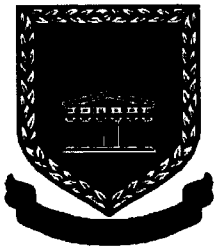
Final Copy



UNIVERSITY *of the*
WESTERN CAPE

SANBI

Ancient Genes in Cancer Gene Expression?

UNIVERSITY of the
WESTERN CAPE

NAME : Sumir Panji

STUDENT NUMBER : 2355015

SUPERVISOR : Professor Winston Hide

A minithesis submitted in partial fulfillment of the requirements for the degree of Magister Scientiae in the Department of Biotechnology, University of the Western Cape.

2004

https://etd.uwc.ac.za/

Keywords

Cancer / Testis genes

Cancer / Testis antigens

The human genome

The X chromosome

Rice's hypothesis

Gene expression

In-silico and wet-lab gene expression assay

Spectrum of gene expression

UNIVERSITY of the
WESTERN CAPE

Comparative genomics

Phylogenetic scope

https://etd.uwc.ac.za/

Abstract.

Background: The Cancer/testis (CT) antigens are a division of germ cell specific genes not expressed in somatic cells, exceptions being placental cells and 20% - 40% of cancer types. The aptitude of CT antigens to elicit humoral immune responses, their restricted expression profile, absence of major histocompatability complex expression in male germline cells have contributed to the emergent attraction of CT antigens as ideal, prospective cancer vaccination candidates.

Motivation: Presently there are 44 CT gene families containing a total of 97 gene products and isoforms. Due to the promulgation in sensitivity and specificity of rapid serological immunodetection assays e.g. serial analysis of recombinant cDNA expression libraries (SEREX), the magnitude of novel CT genes and gene families will increase. Hence, characterization of this unique subset of CT genes is fundamental to our erudition of this rapidly emerging novel subset of genes.

Objectives: The sequencing of the human genome provides a useful biological framework for the categorization and systematization of rapidly accumulating biological information. A genomic approach was used to ascertain the locations of the CT genes in the human genome and determine if the genomic locations of the CT genes is non-random. An *in-silico* expression study was conducted for the CT genes with the aim of establishing if CT gene expression is restricted to the testis. A portion of the human genome housing the largest proportion of the CT genes was selected for analysis in order to determine if the surrounding genomic architecture influences CT gene expression. A comparative genomics approach was used in determining if the CT genes are "ancient genes".

UNIVERSITY of the
WESTERN CAPE

https://etd.uwc.ac.za/

Declaration.

I declare that "*Ancient Genes in Cancer Gene Expression?*" is my own work, that it has not been submitted for any degree or examination in any other university, and that all the sources I have used or quoted have been indicated and acknowledged by complete references.

Name: Sumir Panji **Date:** 21st December 2004.

UNIVERSITY of the
WESTERN CAPE

Signed:.....

https://etd.uwc.ac.za/

Acknowledgments.

I would like to extend my heartfelt thanks to my thesis supervisor Professor Winston Hide (Win) for his constant encouragement and numerous invaluable and often humorous discussions, without which this mini-thesis would not be possible. I would also like to say thank you to Win for his patience and allowing me the freedom to explore many interesting scientific topics, some more pertinent to this mini-thesis then others. It has been an interesting one and a half years that I have been part of Win's lab of which I have benefited by learning how to tackle different scientific issues in depth and learned to think really hard, no matter how cruel a mistress science may be. Thanks Win.

I would like to thank Dr Janet Kelso for her many invaluable and insightful discussions regarding the interpretation of the eVoc results as well as some salient features of gene expression and different gene expression capture technologies. I would like to thank both Simon Cross for providing me with the number of annotated cDNA libraries present in eVoc for various tissue types and Konrad Scheffler for help with the statistical aspects such as the Chi-square tests.

I would like thank the SANBI masters students of 2003; Anelda Boardman, Bukiwe Lupindo, Faisel Mousoval, Mario Jonas and Nothemba Gwija for their many hilarious conversations and making SANBI a fun place to work as well as their constant help and solidarity, especially when I was going through a trying period when I first arrived in South Africa.

I would also like to thank my parents; Usha and Noordin Panji (not least for conceiving me!!!) for their constant encouragement and unwavering, unconditional support through out my academic endeavors, thanks Mum and Dad.

I am grateful to SANBI and the National Bioinformatics Network for providing me with support in the form of a bursary during this degree.

UNIVERSITY of the
WESTERN CAPE

https://etd.uwc.ac.za/

TABLE OF CONTENTS.

UNIVERSITY of the WESTERN CAPE

https://etd.uwc.ac.za/

UNIVERSITY of the WESTERN CAPE

https://etd.uwc.ac.za/

Chapter 1

Introduction.

Cancer is a generic term used to describe over a hundred types of diseases characterized by the uncontrolled, rampant proliferation of, in most instances, somatic cells. The estimated incidence of cancer worldwide is approximately 10 million cases per a year which contributes to 12% (6 million) of worldwide mortalities per annum♣.

Improvements in imaging, screening and molecular diagnostic technologies for cancer detection together with increasing life expectancies will contribute to a rise in cancer incidence. The predicted incidence of cancer cases diagnosed by the year 2020 is expected to be 15 million new cancer cases per annum♣.

Since President Nixon's declaration of the "War on Cancer" in 1970, our understanding of cancer etiology has significantly improved. The emerging consensus is that cancer is not a disease caused by the simple mutation of a small subset of genes [1,2]. Cancer is a complex, dynamic and multi-factorial process involving the interaction of an organism's genes and their environment [1-4]. Additionally, numerous epigenetic events combined with the variability in an individual's inherited genetic background can subtly alter cell physiology without directly compromising the integrity of a cell's DNA or genetic machinery [2,5]. In the post-genomic era, data garnered from the human genome projects [6,7] and model disease organisms [8,9] offers invaluable insights to the biological significance of the plethora of events that occur during oncogenesis. These insights will assist in the interpretation of the accumulating body of cancer knowledge by placing the innumerable events that occur during oncogenesis in their appropriate contexts, thereby facilitating our understanding of the complex events that occur during cancer.

♣ http://www.who.int/cancer/en/

https://etd.uwc.ac.za/

A Genomic Overview of Cancer.

Emerging neoplastic cells amass an array of genetic and epigenetic modifications which culminate in altered genomic expression patterns [10-18]. These genomic aberrations are ultimately expressed resulting in the observed phenotype of neoplastic cells [11,13,15,16,19,20]. Hanahan *et al* [21] have outlined six prerequisite adaptations in cell physiology observed in the bulk of cancerous cells; impunity from growth inhibitory and differentiation signals, evasion of apoptosis, self sufficiency in growth factors, limitless proliferation capacity, angiogenesis, invasion and metastasis [21]. The pool of neoplastic progeny cells are dynamically selected for in a Darwinian fashion by virtue of their genomic complement enabling the survival and expansion of successful neoplastic cells in their respective micro-environments [21,22].

Regulation of gene expression is an exquisitely concerted and tightly coordinated process operating at multiple levels and stages during an organism's lifespan [23]. Hence, deregulation of gene expression resulting in cancer is not likely to be a frequent occurrence [13,23,24]. With approximately 10^14 cells in the human body, an estimated mutation rate of approximately 1 gene in every 2 x 10^7 cell divisions, the emergence of cancer is estimated to occur once in every 3 lifetimes [12,24]. Subversion from the pre-programmed terminal differentiation pathway of somatic cells, a salient characteristic of cancer, occurs predominantly in two ways [2]:

1). A direct alteration to the DNA sequence itself resulting in a gain or loss of function mutation [10,14,15,20,25].

2). An epigenetic modification which does not directly alter the DNA sequence but results in altered gene expression patterns [11,16,18,26].

A multitude of genetic and epigenetic factors, which can be inherited or acquired, influences the progression of cancer. The multi-faceted nature of cancer indicates numerous complex interactions occurring between a diverse range of causative factors ranging from

UNIVERSITY of the WESTERN CAPE

https://etd.uwc.ac.za/

lifestyle choices e.g diet, exposure to known carcinogens, to the genetic variability inherited by an individual [1,2,5,13]. The combinatorial complexity of the potential factors involved in the subversion of a somatic cell's genome during cancer indicates that the likelihood of devising a single "magic bullet" type of cure for cancer is slim [24,27].

A Cure for Cancer ?

A "magic bullet" type of cure for cancer would function by specifically targeting cancerous cells while concurrently ignoring the normal somatic cell population [27,28]. As cancer cells arise from somatic cells, a "magic bullet" would have to counter the apparent homogeneity between cancerous and normal somatic cells, rendering the normal somatic cell population intact and viable. Fortuitously, the subversion of a somatic cell's genome during cancer results in the expression of gene products which are not normally expressed at a particular anatomical site by the normal somatic cell population of that anatomical site.

Advancements in molecular immunogenic methodologies and biotechnology permit detailed topographies of cancerous cells to be obtained [27-29]. Detailed cell topographies and the ability to synthesis biological molecules in pharmacological quantities enable the molecular selection of gene products which are aberrantly expressed during cancer [27-30]. These aberrantly expressed gene products serve as potential beacons for a "magic bullet" type cure to target cancerous cells from normal somatic cells, and hence potentially destroy the emerging cancerous cells leaving the normal somatic cell population intact and viable. An example such aberrantly expressed genes with cell surface antigenic properties are a class of human genes designated as the Cancer/Testis Antigens[27,29-36].

UNIVERSITY of the
WESTERN CAPE

https://etd.uwc.ac.za/

The Cancer / Testis Antigens.

The Cancer / Testis Antigens (CTAs) have been described as a collation of genes which exhibit a testis restricted expression pattern and are immunogenic in cancer patients [30,33,37]. Cancer / Testis (CT) genes are genes which have been identified by bioinformatics analysis of public high throughput gene expression data such as Expressed Sequences Tags (ESTs) and Serial Analysis of Gene Expression (SAGE) [37,38]. Additionally, the cancer / testis restricted expression profiles of the CT genes and CT antigens are determined by Reverse Transcriptase Polymerase Chain Reaction (RT-PCR) of different non-cancerous and cancerous tissue samples [37].

The principle difference between CTAs and CT genes is that the former category has been identified through immunological methods such as serological analysis of recombinant complementary DNA (cDNA) expression libraries (SEREX) and autologous typing methodologies, the latter through gene expression studies of normal and cancerous cells[29,30,37-39]. Although the CT genes may have the propensity to elicit an immune response, their immunogenicity remains to be experimentally validated[37,38]. Due to the isolation and characterization of the CTAs by immunological methods, the immunogenic properties of the CTAs have been experimentally validated[29,30,37,39]. Throughout this study, the subset of cancer / testis genes that have been identified and isolated through immunological methods shall be referred to as CTAs. The subset of cancer / testis genes that have been identified by gene expression methods, but whose immunogenicity remains to be experimentally validated, shall be referred to as CT transcripts. The term "CT genes" shall be used to refer to both, the CTAs and CT transcripts throughout this thesis.

The CT genes are a group of unrelated heterogeneous genes whose key unifying feature is an mRNA expression profile restricted to male germ cells in the testis, fetal ovary, placenta, pancreas and 20% – 40% of a disparate range of cancer types [30,31,33,37]. CTAs have

UNIVERSITY of the WESTERN CAPE

https://etd.uwc.ac.za/

the additional property of being able to elicit a cellular and humoral immune response[30,37,39].

Presently there are 44 CT gene families that have been partly characterized [37]. An expression survey of 43 of the 44 founding constituents of the CT gene families undertaken by Scanlan *et al*[37] has subdivided the CT genes into four categories based on their expression profiles, these four CT gene expression categories are presented below in Table 1. Standardized RT-PCR of a panel of 16 non-cancerous tissues comprising of 13 non-gametogenic and 3 gametogenic tissue types was used to derive the four CT gene expression categories[37].

CT Gene Expression Category	Percentage* of 43 CT Genes in each CT Gene Expression Category
1. Testis – restricted expression.	44%
2. CT gene expression detected in ≤ 2 non-gametogenic tissues.	23%
3. CT gene expression detected in 3-6 non-gametogenic tissues.	21%
4. Ubiquitous CT gene expression ≥ 6 non-gametogenic tissues.	12%

UNIVERSITY of the WESTERN CAPE

*43 out of the 44 founding CT family members were used as HCA661/CT30 could not be amplified [37]

Table 1 : The Percentage of CT genes in each Gene Expression Group as determined by their expression profile in 13 non-gametogenic and 3 gametogenic normal tissues assayed [37].

From the first 2 CT gene expression categories shown in Table 1, 48% of these CT genes have been shown to elicit a cellular and / or humoral immune response, and hence are CTAs [37]. CT gene expression in the testis is confined to immature germ cells such as spermatogonia, the exception being CT 23 (OY-TES-1) which is a precursor for proacrosin binding protein sp32 engaged in acrosin packaging in the acrosome in the late stages of sperm maturation[33,37,40]. Cytotrophoblasts and syncytiotrophoblasts from fetal placenta exhibit CT gene expression while term placenta from the third trimester of pregnancy has

https://etd.uwc.ac.za/

little or no CT gene expression [33] . CT gene expression has also been detected in fetal ovaries and immature germ cells, but not in oocytes quiescent in the primordial follicles[33]. Hence, it would appear that CT gene expression is primarily restricted to germ cells and male gametogenic tissues such as the testis[33]. Apart from CT 23 (OY-TES-1) and the synaptonemal complex protein, CT 8 (SYCP-1) which is involved in homologous chromosome pairing during the meiotic prophase of spermatocytes, vital for the establishment of haploid cells in meiosis I [41], the biological roles of 97.5% CT genes is largely unknown [33,37]

The testis forms an immuno-privileged site as male germ cells do not express MHC class I molecules, and consequently are incapable of expressing antigens recognized by T lymphocytes [31,42]. Thus, the testis restricted expression profile of the CTAs make them promising immunogenic vaccine targets against cancer cells by circumnavigating the possibility of inducing an auto-immunogenic response [31,32,34,36,42]. The ability of the CTAs to elicit an autologous cytolytic T lymphocyte (CTL) mediated immune response against neoplastic cells, coupled with the narrow range of tissues in which the CT genes are expressed in, provides the main impetus for any study conducted regarding the CT genes[31,37,39,42].

UNIVERSITY of the
WESTERN CAPE

CT Gene Expression During Cancer.

Regulation of gene expression is a tightly co-coordinated process operating throughout an organism's lifespan at multiple levels and stages [23]. The aspects by which modulation of gene expression occurs are diverse and amongst others, include transcriptional control, the accessibility of DNA to transcription factors, chromatin structure as well as epigenetic aspects, to mention a few[5,16,23,43-47]. Deregulation of normal gene expression, a hallmark of cancer, can also occur through direct DNA mutations resulting in the abnormal expression of genes and their biological products in neoplastic cells[2,12,13,15,20,21]. Additionally, epigenetic

https://etd.uwc.ac.za/

modifications which result in altered gene expression patterns without directly changing the underlying DNA base composition are also found to occur in neoplastic cells [2,5,11,13,16,21,46].

Epigenetic modifications enable a cell to program its genome permitting the expression of genes at a particular point and under specific stimuli [5,48,49]. Epigenetic modifications are also attributed to the specificity of gene expression modulation during developmental and programmed cell specialization / differentiation[5,48,49]. Amongst other mechanisms, epigenetic modifications include the specific methylation and demethylation of the cytosine nucleic acid bases in a DNA strand resulting in different genome wide methylation configurations[5,48,49,50]. DNA methylation is undertaken by enzymes known as DNA cytosine methyltransferases (DMNT) that catalyse the addition of methyl groups to the 5' position of cytosines in a DNA strand[5,48-50]. The addition or removal of methyl groups to cytosines in a DNA strand results in conformational changes to the major DNA groove to which DNA proteins bind, thereby inhibiting or facilitating the initiation of transcription for a gene[5,48-50]. These acquired or lost methylated epigenetic markers remain fixed within the human genome and any changes to these methylated epigenetic markers are inherited during DNA replication, after DNA synthesis[5,48-50]. The fixed methylated epigenetic markers are only erased and reset at the blastocyst stage of embryogenesis, through global hypomethylation proceeded by global *de novo* methylation[48-51]. Alterations in genome wide DNA methylation patterns are one of the most frequent genomic alterations observed in human cancers [5,52,53]. The genomic alterations in methylation patterns are due to the hypermethylation of GC rich regions associated with the promoter sequences of different genes while concurrently ushering a decrease in overall global levels of DNA methylation in cancer[5,11,18,26,52,53].

CT gene expression in cancer cells is mainly thought to be the result of epigenetic modifications [33,37,42,50,54-56]. A correlation was found between the genome wide demethylation events that occur during cancer and the unrestricted expression of some of the CT genes in

UNIVERSITY of the WESTERN CAPE

https://etd.uwc.ac.za/

20% – 40% of cancer types[33,54-56]. The melanoma antigen1 family (MAGEA) was found to be stochastically expressed in different types of cancers as a result of genome wide hypomethylation[55,56]. De Smet *et al*[55,56] demonstrated that regulation of gene expression for the MAGEA genes in tumor cells occurs at the genomic level and not at the transcriptional level. Tumor cells that did not express the MAGEA genes were found to be proficient for transcription factors that activate the MAGEA promoters[55,56]. Induction of gene expression for the MAGEA genes in these tumor cells was only observed when the tumor cells were treated with the demethylating agent 5-Aza-2'-Deoxycytidine (DAC)[55,56]. The cell clones treated with DAC were additionally found to have stable levels of MAGEA gene expression[55,56]. Thus, De Smet *et al* [55,56] showed a change in the methylation status of the MAGEA genes is both necessary and sufficient for MAGEA gene expression in tumor cells. The GC promoter regions of the MAGEA and LAGE genes were also found to be heavily methylated in normal somatic cells, but unmethylated in germ cells[55,56]. Hence, methylation of the GC rich promoter regions of the MAGEA genes promoters appears to be the primary mechanism controlling the expression of the MAGEA genes in normal and cancerous somatic cells[33,50,55,56]. Similar studies with the CT 4 (GAGE) gene family of CT genes also indicates that their expression is regulated by methylation[33,42,54]. The expression of CT genes in the male germline is attributed to the global demethylation status present in male germ cells[33,50,55,56]. However, the demethylation and subsequent expression of some of the CT genes during cancer appears to be a stochastic phenomenon as some heavily hypomethylated tumor cell lines do not exhibit CT gene expression, and only 20-40% of all cancers exhibit CT gene expression[33,55,56].

UNIVERSITY of the
WESTERN CAPE

https://etd.uwc.ac.za/

Study Outline.

The main approaches towards the characterization of the CT genes have focused on the tissue expression profiles and the antigenicity of the CT genes[30,33,37]. The sequencing of the human genome provides a useful framework within which the characterization of the CT genes with regards to their genomic locations, expression profiles, antigenicity and conservation amongst different metazoan species can be undertaken. All the CT genes have been primarily identified and characterized in humans and hence contribute, in a small part, to the catalogue of genes that ultimately form the human genome. Four pilot studies were conducted, each with the aim of determining and characterizing specific aspects of the CT genes. These four studies are presented as separate chapters, each containing an introduction, a methods section, an analyses and discussion of the results and conclusions with regards to the relevant published literature present.

The first research chapter, Chapter 2, undertakes the genomic mapping of the CT genes to the human genome in order to determine the genomic distribution of the CT genes and ascertain any possible factors that may have led to the observed genomic distribution of the CT genes. Chapter 3 compares the results of an *in-silico* CT gene expression study undertaken with the results of a "wet-lab" study conducted by Scanlan *et al* [37], with an aim to determine if CT gene expression is testis specific. Chapter 4 focuses on a portion of the human genome housing the CT genes in order to determine if the expression of the CT genes may be influenced by their physical location in the human genome. In Chapter 5, a comparative genomic approach is used to determine how well conserved the CT genes are in different metazoan species, and consequently, which CT genes are "ancient".

UNIVERSITY of the
WESTERN CAPE

https://etd.uwc.ac.za/

Chapter 2

Where do the CT Genes Reside in the Human Genome?

The mapping of CT genes to the human genome provides a coherent functioning genomic framework permitting the analysis of the genomic distribution of the CT genes with regards to the CT gene expression profile and antigenicity. The mapping of the CT genes to the human genome was undertaken in order to determine the genomic distribution of the CT genes and what possible factors may influence the genomic distribution of the CT genes. Determination of the genomic locations of the CT genes in the human genome was also undertaken partly due to the lack of functional knowledge of the gene products for 97.5% of the known CT genes as well as which of the CT genes are the true ancestral genes[37]. Additionally, the elucidation of the genomic locations for all of the CT genes facilitates in determining whether CT gene expression may be influenced by their genomic locations.

UNIVERSITY of the
WESTERN CAPE

Methods.

Working accession numbers for the CT genes were obtained through a collaboration with the Ludwig Institute for Cancer Research. The non-redundant curated RefSeq [57,58] database (Release 3) was mined using the appropriate accession numbers. NCBI's Entrez system was used to link out to the gene centric LocusLink [57,58] database (based on RefSeq Release 3). Where possible, the genomic co-ordinates of the CT gene transcripts that were mapped onto Ensembl [59-61] (version 19.34b.2 based on NCBI build 34, 9 Feb 2004), a database providing an automated annotation of the human genome sequence, was obtained via LocusLink. As Ensembl utilizes an automated annotation pipeline[59,62] , any ambiguities in the mapping of the CT genes, e.g multiple CT gene transcripts grouped together and mapped

https://etd.uwc.ac.za/

as a single gene, were resolved using a combination of the WU-BLASTN algorithm[♣] from Ensembl's BLAST homepage, GeneLoc [63] (Version 2.9 based on NCBI build 34) and STS markers. GeneLoc seeks to provide an integrated map of the human genome by eliminating redundancies and comparing gene collections from NCBI and Ensembl when determining which gene entries are distinct, and which entries should be merged[63]. The combination of approaches allowed a manually curated CT gene dataset to be constructed which is presented in the Appendix as Table A, pg 82. The CT gene nomenclature proposed by Scanlan *et al* [37] together with the CT gene names in brackets is used when referring individually to the different CT genes and gene families. The CT nomenclature proposed by Scanlan *et al*[37] was adopted because some of the CT genes have aliases and have yet to be designated official HUGO[64] gene names. A full listing of the CT genes with their relevant CT gene identifiers is presented in the Appendix as Table A.

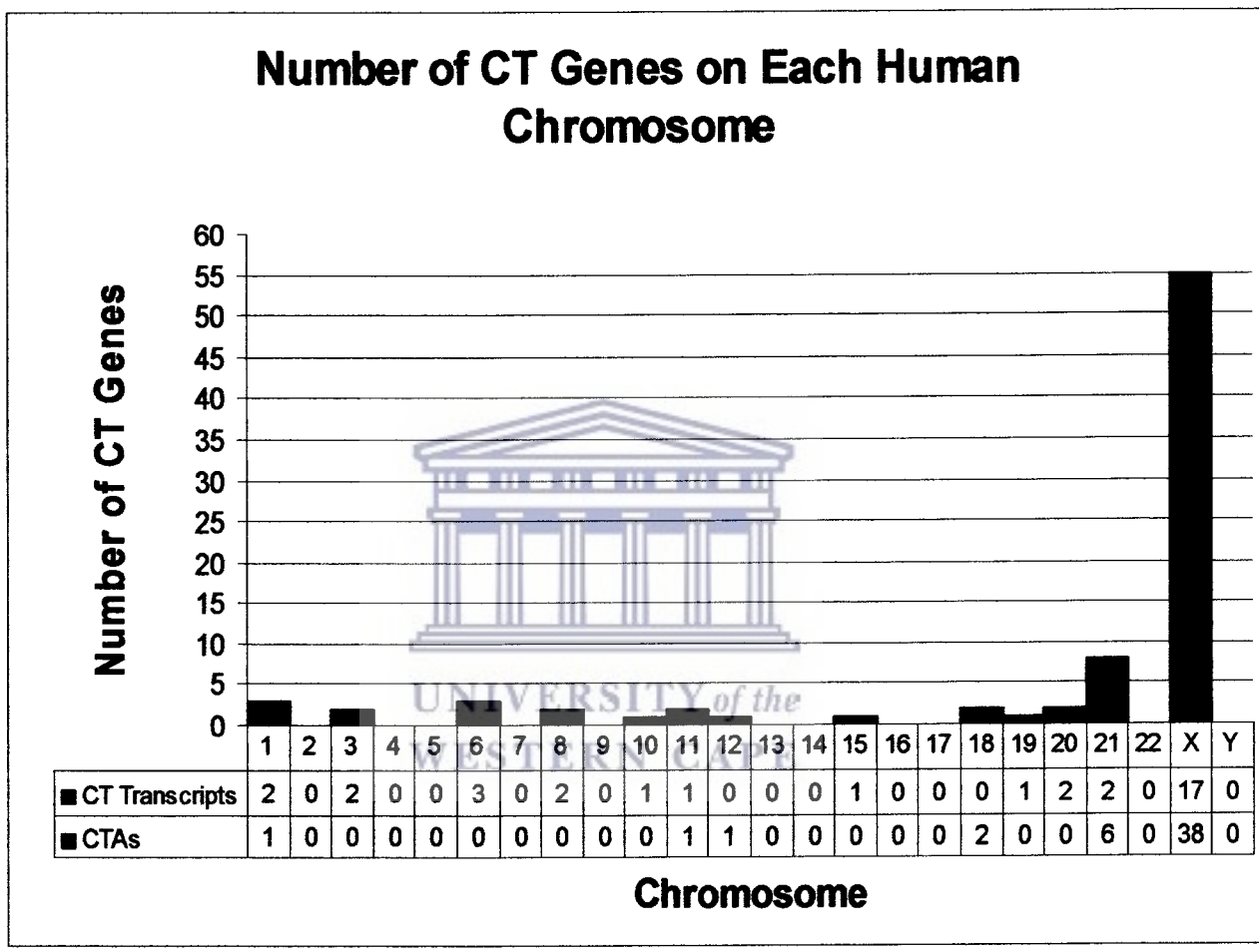
Is the Genomic Distribution of the CT Genes Random?

Table A (Appendix, pg 82) provides a list of the CT genes and their transcript variants together with their genomic locations and immunogenic profiles. The complete dataset containing 97 different transcript variants and isoforms for 83 of the CT genes was filtered to eliminate transcript variants of the same gene for the analysis of genomic distributions of the CT genes. Removal of different transcript variants for a single CT gene eliminates a source of bias when determining the genomic distribution of the CT genes. Although the transcript variants and protein isoforms of a single CT gene differ from each other slightly, they are essentially products of the same gene that occupies a fixed genomic position in the human genome. The inclusion of CT gene transcript variants and isoforms in

[♣] Gish, W. (1996-2004) http://blast.wustl.edu

UNIVERSITY of the WESTERN CAPE
https://etd.uwc.ac.za/

the analysis of the genomic distribution of the CT genes would lead to the "double counting" of the number of CT genes present at a specific genomic locus, hence introducing a source of redundancy. The resulting dataset comprising of 83 CT genes in total was plotted according to the chromosomal locations of the CT genes, the results are presented in Figure 1.



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	X	Y
CT Transcripts	2	0	2	0	0	3	0	2	0	1	1	0	0	0	1	0	0	0	1	2	2	0	17	0
CTAs	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	2	0	0	6	0	38	0

Figure 1 : The genomic location and distribution of the 83 known CTAs and CT transcripts in the human genome based on their chromosomal locations.

Fifty-five (66%) of the 83 CT genes are located on the submetacentric X chromosome (Figure 1). A slightly higher proportion of 31 CT genes (37%) reside on the long arm of the X chromosome and 29% (24 CT genes) on the short arm, Xp. In comparison, the second highest congruence of CT genes map onto the acrocentric chromosome 21 with a total of 8 CT genes (9.6%), 6 of the CT genes are located on the short arm of chromosome 21. Hence, it appears that the X chromosome houses the majority of the CTAs as well as CT transcripts.

https://etd.uwc.ac.za/

However, closer inspection of the X linked CT genes indicates that they are members of different CT gene families, some of which have been shown to have arisen by duplication[37,65-70]. Thus, the observation of a large number of CT genes mapping onto the X chromosome maybe biased by the number of CT gene families with multiple members that have arisen by duplication (Figure 1, pg 14). In order to determine if the CT genes truly exhibit an X chromosomal linkage biasness, the 83 CT gene dataset was further filtered to remove CT gene family members from CT gene families which appear to have arisen by duplication[37,65-70]. The resulting curated dataset is similar to Scanlan *et al*'s [37] and comprises of the 44 CT genes' founding family members, and hence is based on the chronological order these CT genes were discovered, rather then their true evolutionary history [37]. Table 2 provides a listing of the CT gene families which contain more then one member.

CT Identifier	Gene Name	Number of Members	Genomic location
CT 1	MAGEA	12	Xq28
CT 2	BAGE	5	21p11
CT 3	MAGEB	4	Xp21
CT 4	GAGE	8	Xp11.4
CT 5	SSX	4	Xp11
CT 6	NY-ESO-1 / LAGE	2	Xq28
CT 7	MAGEC	2	Xq26
CT 11	SPANX	4	Xq27.1
CT 12	XAGE	4	Xp11.22
CT 21	CTAGE	2	18p11.2
CT 24	CSAGE	2	Xq28
CT 41	TDRD and NY-CO-45*	2	10q26.11 and 6p21.1*

Table 2 : There are 12 CT Gene families that contain more then one member. *As only founding members of the CT gene families are used for analysis, CT 41.2 (NY-CO-45) is excluded from the analysis of the genomic distributions of the CT genes.

UNIVERSITY of the WESTERN CAPE
https://etd.uwc.ac.za/

From the 12 CT gene families identified, 9 of the CT gene families map onto the X chromosome (Table 2, pg 15). In total, 51 CT genes were identified as belonging to a CT gene family, with 42 CT gene family members mapping onto the X chromosome (Table 2 pg 15). The CT gene families on the X chromosome have on average 4.6 members as opposed to an average of 3 members for other chromosomal locations of CT gene families. The genomic distribution of the founding 44 CT genes is presented in Figure 2.

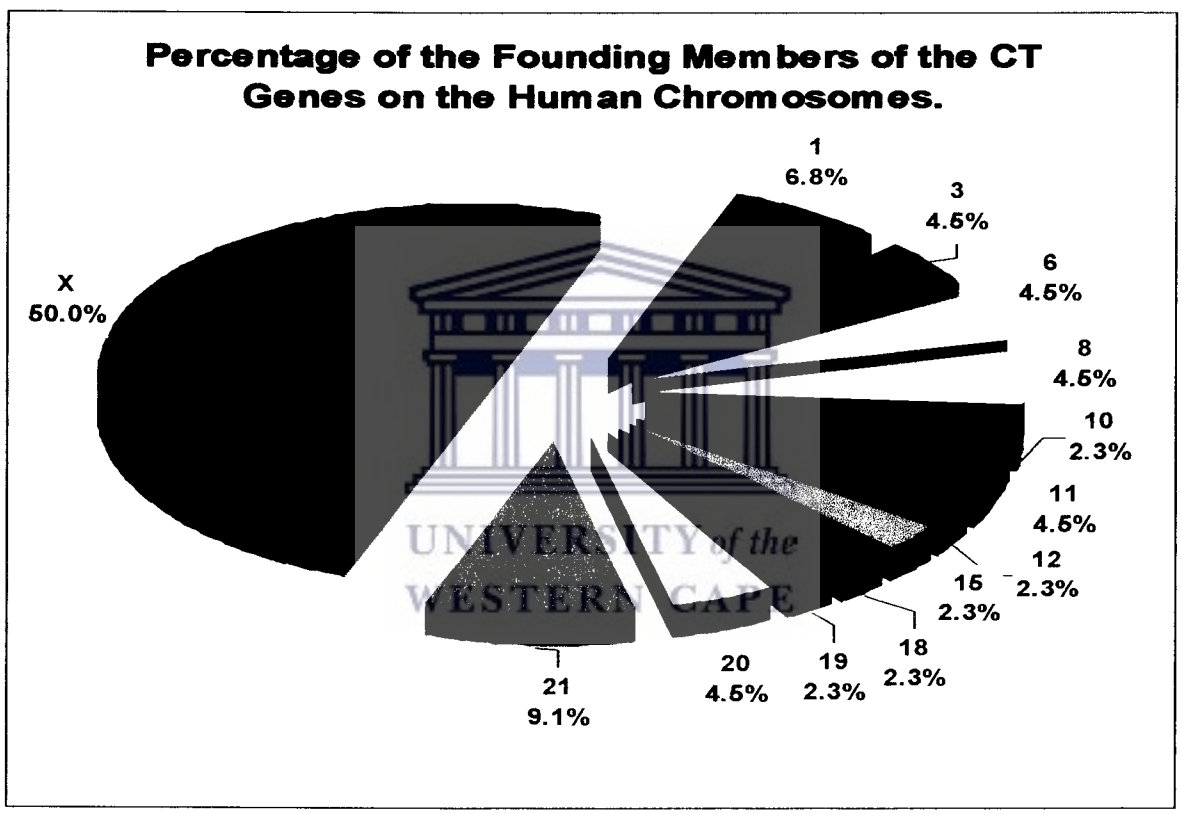


Figure 2 : The proportion of CT genes shown as a percentage (blue) of the forty-four founding members of the CT genes contained on each chromosome (red).

Exclusion of 47% of the full 83 CT gene dataset did not dramatically reduce the number of CT genes that map onto the X chromosome. As 85% of the 51 CT gene family members excluded reside on the X chromosome, one would expect the number of CT genes mapping onto the X chromosome to decrease substantially. However, the highest proportion of CT genes still localize to the X chromosome (Figure 2). Of the 22 CT genes that are X linked (Figure 2), 14 are CTAs and 8 are CT transcripts. In comparison, exclusion of 4 CT

https://etd.uwc.ac.za/

gene family members from chromosome 21 did not result in a large decrease of the percentage (0.5%) of CT genes mapping onto chromosome 21 (Figure 1 pg 14, Figure 2 pg 16). If the 44 founding CT gene family members were randomly distributed in the human genome, one would expect 1 – 2 CT genes on each of the 24 human chromosomes. Only chromosomes 3, 6, 8, 10, 11, 12, 15, 18, 19 and 20 exhibit such a distribution while chromosomes 1, 21 and X do not. The results imply that the X chromosome is enhanced for CT genes. In order to determine why a seemingly heterogeneous subset of human genes that are unified by a cancer / testis expression profile are preferentially located on the X chromosome, the evolutionary history and possible forces which have led to the shaping of an unusual complement of genes on the mammalian sex chromosomes was reviewed.

Evolution of the Mammalian Sex Chromosomes.

In dioecious taxa such as *Homo sapiens*, the karyotype of the gametic sex chromosomes determines the outcome of an individuals' sex. The homogametic karyotype being female (XX), and the heterogametic (XY) being male. Exceptions to rule being *Caenorhabditis elegans* with an XX karyotype being hermaphroditic, and XO being exclusively male. The sex chromosomes have evolved from a pair of autosomes and have arisen independently multiple times amongst dioecious lineages as observed in the ZW (female) and ZZ (male) avian chromosomal sex determining system [71-75].

Before the derivation of mammalian sex chromosomes, approximately 350 million years ago, environmental stimuli such as ambient temperature during embryonic development may have determined the sex of an organism, as observed in reptiles and boney fish [71,74]. Evolution towards a chromosome based sex determining pathway may have occurred when environmental stimuli would have ceased being a useful cue for developmental switching during embryogenesis, which still occurs in modern reptiles[37,71,74].

UNIVERSITY of the WESTERN CAPE

https://etd.uwc.ac.za/

Indeed, the earliest sex chromosome bearers from the lineages leading to birds and mammals are homeotherms [71,74]. Thus, environmental stimuli such as ambient temperature may have ceased being a useful cue for the determination an individual's sex.

The definitive mammalian male specifying gene, Sex determining Region on the Y (SRY), initiates the development of the testis in the bipotential gonad during mammalian embryogenesis and is expressed in the adult testis [71,73,74,76]. The emergence of SRY is thought to herald the cessation of recombination between the ancestral autosomal sex chromosomes, approximately 240-320 million years ago, shortly after the divergence between the mammalian and avian lineages, roughly 350 million years ago [71,73,74,76]. The approximate ages of when recombination between the X and Y chromosome ceased is supported by SRY's closest X-linked gene homologue, SOX3, with both genes forming the oldest X-Y gene pair in *Homo sapiens* [73,74]. The evolutionary forces acting on both the sex chromosomes have a specific, direct impact on the genomic landscape of the X and Y chromosomes and hence, the unusual complement of genes they house.

UNIVERSITY of the

The Genomic Landscape of the Mammalian Sex Chromosomes.

After ~300 million years of evolution, the X and Y chromosomes are genetically and morphologically dissimilar from one another, with a subtle exception. The acrocentric Y chromosome is the smallest chromosome in the human genome and a third the size of the X chromosome (~50 million base pairs as opposed to ~153 million base pairs in *Homo sapiens*). The diminutive size of the Y chromosome is considered to be a result of genetic decay through the loss of most of its ancestral genes [71,73,74]. In contrast, the X chromosome is thought to harbor the ancestral gene complement. Current consensus indicates that X and Y chromosome differentiation arose with the cessation of recombination between both the X and the Y chromosomes [71,73,74,76] . Lahn *et al* have posited that four inversions occurred on

https://etd.uwc.ac.za/

the Y chromosome which prevented the successive recombination of the X and Y chromosomes by disrupting alignment between both the chromosomes [73]. Once recombination had ceased between the X and Y chromosome, the Y chromosome began to decay losing most of the ancestral Y-linked genes [73,76,77]. The X chromosome managed to retain its full ancestral gene complement by recombination in the female germline with its homogametic X chromosome partner [73,76,77]. The Y chromosome does not have a recombining partner in the heterogametic male germline. Indeed, the "addition-attrition" hypothesis is used to explain the silencing of one of the X chromosomes in homogametic somatic mitotic cells. The X-linked homologues are thought to be up regulated in order to compensate for the loss of their Y-linked counterparts [71,73,76]. As embryogenesis is sensitive to gene dosage, one of the X chromosomes is silenced in the homogametic sex facilitating dosage compensation [71,76,78].

Subtle exceptions to the dissimilarity between the X and Y chromosome are the pseudoautosomal regions (PARs) which contribute ~5% of the Y chromosomal sequence, and are divided into two regions situated near the distal and proximal telomeric portions of the Y and X chromosomes [74,79-81]. The PARs on both the X and Y chromosomes are homologous and recombine during male meiosis enabling the correct partitioning of the chromosomes in male gametes[74,79-81]. The X and Y PARs have a higher recombinogenic rate relative to the rest of the human genome[74,79-81]. The known PARs are similar to autosomes in terms of gene density, diversity and base composition[74,79-81]. Like autosomal genes, PAR linked genes are shared freely between males and females [74,79-81]. Elucidation of the genomic locations for the CT genes reveal that none of the currently known CT genes are situated in the characterized PARs on the X chromosome, and none of the CT genes are located on the Y chromosome (Figure 1, pg 14).

UNIVERSITY of the
WESTERN CAPE

https://etd.uwc.ac.za/

The remaining 95% of the Y chromosome comprises of the None Recombining region on the Y (NRY), or as recently renamed, the Male Specific region of the Y chromosome (MSY), and distinguishes the sexes [74,82]. The Y chromosome has been shown to contain numerous spermatogenesis specific genes whose expression is testis restricted [74,76,82,83]. In contrast, the X chromosome houses a diverse assortment of genes, although there does appear to be a higher proportion of genes expressed in female specific tissues such as the placenta and ovary [71,78,80,81,84-86].

Why is the Y Chromosome Enriched for Testis Specific genes?

The universal theme during Y chromosome evolution appears to be the assimilation and incorporation of genes specifically benefiting male fecundity and spermatogenesis [71,74,76,80,81,83]. Ronald Fisher proposed a selective advantage in the appropriation of male specific genes detrimental to females in a male specific portion of a genome [74]. There is experimental evidence of gender specific genes which are beneficial to one sex and detrimental to the other, such genes have been termed as sexually antagonistic [87]. Fisher's model is used to account for the Y-linked ornamental genes in male *Poecilia reticulata*s (guppies) [74,88]. Ornamental genes provide male guppies with a rich tapestry of colours, while females are drab in comparison. The ornamental genes are thought to enhance male guppy attractiveness and hence fecundity. Ornamental genes are considered sexually antagonistic in female guppies as they increase the risk of predation without compensating by enhancing reproductive success [74,88].

Lahn *et al* [37,74] argue a decrease in female fecundity can be alleviated at a low metabolic cost by transcriptionally silencing male specific genes without the need of physically relocating male specific genes to the Y chromosome. The silencing of male specific genes is possible, as evidenced by the transcriptional silencing of the X chromosome

UNIVERSITY of the WESTERN CAPE
https://etd.uwc.ac.za/

Although there is much speculation on how male specific genes have become Y linked, the Y chromosome does house a large number of spermatogenesis genes and hence, testis specific genes [74,76,83]. The CT genes can be considered male specific as their expression profile is predominantly confined to the testis [29,33,35,41]. Surprisingly, none of the currently known CT genes map onto the Y chromosome, and 50% (not including CT gene family members; Figure 1 pg 14, Figure 2 pg 16) map onto the X chromosome. As the genomic location of a gene, its function and expression profile is non-random as observed by Y linked male specific genes and the clustering of housekeeping genes [43,74,88,94,96], the location of the bulk of CT genes on the X chromosome is puzzling.

Why are Testis Specific genes situated on the X Chromosome?

Using a maximum likelihood approach, Saifi and Chandra demonstrated the X chromosome contains an elevated number of genes related to sex and reproduction[81,86]. In comparison to autosomes 1, 2, 3, 4 or 11, the odds of an X linked chromosomal locus being sex or reproduction related are 4.4 times higher [86,94]. Gibson *et al* have also shown the X chromosome to be enriched for variation in sexually antagonistic fitness in *Drosophilia melanogaster* [97].

UNIVERSITY of the
WESTERN CAPE

There have been previous reports of male specific and testis specific genes residing on the X chromosome [95,98,99]. The localization of the CT genes by genomic mapping to the X chromosome strengthens these prior observations. In the aforementioned reports, Rice's hypothesis was invoked to explain this non-random male specific X chromosomal gene distribution. Rice's hypothesis stipulates the preferential accumulation of sexually antagonistic genes on the sex chromosomes. Rice's model is used to account for the abundance of male fecundity genes residing on the Y chromosome [74,98]. A further extension of Rice's hypothesis proposes a sexually antagonistic gene which is beneficial to the

https://etd.uwc.ac.za/

heterogametic sex while detrimental to the homogametic sex will spread through the population, if the sexually antagonistic gene is X linked recessive [94,98,99]. As the female chromosome spends two-thirds of its time in the female germline, the detrimental effects of a recessive gene would be masked from expression by the second wildtype dominant gene on the second X chromosome (Figure 3, pg 21)[94]. As the heterogametic sex is hemizygous for the X chromosome, the full benefit of that sexually recessive antagonistic gene will be realized, as males only have one X chromosome (Figure 3, pg 21). All the genes expressed by the single X chromosome in males would be dominant, regardless if those genes are recessive or dominant in the female germline (Figure 3, pg 21)[94]. If that sexually antagonistic gene had the reverse effect, i.e beneficial to females and detrimental to males, fixation of that gene in a dioecious population would be strongly opposed by selection due to the lack of a second X chromosome in males to mask the expression of that sexually antagonistic gene[94]. A male beneficial sexually antagonistic recessive gene which is X linked will manage to invade and achieve significant penetrance in a dioecious population. The detrimental effects for homozygous females of that sexually antagonistic gene will not be observed until the sexually antagonistic gene has reached a significant frequency in the effective population (Figure 3, pg 21).

UNIVERSITY of the
WESTERN CAPE

If the sexually antagonistic gene was autosomally linked, the gross advantage derived from that gene would need to be greater then gross detrimental fitness caused to the other sex. Only then would an autosomally linked sexually antagonistic gene increase in its frequency in a population[94]. If the sexually antagonistic gene is autosomal recessive, it would be unlikely to spread in a population as its phenotype would be masked from selection. Homozygousity of that autosomal sexually antagonistic would not spread as it would decrease the overall fitness of a specific sex due to the sexually antagonistic detrimental nature of that gene[94]. Hence, by being autosomally linked, a sexually antagonistic gene is

https://etd.uwc.ac.za/

unlikely to be seen in a population as it would be subjected to strong purifying selection pressure [94,98,99].

One of the shortcomings of Rice's hypothesis is it does not take into account genomic imprinting and epigenetic events. Genomic imprinting is the parental specific expression or repression of a gene or chromosome in a progeny with methylation postulated to be one of the main mechanisms regulating imprinting [46,100]. Imprinted genes apparently show a tendency to cluster in the human genome. A known imprinting cluster is located on chromosome 11p15.5, interestingly CT 32 (LDHC) also cytogenetically maps to 11p15.5 [46,94,100]. Whether CT 32 is imprinted or not by virtue of it genomic location is speculative as further experimental validation is required to provide a definitive answer.

Is Rice's Hypothesis Relevant to the X linked CT Genes?

The CT genes exhibit a testis preferential expression profile and can be considered male specific genes. However, a male specific and testis restricted expression profile does not translate to a gene being sexually antagonistic. The CT genes can not presently be labeled as sexually antagonistic for a number of reasons;

UNIVERSITY of the
WESTERN CAPE

1). The biological role of 97.5% of the CT genes is currently unknown which results in the inability to infer if the CT genes are beneficial or detrimental to a specific sex, based on their functional role in the human genome[33,37]. Without knowing the function of the CT genes, it would be imprudent to classify them as sexually antagonistic based simply on their tissue specific expression profile.

2). Some of the X linked genes such as CT 1.1 (MAGEA-1), CT 19 (IL-13Rα), CT 43 (FATE) amongst others, are expressed in the male specific testis, female specific placenta and ovaries as well as shared tissues between the sexes such as the pancreas [37]. Hence, although CT gene expression is primarily found in the testis, the presence of CT gene

https://etd.uwc.ac.za/

expression in other types of tissues indicates that they may not be exclusively male specific. As the CT genes may not be exclusively male specific, they might not be sexually antagonistic.

3). The CT genes are a heterogeneous group of genes unified by their restricted gene expression profiles [37]. Consequently, what would apply for one CT gene will not necessarily apply to another e.g an X linked CT gene might well be sexually antagonistic, but another X linked CT gene may well not be.

4). Rice's hypothesis, as invoked by previous studies, stipulates the accumulation of recessive sexually antagonistic genes [94,98,99]. It is not known if the CT genes are recessive or dominant genes, let alone sexually antagonistic.

5). Rice has also contended that only a few traits, in essence, can be considered sex specific[94].

Rice's hypothesis does provide an appealing model in which to explain the localization of 66% of the CT genes to the X chromosome, and the remaining 37% to twelve autosomal chromosomes. At present though, Rice's hypothesis solely can not sufficiently and necessarily explain, with regards to the CT genes, their genomic bias towards the X chromosome.

UNIVERSITY of the
WESTERN CAPE

Is the X chromosome Truly Enriched for Testis / Male Specific Genes?

The genomic mapping of the CT genes in this study indicates that the X chromosome hosts the majority of CT genes. Previous reports by Lercher *et al* [98] and Wang *et al* [99] have described similar genomic distributions for different subsets of testis and male specific genes, and hence propose an enrichment of male specific genes on the X chromosome, in line with Rice's hypothesis [95,98,99]. Wang *et al* [99] specifically focused on spermatogonial expressed genes in the mouse testis, CT 1.5 (MAGEA-5) was one of the genes used to confirm their

https://etd.uwc.ac.za/

subtractive cDNA hybridization protocol [95,99]. Lercher *et al* [98] did not examine genes present in the germline and used SAGE data from somatic male specific tissues such as the prostrate, to derive a set of male specific genes [98]. Lercher *et al*'s study identified 13 prostrate specific genes which are X linked [98]. Wang *et al*'s study uncovered 9 spermatogonial genes in mice, 7 of which have human orthologues that map to the X chromosome [99]. In total, the three studies have identified 75 human testis and male specific genes that are located on the X chromosome. Using Ensembl's known gene count of 957 genes on the X chromosome (including PARs), the X-linked CT genes together with the testis and male specific genes identified by Wang *et al* [99] and Lercher *et al* [98] represent 7.83% of the total 957 X linked genes. The 75 human testis and male specific genes which map to the X chromosome form 0.32% of the total genes in the human genome (using Ensembl's prediction of 23,531 genes).

The total number of testis and male specific genes not mapping to the X chromosome identified in this study (Figure 1, pg 14) as well as by Wang *et al* [99] and Lercher *et al* [98] is 224. Ensembl's gene count for the Y chromosome is 117 (including PARs), Y-linked genes can be considered male specific as only males carry a Y chromosome (although there are X linked homologues [73,74,82]). In total, 341 testis and male specific genes are not present on the X chromosome. In terms of the human genome, that forms 1.45% of testis and male specific genes that do not map to the X chromosome, approximately a 4.5 fold increase. Hence for every X-linked testis or male specific gene identified as in this study as well as by Wang *et al*[99] and Lercher *et al*[98], there are 4.5 male specific genes which are not X-linked.

Unfortunately, the calculations presented above are a crude approximation. There is no known gene count for the total number of testis and male specific genes present in the human genome. Although there are ongoing large scale efforts to map genes to the human genome, there is no large scale analysis providing approximations of the proportion of testis and male specific genes mapping onto each chromosome. Additionally, the calculations do not take

UNIVERSITY of the WESTERN CAPE

https://etd.uwc.ac.za/

into account the PAR regions on both the X and Y chromosome although genes in these regions are shared freely between the sexes[74,79-81]. As a result, it would appear premature to proclaim that the X chromosome is enriched for testis and male specific genes when only 1.27 % of genes present in the human genome have been used in all three studies.

Although the X chromosome may not be enhanced for testis and male specific genes from a genomic perspective, this study and two previous ones indicate that testis specific and male specific genes do reside on the X chromosome [98,99].

Why Should Testis Specific genes not be Present On the X Chromosome?

As the X chromosome spends two-thirds of its time in the female germline coupled with an asymmetric transmission of a paternal X chromosome to male progeny, there should be a scarcity of male specific genes such as testis specific genes. The "demasculisation" of the X chromosome has been observed for *Drosophilia melanogaster, Anopheles gambiae* and *C. elegans* [51,85]. Positive selection is postulated to lead to the feminization of the X chromosome due to the disproportional time spent by the X chromosome in the female germline. Concurrently, negative selection is also thought to act on the X linked male beneficial genes detrimental to females, further feminizing the X chromosome [51,51,85,85,101].

UNIVERSITY of the WESTERN CAPE

Charlesworth *et al* have suggested translocations of autosomal regions to the X chromosome could account for the presence of male specific genes on the X chromosome [94,102]. The translocations they argue, would be favored if the translocated regions comprised of sexually antagonistic genes[94,102].

Conversely, a recent study by Emerson *et al* has demonstrated an excess of (~300%) retrogenes moving from the human X chromosome to the autosomes [72]. Retrogenes are defined as genes that are derived from the reintegration of reverse-transcribed mature mRNA into the human genome, a process known retroposition[72]. The CT 1 (MAGEA) gene family is

https://etd.uwc.ac.za/

also postulated to have arisen by retroposition and duplication from an ancestral MAGE-D gene[66]. Emerson *et al* estimate that roughly 77% of autosomal retrogenes derived from the X chromosome have a testis expression profile, in contrast to 44% of retrogenes derived from autosomes[72]. The results the authors argue, is consistent with both natural selection in an attempt to achieve a male germline character and a mutational bias[72]. A possible motive postulated by Emerson *et al* for the elevated retroposition of genes from the X chromosome to autosomes is meiosis [72]. During meiosis in the testis, the X chromosome condenses and is transcriptionally silenced [89]. Consequently, X linked genes are not expressed at the onset of male meiosis from stage I in spermatogenesis, the paternal X chromosome is only reactivated after fertilization with an ovum [89]. The transcriptional suppression of X linked genes during meiosis in the male germline is compensated by the expression of autosomal genes [72,89]. In relation to the CT genes, the only two whose functions are known are CT 23 (OY-TES-1) and CT 8 (SYCP) [33,40]. Both CT 23 (OY-TES-1) and CT 8 (SYCP) are actively involved in the male germline during meiosis, CT 23 (OY-TES-1) is engaged in the acrosin packaging of the acrosome in the late stages of sperm maturation and CT 8 (SYCP) is involved in homologous chromosomal pairing during the meiotic prophase of spermatocytes [33,40]. CT 23 (OY-TES-1) and CT 8 (SYCP) map to chromosome 12 and 1 respectively while CT 15 (ADAM2 / Fertilin β), which is also involved in mammalian meiosis is located on chromosome 8 [72]. Hence, it would appear that if a CT gene is involved in meiosis, it is unlikely to be X linked. Interestingly, from Emerson's *et al*'s supplementary data, the only X linked CT gene with an autosomal retrogene located on chromosome 5 in Humans is CT 40 (TAF7L)[72]. These results are consistent with an alternative model, as outlined by Parisi *et al* [85], which predicts that sexually antagonistic genes are likely to be found on the autosomes due to feminization of the X chromosome.

UNIVERSITY of the
WESTERN CAPE

https://etd.uwc.ac.za/

Which Hypothesis is Pertinent to the CT Genes On the X Chromosome?

Khil *et al* seek a reconciliation between both the conflicting hypotheses; Rice's which predicts an enrichment of male specific genes on the X chromosome, and the second model outlined by Parisi *et al* which predicts the demasculisation of the X chromosome [84,85,94,101]. Using micro-array experiments and data, Khil *et al* demonstrated that spermatogenesis genes in mouse testis are depleted on the X chromosome, where as genes expressed in exclusively female tissues such as the ovaries and placenta were preferentially located on the X chromosome [84]. Hence, Khil *et al*'s results supports the feminization of the X chromosome[84,101]. When Khil *et al* repeated their approach using knockout mice for spermatogenesis before the onset meiosis I, they found the opposite[84]. By comparing their contrasting results, Khil *et al* demonstrated that testis specific genes expressed on the X chromosome are abundant before the onset of X chromosome inactivation that occurs during male meiosis[84,89]. Once meiotic sex chromosome inactivation has taken place, the X chromosome becomes depleted for testis specific expressed genes [84]. Hence, Rice's hypothesis would prevail before the onset of meiotic sex chromosome inactivation in the testis, resulting in the X chromosome being enriched for testis specific genes[84,98,99,101]. The second model that predicts the autosomal linkage for testis specific genes and the feminization of the X chromosome, comes into force once meiotic sex chromosome inactivation has occurred[72,84,85,101]. Khil *et al* [84] postulate two opposing evolutionary forces dictating the distribution of male specific genes in the genome. One to remove male specific genes utilized in meiosis from the X chromosome, a second for the sequestering of male specific genes on the X chromosome [84].

Gene expression for the CT genes is derived from the testis and more specifically primordial germ cells such as spermatogonia which are the cell types Wang *et al* [99] used in

UNIVERSITY of the WESTERN CAPE

https://etd.uwc.ac.za/

their study. In this scenario, it would appear that the X chromosome is enhanced for CT genes due to the over representation of testis specific genes before the onset of meiotic sex chromosome inactivation. Hence the spatial and temporal expression profiles of the CT genes appear to be influenced by their genomic locations.

https://etd.uwc.ac.za/

Chapter 3

Is CT Gene Expression Testis Specific?

A recent publication by Scanlan *et al* indicates CT gene expression may not be exclusively testis restricted [37]. In order to investigate these observations, an *in-silico* approach was undertaken to ascertain whether the CT genes exhibit a high degree of tissue tropism.

Gene Expression Capture – A Question of Methods?

A multitude of technologies can be used to analyze gene expression at the post transcriptional level by assaying for messenger ribonucleic acid (mRNA) in a biological sample. These types of gene expression technologies have a broad spectrum of variation in terms of their breadth and depth of gene expression capture. High-throughput technologies include Expressed Sequence Tags (ESTs), Serial Analysis of Gene Expression (SAGE) and micro-array analysis. Small scale specialized gene expression technologies include Reverse Transcriptase Polymerase Chain Reaction (RT-PCR) and Northern Blotting. Each type of technology imposes its own artefact of gene expression capture onto the data. A key difference between the high throughput gene expression technologies and specialized gene expression technologies such as RT-PCR, is their window of coverage and resolution in terms of depth during the capture of gene expression events. High throughput technologies such as ESTs provide an expansive window of gene expression coverage. However, due to this broad but biased coverage of gene expression capture, the resolution of the depth of gene expression is shallower compared to RT-PCR. RT-PCR provides a much narrower window of gene expression coverage due to the explicit sampling of chosen genes by the use of gene specific primers. The depth in resolution provided by RT-PCR principally arises through the quantification of the concentration of mRNA transcripts present in a cell, calculated on the number of PCR cycles.

https://etd.uwc.ac.za/

Methods.

Command-line queries using the eVoc Query Language (EQL) were used to produce normal "virtual gene expression libraries" of genes expressed in different anatomical tissues. When a cDNA library is annotated as "normal" in eVoc, the contributing record for that cDNA library is given as normal tissue, i.e derived from non-diseased tissue. Normal virtual gene expression libraries were produced for 13 non-gametogenic tissues; bone marrow, brain, colon, heart, kidney, liver, lung, pancreas, prostrate, skeletal muscle, small intestine, spleen and thymus as well as 3 gametogenic tissues; testis, ovary and placenta. The 16 tissues assayed for CT gene expression using eVoc were chosen in order to directly compare the results with the RT-PCR results of CT gene expression in the same 16 tissues conducted by Scanlan *et al* [37]. The virtual tissue specific gene libraries comprise of accession numbers from the curated RefSeq database[58]. The EQL query returns a list of transcripts present in each cDNA library which has been manually annotated as normal for the specific tissue types queried. Scanlan *et al*'s[37] study made use of commercially available tissue panels labeled as normal by virtue of being derived from disease-free individuals. All the virtual tissue specific gene libraries were parsed to in order to determine the presence / absence of 43 of the 44 founding CT members using their RefSeq accession identifiers. CT 18 (NA88A) was excluded from the analysis of the results due to its RefSeq pseudogene accession identifier (NR_001559) as well as CT 30 (HCA661) whose tissue expression profiles could not be determined by RT-PCR[37]. The tissue expression results obtained through eVoc for 43 of the CT genes were sorted into the 4 CT gene expression categories devised by Scanlan *et al* based on the detection of CT gene expression by RT-PCR in a panel of 16 tissues[37]. Agreement between the *in-silico* and RT-PCR results was calculated as a percentage based on the presence / absence of CT gene expression by both types of gene expression technologies for 42 CT genes in all 16 tissues.

UNIVERSITY of the
WESTERN CAPE

https://etd.uwc.ac.za/

Comparison of an *In-silico* and a Wet-lab CT Gene Expression Study.

Four CT gene expression categories were devised by Scanlan *et al* [37] that are based on the quantity of tissue types in which mRNAs corresponding to 43 of the founding CT gene members were found present, in a panel of 16 tissue types assayed. The CT gene expression categories, together with number of tissues that constitute a specific expression category are summarized below, in Table 3.

CT Gene Expression Category	Tissue Distribution of the CT genes
1	Testis Restricted.
2	≤ Two non-gametogenic tissues.
3	≤ Six non-gametogenic tissues. Differentially Expressed.
4	≥ Six non-gametogenic tissues. Ubiquitously Expressed.

Table 3 : Four CT gene expression categories were developed by Scanlan et al 37 to reflect the number of tissues from a panel of 16 tissue types in which CT gene expression was detected by RT-PCR.

UNIVERSITY of the
WESTERN CAPE

Tissue specific comparisons of the results between an *in-silico* approach used in this study and an RT-PCR approach undertaken by Scanlan *et al* [37] for determining CT gene tissue tropism is shown overleaf, in Figure 4. Both types of technologies showed an agreement of 73% which is based on the absence (54%) or presence (19%) of CT gene expression for the 42 CT genes in 16 of the tissue types assayed for CT gene expression by RT-PCR [37] and eVoc (Figure 4). Changes in the 4 categories of CT gene expression from the RT-PCR based results[37] to the eVoc based results for each of the 42 CT genes was determined (Figure 4). The changes from the RT-PCR based CT gene expression categories[37] to the eVoc based CT gene expression categories were found to comprise of 4 different types.

https://etd.uwc.ac.za/

6 normal tissue types.

CT Gene	Thymus	CT Gene Expression Category eVoc	CT Gene Expression Category RT - PCR	Change in CT Gene Expression Category.
[redacted]	1	4	4	0
CT 29 D40	1	4	3	1
CT 32 LDHC	1	3	1	2
CT 10 MAGEE1	X	3	1	2
CT 11.3 SPANXC1	X	3	2	1
CT 12.1a GAGED	X	4	3	1
CT 13 HAGE	6	3	2	1
CT 14 SAGE	X	2	1	1
CT 15 ADAM2	8	2	2	0
CT 16 PAGE-5	X	2	3	-1
CT 19 IL-13R-alpha	X	4	4	0
[redacted]	X	3	2	1
CT 20 TSP50	3	3	4	-1
[redacted]	[redacted]	2	1	1
[redacted]	[redacted]	4	4	0
[redacted]	X	3	1	2
CT 25.1a MMA-1	2	2	1	1
[redacted]	X	2	3	-1
CT 27 BORIS	2	3	3	0
[redacted]	[redacted]	1	1	0
[redacted]	[redacted] N/A	2	Not Available	Not Available
CT 31 PLU1	[redacted]	4	4	0
CT 33 MORC	[redacted]	2	1	1
CT 34 SGY-1	[redacted]	2	2	0
CT 35 SPO11	[redacted]	2	1	1
CT 36 TPX1	[redacted]	3	1	2
[redacted]	[redacted]	2	2	0
CT 39.1a NFX2	[redacted]	2	3	-1
[redacted]	[redacted]	2	1	1
CT 40 TAF7L	[redacted]	2	3	-1
CT 41.1 TDRD1	[redacted]	3	2	1
CT 42 TEX15	[redacted]	2	3	-1
[redacted]	[redacted]	2	3	-1
[redacted]	[redacted]	2	1	1
[redacted]	[redacted]	4	1	3
[redacted]	[redacted]	2	2	0
[redacted]	[redacted]	2	2	0
[redacted]	[redacted]	2	2	0
CT 9 BRDT	[redacted]	3	1	2
[redacted]	[redacted]	2	1	1
[redacted]	[redacted]	0	1	-1
[redacted]	[redacted]	0	1	-1
CT 17 LIP1	[redacted]	0	1	-1

UNIVERSITY of the WESTERN CAPE

Figure 4 : The results

The green squares in

a specific tissue type

https://etd.uwc.ac.za/

The 4 different types of changes in the number of CT gene expression categories observed for 42 of the CT genes from the RT-PCR determined CT gene expression categories[37] to the eVoc CT gene expression categories are presented in Table 4.

Change in the Number of CT Gene Expression Categories	Percentage of 42 CT Genes observed in each CT Gene Expression Category change
-1	23.8%
0	28.6%
1	33.3%
≥ 2	14.3%

Table 4 : Four types of changes in the number of CT gene expression categories from RT-PCR to eVoc were observed, together with the corresponding percentage of CT genes present in each type of CT gene expression category change.

Of the 42 CT genes for which a change in the number of CT gene expression categories from RT-PCR to eVoc was determined, 30 (71.4%) exhibit a change in CT gene expression categories (Figure 4, pg 35, Table 4). Changes in the number of CT gene expression categories between the RT-PCR and eVoc results is expected due to the inherent differences in the breadth and depth of gene expression capture by both methods. However, due to the broader sampling of gene expression data derived from the public domain by eVoc, a decrease in a CT gene expression category (-1) from RT-PCR to eVoc of 23.8% of the CT genes is perplexing (Figure 4, pg 35, Table 4). The unraveling of the transcriptome is an ongoing project and although there are vast amounts of gene expression data present in the public domain, it is by no means complete. Some tissue libraries in the public domain do have more cDNA libraries present than others which would affect the eVoc results due to data sampling biases[♣]. The more cDNA libraries sequenced and annotated for a specific

UNIVERSITY of the WESTERN CAPE

[♣] Personal Communication ; Professor W. Hide and Dr. J. Kelso

https://etd.uwc.ac.za/

tissue, the higher the chance of a specific gene transcript being detected in that tissue as the depth of gene expression sampling is likely to increase due to more clone libraries being annotated. The total number of 43 CT genes found present in each of the 16 tissues assayed by eVoc (Figure 4, pg 35), as well as the number of cDNA libraries annotated as "Normal" in eVoc for the 16 tissues assayed for CT gene expression was plotted using a Log_{10} scale and is presented in Figure 5.

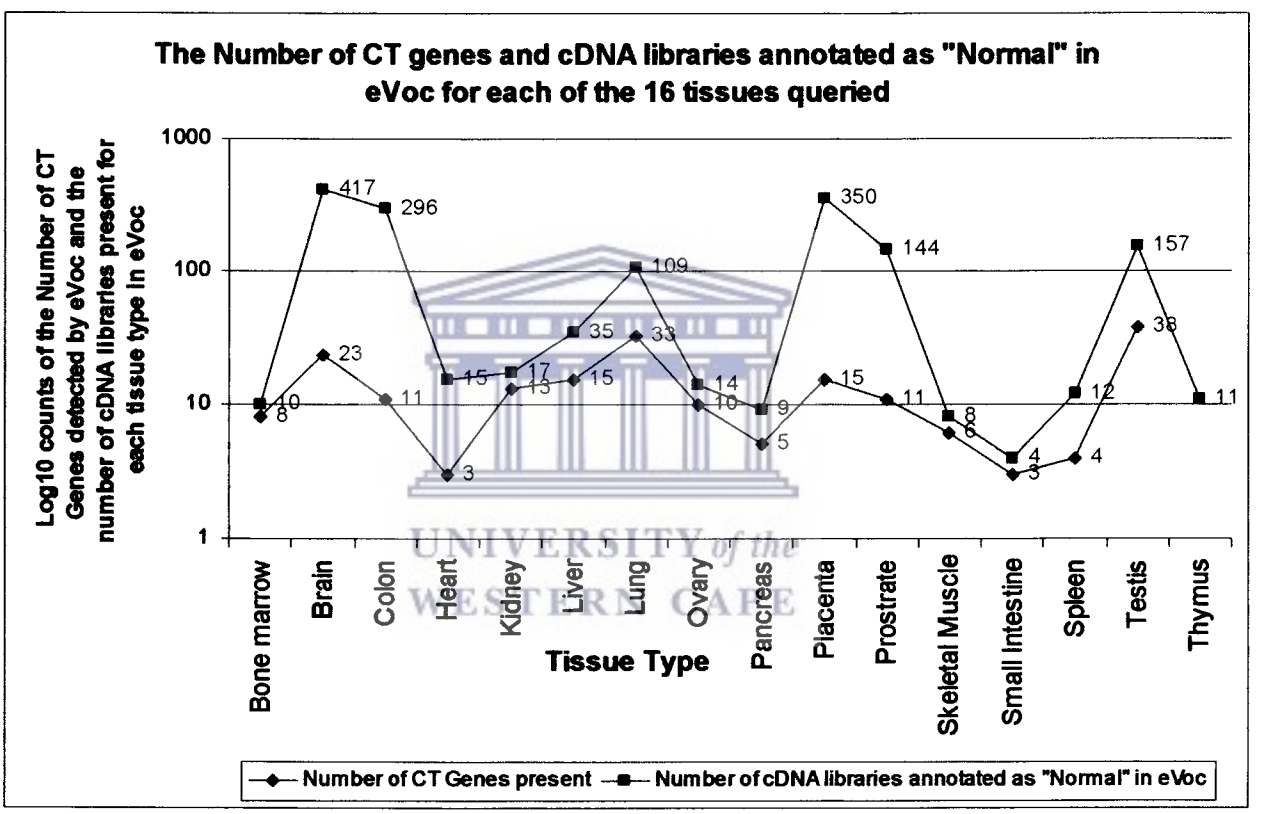


Figure 5 : The number of CT genes that were detected in the 16 tissue types (shown as red) increases as the number of cDNA libraries that have been annotated for a specific tissue in eVoc (shown as blue) increases. The number of cDNA libraries annotated as "Normal" for the 16 tissues in eVoc was kindly provided by Simon Cross, an eVoc developer.

The number of CT genes detected by eVoc in the different tissue types roughly increases in tissues which have a higher number of annotated cDNA libraries, exceptions being the colon, heart, prostrate and thymus, Figure 5. The brain and lung tissue types queried by eVoc have 417 and 109 cDNA libraries which are annotated as "Normal", the total number of 43 CT genes detected in the brain and lung tissues are 23 and 33 respectively

https://etd.uwc.ac.za/

(Figure 4, pg 35, Figure 5, pg 37). In some instances, CT gene expression was detected by RT-PCR and not eVoc in tissues such as the pancreas and the small intestine (Figure 4, pg 35), the number of cDNA libraries annotated as normal present in eVoc for the pancreas and small intestine is 9 and 4 (Figure 5, pg 37). Additionally CT gene expression in some tissues such as the thymus was not found in eVoc, but was found to be present by RT-PCR (Figure 4, pg 35). The thymus has 11 annotated cDNA libraries as compared to the 4 annotated cDNA libraries for the small intestine in eVoc, nevertheless, CT gene expression could only be detected for the small intestine and not the thymus in eVoc (Figure 4, pg 35, Figure 5, pg 37).

Factors that may contribute to the lack of CT gene expression detection by high throughput methods is that cells do not express all the genes in their genome singularly and at a fixed quantity. Thus, it is possible that in some instances the CT genes may represent rare transcripts in different specific tissues and hence, are not detected by some high throughput methods due to their low levels of transient gene expression[♣]. The detection of these rare transcripts by high throughput methods would be hindered by the innumerous diverse transcripts present in a cell as well as the relative scarcity of these rare transcripts. RT-PCR does enable the detection of specific gene expression through the use of gene specific primers enabling the exponential amplification of a specific gene transcript. Hence, in certain cases where the CT genes may form rare transcripts in a specific tissue, CT gene expression in these specific tissues would be better assayed for by RT-PCR which provides a higher resolution of gene expression in terms of depth.

The 28.6% of CT genes which did not exhibit a change in CT gene expression categories from RT-PCR to eVoc do show a general agreement between the tissues in which

---

[♣] Personal Communication ; Professor W. Hide and Dr. J. Kelso

UNIVERSITY of the WESTERN CAPE

https://etd.uwc.ac.za/

the CT gene expression was detected by both types of technologies (Figure 4, pg 35). Examples include CT 31 (PLU1) which is ubiquitously expressed in all tissues tested for, and CT 28 (HOM-TES-85) which displays a testis restricted expression profile by both RT-PCR and eVoc (Figure 4, pg 35). However, in some instances there might be no change in CT gene expression categories (Table 4, pg 36), there are differences in the types of tissues in which CT gene expression was found between eVoc and RT-PCR e.g CT 34 (SGY-1) was found by RT-PCR to be present in the testis, ovary, spleen and pancreas while in eVoc, CT 34 (SGY-1) gene expression was only detected in the testis and brain (Figure 4, pg 35).

From the 42 CT genes for which a change in CT gene expression categories was observed from RT-PCR to eVoc, 33.3% exhibit a change of 1 CT gene expression category (Figure 4, pg 35, Table 4, pg 36). With regards to an increase in 1 CT gene expression category from RT-PCR to eVoc, there are three types of increases that can occur ($n + 1$ where n is a CT gene expression category in Table 3, pg 34). Of the 14 CT genes which were found to have increased by one CT gene expression category from RT-PCR to eVoc, 8 exhibit a change from CT gene expression category 1 to CT gene expression category 2 from RT-PCR to eVoc (Figure 4, pg 35, Table 3, pg 34). Expression for the majority of these 8 CT genes classified as "Testis Restricted" by RT-PCR and placed into CT gene expression category 2 by eVoc was found in the brain, lung and to a lesser degree in the liver and kidney by eVoc (Figure 4, pg 35, Table 3, pg 34).Detection of CT gene expression in the lung and brain is likely to be due to the higher number of annotated cDNA libraries present in eVoc for these tissues, Figure 5, pg 37.

The fourth type of CT gene expression category change observed from RT-PCR to eVoc involves two or more category changes of CT gene expression ($n + 2$ or $n + 3$ where n is a CT gene expression category in Table 3, pg 34).There are 6 CT genes involved in this type of CT gene expression category changes; CT 5.1 (SSX1), CT 9 (BRDT), CT 10

UNIVERSITY of the WESTERN CAPE

https://etd.uwc.ac.za/

(MAGEE1), CT 24.1 (CSAGE), CT 32 (LDHC) and CT 36 (TPX1), Figure 4, pg 35. All 6 CT genes have been placed in CT gene expression category 1 by RT-PCR (Figure 4, pg 35). Although the in depth resolution of RT-PCR did not detect the expression of these 6 CT genes in tissues other then the testis (Figure 4, pg 35), a broader survey of publicly available gene expression data by eVoc indicates that these 6 CT genes are expressed in cDNA libraries annotated as normal, for tissues other then the testis, Figure 4, pg 35. CT gene expression for these 6 CT genes was also detected in the brain and lung by eVoc, Figure 4, pg 35.

Are the CTAs Testis Specific?

The rationale for using CTA as cancer vaccines lies in their ability to elicit an immunological response and their testis restricted expression profile[31,32,34,36,42]. The testis is an immuno-privilaged site as it lacks the expression of major histocompatability complexes and thus unable to present cell surface antigens[31,32,34,36,42]. A testis restricted expression profile coupled with an immunogenic profile in cancerous cells provides an ideal platform for the development of cancer vaccines by circumnavigating the possibility of inducing an autoimmunogenic reaction[31,32,34,36,42].

UNIVERSITY of the
WESTERN CAPE

The 19 CTAs together with their CT gene expression categories determined by RT-PCR and eVoc are shown in Figure 6, page 41. Concordance of CT gene expression categories from RT-PCR to eVoc occurs for 39% of 18 CTAs for which CT gene expression was assayed for by RT-PCR and eVoc (Figure 6, pg 41). Of the 18 CTAs, 28 % show an increase in one CT gene expression category, 11 % show a change in two or more CT gene expression categories and 22 % exhibit a decrease in one CT gene expression category change from RT-PCR to eVoc, Figure 6 pg 41. CT 30 (HCA661) represents the 19th CTA whose CT gene expression category was determined by eVoc but was not included in the comparisons due to the absence of RT-PCR results[37].

https://etd.uwc.ac.za/

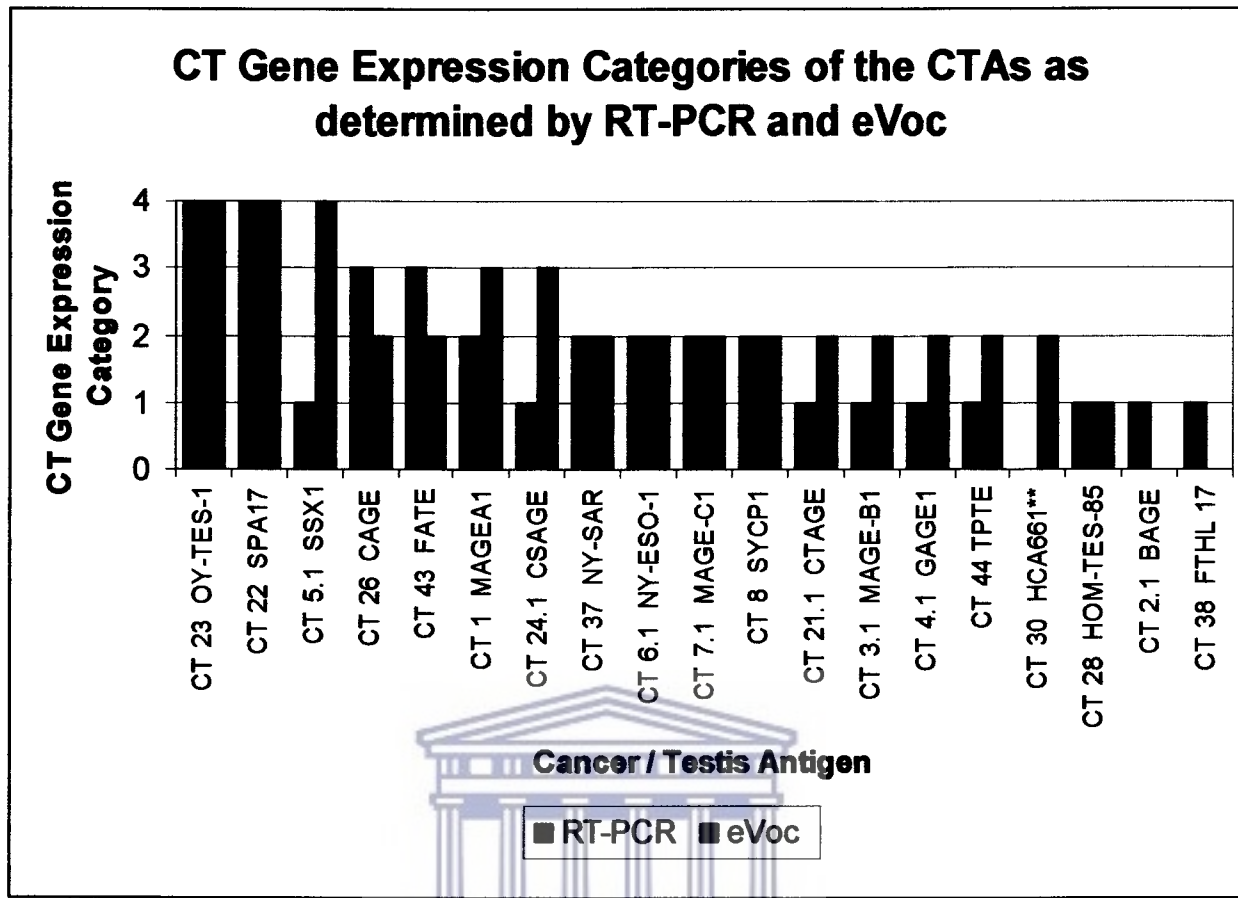


Figure 6 : A comparison of CT gene expression categories for 18 of the 19 founding members of the CTA families as determined by RT-PCR37 and eVoc. **Comparisons in CT gene expression categories between RT-PCR and eVoc results for CT 30 HCA661 were not made due to an absence of RT-PCR results[37].

UNIVERSITY of the WESTERN CAPE

Although 39% of the CTAs show concordance in CT gene expression categories, the tissues in which gene expression for these 7 CTAs is detected by both RT-PCR and eVoc do show variations, Figure 4 pg 35. Four of the 18 CTAs were found to have decreased by one a CT gene expression from the RT-PCR results to the eVoc results, Figure 6. In the case CT 26 (CAGE) and CT 43 (FATE), expression was found by RT-PCR in tissues such as the pancreas and the thymus but was not detected by eVoc (Figure 4, pg 35), hence the decrease in one CT gene expression category from RT-PCR to eVoc (Figure 6). Both CT 2.1 (BAGE) and CT 17 (FTHL) were found to be testis restricted by RT-PCR[37], expression for these two CTAs was not found in any of the 16 tissues assayed for CT gene expression by eVoc

https://etd.uwc.ac.za/

(Figure 4, pg 35) and hence the observed decrease of one CT gene expression category for these 2 CTAs from RT-PCR to eVoc (Figure 6, pg 41). The remaining 7 CTAs show an increase in CT gene expression categories from RT-PCR to eVoc, with 2 CTAs exhibiting a change of more then one CT gene expression category (Figure 6, pg 41). Apart from CT 21.1 (CTAGE), the 6 CTAs which exhibited an increase in CT gene expression categories from RT-PCR to eVoc were found to be expressed in the brain and lung by eVoc (Figure 6, pg 41, Figure 4, pg 35).

Thus, apart from CT 28 (HOM-TES-85), the remaining CTAs have an expression profile that is not exclusively restricted to the testis, as determined by eVoc (Figure 6, pg 41, Figure 4, pg 35). However, considering the broad coverage of gene expression offered by high throughput technologies and the accumulating volume of gene expression data present in the public domain, it is not unfeasible to accept a change in one CT gene expression category from the RT-PCR results to the eVoc results. The reason being that a narrow, in depth high resolution gene expression assay like RT-PCR revels that these CTAs cannot be quantified in those tissues that the CTAs were identified by eVoc. Indeed, CT 6 (NY-ESO-1) which falls into CT gene expression category 2 by eVoc and RT-PCR (Figure 6, pg 41), but shows differences in tissue distribution between both types of technologies (Figure 4, pg 35) has recently been used in clinical trials. The authors found CT 6 (NY-ESO-1) to be safe, well tolerated and strongly immunogenic[104,105].

UNIVERSITY of the
WESTERN CAPE

https://etd.uwc.ac.za/

Conclusions.

The results of the following *in-silico* study are consistent with the results from a wet-lab study conducted by Scanlan *et al* [37] in that CT gene expression can be detected in a variety of non-gametogenic tissues[37]. The following *in-silico* study does indicate that CT gene expression is present in tissues other then the testis (Figure 4, pg 35). Expression for a large proportion of CT genes was found to be present in the brain and lung amongst other tissues by eVoc (Figure 4, pg 35). The large proportion of CT genes whose expression was detected in the lung and brain by eVoc may be attributable to the higher number of annotated cDNA libraries present for these tissues (Figure 5, pg 37). Although CT gene expression may not be exclusive to the testis, CT gene expression does exhibit a marked tissue tropism for the testis as the expression for 90.7% of 43 CT genes was detected in the testis by eVoc (Figure 4, pg 35). The high proportion of CT genes that exhibited increases in CT gene expression categories (47.6% of 42 CT genes, Table 4, pg 36) from the RT-PCR based results to the eVoc results is mainly due to the broader sampling of publicly available gene expression data by eVoc. An in depth CT gene expression survey quantifying the level of CT gene expression in non-gametogenic tissues by RT-PCR indicates that the CT genes are expressed at "exceptionally" low levels in these non-gametogenic tissues[37].

In order to determine the tissue expression profiles of genes, *in-silico* technologies such as eVoc provide a powerful resource for mining the rapidly accumulating gene expression data present in the public domain. Compared to some high throughput methods, wet-lab gene expression technologies such as RT-PCR are advantageous in that it provides a deeper resolution of gene expression, principally due to the explicit sampling and quantification of selected transcripts. An advantage that *in-silico* technologies such as eVoc have compared to wet-lab technologies such as RT-PCR, is that the tissues in which the expression of a gene is being assayed for does not have to be defined *a priori*.

UNIVERSITY of the WESTERN CAPE

https://etd.uwc.ac.za/

Chapter 4.

Is CT Gene Expression Influenced by their "Genomic Neighborhood[44]"?

Modulation of gene expression is a logistically co-ordinated process occurring through the dynamic interplay of a multitude of factors [45,106,107]. Factors influencing the modulation of gene expression can be directly sequence based [23] e.g promoters, enhancers, repressors, or structurally based [45,106,107] e.g chromatin structure influencing the aptness of transcription factors to their cognate binding sites, or epigenetically based [5,46,55,56] e.g methylation, imprinting. Data generated from high throughput gene expression capture methods like EST or SAGE permits the quantification of a gene's expression spectrum at the post transcriptional level [108,109]. A gene's expression spectrum is measured in terms of the range of tissues (breadth), and the level (depth) at which a gene is expressed based on the quantity of mRNAs (transcripts) detected for a specific gene [108,109]. In terms of gene expression breadth, a gene can arbitrarily be placed into two categories depending on the number of tissues a gene's transcript is recorded in [96,108]. Genes that exhibit a broad spectrum of tissue expression are ubiquitously expressed e.g housekeeping genes like ALAS1 [110]. Genes that exhibit a narrow spectrum of tissue expression are differentially expressed e.g the CT genes[37].

The human genome sequence provides a physical map for the locations of genes present in the human genome [6]. High throughput gene expression data coupled with the positional information of genes present in the human genome, enables the delineation of genomic regions based on the spectra of expression exhibited by genes present in those genomic regions [96,108,111,112]. The mapping of gene expression data to the human genome discerned genomic regions housing genes which exhibit similar expression spectra[96,108,111,112].

UNIVERSITY of the WESTERN CAPE

https://etd.uwc.ac.za/

These genomic regions of similarly expressed genes indicate the distribution of gene order in terms of their expression spectra in the human genome is non-random [96,108,111-113].

Caron *et al* [96] determined that some highly expressed genes spatially congregate in the human genome. Approximately fifty genomic regions of increased gene expression (RIDGEs) were identified, thirty of which were subsequently reported as having greater gene densities and a higher GC content than genomic regions harboring genes with a narrower spectra of expression [111]. Through the use of EST data, Bortoluzzi *et al* [114] and Dempsey *et al* [115] observed the positional clustering of genes with a narrow spectrum of tissue expression in the human genome. The skeletal muscle genes studied by Bertoluzi *et al* [114] form clusters on chromosome 17, 19, X and the cardio-vascular specific genes identified by Dempsy *et al* [115] form clusters on chromosomes 21 and 22.

In contrast, utilizing a combination of SAGE and EST data Lercher *et al* [108] found that genes which exhibit a narrow spectrum of tissue expression do not form clusters in the human genome. Instead, genes which exhibited a broad spectrum of tissue expression were found to cluster in the human genome, the broadly expressed clustered genes were also discovered to be expressed at high levels, hence the observation of RIDGEs by Caron *et al* [96,108]. Clustering of genes with similar expression spectra in genomic domains have been discerned in *Drosophilia melanogaster* [43,47]. An estimated 80% of imprinted genes in mammals physically resolve into clusters [46]. Prokaryotic organisms also have similar genomic regions which are transcribed as a unit and gene order is organized into structures termed operons, the best characterized example being the *Lac* operon in the gram negative bacteria *Escherichia coli*.

As posited by the authors in all the studies mentioned thus far, the observed partitioning of genes with similar expression spectrums to genomic domains is likely to be symptomatic of the regulation of those genomic domains, in which the distribution of gene

UNIVERSITY of the WESTERN CAPE

https://etd.uwc.ac.za/

Methods.

The highest congruence of CT genes (66%) occurs on the X chromosome (Figure 1, page 14). On the X chromosome itself, the cytogenetic band Xq28 is the densest CT gene region hosting 31% of all the X linked CT genes. The 17 CT genes on the Xq28 region are subdivivded into four CT gene families [37]; CT 1 (MAGEA), CT 6 (NY-ESO-1/LAGE), CT 24 (CSAGE) and CT 43 (FATE). All the known genes mapped to genomic portion of Xq28 in the Ensembl database [59,60] (version 19.34b.2) were obtained by the use of EnsMart [116] (version 19.2) with the focus on Ensembl genes. The Xq28 dataset attributes selected through EnsMart were known Xq28 linked genes only, the start and end position of each Xq28 linked gene on the X chromosome in base pairs, Ensembl [59], RefSeq [57] and OMIM [117] accessions, an external gene identifier, as well as %GC and strand. The dataset exported by EnsMart is linear i.e the gene order of the Xq28 linked genes is preserved. The resulting dataset comprising of 112 genes was manually curated in order to remove redundancies arising from gene transcripts mapping to multiple locations of the Xq28 region using a combination of WU-BLAST from Ensembl's Blast page and the GeneLoc database [63]. This resulted in a non-redundant set of 102 known Xq28 linked genes whose gene order was preserved (Table B pg 85, Appendix).

UNIVERSITY of the
WESTERN CAPE

The preservation of gene order in the dataset is fundamental because although gene expression is a dynamic process, the physical location of a gene in a genome at a given specific time is inert. The tissue expression spectra for all of the obtained Xq28 linked genes were determined qualitatively through the use of the available published literature and database annotations of the Xq28 linked genes. The OMIM database served as the primary portal in identifying the published literature for the individual Xq28 linked genes [117]. Additionally, database annotations from SwissProt [118] (release 42.11) and GeneLynx [119]

https://etd.uwc.ac.za/

(release 1.99) of the individual Xq28 linked genes were incorporated in establishing the tissue expression spectrum of an Xq28 gene.

Due to the heterogeneity of tissues in which the individual Xq28 linked genes are expressed, a broad classification scheme was devised. If an Xq28 linked gene exhibited a narrowly expressed tissue spectrum i.e the expression of a specific Xq28 linked gene was reported in only a few tissues, it was categorized as "Differentially Expressed". If a wide tissue expression spectrum was recorded for an Xq28 linked gene, it was categorized as "Ubiquitously Expressed". All the Xq28 linked CT genes exhibit a narrow spectrum of tissue expression and hence are classified as differentially expressed [37]. The spectrum of tissue expression for some Xq28 linked genes could not be resolved, these genes were placed in a third category classed as "Expression Undetermined". The 3 expression categories devised are simplistic, but have the advantage of being mutually exclusive. A disadvantage of the classification scheme devised is that it is qualitative and hence provides no physical metric by how much "differentially" or "ubiquitously" a gene is expressed. An attempt to incorporate Gene Ontology (GO) classifications [120] was not viable as the CT genes are placed into three uninformative GO categories; GO:0005554, GO:0000004, GO:0008372 – molecular function unknown, biological process unknown and cellular component unknown, respectively.

UNIVERSITY of the
WESTERN CAPE

In order to determine if there are regions of similar expression in the Xq28 region, gene expression clusters of physically adjacent genes were identified. A gene expression cluster is defined as two or more physically adjacent genes exhibiting the same expression category (either Ubiquitously or Differentially Expressed). The third category, Expression Undetermined, was masked from the subsequent analysis of the Xq28 region. However, the spatial information of genes in the Expression Undetermined category was preserved when establishing if there are clusters of similar gene expression in the Xq28 region. Retention of

https://etd.uwc.ac.za/

the spatial information of genes in the Expression Undetermined category enables the preservation of gene order in the Xq28 region and excludes a potential source of bias. If the spatial information of genes falling into the Expression Undetermined category is removed, there would be a merging of expression clusters which would not be representative of the Xq28 region as gene order is not maintained. An offset of preserving the spatial position of genes in the Expression Undetermined category during the identification of expression clusters is that any clusters of gene expression found would tend to be smaller and more numerous. However, when the tissue expression spectra of Expression Undetermined genes are finally resolved, they can only fall into one of the two mutually exclusive expression categories devised.

The experimental design employed differs from previous studies by using a "genome to expression" approach[121]. Previous studies have used an "expression to genome approach" by mapping specific sets of tissue gene expression data to the genome [43,96,108,111-115]. In the "expression to genome" approach, the gene expression data derived from a fixed number of tissues was defined *a priori* [43,96,108,111-115]. Depending on what sets of tissues expression data and hence genes used, there may or may not be any clustering of genes with similar expression spectra in the genome. The use of a pre-defined genomic region such as Xq28 would overcome the potential variability of mapping heterogeneous tissue expression datasets to the genome as the physical locations of genes in the genome remain relatively static [121]. Like quantitative methods used to determine if the position and observed expression spectra of a gene in a genome is non-random, the qualitative method used does have some drawbacks. Possible drawbacks of quantitative methods using EST data for determining the breadth of gene expression include a shallow sampling of transcripts represented in a cDNA library, the quantity of sequences derived and possible 5' or 3' end sequencing of DNA prejudices [109,122]. A potential source of bias in quantitatively using

UNIVERSITY of the WESTERN CAPE

https://etd.uwc.ac.za/

SAGE data for determining the level of gene expression is the potential over representation of GC rich sequences in some experimental cases, if not corrected for [113,123]. A drawback of the qualitative methodology used in this study is a gene's expression spectra can not be quantified unless SAGE or EST data is utilized. The qualitative classification of the Xq28 linked genes' expression spectra relies heavily on the published literature and database annotations of each of the genes. It is possible that some genes are broadly expressed, but have their tissue expression spectra characterized as differentially expressed, as in the case of some of the CT genes [37]. Some genes which may have been classified as ubiquitously expressed are actually, just widely expressed. The methodology employed is not amendable to the analysis of very large genomic regions comprising of thousands of genes unless some form of text mining is used in conjunction.


UNIVERSITY of the
WESTERN CAPE

https://etd.uwc.ac.za/

A Panorama of the Xq28 region.

The Xq28 region, as defined by Ensembl (version 19.34b.2), originates from the start position of the FMR2 gene at base pair position 146,287,692 proximal to the centromere and terminates at the end position of the ILR9 gene at base pair position 153,672,692, distal from the centromere on the X chromosome. The Xq28 region houses 13.6% of all known X-linked genes and is 7.384621 mega bases (Mb) in size contributing to 0.23% of the human genome sequence. The expression categories, %GC and strand of all the Xq28 linked genes as well as the gene density of the Xq28 region using a 100 kilo base (Kb) window size is shown in Figure 8, page 52. The 102 Xq28 linked genes include 3 pseudoautosomal (PAR) linked genes (SPRY3, SYBL1 and IL9R)[80], genes implicated in mental retardation and the Fragile X syndrome (FMR2) [124], hemophilia A (NM_019863) [125], development of genital phenotype (CXorf6) [126] and CT genes amongst others [69] (Table B, Appendix pg 85).

The average length of an Xq28 linked gene (including introns) is 52,893 base pairs (bp) in size. The average gene length of the Xq28 linked CT gene families CT 1 (MAGEA), CT 6 (NY-ESO-1/LAGE), CT 24 (CSAGE) and CT 43 (FATE) are 11,248 bp, 1,632 bp, 860 bp and 7,159 bp respectively. The Xq28 linked genes (including intronic regions) cover 35.5% of the Xq28 genomic region. The intergenic regions of Xq28 traverse 64.5% of the total Xq28 region with the mean intergenic length being 47,165 bp in size. The average gene density in the Xq28 region is 13.8 genes per Mb, slightly higher then the genome wide average of 11.1 genes per Mb. Gene density on a regional scale using a non-overlapping window size of a 100 Kb peaks from base pair position 150,400,000 to base pair position 153,100,000 on Xq28 (Figure 8, pg 52).

https://etd.uwc.ac.za/

The Known genes, %GC, Strand and Gene Density of the Xq28 region.

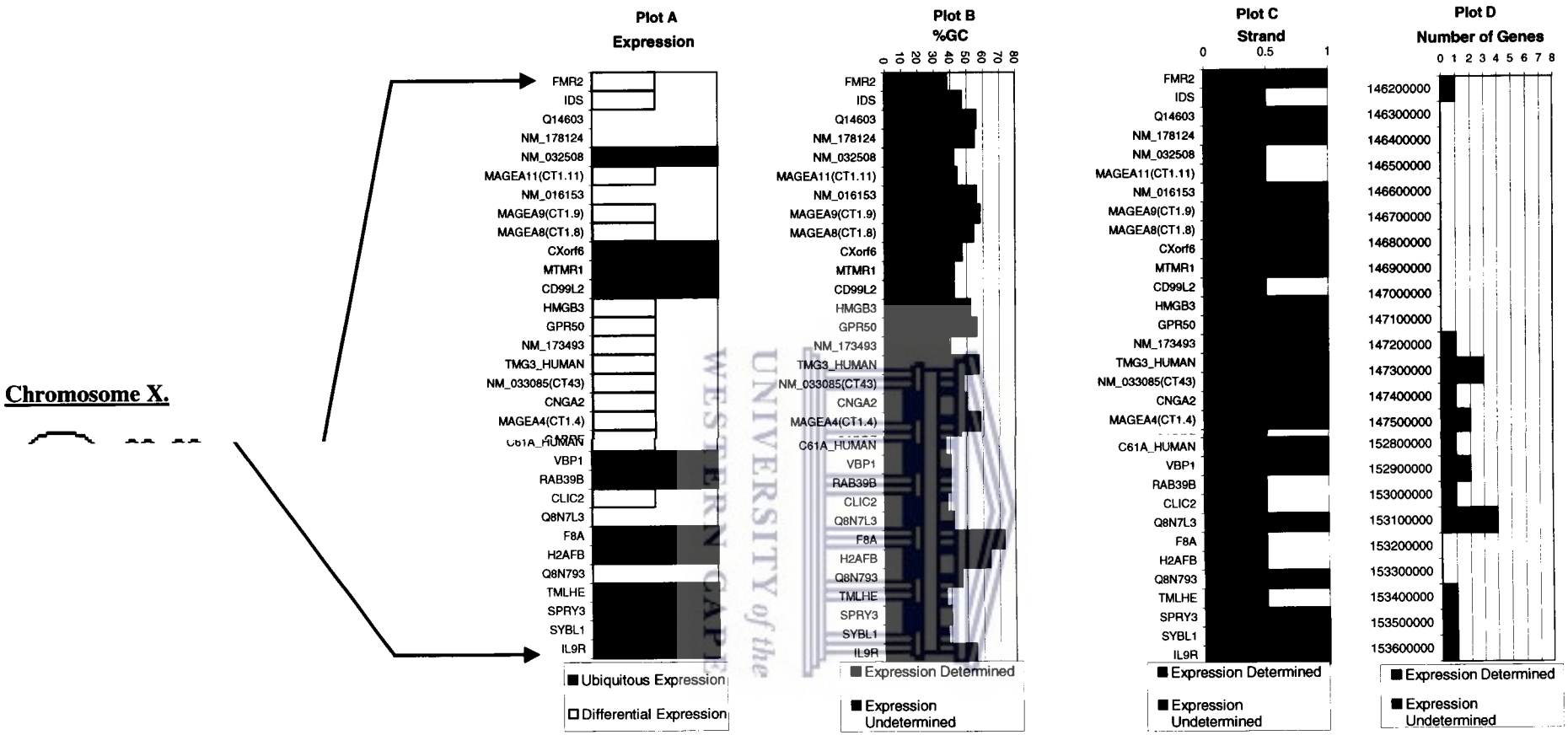


Figure 8 : The 102 human Xq28 linked genes and region exported from Ensembl for analysis. Plot A shows the expression profiles determined for the Xq28 genes, genes whose tissue expression spectra could not be defined are represented by the white gaps and were masked from subsequent analyses. Plot B shows the %GC of the genes exported through EnsMart, genes whose expression was determined is shown in red bars. Plot C shows the strand to which the genes have been mapped to in Ensembl, 0.5 represents the negative strand while 1 represents the positive strand, green bars indicate the genes whose expression has been determined. Plot D shows the gene density of the Xq28 genes exported by EnsMart in using a non-overlapping window size of 100 Kb starting from 142600000bp based on the start co-ordinates of the known Xq28 genes.

The X chromosome figure is taken from : http://www.ensembl.org/Homo_sapiens/mapview?chr=X

https://etd.uwc.ac.za/

The mean %GC content of the Xq28 genes is 52.6% where as the average %GC content for the Human genome sequence is 41% 6. As noted by the International Human Genome Sequencing Consortium (IHGSC) 6, there are regional variations in the GC content of the Human genome and in the Xq28 region these range from 73% (FMR1) to 37% (GABRA5), Figure 8, pg 52. The average %GC of the 16 CT genes present in each of the 4 Xq28 linked CT gene families; CT 1 (MAGEA), CT 6 (NY-ESO-1/LAGE), CT 24 (CSAGE) and CT 43 (FATE) are 54.4%, 64.5%, 51.5% and 48%, accordingly. Gene density and expression has been shown to increase in regions of high GC content and have been linked with short intron length[6,111].

Of the 102 Xq28 linked genes, 83 genes could be classified as either Ubiquitously or Differentially expressed. The number of genes and proportion as a percentage of the total number of Xq28 linked genes placed into each expression category is summarized in Table 5.

Expression Category	Number of Xq28 Linked Genes	Percentage of the Total Number of Xq28 genes.
Differentially Expressed Genes	45	44.14 %
Ubiquitously Expressed Genes	38	37.25 %
Expression Undetermined	19	18.63 %

Table 5 : The number and proportion of the total 102 Xq28 linked genes which have been placed in each of the three mutually exclusive expression categories devised.

From a total of 17 Xq28 linked CT genes, 16 were obtained through the use of EnsMart. As only known Xq28 linked genes were chosen for the analysis of expression breadth, CT 1.7 (MAGEA7) which is a pseudogene (NG_001156) was not exported [6,67,69]. The Xq28 linked CT genes contribute 35.6% to the total of the 45 Differentially Expressed Xq28 linked genes identified.

UNIVERSITY of the WESTERN CAPE
https://etd.uwc.ac.za/

Are there any Clusters of Expression on Xq28?

A cluster of expression is defined as two or more physically adjacent genes that belong to the same mutually exclusive expression categories (excluding the Expression Undetermined category) based on the tissue expression spectra exhibited by those physically adjacent genes. The size of an expression cluster is determined by the number of genes that constitute that cluster. Using these definitions, 20 different gene expression clusters of various sizes could be identified on the Xq28 region as shown in Figure 9 page 55. The 20 observed Xq28 gene clusters can be placed into 5 groups based on their cluster size. The number of gene expression clusters observed for each cluster size group, expression category, percentage of the total 83 Xq28 genes whose expression has been characterized and the percentage each cluster group contributes to the 20 Xq28 expression clusters is summarized in Table 6.

UNIVERSITY of the WESTERN CAPE

Expression Category	Cluster Size	Number of Clusters	% Of 83 Genes whose Expression has been Categorized	% Of the Twenty Expression Clusters Observed
Differentially Expressed	**2**	4	9.64 %	20 %
Ubiquitously Expressed		5	12.05 %	25 %
Differentially Expressed	**3**	1	3.60 %	5 %
Ubiquitously Expressed		3	10.80 %	15 %
Differentially Expressed	**4**	2	9.64 %	10 %
Ubiquitously Expressed		1	4.82 %	5 %
Differentially Expressed	**5**	1	6.02 %	5 %
Ubiquitously Expressed		2	12.04 %	10 %
Differentially Expressed	**10**	1	12.05 %	5 %
Ubiquitously Expressed		0	- - - - -	- - - -

Table 6 : The 20 observed Xq28 gene expression clusters as shown in Figure 9 (overleaf) can be placed into 5 groups based on the number of genes present in an observed expression cluster.

https://etd.uwc.ac.za/

Expression Clusters Identified on Xq28

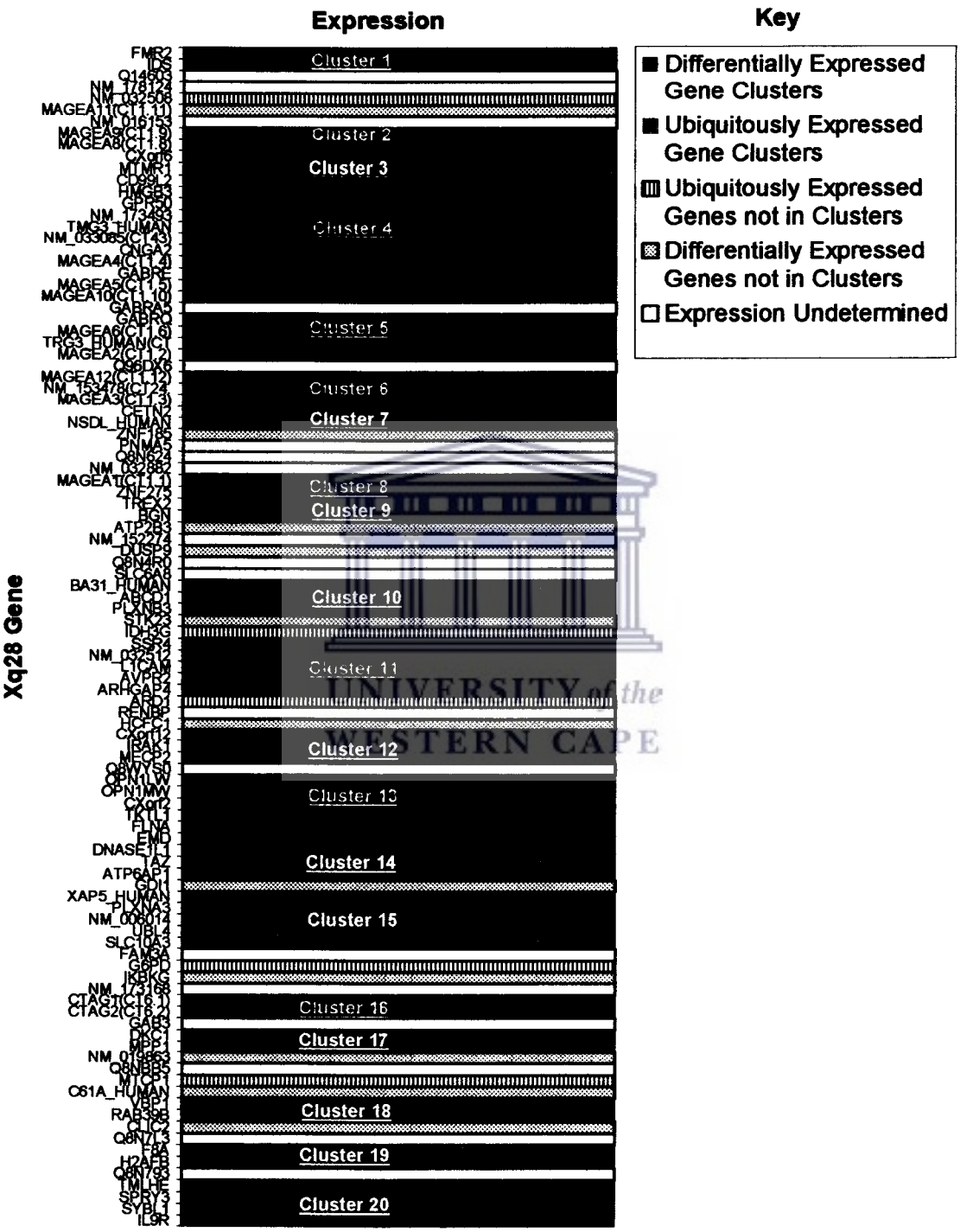


Figure 9 : Collinear genes which belong to the same expression category form 20 physically adjacent clusters of similar gene expression in terms of their spectrum of tissue expression in the Xq28 region.

https://etd.uwc.ac.za/

From the total of 83 Xq28 linked genes whose expression could be characterized as either Ubiquitously or Differentially Expressed, 16 of the genes do not belong to any gene expression cluster and hence, form singletons (Figure 9, pg 55). Eleven of the 16 singletons, which include CT 1.11 (MAGEA11), are categorized as Differentially Expressed while 5 are Ubiquitously Expressed, Figure 9, pg 55. Of the 16 singletons, only 4 genes, (C16A_Human, GDI1, IDH3G and STK3), can unambiguously be excluded from being placed into any expression cluster (Figure 9, pg 55). The 12 remaining singletons can not unambiguously be excluded from or assigned to a potential expression cluster. The ambiguity in assigning or excluding the singletons from an expression cluster is partly due to the lack of information which would enable the categorization of the 19 Expression Undetermined Xq28 linked genes, as well as the preservation of their spatial information (Figure 9, pg 55). As the expression categories derived are mutually exclusive, genes in the Expression Undetermined category can only fall into one of two expression categories; Ubiquitously Expressed or Differentially Expressed depending on their breadth of tissue expression recorded. Hence, the categorization of the tissue expression spectra of the Expression Undetermined genes will either extend, interrupt or define new clusters of gene expression on the Xq28 region (Figure 9, pg 55).

UNIVERSITY of the
WESTERN CAPE

Sixty-seven genes (66 % of the 102 Xq28 linked genes) can be assigned to an expression cluster. The proportion of 16 Xq28 linked CT genes which are assigned to an expression cluster is 93.75%. Clusters 2, 4, 5, 6, 8 and 16 (Figure 9, pg 55) all contain CT genes with cluster 4 housing the highest proportion (25%) of the 16 Xq28 linked CT genes. The high proportion of CT genes being placed in an expression clusters can partly be attributed to fact that the 16 Xq28 linked CT genes are members of 4 CT gene families. Previous studies indicate that some of the CT genes like the CT 1 (MAGEA) family have arisen from the retroposition and duplication of one or more members of the MAGE-D gene

https://etd.uwc.ac.za/

family [66,69]. Hence, the observed clustering of the Xq28 linked CT genes in differentially expressed clusters such as cluster 2 or cluster 16 (Figure 9) is likely to be a mechanistic product of the retroposition and subsequent expansion by duplication of the Xq28 linked CT gene families [66,69].

Although there might be a mechanistic positional biasness for some Xq28 linked CT genes being placed into expression clusters, their physical location on the genome defines the expression spectra measured of that particular physical genomic location they are embedded in. Additionally, if genes arising from duplication in the Xq28 region were removed, gene order would not be preserved. In some cases, the physical genomic preservation of gene order is vital for the regulation of gene expression as exemplified by the physical clustering of the homeotic (Hox) genes [127-129]. The Hox genes are key players in determining anterior-posterior morphological identities of tissues during the development phase of multi-cellular organisms [127-129]. The Hox genes have arisen and expanded through multiple gene duplications [127-129]. Through the subsequent duplications and expansions of the Hox gene family the physical gene order of the Hox genes was maintained [127-129]. The physical clustering of the Hox genes which have arisen through duplication and in which gene order is tightly maintained is postulated to be due to the constraints imposed by the Hox genes' spatial and temporal gene expression profiles [127-129]. The physical gene order of the Xq28 linked genes may not be as tightly constrained as the Hox genes. Nonetheless, due to the lack of functional information for the CT genes as well as some of the Xq28 linked genes, the possibility of physically adjacent Xq28 linked genes arising from duplication exhibiting a similar expression spectrum can not be excluded. For example, all the CT 1 (MAGEA) gene family members exhibit a narrow spectrum of tissue expression even though the ancestral MAGE-D gene from which the CT 1 (MAGEA) family is postulated to have arisen from is ubiquitously expressed[66,67,69,130]. Furthermore, Lercher *et al* [108] confirmed that the previously

UNIVERSITY of the WESTERN CAPE

https://etd.uwc.ac.za/

observed genomic partitioning of highly expressed genes persist, even when gene expression data from duplicated genes such as the Hox gene cluster is corrected for.

Is There any Correlation in the Position or %GC Content Between the Ubiquitous and Differential Xq28 Gene Expression Clusters?

Through the use of SAGE and EST data, Lercher *et al* [108] established that genes which are classified as highly and broadly expressed are non-randomly dispersed in the human genome and form clusters. The observed clustering of genes based on their expression spectra in the human genome was shown to be related to the expression breadth of widely expressed genes rather then the rate of gene expression [108,113].

A qualitative approach towards determining the tissue expression spectra of the Xq28 linked genes indicates that the clustering of broadly expressed genes in the human genome would be observed when using gene expression data quantitatively. The observed clustering of ubiquitously expressed genes is due to the sheer heterogeneity of tissues in which transcripts of differentially expressed genes have been recorded. A quantitative methodology mapping genome wide gene expression data based on the recorded presence / absence of a gene's transcripts for a fixed number of pre-defined tissues will not reveal the physical clustering of tissue specific genes *per se* on the genome, it will reveal genomic regions of broadly expressed genes [96,108,113]. In order to establish if there is a relationship between the expression of genes in the 20 identified Xq28 gene expression clusters and their physical positioning in the Xq28 region, the weighted mean intergenic lengths of Ubiquitously and Differentially expressed Xq28 gene clusters was determined and is presented in Table 7, overleaf.

UNIVERSITY of the WESTERN CAPE

https://etd.uwc.ac.za/

	Ubiquitously Expressed Xq28 Gene Clusters	Differentially Expressed Xq28 Gene Clusters	102 Xq28 Genes
Average Intergenic Length (Base pairs)	30, 893 bp	79,052 bp	47,165 bp

Table 7 : The average weighted intergenic lengths of the Xq28 genes present in the 20 identified Ubiquitously and Differentially Expressed gene clusters as well as the average intergenic length of the 102 Xq28 genes (Figure 8 and 9).

On average, the intergenic lengths between Differentially Expressed gene clusters is two fold greater then the average intergenic lengths of Ubiquitously Expressed gene clusters on the Xq28 region. This would be indicative of Ubiquitously Expressed genes physically forming more compact genes expression clusters as compared to the Differentially Expressed genes on the Xq28 region. However, a chi square test indicates the differences in intergenic lengths between the Ubiquitously Expressed and Differentially Expressed Xq28 gene clusters is not statistically significant (Appendix, pg 90).

UNIVERSITY of the
WESTERN CAPE

The link between GC content and gene expression is more tenuous. Gene density has been shown to increase in genomic regions of high GC content [6,96,111,113]. Genomic regions of high gene expression (RIDGEs) are also characterized by high GC content [96,108,111,113]. However, as gene expression and gene density from a particular genomic location are interlinked, the observation of highly expressed regions being GC rich may reflect the gene order organization of a particular genomic region, rather then its underlying base composition [131]. In a separate study, Lercher *et al* [113] determined the breadth of tissue expression spectra for a gene is correlated to its genomic base composition, widely expressed genes were found to congregate in regions of high GC content. Tissue specific genes were found by Vinogradov to be, on average, GC poorer then ubiquitously expressed genes [132].

https://etd.uwc.ac.za/

The average %GC for the Xq28 linked genes present in the Ubiquitously Expressed and Differentially Expressed gene clusters are 52.67 % and 53.68% respectively. There are variations in the individual %GC content of the individual genes which constitute the expression clusters, but the average %GC for genes in both the Ubiquitously and Differentially expressed gene clusters does not significantly deviate from the mean %GC of all the Xq28 linked genes (52.6 %).

Hence, a positional and %GC content biasness can not be found between genes in the observed gene expression clusters of different expression spectra on the Xq28 region.

Conclusions.

At a regional genomic level, there are differences in the tissue expression spectra of genes on the Xq28 region. As there are differences in the expression spectra measured of genes embedded in the Xq28 region, this indicates there are different genomic "localities" of expression in the Xq28 region. In some cases these localities of expression congregate to form neighborhoods comprising of genes exhibiting similar expression spectra, in some cases they do not. Of the 83 Xq28 linked genes whose spectrum of tissue expression could be determined, 80% can be assigned to an expression cluster and hence form neighborhoods of similar expression spectra. These neighborhoods may partly be an artefact of the qualitative categorization of the Xq28 genes into two mutually exclusive expression categories, but if a gene is highly and broadly expressed its expression spectra will be classified as such and vice versa, regardless of the methodology or categorization scheme used.

The physical clustering of genes based on their expression spectra in the human genome depends on the genes, genomic resolution, tissue expression data and categorization scheme for the genes used. The heterogeneity of tissue types in which the expression for the Xq28 linked genes have been found does not facilitate individual tissue categories to be

UNIVERSITY of the WESTERN CAPE
https://etd.uwc.ac.za/

constructed, unless those categories are numerous. For example, the CT genes are predominantly expressed in the testis, the opsin genes on Xq28 are found in the photoreceptors of retinal cone cells [117,118]. Hence a quantitative classification scheme would need to account for the tissue heterogeneity in expression for all genes present in the human genome.

No differences could be found between the physical positioning of genes and %GC content with regards to their expression spectra. However the Xq28 region examined contributes to 0.23% of the human genome and due to this small sample size, the unique Xq28 features are not representative of the whole human genome.

The bulk of the Xq28 CT genes are located in neighborhoods comprising of genes with a narrow expression spectra. Whether the expression spectra of the CT genes is influenced by their genomic neighborhood can not be conclusively determined. The expression of CT 1 (MAGEA) family in somatic cells has been shown to be caused by the demethylation of their GC rich promoter sequence [55,56]. Transcription factors proficient for the expression of the CT 1 (MAGEA) genes are present in a cell even when the CT 1 (MAGEA) genes are not expressed [55]. Hence regulation of the Xq28 CT 1 (MAGEA) genes occurs epigenetically rather then at the transcriptional level [55,56]. The specific mechanisms regulating gene expression for the remaining Xq28 linked genes was not determined in this study. Whether there is any similarity between the mechanisms regulating gene expression in the Differentially or Ubiquitously Xq28 gene expression clusters remains to be resolved.

UNIVERSITY of the
WESTERN CAPE

https://etd.uwc.ac.za/

Chapter 5

Ancient Genes?

The sequencing of a variety of metazoan genomes, which are at different stages of completion[133], permits the comparison of genomic regions and whole genomes at the nucleotide and protein level between different metazoan species[8,9,134-137]. Comparison of genomic regions e.g exons, introns, promoters and whole genomes of different metazoan species provides insights to the evolutionary and biological processes which are common, as well as unique, to the metazoan species being compared[8,135,137-139]. A principal assumption made when comparing genomic regions or genomes of different metazoan species is that they have descended, with divergence, from the last common ancestor of the metazoan taxa under comparison i.e a cenancestor[137-140]. Hence, the genomic regions under comparison from different metazoan species are homologous to each other by virtue of being derived from a cenancestor[137,139,140].

UNIVERSITY of the
WESTERN CAPE

Homologous genomic regions which are more similar amongst metazoan species being compared are termed as conserved, dissimilar homologous genomic regions are termed as divergent, with regards to the metazoan species being compared[134,135,137,139]. The various degrees of conservation and divergence observed between homologous metazoan genomic regions are attributable to the subsequent evolution, of those metazoan genomic regions from their cenancestor[134,135,137,139,140]. The evolution of genomic sequences in metazoan species from a cenancestral genomic sequence is generally considered to be a fusion of two predominant forces[137-139];

1). Random mutations being created in a genomic sequence by mutational processes.

https://etd.uwc.ac.za/

2). Selection factors which can either have no effect on the random mutation (neutral selection), cause the fixation of a random mutation due to a gain in fitness (positive selection) or purge the random mutation (negative selection).

Functional genomic regions are defined as genomic regions that assist a metazoan organism throughout its lifespan and reproduction, and are likely to be under selection pressures in different metazoan species[135-139]. Functional genomic regions generally tend to be conserved between metazoan species being compared, mainly because any random mutations that occur in functional genomic regions are potentially detrimental to the fitness of an organism[135-139]. As random mutations in functional genomic regions are likely to be selected against, functional genomic regions are inclined to evolve at a slower pace as opposed to less functionally constrained genomic regions, and hence, tend to be conserved between different metazoan species[135-139]. Exceptions include genomic regions involved in spermatogenesis, olfaction and immunity that appear to be under positive selection and thus evolve at elevated rates compared to other functional genomic regions in different metazoan organisms.[6,8,9,90]

UNIVERSITY of the
WESTERN CAPE

The sequencing of a variety of different metazoan genomes additionally provides an increased resolution in the estimation of divergence times between the lineages of different metazoan species[134,135,141]. Divergence times between different metazoan species are estimated from protein and DNA sequence data using a combination of the molecular clock principle and an evaluation of fossil records[134,137]. The molecular clock hypothesis stipulates that the rate of evolution for a given protein or DNA sequence is roughly constant amongst lineages of metazoan species that arise from a cenancestor[134,137,139]. Estimations of divergence times between metazoan species using molecular clock methods tend to be older then their corresponding fossil dates[134]. The younger estimations of divergence times between metazoan species obtained by fossil records are considered to be a minimum

https://etd.uwc.ac.za/

estimate of divergence times[134]. In contrast, divergence time estimates of metazoan species derived from molecular clock methods are thought to measure molecular divergence as soon as two lineages descend from a cenancestor[134]. Estimations of divergence times between different metazoan species from their cenancestors provides a framework of approximate timescales for comparative genomic studies. Based on the degree of conservation or divergence of homologous genomic regions in different metazoan taxa, a rough approximation of how old the homologous genomic regions are can be made.[8,9,134,135,138,139,141]

The CT genes are a heterogeneous subset of human genes which are expressed in gametogenic tissues, germ cells and various types of cancers[29,30,33,37]. The CTAs relatively narrow tissue tropism combined with their ability to elicit an autologous immune response in 20%-40% of cancer types, indicates that all of the CT genes can be considered functional genomic regions in the human genome[29,30,33,37]. Exceptions being CT 1.7 (MAGEA7) and CT 18 (NA88A) which are pseudogenes, (Accession numbers : NG_001156 and NR_001559 respectively), although CT 18 is a Cancer/Testis Antigen[142]. Elucidation of the genomic locations of the all known CT genes conducted in Chapter 2 indicate that, apart from Chromosome 1, 21 and X, the remainder of the CT genes are evenly distributed in the human genome. However, the product functions of 97.5% of the CT genes are presently not known, although the X linked CT genes are unlikely to participate in spermatogenesis beyond the onset of meiosis I due to the initiation of meiotic sex chromosome inactivation[72,84,89,101]. Some of the CT genes have been shown to arise through retroposition and duplication of ancestral genes, events which are postulated to be fairly recent[66]. The possible events leading the emergence of other CT genes have not been investigated.

A pilot comparative genomic study was undertaken to determine the absence or presence of homologues for all the CT genes in a variety of metazoan species belonging to

UNIVERSITY of the WESTERN CAPE

https://etd.uwc.ac.za/

different metazoan phylogenetic taxas. Different phylogenetic scopes based on estimations of divergence times between the metazoan cenancestors relative to *Homo sapiens*, was used to determine the approximate ages of all the CT genes. A comparative genomic approach facilitates in identifying potentially useful CT gene homologues in model organisms for biomedical research. Additionally, a comparative genomic approach enables the elucidation of which CT genes are "ancient", as well as which CT genes are unique, to the metazoan species and the phylogenetic scopes used in this study.

Methods.

Pre-computed homologues for 82 of the CT genes in a variety of metazoan species were mined from the Ensembl (version 26.35.1) and HomoloGene (Build 36) databases using the relevant CT gene accession identifiers (Table A, Appendix pg 82). Both Ensembl and HomoloGene seek to provide automated catalogues of homologous genes in different metazoan species whose genomes have been sequenced[143,144]. Amongst other strategies, Ensembl utilizes the BLASTP algorithm to determine "Reciprocal best hit pairs" for inferring homology[143,145]. Reciprocal best hit pairs are two genomic regions from two different metazoan species under comparison that produce the best complementary BLAST results to each other[143,145]. Genomic regions that produce reciprocal best hit pairs between two metazoan species under comparison are termed as putative homologues[143,145]. HomoloGene utilizes a similar BLASTP matching strategy that is guided by a taxonomic tree when inferring homology between functional genomic regions of distantly related species♣.

The term "homologue" is a higher order hierarchical classification which comprises of three disjoint subtypes based on the origin of the genomic regions under comparison[137,139,140]:

♣ http://www.ncbi.nlm.nih.gov/HomoloGene/HTML/homologene_buildproc.html

https://etd.uwc.ac.za/

1). **Orthologues** are homologous genomic regions which have originated from a cenancestral genomic region by a speciation event. Comparison of orthologous genomic regions reflects the accurate phylogeny of the organisms from which they are obtained[137,139,140].

2). **Paralogues** are homologous genomic regions which have originated by duplication events from a cenancestral genomic region after a speciation event. Comparison of paralogous genomic regions reflects the correct phylogeny of those genomic regions, rather then the true phylogeny of the taxa from which the genomic regions are obtained[137,139,140].

3). **Xenologues** are homologous genomic regions which have originated from the horizontal transfer of genetic material between different species as opposed to their inheritance from a cenancestor. Comparison of xenologous regions does not reflect the true phylogeny of the taxa from which those genomic regions are obtained[140].

Both Ensembl and HomoloGene differentiate between paralogues and orthologues. In this study, only information from putative orthologues was used to determine the presence / absence of CT gene homologues in different metazoan organisms while information regarding putative CT gene paralogues was ignored. An in depth manual characterization of the different subtypes of homologous regions automatically computed by Ensembl and HomoloGene in the different metazoan species sampled was not undertaken. The ability to make functional inferences for any putative CT gene homologues identified in different metazoan species is hindered by the lack of functional knowledge for the role of the biological gene products for 97.5% of the CT genes. Additionally, the aim of this study is to determine whether the CT genes are conserved in a variety metazoan species by the presence of homologous genomic regions, and as a consequence, determine the approximate ages of the CT genes. Hence, similar precomputed genomic regions to the CT genes identified in different metazoan species through Ensembl and Homologene are termed as CT gene homologues in this study.

UNIVERSITY of the WESTERN CAPE

https://etd.uwc.ac.za/

Cooper *et al*[138] have identified two vital factors which impact a comparative genomic study ; the extent of divergence captured between the genomic sequences being compared, and secondly, the phylogenetic scope employed for the comparison of homologous genomic regions.

Homologous genomic regions are more similar between metazoan species that have diverged recently from a cenancestor, and less similar between related metazoan species separated by longer divergence times[134-139]. The degree of conservation between different homologous metazoan functional genomic sequences will depend on the time that has elapsed since the divergence of those genomic regions from a cenancestor[134-139]. Protein sequences are generally not used for identifying conserved homologous functional genomic regions between closely related metazoan species, principally because insufficient time has elapsed for functionally unconstrained sites in the protein sequences to accumulate changes[134-139]. As homologous metazoan genomic regions diverge from a cenancestor in a roughly chronological fashion, conservation of homologous protein sequences in closely related species, such as *H. sapiens* and *Pan troglodytes*, may reflect a lack of divergence time, rather then their actual functional conservation. Hence, conserved homologous functional genomic regions between closely related metazoan species which have diverged ~90 million years ago (mya) from a cenancestor are typically inferred using DNA sequence data[134-139]. DNA sequences are generally considered to be to divergent, as opposed to protein sequences, for identifying conserved functional genomic regions between metazoan species which have more ancient divergence times[134-139]. Protein sequences are generally used for identifying conserved functional genomic regions between metazoan species with older divergence times, e.g *H. sapiens* and *Fugu rubripes*[134-139]. Both the Ensembl♠ and

UNIVERSITY of the
WESTERN CAPE

♠ http://www.ensembl.org/Homo_sapiens/whatsnew

https://etd.uwc.ac.za/

HomoloGene♦ databases take into account the divergence times of different metazoan species when identifying homologues.

A phylogenetic scope, as defined by Cooper *et al*[138] , is the smallest taxanomic group that embodies the genomic sequences being compared. A predefined phylogenetic scope provides a useful organizational framework for the comparison of homologous genomic regions between different metazoan species by virtue of their descent from a cenancestor[134,135,138,140]. The biological gene product function of 97.5% of the CT genes is unknown, but all the CT genes have been primarily characterized in *H. sapiens*. Hence, a series of phylogenetic scopes relative to *H. sapiens* was used as a framework for the sampling of CT gene homologues in different metazoan species. The metazoan species and series of phylogenetic scopes employed for this study, together with the approximate divergence times between the lineages of the metazoan species sampled for CT gene homologues, is depicted as a tree, relative to *H. sapiens*, in Figure 10, pg 70. Each chosen phylogenetic scope (shown as squares in Figure 10, pg 70) starting from primates, is a subset of a broader encompassing phylogenetic scope, relative to *H. sapiens*. Each phylogenetic scope comprises of the different metazoan species sampled for CT gene homologues (Figure 10, pg 70). The age of each phylogenetic scope is based on the estimated divergence times of the lineages leading to the present day metazoan species sampled for CT gene homologues (Figure 10, pg 70). As the phylogenetic scope broadens, the estimated times of divergence between the cenancestors of the metazoan species sampled for CT gene homologues also increases (Figure 10, pg 70). The increasing divergence times between the metazoan species enables a rough estimation in determining which phylogenetic scope the CT genes are conserved in by the presence / absence of CT gene homologues. The increasing estimated

UNIVERSITY of the
WESTERN CAPE

♦ http://www.ncbi.nlm.nih.gov/Web/Newsltr/Spring04/homologene.html

https://etd.uwc.ac.za/

divergence times between the species of the different phylogenetic scopes employed in this study also enables an approximate "dating" of the CT genes based on the presence / absence of CT gene homologues, in the metazoan species sampled.

https://etd.uwc.ac.za/

The Phylogenetic Scopes with their respective species for which CT gene homologues were sampled.

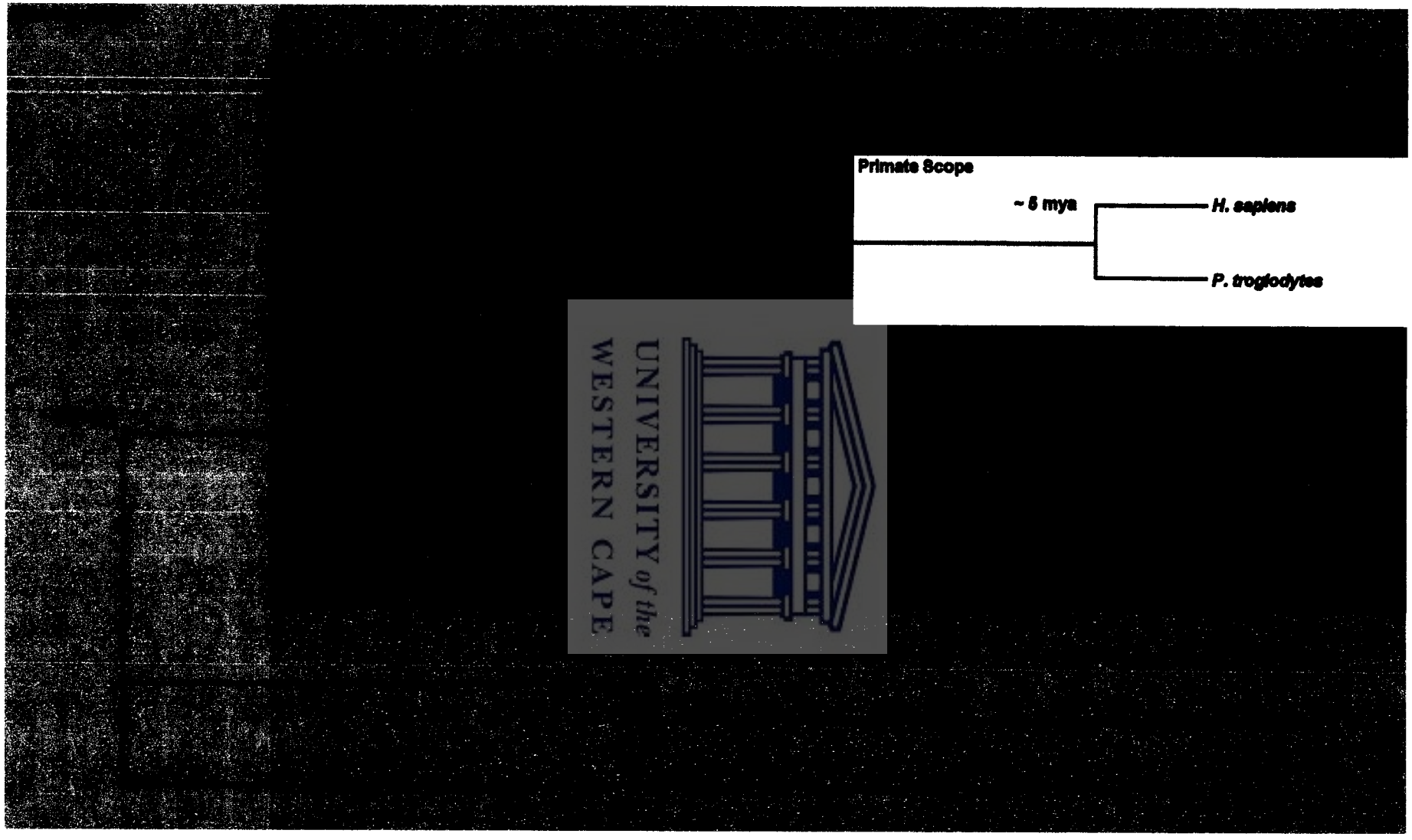


Figure 10 : The scoring scheme used to determine the ages of the CT genes is based on a metazoan phylogenetic taxanomic scope relative to *H. sapiens*. Each phylogentic taxonomic scope (coloured rectangles) starting from the Primate Scope froms a subset of a larger phylogenetic taxonomic scope up till the Metazoan Scope. The age of the CT genes is determined by the presence / absence of CT gene homologues determined from Ensembl and Homologene. The approximate ages of divergence from a cenancestor of each phylogenetic taxominc scope is approximatly given as million years ago (mya). The presence / absence of a CT gene homologue detected in a specific organism sampled whose genome has been sequenced allows the rough quantification of a CT genes' age based on the approximate divergence times of a cenancestor for each phylogenetic taxonomic scope sampled. The Figure has been adapted from reference 138 with the authors' kind permission.

https://etd.uwc.ac.za/

How "Old" are the CT Genes?

The CT genes have been described as a collation of "heterogeneous genes[37]", this is reflected by the conservation of the CT genes in the different metazoan species sampled for CT gene homologues. A full listing of the CT genes, antigenicity, cytogenetic locations, phylogenetic scope and the individual metazoan species, for which CT gene homologues were identified, is presented overleaf, in Figure 11. The CT genes in Figure 11, pg 72 are ordered according to the most recent CT genes, which are defined as having the narrowest phylogenetic scope, to the most ancient CT genes, which have an extensive phylogenetic scope, Figure 10, pg 70. The extent of a phylogenetic scope is determined by the absence / presence of CT gene homologues in the different metazoan species sampled which are representative of a specific phylogenetic scope (Figure 11, pg 72). By default, all the CT genes are present in the primate scope as all the CT genes have been primarily identified and characterized in *H. sapiens*. As the phylogenetic scope broadens, the number of CT gene homologues identified decreases. The number of CT genes present in each phylogenetic scope is presented in Figure 12.

UNIVERSITY of the WESTERN CAPE

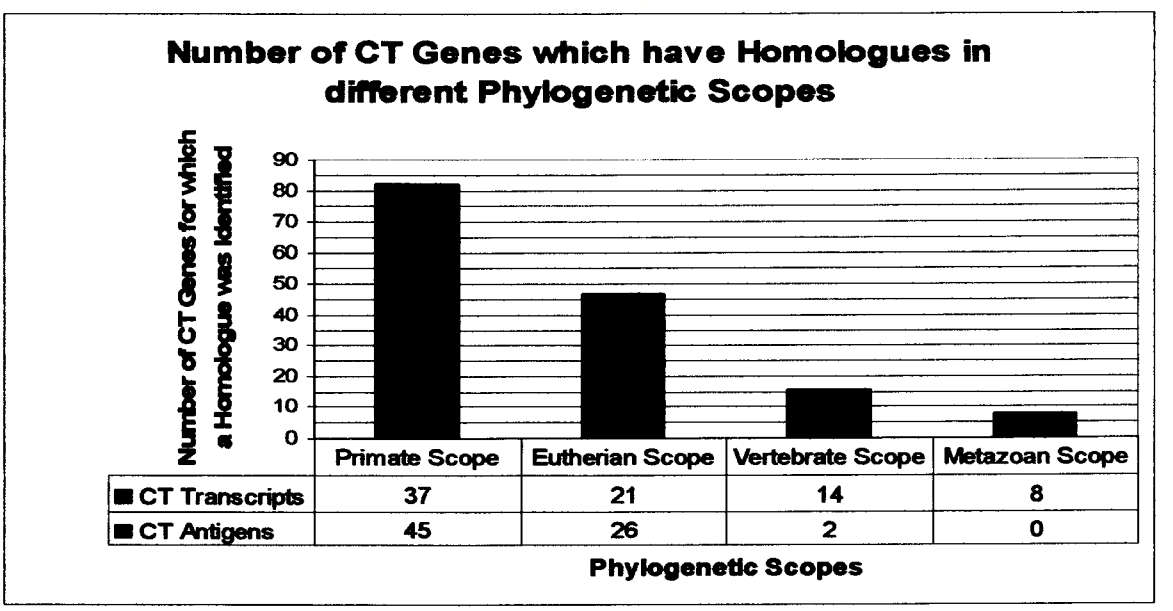
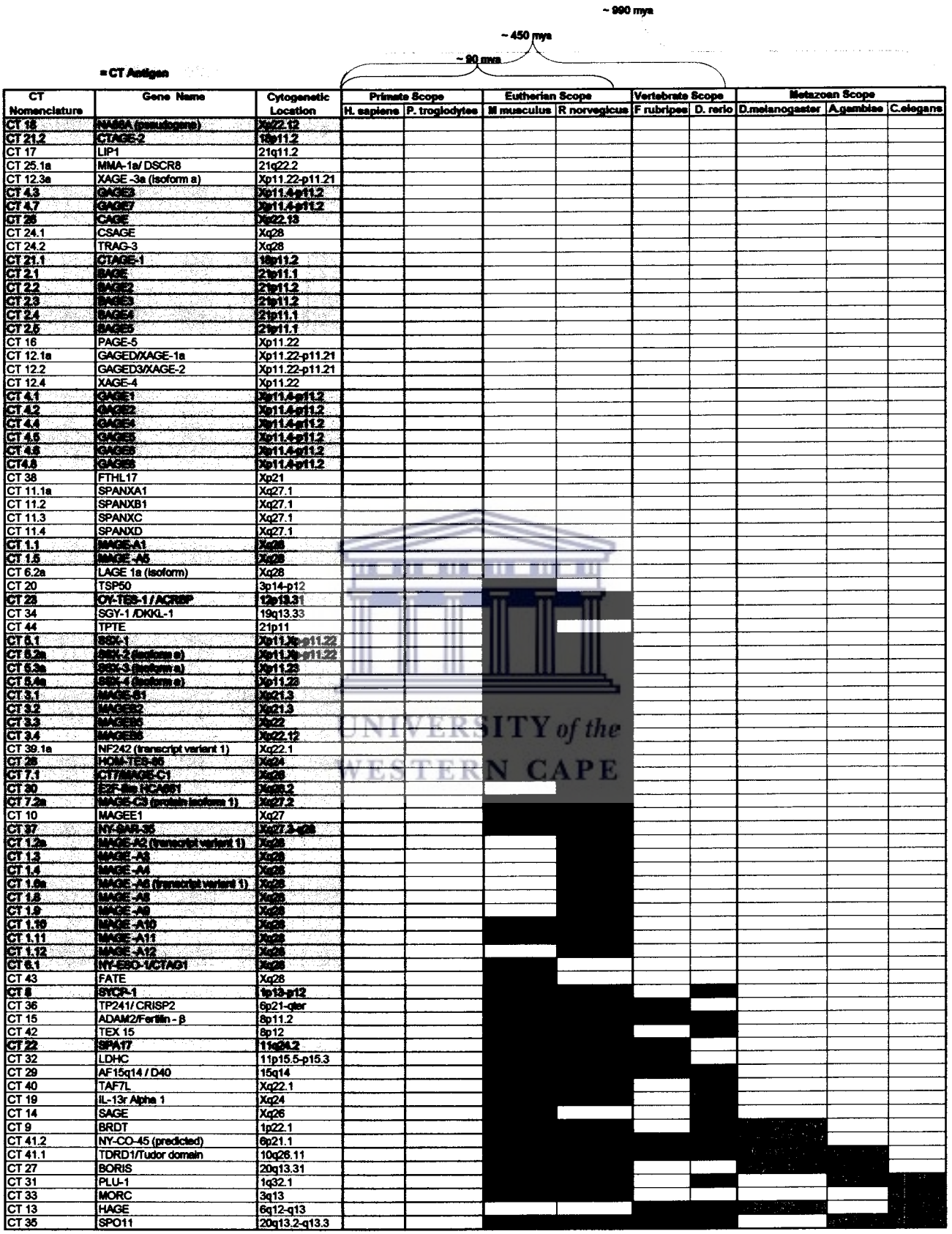


Figure 12 : The number of CT gene homologues identified in different metazoan species (shown in the data table below the bar chart) decreases as the phylogenetic scope increases.

https://etd.uwc.ac.za/

The Phylogenetic Scopes in which CT gene homologues were identified.



CT Nomenclature	Gene Name	Cytogenetic Location	Primate Scope		Eutherian Scope		Vertebrate Scope		Metazoan Scope		
			H. sapiens	P. troglodytes	M musculus	R norvegicus	F rubripes	D. rerio	D.melanogaster	A.gambiae	C.elegans
CT 18	NA88A (pseudogene)	Xp22.12									
CT 21.2	CTAGE-2	18p11.2									
CT 17	LIP1	21q11.2									
CT 25.1a	MMA-1a/ DSCR8	21q22.2									
CT 12.3a	XAGE -3a (isoform a)	Xp11.22-p11.21									
CT 4.3	GAGE3	Xp11.4-p11.2									
CT 4.7	GAGE7	Xp11.4-p11.2									
CT 26	CAGE	Xq22.13									
CT 24.1	CSAGE	Xq28									
CT 24.2	TRAG-3	Xq28									
CT 21.1	CTAGE-1	18p11.2									
CT 2.1	BAGE	21p11.1									
CT 2.2	BAGE2	21p11.2									
CT 2.3	BAGE3	21p11.2									
CT 2.4	BAGE4	21p11.1									
CT 2.5	BAGE5	21p11.1									
CT 16	PAGE-5	Xp11.22									
CT 12.1a	GAGED/XAGE-1a	Xp11.22-p11.21									
CT 12.2	GAGED3/XAGE-2	Xp11.22-p11.21									
CT 12.4	XAGE-4	Xp11.22									
CT 4.1	GAGE1	Xp11.4-p11.2									
CT 4.2	GAGE2	Xp11.4-p11.2									
CT 4.4	GAGE4	Xp11.4-p11.2									
CT 4.5	GAGE5	Xp11.4-p11.2									
CT 4.6	GAGE6	Xp11.4-p11.2									
CT4.8	GAGE8	Xp11.4-p11.2									
CT 38	FTHL17	Xp21									
CT 11.1a	SPANXA1	Xq27.1									
CT 11.2	SPANXB1	Xq27.1									
CT 11.3	SPANXC	Xq27.1									
CT 11.4	SPANXD	Xq27.1									
CT 1.1	MAGE-A1	Xq28									
CT 1.5	MAGE -A5	Xq28									
CT 6.2a	LAGE 1a (isoform)	Xq28									
CT 20	TSP50	3p14-p12									
CT 23	OY-TES-1 / ACRBP	12p13.31									
CT 34	SGY-1 /DKKL-1	19q13.33									
CT 44	TPTE	21p11									
CT 5.1	SSX-1	Xp11.Xp-p11.22									
CT 5.2a	SSX-2 (isoform a)	Xp11.Xp-p11.22									
CT 5.3a	SSX-3 (isoform a)	Xp11.23									
CT 5.4a	SSX-4 (isoform a)	Xp11.23									
CT 3.1	MAGE-C1	Xq21.3									
CT 3.2	MAGEC2	Xq21.3									
CT 3.3	MAGEC3	Xq22									
CT 3.4	MAGEC4	Xq22.12									
CT 39.1a	NF242 (transcript variant 1)	Xq22.1									
CT 28	HOM-TES-85	Xq24									
CT 7.1	CT7/MAGE-C1	Xq28									
CT 30	E2F-like NCA951	Xq28.2									
CT 7.2a	MAGE-C2 (protein isoform 1)	Xq27.2									
CT 10	MAGEE1	Xq27									
CT 37	NY-SAR-35	Xq27.3-q28									
CT 1.2a	MAGE-A2 (transcript variant 1)	Xq28									
CT 1.3	MAGE -A3	Xq28									
CT 1.4	MAGE -A4	Xq28									
CT 1.6a	MAGE -A6 (transcript variant 1)	Xq28									
CT 1.8	MAGE -A8	Xq28									
CT 1.9	MAGE -A9	Xq28									
CT 1.10	MAGE -A10	Xq28									
CT 1.11	MAGE -A11	Xq28									
CT 1.12	MAGE -A12	Xq28									
CT 6.1	NY-ESO-1/CTAG1	Xq28									
CT 43	FATE	Xq28									
CT 8	SYCP-1	1p13-p12									
CT 36	TP241/ CRISP2	6p21-qter									
CT 15	ADAM2/Fertilin - β	8p11.2									
CT 42	TEX 15	8p12									
CT 22	SPA17	11q24.2									
CT 32	LDHC	11p15.5-p15.3									
CT 29	AF15q14 / D40	15q14									
CT 40	TAF7L	Xq22.1									
CT 19	IL-13r Alpha 1	Xq24									
CT 14	SAGE	Xq26									
CT 9	BRDT	1p22.1									
CT 41.2	NY-CO-45 (predicted)	6p21.1									
CT 41.1	TDRD1/Tudor domain	10q26.11									
CT 27	BORIS	20q13.31									
CT 31	PLU-1	1q32.1									
CT 33	MORC	3q13									
CT 13	HAGE	6q12-q13									
CT 35	SPO11	20q13.2-q13.3									

UNIVERSITY of the WESTERN CAPE

Figure 11 : The various phylogenetic scopes and estimated divergence times for each phylogenetic scope in which CT gene homologues were identified. The CT genes are ordered from the most recent which is defined as having the narrowest phylogenetic scope to the oldest CT genes for which a homologue was identified in a broader phylogenetic scope. The CT genes families and cytogenetic locations for the CT are presented in the first three columns, the CTAs are shaded in grey.

https://etd.uwc.ac.za/

The Primate Scope

Homologues for 34 of the CT genes were found to be absent the metazoan species that constitute the different phylogenetic scopes, other then the primate scope (CT 18 (NA88A) to CT 6.2a (LAGE 1), Figure 11, pg 72), and hence are restricted to the primate phylogenetic scope. The 41 % of 82 CT genes that are restricted to the primate phylogenetic scope may indicate that those CT genes have not been conserved in the metazoan species sampled that constitute the various phylogenetic scopes (Figure 10, pg 70, Figure 11, pg 72). However, this scenario seems highly unlikely as it would indicate that all the potential 238 CT gene homologues in the 7 metazoan species belonging to phylogenetic scopes other then the primate scope, underwent extremely rapid divergence (Figure 11, pg 72). The Ensembl and HomoloGene databases use heuristic algorithms such as BLAST rather then more sensitive algorithms such as Smith – Waterman for identifying homologues due to overwhelming quantity of genome data present, and the compute intensive nature of the sensitive algorithms[137,139,143,145]. However, the use of heuristic algorithms in determining CT gene homologues can not necessarily nor sufficiently explain the absence of 238 potential CT gene homologues in the 7 metazoan species of different phylogenetic scopes sampled. An additional source of bias is that some metazoan species have more extensively characterized functional genomic regions and protein coding data available than other metazoan species[137,139]. Hence, even though homologues for any of the CT genes in any of the phylogenetic scopes may have not been identified, this by itself is insufficient to preclude their existence in the different metazoan species sampled.

Alternatively, those 34 CT genes may have arisen recently in the primate lineage after the divergence from a rodent – primate cenancestor and hence, are primate specific. The 34 CT genes which are restricted to the primate phylogenetic scope are members of 15 CT gene families (Figure 11, pg 72). Some of these CT gene families appear to have undergone recent

UNIVERSITY of the WESTERN CAPE

https://etd.uwc.ac.za/

expansions through duplication in the lineages leading to the present day *H. sapiens* and *P. troglodytes* after their divergence from a rodent – primate cenancestor and hence, are subsequently paralogous by nature[65,66,68,70,146]. The CT genes that are restricted to the primate phylogenetic scope would have arisen after the divergence of the primate and rodent lineages from a cenancestor approximately 90 million years ago (mya) (Figure 10, pg 70). Thus, these CT genes can be considered younger then 90 million years as they do not have any identified homologues from the eutherian to the metazoan phylogenetic scopes (Figure 11, pg 72 and Figure 10, pg 70). The primate restricted CT genes are distributed amongst chromosomes; X (25 CT genes), 21 (7 CT genes) and 18 (2 CT genes), Figure 11, pg 72. From the 34 CT genes which are restricted to the primate specific phylogenetic scope, 56% are known CTAs (Figure 11, pg 72 and Figure 12, pg 71).

The Eutherian Scope.

Homologues for 48 of the CT genes, (58.5 % of 82 CT genes), were identified in the eutherian phylogenetic scope, of which the cenancestor between primates and rodents is estimated to be approximately older then 90 million years (CT 20 (TSP50) - CT 35 (SPO11); Figure 11, pg 72, Figure 12, pg 71 and Figure 10, pg 70). Of the 48 CT gene homologues identified in *Mus musculus* and /or *Rattus norvegicus*, 30 are restricted to the eutherian phylogenetic scope (CT 20 (TSP50) - CT 43 (FATE), Figure 11, pg 72). These 30 eutherian restricted CT genes also comprise of 15 CT gene families with the *H. sapiens*' X chromosome hosting 26 of the eutherian restricted CT genes while chromosomes 3, 12, 19 and 21 each house a single CT gene (Figure 11, pg 72).

There are two possible scenarios that may explain the presence of the 30 CT gene homologues in the eutherian phylogenetic scope and not in the vertebrate or metazoan phylogenetic scopes. Either these 30 eutherian restricted CT genes arose in a primate –

UNIVERSITY of the WESTERN CAPE
https://etd.uwc.ac.za/

rodent cenancestor after the divergence of the mammalian and teleost lineages ~450 mya, and hence are not present in a vertebrate phylogenetic scope as they were not present in a vertebrate – eutherian cenancestor (Figure 10, pg 70). Alternatively, these 30 eutherian restricted CT genes may have been present in a vertebrate – eutherian cenancestor, but were only conserved in the eutherian lineage and lost in the teleost lineage, after their divergence ~450 mya (Figure 10, pg 70). However, some of the CT gene families to which these 30 CT genes belong to have been identified as paralogous[65,66,68,70,146]. Hence it would appear more likely that some of these 30 CT genes may have arisen in a primate – rodent cenancestor belonging to the eutherian phylogenetic scope, rather then a eutherian – vertebrate cenancestor belonging to the vertebrate phylogenetic scope (Figure 10, pg 70).

Unfortunately CT gene homologues were not sampled in species representative of a monotreme, avian or amphibian phylogenetic scope. The lineages leading to the monotreme, avian or amphibian species is estimated have diverged from a vertebrate cenancestor ~185 mya, ~310 mya and ~360 mya respectively[134,135,137,141]. The sampling of CT genes in species belonging to these phylogenetic scopes would provide an increased resolution in determining which lineages these 30 eutherian restricted CT genes are conserved in[134,135,137,141]. Presently, the age of these 30 eutherian restricted genes can only be estimated as being approximately older then ~90 mya and less then ~450 mya (Figure 10, pg 70 and Figure 11, pg 72). Of the 30 eutherian specific CT genes, 24 (80% of the 30 CT genes) are identified CTAs, Figure 11, pg 72 and Figure 12, pg 71.

UNIVERSITY of the
WESTERN CAPE

The Vertebrate Scope

From 82 CT genes, 19.5 % have putative homologues present in the vertebrate phylogenetic scope, these include CT 8 (SYCP-1) – CT 41.2 (NY-CO-45), CT 31 (PLU-1), CT 13 (HAGE) and CT 35 (SPO11), Figure 11, pg 72. The 16 CT genes which have homologues identified in the vertebrate scope belong to 15 CT gene families (Figure 11, pg

https://etd.uwc.ac.za/

72). Ten of the 16 CT genes have homologues which are restricted to the vertebrate phylogenetic scope (CT 8 (SYCP-1) – CT 14 (SAGE), Figure 11, pg 72) and belong to 10 different CT gene families, with each CT gene family comprising of a single member (Table A, Appendix pg 82). These 10 CT genes which are restricted to the vertebrate phylogenetic scope are distributed in the human genome amongst chromosomes 1, 6 and 15 which each host a single CT gene, chromosomes 8 and 11 that host 2 CT genes and the X chromosome which hosts 3 CT genes (Figure 11, pg 72). The divergence time between the teleost and mammalian lineages that lead to the present day metazoan species *H. sapiens*, *Fugu rubripes* and *Danio rerio* is estimated to be ~450 mya (Figure 10, pg 70). Hence, the 10 CT genes which are restricted to the vertebrate phylogenetic scope would be older then ~450 mya as they would have been present in a vertebrate – eutherian cenancestor.

As homologues for these 10 CT genes were not identified in any of the species constituting the metazoan phylogenetic scope (Figure 10, pg 70) they may have not been conserved in the arthropod and nematode lineages leading to *Drosophila melanogaster, Anopheles gambiae* and *Caenorhabditis elegans*. The 10 CT genes restricted to the vertebrate phylogenetic scope would have been conserved in the teleost and mammalian lineages leading to the species present in the vertebrate, eutherian and primate phylogenetic scopes (Figure 10, pg 70). It is also possible that the 10 CT genes which are restricted to the vertebrate phylogenetic scope may have arisen in the vertebrate – eutherian cenancestor, rather then a metazoan – vertebrate cenancestor (Figure 10, pg 70). The usage of a chordate phylogenetic scope whose estimated time of divergence from a metazoan cenancestor is ~550 mya would provide an increased resolution in helping to determine whether these 10 vertebrate restricted CT genes were present in a metazoan – chordate, or specific to a vertebrate – eutherian cenancestor[134,137,138].

UNIVERSITY of the
WESTERN CAPE

https://etd.uwc.ac.za/

At present, the 10 CT genes with homologues restricted to the vertebrate phylogenetic scope can only be estimated as being older the ~450 mya and younger then ~990 mya. Of these 10 CT genes which are restricted to the vertebrate phylogenetic scope, 20% are known CTAs, Figure 11, pg 72 and Figure 12, pg 71.

The Metazoan Scope

From the total of 82 CT genes, approximately 10 % have putative homologues identified in the metazoan phylogenetic scope (Figure 12, pg 71 and Figure 11; CT 9 (BRDT) – CT 35 (SPO11), pg 72). Two of the CT genes, CT 27 (BORIS) and CT 33 (MORC), do not have any identified homologues in the vertebrate scope while putative homologues were identified in the metazoan phylogenetic scope (Figure 11, pg 72). The absence of homologues for CT 27 (BORIS) and CT 33 (MORC) in the vertebrate phylogenetic scope may reflect the lack of conservation of those CT genes in the teleost lineage. Conversely, the absence of identified putative homologues for CT 27 (BORIS) and CT 33 (MORC) in the vertebrate phylogenetic scope may be an artefact resulting from the variability of the comprehensiveness of protein coding regions present for the individual metazoan species used in this study[137,139].

UNIVERSITY of the
WESTERN CAPE

The 8 CT genes that have homologues present in the metazoan phylogenetic scope can be roughly categorized as ancient by virtue of being present in the broadest phylogenetic scope used in this study (Figure 10, pg 70 and Figure 11, pg 72). As putative homologues for these 8 CT genes are identified in the metazoan phylogenetic scope, they would have been present in a metazoan cenancestor common to all the metazoan species used in this study and hence, can be inferred as being older then ~990 mya. The 8 CT genes identified in a metazoan phylogenetic scope may not necessarily be a result of these 8 CT genes being the oldest of all the 82 CT genes. It is possible that these 8 CT genes have identifiable putative

https://etd.uwc.ac.za/

homologues in the metazoan phylogenetic scope as they have been conserved amongst the metazoan species sampled for CT gene homologues, where as the CT genes restricted to narrower phylogenetic scopes have not been well conserved.

These 8 ancient CT genes belong to 7 different CT gene families, with only CT 41.1 (TDRD1) and CT 41.2 (NY-CO-45) belonging to the same CT gene family (Figure 11, pg 72 and Table A, Appendix pg 82). The remaining 6 ancient CT genes are all members of 6 different CT gene families, each of which comprise of only a single CT gene representative member (Figure 11, pg 72 and Table A, Appendix pg 82). These 8 ancient CT genes are distributed amongst chromosomes 3 and 10 that each host a single CT gene and chromosomes 1, 6 and 20 each hosting 2 of the remaining ancient 6 CT genes. Of the 8 CT genes which have homologues present in the metazoan phylogenetic scope, all 8 are CT transcripts (Figure 11, pg 72 and Figure 12, pg 71).

UNIVERSITY of the
WESTERN CAPE

https://etd.uwc.ac.za/

Why are the Majority of CT Genes Restricted to the Primate and Eutherian Phylogenetic Scopes?

The bulk of the CT genes appear to be fairly recent as defined by the phylogenetic scopes for which CT gene homologues were found to be present. The majority of the CT genes restricted to the primate and eutherian phylogenetic scopes are CTAs, X-linked and belong to CT gene families which comprise of multiple members, Figure 13.

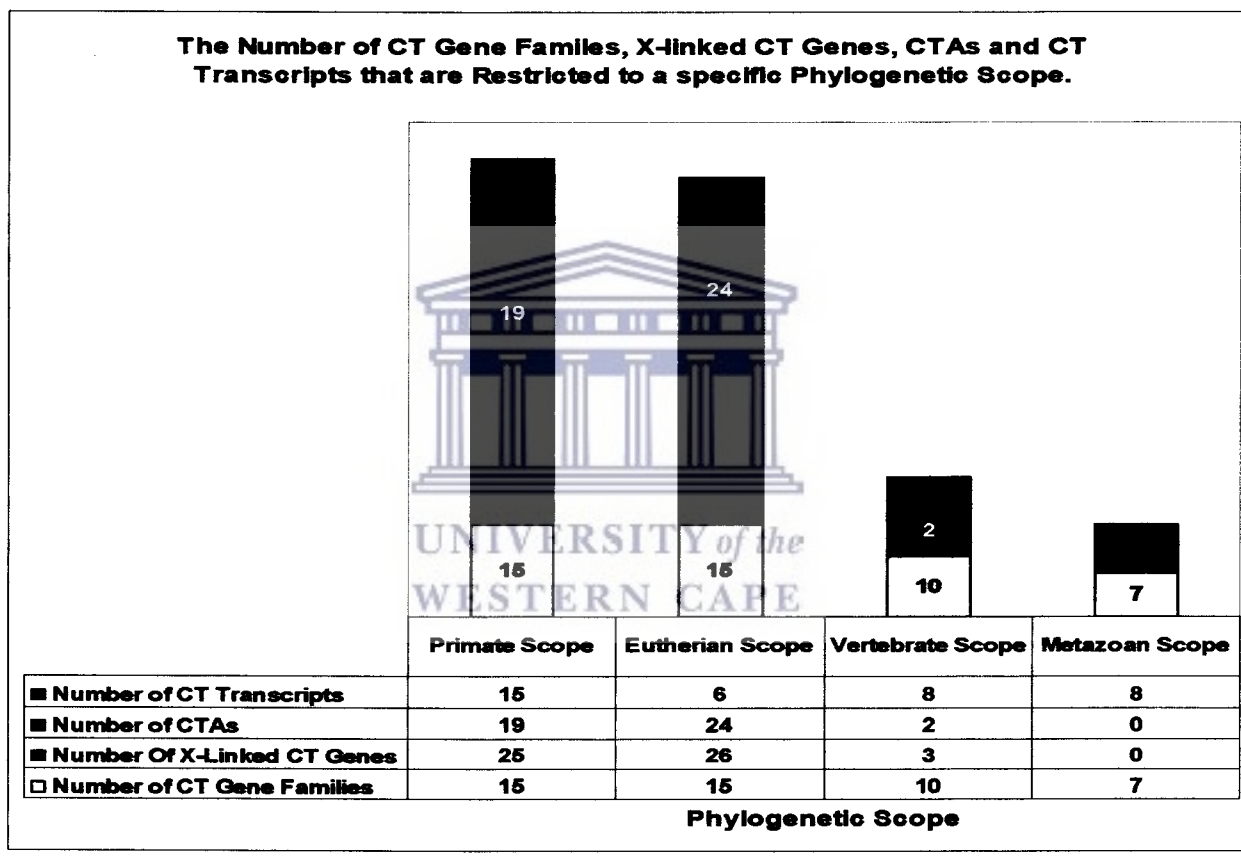


Figure 13 : The majority of the CT genes that are restricted to the Primate and Eutherian phylogenetic scopes are X-linked and belong to CT gene families that are comprised of multiple members as opposed to the CT genes present in the vertebrate and metazoan phylogenetic scopes.

Of the 82 CT genes, 64 are restricted to the primate and eutherian phylogenetic scopes and are members of 27 CT gene families. In some cases, the high number of CT genes which are restricted to the primate and eutherian phylogenetic scopes can be attributed to the recent duplication and expansion of different CT gene families[65,66,68,70,146]. Indeed, some CT gene

https://etd.uwc.ac.za/

families such as CT 1 (MAGEA), CT 2 (BAGE), CT 4 (GAGE), CT 5 (SSX), CT 6 (NY-ESO-1 / LAGE), CT 11 (SPANX) and CT 12 (XAGE / GAGED) comprise of CT genes that have been found to have arisen recently by duplication[65,66,68,70,146].

In some cases, specific CT gene families such as CT 2 (BAGE), CT 6 (NY-ESO-1 / LAGE) and CT 11 (SPANX) have been shown to be specific for the hominidae lineage leading to the present day *H. sapiens*, *P. troglodytes* and *Gorilla gorilla* species[65,68,70]. The separation between the hominidae lineage (human, chimpanzee and gorilla) and the lineage leading to the *Pongo pygmaeus* (orangutan) species is thought have occurred ~12 mya[147]. Hence some of the CT genes are younger then ~12 mya.

For some CT gene families like CT 1 (MAGEA), CT 4 (GAGE), CT 5 (SSX) and CT 12 (XAGE / GAGED) there do appear to be ancestral genes present in the human genome from which some of the CT gene families have arisen through retroposition and subsequent duplications[66,146]. The ancestral genes from which these CT gene families have arisen from would have been present in a primate – rodent cenancestor. However, the present day human CT gene families cannot be considered ancient as they have, in some cases, arisen by duplication after divergence from a primate – rodent cenancestor and hence, are paralogous[66,146].

UNIVERSITY of the
WESTERN CAPE

The human genome is known to have undergone duplication events that are both ancient (~450 mya) as well as recent (~35 mya) with the latter category constituting ~5% of the human genome sequence[6,146,147]. Genomic sequence duplication is thought to have significantly contributed to the evolution of the human genome and in some cases, may have contributed to speciation[6,147-149]. In some cases, duplication of genomic sequences enables the evolution novel gene functions e.g the duplication of the Xq28 opsin genes in an old world monkey – ape cenancestor is postulated to have resulted in trichromatic colour vision and hence the ability to perceive a wider spectrum of discernible light by primates[147,150]. An

https://etd.uwc.ac.za/

analysis of the X chromosome for a class of human duplications known as inverted repeats by Warburton *et al*[151] revealed that most of these duplications contained genes predominantly expressed in the testis, with a sizable proportion housing CT genes. A large number of genes which have undergone duplication in the lineage leading to humans after the divergence of a rodent – primate cenancestor are found to be involved in processes such as immunity, reproduction and olfaction[6,8,9,146,147]. The majority of the CT genes restricted to the primate and eutherian phylogenetic scopes are members of CT gene families that have undergone recent expansions by duplication, appear to be involved in reproduction by virtue of their testis tropic expression profile and in the case of CTAs, are involved in immunity.

https://etd.uwc.ac.za/

Conclusions.

The human genome provides a useful, functioning framework for the systemisation of biological information pertaining to the CT genes. The genomic distribution of the CT genes in the human genome is non-random, with 66% of the 83 CT genes residing on the human X chromosome. Of the 12 CT gene families that comprise of multiple members, 75% are X – linked. In some instances, the genomic distribution of the CT genes appears to be influenced by their spatial – temporal expression profile. The X – linked CT genes are unlikely to participate in biological events past male meiosis I, unless they form long lived transcripts or are not transcriptionally silenced during the onset of male meiosis I [71,72,84,101]. The CT genes known to be involved in male meiosis; CT 8 (SYCP), CT 15 (ADAM2) and CT 23 (OY-TES-1) reside on chromosomes 1, 8 and 12 respectively, as opposed to the X chromosome that is transcriptionally silenced during the onset of male meiosis I.

CT gene expression is not exclusively restricted to the testis and neoplastic cells as was initially thought. Both an *in-silico* and a wet-lab approach indicate that CT gene transcripts can be detected in various non-gametogenic tissues [37]. Although CT gene transcripts can be detected in a variety of non-gametogenic tissues, both types of gene expression capture technologies concur in that the CT genes do exhibit a definite testis tropic expression profile. Expression for some of the CT genes is known to be epigenetically controlled by the methylation and demethylation of their promoter sequences [33,42,55,56]. Expression of these CT genes in the testis and during cancer maybe attributed to the global demethylation status of male germ cells and genome-wide hypomethylation events that occur during cancer [33,42,55,56]. Modulation and specificity of some gene expression events during development, programmed cell specialization / differentiation and oncogenesis are attributable to epigenetic modifications which enable the expression of certain genes at a particular point in

UNIVERSITY of the WESTERN CAPE

https://etd.uwc.ac.za/

time [5,26,49,50,52]. At present, the distribution of CT gene expression in terms of normal tissues is known. However, the particular time points and stages in development, programmed cell specialization / differentiation and oncogenesis that the CT genes are functionally expressed in are at present, not known. Even though CT gene expression is epigenetically controlled and the CT genes are predominately expressed in the testis, their contribution to any developmental, programmed cell specialization / differentiation and oncogenic processes is unclear. The CT genes are expressed in various types of neoplasms with the suggestion that CT genes are more predominantly expressed in malignant neoplasms [33,37]. Whether or not CT gene expression may be conducive to the development and malignancy of neoplasms by participating in a subverted pseudo-developmental pathway that provides a selective advantage to emerging neoplastic cells remains to be resolved.

A large proportion (77% of 83 CT genes) could only be found in a eutherian phylogenetic scope, 51 of the 66 eutherian restricted CT genes are X-linked. The majority of CT genes confined to the eutherian phylogenetic scope are CTAs and are members of 27 CT gene families. Indeed, some of the CT genes are fairly recent and could only be found in a primate phylogenetic scope which forms a subset of the eutherian phylogenetic scope. The human genome is known to have undergone a series of gene duplications, both ancient (~450 mya) and new (~35 mya), with the latter forming approximately 5% of the human genome [6,146,147]. Analyses of the human genome by the International Human Genome Sequencing Consortium indicate a large number of genes that have undergone recent duplications in the lineages leading to *Homo sapiens* after the divergence of a rodent – primate cenancestor, and are involved in processes such as immunity, olfaction and reproduction [6,8,9,146,147]. The majority of the CT genes restricted to the primate and eutherian phylogenetic scopes are members of 27 CT gene families that have undergone recent expansions by duplications and may be involved in reproduction due to their testis tropic and spatial – temporal expression

UNIVERSITY of the WESTERN CAPE

https://etd.uwc.ac.za/

profiles and in the case of the CTAs, are involved in immunity due to their immunogenic properties in 20 – 40 % of cancer types. As the majority of the CT genes appear to be fairly recent due to their fairly narrow taxonomic distribution, it is possible that the CT genes may be involved in certain aspects of development, reproduction and oncogenesis exclusive to the taxonomic species within which the CT genes appear to be confined to.

https://etd.uwc.ac.za/

Appendix.

Table A : A full listing of the CT genes used in this study

CT[1] Identifier	Gene[2] Name	RefSeq[3] ID	Cytogenetic location	Genomic[4] co-ordinates(bp)	Number[4] of exons
CT 1.1	MAGE-A1	NM_004988	Xq28	150912768 - 150917329	3 EXONS
CT 1.2a	MAGE-A2	NM_005361	Xq28	150500732 - 150504708	3 EXONS
CT 1.2b	MAGE-A2	NM_175742	Xq28	150536022 - 150539953	3 EXONS
CT 1.3	MAGE-A3	NM_005362	Xq28	150552266 - 150555853	3 EXONS
CT 1.4	MAGE-A4	NM_002362	Xq28	149702323 - 149711253	3 EXONS
CT 1.5	MAGE-A5	NM_021049	Xq28	149900138 - 149904057	3 EXONS
CT 1.6a	MAGE-A6	NM_005363	Xq28	150484858 - 150488427	3 EXONS
CT 1.6b	MAGE-A6	NM_175868	Xq28	Transcript variant	3 EXONS
CT 1.7	MAGE-A7	NG_001156	Xq28	Psuedogene	
CT 1.8	MAGE-A8	NM_005364	Xq28	147668506 - 147673127	3 EXONS
CT 1.9	MAGE-A9	NM_005365	Xq28	147569263 - 147575058	4 EXONS
CT 1.10	MAGE-A10	NM_021048	Xq28	149920523 - 149924647	4 EXONS
CT 1.11	MAGE-A11	NM_005366	Xq28	147473361 - 147550140	5 EXONS
CT 1.12	MAGE-A12	NM_005367	Xq28	150516915 - 150520785	2 EXONS
CT 2.1	BAGE	NM_001187	21p11.1	10080123 - 10120799	
CT 2.2	BAGE2	NM_182482	21p11.2	10043150 - 10120787	
CT 2.3	BAGE3	NM_182481	21p11.2	10071889 - 10080650	2 EXONS
CT 2.4	BAGE4	NM_181704	21p11.1	10119404 - 10119508	
CT 2.5	BAGE5	NM_182484	21p11.1	10119404 - 10119508	1 EXON
CT 3.1	MAGE-B1	NM_002363	Xp21.3	29623274 - 29631581	2 EXONS
CT 3.2	MAGEB2	NM_002364	Xp21.3	29595103 - 29599630	2 EXONS
CT 3.3	MAGEB5	AF333705	Xp22	27191868 - 27212410	
CT 3.4	MAGEB6	NM_173523	Xp22.12	25571993 - 25575189	2 EXONS
CT 4.1	GAGE1	NM_001468	Xp11.4-p11.2	48272768 - 48290684	5 EXONS
				48252586 - 48288253	6 EXONS
CT 4.2	GAGE2	NM_001472	Xp11.4-p11.2	48214327 - 48290684	
CT 4.3	GAGE3	NM_001473	Xp11.4-p11.2	48262104 - 48288256	
CT 4.4	GAGE4	NM_001474	Xp11.4-p11.2	48252552 - 48259872	
CT 4.5	GAGE5	NM_001475	Xp11.4-p11.2	48252560 - 48259875	
CT 4.6	GAGE6	NM_001476	Xp11.4-p11.2	48252553 - 48259872	
CT 4.7	GAGE7	NM_021123	Xp11.4-p11.2	48252518 - 48259761	5 EXONS
CT 4.8	GAGE8	NM_012196	Xp11.4-p11.2	48214327 - 48278968	5 EXONS
CT 5.1	SSX-1	NM_005635	Xp11.23-p11.22	47160779 - 47172861	8 EXONS
CT 5.2a	SSX-2	(isoform a)	Xp11.23-p11.22	51747000 - 51786141	9 EXONS
CT 5.2b	SSX-2	NM_003147	Xp11.23-p11.22	51639902 - 51703040	8 EXONS
CT 5.3a	SSX-3	NM_021014	Xp11.23	47200867 - 47262238	8 EXONS
CT 5.3b	SSX-3	NM_175711	Xp11.23	47200867 - 47262238	8 EXONS
CT 5.4a	SSX-4	NM_005636	Xp11.23	47307506 - 47317326	8 EXONS
CT 5.4b	SSX-4	NM_175729	Xp11.23	47288950 - 47298767	7 EXONS
CT 6.1	NY-ESO-1	NM_001327	Xq28	152280969 - 152282626	3 EXONS
CT 6.2a	LAGE 1a	NM_020994	Xq28	152347797 - 152349404	
CT 6.2b	LAGE1b	NM_020994	Xq28	152347797 - 152349404	2 EXONS
CT 7.1	MAGE-C1	NM_005462	Xq26	139686796 - 139690221	1 EXON
CT 7.2a	MAGE-C3	NM_138702	Xq27.2	139619704 - 139679220	8 EXONS
CT 7.2b	MAGE-C3	NM_177456	Xq27.2	139676706 - 139679220	3 EXONS

https://etd.uwc.ac.za/

CT[1] Identifier	Gene[2] Name	RefSeq[3] ID	Cytogenetic location	Genomic[4] co-ordinates(bp)	Number[4] of exons
CT 8	SYCP-1	NM_003176	1p13-p12	114696216 - 114836367	32 EXONS
CT 9	BRDT	NM_001726	1p22.1	91886729 - 91951778	20 EXONS
CT 10	MAGEE1	NM_016249	Xq27	139976193 - 139986680	3 EXONS
CT 11.1a	SPANXA1	NM_013453	Xq27.1	139371404 - 139372490	2 EXONS
CT 11.1b	SPANXA2	NM_145662	Xq27.1	139365408 - 139366438	2 EXONS
CT 11.2	SPANXB1	NM_032461	Xq27.1	138790363 - 138791478	2 EXONS
CT 11.3	SPANXC	NM_022661	Xq27.1	139029199 - 139030249	2 EXONS
CT 11.4	SPANXD	NM_032417	Xq27.1	139479171 - 139480257	2 EXONS
CT 12.1a	XAGE-1a	NM_020411	Xp11.22-p11.21	51507847 - 51512824	4 EXONS
CT 12.1b	XAGE-1b	(isoform)			
CT 12.1c	XAGE-1c	NM_133431	Xp11.22-p11.21	51478716 - 51483693	4 EXONS
CT 12.1d	XAGE-1d	NM_133430	Xp11.22-p11.21	51494953 - 51499930	4 EXONS
CT 12.2	XAGE-2	NM_130777	Xp11.22-p11.21	51297145 - 51303887	5 EXONS
CT 12.3a	XAGE-3a	NM_133179	Xp11.22-p11.21	51858351 - 51863890	5 EXONS
CT 12.3b	XAGE-3b	NM_130776	Xp11.22-p11.21	51858351 - 51863890	5 EXONS
CT 12.4	XAGE-4	AJ318895	Xp11.22	54647931 - 54650238	3 EXONS
CT 13	HAGE	NM_018665	6q12-q13	74100069 - 74122668	17 EXONS
CT 14	SAGE	NM_018666	Xq26	133681392 - 133700826	20 EXONS
CT 15	ADAM2	NM_001464	8p11.2	39618625 - 39713143	21 EXONS
CT 16	PAGE-5	NM_130467	Xp11.22	54213577 - 54217327	5 EXONS
CT 17	LIPI	NM_145317	21q11.2		
CT 18	NA88	NR_001559	Xp22.12	25937880 - 25939314	
CT 19	IL-13r	NM_001560	Xq24	116615661 - 116682598	11 EXONS
CT 20	TSP50	NM_013270	3p14-p12	46714199 - 46836177	2 EXONS
CT 21.1	CTAGE-1	NM_022663	18p11.2	18246580 - 18246804	1 EXON
CT 21.2	CTAGE-2	NM_172241	18p11.2		2 EXONS
CT 22	SPA17	NM_017425	11q24.2	124081392 - 124102337	5 EXONS
CT 23	OY-TES-1	NM_032489	12p13.31	6617503 - 6626841	10 EXONS
CT 24.1	CSAGE	NM_153478	Xq28	150526390 - 150527130	2 EXONS
CT 24.2	TRAG-3	NM_004909	Xq28	150494380 - 150495360	2 EXONS
CT 25.1a	MMA-1a	NM_032589	21q22.2	38414822 - 38448892	3 EXONS
CT 25.1b	MMA-1b	NM_203428	21q22.2	Transcript variant.	
CT 26	CAGE	NM_182699	Xp22.13	22379504 - 22381630	1 EXON
CT 27	BORIS	NM_080618	20q13.31	56757645 - 56785584	11 EXONS
CT 28	HOM-TES-85	NM_016383	Xq24	113288572 - 113306372	4 EXONS

UNIVERSITY of the

https://etd.uwc.ac.za/

CT[1] Identifier	Gene[2] Name	RefSeq[3] ID	Cytogenetic location	Genomic[4] co-ordinates(bp)	Number[4] of exons
CT 29	AF15q14	NM_020380	15q14	38602519 - 38670424	23 EXONS
CT 30	E2F-like	HCA661 NM_016521	Xq26.2	131056304 - 131057983	
CT 31	PLU-1	NM_006618	1q32.1	199984790 - 200066100	28 EXONS
CT 32	LDHC	NM_002301	11p15.5-p15.3	18380419 - 18393702	8 EXONS
CT 33	MORC	NM_014429	3q13	109997992 - 110157869	28 EXONS
CT 34	SGY-1	NM_014419	19q13.33	54558974 - 54570182	5 EXONS
CT 35	SPO11	NM_012444	20q13.2-q13.3	56590237 - 56604470	13 EXONS
CT 36	TPX1	NM_003296	6p21-qter	49707322 - 49723745	7 EXONS
CT 37	NY-SAR-35	NM_152578	Xq27.3-q28	145730810 - 145774207	5 EXONS
CT 38	FTHL17	NM_031894	Xp21	30450945 - 30451496	1 EXON
CT 39.1a	NFX2	NM_017809	Xq22.1	100333841 - 100353490	21 EXONS
CT 39.1b	NFX2	NM_022053	Xq22.1	100387174 - 100406871	21 EXONS
CT 40	TAF7L	NM_024885	Xq22.1	99295099 - 99319902	13 EXONS
CT 41.1	TDRD1	NM_198795	10q26.11	115627837 - 115652765	18 EXONS
CT 41.2	NY-CO-45	AF039442	6p21.1	46706512 - 46708750	2 EXONS
CT 42	TEX15	NM_031271	8p12	30746730 - 30764201	4 EXONS
CT 43	FATE	NM_033085	Xq28	149502116 - 149509275	5 EXONS
CT 44	TPTE	NM_013315	21p11	9929055 - 10013188 bp	22 EXONS

1 CT nomenclature as devised by Scanlan *et al*[37].
2 Generic gene name.
3 RefSeq identifiers supplied by the Ludwig Institute for Cancer Research.
4 Based on Ensembl version 19.34b.2 based on NCBI build 34, 9 Feb 2004.
CT Antigens.

UNIVERSITY of the WESTERN CAPE

https://etd.uwc.ac.za/

Table B : The curated Xq28 dataset used for Chapter 4 was exported from Ensembl (19.34b.2 based on NCBI build 34, 9 Feb 2004) using EnsMart (version 19.2). The Xq28 CT are shaded in gray.

External Gene ID	Band	Start Position (bp)	End Position (bp)	% GC content	Strand	Ensembl Gene ID	RefSeq ID	MIM ID
FMR2	Xq28	146287692	146787760	38	1	ENSG00000155966.2	NM_002025	309548
IDS	Xq28	147269524	147292720	47	-1	ENSG00000010404.3	NM_006123	309900
							NM_000202	
Q14603	Xq28	147312281	147313205	56	1	ENSG00000176289.1		
NM_178124	Xq28	147327954	147334620	55	1	ENSG00000155976.4	NM_178124	
NM_032508	Xq28	147383877	147419236	42	-1	ENSG00000155984.2	NM_032508	
MAGEA11	Xq28	147473361	147550140	44	-1	ENSG00000185247.3	NM_005366	300344
NM_016153	Xq28	147561993	147564186	56	1	ENSG00000171116.1	NM_016153	
MAGEA9	Xq28	147569263	147575058	58	1	ENSG00000166008.1	NM_005365	300342
MAGEA8	Xq28	147668506	147673127	54	1	ENSG00000156009.1	NM_005364	300341
CXorf6	Xq28	148254188	148322907	47	1	ENSG00000013619.2	NM_005491	300120
MTMR1	Xq28	148502313	148574036	42	1	ENSG00000063601.3	NM_003828	300171
							NM_176789	
CD99L2	Xq28	148575278	148707647	42	-1	ENSG00000102181.4	NM_031462	
HMGB3	Xq28	148792237	148797802	52	1	ENSG00000029993.3	NM_005342	300193
GPR50	Xq28	148985593	148990377	56	1	ENSG00000102195.1	NM_004224	300207
NM_173493	Xq28	149349705	149462822	39	1	ENSG00000166049.2	NM_173493	
TMG3_HUMAN	Xq28	149486077	149487118	57	1	ENSG00000130032.3	NM_024082	
NM_033085	Xq28	149502116	149509275	48	1	ENSG00000147378.1	NM_033085	300450
CNGA2	Xq28	149529462	149530244	50	1	ENSG00000183862.1		300338
MAGEA4	Xq28	149702323	149711253	58	1	ENSG00000171047.4	NM_002362	300175
GABRE	Xq28	149739210	149760764	46	-1	ENSG00000102287.2	NM_021990	300093
							NM_021984	
							NM_004961	
							NM_021987	
MAGEA5	Xq28	149900138	149904057	54	-1	ENSG00000183686.1	NM_021049	300340
MAGEA10	Xq28	149920523	149924647	55	-1	ENSG00000124260.2	NM_021048	300343
GABRA5	Xq28	149954145	150237443	37	-1	ENSG00000011677.1	NM_000810	137142

https://etd.uwc.ac.za/

External Gene ID	Band	Start Position (bp)	End Position (bp)	% GC content	Strand	Ensembl Gene ID	RefSeq ID	MIM ID
GABRA5							NM_000808	305660
GABRQ	Xq28	150424250	150439357	48	1	ENSG00000147402.2	NM_018558	300349
MAGEA6	Xq28	150484858	150488427	55	1	ENSG00000183305.1	NM_175868	300176
							NM_005363	
TRG3_HUMAN	Xq28	150494380	150495360	50	-1	ENSG00000185377.1	NM_004909	
MAGEA2	Xq28	150500732	150504708	55	1	ENSG00000182895.3	NM_153488	300173
							NM_175742	
							NM_175743	
							NM_005361	
Q96DX6	Xq28	150513593	150514234	52	-1	ENSG00000184324.1	NONE	
MAGEA12	Xq28	150516915	150520785	55	-1	ENSG00000147381.1	NM_005367	300177
NM_153478	Xq28	150526390	150527130	53	1	ENSG00000183956.1	NM_153478	
							NM_153479	
MAGEA3	Xq28	150552266	150555853	55	-1	ENSG00000185913.1	NM_005362	300174
CETN2	Xq28	150613472	150615917	42	-1	ENSG00000147400.3	NM_004344	300006
NSDL_HUMAN	Xq28	150617234	150655885	45	1	ENSG00000147383.1	NM_015922	300275
								308050
ZNF185	Xq28	150704278	150756698	50	1	ENSG00000147394.3	NM_007150	300381
PNMA5	Xq28	150774984	150778371	56	-1	ENSG00000183213.1	NM_052926	
Q8N624	Xq28	150844670	150846435	60	1	ENSG00000183837.1		
NM_032882	Xq28	150858479	150861015	66	1	ENSG00000177475.1	NM_032882	
MAGEA1	Xq28	150912768	150917329	55	-1	ENSG00000126977.1	NM_004988	300016
ZNF275	Xq28	151041461	151045312	53	1	ENSG00000063587.1		
TREX2	Xq28	151178082	151203949	55	-1	ENSG00000183479.2	NM_017518	300370
							NM_080699	
							NM_080700	
							NM_080701	
							NM_007205	
BGN	Xq28	151228315	151242908	62	1	ENSG00000182492.1	NM_001711	301870
ATP2B3	Xq28	151269610	151313660	58	1	ENSG00000067842.3	NM_021949	300014
NM_152274	Xq28	151321289	151332481	55	-1	ENSG00000147382.4	NM_152274	
DUSP9	Xq28	151376838	151384675	63	1	ENSG00000130829.4	NM_001395	300134

https://etd.uwc.ac.za/

External Gene ID	Band	Start Position (bp)	End Position (bp)	% GC content	Strand	Ensembl Gene ID	RefSeq ID	MIM ID
Q8N4R0	Xq28	151403094	151407161	65	-1	ENSG00000130822.2		
SLC6A8	Xq28	151421296	151429944	65	1	ENSG00000130821.2	NM_005629	300036
BA31_HUMAN	Xq28	151433860	151457784	54	-1	ENSG00000185825.1	NM_005745	300398
ABCD1	Xq28	151458227	151478120	58	1	ENSG00000101986.2	NM_000033	300371
								300100
PLXNB3	Xq28	151498886	151512398	66	1	ENSG00000102012.2	NM_005393	300214
STK23	Xq28	151514422	151519088	64	1	ENSG00000184343.1	NM_014370	
IDH3G	Xq28	151519126	151527870	58	-1	ENSG00000067829.4	NM_004135	300089
							NM_174869	
SSR4	Xq28	151526875	151531857	61	1	ENSG00000180879.2	NM_006280	300090
NM_032512	Xq28	151535527	151541957	63	-1	ENSG00000067840.2	NM_032512	
L1CAM	Xq28	151595291	151609215	60	-1	ENSG00000102022.2	NM_000425	308840
							NM_024003	303350
							NM_000425	307000
							NM_024003	
AVPR2	Xq28	151635889	151640544	62	1	ENSG00000126895.3	NM_000054	304800
ARHGAP4	Xq28	151640735	151659618	57	-1	ENSG00000089820.3	NM_001666	300023
ARD1	Xq28	151663281	151667835	56	-1	ENSG00000102030.2	NM_003491	300013
RENBP	Xq28	151668627	151678136	55	-1	ENSG00000102032.2	NM_002910	312420
HCFC1	Xq28	151680918	151704195	59	-1	ENSG00000172534.2	NM_005334	300019
CXorf12	Xq28	151706147	151716548	51	1	ENSG00000177854.1	NM_003492	300059
IRAK1	Xq28	151743864	151753246	61	-1	ENSG00000184216.2	NM_001569	300283
MECP2	Xq28	151755168	151825669	46	-1	ENSG00000169057.3	NM_004992	300005
								312750
Q8WYS0	Xq28	151785853	151789936	45	1	ENSG00000182269.1		
OPN1LW	Xq28	151877662	151892305	52	1	ENSG00000102076.1	NM_020061	303900
OPN1MW	Xq28	151916071	151929435	52	1	ENSG00000147380.1	NM_000513	303800
CXorf2	Xq28	151929866	151946861	50	-1	ENSG00000182242.1	NM_001586	300092
TKTL1	Xq28	152001579	152026603	48	1	ENSG00000007350.3	NM_012253	300044
FLNA	Xq28	152044801	152067547	62	-1	ENSG00000071872.2	NM_001456	300017

https://etd.uwc.ac.za/

UNIVERSITY of the WESTERN CAPE

								300049
External Gene ID	Band	Start Position (bp)	End Position (bp)	% GC content	Strand	Ensembl Gene ID	RefSeq ID	MIM ID
FLNA								304120
								305620
								309350
								311300
EMD	Xq28	152075635	152077797	61	1	ENSG00000102119.1	NM_000117	300384
								310300
DNASE1L1	Xq28	152098015	152108302	54	-1	ENSG00000013563.3	NM_006730	300081
TAZ	Xq28	152107792	152117976	55	1	ENSG00000102125.4	NM_181314	300394
							NM_000116	302060
							NM_181312	300069
							NM_181313	300183
							NM_181311	604169
ATP6AP1	Xq28	152124918	152132766	56	1	ENSG00000071553.2	NM_001183	300197
GDI1	Xq28	152133409	152139714	59	1	ENSG00000102129.2	NM_001493	300104
								309541
XAP5_HUMAN	Xq28	152140396	152146897	60	1	ENSG00000071859.2	NM_004699	
PLXNA3	Xq28	152154527	152169171	63	1	ENSG00000130827.1	NM_017514	300022
NM_006014	Xq28	152174019	152175156	61	-1	ENSG00000126897.2	NM_006014	300060
UBL4	Xq28	152179960	152182842	60	-1	ENSG00000102178.1	NM_014235	312070
SLC10A3	Xq28	152183548	152186895	60	-1	ENSG00000126903.4	NM_019848	312090
FAM3A	Xq28	152202404	152212413	55	-1	ENSG00000071889.4	NM_021806	
G6PD	Xq28	152227157	152241921	55	-1	ENSG00000160211.1	NM_000402	305900
IKBKG	Xq28	152238010	152260811	56	1	ENSG00000073009.2	NM_003639	300248
								300291
								300301
								308300
NM_173168	Xq28	152267030	152267737	46	1	ENSG00000182204.1	NM_173168	
CTAG1	Xq28	152280969	152282626	65	1	ENSG00000183678.2	NM_001327	300156
							NM_139250	

https://etd.uwc.ac.za/

UNIVERSITY of the WESTERN CAPE

CTAG2	Xq28	152347797	152349404	64	-1	ENSG00000126890.4	NM_020994	300396
GAB3	Xq28	152374006	152446883	40	-1	ENSG00000160219.2	NM_080612	

External Gene ID	Band	Start Position (bp)	End Position (bp)	% GC content	Strand	Ensembl Gene ID	RefSeq ID	MIM ID
DKC1	Xq28	152458584	152473507	45	1	ENSG00000130826.2	NM_001363	300126
DKC1								305000
								300240
MPP1	Xq28	152474519	152501314	43	-1	ENSG00000130830.2	NM_002436	305360
NM_019863	Xq28	152532912	152557290	40	-1	ENSG00000185010.1	NM_019863	306700
Q8NBB5	Xq28	152677500	152706160	39	1	ENSG00000165775.4		
MTCP1	Xq28	152712216	152721853	40	-1	ENSG00000182712.3	NM_014221	300116
C61A_HUMAN	Xq28	152722117	152770739	37	1	ENSG00000185515.2	NM_024332	
VBP1	Xq28	152866864	152890410	40	1	ENSG00000155959.2	NM_003372	300133
RAB39B	Xq28	152909841	152916051	40	-1	ENSG00000155961.1	NM_171998	
CLIC2	Xq28	152928770	152986271	38	-1	ENSG00000155962.2	NM_001289	300138
Q8N7L3	Xq28	153009292	153009879	42	1	ENSG00000184853.1		
F8A	Xq28	153118977	153120092	73	-1	ENSG00000185990.1	NM_012151	305423
H2AFB	Xq28	153121072	153121419	64	-1	ENSG00000185978.1		300445
Q8N793	Xq28	153127635	153128048	47	1	ENSG00000185977.1		
TMLHE	Xq28	153151604	153274428	37	-1	ENSG00000185973.1	NM_018196	
SPRY3	Xq28	153429282	153443948	40	1	ENSG00000168939.1	NM_005840	602467
SYBL1	Xq28	153550961	153603634	38	1	ENSG00000124333.2	NM_005638	300053
IL9R	Xq28	153659077	153672313	55	1	ENSG00000124334.4	NM_002186	300007
							NM_176786	

https://etd.uwc.ac.za/

UNIVERSITY of the WESTERN CAPE

CHI-square test – Chapter 4

The chi critical value will be calculated according to the following formula:

$$X^2 = \sum \frac{(O - E)^2}{E}$$

O = Observed averaged intergenic lengths

E = Expected average intergenic lengths

$\sum$ = sum of

^{2} = square

Ho = There is no difference in the intergenic lengths of differentially expressed and ubiquitously expressed gene clusters on Xq28

H1= There is a difference in the intergenic lengths of differentially expressed and ubiquitously expressed genes.

Assumption = cluster size has no impact on the average intergenic length between differentially expressed and ubiquitously expressed gene clusters.

O = weighted average of all intergenic lengths for the different expression categories. This was used as there are an unequal number of genes for each expression category.

E = the overall average length of all interegenic regions regardless of cluster size or expression category.

UNIVERSITY of the
WESTERN CAPE

Expression	**Observed**	**Expected**	**(O - E)^2 / E**
Ubiquitous	30.89271429	57.066261	12.0045458
Differential	79.05204	57.066261	8.4704075

$$X^2 = \frac{12.0045458 + 8.4704075}{57.06626087}$$

$$X^2 = 0.35879262$$

Degrees of freedom is n -1. I have 2 categories hence 2 - 1 = 1 degree of freedom.

d.o.f	**P = 0.01**	**P =0.05**	**P = 0.01**
1	3.84	6.64	10.83

As the calculated value of chi-square is smaller then chi-critical values of all P, **H1** is rejected in favor of **Ho**.

https://etd.uwc.ac.za/

Reference List

1. Peto,J. Cancer epidemiology in the last century and the next decade. *Nature* **411**, 390-395 (2001).

2. Ponder,B.A. Cancer genetics. *Nature* **411**, 336-341 (2001).

3. McCarthy,M.I., Smedley,D. & Hide,W. New methods for finding disease-susceptibility genes: impact and potential. *Genome Biol.* **4**, 119 (2003).

4. Ward,E. *et al.* Cancer disparities by race/ethnicity and socioeconomic status. *CA Cancer J. Clin.* **54**, 78-93 (2004).

5. Jones,P.A. & Laird,P.W. Cancer epigenetics comes of age. *Nat. Genet.* **21**, 163-167 (1999).

6. Lander,E.S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921 (2001).

7. Venter,J.C. *et al.* The sequence of the human genome. *Science* **291**, 1304-1351 (2001).

8. Gibbs,R.A. *et al.* Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428**, 493-521 (2004).

9. Waterston,R.H. *et al.* Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520-562 (2002).

10. Bishop,J.M. Enemies within: the genesis of retrovirus oncogenes. *Cell* **23**, 5-6 (1981).

11. Cui,H., Horon,I.L., Ohlsson,R., Hamilton,S.R. & Feinberg,A.P. Loss of imprinting in normal tissue of colorectal cancer patients with microsatellite instability. *Nat. Med.* **4**, 1276-1280 (1998).

12. Evan,G. & Littlewood,T. A matter of life and cell death. *Science* **281**, 1317-1322 (1998).

13. Hoeijmakers,J.H. Genome maintenance mechanisms for preventing cancer. *Nature* **411**, 366-374 (2001).

14. Parada,L.F., Tabin,C.J., Shih,C. & Weinberg,R.A. Human EJ bladder carcinoma oncogene is homologue of Harvey sarcoma virus ras gene. *Nature* **297**, 474-478 (1982).

15. Perry,M.E. & Levine,A.J. Tumor-suppressor p53 and the cell cycle. *Curr. Opin. Genet. Dev.* **3**, 50-54 (1993).

16. Rainier,S. *et al.* Relaxation of imprinted genes in human cancer. *Nature* **362**, 747-749 (1993).

17. Vogelstein,B. & Kinzler,K.W. The multistep nature of cancer. *Trends Genet.* **9**, 138-

UNIVERSITY of the WESTERN CAPE

https://etd.uwc.ac.za/

141 (1993).

18. Wada,K., Maesawa,C., Akasaka,T. & Masuda,T. Aberrant expression of the maspin gene associated with epigenetic modification in melanoma cells. *J. Invest Dermatol.* **122**, 805-811 (2004).

19. Golub,T.R. *et al.* Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286**, 531-537 (1999).

20. Vogelstein,B. & Kinzler,K.W. p53 function and dysfunction. *Cell* **70**, 523-526 (1992).

21. Hanahan,D. & Weinberg,R.A. The hallmarks of cancer. *Cell* **100**, 57-70 (2000).

22. Liotta,L.A. & Kohn,E.C. The microenvironment of the tumour-host interface. *Nature* **411**, 375-379 (2001).

23. Emerson,B.M. Specificity of gene regulation. *Cell* **109**, 267-270 (2002).

24. Evan,G.I. & Vousden,K.H. Proliferation, cell cycle and apoptosis in cancer. *Nature* **411**, 342-348 (2001).

25. Collins,C. *et al.* Positional cloning of ZNF217 and NABC1: genes amplified at 20q13.2 and overexpressed in breast carcinoma. *Proc. Natl. Acad. Sci. U. S. A* **95**, 8703-8708 (1998).

26. Chen,R.Z., Pettersson,U., Beard,C., Jackson-Grusby,L. & Jaenisch,R. DNA hypomethylation leads to elevated mutation rates. *Nature* **395**, 89-93 (1998).

27. Rosenberg,S.A. Progress in human tumour immunology and immunotherapy. *Nature* **411**, 380-384 (2001).

28. Houshmand,P. & Zlotnik,A. Targeting tumor cells. *Curr. Opin. Cell Biol.* **15**, 640-644 (2003).

29. Sahin,U. *et al.* Human neoplasms elicit multiple specific immune responses in the autologous host. *Proc. Natl. Acad. Sci. U. S. A* **92**, 11810-11813 (1995).

30. Old,L.J. & Chen,Y.T. New paths in human cancer serology. *J. Exp. Med.* **187**, 1163-1167 (1998).

31. Chen,Y.T. *et al.* A testicular antigen aberrantly expressed in human cancers detected by autologous antibody screening. *Proc. Natl. Acad. Sci. U. S. A* **94**, 1914-1918 (1997).

32. Coulie,P.G. *et al.* A monoclonal cytolytic T-lymphocyte response observed in a melanoma patient vaccinated with a tumor-specific antigenic peptide encoded by gene MAGE-3. *Proc. Natl. Acad. Sci. U. S. A* **98**, 10290-10295 (2001).

33. Old,L.J. Cancer/testis (CT) antigens - a new link between gametogenesis and cancer. *Cancer Immun.* **1**, 1 (2001).

UNIVERSITY of the WESTERN CAPE

https://etd.uwc.ac.za/

34. Stockert,E. *et al.* A survey of the humoral immune response of cancer patients to a panel of human tumor antigens. *J. Exp. Med.* **187**, 1349-1354 (1998).

35. Traversari,C. *et al.* Transfection and expression of a gene coding for a human melanoma antigen recognized by autologous cytolytic T lymphocytes. *Immunogenetics* **35**, 145-152 (1992).

36. Van den,E.B. *et al.* A new family of genes coding for an antigen recognized by autologous cytolytic T lymphocytes on a human melanoma. *J. Exp. Med.* **182**, 689-698 (1995).

37. Scanlan,M.J., Simpson,A.J. & Old,L.J. The cancer/testis genes: review, standardization, and commentary. *Cancer Immun.* **4**, 1 (2004).

38. Scanlan,M.J. *et al.* Identification of cancer/testis genes by database mining and mRNA expression analysis. *Int. J. Cancer* **98**, 485-492 (2002).

39. van der,B.P. *et al.* A gene encoding an antigen recognized by cytolytic T lymphocytes on a human melanoma. *Science* **254**, 1643-1647 (1991).

40. Baba,T. *et al.* An acrosomal protein, sp32, in mammalian sperm is a binding protein specific for two proacrosins and an acrosin intermediate. *J. Biol. Chem.* **269**, 10133-10140 (1994).

41. Tureci,O. *et al.* Identification of a meiosis-specific protein as a member of the class of cancer/testis antigens. *Proc. Natl. Acad. Sci. U. S. A* **95**, 5211-5216 (1998).

42. De Backer,O. *et al.* Characterization of the GAGE genes that are expressed in various human cancers and in normal testis. *Cancer Res.* **59**, 3157-3165 (1999).

43. Boutanaev,A.M., Kalmykova,A.I., Shevelyov,Y.Y. & Nurminsky,D.I. Large clusters of co-expressed genes in the Drosophila genome. *Nature* **420**, 666-669 (2002).

44. Oliver,B., Parisi,M. & Clark,D. Gene expression neighborhoods. *J. Biol.* **1**, 4 (2002).

45. Orphanides,G. & Reinberg,D. A unified theory of gene expression. *Cell* **108**, 439-451 (2002).

46. Reik,W. & Walter,J. Genomic imprinting: parental influence on the genome. *Nat. Rev. Genet.* **2**, 21-32 (2001).

47. Spellman,P.T. & Rubin,G.M. Evidence for large domains of similarly expressed genes in the Drosophila genome. *J. Biol.* **1**, 5 (2002).

48. Jones,P.A. & Takai,D. The role of DNA methylation in mammalian epigenetics. *Science* **293**, 1068-1070 (2001).

49. Rideout,W.M., III, Eggan,K. & Jaenisch,R. Nuclear cloning and epigenetic reprogramming of the genome. *Science* **293**, 1093-1098 (2001).

50. Bird,A. DNA methylation patterns and epigenetic memory. *Genes Dev.* **16**, 6-21

UNIVERSITY of the WESTERN CAPE

https://etd.uwc.ac.za/

(2002).

51. Reinke,V. *et al.* A global profile of germline gene expression in C. elegans. *Mol. Cell* **6**, 605-616 (2000).

52. Jones,P.A. The DNA methylation paradox. *Trends Genet.* **15**, 34-37 (1999).

53. Jones,P.A. & Baylin,S.B. The fundamental role of epigenetic events in cancer. *Nat. Rev. Genet.* **3**, 415-428 (2002).

54. Cho,B. *et al.* Identification and characterization of a novel cancer/testis antigen gene CAGE. *Biochem. Biophys. Res. Commun.* **292**, 715-726 (2002).

55. De Smet,C. *et al.* The activation of human gene MAGE-1 in tumor cells is correlated with genome-wide demethylation. *Proc. Natl. Acad. Sci. U. S. A* **93**, 7149-7153 (1996).

56. De Smet,C., Lurquin,C., Lethe,B., Martelange,V. & Boon,T. DNA methylation is the primary silencing mechanism for a set of germ line- and tumor-specific genes with a CpG-rich promoter. *Mol. Cell Biol.* **19**, 7327-7335 (1999).

57. Pruitt,K.D., Katz,K.S., Sicotte,H. & Maglott,D.R. Introducing RefSeq and LocusLink: curated human genome resources at the NCBI. *Trends Genet.* **16**, 44-47 (2000).

58. Pruitt,K.D. & Maglott,D.R. RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res.* **29**, 137-140 (2001).

59. Birney,E. *et al.* An overview of ensembl. *Genome Res.* **14**, 925-928 (2004).

60. Birney,E. *et al.* Ensembl 2004. *Nucleic Acids Res.* **32 Database issue**, D468-D470 (2004).

61. Hubbard,T. *et al.* The Ensembl genome database project. *Nucleic Acids Res.* **30**, 38-41 (2002).

62. Potter,S.C. *et al.* The ensembl analysis pipeline. *Genome Res.* **14**, 934-941 (2004).

63. Rosen,N. *et al.* GeneLoc: exon-based integration of human genome maps. *Bioinformatics.* **19 Suppl 1**, i222-i224 (2003).

64. Wain,H.M. *et al.* Guidelines for human gene nomenclature. *Genomics* **79**, 464-470 (2002).

65. Aradhya,S. *et al.* Multiple pathogenic and benign genomic rearrangements occur at a 35 kb duplication involving the NEMO and LAGE2 genes. *Hum. Mol. Genet.* **10**, 2557-2567 (2001).

66. Chomez,P. *et al.* An overview of the MAGE gene family with the identification of all human members of the family. *Cancer Res.* **61**, 5544-5551 (2001).

UNIVERSITY of the
WESTERN CAPE

https://etd.uwc.ac.za/

67. De Plaen,E. *et al.* Structure, chromosomal localization, and expression of 12 genes of the MAGE family. *Immunogenetics* **40**, 360-369 (1994).

68. Kouprina,N. *et al.* The SPANX gene family of cancer/testis-specific antigens: rapid evolution and amplification in African great apes and hominids. *Proc. Natl. Acad. Sci. U. S. A* **101**, 3077-3082 (2004).

69. Rogner,U.C., Wilke,K., Steck,E., Korn,B. & Poustka,A. The melanoma antigen gene (MAGE) family is clustered in the chromosomal band Xq28. *Genomics* **29**, 725-731 (1995).

70. Ruault,M. *et al.* BAGE genes generated by juxtacentromeric reshuffling in the Hominidae lineage are under selective pressure. *Genomics* **81**, 391-399 (2003).

71. Brown,C.J. & Greally,J.M. A stain upon the silence: genes escaping X inactivation. *Trends Genet.* **19**, 432-438 (2003).

72. Emerson,J.J., Kaessmann,H., Betran,E. & Long,M. Extensive gene traffic on the mammalian X chromosome. *Science* **303**, 537-540 (2004).

73. Lahn,B.T. & Page,D.C. Four evolutionary strata on the human X chromosome. *Science* **286**, 964-967 (1999).

74. Lahn,B.T., Pearson,N.M. & Jegalian,K. The human Y chromosome, in the light of evolution. *Nat. Rev. Genet.* **2**, 207-216 (2001).

75. Nanda,I. *et al.* 300 million years of conserved synteny between chicken Z and human chromosome 9. *Nat. Genet.* **21**, 258-259 (1999).

76. Jegalian,K. & Page,D.C. A proposed path by which genes common to mammalian X and Y chromosomes evolve to become X inactivated. *Nature* **394**, 776-780 (1998).

77. Rice,W.R. Degeneration of a nonrecombining chromosome. *Science* **263**, 230-232 (1994).

78. Meller,V.H. Dosage compensation: making 1X equal 2X. *Trends Cell Biol.* **10**, 54-59 (2000).

79. Charchar,F.J. *et al.* Complex events in the evolution of the human pseudoautosomal region 2 (PAR2). *Genome Res.* **13**, 281-286 (2003).

80. Ciccodicola,A. *et al.* Differentially regulated and evolved genes in the fully sequenced Xq/Yq pseudoautosomal region. *Hum. Mol. Genet.* **9**, 395-401 (2000).

81. D'Esposito,M. *et al.* Differential expression pattern of XqPAR-linked genes SYBL1 and IL9R correlates with the structure and evolution of the region. *Hum. Mol. Genet.* **6**, 1917-1923 (1997).

82. Skaletsky,H. *et al.* The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature* **423**, 825-837 (2003).

UNIVERSITY of the WESTERN CAPE

https://etd.uwc.ac.za/

83. Lahn,B.T. & Page,D.C. Retroposition of autosomal mRNA yielded testis-specific gene family on human Y chromosome. *Nat. Genet.* **21**, 429-433 (1999).

84. Khil,P.P., Smirnova,N.A., Romanienko,P.J. & Camerini-Otero,R.D. The mouse X chromosome is enriched for sex-biased genes not subject to selection by meiotic sex chromosome inactivation. *Nat. Genet.* **36**, 642-646 (2004).

85. Parisi,M. *et al.* Paucity of genes on the Drosophila X chromosome showing male-biased expression. *Science* **299**, 697-700 (2003).

86. Saifi,G.M. & Chandra,H.S. An apparent excess of sex- and reproduction-related genes on the human X chromosome. *Proc. R. Soc. Lond B Biol. Sci.* **266**, 203-209 (1999).

87. Rice,W.R. Sexually antagonistic genes: experimental evidence. *Science* **256**, 1436-1439 (1992).

88. Brooks,R. Negative genetic correlation between male sexual attractiveness and survival. *Nature* **406**, 67-70 (2000).

89. Goto,T. & Monk,M. Regulation of X-chromosome inactivation in development in mice and humans. *Microbiol. Mol. Biol. Rev.* **62**, 362-378 (1998).

90. Swanson,W.J. Adaptive evolution of genes and gene families. *Curr. Opin. Genet. Dev.* **13**, 617-622 (2003).

91. Clark,A.G., Aguade,M., Prout,T., Harshman,L.G. & Langley,C.H. Variation in sperm displacement and its association with accessory gland protein loci in Drosophila melanogaster. *Genetics* **139**, 189-201 (1995).

92. Snook,R.R. & Hosken,D.J. Sperm death and dumping in Drosophila. *Nature* **428**, 939-941 (2004).

93. Swanson,W.J., Clark,A.G., Waldrip-Dail,H.M., Wolfner,M.F. & Aquadro,C.F. Evolutionary EST analysis identifies rapidly evolving male reproductive proteins in Drosophila. *Proc. Natl. Acad. Sci. U. S. A* **98**, 7375-7379 (2001).

94. Hurst,L.D. & Randerson,J.P. An eXceptional chromosome. *Trends Genet.* **15**, 383-385 (1999).

95. Hurst,L.D. Evolutionary genomics. Sex and the X. *Nature* **411**, 149-150 (2001).

96. Caron,H. *et al.* The human transcriptome map: clustering of highly expressed genes in chromosomal domains. *Science* **291**, 1289-1292 (2001).

97. Gibson,J.R., Chippindale,A.K. & Rice,W.R. The X chromosome is a hot spot for sexually antagonistic fitness variation. *Proc. R. Soc. Lond B Biol. Sci.* **269**, 499-505 (2002).

98. Lercher,M.J., Urrutia,A.O. & Hurst,L.D. Evidence that the human X chromosome is enriched for male-specific but not female-specific genes. *Mol. Biol. Evol.* **20**, 1113-

UNIVERSITY of the WESTERN CAPE

https://etd.uwc.ac.za/

1116 (2003).

99. Wang,P.J., McCarrey,J.R., Yang,F. & Page,D.C. An abundance of X-linked genes expressed in spermatogonia. *Nat. Genet.* **27**, 422-426 (2001).

100. Morison,I.M., Paton,C.J. & Cleverley,S.D. The imprinted gene and parent-of-origin effect database. *Nucl. Acids. Res.* **29**, 275-276 (2001).

101. Reinke,V. Sex and the genome. *Nat. Genet.* **36**, 548-549 (2004).

102. Charlesworth,D. & Charlesworth,B. Sex differences in fitness and selection for centric fusions between sex-chromosomes and autosomes. *Genet. Res.* **35**, 205-214 (1980).

103. Kelso,J. *et al.* eVOC: a controlled vocabulary for unifying gene expression data. *Genome Res.* **13**, 1222-1230 (2003).

104. Davis,I.D. *et al.* Recombinant NY-ESO-1 protein with ISCOMATRIX adjuvant induces broad integrated antibody and CD4(+) and CD8(+) T cell responses in humans. *Proc. Natl. Acad. Sci. U. S. A* **101**, 10697-10702 (2004).

105. Chen,Q. *et al.* Immunodominant CD4+ responses identified in a patient vaccinated with full-length NY-ESO-1 formulated with ISCOMATRIX adjuvant. *Proc. Natl. Acad. Sci. U. S. A* **101**, 9363-9368 (2004).

106. Kosak,S.T. & Groudine,M. Gene order and dynamic domains. *Science* **306**, 644-647 (2004).

107. Osborne,C.S. *et al.* Active genes dynamically colocalize to shared sites of ongoing transcription. *Nat. Genet.* **36**, 1065-1071 (2004).

108. Lercher,M.J., Urrutia,A.O. & Hurst,L.D. Clustering of housekeeping genes provides a unified model of gene order in the human genome. *Nat. Genet.* **31**, 180-183 (2002).

109. Sun,M. *et al.* SAGE is far more sensitive than EST for detecting low-abundance transcripts. *BMC. Genomics* **5**, 1 (2004).

110. Bishop,D.F., Henderson,A.S. & Astrin,K.H. Human delta-aminolevulinate synthase: assignment of the housekeeping gene to 3p21 and the erythroid-specific gene to the X chromosome. *Genomics* **7**, 207-214 (1990).

111. Versteeg,R. *et al.* The human transcriptome map reveals extremes in gene density, intron length, GC content, and repeat pattern for domains of highly and weakly expressed genes. *Genome Res.* **13**, 1998-2004 (2003).

112. Yamashita,T. *et al.* Genome-wide transcriptome mapping analysis identifies organ-specific gene expression patterns along human chromosomes. *Genomics* **84**, 867-875 (2004).

113. Lercher,M.J., Urrutia,A.O., Pavlicek,A. & Hurst,L.D. A unification of mosaic structures in the human genome. *Hum. Mol. Genet.* **12**, 2411-2415 (2003).

UNIVERSITY of the
WESTERN CAPE

https://etd.uwc.ac.za/

114. Bortoluzzi,S. *et al.* A comprehensive, high-resolution genomic transcript map of human skeletal muscle. *Genome Res.* **8**, 817-825 (1998).

115. Dempsey,A.A., Pabalan,N., Tang,H.C. & Liew,C.C. Organization of human cardiovascular-expressed genes on chromosomes 21 and 22. *J. Mol. Cell Cardiol.* **33**, 587-591 (2001).

116. Kasprzyk,A. *et al.* EnsMart: a generic system for fast and flexible access to biological data. *Genome Res.* **14**, 160-169 (2004).

117. Hamosh,A. *et al.* Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* **30**, 52-55 (2002).

118. Boeckmann,B. *et al.* The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* **31**, 365-370 (2003).

119. Lenhard,B., Hayes,W.S. & Wasserman,W.W. GeneLynx: a gene-centric portal to the human genome. *Genome Res.* **11**, 2151-2157 (2001).

120. Creating the gene ontology resource: design and implementation. *Genome Res.* **11**, 1425-1433 (2001).

121. Megy,K., Audic,S. & Claverie,J.M. Positional clustering of differentially expressed genes on human chromosomes 20, 21 and 22. *Genome Biol.* **4**, 1 (2003).

122. Stolc,V. *et al.* A gene expression map for the euchromatic genome of Drosophila melanogaster. *Science* **306**, 655-660 (2004).

123. Margulies,E.H., Kardia,S.L. & Innis,J.W. Identification and prevention of a GC content bias in SAGE libraries. *Nucleic Acids Res.* **29**, E60 (2001).

124. Sutherland,G.R. & Baker,E. Characterisation of a new rare fragile site easily confused with the fragile X. *Hum. Mol. Genet.* **1**, 111-113 (1992).

125. Lakich,D., Kazazian,H.H., Jr., Antonarakis,S.E. & Gitschier,J. Inversions disrupting the factor VIII gene are a common cause of severe haemophilia A. *Nat. Genet.* **5**, 236-241 (1993).

126. Laporte,J. *et al.* Cloning and characterization of an alternatively spliced gene in proximal Xq28 deleted in two patients with intersexual genitalia and myotubular myopathy. *Genomics* **41**, 458-462 (1997).

127. Kmita,M. & Duboule,D. Organizing axes in time and space; 25 years of colinear tinkering. *Science* **301**, 331-333 (2003).

128. Patel,N.H. & Prince,V.E. Beyond the Hox complex. *Genome Biol.* **1**, REVIEWS1027 (2000).

129. Patel,N.H. Evolutionary biology: time, space and genomes. *Nature* **431**, 28-29 (2004).

UNIVERSITY of the WESTERN CAPE

https://etd.uwc.ac.za/

130. Lucas,S., Brasseur,F. & Boon,T. A new MAGE gene with ubiquitous expression does not code for known MAGE antigens recognized by T cells. *Cancer Res.* **59**, 4100-4103 (1999).

131. D'Onofrio,G. Expression patterns and gene distribution in the human genome. *Gene* **300**, 155-160 (2002).

132. Vinogradov,A.E. Isochores and tissue-specificity. *Nucleic Acids Res.* **31**, 5212-5220 (2003).

133. Bernal,A., Ear,U. & Kyrpides,N. Genomes OnLine Database (GOLD): a monitor of genome projects world-wide. *Nucleic Acids Res.* **29**, 126-127 (2001).

134. Hedges,S.B. The origin and evolution of model organisms. *Nat. Rev. Genet.* **3**, 838-849 (2002).

135. Hedges,S.B. & Kumar,S. Genomics. Vertebrate genomes compared. *Science* **297**, 1283-1285 (2002).

136. Taylor,M.S. & Semple,C.A. Sushi gets serious: the draft genome sequence of the pufferfish Fugu rubripes. *Genome Biol.* **3**, reviews1025 (2002).

137. Ureta-Vidal,A., Ettwiller,L. & Birney,E. Comparative genomics: genome-wide analysis in metazoan eukaryotes. *Nat. Rev. Genet.* **4**, 251-262 (2003).

138. Cooper,G.M. & Sidow,A. Genomic regulatory regions: insights from comparative sequence analysis. *Curr. Opin. Genet. Dev.* **13**, 604-610 (2003).

139. Frazer,K.A., Elnitski,L., Church,D.M., Dubchak,I. & Hardison,R.C. Cross-species sequence comparisons: a review of methods and available resources. *Genome Res.* **13**, 1-12 (2003).

140. Fitch,W.M. Homology a personal view on some of the problems. *Trends Genet.* **16**, 227-231 (2000).

141. Thomas,J.W. & Touchman,J.W. Vertebrate genome sequencing: building a backbone for comparative genomics. *Trends Genet.* **18**, 104-108 (2002).

142. Moreau-Aubry,A. *et al.* A processed pseudogene codes for a new antigen recognized by a CD8(+) T cell clone on melanoma. *J. Exp. Med.* **191**, 1617-1624 (2000).

143. Clamp,M. *et al.* Ensembl 2002: accommodating comparative genomics. *Nucleic Acids Res.* **31**, 38-42 (2003).

144. Wheeler,D.L. *et al.* Database resources of the National Center for Biotechnology. *Nucleic Acids Res.* **31**, 28-33 (2003).

145. Hammond,M.P. & Birney,E. Genome information resources - developments at Ensembl. *Trends Genet.* **20**, 268-272 (2004).

146. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931-945

UNIVERSITY of the WESTERN CAPE

https://etd.uwc.ac.za/

(2004).

147. Samonte,R.V. & Eichler,E.E. Segmental duplications and the evolution of the primate genome. *Nat. Rev. Genet.* **3**, 65-72 (2002).

148. Marques-Bonet,T. *et al.* Chromosomal rearrangements and the genomic distribution of gene-expression divergence in humans and chimpanzees. *Trends Genet.* **20**, 524-529 (2004).

149. Stankiewicz,P., Shaw,C.J., Withers,M., Inoue,K. & Lupski,J.R. Serial segmental duplications during primate evolution result in complex human genome architecture. *Genome Res.* **14**, 2209-2220 (2004).

150. Nei,M., Zhang,J. & Yokoyama,S. Color vision of ancestral organisms of higher primates. *Mol. Biol. Evol.* **14**, 611-618 (1997).

151. Warburton,P.E., Giordano,J., Cheung,F., Gelfand,Y. & Benson,G. Inverted repeat structure of the human genome: the X-chromosome contains a preponderance of large, highly homologous inverted repeats that contain testes genes. *Genome Res.* **14**, 1861-1869 (2004).

UNIVERSITY of the
WESTERN CAPE

https://etd.uwc.ac.za/