Characterization of TCRβ sequence diversity in colorectal carcinoma

by

Lisa Ann Raeburn

B.Sc. (Hons; Biology), Simon Fraser University, 2007

Thesis Submitted in Partial Fulfillment
of the Requirements for the Degree of
Master of Science

in the
Department of Molecular Biology and Biochemistry
Faculty of Science

SIMON FRASER UNIVERSITY

Summer 2012

Approval

Name: **Lisa Ann Raeburn**

Degree: **Master of Science**

Title of Thesis: ***Characterization of TCRβ sequence diversity in colorectal carcinoma***

Examining Committee:

Chair: Dr. Rosemary Cornell, Professor

Dr. Robert Holt
Senior Supervisor
Associate Professor

Dr. John Webb
Supervisor
Associate Professor (Biochemistry and Microbiology, University of Victoria)

Dr. Jack Chen
Supervisor
Associate Professor

Dr. Jonathan Choy
Internal Examiner
Assistant Professor

Date Defended: June 13, 2012

Partial Copyright Licence

Abstract

T cells can play a critical role in mediating anti-tumour responses in cancer, but thus far have been characterized at low resolution. Each T cell receptor beta subunit (TCRβ) possesses a hypervariable sequence (CDR3) that is a principal site of contact with peptide-MHC complexes on heterologous cells and can be used to characterize TCRβ diversity. Here, I used a sequencing approach developed for interrogating peripheral blood immune repertoires to explore TCRβ sequence diversity associated with colorectal carcinoma (CRC). TCRβ sequences were amplified from RNA isolated from biopsies of tumour and matched control tissue of 43 CRC patients. Amplicons were indexed and survey sequenced on the Illumina platform. Sequence reads were assembled and filtered revealing a diverse sequence repertoire. Some abundant sequences shared between patients were significantly associated with improved overall survival. These results have great potential utility in the design of immunological approaches targeting tumours, and in the design of specific screening approaches for CRC.

Keywords: CDR3; cancer; sequencing; survival; T cell receptor

Dedication

This thesis is dedicated to all of those people who helped me over the steep slopes I encountered over the past few years.

Acknowledgements

First and foremost, I would like to thank my senior supervisor, Dr. Rob Holt, for the privilege to be part of such a great lab and such interesting research. The opportunity has and will continue to shape my life. I would also like to thank my committee, Dr. John Webb and Dr. Jack Chen for all of their advice, and to Dr. Jonathan Choy for acting my public examiner.

I was truly fortunate to have met some wonderful people here at the GSC. I would like to extend an enormous thank you to René Warren and Doug (Yoda) Freeman, both of whom are great mentors. Thank you for your patience and guidance. Without you, this project would not have been possible. I will never forget the endless encouragement from Mauro Castellarin. His words of support truly made the difference. I want to thank Lisa Dreolini for the help and the laughs. Chandra Lebovitz was my cheerleader, never letting me give up and making me stay on task. Thank you to Jillian McKenna whose sense of humour and assistance kept me going. More thanks are for everyone in the Holt Lab, past, present and adopted, and everyone at the GSC and SFU.

To my dear family and friends, thank you for always being there for me and encouraging me to push through. This accomplishment is one that would not have been possible without you.

"I can no other answer make, but thanks, and thanks, and ever thanks."

(Act 3, Scene 3 of the Twelfth Night by William Shakespeare, 1601-1602)

Table of Contents

List of Tables

List of Figures

List of Acronyms

aaCDR3β	CDR3β amino acid sequence
APC	Antigen presenting cell
CDR	Complementary determining region
CIH	Cancer immunoediting hypothesis
CRC	Colorectal carcinoma
CTL	Cytotoxic T cell
DN	Double negative (T cell)
DOC	Depth of coverage
DP	Double positive (T cell)
HLA	Human leukocyte antigen
IET	Intraepithelial T cell
IHC	Immunohistochemistry
GALT	Gut-associated lymphoid tissue
LOD	Limit of detection
MHC	Major histocompatibility complex
MIS	Mucosal immune system
NGS	Next generation sequencing
NK cell	Natural killer cell
NSIS	Non-specific immune stimulation
ntCDR3β	CDR3β nucleotide sequence
PBL	Peripheral blood lymphocyte
pMHC	Peptide antigen presented by a major histcompatibility complex molecule
preTCRα	Pre T cell receptor α chain
preTCRαβ	Pre αβ T cell receptor
qPCR	Quantitative PCR
RACE	Rapid amplification of cDNA ends
RSS	Recombination signal sequence
RTE	Recent thymic emigrant
SP	Single positive (T cell)
TA	Tumour antigen

TCEC	Thymic cortical epithelial cell
TCR	T cell receptor
TIfT (TIL)	Tumour infiltrating T cell (tumour infiltrating lymphocyte)
TRAV	T cell receptor α chain variable gene
TRBC	T cell receptor β chain constant gene
TRBD	T cell receptor β chain diversity gene
TRBJ	T cell receptor β chain joining gene
TRBV	T cell receptor β chain variable gene
Th	Helper T cell
Treg	Regulatory T cell

Glossary

CDR3β sequence	Sequence encoding the hypervariable T cell receptor CDR3β, the site of peptide antigen contact. CDR3β spans the TRBV(D)J gene junctions and is traditionally defined to be between the last conserved cysteine (C) codon of the TRBV-gene and the first conserved phenylalanine (F) codon of the TRBJ gene in the conserved motif FGXG.
Colorectal carcinoma	Cancer that forms in the tissues of the colon (the longest part of the large intestine) and rectum (distal part of the long intestine) (http://www.cancer.gov/cancertopics/types/colon-and-rectal)
Depth of coverage (DOC)	The number of observations of a given TCRβ /CDR3 sequence in a data set.
Immunogenicity	The ability to induce an adaptive immune response.
Phred score	Score given to assess likelihood that base called at a position is the correct base. Phred reads DNA sequence files and analyzes the peaks to call bases, assigning quality scores. Higher Phred values correspond to higher quality bases. The quality scores are logarithmically linked to error probabilities (e.g. Q30 = 99.9 % accurate that base call is correct).
T cell	Lymphocytes derived from progenitors in the bone-marrow that have migrated to the thymus for maturation
TCRβ sequence	The sequences encoding the CDR3β, partial TRBV-gene and full TRBJ-gene, taking into account bases added and deleted at gene junctions.
TCR Spectratyping	Method for TCR profiling involving amplifying target TCRβ sequences with 5′ TRBV-gene specific primers and a common 3′ TRBC gene-specific primer. These amplification products are used as templates for a primer extension reaction using an additional, fluorescently labeled 3′ TRBC-gene primer. The fluorescent products are end-labeled and extended through the hypervariable CDR3 region, and run on polyacrylamide sequencing gels to reveal precise sizes of the CDR3. With the aid of an automated sequencing machine and appropriate software, the size of the peaks corresponding to discrete CDR3 lengths can be analyzed.

1. General Introduction

1.1. The adaptive immune system

The human immune system maintains homeostasis by identifying and removing foreign or mutated agents in the body. The two major arms of the immune system – innate immunity, and adaptive immunity – rely on the activities of leukocytes for this surveillance. While the leukocytes of innate immunity are primarily involved with providing immediate recognition and defence, it is the cells of the adaptive immune system that offer long-lasting protection to the host[1]. The adaptive immune system is comprised predominantly of a highly specialised group of leukocytes, called lymphocytes. These lymphocytes, B cells and T cells, are derived from a common lymphoid progenitor cell in the bone marrow. While B cells remain in the bone marrow to mature, T cells leave the bone marrow and travel to the thymus to mature[1].

1.1.1. Thymic maturation

Thymic maturation is a pivotal process that shapes the repertoire of naïve T cells that end up in the periphery. Steps that occur in the thymus to generate this repertoire are briefly discussed in the sections below:

1.1.1.1. Stages of T cell development

Thymocytes undergo a series of distinct phases during their maturation that are marked by changes in cell surface molecules[1]. Upon entering the thymus, immature thymocytes do not express any typical T cell cell-surface molecules (e.g. CD3, CD4, CD8) and their receptor genes are not yet rearranged. These cells are referred to as double negative (DN) thymocytes[1]. These thymocytes interact with thymic stroma and trigger an initial phase of differentiation along the T cell lineage pathway, followed by proliferation, differentiation and the expression of T cell specific cell surface molecules[1]. There are four stages of development for DN cells, called DN1-DN4; in DN1, thymocytes

express molecules CD44, Notch and Kit[1]. As cells undergo development into DN2 thymocytes, they express CD25 (α chain of the IL-2 receptor)[1]. After this, DN2 cells begin to rearrange their T cell receptor β (TCRβ) locus, lowering their expression of CD44 and Kit to become DN3 thymocytes. DN3 cells also begin to express CD3. DN3 thymocytes are arrested in this stage (CD25, $CD44^{low}$) until they have successfully rearranged the gene segments making up the TCRβ locus (rearrangement discussed in section 1.1.1.2)[1]. Once a productive TCRβ chain has been expressed on the cell surface in conjunction with a pre- T cell receptor α chain (preTCRα) (discussed in section 1.1.1.2), the T cell rapidly proliferates, resulting in loss of expression of CD25 and CD44[1] cell surface markers. These thymocytes are now in the DN4 stage. Once DN4 thymocytes cease proliferation, they express CD4 and CD8 co-receptors and are referred to as double positive (DP) thymocytes[1]. It is in this DP stage that the TCRα locus rearranges and pairs with the TCRβ on the T cell surface (discussed in section 1.1.1.2) [1]. The DP cells that successfully survive selection events (discussed in section 1.1.1.2) lose expression of one of the CD4 or CD8 co-receptors and migrate out of the thymus as single positive (SP) cells[1].

1.1.1.2. Generation of the TCRβ chain

T cells express highly specialised, heterodimeric, membrane-bound T cell receptors (TCRs) on their surfaces that are responsible for surveying antigen ligands that T cells could encounter in their lifetimes[1]. Human TCRs are primarily composed of two transmembrane glycoprotein chains, an alpha (TCRα) and a beta (TCRβ) chain, that are linked by a disulphide bond (TCRαβ)[1] (Fig. 1a). The extracellular portion of each chain contains 2 regions: a variable (V)-region where the antigen binging site is located and a constant (C)-region that anchors the TCRαβ to the cell membrane. Although not the focus of this thesis, there is additionally a minority of less well-described T cells in humans that bear an alternative, but structurally similar heterodimeric receptor composed of a γ and a δ chain (TCRγδ)[1].

The TCRβ locus is located on chromosome 7 at position 7q34 and consists of two diversity gene segments (TRBD), 13 joining gene segments (TRBJ) and > 50 variable gene segments (TRBV)[2] (Fig. 1b). During the DN2-4 stages of development, an enzymatic complex called the V(D)J-recombinase mediates the rearrangement of these

germline gene segments to produce a functional TCRβ chain[1]. Briefly, enzymes in the complex (notably RAG1/2 enzymes) bring together a single TRBD gene segment with a single TRBJ gene segment at special recognition signal sequences (RSSs) flanking each gene segment. DNA repair enzymes of the complex can then remove germline coded nucleotides at the gene junctions, while at the same time enzyme terminal deoxynucleotidyl transferase (TdT) can add non-templated bases at these junctions[1]. The enzyme DNA ligase IV in the complex will then join the modified TRBD and TRBJ gene segments together. Next, a single TRBV segment is joined the TRBD/J by the RAG enzymes, where the addition and deletion of nucleotides occurs again at the TRBV-TRBD gene junction and ligation of the TRBV to the TRBD/J occurs via same processes discussed above[1]. As a final step in the generation of a complete TCRβ chain, the functional TRBV(D)J V exon is transcribed and spliced to join one of two TCR constant (TRBC) genes[1]. The resulting mRNA is translated, yielding an extremely variable TCRβ chain (Fig. 1b). In the DP stage, functional TCRα chains are formed in the same way as the TCRβ, with the exception that the TCRα chain does not have a diversity gene segment. The pairing of TCRα to a TCRβ chain in DP T cells further diversifies the TCRαβ repertoire.

The mechanisms of somatic recombination, addition and deletion of nucleotides at gene junctions and TCRα – TCRβ chain pairing, result in an enormous repertoire of TCRαβs. This repertoire has theoretically been estimated to be composed of 10^{15} distinct TCRαβs[3]. Actual diversity however is restricted by the events of thymic selection (discussed in next section, 1.1.1.3). The number of distinct TCRβ chains in the blood has been directly estimated to be at least 10^6 [4,5]. Given that each of these 10^6 possible TCRβ chains is expected to pair with ~25 different TCRα chains in peripheral blood[4], TCRαβ diversity could theoretically be composed of at least 25 x10^6 different TCRαβs (the upper limit of which is set by the number of TCRα chains that each of the TCRβ chains actually pairs with)[4].

The majority of the sequence diversity of the TCRβ is concentrated in the third complementary determining region (CDR3β) of the V-region, which encompasses the TRBV(D)J recombination junctions and is a primary site of antigen contact[3]. The CDR3β is defined as the region between the last conserved cysteine (C) of the TRBV-gene segment and the first conserved phenylalanine (F) of the TRBJ gene segment in the

conserved motif FGXG[5,6] (Fig. 1b). There are six CDRs in total, belonging to both the TCRα and the TCRβ chains of the receptor (CDR1α, 2α, 3α and CDR1β, 2β, 3β) (Fig. 1a). The CDR3 of the TCRα and the TCRβ come into primary contact with the bound peptide antigen, while the CDR1 and CDR2 of both chains (coded in the TRBV/TRAV gene segments) contact the MHC molecule presenting the peptide[1] (Fig. 1a). Due to allelic exclusion (discussed in section below, 1.1.1.3), a single TCRβ is typically expressed on the surface of a T cell, making CDR3β sequence diversity a good measure of the total T cell diversity in a system[5].

1.1.1.3. Allelic exclusion and selection

Once TCRβ gene rearrangement occurs, TCRβ is expressed on the surface of a DN3 T cell with the invariant preTCRα chain (now called a preTCRαβ) and a CD3 molecule complex[1]. Expression of the pre-TCRαβ and CD3 triggers the phosphorylation and degradation of RAG-2 enzyme, halting TCRβ-chain gene rearrangement, and thus resulting in allelic exclusion (i.e. expression of a single TCRβ variant) at the TCRβ locus[1]. This process induces rapid T cell proliferation and loss of CD25 expression. After the proliferative phase ends, CD4 and CD8 co-receptors are expressed on T cell surfaces[1]. RAG1/2 are then transcribed again and rearrangement of the TCRα chain commences ensuring that a single functional TCRβ can be associated with many TCRα chains[1]. While TCRα rearrangement is occurring, TCRαβ receptors are expressed on DP T cell surfaces and undergo positive selection. Rearrangement of the TCRα will continue until signalling by a self-peptide:self-MHC on a thymic cortical epithelial cell (TCEC) successfully positively selects the TCRαβ variant[1]. This means however that many T cells can have functional rearrangements on both chromosomes and can thus produce two types of TCRα chains (and therefore two functional TCRαβs with differing antigenic specificities) on the T cell surface[1]. It has been estimated that up to one third of mature T cells express two TCRα chains on their surfaces[7]. Therefore, unlike what is typical for TCRβ chains, TCRα chains do not display allelic exclusion[1]. Dual specificity T cells (i.e. T cells that express TCRαβs of differing specificities on their surfaces), can be the result of multi-TCRα chains being expressed on the surface, but have also been attributed to (albeit less commonly) multi-TCRβ chain expression[1,8].

The ligands for TCRs are endogenous and exogenous protein peptide antigens that are routinely broken down and presented on the surfaces of antigen presenting cells (APCs) by specialised glycoproteins called Human Leukocyte Antigens (HLA)[1] (Fig. 1a). HLA molecules are encoded by a large cluster of genes called the Major Histocompatibility Complex (MHC), and are commonly referred to as MHC molecules. If a TCRαβ is able to recognise a self-peptide:self-MHC class 1 complex, it will receive survival and maturation signals and maintain the expression of the CD8 co-receptor and is now referred to as a SP $CD8^+$ T cell[1]. If the T cell is able to recognise a self-peptide:self-MHC class 2 complex, it will receive survival and maturation signals and maintain the expression of CD4 co-receptor and is now referred to as a SP $CD4^+$ T cell[1]. T cells that react too strongly with self-peptide:self-MHC complexes presented primarily by bone marrow derived dendritic cells and macrophages in the thymic cortex or medulla, are deleted. In this way, T cells capable of responding to self -peptide antigens in the periphery are eliminated, in a process known as negative selection[1]. T cells expressing TCRαβs that are unable to recognise either MHC class I or MHC class II molecules fail to receive any survival signals, and die by programmed cell death or apoptosis in the thymus[1]. It is estimated that 2% of the immature T cells arriving at the thymus will survive positive and negative selection and leave the thymus as mature MHC-restricted T cells[1].

1.1.2. TCR activation

T cells that survive thymic selection emerge in the periphery as mature, naive recent thymic emigrants (RTEs)[1]. Naïve RTEs can now migrate throughout the body and survey pMHC complexes on the surfaces of APCs. If a TCR comes into contact with its cognate antigen, and the appropriate co-stimulatory signals are present (discussed below), the T cell can become activated, differentiated, undergo clonal expansion into an effector T cell. The army of effector T cell clones resulting from this activation have identical antigenic specificity to the original T cell. Multiple signals are involved in T cell activation: (1) TCR engagement with a pMHC and CD4 or CD8 interaction with the MHC transmits a signal to the T cell that antigen has been encountered[1]. (2) Effective activation also requires the co-stimulatory molecule CD28 expressed on T cell surfaces to engage with the B7 molecule expressed by activated APCs[1]. This signal ensures survival and proliferation of the T cell. TCR-pMHC engagement without CD28 activation

can lead T cell anergy, whereby T cells are inactivated, but remain alive for an extended period of time in a hyporesponsive state[9]. (3) Finally, cytokine stimulation by the APC can direct T cell differentiation into different effector subsets of T cells[1]. Activation of $CD8^+$ T cell results in a clonal expansion of effector cytotoxic "killer" T cells (CTLs) that can release cytotoxic granules (e.g. perforin, granzyme proteases) onto the surface of peptide presenting cells, inducing apoptosis[1]. Activation of a $CD4^+$ T cell results in a clonal expansion of effector "helper" T cells (e.g. Th1, Th2, Th17) that are involved in promotion of T cell and B cell immune responses[1]. Briefly, $CD4^+$ Th1 T cells can activate macrophages and B cells to produce antibodies. $CD4^+$ Th2 T cells produce cytokines that activate and drive B cells to differentiate and produce different types of antibodies. $CD4^+$ Th17 T cells induce cells to recruit innate immune system cells, neutrophils, to sites of infection. Activated $CD4^+/CD8^+$ T cells can also function as regulators of immune responses, suppressing immune responses, and are called T regulatory cells (Tregs)[1].

1.2. T cell gut mucosal immunity

1.2.1. T cell population in the gut

As one of the largest mucosal surfaces, the gastrointestinal (GI) tract encounters more antigen than any other part of the mucosal immune system (MIS)[10], and therefore not surprisingly, contains one of the largest populations of T cells in the body[1]. The intraepithelial T cell (IET) population of the colon/rectum is composed mostly of αβ T cells (65% αβ T cells in murine models), with the remainder being γδ T cells[11]. Further, intestinal T cells are unlike other T cells in the body, in that the majority of them have an antigen experienced phenotype, commonly expressing CD45RO [10].

1.2.1.1. T cell arrival in the gut

Mature naïve RTEs travel through the blood and can enter organized gut associated lymphoid tissues (GALT) (e.g. Peyers patches, isolated lymphoid follicles) and/or lamina propria, where they can interact with pMHC presented by APCs (e.g. dendritic cells (DCs)[1]. T cells that encounter their cognate antigen and become activated, acquire the expression of gut homing molecules, such as $\alpha_4\beta_7$ integrin and the chemokine receptor CCR9[1]. They then leave organized lymphoid tissues via the afferent

lymphatics and travel to the mesenteric lymph node (MLN), where they differentiate, and later migrate into the blood stream through the thoracic duct[10]. These antigen-experienced effector T cells can re-accumulate in the mucosa as a result of the interaction of their expressed α_4:β_7 integrin with the mucosal vascular adressin, MAdCAM1, expressed on the endothelial cells of blood vessels and high endothelial venules of the intestine[10]. Effector T cells that were originally primed in the large intestine also express the receptor CCR10, the ligand of which is CCL28 produced by colon epithelial cells[1]. Effector T cells that re-enter into the intestinal mucosa re-distribute into unique compartments: $CD4^+$ T cells remain largely in the lamina propria, while CD8+ T cells migrate preferentially to the epithelium, although ~40% of T cells in the lamina propria are also CD8+ T cells (as reviewed in Mowat 2003[10] and Cheroutre and Madakamulti 2004[12]). Lymphocytes that enter the epithelium stop expressing α_4:β_7 and instead express α_E:β_7, the receptor for which is E-cadherin on epithelial cells[1].

1.3. TCR profiling

The TCRαβ repertoire of the adult human intestine has been profiled in the past and has largely been found to be oligoclonal (i.e. made up predominately of a few TCR clonotypes)[13–17]. At the time these studies were undertaken however, the technology available for immunoprofiling was of low resolution and limited specificity. A brief overview of some of these studies and technologies are described in the sections below:

1.3.1. TCR spectratyping

In 1993, Pannetier and colleges developed an approach that allowed for analysis of the T cell repertoire called TCR spectratyping[17]. Briefly, TCR spectratyping involves amplifying target TCRβ sequences with 5′ TRBV-gene specific primers and a common 3′ TRBC gene-specific primer. These amplification products are used as templates for a primer extension reaction using an additional, fluorescently labeled 3′ TRBC-gene primer. The fluorescent products are end-labeled and extended through the hypervariable CDR3 region, and run on polyacrylamide sequencing gels to reveal precise sizes of the CDR3. With the aid of an automated sequencing machine and appropriate software, the size of the peaks corresponding to discrete CDR3 lengths can

be analyzed[17,18]. Using this technology, the researchers concluded that there were at least 2,000 TCRβ distinct transcripts in mice[17]. This estimate was far from what has been directly measured recently using more sensitive approaches (discussed in section 1.3.3). Further, TCR spectratyping does not allow quantification of individual T cell clonotypes, and makes the identification of rare TCR sequences difficult[19]. In addition, TRBV-gene usage cannot be assigned quantitatively due to differential efficiencies between TRBV-gene family-specific PCRs[20]. This technology has therefore largely been replaced by higher throughput DNA sequencing technologies that are of finer resolution and greater specificity[19].

1.3.2. TCR amplification and sequencing

Ostenstad and colleges (1994)[21] investigated (tumour infiltrating lymphocyte) TIL TRBV diversity from the tumours of seven colorectal carcinoma (CRC) patients using TRBV-specific amplification with 19 different TRBV-gene family primers and conserved TRBC primers. They found that in three patients, there was an oligoclonal pattern of TRBV-gene family usage (range of 1 – 4 families used), while in the remaining four patients, there was limited heterogeneity (range of 8 – 20 families used). They concluded that TRBV-gene usage in TILs was limited when compared to that of PBL and lamina propria T cells. While their results were informative on the possible TRBV makeup of the TIL population, they are not informative on any other aspects of the TCR repertoire, including CDR3 sequence diversity, TRBJ gene usage and TCRβ clonotype abundance.

Sanger sequencing offered increased resolution for profiling the TCR repertoire by sequencing TCR-specific amplification products. Blumberg and colleges (1993)[15] found that the majority of intraepithelial T cells (IETs) and lamina propria T cells derived from five healthy patients were expansions of relatively few T cells. They used 20 TRBV-gene family primers and conserved TRBC gene primers to amplify TCRβ sequences from cDNA. In two cases, a patient had a single TRBV-gene family expressed predominantly, while for the three other patients, TRBV-gene usage was more evenly distributed between 14-15 families. They further cloned and Sanger sequenced these products from three donors on a sequencing gel, revealing that most of the IETs were derived from the clonal expansion of relatively few T cells. While higher resolution than

the PCR-based study described above, Sanger sequencing is still of relatively low sensitivity. Further, given the extreme variability of the CDR3, sequencing the TCRβ repertoire from three individuals is not enough to draw conclusions about the TCRβ diversity in the gut.

Li and colleges (2003)[22] sequenced TCRβ transcripts reverse transcribed from the tumour tissues and peripheral blood lymphocytes (PBL) of eight CRC patients. They found an oligoclonal expanded TRBV repertoire both in tumour tissues and peripheral blood lymphocytes (PBL). They sequenced the highly variable CDR3β region, providing a more comprehensive description of TCRβ diversity. They found oligoclonal expansions transcripts of TRBV-genes from TIL and PBL. Additionally, they found TCRβ transcripts derived from TILs of patients that expressed the same TRBV/TRBJ combination and the same CDR3β nucleotide sequence. Further, they were able to identify some TCRβ transcripts that that despite using different TRBV/J genes, had the same CDR3β nucleotide sequence. While more specific than the previous attempts at characterizing TCRβ repertoire in CRC, results from this study were still low sensitivity, using Sanger sequencing and only including the repertoires of eight individuals.

1.3.3. Next generation sequencing

The advent of next generation sequencing (NGS) has provided the tools needed for large-scale sequence immunoprofiling at high resolution and low cost. When coupled with unbiased amplification techniques such as rapid amplification of 5' cDNA ends (5'RACE), rare clonotypes can be detected from many individuals at once, resulting in a more complete characterization of the TCRβ repertoire. The high throughput capability of NGS allows for a more complete characterization of the human TCRβ sequence repertoire to be obtained, evident by profiling the peripheral blood (PB) TCRβ repertoire from hundreds of individuals congruently[23], and by the exhaustive sequencing of the TCRβ repertoire present within a sample of peripheral blood of a single individual[5]. Further, NGS sequencing has allowed for more definitive estimates of the TCRβ diversity and has shown that a single PB sample is not adequate to fully characterize the TCRβ diversity of an individual[24]. While, NGS technologies have shorter read lengths and show considerable base error when compared to other technologies[25], but these limitations can be successfully addressed computationally[5].

1.4. Tumour immunity

At the most basic level, cancer is the result of uncontrolled proliferation of the progeny of a single transformed cell[1]. The immune system can play dual roles in cancer- not only can it suppress tumour growth by destroying or inhibiting cancerous cells (e.g. through the actions of cytotoxic $CD8^+$ T cells), it can also promote tumour progression (e.g. by the actions of $CD4^+$ regulatory T cells)[26].

1.4.1. The body's response to cancer

The way in which the human immune system responds to cancer is called the cancer immunoediting hypothesis (CIH)[26]. Briefly, the CIH encompasses 3 phases: elimination, equilibrium and escape. In the elimination phase, cells of the adaptive and innate immune systems detect and destroy potential tumour cells[26]. The detection of and response to self-TA pMHC complexes however is limited to the small population of T cells that may have escaped negative selection in the thymus and are thus reactive against self antigen. Rare tumour variants that can survive the elimination phase can enter into the equilibrium phase, in which cells of the adaptive immune system prevent tumour cell outgrowth, while at the same time shape the immunogenicity of tumour cells[26]. Tumour cells therefore can lie in a state of dormancy, and may reside for extended periods of time before resuming growth. Tumour cells that acquire the ability to circumvent immune recognition and/or destruction emerge as progressively growing and visible tumours are in the escape phase. These tumours are often of low immunogenicity and can appear invisible to the cells of the adaptive immune system.

During elimination and equilibrium phases (and likely less frequently during the escape phase), T cells can actively survey tumour antigens (TA) on the surfaces of tumour cells and antigen presenting cells (APCs). Several categories of TA that can induce T cell responses in vitro and in vivo are (1) cancer testis antigens - protein antigens normally expressed only in germ cells of testis, and in some human cancers[27], (2) differentiation antigens - protein antigens which are expressed on malignant and normal cells, (3) point mutations of normal genes, (4) self antigens that are overexpressed by malignant cells, and (5) viral antigens[28]. If a T cell engages with a TA,

becomes activated and clonally expands, a population of tumour antigen-specific T cells can be generated.

1.4.2. Colorectal carcinoma

Colorectal carcinoma (CRC) is currently the third most common cancer type and the second and third leading cause of cancer death in men and women respectively, contributing to an estimated 9,200 deaths (12% of all cancer deaths) annually in Canada[29]. While largely spontaneously arising, it is hereditary in ~20% of all CRC cases[30]. The five-year survival rate is only 64%[30], largely owing to late detection after the cancer has already spread to other areas of the body[31]. If detected at an early stage, the survival rate can be increased[30–32]. Current screening approaches are often of low sensitivity, low specificity and/or are highly invasive. These approaches include both physical examinations for colorectal abnormalities (e.g. digital rectal examination, flexible sigmoidoscopy, barium enema, CT colonography, colonoscopy) and molecular tests for biomarkers of CRC (e.g. fecal occult blood test, fecal DNA/RNA/protein analysis, blood DNA/RNA/protein analysis)[33]. CRC often does not present symptoms until it has reached later stages, making an urgent need for a more highly specific and low invasive screening approach for high risk, asymptomatic patients.

1.4.3. Tumour infiltrating T cells

The presence of T cells in solid tumours reflects an ongoing immune response against transformed cells[28], and their presence has been associated with survival in CRC[34–39]. Multiple subtypes of tumour infiltrating T cells (TIfTs) have been found within CRC tumour masses including $CD4^+$ helper T cells[38] and CD8+ cytotoxic T cells (CTLs)[34–37], T regulatory cells[39], as well as the innate immune system's CD56+/57+ natural killer (NK) cells[37]. Those with the most potent anti-tumour effect studied to date are the $CD8^+$ CTLs. In 1998, Naito and colleagues[34] found that activated cytotoxic CD8+ T cells within tumour cell nests of CRC were significantly associated with improved survival of patients by both mono and multivariate analyses. They went on to suggest that their activated phenotype could be a result of antigenic stimulation by B7+ macrophages, possibly displaying tumour antigens, distributed along the invasive margin of colon tumours[40]. They further suggested that after being activated at the margin, T

cells may migrate into cancer cell nests, and exhibit higher proliferative activity and cytotoxic effector function[34,40]. Further, given that the occurrence of metastasis in the liver and lung via a haematogenous route is one of the major causes of death in patients with CRC, they suggested that these activated cytotoxic T cells may act systemically in the liver and lung to suppress metastasis after bring activated in the cancer tissue, thus improving survival.

1.4.4. Cancer immunotherapy

Given the survival advantage conferred by the presence of TIfTs in CRC[34–39] and other cancer types[41], it is not surprising that T cells and other cells of the innate and adaptive immune systems are commonly used in immunotherapeutic approaches against cancer. Strategies to enhance anti-tumour immune responses are called immunotherapies, and a few are briefly overviewed here (reviewed in Ogino et al. 2011[42]): (1) Non-specific immune stimulation (NSIS) - APCs can be activated via the injection of molecules that bind DC receptors. Activated dendritic cells in close proximity to the tumour can then attract T cells, and in the presence of IL-2 and IFNγ, an anti-tumour response can be mounted. Another way to boost the immune system with NSIS is through the injection of weakened bacteria into the area surrounding the tumour. Bacterial presence induces inflammation, resulting in lymphocyte recruitment to the tumour (reviewed in Lesterhuis and Punt[43]). (2) Immune checkpoint blockade - the use of an anti-CTLA4 antibody can block the regulatory CTLA4 molecule on the surface of a T cell that would normally bind B7 in a regulatory role. Blocking it essentially increases the immune response to a given antigen (reviewed in Lesterhuis and Punt[43]). (3) Targeted antigen specific vaccines can be used to direct immune cells specifically to cancer tissue. Vaccines that use peptide variants of TAs bind MHC or tumour specific TCRs with high affinity, stimulating the expansion of often low affinity TA-specific T cells that escape negative selection during development and are sub-optimally activated by native tumour antigens (reviewed in Jordan et al.[44]). Administered vaccines can also include a patient's own irradiated and genetically engineered tumour cells secreting immune system stimulating growth factors, or a patient's own immune cells. In this third case, immature autologous APCs can be matured and loaded with tumour antigen ex vivo. Once reintroduced back into the body, these APCs can help to stimulate the immunes system against the tumour (reviewed in Lesterhuis and Punt[43]). (4) Adoptive T

cell therapy - tumour antigen-specific TCRs from CRC tumour infiltrating lymphocytes, in the peripheral blood of CRC tumour bearing patients[43], or genetically engineered to harbour tumour antigen specific TCRs, can be clonally expanded in vitro and reintroduced back into the body in conjunction with T cell promoting cytokines (e.g. IL-2). Now with an expanded tumour antigen specific T cell army, increased anti-tumour responses can occur.

1.5. Aims of this thesis

Because of the apparent survival advantage for patients with T cell infiltration in CRC tumours[34–39], a better understanding of the T cell response to CRC is of great interest. The CRC TCR repertoire has been explored in previous studies, but with low resolution. In this thesis, I have explored TCRβ sequence diversity from multiple colorectal carcinoma specimens using high resolution NGS technology. To my knowledge, this is the first time such a task has been attempted. In chapter two, I have characterized the TCRβ sequence repertoire present from CRC tumour tissue of patients by examining specific features of the TCR repertoire, namely sequence abundance and diversity, sequence composition and TRBV/J gene usage relative to mucosal control tissue repertoire. In chapter 3, I have investigated the presence of shared CDR3β sequences in CRC, and if their presence in tumour tissues can be associated with improved survival in CRC. The findings presented in this thesis could contribute to our understanding of the T cell adaptive immune response to CRC. Characterising the host T cell response could help to distinguish potential targets of immunotherapy and provide a biological tool for detecting malignancy.

1.6. Contributions

Tissue specimens were collected by skilled physicians and were obtained from Rebecca Barnes and Peter Watson at the BC Cancer Agency Tumour Tissue Repository in Victoria, BC, Canada. Doug Freeman, a skilled senior laboratory technician at the GSC, preformed all of the RNA extractions from patient tissues. After RNA extraction, I followed an Illumina library construction sequencing approach previously developed in

the laboratory of my senior supervisor, Dr. Robert Holt. The sequencing group at the GSC preformed the Illumina paired end sequencing. The GSC Bioinformatics Coordinator, René Warren, developed the TCRβ profiling pipeline I used to process my sequence reads. The Fishers Exact Test script used in data analysis was written by myself, with assistance from René and GSC scientist, Martin Krzywinski. As part of a different study, René developed a computational HLA predictor pipeline, which was used to predict HLA types for a subset of patients in the cohort used in this study. Sarah Munro of the GSC verified predicted HLA types for two of the individuals. PhD. candidate Mauro Castellarin and Co-op student Scott Brown designed the sequence validation approach (TaqMan qPCR) I used. Martin further aided in the design of Circos plots. I preformed all other aspects of this project.

2. TCRβ sequence diversity in colorectal carcinoma

2.1. Background

T cells are key players in immunotherapeutic approaches against cancer because they can recognise and respond to tumour antigens (TAs)[1] and can confer anti-tumour effects[46,47]. Tumour antigen-specific TCRαβs have been found both in the population of CRC tumour infiltrating lymphocytes, as well as in peripheral blood of CRC tumour bearing patients[45]. Antigenic specificity of the TCRβ is conferred in the CDR3β, a primary site of antigen contact. To capture the full diversity of the CDR3β (and thus TCRβ diversity), utilization of non-biased, highly specific molecular techniques are required. Previous studies have used techniques such as rapid amplification of 5'-cDNA ends (5'-RACE) and TCR spectratyping to profile mucosal TCR CDR3β diversity. More recently, next generation sequencing (NGS) approaches have been used[5,23], as they are more sensitive than other approaches, offering more complete characterization of the TCRβ sequence repertoire by having the capability to capture rare sequences and to provide quantitative information regarding sequence abundance.

Currently, the TCRβ sequence repertoire in CRC has not been characterized at high resolution using NGS approaches. It is not known therefore whether or not a TCRβ tumour specific sequence repertoire exists in CRC. Detection of CRC at an early stage, before there are any symptoms of disease, offers the best chance of effective treatment, thus increasing the likelihood of survival for patients[29]. If present, CRC specific TCRβ sequences could be used as biomarkers in highly specific, non-invasive screens for high-risk or asymptomatic individuals, thus increasing the likelihood of early detection. Further, knowledge of a CRC specific TCRβ sequence repertoire could aid in the design of future tools to target CRC tumour antigens.

For this project, it was my aim to characterize the TCRβ sequence repertoire present in CRC tumour tissue. To do this I profiled CRC tumour surgical sections, increasing the likelihood of capturing a possible CRC-associated TCRβ sequence repertoire - in gut tissue, there can be multiple populations of T cells, including those in gut associated lymphoid tissues (GALT), gut epithelium and lamina propria, as well as those deriving from peripheral blood mononuclear cells (PBMCs) and tumour infiltrating T cells (TIfTs). I hypothesized that a tumour-associated TCRβ sequence repertoire would be found in the tumour tissues of CRC patients.

2.2. Materials and Methods

2.2.1. Clinical specimens

Tumour and matched control tissue specimens came from a cohort of 43 CRC patients with differing clinical backgrounds - patients had differing stages of CRC - stage I (n=6), stage II (n=15), stage III (n=15), stage IV (n=3) and some had no stage assigned (n=4) when tissue sections were removed. Patient tumours were localized to differing regions of the colon/rectum, including the ascending colon (n=12), cecum (n=9), sigmoid colon (n=7), rectum (n=5), recto-sigmoid junction (n=3), transverse colon (n=3), descending colon (n=2) and the right hepatic flexure (n=2). Further, patients had undergone differing treatment regimes prior to surgery, including radiotherapies, chemotherapies or combination therapies among others. At the time of TCRβ sequence data collection, 25 individuals were still alive (21 free of disease, 1 with disease, 3 unknown if disease present) and 18 were deceased (12 result of disease, 6 unknown if disease present at death).

Tissue specimens were collected as described in Castellarin et al.[48] and is re-described here: Fresh CRC samples were obtained with informed consent by the BC Cancer Agency Tumour Tissue Repository (BCCA-TTR)[49], which operates as a dedicated biobank with approval from the University of British Columbia–British Columbia Cancer Agency Research Ethics Board (BCCAREB). The BCCA-TTR platform is governed by Standard Operating Procedures (SOPs) that meet or exceed the recommendations of international best practice guidelines for repositories (NCI Office of

Biorepositories and Biospecmen Research, NCI Best Practices for Biospecimen Resources). Specimens were handled with very close attention to maintaining integrity and isolation. Overall average collection time (time from removal from surgical field to cryopreservation in liquid nitrogen) for all colorectal cases in the BCCA-TTR was 31 minutes. Tumour tissues were sectioned from the tumour mass, while control tissues were removed from an area adjacent to the tumour, but deemed by skilled clinicians as being non-tumourous. Tissue specimens were held briefly at -20°C during frozen sectioning, using 100% ethanol to clean the blade between all samples. Clinical pathological and outcomes data were obtained from the BC Cancer Agency clinical chart including tumour features reported according to the American College of Pathologists criteria and the "Protocol for Examination of Specimens from Patients with Primary Carcinoma of the Colon and Rectum"[49].

2.2.2. Illumina library construction

2.2.2.1. RNA extraction

Total RNA was purified from frozen colorectal tumour and matched control tissue sections using the Qiagen's RNeasy Plus kit. Briefly, the kit enables RNA purification with the removal of genomic DNA from of tissue. After homogenization of frozen tissue sections (30 mg per patient) in denaturing buffer, lysate was centrifuged and transferred to a genomic DNA eliminator column. DNase was added to the column, followed by multiple buffer washes. Total RNA was eluted from the column using 50 µl RNase-free water.

2.2.2.2. cDNA synthesis

First strand cDNA synthesis was run on groups of 11 RNA samples (plus one negative control that substituted sterile water for RNA template) at a time, with careful consideration to avoid cross-contamination. Control tissue RNA samples were processed first, and once all had undergone cDNA synthesis, synthesis of tumour cDNA from RNA commenced. First strand cDNA was synthesized as described in Freeman et al. 2009[23] using a published TRBC-gene primer[50] (C6) and a template switching oligo[51] (LTS TSP) (see Appendix D1 for sequences) to provide a 5' template for Rapid Amplification of cDNA Ends (RACE). cDNA synthesis reaction conditions were as follows: 333ng/µl of RNA and 50 µM C6 and LTS oligonucleotides, incubated for 2

minutes at 70°C. 40U/µl RNaseOUT (Invitrogen), 5X first strand buffer (Clontech), 20mM DTT, 10mM dNTP (NEB premix), 600 units of Superscript II (Invitrogen) were added to make a 20 µl reaction volume. Extension was for 90 min at 42°C, followed by enzyme inactivation for 15 minutes at 72°C.

2.2.2.3. 5' Rapid amplification of cDNA Ends (5'RACE)

5'RACE was performed with 0.5µl of first-strand reactions, without normalization in an attempt to preserve any differences in the number of T cells biologically present between patient samples. A combination of 3 oligonucleotides was used: an oligo that anneals to the end of the cDNA template (UPM long), a short oligo (UPM short) and a uniquely designed indexed TRBC-gene specific primer (iGSP). The iGSP was designed to sit on a conserved region of the TRBC1,2 genes, followed by an internal unique hexameric index sequence, a Pac1 restriction site, and a non-consensus tail (Fig. 2a). A unique iGSP was used for the amplification of each individual patient tissue sample (n = 43 tumour tissue samples and n =43 control tissue samples), as well as for negative control reactions (n=10) (see Appendix D2 for sequences). PCR amplification was done for each group of 11 samples (plus the negative control) as in cDNA synthesis, with careful consideration to avoid cross-contamination. Control samples were processed first, and once all 43 cDNA samples had undergone indexing PCR, the workbench and equipment used were sterilized, and processing of tumour cDNA samples occurred. For each group, a known plasmid TCRβ template also underwent primary PCR without iGSP as a control to test for primer contamination between samples. PCR reaction conditions were as follows: 5X Phusion HF buffer, 10 mM dNTP (NEB premix), 10 mM each UPMshort, UPMlong, iGSP, DMSO, 1 unit of Phusion DNA polymerase in a 50 µl reaction volume. Thermal cycling conditions were 30 seconds denaturation at 98°C, 26 cycles of 10 seconds at 98°C, 10 seconds at 61°C, and 20 seconds at 72°C, plus a final extension for 5 minutes at 72°C. For visual confirmation of amplification of the 500-650bp product, PCR product from each sample was run on a 1.5% agarose gel (120V,1.5 hours with 10% Sybr stain).

2.2.2.4. Sample pooling

Following indexing PCR, 10 µl PCR product from each patient tumour (n=43) and control (n=43) sample as well as from negative controls (n=10) was pooled together and

purified using Qiagen's QIAquick PCR Purification kit for purification of PCR products and two rounds 1.5% agarose gel extractions of the 500-650bp TCRβ product (75V, 4 hours with buffer replacement for cooling and 10% Sybr stain) followed by Qiagen mini-elute gel extraction purifications. Purified DNA was eluted into 10 µl EB buffer.

2.2.2.5. Biotinylation PCR

In order to obtain a cleaner product, a semi-nested PCR was performed on 0.5µl of the 500-650bp pooled primary PCR product using a biotinylated primer (Biotin Tail) and a semi-nested primer (SN2) (see Appendix D3 for sequences, Fig. 2b, 2c). Reaction conditions were as follows: 5X Phusion Buffer, 10mM dNTPs (NEB premix), 10 uM of each Biotin Tailed primer and SN2 primer, DMSO, 1 unit Phusion DNA Polymerase in a 50 µl reaction volume. Thermal cycling conditions were 30 seconds denaturation at 98°C, 12 cycles of 10 seconds at 98°C, 10 seconds at 65°C, and 20 seconds at 72°C, plus a final extension for 5 minutes at 72°C. After PCR, the nested product underwent QIAquick PCR Purification followed by a 1% low melting point agarose gel excision (75V, 4 hours with buffer replacement for cooling and 10% Sybr stain) and β–agarose:Phenol purification (and precipitation) to increase yield of captured product. The product was eluted to 60 µl with EB and run out on a 1% agarose gel (80V, 2 hours, 10% Sybr stain) to confirm the presence of the 500-650bp product.

2.2.2.6. Shearing

To ensure double stranded coverage of the TCRβ CDR3 region during sequencing on the Illumuna GAIIx and HiSeq platforms, the 500-650bp product needed to be sheared into 200-300bp fragments (Fig. 2d). Random sharing occurred via sonication on the Covaris E-series sonicator. Briefly, the Covaris E-series uses a transducer to transmit focused acoustic energy to the sample in a small, localized area. At high intensity, the acoustics create a shock wave environment that is capable of DNA fragmentation. Samples underwent sonication under conditions to maximize the number amount of 200-300bp fragments (9 minutes shearing at a duty cycle of 10% and an intensity level of 5).

2.2.2.7. Isolation of TCRβ via streptavidin purification

Following sonication, the fragmented product underwent streptavidin purification to

isolate biotinylated 3' C-region products (Fig. 2e). Reaction conditions were as follows: fragmented product and streptavidin Dynabeads M270 were mixed and incubated at RT for 15 minutes, followed by followed by 3 minutes on a magnet and 3 buffer washes. Breaking of the biotin end labels occurred via addition of BSA buffer and 50U Pac1 enzyme, incubation at 37°C for 2 hours with gentle rocking, and application of a magnet for 3 minutes. The supernatant was removed and incubated at 70°C for 20 minutes, followed by standard EtOH precipitation and re-suspension in 8 µl EB. The 125-175bp TCRβ fragment was visualized on an 8% polyacrylamide gel (PAGE) excised, and precipitated again using standard EtOH precipitation.

2.2.2.8. Preparation for Illumina paired-end sequencing

Preparation for Illumina sequencing continued firstly through blunting of DNA ends. Reaction conditions for blunting were as follows: 1 X NEB blunting buffer, 1 mM dNTPs, Blunting enzyme mix in a 25 µl reaction volume incubated for 30 minutes at 21°C, standard EtOH precipitation and re-suspension using EB buffer. Blunted products were then A-tailed using the following reaction conditions: 1X NEB reaction buffer, 10 mM dATP, 1U of Klenow Exo$^-$ to a total reaction volume of 50 µl and incubation at 37°C for 30 minutes. A-tailed products were purified using phenol-chloroform at re-suspended with EB buffer to 10 µl. Illumina TS adapters (see Appendix D4 for sequences) were ligated onto A-tailed products using the following reaction conditions: 20X NEB ligase buffer, PE adapters, 1000U NEB DNA ligase buffer, incubation at 21°C for 15 minutes and purification using a QIAquick column (Qiagen). The ligated product was eluted with 50 µl of EB buffer.

The ligated product was amplified prior to Illumina sequencing using two paired end Illumina primers (PE PCR primers 1 and 2) (see Appendix D4 for sequences, Fig. 2f). Reaction conditions were as follows: adapter ligated product, 5X Phusion HF buffer, 10 mM dNTP (NEB premix), 10 mM each PE PCR1 and PE PCR2 primers, DMSO, 1 unit of Phusion DNA polymerase in a 50 µl reaction volume. Thermal cycling conditions were 2 minutes denaturation at 98°C, 15 cycles of 10 seconds at 98°C, 15 seconds 65°C, 15 seconds at 72°C, plus a final extension for 5 minutes at 72°C. The amplified product was purified using a QIAquick column (Qiagen), re-suspended in 10 µl EB and quantified on the NanoDrop spectrometer to be 250.8 ng/µl. Further purification occurred via 8% PAGE gel excision and standard EtOH precipitation. The purified product was

quantified using Nanodrop at 187.2ng/µl.10 µl of prepared product was submitted for Illumina next generation sequencing.

2.2.3. *Illumina paired end sequencing*

Following library construction, adapter-ligated DNA template was loaded onto a single lane of an Illumina flow cell (Fig. 2g), and bound to a lawn of oligonucleotide (oligo) anchors grafted to the cell surface. In each PCR cycle, priming occurred by arching of the template molecule such that the adapter at its un-tethered end hybridized to and was primed by a free oligo in the near vicinity on the flow cell surface[25]. This process resulted in a raindrop pattern of clonally amplified templates (ie. clusters). Sequencing by synthesis proceeded in parallel using reversible four-color fluorescence (e.g., a mix of the four bases each labeled with a different cleavable fluorophore, such that they can be used simultaneously rather than sequentially to interrogate a given nucleotide position in the template)[25]. Labeled terminators, primer, and polymerase were then applied to the flow cell. After base extension and recording of the fluorescent signal at each cluster, the sequencing reagents were washed away, labels are cleaved, and the 3' end of the incorporated base is unblocked in preparation for the next nucleotide addition[25]. Paired end sequencing facilitated reading both the forward and reverse template strands of each cluster. A single lane of paired end sequencing (114bp read length) was done using the Illumina GAIIx using SCS 2.8 software, V4 cluster generation and V5 sequencing reagents. Later, an additional lane of sequencing was done using the original library sample using Illumina's HiSeq platform (150bp read length) using HCS1.3.8 software, V2 cluster generation and V1 sequencing reagents.

2.2.4. *Bioinformatic analysis*

2.2.4.1. TCRβ sequence assembly and mining

Illumina paired end sequence reads generated from the Illumina GAIIx and HiSeq runs were processed together through a bioinformatic pipeline specifically designed to profile CDR3β Illumina paired end reads (see Appendix A, Fig. 2h). The pipeline first required that raw Illumina reads were copied from the server into a working directory, where they were decompressed using bunzip2. A microassembler was developed to join overlapping paired-end reads from each sequencing template. Briefly,

the assembler uses the sequence aligner Exonerate[52] to perform gapless alignments between any two mate pairs joining them into a single contiguous sequence (contig) provided the reads align on opposite strands, facing inwards[5]. Each alignment was scrutinized at run-time to resolve base conflicts, whenever applicable, and a consensus base sequence and quality score was generated for each newly formed contig[5]. Agreeing bases on opposite strands were given a consensus score that corresponds to the sum of individual Q scores[5]. Disagreeing bases were assigned an N at that position and a score of zero, unless the base call on one strand was 99.9% accurate or higher (>= Q30) and the discrepant base on the other strand was <99% accurate (< Q20)[5]. In the latter case, the most accurate base was called[5]. The aligned paired end sequence contigs were annotated by aligning contigs to the 3' end of Ensembl TRBV-gene predictions[5] and searched for the presence of 18 consecutive TRBJ segment bases. For a TRBJ segment, any 18-letter word from base positions 1–25 characterized uniquely that segment and allowed the identification of the precise TRBJ segment boundary as well as the number of TRBJ bases deleted. The TRBV segment boundaries and exact number of deleted TRBV bases were inferred by tracing back the alignments in the contig under scrutiny. The CDR3β was defined as the region between the last conserved cysteine (C) codon of the TRBV-gene and the first conserved phenylalanine (F) codon of the TRBJ gene in the conserved motif FGXG[5]. Since the raw data contained reads from multiple indexed samples, the pipeline also binned sequence contigs according to the unique hexameric index sequence applied to each sample during library construction. The pipeline thus allowed us to find specific TCRβ sequences and keep track of patient identity from which the sequences originated.

2.2.4.2. MySQL sequence database

The mined sequence data was outputted to a text file that was used to populate three tables in a MySQL database (see Appendix A); the three tables (named Contig, Run, and Sample), contained information pertaining the TCRβ/CDR3β sequence information, sample identification information. Features of the tables were linked together in command-line queries to facilitate profiling (for examples of commands see Appendix Bi, ii). Command line queries included those specific for certain features of the TCRβ repertoire such as CDR3β sequence identity, sequence depth of coverage (DOC),

predicted amino acid sequence identity, TRBV-gene usage, TRBJ gene usage, functionality (in-frame versus out-of-frame) and patient sample identity.

TCRs are extraordinarily difficult sequencing targets because any given receptor variant may be present at very low abundance and may differ legitimately from other receptor variants by only a single nucleotide[5]. This property makes distinguishing rare TCR sequences from sequencing errors difficult. For this reason, sequences used in analysis were further limited to only include those with a DOC of greater than one (DOC > 1), in the correct reading frame (frameCheck = 1) and with an unambiguous TRBV-gene assignment (vName = 1) (see Appendix Bi for an example of a MySQL search query used).

2.3. Results

All sequence data presented in this chapter (unless otherwise stated and excluding section 2.3.1) represent in-frame sequences with only one identified TRBV-gene and a depth of coverage (DOC) of greater than one. Table 1 shows a summary of this data and additionally, although not addressed in this thesis, shows a summary of the data with varying filters applied.

Further, throughout this chapter, *CDR3β sequence* refers to the sequences present in the dataset encoding the hypervariable CDR3β, the site of antigen contact. (e.g. CDR3β amino acid sequence (aaCDR3β): CSARAPDGNTGELFF; CDR3β nucleotide sequence (ntCDR3β): TGCAGTGCTAGAGCCCCCGACGGTAACACCGGGGAGCTGTTTTTT); *TCRβ sequences* refers to the sequences present in the dataset encoding the CDR3β, partial TRBV-gene and full TRBJ-gene, taking into account bases added and deleted at gene junctions (e.g. TCRβ sequence: TRBV20-1_1_CCCCCGACGGT_2_TRBJ2-2 ; Sequence *depth of coverage (DOC)* refers to the number of observations of a given TCRβ /CDR3β sequence in a data set.

2.3.1. *Sequencing statistics*

Paired end sequence reads were generated from 51,700,780 clusters on a single lane on the Illumina GAIIx platform, from which 36,017,984 contiguous sequences joined. From these, 13,430,222 TCRβ sequences had a TRBV-gene identified. Additionally, paired end sequence reads were generated from 91,282,098 clusters on a single lane of the HiSeq platform, 81,504,381 of which joined into contiguous sequences. From these, 11,775,841 TCRβ sequences had a TRBV-gene identified and were added to the GAIIx sequences for annotation.

2.3.2. *Sequence diversity and abundance in CRC*

2.3.2.1. Sequences present at varying depths of coverage (DOC)

Overall, there were 47,524 unique TCRβ sequences, 35,740 unique ntCDR3β sequences and 23,305 predicted unique aaCDR3β sequences present in my complete data set (Fig. 3). The number of unique sequences present decreased with increasing DOC.

2.3.2.2. Sequence diversity and abundance similar in tumour and control repertoires

Collectively, 79,497 (25,464 unique) TCRβ sequences were found from the tumour repertoire, with a depth of coverage (DOC) of 8,547,637. This corresponded to 19,532 unique CDR3β nucleotide sequences (ntCDR3β), predicting 14,727 unique CDR3β amino acid sequences (aaCDR3β) (Table 1). This was similar to what was observed in the control repertoire where collectively, 76,944 (25,010 unique) TCRβ sequences were found, with a DOC of 7,790,030. This corresponded to 18,882 unique ntCDR3β sequences, predicting 14,266 aaCDR3β sequences (Table 1).

Some TCRβ sequences were present in no-template negative controls after quality filtering, representing possible cross contamination between samples likely at the library construction stage or as a result of sequencing error modifying the hexamer identifier sequence. These sequences were subtracted from the data set.

2.3.2.3. Per patient sequence diversity and abundance similar in tumour and control repertoires

There was considerable variation in the counts of distinct sequences and total abundances of TCRβ/CDR3β sequences present in tissue repertoires both across all and within patients (Table 2 and Fig. 4a). On average, 1,848.77 +/- 1,099.68 (mean +/- standard deviation) (1,719.65 +/- 1,001.32 unique) TCRβ sequences were present in each tumour sample, identified by an average DOC of 198,728.26 +/- 166,751.42. An average of 1,568.37 +/- 878.92 unique ntCDR3β sequences, predicting 1,453.12 +/- 788.58 unique aaCDR3β sequences were present (Table 2, Fig. 4a). Similarly, in control tissue samples, on average 1,789.40 +/- 1,159.83 (1,668.21 +/- 1,053.45 unique) TCRβ sequences were present from each patient sample, identified by an average DOC of 181,163.49 +/- 163755.31. An average of 1,512.05 +/- 913.12 unique ntCDR3β sequences, predicting 1,401.95 +/- 819.80 unique aaCDR3β sequences were present (Table 2, Fig. 4a). Thus, the average number of unique sequences and DOC present in the tumour repertoire did not differ from the control repertoire.

The distribution of ntCDR3β sequence diversity and DOC within each patient sample (regardless of being of tumour or control tissue origin) followed a similar distribution, where a small proportion of unique ntCDR3βs made up the majority of DOC, with the remaining large proportion of unique ntCDR3βs contributing to a minority of the sequence abundance in each sample (Fig. 4b).

2.3.2.4. Sequences observed only in tumour repertoires

In order to isolate any TCRβ/CDR3β sequences that were specific to CRC tumour repertoires, I removed all sequences from the analysis that were found in both tumour and control tissues, and analysed repertoire diversity observed *only* in the tumour repertoire.

A majority of the sequences were specific to *only* the tumour or *only* the control repertoires. Overall, 88%, 86%, and 82% of the unique TCRβ, ntCDR3β and aaCDR3β sequences respectively present in the tumour repertoire, were only present within the tumour repertoire (Fig. 5). Similarly, in the control repertoire, 88%, 86% and 81% of the unique TCRβ, ntCDR3β and aaCDR3β sequences present in the control repertoire were only present within the control repertoire (Fig. 5).

2.3.2.5. CDR3β nucleotide sequences observed only in tumour repertoires of each patient

On average 55% of the unique ntCDR3β sequences present in a patient's tumour repertoire were *only* present in the tumour repertoire. This was similar for control tissues, where 53% of the unique ntCDR3β sequences present in a patient's complete control repertoire were only present in the control repertoire (Table 3).

2.3.3. CDR3β sequence composition in CRC

2.3.3.1. Predicted CDR3β amino acid consensus sequence present in complete tumour repertoire differs slightly from complete control repertoire

It is possible that some of the CDR3β sequences present from the tumour repertoire are expressed on T cells that are specific to CRC tumour derived peptide antigens (TAs). Characterizing the molecular composition of antigen-specific CDR3β sequences could thus provide valuable insight into the makeup of cancer-related antigens. The length of CDR3β in the tumour repertoire ranged from 18-84 nucleotides with a mean and standard deviation of 43.42 +/- 5.34 (or 6-28 predicted amino acid residues with a mean and standard deviation of 14.47 +/- 1.79) (Fig. 6).. This was similar to the control repertoire where, the length of CDR3β ranged from 24-84 nucleotides with a mean and standard deviation of 43.58 +/- 5.46 (or 8-28 predicted amino acid residues with a mean and standard deviation of 14.52 +/-1.83) (Fig. 6). For both tissue types, the most frequently observed sequence length was 42 nucleotides (or 14 predicted amino acid residues).

While the compositions of the 14 residue aaCDR3β sequences in tumour (n = 3,711) and control (n = 3,516) repertoires, were highly similar, there were some differences; The most frequently observed consensus sequence for the complete tumour repertoire was CASSLGGG**G**NEQ**F**F and for the control was CASSLGGG**T**NEQ**Y**F (Table 4, Fig. 7) - In the tumour repertoire, position 9 (bold) was most frequently a non-polar glycine (G) and position 13 (bold) was a non-polar phenylalanine (F), while in the control repertoire, position 9 was a polar theronine (T) and position 13 was a non-polar tyrosine (Y) residue. While these consensus sequences do qualitatively differ, the top three most frequently observed residues at position 9 and 13 are in fact the same, but in

a different order (Fig. 7). When considering differences between tumour and control repertoires other than in the consensus sequence, there were notable residue-usage differences at position 7 and 9. Further, an aspartic acid residue (D) was more commonly used in control than tumour repertoires at position 10 (Fig. 7).

If specific for a TA, a T cell may have undergone antigen specific clonal expansion, represented in this study by a distinct CDR3β sequence present at relatively high DOC. In this study, a DOC > 99 was considered a relatively high DOC. In the DOC > 99 repertoire, 14 residue aaCDR3βs were again the most abundant (tumour n = 347, control n = 332). Further, the molecular composition of these sequences did not differ, resulting in the same consensus sequences (tumour – CASSLGGGGNEQFF, control - CASSLGGGTNEQYF) as described above for the DOC > 1 repertoire (data not shown).

2.3.3.2. Predicted CDR3β amino acid sequence in tumour *only* repertoire, slightly differs from control *only* repertoire

In the tumour *only* repertoire, the length of CDR3βs was similar to the complete tumour repertoire, ranging from 18-84 nucleotides (43.45 +/- 5.37 mean +/- SD), or 6-28 predicted amino acid residues (14.48 +/- 1.81). In the control *only* repertoire, the length of CDR3β ranged from 24-84 nucleotides (43.64 +/- 5.51), or 8-28 predicted amino acid residues (14.54 +/- 1.86) (Fig. 6). For both tissue types, the most frequently observed sequence length was 42 nucleotides (or 14 predicted amino acid residues).

The consensus sequence composition of the 14 residue aaCDR3βs for the tumour only (n = 3,056) and control only (2,861) repertoires were the same as above (section 2.3.3.1), differing again only at positions 9 and 13 (data not shown). When limiting the sequences to those 14 residue sequences present at DOC > 99 (tumour only n= 328, control n = 343) the consensus sequences did not change from those sequences found in section 2.3.3.1 (tumour – CASSLGGGGNEQFF, control - CASSLGGGTNEQYF) (data not shown).

2.3.4. *TCRβ sequence TRBV/J-gene usage in CRC*

2.3.4.1. TRBV/J gene usage highly similar in tumour and control repertoires

Complete characterization of the T cell response in CRC requires investigating TRBV/J gene usage. In the experimental approach used, only a portion of the TRBV-gene is available for assignment[23]. This, together with the often high sequence homology observed between some TRBV-genes makes it impossible to assign a single TRBV-gene to every TCRβ sequence[23]. I was successful however in making an unambiguous assignment for 83% of the in-frame, DOC >1 TCRβ sequences in both the tumour and control repertoires to one of 51 unique TRBV-genes. Usage ranged from 18.0% and 17.8% for TRBV20-1 to 1.26×10^{-3}% and 3.89×10^{-3}% for TRBV17 in tumour and control repertories respectively. All known functional TRBJ genes were unambiguously identified ranging from 19.6% and 19.2% for TRBJ2-1 to 1.16% and 1.27% for TRBJ1-3 in tumour and control repertories respectively. TRBD genes sustain substantial base deletion and overall transformation during formation of the TCRβ chain, so the segments are unrecognizable[23].

2.3.4.2. TRBV/J gene pairings highly similar between complete tumour and control repertoires

From my set of 51 TRBV-genes and 13 TRBJ genes, there were 663 potential pairings. In the entire data set, I found 622 unique TRBV/J-gene pairings, 591 and 597 of which were present in the tumour and control repertoires respectively (Fig. 8). The most frequent pairing was TRBV5-1 / TRBJ2-1, accounting for 4.19% and 3.67% of all pairings in tumour and control repertoires respectively, whereas 37 and 40 pairings occurred only once in tumour and control repertoires respectively (Fig. 8).

2.3.4.3. TRBV/J gene pairings present in tumour and control *only* repertoires

There were 23 TRBV/J gene pairings found *only* in the tumour repertoire (Table 5a): One pairing (TRBV7-7/TRBJ1-2) was present in the tumour repertoires of five patients, two pairings (TRBV7-7/TRBJ1-5, TRBV23-1/TRBJ2-4) were present in three repertoires each, two pairings (TRBV6-8/TRBJ2-5, TRBV5-3/TRBJ1-2) were present in two repertoires each and one pairing (see Table 5 for pairings) was present in each of

18 repertoires. One pairing (TRBV6-9/TRBJ1-4) from a single TCRβ sequence was present at a DOC = 187, but only from the repertoire of a single patient (78-T). All other pairings were found at low DOC (Table 5a).

There were 29 TRBV/J gene pairings found only in the control repertoire (Table 5b): Two pairings (TRBV6-8/TRBJ2-7, TRBV18/TRBJ1-3) were present in the control repertoires of four patients each, one pairing (TRBV5-3/TRBJ1-1) was present in three control repertoires, 10 pairings (see Table 5b) were present in two control repertoires each, and 16 pairings (see Table 5b) were present in a single control repertoire each. One pairing from a single TCRβ sequence (TRBV23-1/TRBJ2-3) was present with a DOC = 445, but only from the repertoire of a single patient (74-N). All other pairings were found at low DOC (Table 5b).

2.4. Discussion

In this study I characterized, for the first time, features of the TCRβ sequence repertoire present in CRC tumour tissues high resolution. I was able to capture TCRβ sequences present at variable DOCs, including some that were likely rare, indicated by a very low DOC from the tumour and control tissues of 43 CRC patients. Overall, there were no major differences in TCRβ/CDR3β sequence diversity, abundance or TRBV/J gene usage in the tumour repertoire when compared to the control repertoire. Where some slight differences emerge however were when considering the residue compositions of the aaCDR3β consensus sequences– the 14 residue aaCDR3β consensus sequence of the tumour repertoire was slightly different in comparison to the consensus sequence of the control repertoire. One of these differences was the more frequent usage of a non-polar residue at position 9 in the tumour repertoire, where a polar residue was used in the control repertoire. This difference however is called into question when considering that the three residues used most frequently used at this position were the same for both tissue types, just in a different stacking order. To confirm any sort of specific TCRβ CDR3 sequence signature in CRC, future studies should use a larger sample size and deeper sequencing. Investigating the compositions of all aaCDR3β lengths from tumour tissue repertoires would also be interesting. The differences I found were strictly qualitative, and future investigation would require use of

a statistical application to quantitatively test residue-usage differences between the different tissue repertoires. One approach for this could be designing a statistical program that randomly assigns residues to each position for a given aaCDR3β length repeatedly, and seeing how many times the observed consensus sequence could be generated randomly. While no specific CDR3β signature was observed for CRC tumour tissue in this dataset, the utility of high throughput sequencing in cancer T cell repertories is obvious - it has allowed for the detailed examination of highly variable repertoires with high sensitivity.

While other studies have found the TCRβ CDR3 repertoire in CRC to be largely oligoclonal in regards to TRBV gene usage, I found them to be of higher diversity. 51 different TRBV genes were found to be recurrent in the tumour repertoires of all patients, the same number as found in the non-cancerous control tissue. This highlights the need for large sample sizes when investigating such a highly variable sequence profile. While there were TRBV-TRBJ pairings found only in the tumour repertoires, there were also TRBV-TRBJ pairings found only in the control repertoire, suggesting that these tissue-specific gene pairings are more likely representative of the sampling of the extreme variability of the TCRβ sequence repertoire rather than a signature of CRC.

Although the exact number of T cells could not be absolutely quantified in this study, it was assumed that the DOC would be proportional to number of T cells expressing that sequence. Depending on the mucosal compartments (e.g. epithelium, lamina propria, GALT) included by chance in resected tissue samples, subsequently sectioned, and used in library construction for each patient, the numbers of T cells could have varied[53], contributing to the variation observed here. Lymphocyte presence in gut tissues can be influenced by many factors including gender, age, level of stress and infection status (such as the common cold)[53], and differing cancer treatment regimes, and may be contributing to the variation present in tissues across all patients. A caveat of this study however was that the depth of sequencing was not exhaustive. Thus, the overlap of distinct sequences seen between repertoires may change with deeper sequencing or even the sample size of patients. Also, by excluding those TCRβ sequences present with a DOC of one, I very likely removed real rare sequences from my CRC patient cohort. Due to the differential abundance of CDR3 sequences however,

it was impossible to tell the difference between a rare sequence and a sequence error, making this DOC cut-off a requirement.

To facilitate earlier detection of CRC, more highly specific, low invasive screening approaches are required. In this study I have shown that there may be specific differences in the tumour CDR3β sequence repertoire that differ when compared to control repertoires of 43 people with CRC. While deeper sequencing, increasing CRC cohort size and peripheral blood TCRβ profiling would be required to confirm this, if differences such as these could be validated, CDR3β sequences could be candidates for use as highly specific biomarkers in non-invasive screening approaches for CRC.

3. Shared CDR3β Sequences and Survival

3.1. Background

Colorectal carcinoma (CRC) often does not present symptoms until a patient is in the later stages of malignancy[33]. Combined with the highly invasive and low specificity of current screening approaches[33], the five-year survival rate of CRC is ~63%[31]. To facilitate earlier detection, and thus improve survival, more highly specific screening approaches are required.

It is well established that the presence of tumour infiltrating T cells (TIfTs) is associated with increased survival in multiple diseases, including CRC[34–39]. T cell receptors (TCRs) have been isolated from TIfTs, and the peripheral blood of CRC tumour bearing patients[45], but have been so far characterized at low resolution. Despite the incredible diversity present in the TCRβ CDR3 repertoire[4], multiple studies have identified TCRβ sequences shared between the peripheral blood repertoires of healthy individuals regardless of the level of human leukocyte antigen (HLA) matching[5,54]. The phenomenon of TCRβ sequence sharing between the repertoires of multiple individuals is referred to as a public T cell response[55], and has been recognised in a range of diseases, including malignancy (reviewed by Miles et al. 2011[20]). The identification of a shared CRC TCR biomarker would have significant implications for the design of highly specific screening, prognostic and immunotherapeutic approaches for CRC.

In this project, it was my aim to (1) determine if there was CDR3β sequence sharing present in CRC and (2) determine if any of the sequences shared between patients were associated with increased overall survival. Using Illumina NGS technology, I have previously characterized TCRβ sequence profiles from tumour and matched normal control tissues of 43 CRC patients (chapter 2 of this thesis). Due to the large number of unique sequences present in the complete tumour repertoire of these patients (>19,000 CDR3β nucleotide sequences (ntCDR3β), predicting almost 15,000 unique

CDR3β amino acid sequences (aaCDR3β)), an automated approach was required to mine for possible shared and survival associated sequences. To do this I utilized mainly bioinformatic approaches to mine through my data set in a fraction of the time of manual annotation. I hypothesized that there would be shared CDR3β sequences present in the repertoires of CRC patients and that some of these sequences would be associated with improved overall survival.

3.2. Materials and Methods

3.2.1. Data collection

TCRβ sequence data used in this chapter was derived from the same libraries as used in chapter 2 (see section 2.2, Appendix A, Fig. 2) for information on CDR3β sequence acquisition, processing, sequencing and quality filtering).

3.2.2. Identification of CDR3β sharing

I queried my data set using command line syntax and MySQL database mining (see section 2.2.4.2) to search for the presence of sequences shared between patient tumour tissues (see Appendix B for an example of a common bioinformatic query used). Sequences were considered shared if they were present with a DOC > 1 in at least 2 patient samples.

3.2.3. HLA types

HLA predictions were made by a computational pipeline (HLAminer[56]) that derived HLA allelic predictions by comparison of sequence reads to a comprehensive database of reference alleles from IMGT/HLA[2]. At the time of the writing of this thesis, 26/43 of the patients in my cohort had undergone HLA prediction, as part of a separate study[56]. Two of these patients (patients 49 and 75) had their predicted HLA types verified in the laboratory using PCR amplification of exons 2 and 3 from HLA-I A, B and C, followed by capillary sequencing (see Warren et al.[56] for a full description of methodology).

3.2.4. *CDR3β sequences associated with improved survival*

To identify any possible sequences associated with increased overall survival in CRC, I queried my data set for predicted aaCDR3β sequences that were shared and present with a DOC > 99 from the tumour (or control) tissues of least two patients. For each of these sequences, I constructed a Kaplan Meier (KM) survival plot using GraphPad Prism 4.0b software[57]. Briefly, KM analysis allowed for the estimation of survival over time, even when there was censored data present. In this study, censored data referred to those patients for whom the event of death had not occurred (i.e. those patients who were still alive or lost to follow-up at the time of data collection). The cohort of patients was broken up into two groups for survival analysis - those who had the sequence present and those who did not have the sequence present. The survival of these two groups was plotted over time, taking into account overall survival status (censored or deceased). In order to compare the distributions of the survival times for these two groups, the nonparametric Mantel Haenszel (log-rank) test was used.

To test the likelihood of selecting survival associated CDR3β sequences by chance, I developed a bootstrapping approach that utilized the Fisher's Exact Test (FET) (Appendix C). As with the log rank test, the FET determined if there was a significant association between the presence and absence of a CDR3β sequence with overall survival, but it was more appropriate for use when samples sizes in each group were small. Do to the large number of sequences that needed to be processed, I wrote a Perl script that preformed a bootstrapped FET test (Appendix C). This script utilized a CPAN (http://www.cpan.org/modules/index.html) module specific for a two tailed FET test. The input file for this script was a text file that contained information on patient identity, survival status and the entire repertoire of unique predicted aaCDR3β sequences from each patient's tumour tissue (Appendix C). The FET program re-organized the data so that for every sequence, a hypothetical 2 x 2 contingency table was created. This table organized the input data into four groups: 1. *Pa* - the number of patients alive *(a)* with a given sequence present *(P)* in their respective repertories, 2. *Pd* - the number of patients deceased *(d)* with the sequence present in their respective repertoires, 3. *Aa* - the number of patients alive with the given sequence absent in their respective repertories *(A)*, and 4. *Ad* - The number of patients deceased with sequence absent from their respective repertoires.

	Sequence present (P)	Sequence absent (A)
Patient alive (a)	Pa	Aa
Patient deceased (d)	Pd	Ad

Values from these four groups were then used to calculate a FET score *(p)* where

$$p = ((Pa + Aa)!\ (Pd + Ad)!\ (Pa + Pd)!\ (Aa + Ad)!) / ((P!)\ (A!)\ (a!)\ (d!)\ (N!))$$

(where 'N' was the total frequency of the table). The FET score calculated for a given sequence was then bootstrapped 1000 times to account for any possible sampling bias. The bootstrapped FET score (FET_{boot}) was generated by randomly assigning the 43 patients as alive (0) or deceased (1), and then re-calculating the FET score for that sequence 1000 times (FET score calculated in each round called FET_n). Each time FET_n was significant ($p < 0.05$) in a given round of the bootstrap (n), a value of 1 was added to a significance score (*Ss*). At the end of the 1000^{th} round of the bootstrap, the *Ss* was divided by 1000 to reveal the FET_{boot} for that sequence. Sequences with a $FET_{boot} < 0.05$ (i.e. in <= 49/1000 rounds of the bootstrap, sequence was associated survival ($FET_n < 0.05$)), were considered as having a significant difference in overall survival between the patient group having the sequence and the patient group not having the sequence in their tissues. This process was repeated for every unique aaCDR3 sequence contained in the input file (see Appendix C).

3.2.5. *Validation of shared sequences*

Given the possibility of cross-contamination during sequencing, sequences appearing to be shared needed to be validated. To do this, I preformed a quantitative real-time PCR (qPCR) using TaqMan primer/probes designed to specifically amplify CDR3β sequences of interest. Probes were designed to anneal to as much of the CDR3β as possible, while forward and reverse primers were designed to anneal as close to the probe as possible, overlapping the CDR3β boundary and including part of TRBV or TRBJ-genes (Fig. 9). TaqMan qPCR was done in triplicate using firstly the original cDNA and then primary indexed PCR product (that went into library construction). Reaction conditions were as follows: 1X final concentration ABI TaqMan

Universal Master Mix, 20X reaction assay (10mM of each forward and reverse primer, and 5 mM probe (see appendix C for sequences)) and PCR water in a 10 µl reaction volume. Amplification and detection of DNA were performed with the ABI 7900HT Sequence Detection System (SDS) using the following thermal cycling parameters: 2 minutes at 50°C, 10 minutes at 95°C, and 45 cycles of 15 seconds at 95°C and 1 minute at 60°C (automated settings for ABI SDS 2.4 software). The qPCR experiment was performed a second time using the same samples and methods as outlined above, for the purpose of replication.

3.3. Results

All data presented in this chapter represent in-frame sequences with only one associated TRBV-gene, and present at a sequencing depth of coverage (DOC) of greater than one.

3.3.1. CDR3β sequence sharing

3.3.1.1. CDR3β sequences were shared between the tumour repertoires of CRC patients

In total, 17% of the predicted unique aaCDR3β sequences present in the tumour repertoire were shared between patients. This was highly similar to what was observed for the control repertoire. The number of sequences shared however decreased as the number of people sharing a sequence increased. There were four predicted aaCDR3β sequences (present at varying DOC) shared between the tumour repertoires of all 43 CRC patients: 1) CASSLGQGAYEQFF, 2) CAISEGQKNIQYF, 3) CSARDRGAENTGELFF and 4) CASSPGGSGATQYF. There were no predicted aaCDR3β sequences were shared by the control repertoires of all patients.

3.3.1.2. The proportion of shared CDR3β sequences between tumour repertoires decreased with increasing DOC

On average, 47.2 +/- 19.8% of ntCDR3β sequences (50.6 +/- 21.1% of aaCDR3β sequences) present with a DOC > 1 were shared between the tumour tissues of at least two patients. When considering those sequences present at increased DOCs, the average proportion of TCRβ sequences shared between patient tumour tissues

decreased (Fig. 10a). The average proportion of shared sequences was highly similar to control tissues, where 47.4 +/- 23.8% of DOC>1 ntCDR3β sequences and 43.4 +/- 22.7% of DOC > 1 aaCDR3β sequences were shared between patients. For all DOC cut-offs, there was no significant difference in the average proportion of tumour sequences shared between patients, nor was there a significant difference in the average proportion of control sequences shared between patients.

3.3.1.3. The proportion of shared ntCDR3β sequences within a patient's own tumour and control repertoires decreased with increasing DOC

On average, 46.8 +/-22.4% of unique CDR3β nucleotide sequences (30.52 +/- 30.15 % of unique amino acid sequences) present at DOC >1 within the tumour repertoire of a patient was shared with the control repertoire of that same patient. This was similar to the control, where on average 46.6 +/-20.9% of unique nucleotide sequences (35.3 +/-30.8% of unique amino acid sequences) present with a DOC >1 within the control repertoire was shared with the tumour repertoire of that same patient (Fig. 10b). As with sharing between patients, the average proportion of CDR3β sequences shared within a patient decreased with increasing DOC (Fig. 10b). For all DOC cut-offs, there was no significant difference in the average proportion of tumour sequences shared with the control of a single patient, nor was there a significant difference in the average proportion of control sequences shared with the tumour of a single patient.

3.3.1.4. HLA type association with CDR3β sequence sharing inconclusive

The majority of patients were matched at at least one computationally predicted HLA allele (Table 6), but remain to have these alleles validated in vitro. For the two patients with in vitro validated HLA types, patients 75 and 49, allele B*08 was matched between them (Table 6). Despite only being matched at a single allele however, these patients shared multiple high DOC TCRβ/CDR3β sequences (data not shown). Until HLA validation has been completed for all 43 CRC patients, associations between sequence sharing and HLA type are inconclusive.

3.3.2. *CDR3β sequences associated with survival*

3.3.2.1. Abundant sequences shared between the tumour repertoires of patients were associated with increased overall survival

14 predicted aaCDR3β sequences were shared between at least two patient's tumour samples, 14 of which were present at a high DOC (DOC > 99) (Fig. 11). After log rank KM analysis, there were six sequences with a significant association with improved overall survival (p-value (p_{km}) < 0.05) when present in tissue. After 1000 rounds of bootstrapping in the FET program, two of these sequences were still significantly associated with survival (p-value (p_{FET}) and bootstrapped p-value (FET_{boot}) < 0.05): 1. CSAPNPSGLLYNEQFF (n = 28, p_{KM} = 0.0029, p_{FET} = 0.0241, FET_{boot} = 0.021) and 2.CSARAPDGNTGELFF (n = 30, p_{KM} = 0.0006, p_{FET} = 0.0225, FET_{boot} = 0.019) (Fig. 12). These two sequences were also present in patient control tissues (1. CSAPNPSGLLYNEQFF n = 29, 2. CSARAPDGNTGELFF n = 30), but at low DOC (DOC < 16), and their occurrence in control tissue was not significantly associated with survival (Fig. 12). For control tissues, there were four amino acid sequences present with a DOC > 99 from at least two patients, none of which were significantly associated with survival (data not shown).

To determine the likelihood that an aaCDR3β sequence could be associated with increased overall survival by chance, I randomly chose 20 predicted aaCDR3β sequences from tumour and 20 from control repertoires, and plotted their survival curves. I used a random number generator in R (http://www.r-project.org/) to choose the sequences. In both cases, 0/20 sequences were associated with increased overall survival (data not shown).

3.3.3. *Validation of shared sequences*

3.3.3.1. Shared/survival associated CDR3β sequence presence could not be validated in vitro

Using a specific TaqMan primer/probe assay, I found that the DOC limit of detection (LOD) using cDNA as template was somewhere between a DOC > 10,000 and a DOC = 20,000 (Fig. 13a). Given that there were no sequences present at a DOC > 10,000 and that were shared between patients, I was unable to use cDNA as template for my qPCR validation of survival associated sequences. Using indexed PCR product

as though template though, I found that I could reliably amplify sequences where the LOD was a DOC > 190 (Fig. 13b). I developed an assay for one of the two sequences (CSAPNPSGLLYNEQFF) statistically shown to be associated with increased overall survival. I was able to reliably detect the sequence (i.e. amplification signal appeared in at least 2/3 triplicates and could be repeated in a subsequent amplification) in the patients that shared it at a DOC > 190 (38T, 44T, 75T, 79T). The average Ct values for samples 38T, 44T, 75T, 79T were 19.20, 20.52, 20.06,15.87 respectively. The DOC of the sequence in the remaining 24 tissue samples that shared this sequence was less than the LOD, and attempts to validate it were unreliable (i.e. amplification signal appeared in less than 2/3 triplicates and/or could not be repeated in a subsequent amplification). One outlier was present in this data set; in sample 9N, the sequence reliably amplified with the strongest signal (average Ct = 16.57), despite only having it present at a DOC of 2 (Fig. 13b). Currently, I do not have an explanation for this observation, but given that other samples with this sequence at higher DOCs (e.g. sample 56T DOC = 6) did not amplify, all that I can say is that sequences with DOCs < 190 in my data set did not reliably validate using qPCR.

3.4. Discussion

In this study, I identified predicted aaCDR3β sequences from the tumour repertoires of multiple CRC patients, sometimes present at a relatively high DOC (DOC>99), that were associated with improved overall survival. I was able to validate the presence of one of these sequences in some patient samples using qPCR under certain conditions (validated from indexed PCR product at a LOD cut-off). Future investigation in this field however should focus on optimizing the assay for rare sequences from cDNA template, as I could not reliably validate sequence presence in cDNA. Additionally, it would be very valuable to further investigate the relationship between sequence DOC and detectability in the laboratory. While I focussed on overall survival, future studies could also investigate sequence associations with other survival types such as disease-free survival or metastasis-free survival.

Although I recovered TCR sequences from tumour tissue, because I did not do immunohistochemistry (IHC), I could not confirm TIfT infiltration into tumour samples.

IHC from frozen tissue would prove very valuable in future studies to confirm TIfT infiltration as well as to detect T cell-APC interaction[40] and T cell activation (CD69[36], CD107a[36], Ki67[34]) within the tumour environment. In future studies, it would also be highly beneficial to isolate active tumour infiltrating T cells from non-frozen and/or frozen[58] tissues and characterize CDR3β sequence diversity of TIfTs on a single-cell basis. Further, IHC could also give insight into T cell subtype(s) (e.g. CD4+, CD8+, CD25+, CD56+ etc...) activated in the tumour environment in response to antigen, allowing for further investigation of T cell subtype with survival.

The CDR3β is a highly variable sequence that if abundant in the tumour tissues of multiple CRC patients, could be indicative of a clonal expansion in response to a common CRC TA. The presence of shared CDR3β sequences however does not definitively indicate that the T cells of different individuals are responding to the same common epitope. It is has been found that that a single TCR can respond to multiple different antigens[1]. In order to gain knowledge about what potential CRC epitopes could look like, a fundamental first step would be to profile the entire TCR paratope (i.e. all of the TCRα and TCRβ CDRs) corresponding to a TCRβ sequence of interest from single cells. Even at that, predicting the correct quaternary structure that a TA responsive TCR would take would not be an easy task. If accomplished however, such knowledge would have great utility in the design of immunotherapeutic tools targeting CRC TAs as well as highly specific bio-indicators and prognostic indicators of the disease in patients.

4. General conclusions and future directions

4.1.1. TCRβ sequence repertoire in CRC patients

Earlier detection and improved treatment strategies are keys to improving the survival rate of patients with colorectal carcinoma (CRC). TCRβ/CDR3β sequences that have an association with CRC have great potential utility for use as highly specific biomarkers of CRC. Their presence in peripheral blood could facilitate earlier detection of the disease, as well as an indication of T cell response to the tumour. In my thesis, through high resolution profiling, I have been able to characterize the TCRβ sequence repertoire from CRC tissue at high resolution (Chapter 2) and most interestingly, have found CDR3β sequences that may be associated with improved survival for CRC patients (Chapter 3). While this finding could have great utility in the detection and/or the treatment of CRC, the results presented here are preliminary, and require future investigation before their association with survival could be confirmed. Regardless, sequence profiling at this resolution has allowed me to capture what are likely rare clonotypes (in addition to the abundant ones), vastly expanding the known TCRβ repertoire present in CRC tumour tissues. Further, NGS profiling has allowed me to capture and compare tumour and control TCRβ profiles from 43 CRC patients at once, a task that would otherwise have been incredibly time consuming and expensive using labour intensive techniques such as spectratyping (chapter 1).

The abundance of data generated using the methodology outlined in chapter 2 allowed me to consider aspects of TCRβ /CDR3β sequence diversity and survival for a cohort of CRC patients. To gain further insight into the survival of CRC patients, a multivariate analysis (such as a Cox proportional hazards regression) should be done to determine if there were possible clinical covariates contributing to or possibly even biasing the presence of a predicted CDR3β sequence association with survival (e.g. stage of cancer, presence of metastasis, location of primary tumour etc...). As mentioned in chapter two however, the dataset presented here does not represent

exhaustive sequencing, and I would expect that with deeper sequencing some of the instances of sequence sharing sequencing could change. Other factors possibly affecting repertoire diversity and sequence sharing include sequencing error - while the parameters I used to filter the data were quite stringent, it is possible that some high quality sequence errors could have made it into my database. This fact highlights the need for sequence validation in the laboratory. The technology behind NGS is continually improving however, reducing the occurrence of sequencing errors. Further, bioinformatic tools for immunoprofiling are constantly being developed and fine-tuned. Therefore I feel confident in the use of NGS technology for immunological profiling in such a way as was done here. Some improvements I would consider for future work in this field would be (1) the use of longer indexes in primers (possibly decamers) to decrease the likelihood that multiple sequencing errors could occur in the barcode causing a mis-assignment to a patient's repertoire; (2) more clean-up steps between stages of library construction to remove possible indexing primer contaminants; and (3) spiking in a certain amount of T cells from a known TCRβ sequence clonotype at the beginning of library construction to give a better idea about the amount of sequence loss incurred throughout the process, as well as a better idea of how DOC correlates to T cell number.

I was able to find two predicted aaCDR3β sequences that appeared to be expanded (i.e. present at a DOC > 99) from the tumour repertoires of multiple CRC patients that were significantly associated with survival. To my knowledge, this was the first study to use next generation sequencing (NGS) techniques to profile the CDR3β repertoire in CRC and associate sequences with survival. Therefore, work needs to be done before confirming that these exact sequences are actually prognostic factors or could be used as indicators of CRC. Future studies should focus on (1) increasing the sample size of CRC patients used in analysis, (2) sequencing repertoires more deeply, (3) validating sequence results on multiple platforms (i.e. capillary sequencing in addition to NGS), (4) sequencing profiling peripheral blood of CRC tumour bearing patients to confirm sequence presence and comparing this to the peripheral blood of healthy patients, (5) considering other clinical factors that could be confounding any survival association that is present.

4.1.2. *Other T cell considerations*

4.1.2.1. The TCRγδ repertoire

To more completely characterize a possible specific T cell response to CRC, the γδ T cell subset should be investigated (e.g. through NGS profiling, qPCR, IHC etc...) in addition to the αβ T cell population. It is known that there are a sizable proportion of γδ T cells present in the colorectal epithelium of mice. While their function is still not well defined (reviewed in Chien et al. 2006[59]), clinical studies have shown that γδ T cells can infiltrate in the tumours of multiple cancers (reviewed in Lamb, 2012[60]) and have been shown to have anti-tumour effects in CRC[61]. Paradoxically however, other studies have found γδ T cells to have tumour-promoting effects in the body[62], highlighting the need to more definitively characterize their roles in cancer. γδ T cells only make up ~1-5% of the circulating lymphocytes in peripheral blood of adults however (as stated in Chien et al. 2006[59]), so while their role in cancer surveillance could be important, their use in non-invasive screening approaches less desirable than the αβ T cell population.

4.1.2.2. TCRα

In this study I have characterized the TCRβ at sequence level resolution and found some TCR CDR3βs that are associated with improved survival. In order to have utility as a tool targeting CRC TAs however, knowledge of the TCRα paired with the TCRβ of interest is required. As discussed in the introduction chapter, the TCR is a dimeric receptor, contacting the peptide epitope using both the CDR3 of the α chain and of the β chain of the TCR. It is vital that future investigation in this field includes profiling and looking for survival associations of paired TCRα-TCRβ sequences. This task however is not a simple one, as it requires linking paired TCRα-TCRβ sequences from single T cells in vitro. More fully characterizing TCR paratopes could have utility in modeling the identity of possible tumour antigens of CRC tumours that could be targeted in immunotherapy.

4.1.3. *CRC tumour antigens*

A major goal of cancer immunotherapy is the activation of T cell responses against specific TAs. T cell antigen discovery has proved difficult however because of several features of TCR molecules[58]: (1), The affinities of TCRs to pMHC complexes

are generally low, meaning that techniques requiring high affinity binding could be precluded[58]. (2) Most TCRs are polyspecific[58], meaning they have the ability to recognise and be activated by multiple distinct pMHC ligands. Third, in contrast to antibodies that bind their ligands directly, TCRs recognise antigenic pMHC complexes generated by a complex intracellular antigen-processing [63]. The processing pattern of the APC used in vitro must be identical to that of the original APCs of the target tissue to result in the same peptide[64]. Screening cDNA libraries have helped to identify TAs[65,66] but their use is limited because they must contain full-length in-frame cDNA, that should originate from affected tissue, which is not always available, and correct antigen processing is required[58]. Another method uses randomized peptide libraries displayed by insect cells, which are screened by oligomerized soluble TCR molecules[67]. This approach is limited however by the notoriously low affinities of TCRs to their MHC-peptide ligands and by complicated library cloning procedures[58].

Recently, Siewert and colleges (2012) have developed an approach for the specific identification of target antigens of $CD8^+$ T cells[58]. Briefly, this technology used T hybridoma cells transfected with CD8α, CD8β, super-GFP (sGFP) under the control of the nuclear factor of activated T cells (NFAT) and TCRα and TCRβ chains originating from an influenza specific public TCR cell line. These transfections led to the expression of TCRα and TCRβ chains together with CD3, CD8αβ molecules on the surfaces of the hybridomas. APCs were transfected with an appropriate HLA molecule and with a plasmid-encoded combinatorial peptide library. This led to the expression of HLA molecules that had bound short antigenic peptides (pMHCs) on the surfaces of APCs. Transfected pMHC expressing APCs were grown in a tissue culture dish and overlaid with TCR transfectants. TCR-transfected hybridoma cells that were in direct contact with a peptide-presenting APC became activated and fluoresced green due to synthesis of sGFP by NFAT. After isolating the T hybridoma cell in complex with the APC presenting peptide, the plasmid coding for the peptide was then cloned and sequenced[58]. The identification of several peptides with converging motifs facilitated the detection of the parent peptide antigen. This approach is highly specific, as the authors found that single APCs carrying the specific pMHCs could be detected from millions of APCs with irrelevant peptides. Further, the use of short peptides would bypass many of the complexities of intracellular antigen processing[58]. Siewert and colleagues[58] go on to

suggest that their method would be suited for use with single T cells isolated by laser microdissection from frozen tissue specimens[58]. They outline a potential strategy as (1) isolating single T cells directly from tissue, (2) cloning and expressing their paired TCR α and β chains and (3) characterizing their antigens using the approach described in their study[58] (above). While their approach is highly targeted, it requires prior knowledge of TCR clonotypes of interest, which are not always known. The high-resolution approach employed in my thesis could be used to help identify TCRβs of interest, which could then be used as tools in CRC TA discovery.

Identifying CRC TCR clonotypes of interest for use as biomarkers in disease detection, tools to target tumours, or prognostic indicators of survival, will likely require a combination of molecular approaches. My thesis research has shown that high resolution profiling approaches can successfully be used to characterize the hypervariable TCRβ sequence repertoire with possible implications for improving the survivability of a devastating disease like CRC. While my thesis focused on CRC in particular, the methodology employed here is highly transferable, and could lead to advancements of biological tools and survivability in many diseases.

5. Tables

Table 1. Sequence counts and depth of coverage similar in CRC tumour and control tissues

	No filters	In-frame	1 V -gene	DOC>1	In-frame, 1-Vgene	In-frame, 1-Vgene, DOC > 1 (presented in thesis)
Tumour						
DOC	10,145,921	10,039,268	8,706,608	10,035,973	8,626,330	8,547,637
TCRβ sequences	209,358	194,628	169,809	99,410	158,190	79,497
Unique TCRβ sequences	87,908	77,360	73,572	32,350	65,112	25,464
Unique ntCDR3β	73,648	64,127	63,437	23,924	55,668	19,532
Unique aaCDR3β	51,468	46,926	44,308	17,216	40,842	14,727
Control						
DOC	9,552,942	9,324,708	8,050,887	9,444,426	7,867,162	7,790,030
TCRβ sequences	205,434	190,079	166,220	96,918	154,076	76,944
Unique TCRβ sequences	88,014	76,912	73,587	32,128	64,642	25,010
Unique ntCDR3β	73,428	63,564	63,131	23,449	55,025	18,882
Unique aaCDR3β	51,081	46,481	43,937	16,885	40,414	14,266

Table compares the counts of sequences present in collectively in tumour (top) and control (bottom) repertoires after processing in the CDR3β pipeline. Each column shows sequence data present for a given filter. Far right column shows data presented in this thesis. Filters include: **No filters**: all sequences processed through TCRβ pipeline; **In-frame**: count of all sequences processed through the TCRβ pipeline, in the correct reading frame; **1 V-gene**: count of all sequences processed through the TCRβ pipeline with one an unambiguously identified TRBV-gene; **DOC>1**: count of all sequences processed through the TCRβ pipeline present at a depth of coverage of greater than one; **In-frame, 1-Vgene:** count of all sequences processed through the TCRβ pipeline that were in the correct reading frame and with one unambiguously identified TRBV-gene; **In-frame, 1-Vgene, DOC > 1:** count of all sequences processed through the TCRβ pipeline that were in the correct reading frame, with one unambiguously identified TRBV-gene and were found at a depth of greater than one. DOC – depth of coverage; TCRβ sequences - nucleotide sequences in the dataset encoding the CDR3β, partial TRBV-gene and full TRBJ-gene, taking into account bases added and deleted at gene junctions; CDR3β sequences - sequences present in the dataset encoding the hypervariable CDR3β the site of antigen contact; ntCDR3β – nucleotide sequences encoding the CDR3β; aaCDR3β – amino acid sequences that make up the CDR3β.

Table 2. Average sequence counts and depths of coverage similar in CRC tumour and control tissues

	Tumour					Control				
Patient	DOC	TCRβ sequences	Unique TCRβ sequences	Unique ntCDR3β	Unique aaCDR3β	DOC	TCRβ sequences	Unique TCRβ sequences	Unique ntCDR3β	Unique aaCDR3β
22	737,572	4,951	4,464	3,879	3,449	28,236	401	378	338	322
14	665,835	3,836	3,519	3,201	2,912	72,585	859	808	776	730
56	466,115	3,529	3,201	2,907	2,593	567,066	4,317	3,956	3,346	3,014
69	463,278	3,687	3,398	2,999	2,713	133,175	1,428	1,338	1,249	1,167
06	431,995	3,401	3,157	2,777	2,527	30,549	363	339	327	305
34	400,198	3,035	2,805	2,507	2,322	215,582	2,455	2,324	2,130	1,983
52	353,446	3,112	2,889	2,651	2,449	33,582	425	399	347	327
33	330,842	3,024	2,816	2,513	2,333	278,483	2,715	2,531	2,352	2,171
43	321,216	2,981	2,764	2,416	2,228	126,323	1,342	1,251	1,166	1,091
05	306,508	2,703	2,499	2,309	2,128	214,761	1,928	1,785	1,707	1,581
45	299,910	2,614	2,424	2,327	2,139	667,669	4,736	4,290	3,664	3,258
03	288,437	2,795	2,592	2,317	2,154	65,125	946	898	823	800
32	235,390	2,492	2,321	2,067	1,916	382,464	2,950	2,706	2,446	2,234
28	230,078	2,268	2,093	1,912	1,778	107,166	1,266	1,201	1,102	1,035
37	224,026	2,579	2,427	2,060	1,936	78,849	1,107	1,063	998	961
24	204,988	2,093	1,946	1,811	1,659	164,284	2,026	1,913	1,783	1,676
63	187,897	2,002	1,890	1,703	1,581	204,114	2,268	2,135	1,929	1,790
79	162,992	1,847	1,726	1,549	1,459	120,554	1,463	1,384	1,344	1,255
50	160,012	1,773	1,663	1,564	1,441	27,876	443	422	398	381
75	145,713	1,569	1,453	1,390	1,306	311,127	2,802	2,577	2,338	2,147
38	132,970	1,529	1,432	1,366	1,268	269,627	2,593	2,412	2,135	1,951
10	130,837	1,499	1,401	1,286	1,196	135,879	1,737	1,670	1,531	1,449
31	119,193	1,384	1,313	1,222	1,155	215,466	2,541	2,380	2,193	2,053
47	116,480	1,104	1,009	991	922	4	2	2	2	2
30	112,157	1,174	1,091	1,031	959	51,625	612	570	516	484
09	111,844	1,586	1,515	1,531	1,466	195,403	1,796	1,697	1,626	1,525
21	107,862	1,397	1,327	1,264	1,191	7,250	100	97	86	84
49	103,112	1,132	1,066	971	920	155,218	1,768	1,668	1,523	1,438

	Tumour					Control				
Patient	DOC	TCRβ sequences	Unique TCRβ sequences	Unique ntCDR3β	Unique aaCDR3β	DOC	TCRβ sequences	Unique TCRβ sequences	Unique ntCDR3β	Unique aaCDR3β
35	100,401	1,213	1,151	1,090	1,023	80,320	1,080	1,014	962	912
36	95,877	1,157	1,096	991	942	67,080	956	926	841	799
44	92,728	1,218	1,147	1,060	1,005	179,740	2,089	1,957	1,769	1,667
71	84,248	1,033	989	893	851	203,645	2,284	2,129	1,891	1,758
11	82,984	966	913	860	821	122,358	1,758	1,684	1,528	1,449
74	82,628	847	800	742	713	584,753	4,383	3,960	3,470	3,111
04	79,729	1,026	977	924	879	22,183	469	453	444	427
02	78,489	1,029	970	899	855	168,584	2,063	1,950	1,868	1,765
12	73,139	968	908	819	781	57,631	876	837	815	778
08	59,259	853	810	781	746	153,679	1,761	1,663	1,478	1,401
78	49,285	504	477	440	420	102,787	1,498	1,406	1,214	1,127
59	46,941	511	482	444	417	169,367	1,701	1,589	1,482	1,385
07	40,114	565	543	532	503	622,667	4,413	3,996	3,486	3,136
40	28,449	385	356	318	302	152,258	1,716	1,618	1,498	1,402
01	2,463	126	125	126	126	242,936	2,508	2,357	2,097	1,953
Mean +/- stdev	198,782.26 +/- 166,751.42	1,848.77 +/- 1,099.68	1,719.65 +/- 1,001.32	1,568.37 +/- 878.92	1,453.12 +/- 788.58	181,163.49 +/- 163,755.31	1,789.40 +/- 1,159.83	1,668.21 +/- 1,053.45	1,512.05 +/- 913.12	1,401.95 +/- 819.80

Table 3. Average proportions of ntCDR3β sequences found only in CRC tumour tissue similar to proportions found only in control tissues

	Tumour			Control		
Patient	Unique ntCDR3β sequences	Tumour only ntCDR3β	Proportion of ntCDR3β only in tumour	Unique ntCDR3β sequences	Control only ntCDR3β	Proportion of ntCDR3β only in control
1	126	23	0.18	2097	1994	0.95
2	899	271	0.30	1868	1240	0.66
3	2317	1683	0.73	823	189	0.23
4	924	695	0.75	444	215	0.48
5	2309	1284	0.56	1707	682	0.40
6	2777	2541	0.92	327	91	0.28
7	532	100	0.19	3486	3054	0.88
8	781	781	1.00	1478	1478	1.00
9	1531	666	0.44	1626	761	0.47
10	1286	573	0.45	1531	207	0.14
11	860	354	0.41	1528	1022	0.67
12	819	511	0.62	815	507	0.62
14	3201	2597	0.81	776	172	0.22
21	1264	1219	0.96	86	41	0.48
22	3879	3602	0.93	338	61	0.18
24	1811	854	0.47	1783	826	0.46
28	1912	1249	0.65	1102	439	0.40
30	1031	804	0.78	516	289	0.56
31	1222	354	0.29	2193	1325	0.60
32	2067	782	0.38	2446	1161	0.48
33	2513	1067	0.43	2352	906	0.39
34	2507	1071	0.43	2130	694	0.33
35	1090	650	0.60	962	522	0.54
36	991	620	0.63	841	470	0.56
37	2060	1341	0.65	998	279	0.28
38	1366	519	0.38	2135	1288	0.60
40	318	135	0.43	1498	1315	0.88
43	2416	1630	0.68	1166	380	0.33
44	1060	398	0.38	1769	1107	0.63
45	2327	680	0.29	3664	2017	0.55
47	991	991	1.00	2	2	1.00
49	971	394	0.41	1523	946	0.62
50	1564	1318	0.84	398	152	0.38
52	2651	2391	0.90	347	87	0.25
56	2907	1127	0.39	3346	1566	0.47
59	444	200	0.45	1482	1238	0.84

	Tumour			Control		
Patient	Unique ntCDR3β sequences	Tumour only ntCDR3β	Proportion of ntCDR3β only in tumour	Unique ntCDR3β sequences	Control only ntCDR3β	Proportion of ntCDR3β only in control
63	1703	666	0.39	1929	892	0.46
69	2999	2087	0.70	1249	337	0.27
71	893	309	0.35	1891	1307	0.69
74	742	156	0.21	3470	2884	0.83
75	1390	496	0.36	2338	1444	0.62
78	440	195	0.44	1214	969	0.80
79	1549	834	0.54	1344	629	0.47
μ +/- σ	1568.37 +/- 878.92	935.30 +/- 768.28	0.55 +/- 0.23	1512.05 +/- 913.12	864.77 +/- 713.97	0.53 +/- 0.23

Bottom row represents mean (μ) +/- standard deviation (σ).

Table 4. ***Consensus CDR3β amino acid sequence composition of most frequently observed length (14 residues) in tumour and control repertoires***

	Tumour	Control
Complete repertoire, DOC > 1	CASSLGGGGNEQFF	CASSLGGGTNEQYF
Complete repertoire, DOC > 99	CASSLGGGGNEQFF	CASSLGGGTNEQYF
Tissue only repertoire, DOC > 1	CASSLGGGGNEQFF	CASSLGGGTNEQYF
Tissue only repertoire, DOC > 99	CASSLGGGGNEQFF	CASSLGGGTNEQYF

Table 5. TRBV/J-gene pairings found only in CRC tumour and only in control tissues

5a. Tumour only repertoire TRBV/J gene pairings

	vName	jName	#	ntCDR3β	DOC	Patient
1	TRBV7-7	TRBJ1-2	1	TGTGCCAGCAGCTTAtcgacagggccggACTATGGCTACACCTTC	2	22
			2	TGTGCCAGCAGTTCAGgggcacgaGGCTACACCTTC	2	14
			3	TGTGCCAGCAGTTcctctctgggggACTATGGCTACACCTTC	2	36
			4	TGTGCCAGCAtaccacaggggcatttcggGGCTACACCTTC	2	32
			5	TGTGCTAGCAGCTTAGaacccgggcagAACTATGGCTACACCTTC	7	43
2	TRBV7-7	TRBJ1-5	1	TGTGCCAGCACCacttaTAGCAATCAGCCCCAGCATTTT	2	03
			2	TGTGCCAGCAGCTTAGCgcaggggAATCAGCCCCAGCATTTT	2	30
			3	TGTGCCAGCAGttacggaacaggggcggcgAGCCCCAGCATTTT	2	32
3	TRBV23-1	TRBJ2-4	1	TGCGCCAGCAGTtccctactagcgggaggggAAACATTCAGTACTTC	4, 2	75 , 79
			2	TGCGCCAGCAGTCcccttatgggcttaaCCAAAAACATTCAGTACTTC	4	22
4	TRBV6-8	TRBJ2-5	1	TGTGCCACCAGcagagataggccagggagtCAAGAGACCCAGTACTTC	6	06
			2	TGTGCCACCAGcagagatcctggAGAGACCCAGTACTTC	3	56
5	TRBV5-3	TRBJ1-2	1	TGTGCCAGccaacgaccggactctTATGGCTACACCTTC	5	32
			2	TGTGCCAGTAGCTTtaaaggtccatgggATGGCTACACCTTC	2	47
6	TRBV6-9	TRBJ1-4	1	TGTGCCAGCAGTCCTTggacaggtgATGAAAAACTGTTTTTT	187	78
			2	TGTGCCCGCAGTCCTTggacaggtgATGAAAAACTGTTTTTT	2	78
7	TRBV5-7	TRBJ2-1	1	TGTGCCAGCAGatccactagcggggCCTACAATGAGCAGTTCTTC	10	14
			2	TGTGCCAGCAGCTTGGttgggggcttagatAATGAGCAGTTCTTC	4	14
8	TRBV12-5	TRBJ1-5	1	TGTGCTAGGGGaaaTAGCAATCAGCCCCAGCATTTT	9	33
9	TRBV5-7	TRBJ2-2	1	TGTGCCAGCAGCTTGGtagtCGGGGAGCTGTTTTTT	9	02
10	TRBV5-8	TRBJ1-3	1	TGTGCCAGCAGCTTaaatgtggcagtgggCACCATATATTTT	7	28
11	TRBV5-3	TRBJ2-3	1	TGTGCCAGcgctacaggggtagGCACAGATACGCAGTATTTT	6	02
12	TRBV6-8	TRBJ1-2	1	TGTGCCACCAGcagatccccgacaggtcctagggaaGGCTACACCTTC	6	69
13	TRBV5-3	TRBJ2-6	1	TGTGCCAGTACGTTccgggCTGGGGCCAACGTCCTGACTTTC	5	06
14	TRBV6-8	TRBJ1-4	1	TGTGCCACCAGcagaggagcggggacagggAATGAAAAACTGTTTTTT	4	05
15	TRBV17	TRBJ2-3	1	TGCAGCGGTGGAGtaAGCACAGATACGCAGTATTTT	3	56
16	TRBV23-1	TRBJ1-4	1	TGCGCCAGCAGTgctgggacaggggacgTGAAAAACTGTTTTTT	3	32
17	TRBV5-3	TRBJ1-4	1	TGTGCCTGGAatcggggagtcAATGAAAAACTGTTTTTT	3	69
18	TRBV16	TRBJ1-4	1	TGTGCCAGCAGggacaggattttcGAAAAACTGTTTTTT	2	37
19	TRBV16	TRBJ1-5	1	TGTGCCATCAGtgagtcggaagggatttcTCAGCCCCAGCATTTT	2	22
20	TRBV16	TRBJ2-4	1	TGCGCCAGCAGCttgagcgagGCCAAAAACATTCAGTACTTC	2	56
21	TRBV6-8	TRBJ1-1	1	TGTGCCAGCAGTTtcatcccggccggTGAAGCTTTCTTT	2	33
22	TRBV7-7	TRBJ2-6	1	TGTGCTAGCccgggacagggaacCTCTGGGGCCAACGTCCTGACTTTC	2	21
23	TRBV10-1	TRBJ2-6	1	TGCGCCAGCAGcgcaccaactatcggcaCTGGGGCCAACGTCCTGACTTTC	2	43

5b. Control only repertoire TRBV/J gene pairings						
	vName	jName	#	ntCDR3β sequence	DOC	Patient
1	TRBV6-8	TRBJ2-7	1	TGTGCCACCAGcagaggaaccgggactagcgggagagagtatGAGCAGTACTTC	3	38
			2	TGTGCCACCAGcagcccctccgggacacggTCCTACGAGCAGTACTTC	4	08
			3	TGTGCCACCAGccctgggacagggctccACGAGCAGTACTTC	7	45
			4	TGTGCCAGCAGTTAtgggggactagcgggagggcccccCGAGCAGTACTTC	2	07
2	TRBV18	TRBJ1-3	1	TGTGCCAGCaagttggacagCTCTGGAAACACCATATATTTT	2	32
			2	TGTGCCAGCagcttcgacggcacCTCTGGAAACACCATATATTTT	2	69
			3	TGTGCCAGCagtcttccagggacagtCTCTGGAAACACCATATATTTT	6	45
			4	TGTGCCAGCagttacgtggaaccccggAACACCATATATTTT	2	10
3	TRBV5-3	TRBJ1-1	1	TGTGCCAGCAGCCTcagggtgggCACTGAAGCTTTCTTT	2	32
			2	TGTGCCAGTAGtatacccagGAACACTGAAGCTTTCTTT	4	07
			3	TGTGCCTGGAGaggaaagggagggtgGAACACTGAAGCTTTCTTT	2	05
4	TRBV23-1	TRBJ2-3	1	TGCGCCAGCAGTCcccaggggacagggaccgacttgaCACAGATACGCAGTATTTT	2	74
			2	TGCGCCAGCAGTCcccaggggacagggaccgacttgGCACAGATACGCAGTATTTT	2, 445	45, 74
5	TRBV6-4	TRBJ1-3	1	TGTGCCAGCACGGACcggacgttcGGAAACACCATATATTTT	6	74
			2	TGTGCCAGCAGCGTCTCgacccgtctCTCTGGAAACACCATATATTTT	7	43
6	TRBV21-1	TRBJ2-6	1	TGTGCCAGCAGCAAtgaacggggacagggctttcgaGGGGCCAACGTCCTGACTTTC	6	37
			2	TGTGCCAGCAGCACAGacagggttaCTGGGGCCAACGTCCTGACTTTC	2	02
7	TRBV21-1	TRBJ1-3	1	TGTGCCAGCAGttacgtggaaccccggAACACCATATATTTT	3	10
			2	TGTGCCAGTACCAAAagggatttCTCTGGAAACACCATATATTTT	2	07
8	TRBV15	TRBJ1-3	1	TGCGCCAGCAcggaccggacgttcGGAAACACCATATATTTT	2	74
			2	TGTGCCACCAGCAGAGttgggggtacaaaCTCTGGAAACACCATATATTTT	16	01
9	TRBV6-8	TRBJ2-3	1	TGTGCCACCAGcagagatcctccgggacaggggACAGATACGCAGTATTTT	2	56
			2	TGTGCCACCAGcggtgacagggggtaCACAGATACGCAGTATTTT	6	11
			3	TGTGCCAGCAacaacaggggtggAGCACAGATACGCAGTATTTT	2	56
10	TRBV16	TRBJ1-1	1	TGTGCCAGCAGCcggggatCTGAAGCTTTCTTT	2	38
			2	TGTGCCAGCAGgcgacaggggaTGAACACTGAAGCTTTCTTT	3	45
11	TRBV17	TRBJ1-1	1	TGCAGCAtttccgggactttgcgCACTGAAGCTTTCTTT	2	59
			2	TGCAGCGTAGGGGGGAACACTGAAGCTTTCTTT	4	75
12	TRBV5-6	TRBJ1-3	1	TGTGCCAGCAGactcggcgggacagCTGGAAACACCATATATTTT	2	34
			2	TGTGCCAGCAGtcttccagggacagtCTCTGGAAACACCATATATTTT	2	45
13	TRBV7-6	TRBJ2-4	1	TGTGCCAGCAGCccaaacagggACATTCAGTACTTC	2	45
			2	TGTGCCAGCAGttatagccgggacagagtaAAAAACATTCAGTACTTC	2	08
14	TRBV7-1	TRBJ1-3	1	TGTGCCAGCAGCCCCGCttgggccgccGGAAACACCATATATTTT	17	56
15	TRBV6-8	TRBJ1-5	1	TGTGCCACCAGCATCTCagacagTAGCAATCAGCCCCAGCATTTT	16	45
16	TRBV5-3	TRBJ1-5	1	TGTGCCAGTAGgaaccgaggaataagctcTCAGCCCCAGCATTTT	4	34
17	TRBV6-7	TRBJ2-4	1	TGTGCCAGCAGTTtcgagggacagctAGCCAAAAACATTCAGTACTTC	4	74
18	TRBV23-1	TRBJ2-6	1	TGCGCCAGCAGTCtgacagaactCTCTGGGGCCAACGTCCTGACTTTC	3	49
19	TRBV5-7	TRBJ1-2	1	TGTGCCAGCAGCTTGGatgcagggttgTATGGCTACACCTTC	3	40

5b. Control only repertoire TRBV/J gene pairings						
	vName	jName	#	ntCDR3β sequence	DOC	Patient
20	TRBV15	TRBJ1-6	1	TGTGCCAGCAGTGGCGAcagggtaCTCCACTTT	2	31
21	TRBV15	TRBJ2-6	1	TGTGCCAGCAGCctcgttagcGGGGCCAACGTCCTGACTTTC	2	31
22	TRBV16	TRBJ2-2	1	TGTGCCAGCAGtcagggggggcctACACCGGGGAGCTGTTTTTT	2	56
23	TRBV16	TRBJ2-3	1	TGTGCCAGCAGggggcagacCACAGATACGCAGTATTTT	2	34
24	TRBV17	TRBJ2-1	1	TGCAGCGGCGGTGGgggcggtaagAATGAGCAGTTCTTC	2	01
25	TRBV5-7	TRBJ2-4	1	TGTGCCAGCAGCTTtcaggagGCCAAAAACATTCAGTACTTC	2	45
26	TRBV6-7	TRBJ1-3	1	TGTGCCAGCAGTTACgtggaaccccggAACACCATATATTTT	2	10
27	TRBV6-8	TRBJ2-4	1	TGTGCCACCAGcggggactccgtgGCCAAAAACATTCAGTACTTC	2	56
28	TRBV7-7	TRBJ1-6	1	TGTGCCAGCAGCcgacagggttatgATTCACCCCTCCACTTT	2	56
29	TRBV12-5	TRBJ1-2	1	TGCGCCAGTGCTCGGGgcggTAACTATGGCTACACCTTC	2	07

TRBV/J-gene pairings for TCRβ sequences found only in tumour tissues (23/591 pairings) (table 5a) and those that are found only in control tissue tissues (29/597 pairings) (table 5b). # refers to ntCDR3β sequence associated with TRBV/J gene pairing. ntCDR3β sequence – Uppercase refers letters to bases coding TRBV gene (left), and TRBJ gene (right). Lower case letters represent non-TRBV, non-TRBJ bases. They comprise a mixture of TRBD and/or non-templated bases added back by the terminal deoxynucleotidyl transferase (TdT). DOC = depth of coverage.

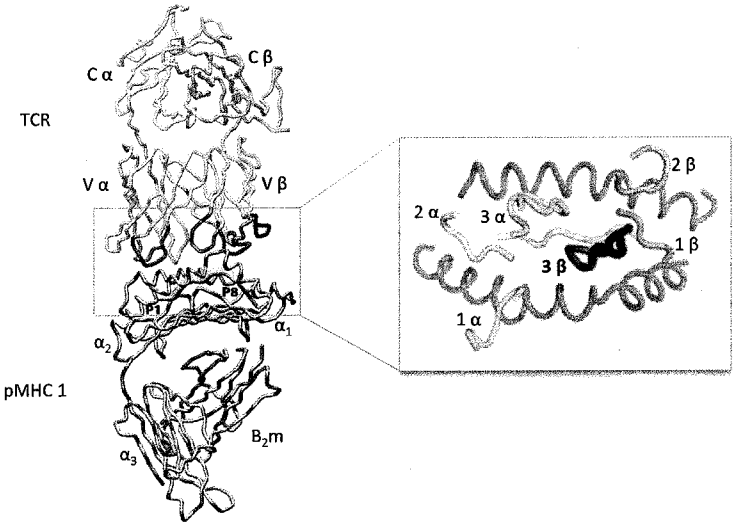
Table 6. HLA prediction for 23/43 CRC patients

Patient	Predicted HLA-1 alleles										
75	A*02	A*03	**B*39**	**B*08**	C*07	C*05				E*01	F*01
49	A*36	**A*01**	**B*08**	**B*57**	**C*06**	**C*07**					
14	A*24	A*03	B*44	B*14	C*02	C*08				E*01	
22	A*03	A*02	B*35	C*04							
28	A*02	B*51									
30	A*24	A*02	B*40	C*01	C*07	C*03	C*06			E*01	
31	A*24	B*15	B*57	B*08	C*01	C*07	C*03				
32	A*02	B*07	C*07								
33	A*01	B*08	B*56	C*01	C*07					E*01	
34	A*02	B*15	B*44	C*03	C*05	C*08				E*01	
35	A*01	A*03	B*08	C*07							
36	A*25	A*02	A*26	B*44	B*57	C*06	C*05				
37	A*02	A*24	B*46	B15	B*40	C*07	C*01			E*01	
40	A*32	A*02	B*27	B*51	C*01	C*02					F*01
43	A*02	A*36	A*01	B*44	B*49	C*07	C*05			E*01	F*01
44	A*34	B*08	B*14	C*05	C*08	C*07					
47	A*31	B*40	C*03								
50	A*24	B*40	B*35	B*41	B*42	B*08	C*16	C*15	C*02		
52	A*68	A*02	B*15	B*44	C*07	C*03				E*01	
56	A*68	A*24	A*02	B*15	C*03	C*01				E*01	F*01
59	A*26	A*01	B*51	B*08	C*07	C*01				E*01	F*01
63	A*29	A*01	B*08	B*40	C*03	C*07				E*01	F*01
69	A*02	B*27	B*37	B*07	B*55	B*56	C*02			E*01	
71	A*32	A*02	B*44	C*05						E*01	
74	A*33	A*03	B*14	B*51	B*52	B*35	C*08	C*02	C*05	E*01	F*01
79	A*03	A*31	A*33	A*02	B*14	C*08				E*01	F*01

HLA-1 alleles predicted using HLAmine[56] for 26/43 CRC patients in my cohort. Bolded alleles represent PCR and capillary sequencing verified alleles. Red represents match at a single HLA-B allele (B*08) between two patients (patients 49 and 75).

6. Figures

1a.



1b.

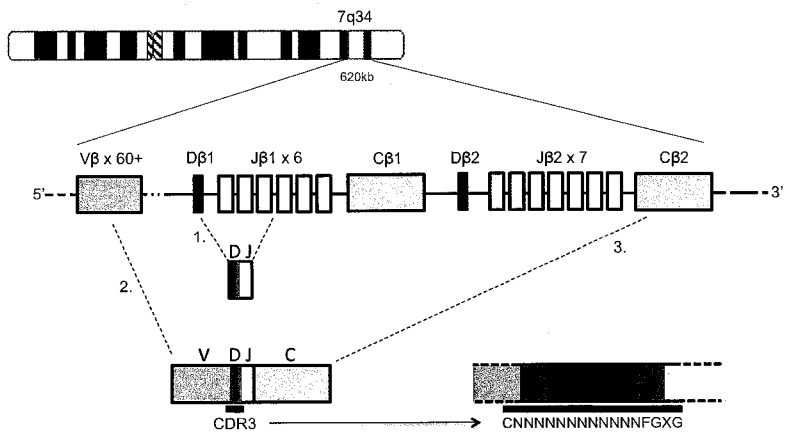


Figure 1. TCRαβ pMHC 1 interaction and formation of a functional TCRβ chain

(a) Interaction of TCRαβ with self peptide-MHC complex (Protein Data Bank - http://www.rcsb.org/pdb/home/home.do; accession number, 2CKB)[68], with a 'footprint' view showing the stereotyped polarity of the TRAV and TRBV CDR loops on pMHC (adapted from Garcia et al. 2009[69]). (b) Representation of the TCRβ locus at human chromosome 7q34 (adapted from Freeman et al.[23]). The TCRβ locus spans 620kb and includes over 60 TRBV-genes (green). There are two TRBC-genes (light blue) each downstream from a TRBD (purple) and six or seven TRBJs (yellow). The steps of recombination to produce a TCRβ chain are numbered 1-3: (1). Recombination first occurs between TRBJ and TRBD genes, followed by (2). recombination to a TRBV-gene. Addition of non-templated bases occurs at gene junctions during recombination events (red). (3). After transcription, intervening sequences are spliced out so that a TRBC is adjacent to the recombined TRBV(D)J sequence. CDR3β encompasses the highly diverse TRBV(D)J gene junctions and is located between the last conserved cysteine (C) codon of the TRBV-gene and the first conserved phenylalanine (F) codon of the TRBJ-gene in the conserved motif FGXG. Gene width and distances are not to scale. A detailed locus map can be obtained from IMGT[2]).

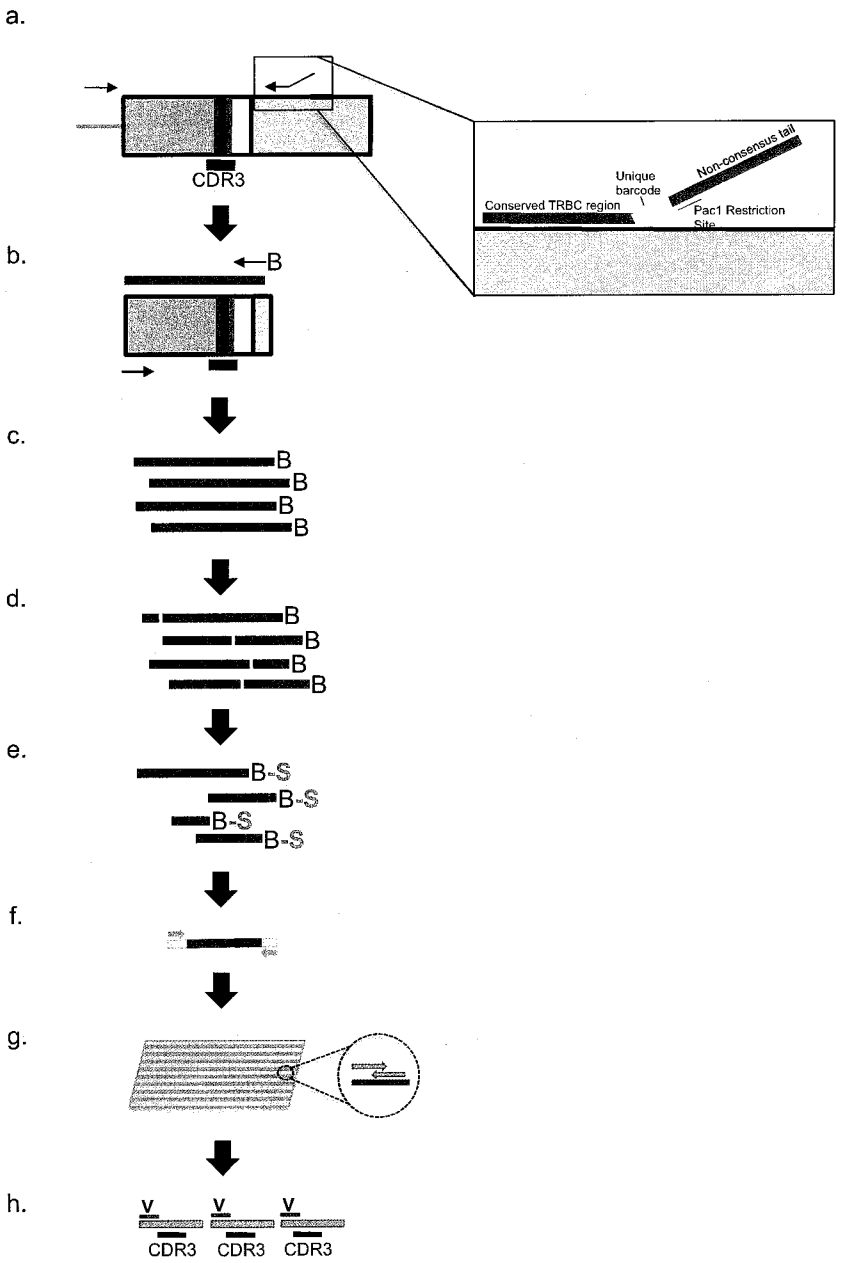


Figure 2. Amplification strategy overview for TCRβ sequences

(a) Indexed TCRβ specific primer used to amplify CDR3β region of interest composed of four sections: TRBC gene consensus site, hexameric indexed site, Pac1 restriction site and non-consensus tail site. (b) Indexed product further amplified with biotinylated primer. (c) Products of biotinylation PCR biotinylated on 3' ends. (d) Biotinylated products randomly sheared into 125-175bp range. (e) 3' biotinylated sheared region captured using streptavidin beads. (f) Product A-tailed and Illumina adapters ligated onto the 5' and 3' ends. Pre-amplified prior to sequencing. (g) Paired-end sequenced in a single lane of Illumina flow cell. (h) Computational TCRβ sequences mined from raw data. Gene widths and distances and amplicon sizes not to scale.

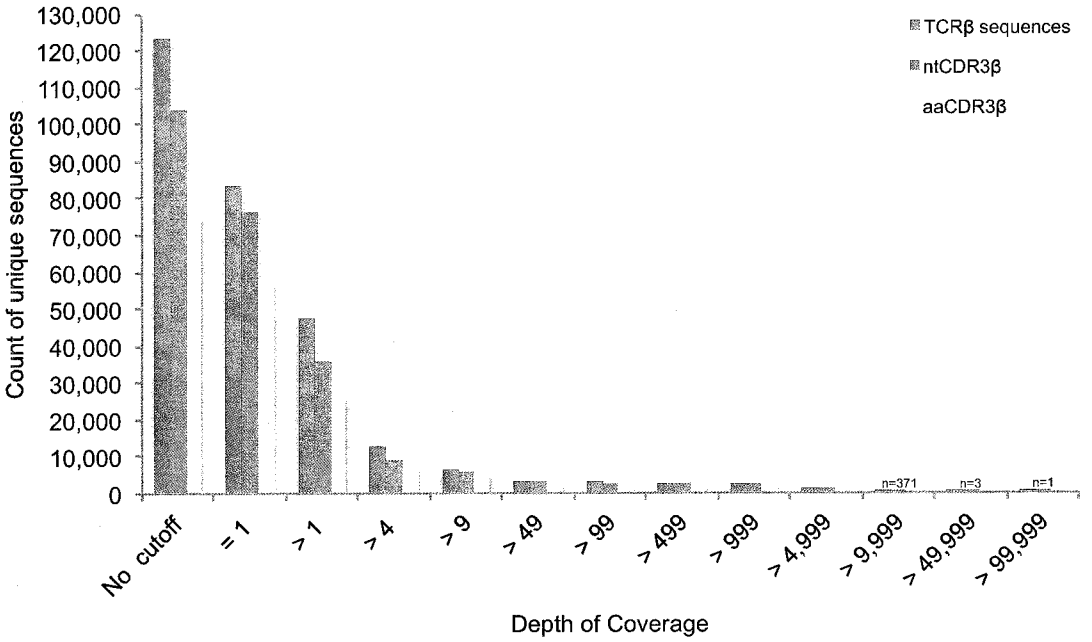
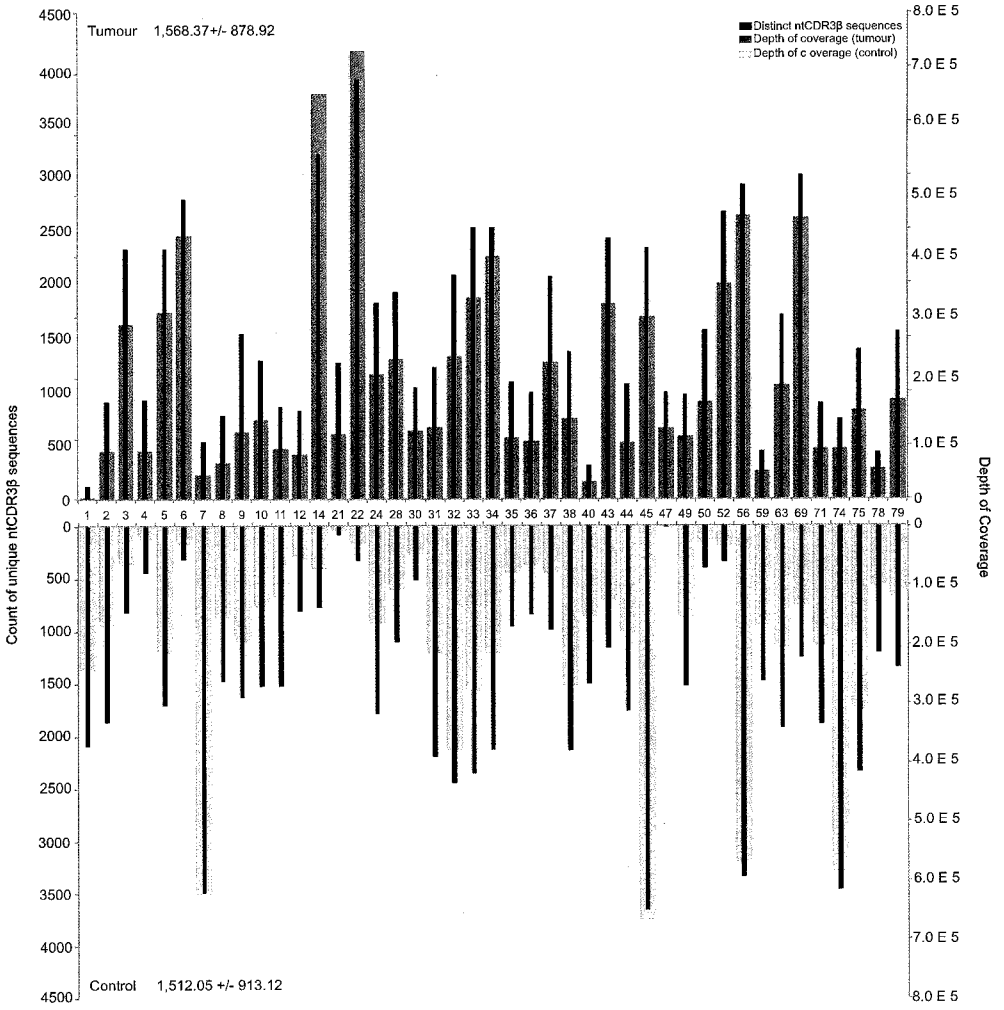


Figure 3. Sequences present at varying depths of coverage

Counts of in-frame unique TCRβ sequences (yellow), unique ntCDR3β sequences (purple) and unique predicted aaCDR3β sequences (yellow) with one possible TRBV gene assignment present at different DOC cutoffs in the complete dataset. Note that *DOC > 9,999* = 371, *DOC > 49,999* = 3, and *DOC > 99,999* = 1. *No cutoff* shows sequence counts for entire dataset, regardless of DOC. *DOC = 1* shows sequence counts for those sequences present in the dataset only once. Sequence counts used in this thesis were present at a *DOC > 1*.

4a.



4b.

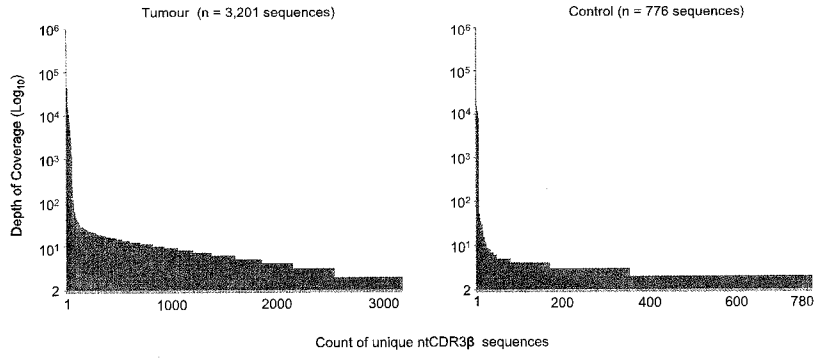


Figure 4. ***CDR3β sequence abundance and diversity of CRC tumour and control tissues variable***

(a) CDR3β nucleotide sequence diversity and abundance across all tumour tissue samples (top plot) and control tissues (bottom plot). Black bars (left y-axis): number of unique CDR3β nucleotide sequences present in each sample. Pink/blue bars (right y-axis): depth of coverage (DOC) of all sequences in a given sample. Note that the scales are different on left and right axes. Statistics given in plots are mean + / standard deviation of distinct ntCDR3b sequences present per tissue type. (b) Shape of sequence distribution observed for patient 14-T shown. Shape of distribution is same for all samples. DOC plotted on a log10 scale.

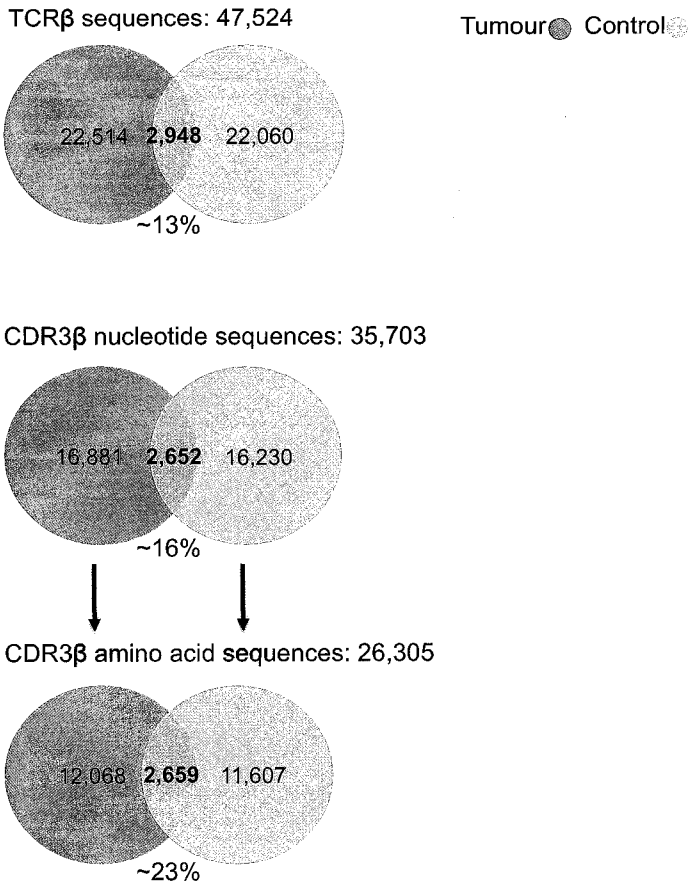


Figure 5. Counts of sequences found in tumour and control tissues

Venn diagrams illustrate the number of unique sequences captured *only* in tumour tissues (red), *only* in control repertoires (blue) and those found in both repertoires (overlap) for TCRβ sequences (top), ntCDR3β sequences (middle) and for predicted aaCDR3β sequences (bottom). Counts given above circles indicate total number of unique sequences in tissue. Counts given within the circles indicate unique sequences only found in a tissue type. Percentage under overlapping sections represents proportion of unique sequences shared between tumour and control repertoires. Amino acid sequences in dataset predicted from nucleotide sequences (arrows). Note that overlaps reflect sequence data captured in this dataset. These overlaps may change depending on the depth of sequencing preformed in future.

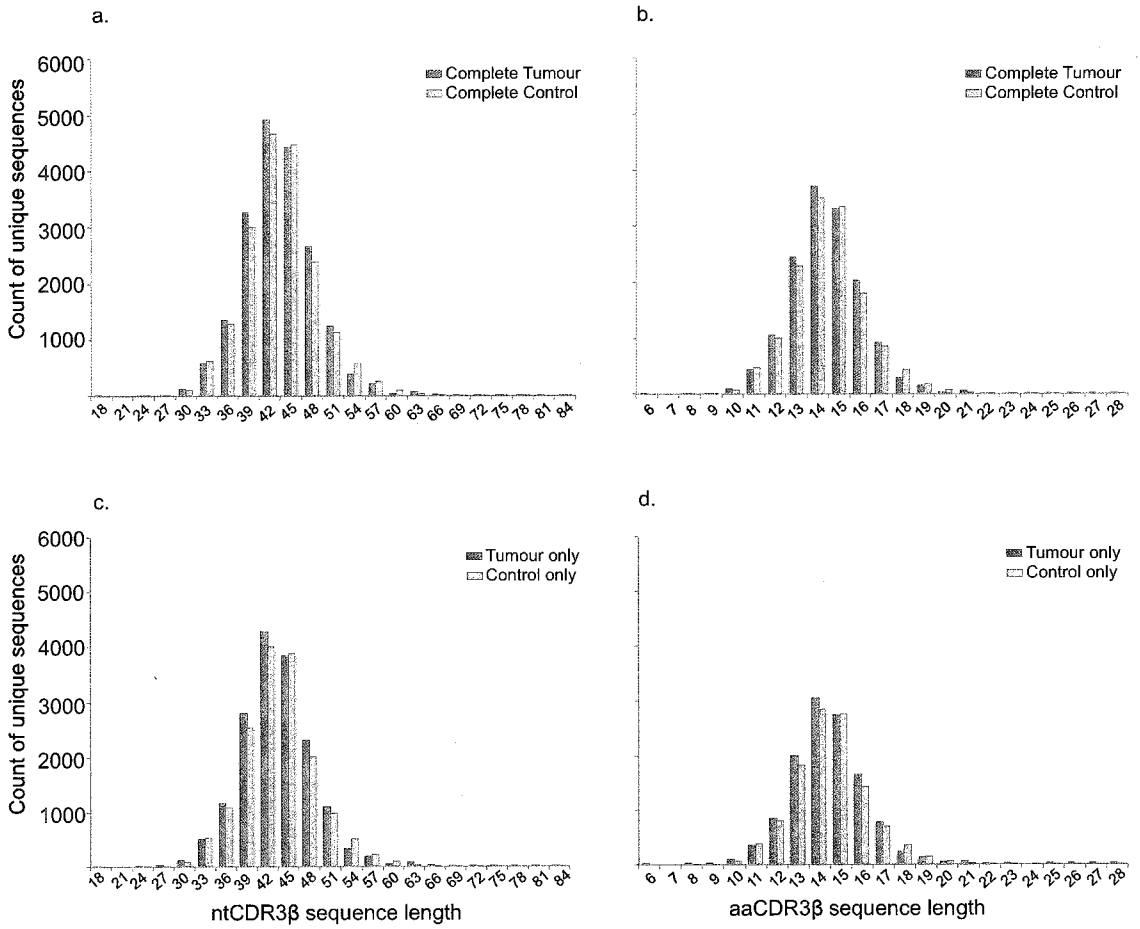


Figure 6. CDR3β sequence length distribution similar for tumour and control tissues

Distribution of ntCDR3β (a, c) and aaCDR3β (b, d) sequence length for tumour (red, left bars on each plot) and matched normal control (blue, right bars on each plot) repertoires. (a, b) represent entire repertoire, while (c, d) represent tumour and control-only repertoires.

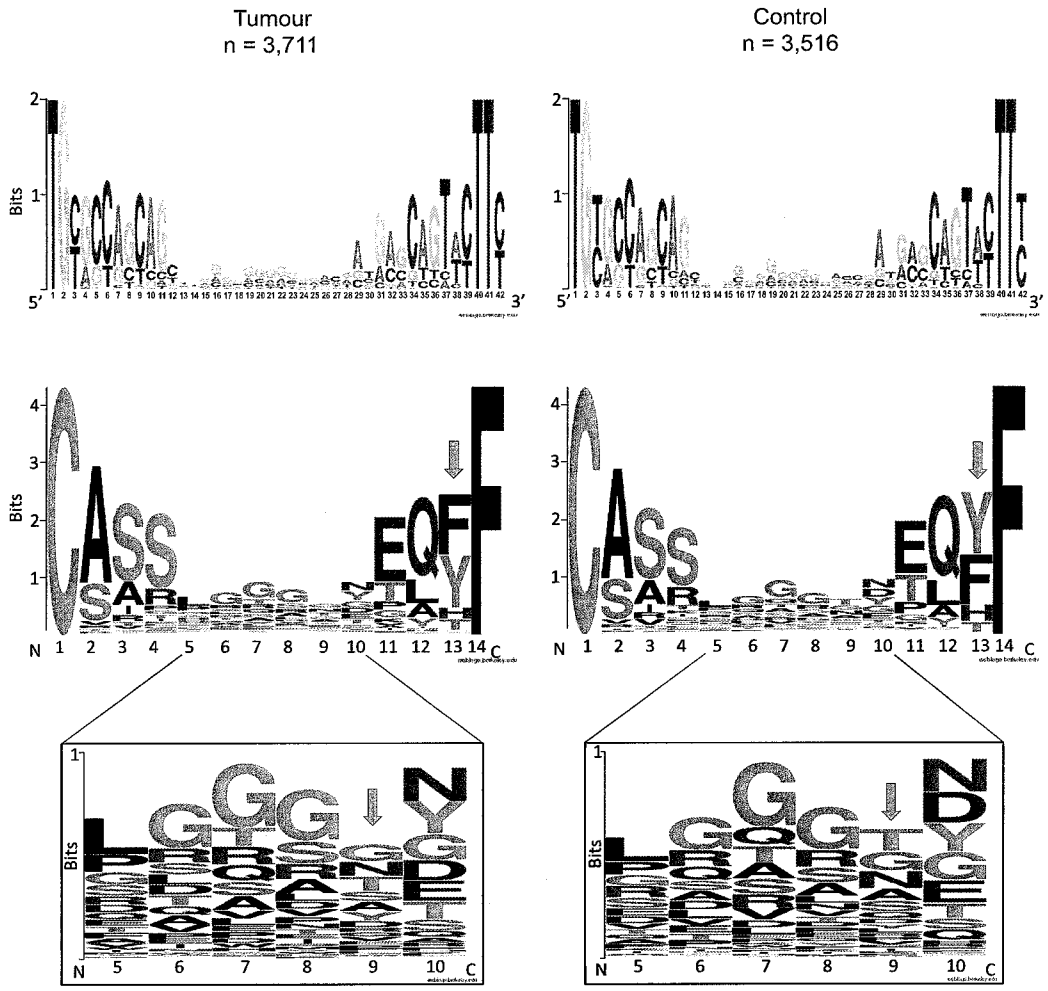


Figure 7. ***Tumour CDR3β amino acid consensus sequence composition differs slightly from control CDR3β amino acid consensus sequence composition***

To explore composition of the most frequently observed aaCDR3β sequence length found in the tumour (left) and control (right) repertoires, I used a precise length criterion, defined as all residues between the last conserved cysteine codon of TRBV and the first conserved phenylalanine in the TRBJ in the conserved motif FGXG. I used using WebLogo[70] to create these logos. At each position the bases/residues are arranged in order of predominance from top to bottom, with the highest frequency based located on the top of the stack[70]. Therefore, the general consensus sequence can be read by reading the top base/residue at every position. In addition, the relative size of individual bases/residues shows the relative frequency of the bases/residues at that position (i.e. if a letter is large at its position, then its frequency at that position is high)[71]. The overall height of each stack is proportional to the sequence conservation, measured in Bits, at that position[70,71]. Amino acid residues are coloured according to their chemical properties; polar amino acids (G, S, T, Y, C, Q, N) show as green, basic (K, R, H) blue, acidic (D, E) red, and hydrophobic (A, V, L, I, P, W, F, M) amino acids as black. Default colours for nucleotide sequences are G, orange; T and U, red; C, blue; and A, green[71]. aaCDR3β sequences are predicted from ntCDR3β sequences. Grey arrows point to positions where there are residue differences in consensus sequences between tumour and control repertories.

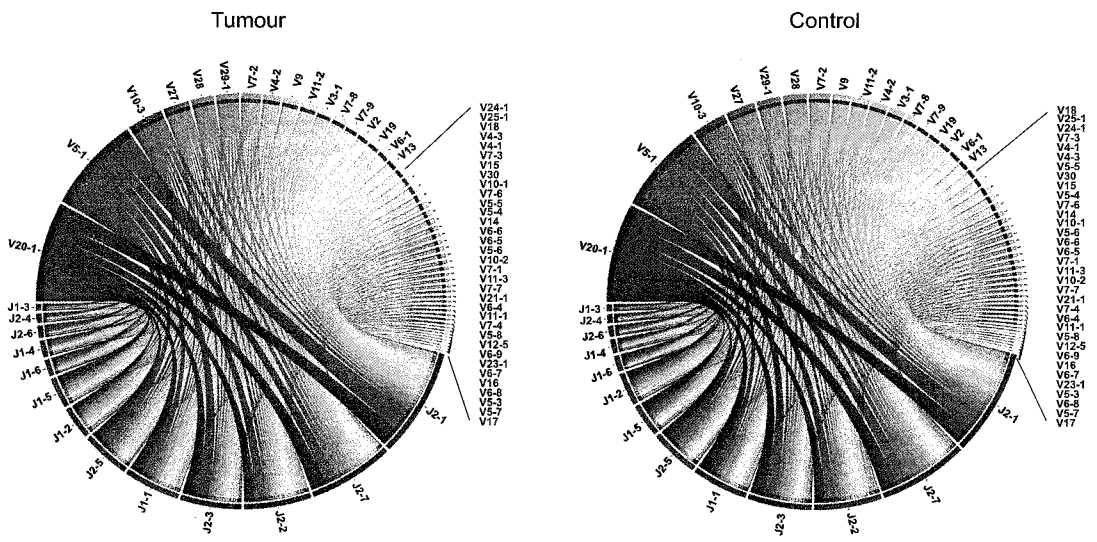


Figure 8. TRBV/J-gene usage similar in tumour and control tissues

Plots created from TCRβ sequence repertoires from tumour (left) and control (right) repertories of 43 CRC patietns. The width of the bands is proportional to the number of times the TRBV and TRBJ genes (connected by the band) co-occur in TCRβ sequences. TRBV and TRBJ segments are arranged left to right and right to left respectively, and ordered by total pairing links that they share. This figure was generated using the Circos software package[72]

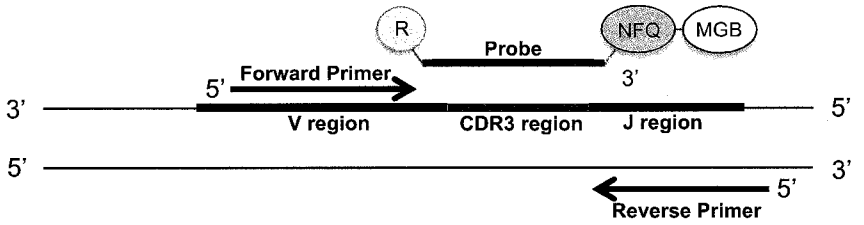
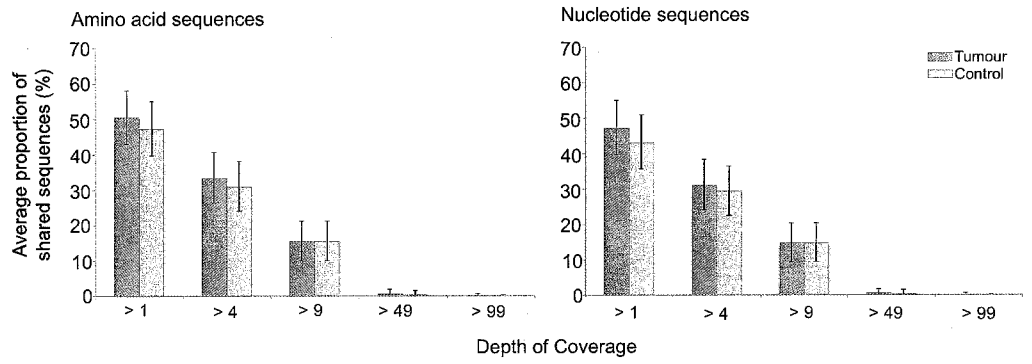


Figure 9. qPCR TaqMan primer/probe assay design for validation of CDR3β sequences

TaqMan probe designed to anneal to the CDR3β region of the TCRβ. Forward and reverse primers designed to sit as closely to probe as possible on the TRBV, and TRBJ-genes respectively. Primer/probe pairs designed using Primer3[73(p3)].

10a.



10b.

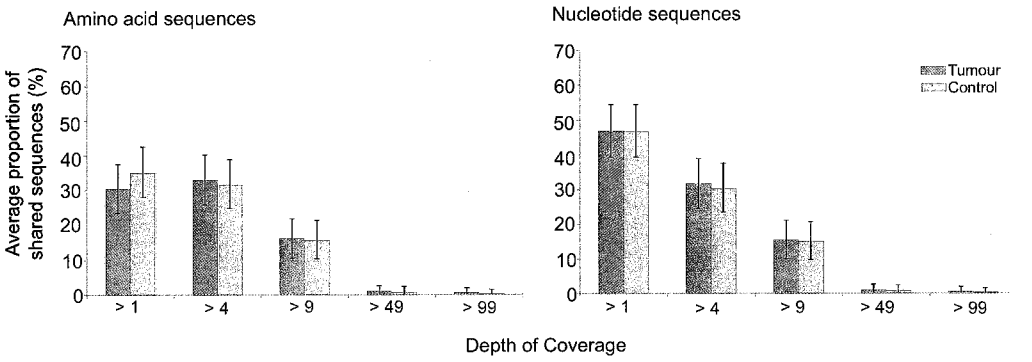


Figure 10. Average proportions of shared CDR3β sequences at different DOCs

Figure shows average +/- standard error proportions of shared aaCDR3β (left plots) and ntCDR3β (right plots) sequences in the tumour (red) and control (blue) repertoires both (a) between and (b) within tumour and control repertoires of patients at differing DOCs.

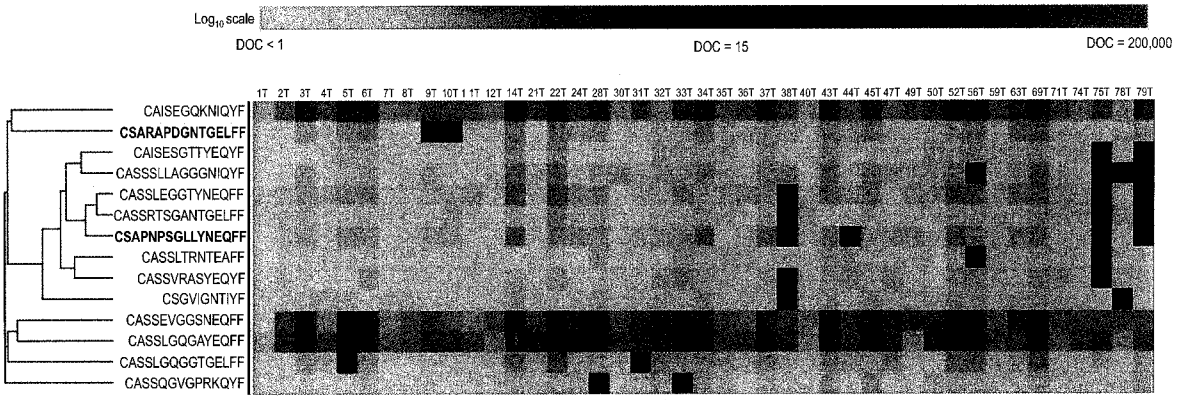


Figure 11. aaCDR3β sequences shared between multiple tumour samples at DOC > 99

Green to red transition indicates lower to higher depth of aaCDR3 sequences. Each branch tip/row represents a unique CDR3 amino acid sequence. Clustering tree was generated according to aaCDR3 hierarchical depth in tumour samples. Each column is unique to the listed sample ID. Data is plotted on a $\log_{10}$ scale. Heat map generated using Mayday software[74]. Highlighted sequences (yellow) represent sequences with significant associations with survival (see figure 12 for survival curves).

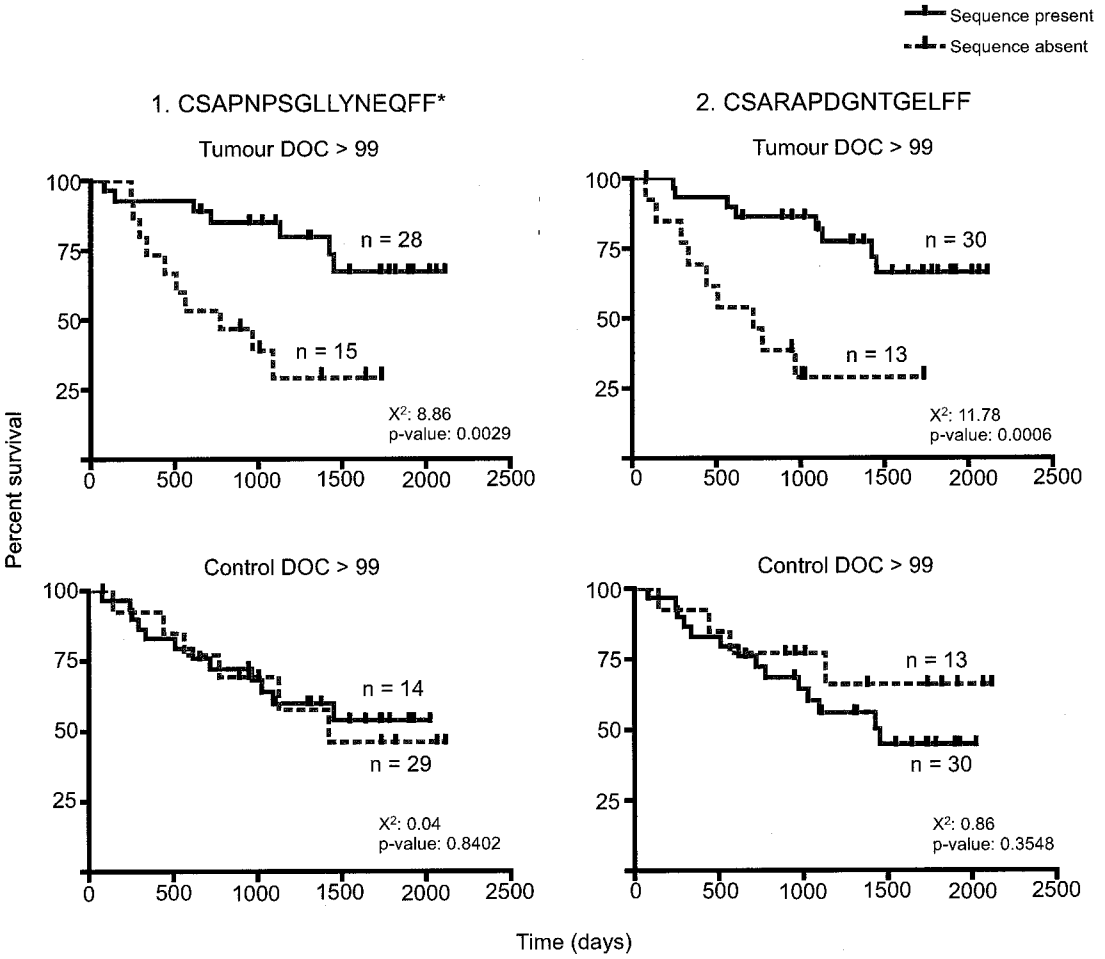
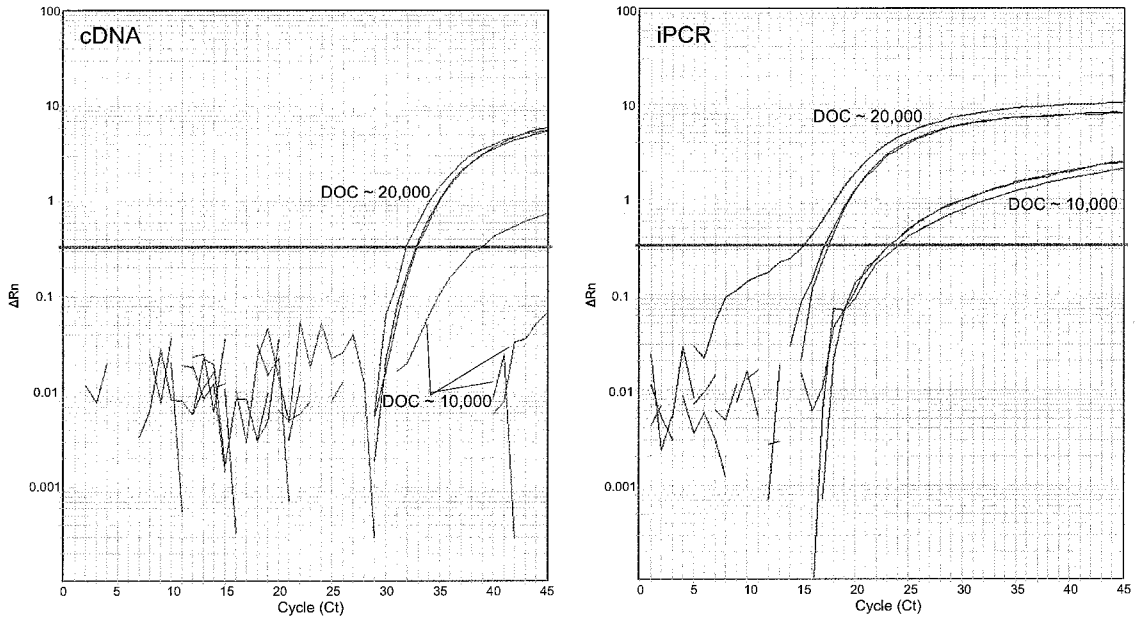


Figure 12. aaCDR3β sequences present in tumour repertoires associated with increased overall survival

Kaplan Meier plots for aaCDR3β amino acid sequences associated with increased overall survival when present in tumour tissues. Plots show percent survival of patients with (red) and without (blue, dashed) amino acid sequence present. Top plots show survival for patients with these sequences in their tumour tissues. Bottom plots show survival for people with these sequences in their control tissues. Asterisk (*) represents the sequence validated using qPCR in patient samples in the lab. Kaplan Meier plot constructed using GraphPad Prism 4.0b[57]

13a. Limit of detection



13b. qPCR for survival associated sequence

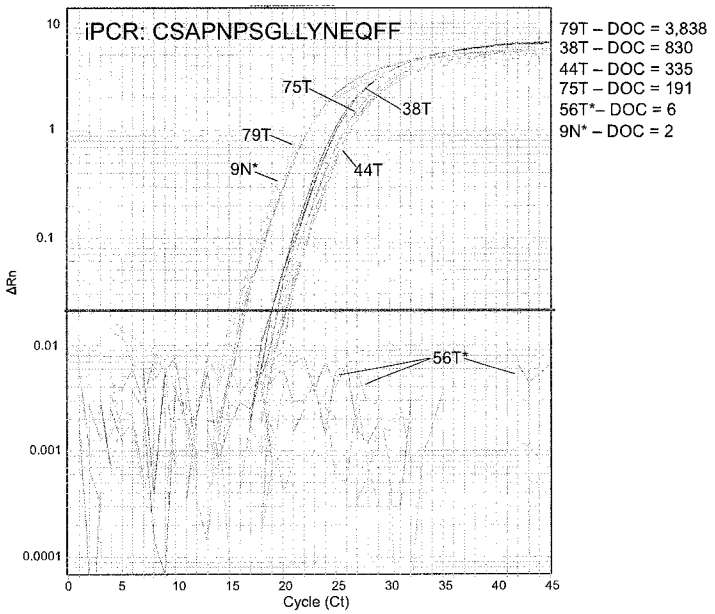


Figure 13. qPCR limit of detection and amplification for validation of aaCDR3b sequence presence in samples

(a) TaqMan qPCR raw data showing limit of detection (LOD) of ntCDR3β sequence at a DOC of 20,000, and a separate aaCDR3β sequence present with a DOC of 10,000 when cDNA (left plot) and indexed PCR product (iPCR) (right plot) from a single sample (14T) are used as template. (b) TaqMan qPCR amplification of aaCDR3β sequence CSAPNPSGLLYNEQFF from four tumour samples with a DOC > 190, and from 2 samples with DOC < 190, illustrating the unreliability of amplification from samples with DOC's less than the LOD. Asterisks (*) highlight samples with unreliable amplification signals relative to DOC. Baseline is background noise between ~ cycle 3-15. Threshold (green horizontal line) - level of signal distinguishes relevant amplification signal over the background. X-axis: Ct – Threshold cycle (Ct) – the cycle number at which the fluorescent signal of the reaction crosses the threshold. Ct is inversely related to the amount of starting template. Y-axis: Rn – the fluorescence of the probe reporter dye (FAM) divided by the fluorescence of a passive reference dye (ROX). Amplification curves generated on 7900HT Fast Real-Time PCR system using SDS 2.4 software (http://www.appliedbiosystems.com/).

7. References

1. Kenneth Murphy, Paul Travers, Mark Walport. *Janeway's Immunobiology*. 7th ed. New York, NY: Garland Science; 2008.

2. Lefranc M-P. IMGT, the international ImMunoGeneTics database. *Nucleic Acid Research*. 2003;31(1):307–310.

3. Davis MM, Bjorkman PJ. T-cell antigen receptor genes and T-cell recognition. *Nature*. 1988;334(6181):395–402.

4. Arstila TP, Casrouge A, Baron V, et al. A direct estimate of the human Aβ T cell receptor diversity. *Science*. 1999;286(5441):958–961.

5. Warren RL, Freeman JD, Zeng T, et al. Exhaustive T-cell repertoire sequencing of human peripheral blood samples reveals signatures of antigen selection and a directly measured repertoire size of at least 1 million clonotypes. *Genome Res.* 2011;21(5):790–797.

6. Monod MY, Giudicelli V, Chaume D, Lefranc M-P. IMGT/JunctionAnalysis: the first tool for the analysis of the immunoglobulin and T cell receptor complex V-J and V-D-J JUNCTIONs. *Bioinformatics*. 2004;20(1):379–385.

7. Padovan E, Casorati G, Dellabona P, et al. Expression of two T cell receptor alpha chains: dual receptor T cells. *Science*. 1993;262(5132):422–424.

8. Balomenos D, Balderas RS, Mulvany KP, et al. Incomplete T cell receptor V beta allelic exclusion and dual V beta-expressing cells. *J. Immunol.* 1995;155(7):3308–3312.

9. Schwartz RH. T Cell Anergy. *Annual Review of Immunology*. 2003;21(1):305–334.

10. Mowat AM. Anatomical basis of tolerance and immunity to intestinal antigens. *Nat. Rev. Immunol.* 2003;3(4):331–341.

11. Suzuki H. Differences in intraepithelial lymphocytes in the proximal, middle, distal parts of small intestine, cecum, and colon of mice. *Immunol. Invest.* 2009;38(8):780–796.

12. Cheroutre H, Madakamutil L. Acquired and natural memory T cells join forces at the mucosal front line. *Nat Rev Immunol*. 2004;4(4):290–300.

13. Regnault A, Kourilsky P, Cumano A. The TCR-β chain repertoire of gut-derived T lymphocytes. *Seminars in Immunology*. 1995;7(5):307–319.

14. Gross GG, Schwartz VL, Stevens C, et al. Distribution of dominant T cell receptor beta chains in human intestinal mucosa. *J. Exp. Med.* 1994;180(4):1337–1344.

15. Blumberg RS, Yockey CE, Gross GG, Ebert EC, Balk SP. Human intestinal intraepithelial lymphocytes are derived from a limited number of T cell clones that utilize multiple V beta T cell receptor genes. *J. Immunol.* 1993;150(11):5144–5153.

16. Probert CSJ, Saubermann LJ, Balk S, Blumberg RS. Repertoire of the alpha beta T-cell receptor in the intestine. *Immunol. Rev.* 2007;215:215–225.

17. Pannetier, Cochet M, Darche S, et al. The sizes of the CDR3 hypervariable regions of the murine T-cell receptor beta chains vary as a function of the recombined germ-line segments. *Proceedings of the National Academy of Sciences of the United States of America*. 1993;90(9):4319–4323.

18. Currier JR, Robinson MA. Spectratype/Immunoscope Analysis of the Expressed TCR Repertoire. In: Coligan JE, Bierer BE, Shevach EM, Strober W, eds. *Current Protocols in Immunology*. Hoboken, NJ, USA: John Wiley & Sons, Inc. 2001.

19. Warren RL, Nelson BH, Holt RA. Profiling model T-cell metagenomes with short reads. *Bioinformatics*. 2009;25(4):458–464.

20. Miles JJ, Douek DC, Price DA. Bias in the αβ T-cell repertoire: implications for disease pathogenesis and vaccination. *Immunol. Cell Biol.* 2011;89(3):375–387.

21. Ostenstad B, Sioud M, Lea T, Schlichting E, Harboe M. Limited heterogeneity in the T-cell receptor V-gene usage in lymphocytes infiltrating human colorectal tumours. *Br J Cancer*. 1994;69(6):1078–1082.

22. Li H-F, Wan Y-L, Liu Y-C, et al. [Colorectal cancer patients have oligoclonal proliferation of T cells in blood]. *Beijing Da Xue Xue Bao*. 2004;36(1):66–69.

23. Freeman JD, Warren RL, Webb JR, Nelson BH, Holt RA. Profiling the T-cell receptor beta-chain repertoire by massively parallel sequencing. *Genome Research*. 2009;19:1817–1824.

24. Warren RL, Freeman JD, Zeng T, et al. Exhaustive T-cell repertoire sequencing of human peripheral blood samples reveals signatures of antigen selection and a directly measured repertoire size of at least 1 million clonotypes. *Genome Research*. 2011.

25. Holt RA, Jones SJM. The new paradigm of flow cell sequencing. *Genome Research*. 2008;18:839–846.

26. Schreiber RD, Old LJ, Smyth MJ. Cancer Immunoediting: Integrating Immunity's Roles in Cancer Suppression and Promotion. *Science*. 2011;331(6024):1565–1570.

27. Chen Y-T, Ross DS, Chiu R, et al. Multiple cancer/testis antigens are preferentially expressed in hormone-receptor negative and high-grade breast cancers. *PLoS ONE*. 2011;6(3):e17876.

28. Jager D, Jager E, Knuth A. Immune responses to tumour antigens: implications for antigen specific immunotherapy of cancer. *J Clin Pathol.* 2001;54(9):669–674.

29. Canadian Cancer Society's Steering Committee on Cancer Statistics. *Canadian Cancer Statistics 2012.* Toronto, ON: Canadian Cancer Society; 2012.

30. Lynch HT, de la Chapelle A. Hereditary colorectal cancer. *N. Engl. J. Med.* 2003;348(10):919–932.

31. Jemal A, Bray F, Center MM, et al. Global cancer statistics. *CA: A Cancer Journal for Clinicians.* 2011;61(2):69–90.

32. Kim H-J, Yu M-H, Kim H, Byun J, Lee C. Noninvasive molecular biomarkers for the detection of colorectal cancer. *BMB Rep.* 2008;41(10):685–692.

33. Bosch LJW, Carvalho B, Fijneman RJA, et al. Molecular tests for colorectal cancer screening. *Clin Colorectal Cancer.* 2011;10(1):8–23.

34. Naito Y, Saito K, Shiiba K, et al. CD8+ T cells infiltrated within cancer cell nests as a prognostic factor in human colorectal cancer. *Cancer Res.* 1998;58(16):3491–3494.

35. Deschoolmeester V, Baay M, Van Marck E, et al. Tumor infiltrating lymphocytes: an intriguing player in the survival of colorectal cancer patients. *BMC Immunol.* 2010;11:19.

36. Koch M, Beckhove P, op den Winkel J, et al. Tumor infiltrating T lymphocytes in colorectal cancer: tumor-selective activation and cytotoxic activity in situ. *Annals of surgery.* 2006;244(6):986.

37. Menon AG, Janssen-van Rhijn CM, Morreau H, et al. Immune system and prognosis in colorectal cancer: a detailed immunohistochemical analysis. *Lab. Invest.* 2004;84(4):493–501.

38. Pagès F, Berger A, Camus M, et al. Effector memory T cells, early metastasis, and survival in colorectal cancer. *N. Engl. J. Med.* 2005;353(25):2654–2666.

39. Salama P, Phillips M, Grieu F, et al. Tumor-Infiltrating FOXP3+ T Regulatory Cells Show Strong Prognostic Significance in Colorectal Cancer. *JCO.* 2009;27(2):186–192.

40. Ohtani H, Naito Y, Saito K, Nagura H. Expression of costimulatory molecules B7-1 and B7-2 by macrophages along invasive margin of colon cancer: a possible antitumor immunity? *Lab. Invest.* 1997;77(3):231–241.

41. Zhang L, Conejo-Garcia JR, Katsaros D, et al. Intratumoral T cells, recurrence, and survival in epithelial ovarian cancer. *N. Engl. J. Med.* 2003;348(3):203–213.

42. Ogino S, Galon J, Fuchs CS, Dranoff G. Cancer immunology--analysis of host and tumor factors for personalized medicine. *Nat Rev Clin Oncol.* 2011;8(12):711–719.

43. W.J. Lesterhuis, C.J. Punt. Harnessing the immune system to combat cancer (Poster). *Nature Reviews Drug Discovery.* 2012;11.

44. Jordan KR, Buhrman JD, Sprague J, et al. TCR hypervariable regions expressed by T cells that respond to effective tumor vaccines. *Cancer immunology, immunotherapy: CII*. 2012.

45. Li H, Wan Y, Liu Y, Wu T, Zhu P. TCRβ repertoire in TIL and PBL of patients with colorectal cancer. *Chinese Journal of Cancer Research*. 2003;15(4):277–281.

46. Dudley ME, Wunderlich JR, Robbins PF, et al. Cancer regression and autoimmunity in patients after clonal repopulation with antitumor lymphocytes. *Science*. 2002;298(5594):850–854.

47. Rosenberg SA. Cancer regression in patients with metastatic melanoma after the transfer of autologous antitumor lymphocytes. *Proceedings of the National Academy of Sciences*. 2004;101(suppl_2):14639–14645.

48. Castellarin M, Warren RL, Freeman JD, et al. Fusobacterium nucleatum infection is prevalent in human colorectal carcinoma. *Genome Research*. 2011;22(2):299–306.

49. Watson PH. Canadian Tumour Repository Network. *Biopreserve Biobank*. 2010;8:181–185.

50. Ozawa T, Tajiri K, Kishi H, Muraguchi A. Comprehensive analysis of the functional TCR repertoire at the single-cell level. *Biochem. Biophys. Res. Commun.* 2008;367(4):820–825.

51. Peters DG, O'Hare EH, Ferrell RE, et al. Comprehensive transcript analysis in small quantities of mRNA by SAGE-Lite. *Nucleic Acids Research*. 1999;27(24):e39 –e44.

52. Slater GSC, Birney E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*. 2005;6:31.

53. Brandtzaeg P, Bjerke K. Immunomorphological characteristics of human Peyer's patches. *Digestion*. 1990;46 Suppl 2:262–273.

54. Robins HS, Srivastava SK, Campregher PV, et al. Overlap and Effective Size of the Human CD8+ T Cell Receptor Repertoire. *Sci Transl Med*. 2010;2(47):47ra64–47ra64.

55. Venturi V, Price DA, Douek DC, Davenport MP. The molecular basis for public T-cell responses? *Nat. Rev. Immunol.* 2008;8(3):231–238.

56. Warren R, Gina C, Castellarin M, et al. HLAminer: Derivation of HLA types from shotgun sequence datasets. *Nature Methods*. 2012;In review.

57. Anon. *GraphPad Prism*. San Diego California USA Available at: www.graphpad.com.

58. Siewert K, Malotka J, Kawakami N, et al. Unbiased identification of target antigens of CD8(+) T cells with combinatorial libraries coding for short peptides. *Nature medicine*. 2012.

59. Chien Y, Bonneville M. Gamma delta T cell receptors. *Cellular and Molecular Life Sciences*. 2006;63(18):2089–2094.

60. Lamb LS. γδ T Cells in Cancer Wang R, ed. *Innate Immune Regulation and Cancer Immunotherapy*. 2012:23–38.

61. Corvaisier M, Moreau-Aubry A, Diez E, et al. V gamma 9V delta 2 T cell response to colon carcinoma cells. *J. Immunol.* 2005;175(8):5481–5488.

62. Wakita D, Sumida K, Iwakura Y, et al. Tumor-infiltrating IL-17-producing gammadelta T cells support the progression of tumor by promoting angiogenesis. *Eur. J. Immunol.* 2010;40(7):1927–1937.

63. Wucherpfennig KW, Allen PM, Celada F, et al. Polyspecificity of T cell and B cell receptor recognition. *Seminars in Immunology*. 2007;19(4):216–224.

64. Vyas JM, Veen AGV der, Ploegh HL. The known unknowns of antigen processing and presentation. *Nature Reviews Immunology*. 2008;8(8):607–618.

65. Smith ES, Mandokhot A, Evans EE, et al. Lethality-based selection of recombinant genes in mammalian cells: application to identifying tumor antigens. *Nat. Med.* 2001;7(8):967–972.

66. Santegoets SJAM, Schreurs MWJ, Reurs AW, et al. Identification and characterization of ErbB-3-binding protein-1 as a target for immunotherapy. *J. Immunol.* 2007;179(3):2005–2012.

67. Crawford F, Jordan KR, Stadinski B, et al. Use of baculovirus MHC/peptide display libraries to characterize T-cell receptor ligands. *Immunol. Rev.* 2006;210:156–170.

68. Garcia KC, Degano M, Pease LR, et al. Structural basis of plasticity in T cell receptor recognition of a self peptide-MHC antigen. *Science*. 1998;279(5354):1166–1172.

69. Christopher Garcia K, Adams JJ, Feng D, Ely LK. The molecular basis of TCR germline bias for MHC is surprisingly simple. *Nat Immunol*. 2009;10(2):143–147.

70. Crooks GE, Hon G, Chandonia J-M, Brenner SE. WebLogo: A Sequence Logo Generator. 2004:1188–1190.

71. Schneider TD, Stephens RM. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.* 1990;18(20):6097–6100.

72. Krzywinski M, Schein J, Birol İ, et al. Circos: An information aesthetic for comparative genomics. *Genome Research*. 2009;19(9):1639 –1645.

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

74. Battke F, Symons S, Nieselt K. Mayday--integrative analytics for expression data. *BMC Bioinformatics*. 2010;11:121.

Appendices

Appendix A. TCRβ profiling with Illumina reads pipeline

Steps to extract TCRβ sequence data from raw Illumina data follows on next page. Pipeline created by René Warren, BCGSC, 2010.

* This manual assumes that you have a working knowledge of unix, MySQL and some shell and perl scripting experience

NAME

TCRB profiling pipeline

CONTENTS

SYNOPSIS

1. copy illumina reads *qseq.txt to a directory of your choice

```
xhost10>cp –rf *qseq.txt .
```

2. uncompress all qseq.txt

```
xhost10>bunzip2 *.bz2
```

3. make a shell script to run joinMates.pl (microassembler) on compute cluster

```
ls -la | perl -ne 'if(/s_(\d)_1_(\d+)\_qseq.txt/){
print "cd /tmp/;mkdir lr;cd lr/;mkdir $1-$2;cd $1-$2/;
/projects/02/rwarren/solexa/TCRb/participants/joinMates.pl
/FULLPATH_TO_QSEQ_1/s_$1_1_$2_qseq.txt
/FULLPATH_TO_QSEQ_2/s_$1_2_$2_qseq.txt
SUFFIX;
rm -rf /tmp/lr/$1-$2;\n";}' > runJM.sh
```

4. run the microassembler on the compute cluster by executing the shell script

```
apollo> mqsub --file runJM.sh
```

5. make a shell script to mine sequence contigs for TRBV

```
xhost10> ls -la | perl -ne 'if(/(s_\d_1_\d+\_qseq.txt.fa)/){
print "/projects/02/rwarren/solexa/TCRb/participants/EScon.pl $1 RW
serial\n";}' > runES.sh
```

6. execute the TRBV mining script on a shell server

```
xhost10> ./runES.sh
```

7. mine candidates contigs for TCRB rearrangements

```
xhost10>./TCRmine.pl CANDIDATES.txt AG 9 9 1
```

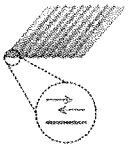
OVERVIEW

A) One sample/library per lane:

Retrieve data location info

Find information about your runs:
http://www.bcgsc.ca/data/sbs/viewer/ and
http://lims02.bcgsc.bc.ca/SDB/cgi-bin/barcode.pl

Copy the flowcell data



Copy :: cp *_qseq.txt.bz2 .
Decompress :: bunzip2 *.bz2

Prepare to run the microassembler on each tile

Make perl one-liner to create shell script

```
ls -la | perl -ne 'if(/s_(\d)_1_([id+])\_qseq.txt/){print "cd /tmp/;mkdir rw;cd rw/;mkdir $1-$2;cd $1-$2/;/joinMates.pl /myIlluminaDir/s_$1_1_$2_qseq.txt SUFFIX /myIlluminaDir/s_$1_2_$2_qseq.txt;rm -rf /tmp/rw/$1-$2;\n";}' > runJM.sh
```

Join mates on the compute cluster

*Go to head cluster node :: ssh apollo
Farm your jobs :: apollo>mqsub --file runJM.sh

Mine for TRBV and orient contigs

Make perl one-liner to create shell script

```
ls -la | perl -ne 'if(/s_\d_1_\d+\_qseq.txt.fa/){print "EScon.pl $1 RW serial\n";}' > runES.sh
```

Run on shell server :: xhost06>./runES.sh

Sequence profile TCR, produce output files

*TCR sequence profile
xhost06>TCRmine.pl CANDIDATES.txt AG 9 9 1

*ensure you have permission to run shell scripts (chmod 755 *.sh) and that alignment.config & freqex.config are copied into your working directory

Rene Warren :: 2010 BC Cancer Agency :: rwarren@bcgsc.ca

B) Multiple indexed samples per lane:

Retrieve data location info

Find information about your runs:
http://www.bcgsc.ca/data/sbs/viewer/ and
http://lims02.bcgsc.bc.ca/SDB/cgi-bin/barcode.pl

Copy the flowcell data

Copy :: cp *_qseq.txt.bz2 .
Decompress :: bunzip2 *.bz2

Prepare to run the microassembler on each tile

Make perl one-liner to create shell script

```
ls -la | perl -ne 'if(/s_(\d)_1_([id+])\_qseq.txt/){print "cd /tmp/;mkdir rw;cd rw/;mkdir $1-$2;cd $1-$2/; /joinMates.pl /myIlluminaDir/s_$1_1_$2_qseq.txt /myIlluminaDir/s_$1_2_$2_qseq.txt SUFFIX;rm -rf /tmp/rw/$1-$2;\n";}' > runJM.sh
```

Join mates on the compute cluster

*Go to head cluster node :: ssh apollo
Farm your jobs :: apollo>mqsub --file runJM.sh

Mine for TRBV and orient contigs

Make perl one-liner to create shell script

```
ls -la | perl -ne 'if(/s_\d_1_\d+\_qseq.txt.fa/){print "EScon.pl $1 RW serial\n";}' > runES.sh
```

Run on shell server :: xhost06>./runES.sh

Bin contigs

Segregate contigs using barcodes ::
./binCandidatesContigs.pl <CANDIDATES.txt> <barcode-sample.tsv>

Sequence profile TCR, produce output files

*TCR sequence profile
xhost06>TCRmine.pl CANDIDATES-sample.txt AG 9 9 1

*ensure you have permission to run shell scripts (chmod 755 *.sh) and that alignment.config & freqex.config are copied into your working directory

Rene Warren :: 2010 BC Cancer Agency :: rwarren@bcgsc.ca

DESCRIPTION

The TCRB mining pipeline comprises a series of PERL one-liners, shell scripts and PERL scripts that run in parallel or sequentially to mine illumina-sequenced TCRB amplicons for specific V-D-J rearrangements. It keeps track of the rearrangements, their depth of coverage, the frame, the number of possible TRBV identified, the identity of both the V and J segments, the number of bases deleted at the 3' and 5' end of these segments, respectively, the boundary of the the CDR3-encoding region as well as the predicted CDR3 amino acid sequence and only reports the highest quality sequences. The mined data is conveniently outputted to a MySQL text file that can be use to populate a MySQL database.

Scripts available at: /home/rwarren/code/TCRb/bin
Configuration files available at: /home/rwarren/code/TCRb/config
Examples at: /projects/02/rwarren/solexa/TCRb/participants

Ensure you set appropriate file permissions to execute shell scripts (chmod 755 *.sh) and that alignment.config & Jregex.config are copied into your working directory

COMMANDS AND OPTIONS

1. copy illumina reads *qseq.txt to a directory of your choice using "cp"

xhost10>cp <source> <destination>
e.g.
xhost10>cp *qseq.txt .

2. uncompress all qseq.txt

xhost10>bunzip2 *.bz2

3. make a shell script to run joinMates.pl (microassembler) on compute cluster

All on one line, without carriage return type:

```
ls -la | perl -ne 'if(/s_(\d)_1_(\d+)\_qseq.txt/){print "cd /tmp/;mkdir lr;cd lr/;mkdir
$1-$2;cd $1-$2/;/projects/02/rwarren/solexa/TCRb/participants/joinMates.pl
/FULLPATH_TO_QSEQ_1/s_$1_1_$2_qseq.txt
/FULLPATH_TO_QSEQ_2/s_$1_2_$2_qseq.txt
SUFFIX;rm -rf /tmp/lr/$1-$2;\n";}' > runJM.sh
```

REPLACE "FULLPATH_TO_QSEQ_1" and "FULLPATH_TO_QSEQ_2" with the full path directory where your qseq.txt files are located.

REPLACE "rw" by a personal name that will become YOUR workspace on Apollo.

DEALING WITH ABUNDANT NGS DATA:

3a) Make shell script to split tiles (using PERL one liner)

```
ls -la | perl -ne 'if(/s_(\d)_1_(\d+)\_qseq.txt/){print
"/home/rwarren/perl/Development/SeqDev/bin/splitFile.pl
s_$1_1_$2_qseq.txt
10\n/home/rwarren/perl/Development/SeqDev/bin/splitFile.pl
s_$1_2_$2_qseq.txt 10\n";}' > splittiles.sh
```

Where each tile is split into 10.

3b) Split tiles

```
chmod 755 splittiles.sh
./splittiles.sh
```

3c) Make a new runJM.sh shell script with this modified command:

```
ls -la | perl -ne 'if(/s_(\d)_1_(\d+)\_qseq.txt.(\d+)/){print "cd /tmp/;mkdir qry999;cd
qry999/;mkdir $1-$2-$3;cd $1-$2-
$3/;/projects/02/rwarren/solexa/TCRb/participants/joinMates.pl
/projects/02/rwarren/solexa/TCRb/participants/lane5/raw/s_$1_1_$2_qseq.txt.$3
/projects/02/rwarren/solexa/TCRb/participants/lane5/raw/s_$1_2_$2_qseq.txt.$3 MC;rm
-rf /tmp/qry999/$1-$2-$3; \n";}' > runJM.sh
```

Notice the new regex (split tiles will have a dot followed by a number e.g. s_1_2_1230_qseq.txt.3

Running joinMates on Genesis

3i) SPLIT TILES

```
xhost09>ls -la | perl -ne 'if(/s_(\d)_1_(\d+)\_qseq.txt/){print
"/home/rwarren/perl/Development/SeqDev/bin/splitFile.pl s_$1_1_$2_qseq.txt
10\n/home/rwarren/perl/Development/SeqDev/bin/splitFile.pl
s_$1_2_$2_qseq.txt 10\n";}' > splittiles.sh
xhost09>chmod 755 splittiles.sh
xhost09>./splittiles.sh &
```

3ii) MAKE SHELL SCRIPT TO JOIN MATES

```
xhost09>ls -la | perl -ne 'if(/s_(\d)_1_(\d+)\_qseq.txt.(\d+)/){print "cd
/genesis/scratch/rwarren_prj/test;mkdir qry999;cd qry999/;mkdir $1-$2-$3;cd
$1-$2-$3/;/genesis/scratch/rwarren_prj/perl/joinMates.pl
/genesis/scratch/rwarren_prj/test/s_$1_1_$2_qseq.txt.$3
/genesis/scratch/rwarren_prj/test/s_$1_3_$2_qseq.txt.$3 RW;rm -rf /qry999/$1-
$2-$3;\n";}' > runJM.sh
```

*** MAKE SURE YOU HAVE THE SAME # LINES IN runJM.sh as YOU HAVE TILES!!!!!!

3iii) JOIN MATES

```
ssh genesis
cd /genesis/scratch/rwarren_prj/test
mqsub --file runJM.sh --name TRW
```

---

e.g.

```
ls -la | perl -ne 'if(/s_(\d)_1_(\d+)\_qseq.txt/){print "cd /tmp/;mkdir rw;cd
rw/;mkdir $1-$2;cd $1-
$2/;/projects/02/rwarren/solexa/TCRb/participants/joinMates.pl
/projects/02/rwarren/solexa/TCRb/participants/lane5/raw/s_$1_1_$2_qseq.txt
/projects/02/rwarren/solexa/TCRb/participants/lane5/raw/s_$1_2_$2_qseq.txt
RW;rm -rf /tmp/rw/$1-$2;\n";}' > runJM.sh
```

joinMates.pl

Usage: ./joinMates.pl <file1> <file2> <suffix?> <print singlets y/n default=n> <verbose (optional)>

<file1>	first illumina qseq.txt file (one tile)
<file2>	second (pair) illumina qseq.txt (file for one tile, must match the tile in file1)
<suffix>	any characters that identifies a lane/run uniquely
<print singlets y/n>	if set to yes, it will print reads that are no joined (optional)
<verbose>	if set to 1, will run in verbose mode (for debugging)

joinMates.pl produces four files:

s_$1_1_$2_qseq.txt.fa	overlapping reads joined into a sequence contig for each cluster, fasta format
s_$1_1_$2_qseq.txt.qua	corresponding quality scores, phred+64 ascii encode
s_$1_1_$2_qseq.txt.cov	corresponding coverage file matching the consensus (1=single strand 2=double strand)
s_$1_1_$2_qseq.txt.log	This is a csv file that constains stats on contiguing success/failure: #merged pairs, total discrepant bases, total unresolved base discrepancies, #alignment error, unaligned pairs (no overlap between mate pairs), #merged pairs with at least 1 low-quality discrepant base between strands, #merged pairs with at least 1 low-quality bases (discrepant or not)

4. run the microassembler on the compute cluster by executing the shell script

i) change file permissions (chmod 755 runJM.sh)
ii) ssh to head node (apollo as of 03/2010)
iii) run:
apollo>mqsub --file /FULLPATH_TO_YOUR_WORKING_DIRECTORY/runJM.sh

This is lengthy and should take ~12h to complete.

5. make a shell script to mine sequence contigs for TRBV

A) First, make sure all *.log and *.fa files exists for all lane (~120 files for each) when the run is done:

```
ls -la | grep -c "log"
this should return 120 and make sure the #byte <>0 for each 120
```

B) Make shell script:

```
ls -la | perl -ne 'if(/(s_\d_1_\d+\_qseq.txt.fa)/){print
"/projects/02/rwarren/solexa/TCRb/participants/EScon.pl $1 RW serial\n";}' >
runES.sh
```

EScon.pl

Usage: ./EScon.pl <query fasta> <base name for files> <serial/parallel> <pipeline name (optional) -clean,finish- >

<query fasta>	this is the fasta file produced for each tile and comprises joined pairs
<base name for files>	this is used by the script as a unique job identifier. It can

	be anything you would like (but short, 2-3 characters).
<serial/parallel>	if set to serial, will run on the shell server you're logged into (recommended). If set to parallel, will attempt to run on compute cluster (not supported)
<pipeline name>	please leave empty

*A single file named "CANDIDATES.txt" will be created. Before running below, make sure that no such file exists in your working directory because it will be appended to.
*The "CANDIDATES.txt" file is a text file whose structure is akin of a fasta file, with a

>header
SEQUENCE:ASCII+66(ILLUMINA)ENCODED QUALITY:COVERAGE:TRBV segments alignments (CIGAR FORMAT)

It is not meant to be human readable: it is used as input for TCRmine.pl

6. execute the TRBV mining script on a shell server

change permissions (chmod 755 runES.sh), ssh to your favorite shell server and run:
xhost10>screen ./runES.sh

That script will create a file named "CANDIDATES.txt" which contains each contig with at least 1 detectable TRBV

INDEXED SAMPLES USING BARCODES

Run scripts below if you are processing a PCR-indexed library, otherwise ignore this section and forward to step 7)

i) bin TCRB sequence candidates using barcodes

binCandidatesContigs.pl

Usage: ./binCandidatesContigs.pl <CANDIDATES.txt> <barcode-sample.tsv> <5'-3' sequence upstream barcode (expected just before barcode) > <quality filtering? 1=yes 0=no>

e.g.
xhost10> ./binCandidatesContigs.pl CANDIDATES.txt barcode-sample.tsv GTCGCT 0
Format from barcode-sample.tsv file ::barcode_sequence<space>sample

<CANDIDATES.txt>	see above description

*If no specific barcodes were discovered in a contig, the files insert_<run/sample/contig>SAMPLE.sql will be missing

7. mine candidates contigs for TCRB rearrangements
* ignore if you ran an indexed library with barcodes (above)

./TCRmine.pl CANDIDATES.txt AG 9 9 1

TCRmine.pl
<CANDIDATES.txt> <First 2 bases of C segment (AG)> <sample id> <run id> <adjust V end? (optional)>

<CANDIDATES.txt>	See description above
<First 2 bases of C segment (AG)>	The first 2 bases of TRBC (AG)—used to calculate the frame
<sample id>	The MySQL sample ID associated with this data
<run id>	The MySQL run id associated with this data
<adjust V end?>	If set to 1, the script will try a range of regex to narrow in on the TRBV boundary that may have been erroneously set by Exonerate alignments (highly recommended)

TCRmine.pl will output:

insert_contig.sql	contains all the unique rearrangements on one line, with tons of additional information (see format below)
insert_run.sql	contains the run id and limited information it is up to the user to enter the run into MySQL (see format below)
insert_sample.sql	contains the sample id and limited information it is up to the user to enter the run into MySQL (see format below)
trackContigs.csv	a coma-separated file (.csv) that lists each contig and rearrangement contained within it (see format below)

*If no specific barcodes were discovered in a contig, the files insert_<run/sample/contig>SAMPLE.sql will be missing

7. mine candidates contigs for TCRB rearrangements
* ignore if you ran an indexed library with barcodes (above)

./TCRmine.pl CANDIDATES.txt AG 9 9 1

TCRmine.pl
<CANDIDATES.txt> <First 2 bases of C segment (AG)> <sample id> <run id> <adjust V end? (optional)>

<CANDIDATES.txt>	See description above
<First 2 bases of C segment (AG)>	The first 2 bases of TRBC (AG)—used to calculate the frame
<sample id>	The MySQL sample ID associated with this data
<run id>	The MySQL run id associated with this data
<adjust V end?>	If set to 1, the script will try a range of regex to narrow in on the TRBV boundary that may have been erroneously set by Exonerate alignments (highly recommended)

TCRmine.pl will output:

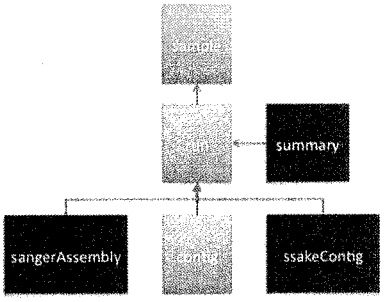
insert_contig.sql	contains all the unique rearrangements on one line, with tons of additional information (see format below)
insert_run.sql	contains the run id and limited information it is up to the user to enter the run into MySQL (see format below)
insert_sample.sql	contains the sample id and limited information it is up to the user to enter the run into MySQL (see format below)
trackContigs.csv	a coma-separated file (.csv) that lists each contig and rearrangement contained within it (see format below)

8. MySQL and output files

Not part of the pipeline per se, but useful information:

i) To build a brand new MySQL database, run:

xhost10>mysql -u USERNAME -pPASSWORD -h HOST -D DATABASE_NAME < /home/rwarren/code/TCRb/mysql/iSSAKE.sql



iSSAKE MySQL database

ii) insert_contig.sql file structure:

Field	Type	Null	Key	Default	Extra
id	int(10) unsigned		PRI	NULL	auto_increment
FK_run__id	int(10) unsigned		MUL	0	
ntSeq	varchar(200)	YES	MUL	NULL	
depth	mediumint(8) unsigned			0	
rearrangement	varchar(200)	YES	MUL	NULL	
ntCDR3	varchar(100)				
aaCDR3	varchar(100)				
ntSeqShort	varchar(200)	YES		NULL	
aaSeqShort	varchar(200)	YES		NULL	
ntSeqLong	varchar(250)	YES		NULL	
aaSeqLong	varchar(250)	YES		NULL	
vName	varchar(15)	YES		NULL	
vDeleted	smallint(5) unsigned	YES		NULL	
vEnd	smallint(5) unsigned	YES		NULL	
vFrame	smallint(5) unsigned	YES		NULL	
jName	varchar(15)	YES		NULL	
jDeleted	smallint(5) unsigned	YES		NULL	
jRestSeq	varchar(200)	YES		NULL	
frameCheck	smallint(5) unsigned			0	
vPossible	mediumint(8) unsigned	YES		NULL	
ntTCRB	text				
aaTCRB	text				

*some of these fields were created in anticipation of certain data sets and are no longer in used. Do not be surprised if there are no corresponding data in insert_contig.sql : it is intentional

Provided you have a MySQL database set up (see i), you can add the data to it by typing:

xhost10>mysql -u rwarren -pviewer -h athena -D issake < insert_contigCANDIDATES.sql

if you have other .sql files, you could run the above line accordingly.

iii) insert_run.sql file structure:

Field	Type	Null	Key	Default	Extra
id	int(11)		PRI	NULL	auto_increment
FK_sample__id	smallint(5) unsigned		MUL	0	
parameters	varchar(80)	YES		NULL	
description	text	YES		NULL	
date	datetime			0000-00-00 00:00:00	

iv) insert_sample.sql file structure:

Field	Type	Null	Key	Default	Extra
id	int(10) unsigned		PRI	NULL	auto_increment
name	varchar(40)		UNI		
source	varchar(80)	YES		NULL	
description	text	YES		NULL	
date	datetime	YES		NULL	

iv) trackContigs.csv file structure

Rearrangement,ntCDR3,aaCDR3,ContigName,ContigSeq,TRBV,VbaseDel,ValignEnd,TRBJ,JbaseDel,FrameCheck,vPossible

e.g.
*_*_ggttcaggg_7_TRBJ1-2,TGTGCCAGCAGTTTggttcagggTGGCTACACCTTC,CASSlvqgGYTF,5-1-1063-

10592j,TGGACTCAGCTGTGTACTTCTGTGCCAGCAGTTTGGTTCAGGGTGGCTACACCTT
CGGTT
CGGGGACCAGGTTAACCGTTGTAGAGGACCTGAACAAGGTGTTCCCACCCGTTAA,*,*,*,T
RBJ1-2,7,1,99,

AUTHOR

Rene Warren 2008-2011

SEE ALSO

Warren RL, Freeman JD, Zeng T, Choe G, Munro S, Moore R, Webb JR, Holt RA. 2011. Exhaustive T-cell repertoire sequencing of human peripheral blood samples reveals signatures of antigen selection and a directly measured repertoire size of at least 1 million clonotypes. Genome Res. doi:10.1101/gr.115428.110

Freeman JD, Warren RL, Webb JR, Nelson BH, Holt RA. 2009. Profiling the T-cell receptor beta-chain repertoire by massively parallel sequencing. Genome Res. 19:1817-1824

Appendix B. TCRβ sequence mining

Included in this appendix are examples of commands used to mine the TCRβ sequence repertoire:

i. TCRβ sequence diversity and abundance present in a single patient: (MySQL database)

Access MySQL database:

[lraeburn@xhost09 ~]$ mysql -u lraeburn –p password -h host -D database

Example of commonly used command and output from command:

mysql> select unique(ntCDR3), sum(depth) from run,contig,sample where FK_run__id like run.id and FK_sample__id like sample.id and frameCheck = 1 and vpossible = 1 and depth > 1 and name = "01-T" group by ntCDR3 ;

ntCDR3	sum(depth)
TGCAGCGTTGAAGggacaaaaAACACCGGGGAGCTGTTTTTT	3
TGCAGCGTTGcccctctgcttcagggggACACTGAAGCTTTCTTT	2
TGCAGgaagagtccgggacaggggtcTGGCTACACCTTC	2
TGCAGTGCAACCGAtagcgggagctggggCTACGAGCAGTACTTC	2
TGCAGTGCCAGtagTAACTATGGCTACACCTTC	2
TGCAGTGCccatctagcgggatCGAACACCGGGGAGCTGTTTTTT	2
TGCAGTGCcccgggactaaataCCTACAATGAGCAGTTCTTC	2
TGCAGTGCcctcggggagacaggggCGAGCAGTACTTC	2042
TGCAGTGCcctctgggagacaggggCGAGCAGTACTTC	3
TGCAGTGCcctttctggtatAGATACGCAGTATTTT	2
TGCAGTGCTAaaacaggggtaCAAGAGACCCAGTACTTC	4
.	.
.	.
.	.
TGTGCCATCAGTGAGgacagAAAAACATTCAGTACTTC	7
TGTGCCATCAGTGAGTCgcgagatggTGGCTACACCTTC	7
TGTGCCATCAGTGAGTCggctaacaggggctATGAGCAGTTCTTC	2
TGTGCCATCAGTGgcgagcgggagaaCTACGAGCAGTACTTC	2
TGTGCCATCAGTGtagggggggACGAGCAGTACTTC	2
TGTGCCATCAtcggacagtTGAACACTGAAGCTTTCTTT	2
126 rows in set (4.89 sec)	

Upper case letters are nucleotide sequences belonging to the TRBV-gene (left) and TRBJ gene (right) of a given CDR3β sequence and coding from the last conserved cysteine of the TRBV-gene to the first conserved phenylalanine og the TRBJ gene in the motif FXFG. Lower case letters represent non-TRBV, non-TRBJ bases. They comprise a mixture of TRBD and/or non-templated bases added back by the Terminal deoxynucleotidyl transferase (TdT). Note only 17 of 126 sequences shown from output.

ii. List all CDR3β amino acid sequences where depth is 100 or more in 2 or more samples:

Example of command, followed by output:

File path to working directory: cd projects/lraeburn/.../qseqshare

[lraeburn@xhost09 qseqshare]$ cat aaCDR3.csv |perl -ne 'chomp;@a=split(/,/);my $ct=0;@l;for(my$x==1;$x<=43;$x++){if($a[$x]>=100){$ct++;push @l,$x;}}print "$_\n" if($ct>=2); '

Sequence, **Patient**
1,2,3,4,5,6,7,8,9,10,11,12,14,21,22,24,28,30,31,32,33,34,35,36,37,38,40,43,45,47,49,50,52,56,59,63,69,71,74,75,78,79

CASSLGQGAYEQFF,5,30,88,24,95,73,17,22,40,33,20,10,202589,36,154,53,55,24,32,48,71,85,25,28,76,33,5,66,25,57,33,5,43,110,112,13,23,111,11,21,39,8,29,

CAISEGQKNIQYF,7,17,50,18,69,60,13,16,29,26,24,14,102,27,121,25,45,9,24,10,42,76470,17,23,58,21,7,73,16,52,14,22,29,79,58,17,32,82,19,11,29,7,60,

CSARAPDGNTGELFF,0,0,6,0,9,12,2,0,11180,320,2,2,11,0,10,5,6,0,0,3,3,9,3,2,5,3,0,8,4,7,0,0,3,5,7,0,7,14,2,0,0,2,6,

CASSQGVGPRKQYF,0,2,3,0,3,5,0,2,2,0,0,2,5,3,5,6,288,0,3,6,8088,6,3,0,4,0,0,6,0,6,0,0,5,3,3,0,0,3,0,0,0,0,5,

CASSLGQGGTGELFF,0,0,7,3,2715,7,0,0,2,0,0,2,11,4,20,3,4,2,6973,6,4,6,5,2,7,2,0,8,5,3,7,0,0,10,11,0,2,14,0,0,3,0,6,

CASSSLLAGGGNIQYF,0,0,7,0,4,7,0,2,5,0,2,2,14,4,15,2,4,6,3,6,5,9,3,2,7,6,0,10,2,7,4,2,3,6,361,0,4,14,2,0,5882,3744,2207,

CASSLEGGTYNEQFF,2,3,7,4,7,10,0,2,3,3,5,2,23,2,24,7,7,0,3,3,9,5,4,5,6,4978,2,10,4,7,0,5,4,15,14,0,9,14,3,3,336,0,2878,

CSAPNPSGLLYNEQFF,0,2,5,0,9,8,0,0,4,4,0,0,19,0,10,9,5,0,3,4,5,12,0,2,11,830,0,8,335,2,3,4,0,13,6,0,5,9,0,0,191,0,3842,

CASSVRASYEQYF,0,3,0,0,0,6,0,0,0,0,0,0,4,0,5,0,0,0,0,4,6,3,2,0,3,1538,0,3,5,0,4,0,0,3,4,0,0,5,3,0,2723,2,0,

CAISESGTTYEQYF,0,0,0,2,4,4,2,0,0,0,0,0,8,0,7,0,0,0,0,3,3,4,3,0,5,0,0,6,0,0,0,0,4,2,3,0,2,4,0,2,1850,0,1097,

CASSRTSGANTGELFF,2,2,3,0,2,3,0,0,0,4,0,0,4,0,11,4,3,0,0,2,2,4,3,0,3,636,3,0,0,2,0,0,0,2,2,0,0,4,2,2,514,3,1800,

CSGVIGNTIYF,0,0,2,2,3,0,0,2,0,0,0,0,4,0,0,0,2,0,0,2,0,3,0,0,0,552,0,2,0,3,0,0,0,3,2,0,0,2,0,0,0,1636,0,

CASSLTRNTEAFF,0,0,0,0,2,3,0,0,0,0,0,0,0,0,2,0,4,0,0,0,3,2,2,2,0,0,0,4,3,0,0,0,0,2,638,0,0,2,0,0,943,0,0,

CASSEVGGSNEQFF,0,20,81,10,72,86,8,18,30,17,9,21,141,24,135,37,40,12,26,46,72,62,17,12,58,26,8,63,29,56,32,24,26,91,72,9,35,123,17,19,28,18,19,

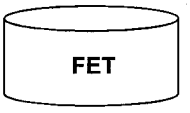
Appendix C. Fisher's Exact Test Perl Script

Perl script was written to perform a bootstrapped Fishers Exact Test (FET) on sequence data from all samples in the CRC cohort. Brief overview of how script works shown on pages 88-89, followed by actual Perl script to run FET on pages 90-92.

1. Configure Input File

aaCDR3	Patient ID	Alive Status (1 = deceased)
CASsrqtghTGELFF	30	1
CSalgetggEQYF	30	1
CSARvpqgayQPQHF	30	1
CAIkrgqeTDTQYF	30	1
CAIlgdSGANVLTF	30	1
CAISEsagGELFF	31	0
CAISEttggtNEKLFF	31	0
CAISpktgqrgTGELFF	31	0
CASatgvgTDTQYF	31	0
CASGyweETQYF	31	0
CAShtsnNQPQHF	31	0
CASkrgeaGNTIYF	31	0
CASnertgmNTEAFF	31	0
CASNTdyrkKLFF	31	0
CASregtggGYTF	31	0
CASRKsNEKLFF	31	0
CASrsprtggeNTEAFF	31	0
CASSaeqkrgrTDTQYF	31	0
CASSAgdsGELFF	32	1
CASSardrgpaYNEQFF	32	1
CASSEgsyQETQYF	32	1
CASSfsqggETQYF	32	1

2. Run through FET program (with bootstrapping)



3. Output file

CDR3 (aa)	Number Present	Original Fisher P-value	Bootstrapped Fisher P-value
caakagtantgelff	3	1.000	0.887
caaragpnteaff	9	0.051	0.040*
caarpigalinygytf	1	1.000	0.625
caartvfvgntiyf	1	1.000	0.610
cacdssganvltf	2	0.506	0.447
.			
.			
.			

* = significant association with survival (p-value < 0.05)

. 1. Configure the input file from data in MySQL database. All diversity (all unique sequences) from each patient was compiled into an input file alongside the overall alive status of that patient (1= deceased, 0 = alive). All sequences included in the input file were in-frame, with one V-gene and were found at a depth of coverage of > 1. 2. Input file loaded into FET program. 3. Output from FET program run included all unique sequences, number of patients sequence was present in, original FET p-value based on actual data and bootstrapped FET p-value based on randomly assigning alive or dead to each patient and running script 1000 times.

Script of Fishers Exact Test Program (written in PERL):

```
###Location of script:  /projects/Iraeburn_prj/FishersTests/fisherdev.pl

#!/home/martink/bin/perl

###Define operating context
use strict ;

###Define Fisher's Exact Test Module to use
##(CPAN: http://search.cpan.org/dist/Text-NSP/lib/Text/NSP/Measures/2D/Fisher/twotailed.pm)
use Text::NSP::Measures::2D::Fisher::twotailed;

###Usage of script                    #ARGV - list of arguments being passed into the program
if($#ARGV<1){
   die "Usage: $0 <*.txt> <bootstrap n times>\n";
#eg. fisherdev.pl inputfile.txt 1000
}

###Specification of lines to input
my $input_file = $ARGV[0];            #specify files as input
my $nboot = $ARGV[1];
if(! $nboot) {
   die " No bootstrap value has been specified\n";
}

###Open file
open(IN,$input_file) || die "Can't open $input_file -- fatal.\n";

my %aainsample;
my %salive;
my %aatable;
my $aa ;

###Read each line of the file
while(<IN>) {
   chomp;
   my $line   = $_;
   my @fields = split(" ",$line);
   my ($aa,$sid,$alive) = @fields[0,1,2];
   $aa = lc $aa;

   $aatable{$aa}++;

   $aainsample{ $sid } { $aa } ++;
##Kill the script if there is inconsistency in the alive column for each patient
   if(defined $salive{$sid}) {
      if($salive{$sid} != $alive) {
         die "problem with consistency of alive record for $sid [$alive] [$salive{$sid}]";
      }
   } else {
      $salive{$sid} = $alive;
   }
}
close IN;
```

```
###Print out headings
print "CDR3 (aa), Number Present, Original Fisher P-value, Bootstrapped Fisher P-value\n";

###Calculate Fisher Stats
for my $aa (sort keys %aatable) {        #sorts the keys from %aatable hash

my $fisherpval = calculateFisher($aa,\%aainsample,\%salive, 1);
#if($nboot){                             #remove this hash to only bootstrap for p-values < 0.05

#if($fisherpval<=.05){                   #remove this hash to only bootstrap for p-values < 0.05

###Bootstraping
 my $ct=0;
       for(my $i=1;$i<=$nboot;$i++){     # bootstrapping loop

       my %rdm_salive = ();

       my $range = 2 ;                   #randomly apply 0 or1 to rows of a patient for bootstrap
       foreach my $sample_id (keys %salive){
              my $random_alive = int(rand($range));
        $rdm_salive{$sample_id} = $random_alive;
     }

    my $nfisherpval = calculateFisher($aa,\%aainsample,\%rdm_salive, 0);

        if($nfisherpval <= $fisherpval){
          $ct++;
        }
    }
         my $fisherbootpval = $ct/$nboot;

printf "$aa, $aatable{$aa}, $fisherpval, $fisherbootpval\n"  ;

 #    }                                  #remove this hash to only bootstrap for p-values < 0.05
#}                                       #remove this hash to only bootstrap for p-values < 0.05
}

#####SUBROUTINE: FET CALCULATION#####

sub calculateFisher  {
   my ($aa,$aainsample,$salive,$null) = @_;
   my $aalookup = ($aa);
   my %stats;
   my @statnames = qw(pa pd aa ad);
   map { $stats{$_} = 0 } @statnames;
   for my $sid (keys %$aainsample) {
      if( $aainsample->{$sid}{$aalookup} ) {
         if($salive->{$sid}) {
            $stats{pa}++;
         } else {
            $stats{pd}++;          }
      } else {
         if($salive->{$sid}) {
            $stats{aa}++;
```

```
} else {
            $stats{ad}++;
          }
       }
    }

    my $n11 = $stats{pa};
    my $n21 = $stats{pd};
    my $n12 = $stats{aa};
    my $n22 = $stats{ad};

    my $n1p = $n11 + $n12 ;
    my $n2p = $n21 + $n22 ;
    my $np1 = $n11 + $n21 ;
    my $np2 = $n22 + $n12 ;
    my $npp = $n1p+ $n2p ;


 my  $fisherpval = calculateStatistic( n11=>$n11,
                                n1p=>$n1p,
                                np1=>$np1,
                                npp=>$npp);
}
```

Appendix D. Primer and adapter sequences

Table D1 Reverse Transcription Primers

Primer	Sequence (5' – 3')
C6	CACGTGGTCGGGGWAGAAGC[57]
LTS (TSP)	AAGCAGTGGTAACAACGCAGAGTACGCGGG[57]

Table D2 Indexed Primers used in library construction

Primer	Sequence (5' – 3')
UPM Short	CTAATACGACTCACTATAGGGC
UPM Long	CTAATACGACTCACTATAGGGCAAGCAGTGGTAACAACGCAGAGT
iGSPs	
C-IDX_01	GGCCACGCGTCGACTAGTTAATTAACGTGATAGCGACCTCGGGTGGGAACA
C-IDX_02	GGCCACGCGTCGACTAGTTAATTAACTGATCAGCGACCTCGGGTGGGAACA
C-IDX_03	GGCCACGCGTCGACTAGTTAATTAAGGGGTTAGCGACCTCGGGTGGGAACA
C-IDX_04	GGCCACGCGTCGACTAGTTAATTAACTGGGTAGCGACCTCGGGTGGGAACA
C-IDX_05	GGCCACGCGTCGACTAGTTAATTAAAGCGCTAGCGACCTCGGGTGGGAAA
C-IDX_06	GGCCACGCGTCGACTAGTTAATTAACTTTTGAGCGACCTCGGGTGGGAACA
C-IDX_07	GGCCACGCGTCGACTAGTTAATTAATGTTGGAGCGACCTCGGGTGGGAACA
C-IDX_08	GGCCACGCGTCGACTAGTTAATTAAAGCTAGAGCGACCTCGGGTGGGAACA
C-IDX_09	GGCCACGCGTCGACTAGTTAATTAAAGCATCAGCGACCTCGGGTGGGAACA
C-IDX_10	GGCCACGCGTCGACTAGTTAATTAACGATTAAGCGACCTCGGGTGGGAACA
C-IDX_11	GGCCACGCGTCGACTAGTTAATTAACATTCAAGCGACCTCGGGTGGGAACA
C-IDX_12	GGCCACGCGTCGACTAGTTAATTAAGGAACTAGCGACCTCGGGTGGGAACA
C-IDX_13	GGCCACGCGTCGACTAGTTAATTAAACATCGAGCGACCTCGGGTGGGAACA
C-IDX_14	GGCCACGCGTCGACTAGTTAATTAAAAGCTAAGCGACCTCGGGTGGGAACA
C-IDX_15	GGCCACGCGTCGACTAGTTAATTAACAAGTTAGCGACCTCGGGTGGGAACA
C-IDX_16	GGCCACGCGTCGACTAGTTAATTAAGCCGGTAGCGACCTCGGGTGGGAACA
C-IDX_17	GGCCACGCGTCGACTAGTTAATTAACGGCCTAGCGACCTCGGGTGGGAACA
C-IDX_18	GGCCACGCGTCGACTAGTTAATTAATAGTTGAGCGACCTCGGGTGGGAACA
C-IDX_19	GGCCACGCGTCGACTAGTTAATTAAGCGTGGAGCGACCTCGGGTGGGAACA

Primer	Sequence (5' – 3')
C-IDX_20	GGCCACGCGTCGACTAGTTAATTAAGTATAGAGCGACCTCGGGTGGGAACA
C-IDX_21	GGCCACGCGTCGACTAGTTAATTAACCTTGCAGCGACCTCGGGTGGGAACA
C-IDX_22	GGCCACGCGTCGACTAGTTAATTAAGCTGTAAGCGACCTCGGGTGGGAACA
C-IDX_23	GGCCACGCGTCGACTAGTTAATTAAATGGCAAGCGACCTCGGGTGGGAACA
C-IDX_24	GGCCACGCGTCGACTAGTTAATTAATGACATAGCGACCTCGGGTGGGAACA
C-IDX_25	GGCCACGCGTCGACTAGTTAATTAAGCCTAAAGCGACCTCGGGTGGGAACA
C-IDX_26	GGCCACGCGTCGACTAGTTAATTAAGTAGCCAGCGACCTCGGGTGGGAACA
C-IDX_27	GGCCACGCGTCGACTAGTTAATTAAAGTCTTAGCGACCTCGGGTGGGAACA
C-IDX_28	GGCCACGCGTCGACTAGTTAATTAATATCGTAGCGACCTCGGGTGGGAACA
C-IDX_29	GGCCACGCGTCGACTAGTTAATTAAAATTATAGCGACCTCGGGTGGGAACA
C-IDX_30	GGCCACGCGTCGACTAGTTAATTAACCGGTGAGCGACCTCGGGTGGGAACA
C-IDX_31	GGCCACGCGTCGACTAGTTAATTAACATGGGAGCGACCTCGGGTGGGAACA
C-IDX_32	GGCCACGCGTCGACTAGTTAATTAATCTGAGAGCGACCTCGGGTGGGAACA
C-IDX_33	GGCCACGCGTCGACTAGTTAATTAAAAGTGCAGCGACCTCGGGTGGGAACA
C-IDX_34	GGCCACGCGTCGACTAGTTAATTAAATTATAAGCGACCTCGGGTGGGAACA
C-IDX_35	GGCCACGCGTCGACTAGTTAATTAACCAGCAAGCGACCTCGGGTGGGAACA
C-IDX_36	GGCCACGCGTCGACTAGTTAATTAAGGACGGAGCGACCTCGGGTGGGAACA
C-IDX_37	GGCCACGCGTCGACTAGTTAATTAATGGTCAAGCGACCTCGGGTGGGAACA
C-IDX_38	GGCCACGCGTCGACTAGTTAATTAATACAAGAGCGACCTCGGGTGGGAACA
C-IDX_39	GGCCACGCGTCGACTAGTTAATTAATCGCTTAGCGACCTCGGGTGGGAACA
C-IDX_40	GGCCACGCGTCGACTAGTTAATTAAGAGAGTAGCGACCTCGGGTGGGAACA
C-IDX_41	GGCCACGCGTCGACTAGTTAATTAACCGTATAGCGACCTCGGGTGGGAACA
C-IDX_42	GGCCACGCGTCGACTAGTTAATTAAATCGTGAGCGACCTCGGGTGGGAACA
C-IDX_43	GGCCACGCGTCGACTAGTTAATTAACCACTCAGCGACCTCGGGTGGGAACA
C-IDX_44	GGCCACGCGTCGACTAGTTAATTAACAGCAGAGCGACCTCGGGTGGGAACA
C-IDX_45	GGCCACGCGTCGACTAGTTAATTAACGCGGCAGCGACCTCGGGTGGGAACA
C-IDX_46	GGCCACGCGTCGACTAGTTAATTAAGAATGAAGCGACCTCGGGTGGGAACA
C-IDX_47	GGCCACGCGTCGACTAGTTAATTAAGCGCCAAGCGACCTCGGGTGGGAACA
C-IDX_48	GGCCACGCGTCGACTAGTTAATTAACTCTACAGCGACCTCGGGTGGGAACA
C-IDX_49	GGCCACGCGTCGACTAGTTAATTAACACTGTAGCGACCTCGGGTGGGAACA
C-IDX_50	GGCCACGCGTCGACTAGTTAATTAAATGTTTAGCGACCTCGGGTGGGAACA

Primer	Sequence (5' – 3')
C-IDX_51	GGCCACGCGTCGACTAGTTAATTAAGTCCTTAGCGACCTCGGGTGGGAACA
C-IDX_52	GGCCACGCGTCGACTAGTTAATTAAATCAGTAGCGACCTCGGGTGGGAACA
C-IDX_53	GGCCACGCGTCGACTAGTTAATTAATAGGATAGCGACCTCGGGTGGGAACA
C-IDX_54	GGCCACGCGTCGACTAGTTAATTAATGAGTGAGCGACCTCGGGTGGGAACA
C-IDX_55	GGCCACGCGTCGACTAGTTAATTAATTGCGGAGCGACCTCGGGTGGGAACA
C-IDX_56	GGCCACGCGTCGACTAGTTAATTAAGGTTTCAGCGACCTCGGGTGGGAACA
C-IDX_57	GGCCACGCGTCGACTAGTTAATTAATAAGGCAGCGACCTCGGGTGGGAACA
C-IDX_58	GGCCACGCGTCGACTAGTTAATTAATCGGGAAGCGACCTCGGGTGGGAACA
C-IDX_59	GGCCACGCGTCGACTAGTTAATTAATTCGAAAGCGACCTCGGGTGGGAACA
C-IDX_60	GGCCACGCGTCGACTAGTTAATTAAGCGGACAGCGACCTCGGGTGGGAACA
C-IDX_61	GGCCACGCGTCGACTAGTTAATTAAATTGGCAGCGACCTCGGGTGGGAACA
C-IDX_62	GGCCACGCGTCGACTAGTTAATTAATGCTTTAGCGACCTCGGGTGGGAACA
C-IDX_63	GGCCACGCGTCGACTAGTTAATTAACCTATTAGCGACCTCGGGTGGGAACA
C-IDX_64	GGCCACGCGTCGACTAGTTAATTAATCTTCTAGCGACCTCGGGTGGGAACA
C-IDX_65	GGCCACGCGTCGACTAGTTAATTAAATAGATAGCGACCTCGGGTGGGAACA
C-IDX_66	GGCCACGCGTCGACTAGTTAATTAACGCCTGAGCGACCTCGGGTGGGAACA
C-IDX_67	GGCCACGCGTCGACTAGTTAATTAACTAAGGAGCGACCTCGGGTGGGAACA
C-IDX_68	GGCCACGCGTCGACTAGTTAATTAATTATTCAGCGACCTCGGGTGGGAACA
C-IDX_69	GGCCACGCGTCGACTAGTTAATTAATGGAGCAGCGACCTCGGGTGGGAACA
C-IDX_70	GGCCACGCGTCGACTAGTTAATTAACTTCGAAGCGACCTCGGGTGGGAACA
C-IDX_71	GGCCACGCGTCGACTAGTTAATTAAGGAGAAAGCGACCTCGGGTGGGAACA
C-IDX_72	GGCCACGCGTCGACTAGTTAATTAATTTCACAGCGACCTCGGGTGGGAACA
C-IDX_73	GGCCACGCGTCGACTAGTTAATTAAGATCTGAGCGACCTCGGGTGGGAACA
C-IDX_74	GGCCACGCGTCGACTAGTTAATTAAGCATTTAGCGACCTCGGGTGGGAACA
C-IDX_75	GGCCACGCGTCGACTAGTTAATTAAGTTTGTAGCGACCTCGGGTGGGAACA
C-IDX_76	GGCCACGCGTCGACTAGTTAATTAACTATCTAGCGACCTCGGGTGGGAACA
C-IDX_77	GGCCACGCGTCGACTAGTTAATTAAGCTCATAGCGACCTCGGGTGGGAACA
C-IDX_78	GGCCACGCGTCGACTAGTTAATTAAGCCATGAGCGACCTCGGGTGGGAACA
C-IDX_79	GGCCACGCGTCGACTAGTTAATTAATTCTCGAGCGACCTCGGGTGGGAACA
C-IDX_80	GGCCACGCGTCGACTAGTTAATTAATCCGTCAGCGACCTCGGGTGGGAACA
C-IDX_81	GGCCACGCGTCGACTAGTTAATTAATGTGCCAGCGACCTCGGGTGGGAACA

Primer	Sequence (5' – 3')
C-IDX_82	GGCCACGCGTCGACTAGTTAATTAATGCCGAAGCGACCTCGGGTGGGAACA
C-IDX_83	GGCCACGCGTCGACTAGTTAATTAAAAACCTAGCGACCTCGGGTGGGAACA
C-IDX_84	GGCCACGCGTCGACTAGTTAATTAAGGCCACAGCGACCTCGGGTGGGAACA
C-IDX_85	GGCCACGCGTCGACTAGTTAATTAATCAAGTAGCGACCTCGGGTGGGAACA
C-IDX_86	GGCCACGCGTCGACTAGTTAATTAACGTACGAGCGACCTCGGGTGGGAACA
C-IDX_87	GGCCACGCGTCGACTAGTTAATTAAAGATGTAGCGACCTCGGGTGGGAACA
C-IDX_88	GGCCACGCGTCGACTAGTTAATTAAGATGCTAGCGACCTCGGGTGGGAACA
C-IDX_89	GGCCACGCGTCGACTAGTTAATTAAAGGAATAGCGACCTCGGGTGGGAACA
C-IDX_90	GGCCACGCGTCGACTAGTTAATTAAAAAATGAGCGACCTCGGGTGGGAACA
C-IDX_91	GGCCACGCGTCGACTAGTTAATTAAATTCCGAGCGACCTCGGGTGGGAACA
C-IDX_92	GGCCACGCGTCGACTAGTTAATTAATATATCAGCGACCTCGGGTGGGAACA
C-IDX_93	GGCCACGCGTCGACTAGTTAATTAACAGGCCAGCGACCTCGGGTGGGAACA
C-IDX_94	GGCCACGCGTCGACTAGTTAATTAAGGTAGAAGCGACCTCGGGTGGGAACA
C-IDX_95	GGCCACGCGTCGACTAGTTAATTAATTGACTAGCGACCTCGGGTGGGAACA
C-IDX_96	GGCCACGCGTCGACTAGTTAATTAACGAAACAGCGACCTCGGGTGGGAACA

Underlined portion shown in C-IDX_01 shows location unique index sequence.

Table D3 Nested PCR Primers

Primer	Sequence (5' – 3')
BiotinTail	/5Biosg/GGCCACGCGTCGACTAGTTA
SN2	ACGACTCACTATAGGGCAAGCAG

Table D4 Adapters and Primers for Illumina PE Sequencing

Primer	Sequence
PE adapters	5' ACACTCTTTCCCTACACGACGCTCTTCCGATCT
	3' GAGCCGTAAGGACGACTTGGCGAGAAGGCTAG-5Phos
PE PCR Primers 1 and 2	5' AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT
	5' CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCT
PE Sequencing Primers	5' ACACTCTTTCCCTACACGACGCTCTTCCGATCT
	5' CGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCT

Table D5 qPCR Primer/probe sequences

Probe	TCRβ sequence	Primer/probe placement on nucleotide sequence
CSapnpsgllYNEQFF	TRBV20-1_6_cccgaaccctagcgggctctta_4_TRBJ2-1	5'tcct**{gaagacagcagcttctacatcTG}** CAGTGC**[cccgaaccctagcgggctctt]{aTACAATGAGCAGTTCTTCggg }**ccagggacacggctcaccgtgctag3'

Sequence (black, bold) within ' {}' represents locations where primers will anneal. Sequence (red, bold) within '[]' represents location where probe will anneal