BC Cancer Agency
CARE & RESEARCH
An agency of the Provincial Health Services Authority

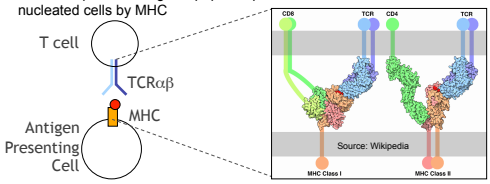
Profiling T Cell Metagenomes with iSSAKE

René L Warren and Robert A Holt

Canada's Michael Smith
Genome Sciences Centre

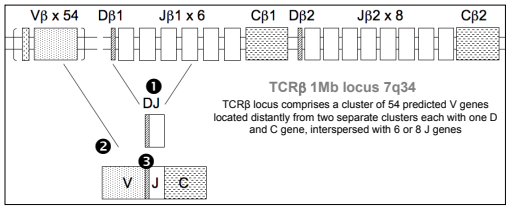
T Cells

- T cells are central to cell-mediated immunity. They are distinguished from other lymphocytes, such as B cells, by the presence of ***the T cell receptor (TCR)*** on its cell surface
- T cells respond to antigenic peptides presented at the surface of nucleated cells by MHC



T Cell Receptors (TCR)

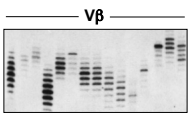
- TCR is a heterodimer and each chain includes a variable (V), joining (J) and a constant (C) segment. TCRβ also has a diversity (D) domain
- Structural TCR diversity necessary for recognition of enormous number of potential antigens generated by:
  1. – DJ gene recombination at the DNA level
  2. – One of the V genes joins DJ and the intermediary DNA is deleted
  3. – During the gene rearrangements, random base addition at the V-D-J junction and frequent base deletion V-3', 5'-J and either side of D yield the CDR3
- At V-(D)-J junction, the 3rd complementarity-determining region (CDR3) of TCR interacts directly with MHC-bound antigenic peptides
- CDR3 is the most critical TCR structure in epitope recognition
  - – $>10^{18}$ theoretically possible αβTCRs and est. ~10^7 T cell clonotypes in a given individual at a given time (Arstila *et al.* 1999)



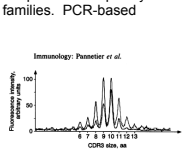
Genomic re-arrangement creates a diverse T cell metagenome in every individual

Current TCR Profiling

- Spectratyping: Analysis of TCR β-chain repertoire complexity based on CDR3 length diversity within Vβ gene families. PCR-based



Spectratype from adult blood lymphocytes (Gorski *et al.* 1994)

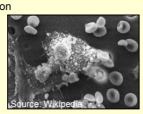


Immunoscope from mouse thymus (Pannetier *et al.* 1993)

- Spectratypes (immunoscopes) provide two levels of information: band number and band (fluorescence) intensity pattern
- Do not allow specific identification of a single T cell clone
- TCR sequence repertoire previously inaccessible due to cost of sequencing

Objectives

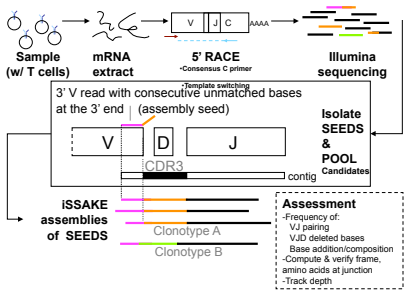
- Develop an approach for sequence profiling entire human T cell repertoire from blood
  - – Deep Illumina sequencing of T cell receptor variable region
  - – *De novo* assembly using SSAKE (Warren *et al.* 2007)



Motivations

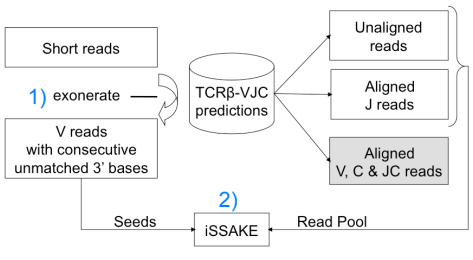
- Current profiling methods not quantitative
- Current sequence profiling are low-scale and costly
  - – hundreds clonotypes Vs. ~10^7 possible
  - – 10-fold coverage of 1 Million 300 bp target sequences: $3M Vs. $3K (Sanger Vs. Illumina)
- Applications include profiling cellular immune response to cancer, infection, vaccination, etc.
  - – Compare profiles in sick Vs. healthy / recurrent Vs. non-recurrent tumors
  - – Reconstitution of immune system after bone marrow replacement
- Next-generation sequencing make large-scale sequence profiling of T cell metagenomes a possibility, not without challenges due to read error and size

TCR mRNA Sequencing



TCR Assembly Strategy

- Designed a strategy for TCR sequence-profiling with short-read data
  - – Method relies on: *1) sequence alignment* *2) de novo assembly*
- Reads aligning to the end of known TRBV genes and having consecutive unmatched bases in the adjacent CDR3 are used to seed iSSAKE *de novo* assemblies of non-templated CDR3



Modeling

- Models of sequence diversity for the TCR beta-chain CDR3 region were built using empirical data and used to simulate, at random, distinct TCR clonotypes at 1-20 parts per million

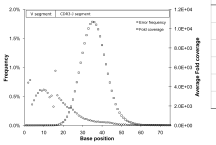
Simulation

- Using simulated TCRβ (sTCRβ) sequences, we randomly created 20 million 36nt reads having 1%-2% random error, 20 million 42nt or 50nt reads having 1% random error and 20 million 36nt reads with 1% error modeled on real short read data

Results

- With assembled 36nt reads we detect over 51% and 63% of rare (1 p.p.m.) clonotypes using a random or modeled error distribution

Over 99% of the more abundant clonotypes (6 p.p.m. or higher) can be detected using either error distribution



Average base error frequency and mean contig base coverage per position on trimmed, normalized, unambiguous contigs from 1% random error read assemblies. Assembly base mismatch frequency (i.e. assembly errors) reaches a maximum within the seed portion and decreases as coverage of J increases substantially

TCRβ reconstructions - general stats

Bases	Error (%) †	Mean seeds	Number of iSSAKE contigs from triplicates (36nt) or duplicates (42 and 50nt) experiments: Short ambiguous *	Long ambiguous ¥	Un-ambiguous §	Unambiguous, but sub-optimal ¶: A	B
36	1.0‡	1,599,140 ± 961	498,583 ± 822	36,352 ± 261	1,064,399 ± 400	1,098 ± 185	2,447 ± 89
36	1.0	1,755,437 ± 404	265,618 ± 245	11,520 ± 500	1,478,299 ± 226	157 ± 19	569 ± 5
36	1.5	1,718,470 ± 890	359,425 ± 882	16,460 ± 325	1,342,585 ± 1558	228 ± 44	827 ± 22
36	2.0	1,678,905 ± 1018	441,317 ± 712	22,018 ± 307	1,215,570 ± 385	302 ± 9	1,158 ± 53
42	1.0	2,448,351 ± 1298	369,847 ± 998	21,956 ± 315	2,056,549 ± 614	239 ± 14	818 ± 3
50	1.0	3,119,789 ± 1150	350,076 ± 578	341,628 ± 653	2,428,086 ± 2380	174 ± 12	459 ± 8

A Misassembled contigs

B Contigs having 5 or more mismatched bases

† Random error distribution generated using simulators from Dohm *et al.* (2007), unless otherwise specified

* Too short to unambiguously decipher J and thus, CDR3

¥ Contigs >= 45 nt, sufficiently long to contain the first 15 bases of J, but base errors/polymorphisms prevent proper identification of the J-segment

§ Captured CDR3 and first 15 bases of J unambiguously

¶ Misassembled contigs are defined here as contigs comprised of reads that belong to distinct sTCRβ. They are identified by looking at discontinuity in the sequence alignment between the contigs and sTCRβ. Contigs having 5 or more mismatches bases with the closest sTCRβ are built with erroneous reads that often yield misassembled contigs

‡ Error distribution modeled using pItch174 Illumina sequence data as the training set (Hong *et al.*, 2008).

Sensitivity, accuracy and clonotype frequency estimates

Error Distribution:

ppm	Number of sTCRβ-CDR3 characterized by iSSAKE contigs* (sensitivity) Error (%) 1.0 random	1.0 modeled	Accuracy (%) Error (%) 1.0 random	1.0 modeled	Average number of contigs characterizing each sTCRβ Error (%) 1.0 random	1.0 modeled
1	70,240	56,564	99.68	99.01	2.0	1.8
2	9,111	8,100	99.90	99.34	3.7	2.5
3	9,747	9,295	99.96	99.64	4.6	3.4
4	9,883	9,721	99.98	99.80	6.0	4.4
5	9,932	9,874	99.99	99.90	7.6	5.5
6	9,936	9,913	99.99	99.94	9.1	6.6
7	9,935	9,936	99.99	99.96	10.6	7.7
8	9,939	9,948	99.99	99.97	12.2	8.9
9	9,948	9,955	99.98	99.97	13.7	10.0
10	9,956	9,958	99.99	99.98	15.2	11.2
15	9,972	9,975	99.99	99.98	23.0	16.8
20	9,955	9,958	99.98	99.98	30.7	22.1

Read Length:

ppm	Number of sTCRβ-CDR3 characterized by iSSAKE contigs* (sensitivity) Read length (nt) 42	50	Accuracy (%) Read length (nt) 42	50	Average number of contigs characterizing each sTCRβ Read length (nt) 42	50
1	90,210	104,155	99.74	99.75	2.4	2.9
2	9,737	9,914	99.93	99.95	4.3	5.4
3	9,911	9,959	99.98	99.98	6.4	8.0
4	9,937	9,963	99.99	99.98	8.5	10.7
5	9,948	9,966	99.99	99.98	10.7	13.2
6	9,944	9,966	99.98	99.98	12.8	15.8
7	9,941	9,974	99.98	99.97	14.8	18.3
8	9,948	9,975	99.98	99.97	17.0	20.8
9	9,954	9,979	99.98	99.98	19.1	23.2
10	9,960	9,983	99.98	99.98	21.1	25.7
15	9,973	9,985	99.99	99.98	31.1	37.2
20	9,958	9,976	99.98	99.98	40.0	47.4

Longer reads improve sensitivity, with assembled 42nt and 50nt reads identifying 82.0% and 94.7% of rare 1 p.p.m. clonotypes

Acknowledgements

Funding: GenomeBritishColumbia

References:

- Arstila PT *et al.* (1999) A direct estimate of the human T cell receptor diversity. *Science*, 286, 958-961.
- Gorski J *et al.* (1994) Circulating T cell repertoire complexity in normal individuals and bone marrow recipients analyzed by CDR3 size spectratyping. Correlation with immune status. *J. Immunol.*, 152, 5109-5119.
- Pannetier C *et al.* (1993) 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. *Proc. Natl. Acad. Sci. USA*, 90, 4319-4323.
- Warren RL *et al.* (2009) Profiling model T cell metagenomes with short reads. *Bioinformatics*. doi:10.1093/bioinformatics/btp010
- Warren RL *et al.* (2009) Assembling millions of short DNA sequences with SSAKE. *Bioinformatics*. 23:500-1 (epub dec06)