras JK, Pizzolitto S, Maglio GD, Bakker MAD, Francesco LD, Kalil RK *et al*: **Brachyury Expression in Extra-axial Skeletal and Soft Tissue Chordomas : A Marker that Distinguishes Chordoma From Mixed Tumor / Myoepithelioma / Parachordoma in Soft Tissue**. *American Journal of Surgical Pathology, The* 2008, **32**:572-580.

5. Las Heras F, Pritzker KP, Colgan TJ: **Chordoma arising in a mature cystic teratoma of the ovary: a case report**. *Pathol Res Pract* 2007, **203**(6):467-471.

6. McMaster ML, Goldstein AM, Bromley CM, Ishibe N, Parry DM: **Chordoma: incidence and survival patterns in the United States, 1973-1995**. *Cancer Causes Control* 2001, **12**(1):1-11.

7. Ashwood N, Hoskin PJ, Saunders MI: **Metastatic chordoma: pattern of spread and response to chemotherapy**. *Clinical oncology* 1994, **6**(5):341-342.

8. McMaster ML, Goldstein AM, Parry DM: **Clinical features distinguish childhood chordoma associated with tuberous sclerosis complex**

(TSC) from chordoma in the general paediatric population. *J Med Genet* 2011, **48**(7):444-449.

9. Schroeder BA, Wells RG, Starshak RJ, Sty JR: **Clivus chordoma in a child with tuberous sclerosis: CT and MR demonstration**. *J Comput Assist Tomogr* 1987, **11**(1):195-196.

10. Miozzo M, Dalpra L, Riva P, Volonta M, Macciardi F, Pericotti S, Tibiletti MG, Cerati M, Rohde K, Larizza L *et al*: **A tumor suppressor locus in familial and sporadic chordoma maps to 1p36**. *International journal of cancer Journal international du cancer* 2000, **87**(1):68-72.

11. Dalpra L, Malgara R, Miozzo M, Riva P, Volonte M, Larizza L, Fuhrman Conti AM: **First cytogenetic study of a recurrent familial chordoma of the clivus**. *International journal of cancer Journal international du cancer* 1999, **81**(1):24-30.

12. Kelley MJ, Korczak JF, Sheridan E, Yang X, Goldstein AM, Parry DM: **Familial chordoma, a tumor of notochordal remnants, is linked to chromosome 7q33**. *Am J Hum Genet* 2001, **69**(2):454-460.

13. Yang XR, Beerman M, Bergen AW, Parry DM, Sheridan E, Liebsch NJ, Kelley MJ, Chanock S, Goldstein AM: **Corroboration of a familial chordoma locus on chromosome 7q and evidence of genetic heterogeneity using single nucleotide polymorphisms (SNPs)**. *Int J Cancer* 2005, **116**(3):487-491.

14. Yang XR, Ng D, Alcorta DA, Liebsch NJ, Sheridan E, Li S, Goldstein AM, Parry DM, Kelley MJ, America N: **T (brachyury) gene duplication confers major susceptibility to familial chordoma**. *Nature Genetics* 2009, **41**:1176-1178.

15. Yang XR, Ng D, Alcorta DA, Liebsch NJ, Sheridan E, Li S, Goldstein AM, Parry DM, Kelley MJ: **T (brachyury) gene duplication confers major susceptibility to familial chordoma**. *Nat Genet* 2009, **41**(11):1176-1178.

16. Chetty R, Levin CV, Kalan MR: **Chordoma: a 20-year clinicopathologic review of the experience at Groote Schuur Hospital, Cape Town**. *J Surg Oncol* 1991, **46**(4):261-264.

17. Presneau N, Shalaby A, Ye H, Pillay N, Halai D, Idowu B, Tirabosco R, Whitwell D, Jacques TS, Kindblom LG *et al*: **Role of the transcription factor T (brachyury) in the pathogenesis of sporadic chordoma: a genetic and functional-based study**. *J Pathol* 2011, **223**(3):327-335.

18. Roberti F, Sekhar LN, Jones RV, Wright DC: **Intradural cranial chordoma: a rare presentation of an uncommon tumor. Surgical experience and review of the literature**. *J Neurosurg* 2007, **106**(2):270-274.

19. Yonemoto T, Tatezaki S, Takenouchi T, Ishii T, Satoh T, Moriya H: **The surgical management of sacrococcygeal chordoma**. *Cancer* 1999, **85**(4):878-883.

20. Hsieh PC, Xu R, Sciubba DM, McGirt MJ, Nelson C, Witham TF, Wolinksy JP, Gokaslan ZL: **Long-term clinical outcomes following en bloc resections for sacral chordomas and chondrosarcomas: a series of twenty consecutive patients**. *Spine* 2009, **34**(20):2233-2239.

21. Stener B, Gunterberg B: **High amputation of the sacrum for extirpation of tumors. Principles and technique**. *Spine* 1978, **3**(4):351-366.

22. Samson IR, Springfield DS, Suit HD, Mankin HJ: **Operative treatment of sacrococcygeal chordoma. A review of twenty-one cases**. *The Journal of bone and joint surgery American volume* 1993, **75**(10):1476-1484.

23. Walcott BP, Nahed BV, Mohyeldin A, Coumans J-V, Kahle KT, Ferreira MJ: **Chordoma: current concepts, management, and future directions**. *The Lancet Oncology* 2012, **13**(2):e69-e76.

24. Amichetti M, Cianchetti M, Amelio D, Enrici RM, Minniti G: **Proton therapy in chordoma of the base of the skull: a systematic review**. *Neurosurg Rev* 2009, **32**(4):403-416.

25. Nguyen QN, Chang EL: **Emerging role of proton beam radiation therapy for chordoma and chondrosarcoma of the skull base**. *Curr Oncol Rep* 2008, **10**(4):338-343.

26. Casali PG, Stacchiotti S, Sangalli C, Olmi P, Gronchi A: **Chordoma**. *Current opinion in oncology* 2007, **19**(4):367-370.

27. Fleming GF, Heimann PS, Stephens JK, Simon MA, Ferguson MK, Benjamin RS, Samuels BL: **Dedifferentiated chordoma. Response to aggressive chemotherapy in two cases**. *Cancer* 1993, **72**(3):714-718.

28. Stacchiotti S, Longhi A, Ferraresi V, Grignani G, Comandone A, Stupp R, Bertuzzi A, Tamborini E, Pilotti S, Messina A *et al*: **Phase II study of imatinib in advanced chordoma**. *J Clin Oncol* 2012, **30**(9):914-920.

29. Stacchiotti S, Marrari A, Tamborini E, Palassini E, Virdis E, Messina A, Crippa F, Morosi C, Gronchi A, Pilotti S *et al*: **Response to imatinib plus sirolimus in advanced chordoma**. *Ann Oncol* 2009, **20**(11):1886-1894.

30. Casali PG, Messina A, Stacchiotti S, Tamborini E, Crippa F, Gronchi A, Orlandi R, Ripamonti C, Spreafico C, Bertieri R *et al*: **Imatinib mesylate in chordoma**. *Cancer* 2004, **101**(9):2086-2097.

31. Singhal N, Kotasek D, Parnis FX: **Response to erlotinib in a patient with treatment refractory chordoma**. *Anticancer Drugs* 2009, **20**(10):953-955.

32. Launay SG, Chetaille B, Medina F, Perrot D, Nazarian S, Guiramand J, Moureau-Zabotto L, Bertucci F: **Efficacy of epidermal growth factor receptor targeting in advanced chordoma: case report and literature review**. *BMC Cancer* 2011, **11**:423.

33. Schwab J, Antonescu C, Boland P, Healey J, Rosenberg A, Nielsen P, Iafrate J, Delaney T, Yoon S, Choy E *et al*: **Combination of PI3K/mTOR inhibition demonstrates efficacy in human chordoma**. *Anticancer Res* 2009, **29**(6):1867-1871.

34. Yang C, Schwab JH, Schoenfeld AJ, Hornicek FJ, Wood KB, Nielsen GP, Choy E, Mankin H, Duan Z: **A novel target for treatment of chordoma: signal transducers and activators of transcription 3**. *Molecular cancer therapeutics* 2009, **8**(9):2597-2605.

35. Yang C, Hornicek FJ, Wood KB, Schwab JH, Choy E, Mankin H, Duan Z: **Blockage of Stat3 with CDDO-Me inhibits tumor cell growth in chordoma**. *Spine (Phila Pa 1976)* 2010, **35**(18):1668-1675.

36. Hanna SA, Tirabosco R, Amin A, Pollock RC, Skinner JA, Cannon SR, Saifuddin A, Briggs TW: **Dedifferentiated chordoma: a report of four cases arising 'de novo'**. *J Bone Joint Surg Br* 2008, **90**(5):652-656.

37. Batsakis JG, Solomon AR, Rice DH: **The pathology of head and neck tumors: neoplasms of cartilage, bone, and the notochord, part 7**. *Head Neck Surg* 1980, **3**(1):43-57.

38. Mirra JM, Brien EW: **Giant notochordal hamartoma of intraosseous origin: a newly reported benign entity to be distinguished from chordoma. Report of two cases**. *Skeletal radiology* 2001, **30**(12):698-709.

39. Mobley BC, McKenney JK, Bangs CD, Callahan K, Yeom KW, Schneppenheim R, Hayden MG, Cherry AM, Gokden M, Edwards MS *et al*: **Loss of SMARCB1/INI1 expression in poorly differentiated chordomas**. *Acta Neuropathol* 2010, **120**(6):745-753.

40. Tirabosco R, Mangham DC, Rosenberg AE, Vujovic S, Bousdras K, Pizzolitto S, De Maglio G, den Bakker MA, Di Francesco L, Kalil RK *et al*: **Brachyury expression in extra-axial skeletal and soft tissue chordomas: a marker that distinguishes chordoma from mixed**

tumor/myoepithelioma/parachordoma in soft tissue. *Am J Surg Pathol* 2008, **32**(4):572-580.

41. Al-Adnani M, Cannon SR, Flanagan AM: **Chordomas do not express CD10 and renal cell carcinoma (RCC) antigen: an immunohistochemical study**. *Histopathology* 2005, **47**(5):535-537.

42. Ooi A, Wong JC, Petillo D, Roossien D, Perrier-Trudova V, Whitten D, Min BW, Tan MH, Zhang Z, Yang XJ *et al*: **An antioxidant response phenotype shared between hereditary and sporadic type 2 papillary renal cell carcinoma**. *Cancer Cell* 2011, **20**(4):511-523.

43. Nelson AC, Pillay N, Henderson S, Presneau N, Tirabosco R, Halai D, Berisha F, Flicek P, Stemple DL, Stern C *et al*: **An integrated functional genomics approach identifies the regulatory network directed by brachyury (T) in chordoma**. In: *J Pathol.* 2012/08/01 edn; 2012.

44. Naka T, Kuester D, Boltze C, Scheil-Bertram S, Samii A, Herold C, Ostertag H, Krueger S, Roessner A: **Expression of hepatocyte growth factor and c-MET in skull base chordoma**. *Cancer* 2008, **112**(1):104-110.

45. Naka T, Boltze C, Samii A, Samii M, Herold C, Ostertag H, Iwamoto Y, Oda Y, Tsuneyoshi M, Kuester D *et al*: **Expression of c-MET, low-molecular-weight cytokeratin, matrix metalloproteinases-1 and -2 in spinal chordoma**. *Histopathology* 2009, **54**(5):607-613.

46. Gao Z, Zhang Q, Kong F, Chen G, Li M, Guo H, Liang J, Bao Y, Ling F: **Fascin expression in skull base chordoma: correlation with tumor recurrence and dura erosion**. *Med Oncol* 2012.

47. Ji Z, Long H, Hu Y, Qiu X, Chen X, Li Z, Fan D, Ma B, Fan Q: **Expression of MDR1, HIF-1alpha and MRP1 in sacral chordoma and chordoma cell line CM-319**. *J Exp Clin Cancer Res* 2010, **29**:158.

48. Li X, Ji Z, Ma Y, Qiu X, Fan Q, Ma B: **Expression of hypoxia-inducible factor-1alpha, vascular endothelial growth factor and matrix**

metalloproteinase-2 in sacral chordomas. *Oncol Lett* 2012, **3**(6):1268-1274.

49. Scheipl S, Froehlich EV, Leithner A, Beham A, Quehenberger F, Mokry M, Stammberger H, Varga PP, Lazary A, Windhager R *et al*: **Does insulin-like growth factor 1 receptor (IGF-1R) targeting provide new treatment options for chordomas? A retrospective clinical and immunohistochemical study**. *Histopathology* 2012, **60**(6):999-1003.

50. Tamborini E, Miselli F, Negri T, Lagonigro MS, Staurengo S, Dagrada GP, Stacchiotti S, Pastore E, Gronchi A, Perrone F *et al*: **Molecular and biochemical analyses of platelet-derived growth factor receptor (PDGFR) B, PDGFRA, and KIT receptors in chordomas**. *Clinical cancer research : an official journal of the American Association for Cancer Research* 2006, **12**(23):6920-6928.

51. Klingler L, Trammell R, Allan DG, Butler MG, Schwartz HS: **Clonality studies in sacral chordoma**. *Cancer genetics and cytogenetics* 2006, **171**(1):68-71.

52. Klingler L, Shooks J, Fiedler PN, Marney A, Butler MG, Schwartz HS: **Microsatellite instability in sacral chordoma**. *Journal of surgical oncology* 2000, **73**(2):100-103.

53. Scheil S, Bruderlein S, Liehr T, Starke H, Herms J, Schulte M, Moller P: **Genome-wide analysis of sixteen chordomas by comparative genomic hybridization and cytogenetics of the first human chordoma cell line, U-CH1**. *Genes Chromosomes Cancer* 2001, **32**(3):203-211.

54. Hallor KH, Staaf J, Jönsson G, Heidenblad M, {Vult von Steyern} F, Bauer HCF, Ijszenga M, Hogendoorn PCW, Mandahl N, Szuhai K *et al*: **Frequent deletion of the CDKN2A locus in chordoma: analysis of chromosomal imbalances using array comparative genomic hybridisation.** *British journal of cancer* 2008, **98**:434-442.

55. Diaz RJ, Guduk M, Romagnuolo R, Smith CA, Northcott P, Shih D, Berisha F, Flanagan A, Munoz DG, Cusimano MD *et al*: **High-resolution Whole-Genome Analysis of Skull Base Chordomas Implicates FHIT Loss in Chordoma Pathogenesis**. *Neoplasia* 2012, **14**(9):788-798.

56. Henderson SR, Guiliano D, Presneau N, McLean S, Frow R, Vujovic S, Anderson J, Sebire N, Whelan J, Athanasou N *et al*: **A molecular map of mesenchymal tumors**. *Genome Biol* 2005, **6**(9):R76.

57. Bruderlein S, Sommer JB, Meltzer PS, Li S, Osada T, Ng D, Moller P, Alcorta DA, Kelley MJ: **Molecular characterization of putative chordoma cell lines**. *Sarcoma* 2010, **2010**:630129.

58. Schwab JH, Boland PJ, Agaram NP, Socci ND, Guo T, Toole GCO, Wang X, Ostroumov E, Hunter CJ, Block JA *et al*: **Chordoma and chondrosarcoma gene profile : implications for immunotherapy**. *Cancer* 2009:339-349.

59. Vujovic S, Henderson S, Presneau N, Odell E, Jacques TS, Tirabosco R, Boshoff C, Flanagan AM: **Brachyury, a crucial regulator of notochordal development, is a novel biomarker for chordomas**. *J Pathol* 2006, **209**(2):157-165.

60. Palena C, Polev DE, Tsang KY, Fernando RI, Litzinger M, Krukovskaya LL, Baranova AV, Kozlov AP, Schlom J: **The human T-box mesodermal transcription factor Brachyury is a candidate target for T-cell-mediated cancer immunotherapy**. *Clin Cancer Res* 2007, **13**(8):2471-2478.

61. Roselli M, Fernando RI, Guadagni F, Spila A, Alessandroni J, Palmirotta R, Costarelli L, Litzinger M, Hamilton D, Huang B *et al*: **Brachyury, a Driver of the Epithelial-Mesenchymal Transition, Is Overexpressed in Human Lung Tumors: An Opportunity for Novel Interventions against Lung Cancer**. *Clin Cancer Res* 2012, **18**(14):3868-3879.

62. Risbud MV, Schaer TP, Shapiro IM: **Toward an understanding of the role of notochordal cells in the adult intervertebral disc: From discord to accord**. *Dev Dyn* 2010, **239**(8):2141-2148.

63. Shalaby AaE, Presneau N, Idowu BD, Thompson L, Briggs TRW, Tirabosco R, Diss TC, Flanagan AM: **Analysis of the fibroblastic growth factor receptor-RAS/RAF/MEK/ERK-ETS2/brachyury signalling pathway in chordomas.** *Modern pathology : an official journal of the United States and Canadian Academy of Pathology, Inc* 2009, **22**:996-1005.

64. Le LP, Nielsen GP, Rosenberg AE, Thomas D, Batten JM, Deshpande V, Schwab J, Duan Z, Xavier RJ, Hornicek FJ *et al*: **Recurrent chromosomal copy number alterations in sporadic chordomas**. *PLoS One* 2011, **6**(5):e18846.

65. Forbes SA, Bhamra G, Bamford S, Dawson E, Kok C, Clements J, Menzies A, Teague JW, Futreal PA, Stratton MR: **The Catalogue of Somatic Mutations in Cancer (COSMIC)**. *Curr Protoc Hum Genet* 2008, **Chapter 10**:Unit 10 11.

66. Tamborini E, Negri T, Miselli F, Lagonigro MS, Pricl S, Pilotti S: **Re: Response of a KIT-Positive Extra-abdominal Fibromatosis to Imatinib Mesylate and KIT Genetic Analysis**. *Journal of the National Cancer Institute* 2006, **98**(21):1583-1584.

67. Walter BA, Begnami M, Valera VA, Santi M, Rushing EJ, Quezado M: **Gain of chromosome 7 by chromogenic in situ hybridization (CISH) in chordomas is correlated to c-MET expression**. *J Neurooncol* 2011, **101**(2):199-206.

68. Shalaby A, Presneau N, Ye H, Halai D, Berisha F, Idowu B, Leithner A, Liegl B, Briggs TR, Bacsi K *et al*: **The role of epidermal growth factor receptor in chordoma pathogenesis: a potential therapeutic target**. *J Pathol* 2011, **223**(3):336-346.

69. Mitelman F JBaMF: **Mitelman Database of Chromosome Aberrations and Gene Fusions in Cancer** In.; 2012.

70. Longoni M, Orzan F, Stroppi M, Boari N, Mortini P, Riva P: **Evaluation of 1p36 markers and clinical outcome in a skull base chordoma study**. *Neuro Oncol* 2008, **10**(1):52-60.

71. Bayrakli F, Guney I, Kilic T, Ozek M, Pamir MN: **New candidate chromosomal regions for chordoma development**. *Surg Neurol* 2007, **68**(4):425-430; discussion 430.

72. Fearon ER, Vogelstein B: **A genetic model for colorectal tumorigenesis**. *Cell* 1990, **61**(5):759-767.

73. Stephens PJ, Greenman CD, Fu B, Yang F, Bignell GR, Mudie LJ, Pleasance ED, Lau KW, Beare D, Stebbings LA *et al*: **Massive genomic rearrangement acquired in a single catastrophic event during cancer development**. *Cell* 2011, **144**(1):27-40.

74. Choi KS, Harfe BD: **Hedgehog signaling is required for formation of the notochord sheath and patterning of nuclei pulposi within the intervertebral discs**. *Proceedings of the National Academy of Sciences of the United States of America* 2011, **108**(23):9484-9489.

75. Maier JA, Harfe BD: **Nuclei pulposi formation from the embryonic notochord occurs normally in GDF-5-deficient mice**. *Spine* 2011, **36**(24):E1555-1561.

76. Choi KS, Cohn MJ, Harfe BD: **Identification of nucleus pulposus precursor cells and notochordal remnants in the mouse: implications for disk degeneration and chordoma formation**. *Developmental dynamics : an official publication of the American Association of Anatomists* 2008, **237**(12):3953-3958.

77. Cates JM, Itani DM, Coffin CM, Harfe BD: **The sonic hedgehog pathway in chordoid tumours**. *Histopathology* 2010, **56**(7):978-979.

78. Yamaguchi T, Suzuki S, Ishiiwa H, Ueda Y: **Intraosseous benign notochordal cell tumours : overlooked precursors of classic chordomas ?** *Histopathology* 2004:597-602.

79. Herrmann BG: **Expression pattern of the Brachyury gene in whole-mount embryos mutant**. *Most* 1991, **917**:913-917.

80. Kispert A, Koschorz B, Herrmann BG: **The T protein encoded by Brachyury is a tissue-specific transcription factor**. *EMBO J* 1995, **14**(19):4763-4772.

81. Naiche LA, Harrelson Z, Kelly RG, Papaioannou VE: **T-box genes in vertebrate development**. *Annu Rev Genet* 2005, **39**:219-239.

82. Jacobs-Cohen RJ, Spiegelman M, Bennett D: **Abnormalities of cells and extracellular matrix of T/T embryos**. *Differentiation* 1983, **25**(1):48-55.

83. Dobrovolskaia-Zavadskaia N: **Preliminary Symbols for a Tail-Mutation in Mice**. *The American Naturalist* 1933, **67**(709):186-188.

84. Herrmann BG: **Expression pattern of the Brachyury gene in whole-mount TWis/TWis mutant embryos**. *Development* 1991, **113**(3):913-917.

85. Schulte-Merker S, van Eeden FJ, Halpern ME, Kimmel CB, Nusslein-Volhard C: **no tail (ntl) is the zebrafish homologue of the mouse T (Brachyury) gene**. *Development* 1994, **120**(4):1009-1015.

86. Herrmann BG: **Brachyury (T)**. 1995, **6**:385-394.

87. Yamaguchi H, Niimi T, Kitagawa Y, Miki K: **Brachyury (T) expression in embryonal carcinoma P19 cells resembles its expression in primitive streak and tail-bud but not that in notochord**. *DNA Sequence* 1999:253-264.

88. Martin BL, Kimelman D: **Regulation of canonical Wnt signaling by Brachyury is essential for posterior mesoderm formation**. *Dev Cell* 2008, **15**(1):121-133.

89. Clements D, Taylor HC, Herrmann BG, Stott D: **Distinct regulatory control of the Brachyury gene in axial and non-axial mesoderm suggests separation of mesoderm lineages early in mouse gastrulation**. *Mech Dev* 1996, **56**(1-2):139-149.

90. Calvo R, Drabkin HA: **Embryonic genes in cancer**. *Ann Oncol* 2000, **11 Suppl 3**:207-218.

91. Monk M, Holding C: **Human embryonic genes re-expressed in cancer cells**. *Oncogene* 2001, **20**(56):8085-8091.

92. Herrmann BG, Lehrach H: **From phenotype to gene: molecular cloning in the Brachyury (T) locus region**. *Curr Top Microbiol Immunol* 1988, **137**:77-81.

93. Edwards YH, Putt W, Lekoape KM, Stott D, Fox M, Hopkinson DA, Sowden J: **The human homolog T of the mouse T(Brachyury) gene; gene structure, cDNA sequence, and assignment to chromosome 6q27**. *Genome Res* 1996, **6**(3):226-233.

94. Papapetrou C, Edwards YH, Sowden JC: **The T transcription factor functions as a dimer and exhibits a common human polymorphism Gly-177-Asp in the conserved DNA-binding domain**. *FEBS Lett* 1997, **409**(2):201-206.

95. Muller CW, Herrmann BG: **Crystallographic structure of the T domain-DNA complex of the Brachyury transcription factor**. *Nature* 1997, **389**(6653):884-888.

96. Kispert A, Herrmann BG: **The Brachyury gene encodes a novel DNA binding protein**. *EMBO J* 1993, **12**(8):3211-3220.

97. Messenger NJ, Kabitschke C, Andrews R, Grimmer D, {Núñez Miguel} R, Blundell TL, Smith JC, Wardle FC: **Functional specificity of the Xenopus T-domain protein Brachyury is conferred by its ability to interact with Smad1.** *Developmental cell* 2005, **8**:599-610.

98. Tada M, Smith JC: **T-targets: clues to understanding the functions of T-box proteins**. *Dev Growth Differ* 2001, **43**(1):1-11.

99. Morley RH, Lachani K, Keefe D, Gilchrist MJ, Flicek P, Smith JC, Wardle FC: **A gene regulatory network directed by zebrafish No tail accounts for its roles in mesoderm formation**. *Proc Natl Acad Sci U S A* 2009, **106**(10):3829-3834.

100. Evans AL, Faial T, Gilchrist MJ, Down T, Vallier L, Pedersen RA, Wardle FC, Smith JC: **Genomic targets of Brachyury (T) in differentiating mouse embryonic stem cells**. *PLoS One* 2012, **7**(3):e33346.

101. Hotta K, Yamada S, Ueno N, Satoh N, Takahashi H: **Brachyury-downstream notochord genes and convergent extension in Ciona intestinalis embryos**. *Dev Growth Differ* 2007, **49**(5):373-382.

102. Li QY, Newbury-Ecob RA, Terrett JA, Wilson DI, Curtis AR, Yi CH, Gebuhr T, Bullen PJ, Robson SC, Strachan T *et al*: **Holt-Oram syndrome is caused by mutations in TBX5, a member of the Brachyury (T) gene family**. *Nature genetics* 1997, **15**(1):21-29.

103. Vance KW, Carreira S, Brosch G, Goding CR: **Tbx2 is overexpressed and plays an important role in maintaining proliferation and suppression of senescence in melanomas**. *Cancer Res* 2005, **65**(6):2260-2268.

104. Yu J, Ma X, Cheung KF, Li X, Tian L, Wang S, Wu CW, Wu WK, He M, Wang M *et al*: **Epigenetic inactivation of T-box transcription factor 5, a novel tumor suppressor gene, is associated with colon cancer**. *Oncogene* 2010, **29**(49):6464-6474.

105. Pillay N, Plagnol V, S.Tarpey P, B.Lobo S, Presneau N, Szuhai K, Halai D, Berisha F, R.Cannon S, Mead S *et al*: **A common single nucleotide variant in T is strongly associated with chordoma** *Nature genetics* 2012.

106. Sarkar D, Shields B, Davies ML, Muller J, Wakeman JA: **BRACHYURY confers cancer stem cell characteristics on colorectal cancer cells**. *Int J Cancer* 2012, **130**(2):328-337.

107. Merscher S, Funke B, Epstein JA, Heyer J, Puech A, Lu MM, Xavier RJ, Demay MB, Russell RG, Factor S *et al*: **TBX1 is responsible for cardiovascular defects in velo-cardio-facial/DiGeorge syndrome**. *Cell* 2001, **104**(4):619-629.

108. Jacobs JJ, Keblusek P, Robanus-Maandag E, Kristel P, Lingbeek M, Nederlof PM, van Welsem T, van de Vijver MJ, Koh EY, Daley GQ *et al*: **Senescence bypass screen identifies TBX2, which represses Cdkn2a (p19(ARF)) and is amplified in a subset of human breast cancers**. *Nature genetics* 2000, **26**(3):291-299.

109. Borozdin W, Bravo-Ferrer Acosta AM, Seemanova E, Leipoldt M, Bamshad MJ, Unger S, Kohlhase J: **Contiguous hemizygous deletion of TBX5, TBX3, and RBM19 resulting in a combined phenotype of Holt-Oram and ulnar-mammary syndromes**. *Am J Med Genet A* 2006, **140A**(17):1880-1886.

110. Rodriguez M, Aladowicz E, Lanfrancone L, Goding CR: **Tbx3 Represses E-Cadherin Expression and Enhances Melanoma Invasiveness**. *Cancer research* 2008, **68**(19):7872-7881.

111. Fan W, Huang X, Chen C, Gray J, Huang T: **TBX3 and its isoform TBX3+2a are functionally distinctive in inhibition of senescence and are overexpressed in a subset of breast cancer cell lines.** *Cancer research* 2004, **64**:5132-5139.

112. Morita M, Watanabe M, Inoue N, Inaoka C, Akamizu T, Tatsumi KI, Hidaka Y, Iwatani Y: **Functional polymorphisms in TBX21 and HLX are associated with development and prognosis of Graves' disease**. *Autoimmunity* 2012, **45**(2):129-136.

113. Ge M, Zheng Y, Li X, Shi J, Ge J, Li H, Feng S: **The polymorphisms of T cell-specific TBX21 and STAT4 genes may contribute to the susceptibility of Chinese individuals to aplastic anemia**. *Human immunology* 2012, **73**(1):118-121.

114. Marcano AC, Doudney K, Braybrook C, Squires R, Patton MA, Lees MM, Richieri-Costa A, Lidral AC, Murray JC, Moore GE *et al*: **TBX22 mutations are a frequent cause of cleft palate**. *Journal of Medical Genetics* 2004, **41**(1):68-74.

115. Ashktorab H, Schäffer AA, Daremipouran M, Smoot DT, Lee E, Brim H: **Distinct Genetic Alterations in Colorectal Cancer**. *PLoS One* 2010, **5**(1):e8879.

116. Stemple DL: **Structure and function of the notochord: an essential organ for chordate development**. *Development* 2005, **132**(11):2503-2512.

117. Stemple DL, Solnica-Krezel L, Zwartkruis F, Neuhauss SC, Schier AF, Malicki J, Stainier DY, Abdelilah S, Rangini Z, Mountcastle-Shah E *et al*: **Mutations affecting development of the notochord in zebrafish**. *Development* 1996, **123**:117-128.

118. Smits P, Lefebvre V: **Sox5 and Sox6 are required for notochord extracellular matrix sheath formation, notochord cell survival and development of the nucleus pulposus of intervertebral discs**. *Development* 2003, **130**(6):1135-1148.

119. Cleaver O, Seufert DW, Krieg PA: **Endoderm patterning by the notochord: development of the hypochord in Xenopus**. *Development* 2000, **127**(4):869-879.

120. Cleaver O, Krieg PA: **Notochord patterning of the endoderm**. *Developmental biology* 2001, **234**(1):1-12.

121. Goldstein AM, Fishman MC: **Notochord regulates cardiac lineage in zebrafish embryos**. *Developmental biology* 1998, **201**(2):247-252.

122. Passamaneck YJ, Katikala L, Perrone L, Dunn MP, Oda-Ishii I, {Di Gregorio} A: **Direct activation of a notochord cis-regulatory module by Brachyury and FoxA in the ascidian Ciona intestinalis.** *Development (Cambridge, England)* 2009, **136**:3679-3689.

123. Yamaguchi T, Watanabe-Ishiiwa H, Suzuki S, Igarashi Y, Ueda Y: **Incipient chordoma: a report of two cases of early-stage chordoma arising from benign notochordal cell tumors**. *Mod Pathol* 2005, **18**(7):1005-1010.

124. Kyriakos M: **Benign notochordal lesions of the axial skeleton: a review and current appraisal**. *Skeletal Radiol* 2011, **40**(9):1141-1152.

125. Faust DB, Gilmore HR, Mudgett CS: **CHORDOMATA: A REVIEW OF THE LITERATURE, WITH REPORT OF A SACROCOCCYGEAL CASE***. *Annals of Internal Medicine* 1944, **21**(4):678-698.

126. Salisbury JR, Deverell MH, Cookson MJ, Whimster WF: **Three-dimensional reconstruction of human embryonic notochords: clue to the pathogenesis of chordoma**. *J Pathol* 1993, **171**(1):59-62.

127. Choi KS, Harfe BD: **Hedgehog signaling is required for formation of the notochord sheath and patterning of nuclei pulposi within the intervertebral discs**. *Proc Natl Acad Sci U S A* 2011, **108**(23):9484-9489.

128. Jambhekar NA, Rekhi B, Thorat K, Dikshit R, Agrawal M, Puri A: **Revisiting chordoma with brachyury, a "new age" marker: analysis of a validation study on 51 cases**. *Arch Pathol Lab Med* 2010, **134**(8):1181-1187.

129. Fernando RI, Litzinger M, Trono P, Hamilton DH, Schlom J, Palena C: **The T-box transcription factor Brachyury promotes epithelial-mesenchymal transition in human tumor cells**. *J Clin Invest* 2010, **120**(2):533-544.

130. Roselli M, Fernando RI, Guadagni F, Spila A, Alessandroni J, Palmirotta R, Costarelli L, Litzinger M, Hamilton D, Huang B *et al*: **Brachyury, a Driver of the Epithelial–Mesenchymal Transition, Is Overexpressed in Human Lung Tumors: An Opportunity for Novel Interventions against Lung Cancer**. *Clinical Cancer Research* 2012, **18**(14):3868-3879.

131. Vermeulen L, De Sousa EMF, van der Heijden M, Cameron K, de Jong JH, Borovski T, Tuynman JB, Todaro M, Merz C, Rodermond H *et al*: **Wnt activity defines colon cancer stem cells and is regulated by the microenvironment**. *Nat Cell Biol* 2010, **12**(5):468-476.

132. Kilic N, Feldhaus S, Kilic E, Tennstedt P, Wicklein D, Wasielewski R, Viebahn C, Kreipe H, Schumacher U: **Brachyury expression predicts poor prognosis at early stages of colorectal cancer**. *Eur J Cancer* 2011, **47**(7):1080-1085.

133. He S, Nakada D, Morrison SJ: **Mechanisms of stem cell self-renewal**. *Annu Rev Cell Dev Biol* 2009, **25**:377-406.

134. Drazin D, Rosner J, Avalos P, Acosta F: **Stem cell therapy for degenerative disc disease**. *Adv Orthop* 2012, **2012**:961052.

135. Burkert J, Wright NA, Alison MR: **Stem cells and cancer: an intimate relationship**. *J Pathol* 2006, **209**(3):287-297.

136. Ravandi F, Estrov Z: **Eradication of leukemia stem cells as a new goal of therapy in leukemia**. *Clin Cancer Res* 2006, **12**(2):340-344.

137. Moitra K, Lou H, Dean M: **Multidrug efflux pumps and cancer stem cells: insights into multidrug resistance and therapeutic development**. *Clin Pharmacol Ther* 2011, **89**(4):491-502.

138. Wu X, Goodyear SM, Tobias JW, Avarbock MR, Brinster RL: **Spermatogonial stem cell self-renewal requires ETV5-mediated downstream activation of Brachyury in mice**. *Biol Reprod* 2011, **85**(6):1114-1123.

139. Arnold SJ, Stappert J, Bauer A, Kispert A, Herrmann BG, Kemler R: **Brachyury is a target gene of the Wnt/beta-catenin signaling pathway**. *Mech Dev* 2000, **91**(1-2):249-258.

140. Powell SM, Zilz N, Beazer-Barclay Y, Bryan TM, Hamilton SR, Thibodeau SN, Vogelstein B, Kinzler KW: **APC mutations occur early during colorectal tumorigenesis**. *Nature* 1992, **359**(6392):235-237.

141. Aydemir E, Bayrak OF, Sahin F, Atalay B, Kose GT, Ozen M, Sevli S, Dalan AB, Yalvac ME, Dogruluk T *et al*: **Characterization of cancer stem-like cells in chordoma**. *J Neurosurg* 2012, **116**(4):810-820.

142. Miller GJ, Bauer KA, Barzegar S, Foley AJ, Mitchell JP, Cooper JA, Rosenberg RD: **The effects of quality and timing of venepuncture on markers of blood coagulation in healthy middle-aged men**. *Thromb Haemost* 1995, **73**(1):82-86.

143. Salasznyk RM, Williams WA, Boskey A, Batorsky A, Plopper GE: **Adhesion to Vitronectin and Collagen I Promotes Osteogenic Differentiation of Human Mesenchymal Stem Cells**. *J Biomed Biotechnol* 2004, **2004**(1):24-34.

144. Livak KJ, Schmittgen TD: **Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method**. *Methods* 2001, **25**(4):402-408.

145. Luo B, Cheung HW, Subramanian A, Sharifnia T, Okamoto M, Yang X, Hinkle G, Boehm JS, Beroukhim R, Weir BA *et al*: **Highly parallel identification of essential genes in cancer cells**. *Proc Natl Acad Sci U S A* 2008, **105**(51):20380-20385.

146. Lo HS, Wang Z, Hu Y, Yang HH, Gere S, Buetow KH, Lee MP: **Allelic variation in gene expression is common in the human genome**. *Genome Res* 2003, **13**(8):1855-1862.

147. Herfs M, Hubert P, Suarez-carmona M, Reschner A, Saussez S, Berx G, Savagner P, Boniver J, Delvenne P: **Regulation of p63 Isoforms by Snail and Slug Transcription Factors in Human Squamous Cell Carcinoma**. *Pathology* 2010, **176**:1-9.

148. Rozen S, Skaletsky H: **Primer3 on the WWW for general users and for biologist programmers**. *Methods Mol Biol* 2000, **132**:365-386.

149. Coffey AJ, Kokocinski F, Calafato MS, Scott CE, Palta P, Drury E, Joyce CJ, Leproust EM, Harrow J, Hunt S *et al*: **The GENCODE exome: sequencing the complete human exome**. *Eur J Hum Genet* 2011, **19**(7):827-831.

150. Mamanova L, Coffey AJ, Scott CE, Kozarewa I, Turner EH, Kumar A, Howard E, Shendure J, Turner DJ: **Target-enrichment strategies for next-generation sequencing**. *Nat Methods* 2010, **7**(2):111-118.

151. Li H, Durbin R: **Fast and accurate short read alignment with Burrows-Wheeler transform**. *Bioinformatics* 2009, **25**(14):1754-1760.

152. Wardle FC, Odom DT, Bell GW, Yuan B, Danford TW, Wiellette EL, Herbolsheimer E, Sive HL, Young RA, Smith JC: **Zebrafish promoter microarrays identify actively transcribed embryonic genes**. *Genome Biol* 2006, **7**(8):R71.

153. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J *et al*: **Bioconductor: open software development for computational biology and bioinformatics**. *Genome Biol* 2004, **5**(10):R80.

154. Langmead B, Trapnell C, Pop M, Salzberg SL: **Ultrafast and memory-efficient alignment of short DNA sequences to the human genome**. *Genome Biol* 2009, **10**(3):R25.

155. Giardine B, Riemer C, Hardison RC, Burhans R, Elnitski L, Shah P, Zhang Y, Blankenberg D, Albert I, Taylor J *et al*: **Galaxy: a platform for interactive large-scale genome analysis**. *Genome Res* 2005, **15**(10):1451-1455.

156. Xu H, Handoko L, Wei X, Ye C, Sheng J, Wei CL, Lin F, Sung WK: **A signal-noise model for significance analysis of ChIP-seq with negative control**. *Bioinformatics* 2010, **26**(9):1199-1204.

157. Chi P, Chen Y, Zhang L, Guo X, Wongvipat J, Shamu T, Fletcher JA, Dewell S, Maki RG, Zheng D *et al*: **ETV1 is a lineage survival factor that cooperates with KIT in gastrointestinal stromal tumours**. *Nature* 2010, **467**(7317):849-853.

158. Flicek P, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fairley S, Fitzgerald S *et al*: **Ensembl 2012**. *Nucleic Acids Res* 2012, **40**(Database issue):D84-90.

159. Pavesi G, Mereghetti P, Mauri G, Pesole G: **Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes**. *Nucleic Acids Res* 2004, **32**(Web Server issue):W199-203.

160. Mahony S, Benos PV: **STAMP: a web tool for exploring DNA-binding motif similarities**. *Nucleic Acids Res* 2007, **35**(Web Server issue):W253-258.

161. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES *et al*: **Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles**. *Proc Natl Acad Sci U S A* 2005, **102**(43):15545-15550.

162. Basso K, Margolin AA, Stolovitzky G, Klein U, Dalla-Favera R, Califano A: **Reverse engineering of regulatory networks in human B cells**. *Nat Genet* 2005, **37**(4):382-390.

163. Margolin AA, Wang K, Lim WK, Kustagi M, Nemenman I, Califano A: **Reverse engineering cellular networks**. *Nat Protoc* 2006, **1**(2):662-671.

164. Reich M, Liefeld T, Gould J, Lerner J, Tamayo P, Mesirov JP: **GenePattern 2.0**. *Nat Genet* 2006, **38**(5):500-501.

165. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD: **The sva package for removing batch effects and other unwanted variation in high-throughput experiments**. *Bioinformatics* 2012, **28**(6):882-883.

166. Presneau N, Shalaby A, Idowu B, Gikas P, Cannon SR, Gout I, Diss T, Tirabosco R, Flanagan AM: **Potential therapeutic targets for chordoma: PI3K/AKT/TSC1/TSC2/mTOR pathway**. *Br J Cancer* 2009, **100**(9):1406-1414.

167. Liu X, Nielsen GP, Rosenberg AE, Waterman PR, Yang W, Choy E, Sassi S, Yang S, Harmon DC, Yang C *et al*: **Establishment and characterization of a novel chordoma cell line: CH22**. *J Orthop Res* 2012.

168. Rinner B, Froehlich EV, Buerger K, Knausz H, Lohberger B, Scheipl S, Fischer C, Leithner A, Guelly C, Trajanoski S *et al*: **Establishment and detailed functional and molecular genetic characterisation of a novel sacral chordoma cell line, MUG-Chor1**. *Int J Oncol* 2012, **40**(2):443-451.

169. Hsu W, Mohyeldin A, Shah SR, ap Rhys CM, Johnson LF, Sedora-Roman NI, Kosztowski TA, Awad OA, McCarthy EF, Loeb DM *et al*: **Generation of chordoma cell line JHC7 and the identification of Brachyury as a novel molecular target**. *J Neurosurg* 2011, **115**(4):760-769.

170. DeComas AM, Penfornis P, Harris MR, Meyer MS, Pochampally RR: **Derivation and characterization of an extra-axial chordoma cell line**

(EACH-1) from a scapular tumor. *J Bone Joint Surg Am* 2010, **92**(5):1231-1240.

171. Fire A, Xu S, Montgomery MK, Kostas SA, Driver SE, Mello CC: **Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans**. *Nature* 1998, **391**(6669):806-811.

172. Bergler W, Petroianu G, Schadel A: **Feasibility of proliferation studies using the BrdU and MTT assays with a head and neck carcinoma cell line**. *ORL J Otorhinolaryngol Relat Spec* 1993, **55**(4):230-235.

173. Roussos ET, Condeelis JS, Patsialou A: **Chemotaxis in cancer**. *Nat Rev Cancer* 2011, **11**(8):573-587.

174. Esmaeil Shalaby AA: **Molecular analysis of chordomas and identification of therapeutic targets**. London: UCL (University College London)
2010.

175. Rieder CL, Maiato H: **Stuck in Division or Passing through: What Happens When Cells Cannot Satisfy the Spindle Assembly Checkpoint**. *Developmental cell* 2004, **7**(5):637-651.

176. De Souza CP, Hashmi SB, Yang X, Osmani SA: **Regulated inactivation of the spindle assembly checkpoint without functional mitotic spindles**. *EMBO J* 2011, **30**(13):2648-2661.

177. Yamada KM: **Cell Adhesion**. In: *Current Protocols in Cell Biology.* John Wiley & Sons, Inc.; 2001.

178. Hatta K, Takeichi M: **Expression of N-cadherin adhesion molecules associated with early morphogenetic events in chick development**. *Nature* 1986, **320**(6061):447-449.

179. Hood JD, Cheresh DA: **Role of integrins in cell invasion and migration**. *Nat Rev Cancer* 2002, **2**(2):91-100.

180. Alam N, Goel HL, Zarif MJ, Butterfield JE, Perkins HM, Sansoucy BG, Sawyer TK, Languino LR: **The integrin—growth factor receptor duet**. *Journal of Cellular Physiology* 2007, **213**(3):649-653.

181. Orlando V: **Mapping chromosomal proteins in vivo by formaldehyde-crosslinked-chromatin immunoprecipitation**. *Trends in biochemical sciences* 2000, **25**(3):99-104.

182. Massie CE, Mills IG: **ChIPping away at gene regulation**. *EMBO Rep* 2008, **9**(4):337-343.

183. The EPC: **A User's Guide to the Encyclopedia of DNA Elements (ENCODE)**. *PLoS Biol* 2011, **9**(4):e1001046.

184. Sandelin A, Alkema W, Engström P, Wasserman WW, Lenhard B: **JASPAR: an open-access database for eukaryotic transcription factor binding profiles**. *Nucleic Acids Res* 2004, **32**(Database issue):D91-94.

185. Matys V, Fricke E, Geffers R, Gössling E, Haubrock M, Hehl R, Hornischer K, Karas D, Kel AE, Kel-Margoulis OV *et al*: **TRANSFAC: transcriptional regulation, from patterns to profiles**. *Nucleic Acids Res* 2003, **31**(1):374-378.

186. Farnham PJ: **Insights from genomic profiling of transcription factors**. *Nat Rev Genet* 2009, **10**(9):605-616.

187. Carroll JS, Meyer CA, Song J, Li W, Geistlinger TR, Eeckhoute J, Brodsky AS, Keeton EK, Fertuck KC, Hall GF *et al*: **Genome-wide analysis of estrogen receptor binding sites**. *Nat Genet* 2006, **38**(11):1289-1297.

188. Visel A, Prabhakar S, Akiyama JA, Shoukry M, Lewis KD, Holt A, Plajzer-Frick I, Afzal V, Rubin EM, Pennacchio LA: **Ultraconservation identifies a small subset of extremely constrained developmental enhancers**. *Nat Genet* 2008, **40**(2):158-160.

189. Odom DT, Dowell RD, Jacobsen ES, Gordon W, Danford TW, MacIsaac KD, Rolfe PA, Conboy CM, Gifford DK, Fraenkel E: **Tissue-specific transcriptional regulation has diverged significantly between human and mouse**. *Nat Genet* 2007, **39**(6):730-732.

190. Wilson MD, Odom DT: **Evolution of transcriptional control in mammals**. *Curr Opin Genet Dev* 2009, **19**(6):579-585.

191. Schmidt D, Wilson MD, Ballester B, Schwalie PC, Brown GD, Marshall A, Kutter C, Watt S, Martinez-Jimenez CP, Mackay S *et al*: **Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding**. *Science* 2010, **328**(5981):1036-1040.

192. Wilson MD, Barbosa-Morais NL, Schmidt D, Conboy CM, Vanes L, Tybulewicz VLJ, Fisher EMC, Tavaré S, Odom DT: **Species-Specific Transcription in Mice Carrying Human Chromosome 21**. *Science* 2008, **322**(5900):434-438.

193. Kasowski M, Grubert F, Heffelfinger C, Hariharan M, Asabere A, Waszak SM, Habegger L, Rozowsky J, Shi M, Urban AE *et al*: **Variation in transcription factor binding among humans**. *Science* 2010, **328**(5975):232-235.

194. Johnson R, Richter N, Bogu GK, Bhinge A, Teng SW, Choo SH, Andrieux LO, de Benedictis C, Jauch R, Stanton LW: **A genome-wide screen for genetic variants that modify the recruitment of REST to its target genes**. *PLoS Genet* 2012, **8**(4):e1002624.

195. Habets PE, Moorman AF, Clout DE, van Roon MA, Lingbeek M, van Lohuizen M, Campione M, Christoffels VM: **Cooperative action of Tbx2 and Nkx2.5 inhibits ANF expression in the atrioventricular canal: implications for cardiac chamber formation**. *Genes Dev* 2002, **16**(10):1234-1246.

196. Hiroi Y, Kudoh S, Monzen K, Ikeda Y, Yazaki Y, Nagai R, Komuro I: **Tbx5 associates with Nkx2-5 and synergistically promotes cardiomyocyte differentiation**. *Nat Genet* 2001, **28**(3):276-280.

197. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdottir H, Tamayo P, Mesirov JP: **Molecular signatures database (MSigDB) 3.0**. *Bioinformatics* 2011, **27**(12):1739-1740.

198. Fernando RI, Castillo MD, Litzinger M, Hamilton DH, Palena C: **IL-8 signaling plays a critical role in the epithelial-mesenchymal transition of human carcinoma cells**. *Cancer Res* 2011, **71**(15):5296-5306.

199. Matsunaga T, Wada Y, Endo S, Soda M, El-Kabbani O, Hara A: **Aldo-Keto Reductase 1B10 and Its Role in Proliferation Capacity of Drug-Resistant Cancers**. *Front Pharmacol* 2012, **3**:5.

200. Yamaguchi H, Tanaka K, Kitagawa Y, Miki K: **A PEA3 site flanked by SP1, SP4, and GATA sites positively regulates the differentiation-dependent expression of Brachyury in embryonal carcinoma P19 cells**. *Biochem Biophys Res Commun* 1999, **254**(3):542-547.

201. Budry L, Couture C, Balsalobre A, Drouin J: **The Ets factor Etv1 interacts with Tpit protein for pituitary pro-opiomelanocortin (POMC) gene transcription**. *J Biol Chem* 2011, **286**(28):25387-25396.

202. Fujikawa T, Takatori N, Kuwajima M, Kim GJ, Nishida H: **Tissue-specific regulation of the number of cell division rounds by inductive cell interaction and transcription factors during ascidian embryogenesis**. *Dev Biol* 2011, **355**(2):313-323.

203. Jacobs JJ, Keblusek P, Robanus-Maandag E, Kristel P, Lingbeek M, Nederlof PM, van Welsem T, van de Vijver MJ, Koh EY, Daley GQ *et al*: **Senescence bypass screen identifies TBX2, which represses Cdkn2a (p19(ARF)) and is amplified in a subset of human breast cancers**. *Nat Genet* 2000, **26**(3):291-299.

204. Raemaekers T, Ribbeck K, Beaudouin J, Annaert W, Van Camp M, Stockmans I, Smets N, Bouillon R, Ellenberg J, Carmeliet G: **NuSAP, a novel microtubule-associated protein involved in mitotic spindle organization**. *J Cell Biol* 2003, **162**(6):1017-1029.

205. Hussain S, Benavente SB, Nascimento E, Dragoni I, Kurowski A, Gillich A, Humphreys P, Frye M: **The nucleolar RNA methyltransferase Misu (NSun2) is required for mitotic spindle stability**. *J Cell Biol* 2009, **186**(1):27-40.

206. Kleinschmidt MA, Wagner TU, Liedtke D, Spahr S, Samans B, Gaubatz S: **lin9 is required for mitosis and cell survival during early zebrafish development**. *J Biol Chem* 2009, **284**(19):13119-13127.

207. Fujiwara T, Harigae H, Okitsu Y, Takahashi S, Yokoyama H, Yamada MF, Ishizawa K, Kameoka J, Kaku M, Sasaki T: **Expression analyses and transcriptional regulation of mouse nucleolar spindle-associated protein gene in erythroid cells: essential role of NF-Y**. *Br J Haematol* 2006, **135**(4):583-590.

208. Vanden Bosch A, Raemaekers T, Denayer S, Torrekens S, Smets N, Moermans K, Dewerchin M, Carmeliet P, Carmeliet G: **NuSAP is essential for chromatin-induced spindle formation during early embryogenesis**. *J Cell Sci* 2010, **123**(Pt 19):3244-3255.

209. Ryu B, Kim DS, Deluca AM, Alani RM: **Comprehensive expression profiling of tumor cell lines identifies molecular signatures of melanoma progression**. *PLoS One* 2007, **2**(7):e594.

210. Marie SK, Okamoto OK, Uno M, Hasegawa AP, Oba-Shinjo SM, Cohen T, Camargo AA, Kosoy A, Carlotti CG, Jr., Toledo S *et al*: **Maternal embryonic leucine zipper kinase transcript abundance correlates with malignancy grade in human astrocytomas**. *Int J Cancer* 2008, **122**(4):807-815.

211. Lauss M, Kriegner A, Vierlinger K, Visne I, Yildiz A, Dilaveroglu E, Noehammer C: **Consensus genes of the literature to predict breast cancer recurrence**. *Breast Cancer Res Treat* 2008, **110**(2):235-244.

212. Homminga I, Pieters R, Langerak AW, de Rooi JJ, Stubbs A, Verstegen M, Vuerhard M, Buijs-Gladdines J, Kooi C, Klous P *et al*: **Integrated transcript and genome analyses reveal NKX2-1 and MEF2C as potential oncogenes in T cell acute lymphoblastic leukemia**. *Cancer Cell* 2011, **19**(4):484-497.

213. Lewis TB, Robison JE, Bastien R, Milash B, Boucher K, Samlowski WE, Leachman SA, Dirk Noyes R, Wittwer CT, Perreard L *et al*: **Molecular classification of melanoma using real-time quantitative reverse transcriptase-polymerase chain reaction**. *Cancer* 2005, **104**(8):1678-1686.

214. Vance KW, Shaw HM, Rodriguez M, Ott S, Goding CR: **The retinoblastoma protein modulates Tbx2 functional specificity**. *Mol Biol Cell* 2010, **21**(15):2770-2779.

215. Baker DJ, Jin F, Jeganathan KB, van Deursen JM: **Whole chromosome instability caused by Bub1 insufficiency drives tumorigenesis through tumor suppressor gene loss of heterozygosity**. *Cancer Cell* 2009, **16**(6):475-486.

216. Erwin WM: **The Notochord, Notochordal cell and CTGF/CCN-2: ongoing activity from development through maturation**. *J Cell Commun Signal* 2008, **2**(3-4):59-65.

217. Yu X, Miyamoto S, Mekada E: **Integrin alpha 2 beta 1-dependent EGF receptor activation at cell-cell contact sites**. *J Cell Sci* 2000, **113 (Pt 12)**:2139-2147.

218. Dhar A, Ray A: **The CCN family proteins in carcinogenesis**. *Exp Oncol* 2010, **32**(1):2-9.

219. Yamaguchi TP, Takada S, Yoshikawa Y, Wu N, McMahon AP: **T (Brachyury) is a direct target of Wnt3a during paraxial mesoderm specification**. *Genes Dev* 1999, **13**(24):3185-3190.

220. Sleijfer S, Wagner AJ: **The challenge of choosing appropriate end points in single-arm phase II studies of rare diseases**. *J Clin Oncol* 2012, **30**(9):896-898.

221. Dei Tos AP: **Unveiling the molecular pathogenesis of chordoma: a new paradigm for molecular targeting of rare cancers**. *J Pathol* 2011, **223**(5):565-566.

222. Frazer KA, Murray SS, Schork NJ, Topol EJ: **Human genetic variation and its contribution to complex traits**. *Nat Rev Genet* 2009, **10**(4):241-251.

223. Lehner B: **Molecular mechanisms of epistasis within and between genes**. *Trends Genet* 2011, **27**(8):323-331.

224. Bell CG, Beck S: **Advances in the identification and analysis of allele-specific expression**. *Genome Med* 2009, **1**(5):56.

225. **The International HapMap Project**. *Nature* 2003, **426**(6968):789-796.

226. Buchanan CC, Torstenson ES, Bush WS, Ritchie MD: **A comparison of cataloged variation between International HapMap Consortium and 1000 Genomes Project data**. *J Am Med Inform Assoc* 2012, **19**(2):289-294.

227. **Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls**. *Nature* 2007, **447**(7145):661-678.

228. Hindorff La, Sethupathy P, Junkins Ha, Ramos EM, Mehta JP, Collins FS, Manolio Ta: **Potential etiologic and functional implications of genome-wide association loci for human diseases and traits.** *Proceedings of the National Academy of Sciences of the United States of America* 2009, **106**:9362-9367.

229. Maher B: **Personal genomes: The case of the missing heritability**. *Nature* 2008, **456**(7218):18-21.

230. Bodmer W: **Familial adenomatous polyposis (FAP) and its gene, APC**. *Cytogenet Cell Genet* 1999, **86**(2):99-104.

231. Fearnhead NS, Wilding JL, Winney B, Tonks S, Bartlett S, Bicknell DC, Tomlinson IP, Mortensen NJ, Bodmer WF: **Multiple rare variants in different genes account for multifactorial inherited susceptibility to colorectal adenomas**. *Proceedings of the National Academy of Sciences of the United States of America* 2004, **101**(45):15992-15997.

232. Cirulli ET, Goldstein DB: **Uncovering the roles of rare variants in common disease through whole-genome sequencing**. *Nature reviews Genetics* 2010, **11**(6):415-425.

233. McCarroll SA, Altshuler DM: **Copy-number variation and association studies of human disease**. *Nat Genet* 2007, **39**(7 Suppl):S37-42.

234. Shlien A, Malkin D: **Copy number variations and cancer**. *Genome Med* 2009, **1**(6):62.

235. Craddock N, Hurles ME, Cardin N, Pearson RD, Plagnol V, Robson S, Vukcevic D, Barnes C, Conrad DF, Giannoulatou E *et al*: **Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls**. *Nature* 2010, **464**(7289):713-720.

236. McCarroll SA: **Copy number variation and human genome maps**. *Nat Genet* 2010, **42**(5):365-366.

237. Alkan C, Coe BP, Eichler EE: **Genome structural variation discovery and genotyping**. *Nat Rev Genet* 2011, **12**(5):363-376.

238. Wren JD, Forgacs E, Fondon JW, Pertsemlidis a, Cheng SY, Gallardo T, Williams RS, Shohet RV, Minna JD, Garner HR: **Repeat polymorphisms within gene regions: phenotypic and evolutionary implications.** *American journal of human genetics* 2000, **67**:345-356.

239. Erkko H, Xia B, Nikkila J, Schleutker J, Syrjakoski K, Mannermaa A, Kallioniemi A, Pylkas K, Karppinen S-M, Rapakko K *et al*: **A recurrent mutation in PALB2 in Finnish cancer families**. *Nature* 2007, **446**(7133):316-319.

240. Tischkowitz M, Xia B, Sabbaghian N, Reis-Filho JS, Hamel N, Li G, van Beers EH, Li L, Khalil T, Quenneville LA *et al*: **Analysis of PALB2/FANCN-associated breast cancer families**. *Proc Natl Acad Sci U S A* 2007, **104**(16):6788-6793.

241. Rahman N, Seal S, Thompson D, Kelly P, Renwick A, Elliott A, Reid S, Spanova K, Barfoot R, Chagtai T *et al*: **PALB2, which encodes a BRCA2-interacting protein, is a breast cancer susceptibility gene**. *Nat Genet* 2007, **39**(2):165-167.

242. Jones S, Hruban RH, Kamiyama M, Borges M, Zhang X, Parsons DW, Lin JC, Palmisano E, Brune K, Jaffee EM *et al*: **Exomic sequencing identifies PALB2 as a pancreatic cancer susceptibility gene**. *Science* 2009, **324**(5924):5.

243. Cowper-Sal Lari R, Zhang X, Wright JB, Bailey SD, Cole MD, Eeckhoute J, Moore JH, Lupien M: **Breast cancer risk-associated SNPs modulate the affinity of chromatin for FOXA1 and alter gene expression**. *Nature genetics* 2012, **44**(11):1191-1198.

244. Schodel J, Bardella C, Sciesielski LK, Brown JM, Pugh CW, Buckle V, Tomlinson IP, Ratcliffe PJ, Mole DR: **Common genetic variants at the 11q13.3 renal cancer susceptibility locus influence binding of HIF to an enhancer of cyclin D1 expression**. *Nature genetics* 2012, **44**(4):420-425, S421-422.

245. Musgrove EA, Caldon CE, Barraclough J, Stone A, Sutherland RL: **Cyclin D as a therapeutic target in cancer**. *Nature reviews Cancer* 2011, **11**(8):558-572.

246. Freedman ML, Monteiro AN, Gayther SA, Coetzee GA, Risch A, Plass C, Casey G, De Biasi M, Carlson C, Duggan D *et al*: **Principles for the post-GWAS functional characterization of cancer risk loci**. *Nat Genet* 2011, **43**(6):513-518.

247. McCarroll SA, Altshuler DM: **Copy-number variation and association studies of human disease**. *Nat Genet* 2007.

248. McCarroll SA, Huett A, Kuballa P, Chilewski SD, Landry A, Goyette P, Zody MC, Hall JL, Brant SR, Cho JH *et al*: **Deletion polymorphism upstream of IRGM associated with altered IRGM expression and Crohn's disease**. *Nature genetics* 2008, **40**(9):1107-1112.

249. De Cid R, Riveira-Munoz E, Zeeuwen PLJM, Robarge J, Liao W, Dannhauser EN, Giardina E, Stuart PE, Nair R, Helms C *et al*: **Deletion of the late cornified envelope LCE3B and LCE3C genes as a susceptibility factor for psoriasis**. *Nature genetics* 2009, **41**(2):211-215.

250. Nguyen DQ, Webber C, Hehir-Kwa J, Pfundt R, Veltman J, Ponting CP: **Reduced purifying selection prevails over positive selection in human copy number variant evolution**. *Genome Research* 2008, **18**(11):1711-1723.

251. Petrij-Bosch A, Peelen T, Van Vliet M, Van Eijk R, Olmer R, Drüsedau M, Hogervorst FBL, Hageman S, Arts PJW, Ligtenberg MJL *et al*: **BRCA1 genomic deletions are major founder mutations in Dutch breast cancer patients**. *Nature genetics* 1997, **17**(3):341-345.

252. Richards FM, Phipps ME, Latif F, Yao M, Crossey PA, Foster K, Linehan WM, Affara NA, Lerman MI, Zbar B *et al*: **Mapping the Von Hippel - Lindau disease tumour suppressor gene: Identification of germline deletions by pulsed field gel electrophoresis**. *Human Molecular Genetics* 1993, **2**(7):879-882.

253. Michils G, Tejpar S, Thoelen R, Van Cutsem E, Vermeesch JR, Fryns JP, Legius E, Matthijs G: **Large deletions of the APC gene in 15% of mutation-negative patients with classical polyposis (FAP): A Belgian study**. *Human mutation* 2005, **25**(2):125-134.

254. Lichtenstein P, Holm NV, Verkasalo PK, Iliadou A, Kaprio J, Koskenvuo M, Pukkala E, Skytthe A, Hemminki K: **Environmental and heritable factors in the causation of cancer--analyses of cohorts of twins from Sweden, Denmark, and Finland**. *N Engl J Med* 2000, **343**(2):78-85.

255. Satagopan JM, Offit K, Foulkes W, Robson ME, Wacholder S, Eng CM, Karp SE, Begg CB: **The lifetime risks of breast cancer in Ashkenazi Jewish carriers of BRCA1 and BRCA2 mutations**. *Cancer Epidemiol Biomarkers Prev* 2001, **10**(5):467-473.

256. Groden J, Thliveris A, Samowitz W, Carlson M, Gelbert L, Albertsen H, Joslyn G, Stevens J, Spirio L, Robertson M *et al*: **Identification and characterization of the familial adenomatous polyposis coli gene**. *Cell* 1991, **66**(3):589-600.

257. Chung CC, Magalhaes WC, Gonzalez-Bosquet J, Chanock SJ: **Genome-wide association studies in cancer--current and future directions**. *Carcinogenesis* 2010, **31**(1):111-120.

258. Easton DF, Eeles RA: **Genome-wide association studies in cancer**. *Hum Mol Genet* 2008, **17**(R2):R109-115.

259. Rapley EA, Turnbull C, Al Olama AA, Dermitzakis ET, Linger R, Huddart RA, Renwick A, Hughes D, Hines S, Seal S *et al*: **A genome-wide association study of testicular germ cell tumor**. *Nat Genet* 2009, **41**(7):807-810.

260. Gudmundsson J, Sulem P, Gudbjartsson DF, Jonasson JG, Sigurdsson A, Bergthorsson JT, He H, Blondal T, Geller F, Jakobsdottir M *et al*:

Common variants on 9q22.33 and 14q13.3 predispose to thyroid cancer in European populations. *Nat Genet* 2009, **41**(4):460-464.

261. Gibson G: **Rare and common variants: twenty arguments**. *Nat Rev Genet* 2011, **13**(2):135-145.

262. Ritchie MD: **Using biological knowledge to uncover the mystery in the search for epistasis in genome-wide association studies**. *Ann Hum Genet* 2011, **75**(1):172-182.

263. Usdin K: **The biological effects of simple tandem repeats: Lessons from the repeat expansion diseases**. *Genome Research* 2008, **18**(7):1011-1019.

264. Bemis LT, Chen R, Amato CM, Classen EH, Robinson SE, Coffey DG, Erickson PF, Shellman YG, Robinson WA: **MicroRNA-137 Targets Microphthalmia-Associated Transcription Factor in Melanoma Cell Lines**. *Cancer research* 2008, **68**(5):1362-1368.

265. Lancaster CA, Peat N, Duhig T, Wilson D, Taylor-Papadimitriou J, Gendler SJ: **Structure and expression of the human polymorphic epithelial mucin gene: an expressed VNTR unit**. *Biochem Biophys Res Commun* 1990, **173**(3):1019-1029.

266. Thorisson GA, Smith AV, Krishnan L, Stein LD: **The International HapMap Project Web site**. *Genome Res* 2005, **15**(11):1592-1593.

267. Johnson AD, Handsaker RE, Pulit SL, Nizzari MM, O'Donnell CJ, de Bakker PI: **SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap**. *Bioinformatics* 2008, **24**(24):2938-2939.

268. Tuch BB, Laborde RR, Xu X, Gu J, Chung CB, Monighetti CK, Stanley SJ, Olsen KD, Kasperbauer JL, Moore EJ *et al*: **Tumor transcriptome sequencing reveals allelic expression imbalances associated with copy number alterations**. *PLoS One* 2010, **5**(2):e9317.

269. Hinds DA, Kloek AP, Jen M, Chen X, Frazer KA: **Common deletions and SNPs are in linkage disequilibrium in the human genome**. *Nat Genet* 2006, **38**(1):82-85.

270. Latinkic BV, Umbhauer M, Neal Ka, Lerchner W, Smith JC, Cunliffe V: **The Xenopus Brachyury promoter is activated by FGF and low concentrations ofactivinandsuppressed by high concentrationsof activin and by paired-type homeodomain proteins**. *Genes & Development* 1997, **11**:3265-3276.

271. Schulte-Merker S, Smith JC: **Mesoderm formation in response to Brachyury requires FGF signalling**. *Curr Biol* 1995, **5**(1):62-67.

272. Cho MSEA: **DeltaNp63 transcriptionally regulates brachyury, a gene with diverse roles in limb development , tumorigenesis and metastasis**. *Cell Cycle* 2010, **9**:1-8.

273. Crum CP, Mckeon FD: **p63 in Epithelial Survival , Germ Cell Surveillance , and Neoplasia**. *Neoplasia*:349-373.

274. Tomkova K, Tomka M, Zajac V: **Contribution of p53 , p63 , and p73 to the developmental diseases and cancer**. *International Journal of Cancer* 2008:177-181.

275. Biase DD, Morandi L, Esposti RD, Ligorio C, Pession A, Foschini MP, Eusebi V: **p63 short isoforms are found in invasive carcinomas only and not in benign breast conditions**. *Cycle* 2010:395-401.

276. Mangiulli M, Valletti A, Caratozzolo MF, Tullo A, Sbisà E, Pesole G, D'Erchia AM: **Identification and functional characterization of two new transcriptional variants of the human p63 gene.** *Nucleic acids research* 2009, **37**:6092-6104.

277. Urist MJ, {Di Como} CJ, Lu M-L, Charytonowicz E, Verbel D, Crum CP, Ince Ta, McKeon FD, Cordon-Cardo C: **Loss of p63 expression is associated with tumor progression in bladder cancer.** *The American journal of pathology* 2002, **161**:1199-1206.

278. Koga F, Kawakami S, Kumagai J, Takizawa T, Ando N, Arai G, Kageyama Y, Kihara K: **Impaired Delta Np63 expression associates with reduced beta-catenin and aggressive phenotypes of urothelial neoplasms.** *British journal of cancer* 2003, **88**:740-747.

279. Liu C, Nakamura E, Knezevic V, Hunter S, Thompson K, Mackem S: **A role for the mesenchymal T-box gene Brachyury in AER formation during limb development**. *Development* 2003, **130**(7):1327-1337.

280. Lin Y-l, Sengupta S, Gurdziel K, Bell GW, Jacks T, Flores ER: **p63 and p73 Transcriptionally Regulate Genes Involved in DNA Repair**. *DNA Repair* 2009, **5**:1-13.

281. Forest-divonne D, Aberdam D, Rouleau M: **DNp63 Is Essential for Epidermal Commitment of Embryonic Stem Cells**. *October* 2008, **3**.

282. Thurfjell N, Coates PJ, Boldrup L, Lindgren B, Backlund B, Uusitalo T, Mahani D, Dabelsteen E, Dahlqvist A, Sjostrom B *et al*: **Function and importance of p63 in normal oral mucosa and squamous cell carcinoma of the head and neck**. *Adv Otorhinolaryngol* 2005, **62**:49-57.

283. Kendall J, Liu Q, Bakleh A, Krasnitz A, Nguyen KC, Lakshmi B, Gerald WL, Powers S, Mu D: **Oncogenic cooperation and coamplification of developmental transcription factor genes in lung cancer**. *Proc Natl Acad Sci U S A* 2007, **104**(42):16663-16668.

284. Fisher C: **Parachordoma exists--but what is it?** *Adv Anat Pathol* 2000, **7**(3):141-148.

285. Rekhi B, Amare P, Gulia A, Baisane C, Patil A, Agarwal S, Puri A, Jambhekar NA: **Primary intraosseous myoepithelioma arising in the iliac bone and displaying trisomies of 11, 15, 17 with del (16q) and del (22q11)--A rare case report with review of literature**. *Pathol Res Pract* 2011, **207**(12):780-785.

286. Jo VY, Fletcher CD: **p63 immunohistochemical staining is limited in soft tissue tumors**. *Am J Clin Pathol* 2011, **136**(5):762-766.

287. Kallen M, Sanders M, Gonzalez A, Black J, Keedy V, Hande K, Homlar K, Halpern J, Holt G, Schwartz H *et al*: **Nuclear p63 expression in osteoblastic tumors**. *Tumor Biology*:1-6.

Appendix

Gene list: Integration of GEM and ChIP-seq datasets

Ensembl_Gene_ID	Gene_name	logFC	AveExpr	t	P.Value	Class	Column1	Column2
ENSG00000061492	WNT8A	2.539668	7.0121769	21.13721	3.25E-07	none		
ENSG00000184408	KCND2	2.333325	7.4717702	17.04284	1.28E-06	enhancer_only		
ENSG00000168530	MYL1	2.088722	8.8399043	15.69395	2.16E-06	none		
ENSG00000083782	EPYC	3.027993	8.81604	15.00976	2.86E-06	promoter_and_enhancer		
ENSG00000136237	RAPGEF5	1.725944	7.3919893	13.31006	6.10E-06	promoter_and_enhancer		
ENSG00000163106	HPGDS	3.103206	8.6153695	13.16593	6.53E-06	enhancer_only		
ENSG00000119125	GDA	2.296694	8.7270224	12.19511	1.05E-05	enhancer_only		
ENSG00000164458	T	1.725551	9.3644373	11.76596	1.32E-05	enhancer_only		
ENSG00000147724	FAM135B	1.456009	6.3436708	11.64971	1.40E-05	enhancer_only		
ENSG00000116983	HPCAL4	1.510819	6.8195878	11.55065	1.48E-05	enhancer_only		
ENSG00000169213	RAB3B	1.801501	9.6708702	11.45903	1.55E-05	enhancer_only		
ENSG00000149573	MPZL2	1.351696	8.3830089	10.53883	2.60E-05	none		
ENSG00000188991	SLC15A5	2.29457	8.4339	10.51229	2.64E-05	promoter_and_enhancer		
ENSG00000111247	RAD51AP1	1.311603	7.494889	10.21505	3.15E-05	none		
ENSG00000115339	GALNT3	1.245073	8.7714864	10.19487	3.19E-05	promoter_and_enhancer		
ENSG00000082482	KCNK2	1.609324	8.0722687	10.00286	3.59E-05	enhancer_only		

ENSG00000142149	HUNK	1.592568	8.8476353	9.969765	3.66E-05	enhancer_only
ENSG00000168329	CX3CR1	2.082249	7.7313561	9.90636	3.80E-05	enhancer_only
ENSG00000073756	PTGS2	1.550378	8.2438394	9.792944	4.08E-05	none
ENSG00000227167		1.355215	4.8626886	9.682778	4.37E-05	none
ENSG00000186367	KIAA1024L	2.224013	6.5831223	9.678995	4.38E-05	none
ENSG00000171243	SOSTDC1	1.298781	5.7703516	9.497935	4.92E-05	none
ENSG00000153707	PTPRD	1.184719	9.2824449	9.423024	5.16E-05	enhancer_only
ENSG00000186213	NA	2.042858	5.9917245	9.370578	5.34E-05	none
ENSG00000157766	ACAN	1.758714	8.1931906	9.356346	5.39E-05	enhancer_only
ENSG00000169429	IL8	1.264186	8.1703491	9.328206	5.49E-05	promoter_and_enhancer
ENSG00000172020	GAP43	1.241586	6.7353001	9.184077	6.03E-05	enhancer_only
ENSG00000102384	CENPI	1.374042	7.0051081	9.17848	6.06E-05	enhancer_only
ENSG00000112964	GHR	1.23345	8.2852772	9.067813	6.52E-05	enhancer_only
ENSG00000172296	SPTLC3	1.343859	9.7701011	8.96848	6.97E-05	enhancer_only
ENSG00000248286		1.042799	9.8769983	8.938576	7.11E-05	none
ENSG00000165186	PTCHD1	1.072935	7.313738	8.708785	8.32E-05	enhancer_only
ENSG00000139318	DUSP6	1.231971	6.618302	8.396322	0.0001036	none
ENSG00000228262		1.744535	7.7284493	8.388291	0.0001042	none
ENSG00000145681	HAPLN1	1.66989	9.5064645	8.381934	0.0001047	promoter_and_enhancer
ENSG00000069122	GPR116	1.721659	6.3397537	8.348249	0.0001072	promoter_and_enhancer
ENSG00000087301	TXNDC16	1.332461	6.5269907	8.32148	0.0001093	enhancer_only
ENSG00000179826	MRGPRX3	1.76379	9.5500135	8.307573	0.0001104	promoter_and_enhancer
ENSG00000171476	HOPX	0.975433	7.0988175	8.237192	0.0001161	promoter_and_enhancer
ENSG00000003987	MTMR7	2.173197	7.7995667	8.168465	0.0001221	promoter_and_enhancer
ENSG00000006468	ETV1	1.370707	8.5029449	8.153982	0.0001234	enhancer_only

ENSG00000174371	EXO1	1.005844	8.4205429	8.123301	0.0001262	none
ENSG00000231297		0.977864	5.9141922	7.984209	0.0001399	none
ENSG00000153246	PLA2R1	1.52301	8.9313276	7.964836	0.0001419	enhancer_only
ENSG00000215866		0.951072	8.305419	7.899981	0.000149	none
ENSG00000229233		1.668054	7.2274869	7.770461	0.0001643	none
ENSG00000127951	FGL2	1.757378	9.9859646	7.737489	0.0001685	none
ENSG00000040731	CDH10	1.084142	7.4972474	7.605088	0.0001865	none
ENSG00000073910	FRY	1.088835	7.4649177	7.565124	0.0001924	enhancer_only
ENSG00000135298	BAI3	0.940272	5.4487624	7.531976	0.0001975	enhancer_only
ENSG00000151789	ZNF385D	1.028848	6.8316767	7.475823	0.0002063	enhancer_only
ENSG00000147257	GPC3	1.42937	6.1742792	7.45917	0.0002091	none
ENSG00000170624	SGCD	1.314605	8.0203641	7.454003	0.0002099	promoter_and_enhancer
ENSG00000168078	PBK	1.263488	9.1510021	7.331718	0.0002313	enhancer_only
ENSG00000205809	KLRC2	1.479875	5.1956848	7.3044	0.0002364	promoter_and_enhancer
ENSG00000218336	ODZ3	1.057187	8.1457468	7.28941	0.0002393	enhancer_only
ENSG00000225298	NCRNA00113	1.474922	5.0741926	7.253297	0.0002463	none
ENSG00000156970	BUB1B	1.066138	8.9856339	7.196668	0.0002578	none
ENSG00000156140	ADAMTS3	1.379364	8.3639595	7.195075	0.0002582	none
ENSG00000160654	CD3G	1.215256	5.9572473	7.167629	0.000264	none
ENSG00000250961		1.252033	5.4379385	7.1504	0.0002677	none
ENSG00000068489	PRR11	1.209257	8.7615755	7.148146	0.0002682	none
ENSG00000102837	OLFM4	0.983941	5.8820581	7.104033	0.0002781	enhancer_only
ENSG00000143127	ITGA10	1.273262	6.6667629	7.103505	0.0002782	enhancer_only
ENSG00000163359	COL6A3	0.894269	10.082543	7.094917	0.0002801	enhancer_only
ENSG00000104432	IL7	2.078885	8.8699516	7.089327	0.0002814	enhancer_only

ENSG00000113319	RASGRF2	0.891114	8.4357822	7.078311	0.000284	enhancer_only
ENSG00000168079	SCARA5	1.546136	7.9681168	7.067537	0.0002865	enhancer_only
ENSG00000231858		1.106391	6.2449655	7.057521	0.0002889	none
ENSG00000140525	FANCI	0.929574	8.4432784	7.057482	0.0002889	none
ENSG00000134765	DSC1	1.486243	6.4538652	7.024845	0.0002968	enhancer_only
ENSG00000137804	NUSAP1	0.89904	9.7100748	7.009698	0.0003005	enhancer_only
ENSG00000108984	MAP2K6	1.132846	6.7792281	6.99639	0.0003039	none
ENSG00000171488	LRRC8C	0.849321	6.7159014	6.984685	0.0003068	promoter_and_enhancer
ENSG00000152270	PDE3B	1.115363	7.699673	6.954008	0.0003148	none
ENSG00000146426	TIAM2	0.924788	7.0729386	6.932646	0.0003204	enhancer_only
ENSG00000129514	FOXA1	0.8071	7.925142	6.919997	0.0003238	none
ENSG00000171241	SHCBP1	0.9874	8.1976969	6.918737	0.0003242	none
ENSG00000085563	ABCB1	0.947868	7.23331	6.892674	0.0003313	enhancer_only
ENSG00000144331	ZNF385B	1.02158	7.1872821	6.888537	0.0003325	promoter_and_enhancer
ENSG00000229654		1.552833	5.8692503	6.874406	0.0003364	none
ENSG00000197299	BLM	0.960994	6.3204958	6.857654	0.0003412	enhancer_only
ENSG00000166803	KIAA0101	1.110265	6.4828307	6.855084	0.000342	none
ENSG00000018236	CNTN1	1.475253	7.3012571	6.854993	0.000342	promoter_and_enhancer
ENSG00000105889		1.012005	7.0723214	6.816784	0.0003532	none
ENSG00000135678	CPM	1.338639	8.7051523	6.795958	0.0003595	enhancer_only
ENSG00000146918	NCAPG2	1.054259	7.0856369	6.789139	0.0003616	none
ENSG00000122986	HVCN1	1.900854	8.5335375	6.727345	0.0003811	enhancer_only
ENSG00000145721	LIX1	1.032949	5.3123164	6.719902	0.0003836	none
ENSG00000056277	ZNF280C	0.910111	9.3424183	6.711285	0.0003864	enhancer_only
ENSG00000173698	GPR64	1.217139	9.1518613	6.686966	0.0003946	none

ENSG00000169570	DTWD2	0.999642	9.0185952	6.68463	0.0003954	none
ENSG00000152127	MGAT5	1.118845	9.4740806	6.660705	0.0004036	enhancer_only
ENSG00000176562	NA	1.517019	5.4173387	6.659941	0.0004038	none
ENSG00000164171	ITGA2	0.89301	6.408267	6.6525	0.0004064	none
ENSG00000176734	TRIL	1.386852	7.63966	6.646148	0.0004087	none
ENSG00000247345		1.11158	4.6688832	6.641798	0.0004102	none
ENSG00000165959	CLMN	0.924717	8.7867901	6.636222	0.0004122	enhancer_only
ENSG00000109674	NEIL3	1.075366	8.4778326	6.619634	0.0004182	enhancer_only
ENSG00000222051		0.917759	3.8884189	6.619364	0.0004183	none
ENSG00000156395	SORCS3	1.147745	6.4822052	6.588247	0.0004297	enhancer_only
ENSG00000229695		1.291459	6.3418296	6.58094	0.0004324	none
ENSG00000226081	USP12PX	1.057963	4.0498134	6.562273	0.0004395	none
ENSG00000115525	ST3GAL5	0.77108	7.224935	6.550014	0.0004443	enhancer_only
ENSG00000164983	TMEM65	1.000691	8.7813061	6.520069	0.0004561	none
ENSG00000139133	ALG10	1.480688	5.2898226	6.454061	0.0004833	none
ENSG00000177822		1.171697	6.947047	6.437709	0.0004904	none
ENSG00000223181		0.807216	3.8684779	6.376685	0.0005177	none
ENSG00000003989	SLC7A2	1.028989	9.2070074	6.318683	0.0005452	none
ENSG00000079215	SLC1A3	0.910885	7.9826824	6.317227	0.000546	enhancer_only
ENSG00000070214	SLC44A1	0.804571	9.303478	6.291764	0.0005586	enhancer_only
ENSG00000168952	STXBP6	0.888005	7.2208828	6.280275	0.0005644	none
ENSG00000091409	ITGA6	0.776757	10.249855	6.250088	0.00058	enhancer_only
ENSG00000242759		0.757226	5.2671669	6.224641	0.0005935	none
ENSG00000235609		0.937005	6.8734356	6.219296	0.0005964	none
ENSG00000111859	NEDD9	0.748875	6.2854152	6.214204	0.0005992	enhancer_only

ENSG00000171320	ESCO2	1.433864	7.3755136	6.198535	0.0006078	enhancer_only
ENSG00000157456	CCNB2	0.737242	9.065518	6.195744	0.0006093	none
ENSG00000118257	NRP2	0.739765	8.9863274	6.158178	0.0006306	enhancer_only
ENSG00000152256	PDK1	0.837776	8.9370355	6.155849	0.000632	none
ENSG00000162599	NFIA	0.899731	7.8619565	6.154682	0.0006326	enhancer_only
ENSG00000243658		0.797675	4.6400318	6.13912	0.0006417	none
ENSG00000119862		0.712492	10.368128	6.134192	0.0006446	enhancer_only
ENSG00000086717	PPEF1	0.820223	8.5505951	6.133763	0.0006449	promoter_and_enhancer
ENSG00000136155	SCEL	1.532862	5.8211321	6.115378	0.0006559	enhancer_only
ENSG00000132436	FIGNL1	0.926707	7.0588698	6.098082	0.0006664	none
ENSG00000197415	VEPH1	0.920944	5.7802173	6.092319	0.0006699	none
ENSG00000140534	C15orf42	0.753174	7.3666992	6.053629	0.0006943	none
ENSG00000165304	MELK	0.868723	8.3553005	5.997615	0.0007315	none
ENSG00000105894	PTN	0.99185	6.5473143	5.988476	0.0007377	enhancer_only
ENSG00000149212	SESN3	0.734213	10.068148	5.988434	0.0007378	none
ENSG00000111696	NT5DC3	0.816848	8.2587252	5.984158	0.0007407	enhancer_only
ENSG00000077152	UBE2T	0.837846	7.7296911	5.972581	0.0007488	none
ENSG00000111181	SLC6A12	1.14226	7.5129313	5.952452	0.000763	promoter_only
ENSG00000199849		0.75936	4.3165922	5.94511	0.0007683	none
ENSG00000117868	ESYT2	1.039071	10.20129	5.935516	0.0007753	none
ENSG00000246528		1.018083	8.6152065	5.923229	0.0007843	none
ENSG00000128656	CHN1	0.71379	7.5914051	5.871761	0.0008234	enhancer_only
ENSG00000234494		0.776651	6.812741	5.847063	0.0008429	none
ENSG00000166833	NAV2	0.978937	7.7845698	5.838578	0.0008497	promoter_and_enhancer
ENSG00000251358	NA	1.062937	5.6350526	5.830442	0.0008563	none

ENSG00000120549	KIAA1217	0.770343	7.9249964	5.789744	0.0008903	promoter_and_enhancer
ENSG00000111907	TPD52L1	0.797266	8.9644726	5.780962	0.0008978	promoter_and_enhancer
ENSG00000218014		0.900434	7.3578509	5.766762	0.0009101	none
ENSG00000005187	ACSM3	1.076271	6.7530089	5.75456	0.0009209	enhancer_only
ENSG00000078596	ITM2A	1.189674	6.4240028	5.746866	0.0009277	none
ENSG00000156298	TSPAN7	1.628024	11.133197	5.746271	0.0009282	enhancer_only
ENSG00000234890		0.700698	4.0924732	5.733689	0.0009395	none
ENSG00000134247	PTGFRN	0.875451	9.4309322	5.700204	0.0009704	enhancer_only
ENSG00000198056	PRIM1	1.137634	7.8084455	5.691152	0.000979	none
ENSG00000178031	ADAMTSL1	0.741724	6.9142618	5.673451	0.0009959	promoter_and_enhancer
ENSG00000186871	ERCC6L	1.038075	6.4243062	5.669401	0.0009999	none
ENSG00000131068	DEFB118	1.059863	5.4273205	5.660064	0.001009	promoter_only
ENSG00000144554	FANCD2	0.876826	6.5366326	5.656357	0.0010126	none
ENSG00000198879	SFMBT2	0.708322	7.5316603	5.654457	0.0010145	enhancer_only
ENSG00000101746	NOL4	0.867009	5.7508035	5.637929	0.001031	enhancer_only
ENSG00000186377	CYP4X1	1.003317	5.6592114	5.63744	0.0010315	none
ENSG00000188828	GLRA4	0.887669	6.1723002	5.632547	0.0010364	enhancer_only
ENSG00000249752		0.653269	5.1697954	5.626598	0.0010425	none
ENSG00000249084		1.53235	6.43525	5.623441	0.0010457	none
ENSG00000232671		0.883654	4.5764226	5.620474	0.0010487	none
ENSG00000088836	SLC4A11	1.214613	9.2835493	5.611374	0.0010581	none
ENSG00000134376	CRB1	1.505515	5.6418343	5.595314	0.0010749	enhancer_only
ENSG00000228093		1.356997	7.3401555	5.576757	0.0010946	none
ENSG00000113263	ITK	0.747362	6.5141098	5.574022	0.0010976	enhancer_only
ENSG00000171345	KRT19	1.384672	8.8123429	5.565288	0.0011071	enhancer_only

ENSG00000253085		0.730648	3.6900032	5.561501	0.0011112	none
ENSG00000251236		0.893172	3.4157595	5.56148	0.0011112	none
ENSG00000170961	HAS2	1.12418	8.5447106	5.554957	0.0011184	enhancer_only
ENSG00000184445	KNTC1	0.961815	7.4327586	5.539772	0.0011353	none
ENSG00000222806		1.172271	6.3590292	5.536966	0.0011384	none
ENSG00000241207		0.963274	4.3145902	5.535849	0.0011397	none
ENSG00000175305	CCNE2	0.850196	7.2911467	5.481133	0.0012032	none
ENSG00000123485	HJURP	0.669981	8.1351559	5.455833	0.001234	none
ENSG00000122966	CIT	0.807253	7.7191595	5.452144	0.0012385	enhancer_only
ENSG00000163235	TGFA	1.226031	8.424087	5.449252	0.0012421	enhancer_only
ENSG00000224956		1.362735	4.8011506	5.44917	0.0012422	none
ENSG00000140937	CDH11	1.096392	9.2072861	5.428149	0.0012686	enhancer_only
ENSG00000251191	C8orf75	1.302138	6.0275631	5.412945	0.0012881	none
ENSG00000251310		0.792723	5.4305621	5.408866	0.0012934	none
ENSG00000065413	ANKRD44	0.868348	6.5629379	5.398154	0.0013074	enhancer_only
ENSG00000238342		1.37039	4.6830203	5.389249	0.0013191	enhancer_only
ENSG00000187678	SPRY4	0.859624	6.5554999	5.385721	0.0013238	none
ENSG00000235350		0.819094	7.9235599	5.371628	0.0013428	none
ENSG00000198796	ALPK2	1.025507	9.9255821	5.36921	0.0013461	promoter_and_enhancer
ENSG00000237316		0.750221	5.1900927	5.358391	0.0013608	none
ENSG00000121152	NCAPH	0.858528	7.6890834	5.356743	0.0013631	none
ENSG00000164292	RHOBTB3	0.780052	9.5112346	5.32814	0.0014032	promoter_and_enhancer
ENSG00000247556		0.645557	8.0338987	5.315701	0.001421	none
ENSG00000249174		0.746629	4.3755943	5.309901	0.0014294	none
ENSG00000099219	ERMP1	0.803091	8.6050404	5.299528	0.0014446	none

ENSG00000144406	UNC80	0.918246	6.8218119	5.280825	0.0014724	promoter_only
ENSG00000150637	CD226	1.504273	6.1088067	5.272262	0.0014853	enhancer_only
ENSG00000227929		0.968289	5.2897688	5.253744	0.0015137	none
ENSG00000182158	CREB3L2	0.751594	9.0528443	5.252493	0.0015157	promoter_and_enhancer
ENSG00000140835	CHST4	1.355177	7.651124	5.250617	0.0015186	promoter_only
ENSG00000185477	GPRIN3	0.874884	8.3948288	5.23967	0.0015357	none
ENSG00000238144		1.066023	4.7356614	5.205872	0.00159	none
ENSG00000164520	RAET1E	0.811093	6.3019904	5.187366	0.0016207	none
ENSG00000184005	ST6GALNAC3	0.722481	6.6852115	5.184704	0.0016252	enhancer_only
ENSG00000222889		0.956857	5.9417145	5.178033	0.0016364	enhancer_only
ENSG00000153162	BMP6	0.78113	9.4825336	5.16031	0.0016667	enhancer_only
ENSG00000162999	DUSP19	0.63239	5.2093521	5.150259	0.0016842	enhancer_only
ENSG00000112029	FBXO5	0.879559	7.9002611	5.148339	0.0016875	none
ENSG00000250416		0.792089	4.0279593	5.145393	0.0016927	none
ENSG00000208038	MIR492	1.033912	6.9701364	5.123648	0.0017314	none
ENSG00000163492	CCDC141	1.376414	7.6751295	5.113799	0.0017493	enhancer_only
ENSG00000164106	SCRG1	1.063185	5.3170083	5.103424	0.0017683	enhancer_only
ENSG00000166292	TMEM100	0.992647	6.8145395	5.090311	0.0017927	promoter_only
ENSG00000113448	PDE4D	0.768141	6.4122257	5.085917	0.001801	enhancer_only
ENSG00000156535	CD109	0.796988	10.723871	5.075247	0.0018212	enhancer_only
ENSG00000146263	MMS22L	0.921891	7.0180456	5.071052	0.0018292	none
ENSG00000144481	TRPM8	1.99509	7.176836	5.05737	0.0018557	enhancer_only
ENSG00000113578	FGF1	0.728752	6.513888	5.055284	0.0018598	enhancer_only
ENSG00000184349	EFNA5	0.658579	9.5148634	5.048101	0.0018739	enhancer_only
ENSG00000143476	DTL	0.778461	9.0741927	5.033241	0.0019034	enhancer_only

ENSG00000149054	ZNF215	0.833592	6.8034911	5.03101	0.0019079	none
ENSG00000154080	CHST9	1.499884	7.1906445	5.027521	0.0019149	enhancer_only
ENSG00000163032	VSNL1	0.976384	7.5030036	5.02476	0.0019205	enhancer_only
ENSG00000132849	INADL	0.699513	7.1916153	5.023158	0.0019238	enhancer_only
ENSG00000168497	SDPR	0.637965	8.0757162	5.010052	0.0019506	enhancer_only
ENSG00000227550	TRBV7-5	0.740138	5.5580904	5.002095	0.001967	none
ENSG00000223169		0.801891	4.0655148	5.001508	0.0019683	none
ENSG00000198756	GLT25D2	0.711757	7.7406265	4.994704	0.0019825	enhancer_only
ENSG00000185070	FLRT2	0.687097	9.5242379	4.981106	0.0020113	enhancer_only
ENSG00000252729		0.700193	3.5231635	4.972565	0.0020296	none
ENSG00000252867		1.314937	7.6516824	4.97167	0.0020315	none
ENSG00000235823	NCRNA00263	0.746337	7.599323	4.959652	0.0020576	none
ENSG00000038295	TLL1	0.772366	6.0082833	4.953126	0.002072	none
ENSG00000228421		1.000986	7.327109	4.949988	0.0020789	none
ENSG00000252593		0.94442	3.3188608	4.942255	0.0020961	none
ENSG00000152527	PLEKHH2	1.899534	8.7496679	4.94089	0.0020991	enhancer_only
ENSG00000174579	MSL2	0.582707	9.030351	4.936893	0.0021081	enhancer_only
ENSG00000057294	PKP2	0.817973	8.7997057	4.936592	0.0021088	enhancer_only
ENSG00000188033	ZNF490	0.732944	8.1455479	4.927063	0.0021303	enhancer_only
ENSG00000213809	KLRK1	1.145226	5.4495046	4.926774	0.002131	enhancer_only
ENSG00000174469	CNTNAP2	0.77298	5.8537989	4.923983	0.0021374	enhancer_only
ENSG00000243655		0.739188	5.2573118	4.921268	0.0021436	none
ENSG00000156103	MMP16	1.159309	7.1853654	4.92115	0.0021438	none
ENSG00000207361		0.941319	7.6871349	4.920408	0.0021456	enhancer_only
ENSG00000196569	LAMA2	1.070807	8.6025984	4.896814	0.0022004	promoter_and_enhancer

ENSG00000197977	ELOVL2	0.861327	9.0873262	4.894896	0.002205	none
ENSG00000229459		1.105279	4.2747629	4.889644	0.0022174	none
ENSG00000249539	MRPS36P2	0.612129	3.8822873	4.881776	0.0022362	none
ENSG00000173848	NET1	0.630832	8.4015066	4.879085	0.0022427	none
ENSG00000006747	SCIN	0.808022	9.6268131	4.863772	0.00228	promoter_and_enhancer
ENSG00000151834	GABRA2	1.20248	8.9649017	4.86368	0.0022802	enhancer_only
ENSG00000166845	C18orf54	1.391878	6.9319674	4.854464	0.002303	none
ENSG00000228650		0.589641	3.7234731	4.846941	0.0023217	none
ENSG00000138336	TET1	1.094976	7.0980357	4.841181	0.0023362	none
ENSG00000188921	PTPLAD2	0.711878	6.2641095	4.840991	0.0023367	none
ENSG00000114854	TNNC1	0.60014	7.0830047	4.838455	0.0023431	none
ENSG00000202175		0.711945	4.3559277	4.804626	0.0024305	none
ENSG00000198826	ARHGAP11A	1.523078	7.9574486	4.801963	0.0024375	none
ENSG00000076382	SPAG5	0.767548	8.2800669	4.798873	0.0024457	none
ENSG00000132294	EFR3A	0.588498	10.51885	4.790865	0.0024671	promoter_and_enhancer
ENSG00000147251	DOCK11	1.518559	9.3860212	4.790346	0.0024685	enhancer_only
ENSG00000156802	ATAD2	1.231408	8.3521736	4.790187	0.0024689	none
ENSG00000112992	NNT	0.588291	9.1072622	4.787718	0.0024756	enhancer_only
ENSG00000187715	KBTBD12	0.705597	5.6440673	4.78725	0.0024768	none
ENSG00000173674	EIF1AX	0.669314	9.6734605	4.781541	0.0024923	none
ENSG00000150722	PPP1R1C	0.569843	5.6352764	4.780906	0.002494	none
ENSG00000138160	KIF11	1.670247	7.6744165	4.779429	0.002498	none
ENSG00000137809	ITGA11	0.94493	7.1308425	4.771479	0.0025197	enhancer_only
ENSG00000150556	LYPD6B	0.687213	6.2044348	4.768315	0.0025284	none
ENSG00000035499	DEPDC1B	0.681669	6.3991968	4.764057	0.0025402	enhancer_only

ENSG00000171885	AQP4	0.83346	5.1331309	4.762043	0.0025458	none
ENSG00000110693	SOX6	0.561636	6.7600754	4.760857	0.0025491	enhancer_only
ENSG00000204634	TBC1D8	0.599374	8.565186	4.757721	0.0025578	promoter_only
ENSG00000233834		0.660623	5.325321	4.755583	0.0025638	none
ENSG00000166575	TMEM135	0.71559	7.2750814	4.747741	0.0025858	enhancer_only
ENSG00000183542	KLRC4	0.892237	4.5656327	4.744796	0.0025942	enhancer_only
ENSG00000044524	EPHA3	0.997343	6.5299911	4.734513	0.0026235	none
ENSG00000196761	NA	0.922382	8.0377003	4.721258	0.0026619	none
ENSG00000240703		0.574375	3.8084156	4.720339	0.0026646	none
ENSG00000137571	SLCO5A1	1.235678	7.4795581	4.712664	0.0026872	enhancer_only
ENSG00000227336		1.358362	4.952664	4.711407	0.0026909	none
ENSG00000116106	EPHA4	0.884287	8.7182184	4.706521	0.0027054	promoter_only
ENSG00000145358	DDIT4L	0.811182	7.5922508	4.703439	0.0027145	none
ENSG00000176597	B3GNT5	0.612406	7.4330403	4.70065	0.0027229	none
ENSG00000139278	GLIPR1	1.488319	9.8084859	4.687824	0.0027616	enhancer_only
ENSG00000124693	HIST1H3B	1.073748	10.009074	4.681589	0.0027806	promoter_only
ENSG00000148773	MKI67	1.065802	9.2263205	4.668865	0.0028199	none
ENSG00000114541	FRMD4B	0.593346	6.1693767	4.661532	0.0028429	enhancer_only
ENSG00000145604	SKP2	0.612229	9.643826	4.651465	0.0028747	none
ENSG00000174804	FZD4	0.689869	8.3376385	4.646373	0.002891	promoter_and_enhancer
ENSG00000188783	PRELP	0.898561	10.423062	4.645339	0.0028943	enhancer_only
ENSG00000201707		0.644315	5.9532662	4.643936	0.0028988	none
ENSG00000164761	TNFRSF11B	0.608603	10.428338	4.641418	0.0029069	enhancer_only
ENSG00000230154		1.050929	5.6052669	4.640836	0.0029088	none
ENSG00000007174	DNAH9	0.711478	6.3849365	4.64003	0.0029114	promoter_and_enhancer

ENSG00000134909	ARHGAP32	0.580825	7.8050417	4.639167	0.0029141	promoter_and_enhancer
ENSG00000133110	POSTN	1.719766	7.1057214	4.636841	0.0029217	none
ENSG00000236896		1.057803	5.6995815	4.632726	0.002935	none
ENSG00000242584		0.620163	3.8593919	4.63109	0.0029404	none
ENSG00000122884	P4HA1	0.575874	9.6730515	4.628578	0.0029486	none
ENSG00000248092		0.747483	6.8686462	4.62824	0.0029497	none
ENSG00000168291	PDHB	0.599228	8.548891	4.62388	0.002964	enhancer_only
ENSG00000168301	KCTD6	0.955264	7.0905341	4.616932	0.002987	enhancer_only
ENSG00000126010	GRPR	1.077542	6.4624441	4.61251	0.0030017	none
ENSG00000100526	CDKN3	0.740968	8.2379965	4.61229	0.0030024	none
ENSG00000058091	CDK14	0.563462	7.4424317	4.60331	0.0030326	enhancer_only
ENSG00000242309	NA	0.613954	3.2156115	4.601127	0.00304	none
ENSG00000124429	POF1B	0.898923	7.7372754	4.599248	0.0030464	none
ENSG00000158966	CACHD1	0.616127	7.7649373	4.594329	0.0030631	enhancer_only
ENSG00000176641	RNF152	0.67343	6.0948664	4.594118	0.0030638	enhancer_only
ENSG00000240660		0.598778	4.8668962	4.594062	0.003064	none
ENSG00000125247	TMTC4	0.630458	7.3323846	4.590367	0.0030767	none
ENSG00000230399		0.700823	4.8722133	4.582333	0.0031044	none
ENSG00000111837	MAK	0.947838	6.4927199	4.581266	0.0031081	enhancer_only
ENSG00000064651	SLC12A2	0.66327	8.8126027	4.57962	0.0031138	none
ENSG00000092470	WDR76	0.737927	7.3476988	4.57894	0.0031162	none
ENSG00000223522		0.651682	6.2673832	4.571215	0.0031432	none
ENSG00000068615	REEP1	0.617505	7.3966957	4.564142	0.0031682	enhancer_only
ENSG00000151623	NR3C2	0.552408	6.6134509	4.556634	0.0031949	enhancer_only
ENSG00000238694		0.650858	3.7995508	4.555381	0.0031994	none

ENSG00000149294	NCAM1	0.830863	7.6144915	4.550202	0.003218	none
ENSG00000196678	ERI2	0.680641	7.7708476	4.550195	0.0032181	enhancer_only
ENSG00000111341	MGP	0.995265	10.438716	4.550007	0.0032188	none
ENSG00000237255	NA	0.725277	4.0692256	4.543577	0.0032421	none
ENSG00000214719	LRRC37BP1	0.559245	6.3882093	4.542987	0.0032442	none
ENSG00000064042	LIMCH1	0.560426	8.1232806	4.542124	0.0032473	enhancer_only
ENSG00000238178		1.048865	6.4049476	4.540903	0.0032518	none
ENSG00000231918		0.829784	6.0680276	4.539941	0.0032553	none
ENSG00000155561	NUP205	0.564169	8.9745621	4.536417	0.0032682	none
ENSG00000182795	C1orf116	0.625882	6.6126267	4.532868	0.0032813	promoter_only
ENSG00000132938	MTUS2	0.556529	6.0417007	4.530327	0.0032907	promoter_and_enhancer
ENSG00000198074	AKR1B10	1.567271	7.6098373	4.529355	0.0032943	none
ENSG00000184564	SLITRK6	0.86774	6.0536791	4.528891	0.003296	none
ENSG00000163380	LMOD3	0.898845	5.9277831	4.527674	0.0033005	none
ENSG00000184661	CDCA2	0.952097	7.8627442	4.514019	0.0033516	none
ENSG00000238202		0.793731	4.9592579	4.50682	0.0033789	none
ENSG00000232193		0.937762	5.1422127	4.505839	0.0033827	none
ENSG00000184156	KCNQ3	0.777531	7.5966778	4.502959	0.0033937	enhancer_only
ENSG00000120756	PLS1	0.659759	6.4171233	4.500722	0.0034023	none
ENSG00000163624	CDS1	0.531878	5.8042534	4.495967	0.0034206	none
ENSG00000184825	HIST1H2AH	0.778246	5.1653768	4.495527	0.0034223	none
ENSG00000168298	HIST1H1E	0.992771	8.4550685	4.490503	0.0034418	none
ENSG00000084710	EFR3B	0.834559	7.3509953	4.48556	0.003461	enhancer_only
ENSG00000164303	ENPP6	1.530873	6.2466353	4.482785	0.0034719	enhancer_only
ENSG00000169607	CKAP2L	1.282265	7.6289298	4.476599	0.0034963	none

ENSG00000205221	VIT	0.935305	6.0271544	4.472021	0.0035145	promoter_only
ENSG00000085840	ORC1	0.581826	6.1306724	4.470444	0.0035208	enhancer_only
ENSG00000183742	MACC1	0.62908	4.5618944	4.463539	0.0035484	none
ENSG00000109654	TRIM2	0.745846	7.5417022	4.463072	0.0035503	enhancer_only
ENSG00000071991	CDH19	1.355875	7.7406747	4.456537	0.0035767	enhancer_only
ENSG00000249945		0.547359	6.7284147	4.455836	0.0035796	none
ENSG00000239864	NA	0.581304	6.7426877	4.449278	0.0036064	none
ENSG00000205133	C8orf83	0.559722	9.1340028	4.445723	0.003621	none
ENSG00000165490	C11orf82	0.647803	5.5924574	4.438678	0.0036501	none
ENSG00000166532	RIMKLB	0.804403	8.1354853	4.436788	0.0036579	none
ENSG00000205810	KLRC3	1.182814	5.0503156	4.436141	0.0036606	promoter_and_enhancer
ENSG00000243831		0.883143	5.4492065	4.431261	0.003681	none
ENSG00000163808	KIF15	1.40409	6.3384026	4.4283	0.0036935	none
ENSG00000234828		1.009402	5.2224732	4.424445	0.0037097	enhancer_only
ENSG00000207209		0.588103	3.4971115	4.419235	0.0037318	none
ENSG00000167524		0.720254	7.5312327	4.413958	0.0037544	promoter_only
ENSG00000178690	C18orf26	0.78594	6.384016	4.412308	0.0037614	none
ENSG00000229589		0.749191	5.0387074	4.401523	0.0038081	none
ENSG00000145386	CCNA2	0.63271	8.3320015	4.394991	0.0038366	enhancer_only
ENSG00000203760	CENPW	0.646065	6.9236703	4.393707	0.0038423	none
ENSG00000225647		0.585157	5.4365502	4.388814	0.0038638	none
ENSG00000119938	PPP1R3C	0.541862	10.727031	4.379349	0.003906	promoter_and_enhancer
ENSG00000129173	E2F8	0.996022	6.509478	4.378894	0.003908	none
ENSG00000142731	PLK4	1.141508	6.7966312	4.366282	0.003965	none
ENSG00000108506	INTS2	0.544712	8.678987	4.36325	0.0039788	none

ENSG00000120539	MASTL	0.523747	7.8279713	4.359274	0.003997	none
ENSG00000213463	SYNJ2BP	0.757759	8.5268855	4.357856	0.0040036	none
ENSG00000169679	BUB1	0.768432	8.5860224	4.356736	0.0040087	none
ENSG00000186908	ZDHHC17	0.797986	7.6471597	4.356218	0.0040111	enhancer_only
ENSG00000241073		0.905845	5.3813286	4.353544	0.0040235	none
ENSG00000248731		0.757602	5.3137017	4.352848	0.0040267	none
ENSG00000144354	CDCA7	0.531057	7.0409659	4.349827	0.0040407	none
ENSG00000242575		0.676967	8.2474938	4.346668	0.0040554	none
ENSG00000251548		0.876629	7.9067734	4.33666	0.0041025	none
ENSG00000102547	CAB39L	0.986028	7.0849387	4.33557	0.0041076	promoter_and_enhancer
ENSG00000145808	ADAMTS19	1.038335	7.6274801	4.334144	0.0041144	enhancer_only
ENSG00000118292	C1orf54	0.905035	6.5248319	4.330147	0.0041334	none
ENSG00000112984	KIF20A	0.688859	8.7925405	4.330138	0.0041334	none
ENSG00000185585	OLFML2A	0.72508	8.3325172	4.328314	0.0041421	enhancer_only
ENSG00000173281	PPP1R3B	0.736266	8.1366558	4.327191	0.0041475	enhancer_only
ENSG00000242996		0.660067	3.7420189	4.317662	0.0041934	none
ENSG00000197594	ENPP1	0.727533	9.074377	4.313016	0.004216	promoter_only
ENSG00000205562		0.629816	6.8207142	4.306797	0.0042464	enhancer_only
ENSG00000240693		0.96261	4.3908303	4.302372	0.0042682	none
ENSG00000144136	SLC20A1	0.49713	9.7256895	4.297576	0.004292	none
ENSG00000247448		0.891215	7.0089739	4.294336	0.0043082	none
ENSG00000198589	LRBA	0.862409	8.425343	4.292716	0.0043163	enhancer_only
ENSG00000164611	PTTG1	0.546025	9.8150564	4.289935	0.0043302	none
ENSG00000120594	PLXDC2	0.782617	9.3259175	4.288259	0.0043386	enhancer_only
ENSG00000134755	DSC2	0.587372	8.0212955	4.283896	0.0043606	none

ENSG00000163428	LRRC58	0.739608	8.2598992	4.283464	0.0043628	none
ENSG00000147459	DOCK5	0.512865	9.1000777	4.276973	0.0043958	promoter_and_enhancer
ENSG00000213853	EMP2	0.988441	8.4216783	4.274154	0.0044103	none
ENSG00000164023	SGMS2	0.639917	7.9614653	4.267518	0.0044444	enhancer_only
ENSG00000246130		0.595703	5.241945	4.264424	0.0044604	none
ENSG00000118596	SLC16A7	0.886426	8.9070443	4.260438	0.0044812	none
ENSG00000155011	DKK2	0.841827	5.6163744	4.248425	0.0045443	none
ENSG00000179331	RAB39	0.56168	5.2264989	4.247236	0.0045506	none
ENSG00000200521		0.832605	7.0060383	4.245955	0.0045574	none
ENSG00000127324	TSPAN8	0.736962	5.6010553	4.244996	0.0045625	enhancer_only
ENSG00000109107	ALDOC	0.62795	7.9226713	4.242895	0.0045737	none
ENSG00000151893	C10orf46	0.557852	7.8427204	4.242712	0.0045747	enhancer_only
ENSG00000099194	SCD	0.938634	11.038375	4.242559	0.0045755	promoter_only
ENSG00000120802	TMPO	1.127238	8.3650189	4.239077	0.0045942	promoter_and_enhancer
ENSG00000134532	SOX5	0.943007	8.8397917	4.232586	0.0046291	enhancer_only
ENSG00000150625	GPM6A	0.644207	6.2625381	4.224311	0.0046741	none
ENSG00000158352	SHROOM4	0.553924	6.8039244	4.21989	0.0046984	enhancer_only
ENSG00000136960	ENPP2	0.628175	11.213473	4.218389	0.0047066	none
ENSG00000206782		0.543513	3.1207396	4.215699	0.0047215	none
ENSG00000164265	SCGB3A2	0.553834	5.8043449	4.214801	0.0047264	none
ENSG00000241399	CD302	0.857285	9.075849	4.214208	0.0047297	none
ENSG00000233224	HIST1H2AM	0.916224	7.3525131	4.211548	0.0047445	none
ENSG00000246898		0.494501	6.4078301	4.206761	0.0047712	none
ENSG00000134330	IAH1	0.522921	8.2713975	4.197439	0.0048236	none
ENSG00000249321	OR5H5P	0.766431	4.7411088	4.190296	0.0048642	none

ENSG00000048540	LMO3	0.839088	9.4599949	4.180671	0.0049196	none
ENSG00000244138		0.918525	5.146736	4.180157	0.0049226	none
ENSG00000116667	C1orf21	0.499223	7.1226666	4.175111	0.0049519	enhancer_only
ENSG00000138767	CNOT6L	0.645684	7.045599	4.174929	0.0049529	enhancer_only
ENSG00000241103		0.494723	4.1625738	4.168611	0.0049899	none
ENSG00000169228	RAB24	0.633202	6.4924212	4.166607	0.0050017	none
ENSG00000136492	BRIP1	1.276939	7.4004766	4.164419	0.0050146	none
ENSG00000241032		0.755775	5.055304	4.158946	0.0050471	none
ENSG00000138376	BARD1	0.711295	7.5470669	4.149958	0.0051009	promoter_only
ENSG00000169255	B3GALNT1	0.627994	8.0134517	4.148424	0.0051102	none
ENSG00000183780	SLC35F3	0.545944	6.8682433	4.140363	0.0051591	enhancer_only
ENSG00000139132	FGD4	0.842977	6.9745701	4.134736	0.0051935	promoter_and_enhancer
ENSG00000133138	TBC1D8B	0.996504	7.4976916	4.133899	0.0051986	none
ENSG00000156162	DPY19L4	0.575582	8.4493189	4.131817	0.0052115	none
ENSG00000120868	APAF1	0.655578	7.5844213	4.123923	0.0052604	none
ENSG00000181092	ADIPOQ	0.561884	5.0492442	4.122957	0.0052664	none
ENSG00000106772	PRUNE2	0.626224	9.2144344	4.121554	0.0052752	promoter_and_enhancer
ENSG00000111727	HCFC2	0.844931	8.7121703	4.120719	0.0052804	none
ENSG00000248142		0.516421	7.5295648	4.117751	0.005299	none
ENSG00000235951		0.779472	4.6510514	4.116588	0.0053063	none
ENSG00000228973		0.609121	6.0015166	4.108398	0.0053582	none
ENSG00000067992	PDK3	0.599672	7.8025691	4.106281	0.0053716	none
ENSG00000185480	C12orf48	0.752909	7.4117579	4.10352	0.0053893	none
ENSG00000164038	NHEDC2	0.633445	8.2449351	4.102589	0.0053952	enhancer_only
ENSG00000151572	ANO4	0.515509	6.1976637	4.100247	0.0054103	enhancer_only

ENSG00000163251	FZD5	0.524816	7.7525105	4.094093	0.00545	enhancer_only
ENSG00000196611	MMP1	1.138997	6.2870658	4.092137	0.0054627	none
ENSG00000117151	CTBS	0.502996	8.4177869	4.088533	0.0054862	none
ENSG00000205269	TMEM170B	0.730364	7.454073	4.087695	0.0054916	promoter_only
ENSG00000213238		0.723717	3.4742599	4.081365	0.0055332	none
ENSG00000226413	OR8T1P	0.494195	5.2372789	4.071275	0.0056001	none
ENSG00000168242	HIST1H2BI	1.208705	5.9061192	4.067461	0.0056257	none
ENSG00000163297	ANTXR2	0.493416	7.9741136	4.062627	0.0056582	none
ENSG00000168772	CXXC4	0.571555	5.8347294	4.060695	0.0056713	enhancer_only
ENSG00000126787	DLGAP5	1.322057	8.2231455	4.060468	0.0056728	none
ENSG00000188691	OR56A5	0.829164	4.9627861	4.058861	0.0056837	enhancer_only
ENSG00000162409	PRKAA2	0.702747	6.8636292	4.053057	0.0057233	enhancer_only
ENSG00000135406	PRPH	0.616287	7.6488885	4.051531	0.0057337	none
ENSG00000111145	ELK3	1.191595	8.4795022	4.045602	0.0057745	none
ENSG00000137760	ALKBH8	0.667207	7.9479601	4.04447	0.0057824	enhancer_only
ENSG00000170522	ELOVL6	0.574273	6.7105695	4.041228	0.0058048	none
ENSG00000115159	GPD2	0.556229	7.7366221	4.034224	0.0058538	none
ENSG00000166881	TMEM194A	0.7833	8.6000902	4.031614	0.0058721	enhancer_only
ENSG00000112303	VNN2	0.706467	4.9562467	4.025521	0.0059151	none
ENSG00000155099	TMEM55A	0.541117	9.5760812	4.025258	0.005917	none
ENSG00000185621	LMLN	0.543983	7.4473257	4.024727	0.0059208	none
ENSG00000206739		0.894232	7.0614338	4.024259	0.0059241	none
ENSG00000197254		1.980409	5.0993092	4.017829	0.00597	none
ENSG00000225307	NA	1.054168	8.6303272	4.016921	0.0059765	none
ENSG00000198105	ZNF248	0.742541	5.995472	4.015361	0.0059877	none

ENSG00000108423	TUBD1	0.568531	7.8898362	4.014638	0.0059929	none
ENSG00000228398	HMGN2P25	0.923493	5.5193619	4.014616	0.0059931	none
ENSG00000118523	CTGF	1.026658	9.7866435	4.013701	0.0059997	promoter_and_enhancer
ENSG00000207407		0.874327	4.5484059	4.013463	0.0060014	none
ENSG00000248846		0.782451	6.5730873	4.01228	0.0060099	none
ENSG00000252336		0.621246	6.2925158	3.997931	0.0061145	none
ENSG00000128606	LRRC17	0.577331	5.9285489	3.997424	0.0061183	none
ENSG00000198203	SULT1C2	0.817616	6.8879011	3.997181	0.0061201	none
ENSG00000134901	KDELC1	0.828403	6.7415258	3.990656	0.0061684	none
ENSG00000211798		0.509586	5.4119447	3.987734	0.0061901	none
ENSG00000222493		1.269902	4.7799628	3.987359	0.0061929	none
ENSG00000134748	PRPF38A	0.499433	9.058127	3.986368	0.0062003	enhancer_only
ENSG00000196950	SLC39A10	0.586449	9.1544836	3.983477	0.006222	enhancer_only
ENSG00000172292	LASS6	0.489318	9.9960028	3.983463	0.0062221	enhancer_only
ENSG00000088387	DOCK9	0.821313	8.1539596	3.981103	0.0062398	enhancer_only
ENSG00000104497	SNX16	0.778632	8.1616329	3.976947	0.0062712	none
ENSG00000231814	NCRNA00210	0.493653	4.6376936	3.973999	0.0062935	none
ENSG00000090889	KIF4A	0.686648	8.3565723	3.973794	0.0062951	none
ENSG00000138798	EGF	0.647004	7.3634949	3.973543	0.006297	promoter_and_enhancer
ENSG00000233942		0.779565	4.536522	3.973393	0.0062981	none
ENSG00000166450	PRTG	0.753298	7.1873427	3.969849	0.0063252	none
ENSG00000198920	KIAA0753	0.596518	8.0081981	3.966691	0.0063493	enhancer_only
ENSG00000250580		0.718236	7.0827901	3.962738	0.0063798	none
ENSG00000163110	PDLIM5	0.566846	7.0938915	3.962552	0.0063812	enhancer_only
ENSG00000245793	NA	0.960328	8.6302054	3.962091	0.0063847	none

ENSG00000131773	KHDRBS3	0.513175	9.3289554	3.962051	0.0063851	none
ENSG00000135040	NAA35	0.58993	9.5074313	3.961946	0.0063859	enhancer_only
ENSG00000226852		0.554616	4.9406934	3.959831	0.0064022	none
ENSG00000237649	KIFC1	0.491486	6.4701421	3.943753	0.0065281	none
ENSG00000060982	BCAT1	0.824153	8.7050111	3.939152	0.0065646	none
ENSG00000104313	EYA1	0.463707	6.9877423	3.937779	0.0065755	enhancer_only
ENSG00000213949	ITGA1	0.866268	7.971017	3.932049	0.0066214	enhancer_only
ENSG00000198015	MRPL42	0.626592	8.7915683	3.928858	0.0066471	none
ENSG00000235638	MTND6P14	0.671893	5.3758667	3.9254	0.0066751	none
ENSG00000222257		0.521736	3.2388091	3.924708	0.0066807	none
ENSG00000070778	PTPN21	0.560646	8.4947151	3.924631	0.0066813	none
ENSG00000178343	SHISA3	1.297184	6.7261491	3.923939	0.0066869	none
ENSG00000236244		1.828517	6.6371694	3.918868	0.0067283	none
ENSG00000128833	MYO5C	0.834492	9.0656411	3.918616	0.0067303	none
ENSG00000170381	SEMA3E	0.716456	9.1511552	3.917523	0.0067393	enhancer_only
ENSG00000226556		0.461839	6.020389	3.898611	0.0068963	none
ENSG00000225826		0.928523	5.4769575	3.889143	0.0069764	none
ENSG00000249643		1.007374	5.3205089	3.882138	0.0070363	none
ENSG00000083067	TRPM3	0.70453	5.6281522	3.881277	0.0070437	enhancer_only
ENSG00000092295	TGM1	0.625483	6.1846516	3.875696	0.0070919	none
ENSG00000225706		0.979603	6.1301989	3.875408	0.0070944	none
ENSG00000176244	ACBD7	0.92942	6.0819375	3.87124	0.0071306	enhancer_only
ENSG00000252815		0.88316	3.2683668	3.869472	0.0071461	none
ENSG00000162694	EXTL2	0.536275	6.2200755	3.867956	0.0071593	none
ENSG00000212673		0.89317	5.2789451	3.858963	0.0072386	none

ENSG00000183856	IQGAP3	0.532544	7.051343	3.850224	0.0073165	enhancer_only
ENSG00000107679	PLEKHA1	0.555069	8.5978135	3.845522	0.0073589	enhancer_only
ENSG00000055163	CYFIP2	0.703692	9.5204305	3.83957	0.0074128	enhancer_only
ENSG00000137463	C4orf49	0.630479	10.1025	3.836474	0.0074411	promoter_only
ENSG00000166068	SPRED1	0.752948	8.598311	3.832959	0.0074733	enhancer_only
ENSG00000178642		0.802207	5.6366202	3.83147	0.007487	none
ENSG00000185418	TARSL2	0.495282	8.1426523	3.831082	0.0074905	enhancer_only
ENSG00000070018	LRP6	0.680887	8.7989233	3.829866	0.0075017	none
ENSG00000144642	RBMS3	0.747999	7.5189664	3.828592	0.0075135	enhancer_only
ENSG00000234281		0.897124	5.4165202	3.826348	0.0075343	none
ENSG00000006576	PHTF2	0.692871	7.8804996	3.826081	0.0075367	none
ENSG00000204899	MZT1	0.664598	7.6895195	3.824634	0.0075502	none
ENSG00000235162	C12orf75	0.592874	10.66954	3.822407	0.0075709	none
ENSG00000173166	RAPH1	0.557481	9.4590923	3.820139	0.007592	promoter_and_enhancer
ENSG00000175395	ZNF25	0.811766	7.1016487	3.81204	0.0076681	enhancer_only
ENSG00000144962	SPATA16	1.093907	5.2840264	3.81127	0.0076754	none
ENSG00000249648	NA	0.499839	8.2653666	3.81037	0.0076839	none
ENSG00000171914	TLN2	0.547501	7.3927676	3.809844	0.0076889	enhancer_only
ENSG00000123983	ACSL3	0.688136	10.031229	3.809137	0.0076956	promoter_only
ENSG00000251183		0.649731	5.0075914	3.807952	0.0077069	none
ENSG00000081026	MAGI3	0.701938	7.3707862	3.805078	0.0077342	none
ENSG00000197302	ZNF720	0.723375	7.3726911	3.801626	0.0077672	none
ENSG00000167005	NUDT21	0.503527	9.5961502	3.797181	0.0078099	enhancer_only
ENSG00000198901	PRC1	0.571343	9.863303	3.796156	0.0078198	enhancer_only
ENSG00000235455		0.614547	5.0930961	3.795383	0.0078273	none

ENSG00000226025		0.667343	5.7594223	3.786406	0.0079146	none
ENSG00000075303	SLC25A40	0.732509	7.6711494	3.785158	0.0079268	none
ENSG00000241351	IGKV3-11	0.804139	5.0786256	3.783899	0.0079391	none
ENSG00000199306		0.589638	4.7711537	3.782918	0.0079488	enhancer_only
ENSG00000226605		0.536238	6.1239371	3.769516	0.0080817	none
ENSG00000175161	CADM2	0.746433	6.4593539	3.762231	0.008155	none
ENSG00000225511	C9orf44	0.62735	6.1438785	3.761761	0.0081598	none
ENSG00000229647		0.630804	5.2948996	3.755764	0.0082207	none
ENSG00000241336		0.665753	4.2955695	3.755389	0.0082245	none
ENSG00000153044	CENPH	0.51542	4.9083454	3.754967	0.0082288	none
ENSG00000198554	WDHD1	0.902281	8.453946	3.75408	0.0082379	none
ENSG00000123473	STIL	0.612257	6.8950771	3.753207	0.0082468	none
ENSG00000215808		0.499208	9.8122388	3.747351	0.008307	none
ENSG00000165891	E2F7	0.767094	8.3970422	3.737477	0.0084095	none
ENSG00000199348		0.93912	4.187526	3.737368	0.0084107	none
ENSG00000139323	POC1B	0.452638	8.4206047	3.736832	0.0084163	enhancer_only
ENSG00000126773	C14orf135	0.472242	8.9585854	3.736596	0.0084188	promoter_only
ENSG00000225712	ATP5G2P1	0.60535	4.1177353	3.735487	0.0084304	none
ENSG00000231762		0.560367	4.0491166	3.73298	0.0084567	none
ENSG00000201330	SNORD32B	0.60063	4.6266299	3.728645	0.0085025	none
ENSG00000198398	TMEM207	0.554664	5.2790328	3.727944	0.0085099	promoter_only
ENSG00000152078	TMEM56	0.60967	7.8804655	3.724391	0.0085476	none
ENSG00000091428	RAPGEF4	0.523323	5.6076389	3.723348	0.0085587	enhancer_only
ENSG00000145332	KLHL8	0.536229	8.2599216	3.719306	0.0086019	enhancer_only
ENSG00000178202	KDELC2	0.772639	9.0084216	3.717103	0.0086256	none

ENSG00000169760	NLGN1	0.664343	4.9375624	3.716732	0.0086296	promoter_only
ENSG00000180663	VN1R3	0.59507	5.7160618	3.710035	0.0087019	none
ENSG00000146842	TMEM209	0.446591	9.3194925	3.709352	0.0087094	enhancer_only
ENSG00000252063		0.80202	4.3356567	3.70818	0.0087221	none
ENSG00000241847		0.599965	5.2878455	3.707254	0.0087322	none
ENSG00000087470	DNM1L	0.619779	9.7193935	3.701594	0.0087941	promoter_and_enhancer
ENSG00000207022		0.824045	4.105769	3.701344	0.0087969	none
ENSG00000222796		0.723529	4.5003441	3.700722	0.0088037	none
ENSG00000225179		0.448492	4.1481129	3.700383	0.0088074	none
ENSG00000177917	ARL6IP6	0.462058	6.4557888	3.699536	0.0088168	none
ENSG00000230829		0.467486	6.6549191	3.697229	0.0088422	none
ENSG00000250603		0.54957	4.2088691	3.695784	0.0088582	none
ENSG00000228056	CFLP3	0.531226	6.2842896	3.695688	0.0088592	none
ENSG00000102098	SCML2	0.632938	6.1516054	3.69564	0.0088598	none
ENSG00000250064		0.438546	4.5934898	3.687774	0.0089474	none
ENSG00000223549		0.481028	7.4042607	3.687333	0.0089523	none
ENSG00000159167	STC1	1.641764	7.8830907	3.686336	0.0089635	enhancer_only
ENSG00000232110		0.509765	6.722359	3.686131	0.0089658	none
ENSG00000124406	ATP8A1	0.748481	6.2741133	3.682962	0.0090014	none
ENSG00000142892	PIGK	0.584325	8.7035515	3.682474	0.0090069	enhancer_only
ENSG00000126785	RHOJ	1.130264	5.7407561	3.681703	0.0090156	enhancer_only
ENSG00000180947	NA	0.5847	5.3982612	3.681311	0.00902	none
ENSG00000186088	PION	0.631454	5.9043493	3.679518	0.0090403	promoter_and_enhancer
ENSG00000127863	TNFRSF19	0.943068	6.4168474	3.675735	0.0090832	enhancer_only
ENSG00000205403	CFI	0.899366	6.4108651	3.673525	0.0091084	none

ENSG00000065328	MCM10	0.774203	6.5810905	3.672106	0.0091246	enhancer_only
ENSG00000241288		0.430401	5.8157086	3.670471	0.0091434	none
ENSG00000129810	SGOL1	1.216274	7.1271414	3.670294	0.0091454	none
ENSG00000156234	CXCL13	0.590485	4.6781829	3.662829	0.0092314	none
ENSG00000101574	METTL4	0.45462	6.6221929	3.662557	0.0092346	none
ENSG00000138658	C4orf21	0.870856	5.3765159	3.662219	0.0092385	none
ENSG00000078098	FAP	1.568001	5.2277845	3.659627	0.0092686	none
ENSG00000174225	ARL13A	0.464466	4.530356	3.659148	0.0092742	none
ENSG00000231792		0.507477	5.5381779	3.658631	0.0092802	none
ENSG00000221574		1.148813	4.4274721	3.658526	0.0092814	none
ENSG00000080839	RBL1	0.685459	8.1097344	3.654127	0.0093328	none
ENSG00000113083	LOX	0.588594	8.0808138	3.653716	0.0093377	none
ENSG00000131725	WDR44	0.466453	8.6236375	3.651248	0.0093666	promoter_and_enhancer
ENSG00000187939	DOC2B	0.567355	7.0726003	3.651076	0.0093687	none
ENSG00000120784	ZFP30	0.706614	6.7930585	3.644709	0.009444	enhancer_only
ENSG00000113361	CDH6	1.059144	5.6994977	3.643808	0.0094547	enhancer_only
ENSG00000112902	SEMA5A	0.493677	7.6782224	3.643613	0.009457	enhancer_only
ENSG00000140479	PCSK6	0.476561	7.0673264	3.642366	0.0094718	enhancer_only
ENSG00000105854	PON2	0.611892	9.447619	3.639737	0.0095032	enhancer_only
ENSG00000227735		0.562773	5.2085062	3.630252	0.0096174	none
ENSG00000117650	NEK2	0.550557	7.649636	3.629533	0.0096261	none
ENSG00000247034	NA	1.099677	9.1759307	3.629189	0.0096303	none
ENSG00000236199		0.946972	5.0777415	3.620267	0.0097392	none
ENSG00000222761		0.605973	2.9756533	3.618036	0.0097666	none
ENSG00000230297	NA	0.495669	3.7029364	3.618011	0.009767	none

ENSG00000082497	SERTAD4	1.374667	5.8748061	3.612062	0.0098406	none
ENSG00000088451	TGDS	0.443883	6.8358014	3.609482	0.0098727	none
ENSG00000109084	TMEM97	0.601202	10.787283	3.607432	0.0098983	none
ENSG00000241808		0.489315	5.3887329	3.606801	0.0099062	none
ENSG00000136122	C13orf34	0.473641	7.6378999	3.604451	0.0099356	none
ENSG00000124788	ATXN1	0.439381	6.9416161	3.604101	0.00994	enhancer_only
ENSG00000183504		0.521102	4.5373637	3.602413	0.0099612	none