BIOINFORMATICS ORIGINAL PAPER Vol. 25 no. 4 2009, pages 458–464
doi:10.1093/bioinformatics/btp010

Sequence analysis

Profiling model T-cell metagenomes with short reads

René L. Warren[1,*], Brad H. Nelson[2] and Robert A. Holt[1]

[1]BC Cancer Agency, Michael Smith Genome Sciences Centre, 675 West 10th Avenue, Vancouver, BC V5Z 1L3 Canada and [2]BC Cancer Agency, Deeley Research Centre, 2410 Lee Ave, Victoria, BC V8R 6V5 Canada

Received on September 26, 2008; revised on November 28, 2008; accepted on January 1, 2009

Advance Access publication January 9, 2009

Associate Editor: Alfonso Valencia

ABSTRACT

Motivation: T-cell receptor (TCR) diversity in peripheral blood has not yet been fully profiled with sequence level resolution. Each T-cell clonotype expresses a unique receptor, generated by somatic recombination of TCR genes and the enormous potential for T-cell diversity makes repertoire analysis challenging. We developed a sequencing approach and assembly software (immuno-SSAKE or iSSAKE) for profiling T-cell metagenomes using short reads from the massively parallel sequencing platforms.

Results: Models of sequence diversity for the TCR β-chain CDR3 region were built using empirical data and used to simulate, at random, distinct TCR clonotypes at 1–20 p.p.m. Using simulated TCRβ (sTCRβ) sequences, we randomly created 20 million 36 nt reads having 1–2% random error, 20 million 42 or 50 nt reads having 1% random error and 20 million 36 nt reads with 1% error modeled on real short read data. Reads aligning to the end of known TCR variable (V) genes and having consecutive unmatched bases in the adjacent CDR3 were used to seed iSSAKE *de novo* assemblies of CDR3. With assembled 36 nt reads, we detect over 51% and 63% of rare (1 p.p.m.) clonotypes using a random or modeled error distribution, respectively. We detect over 99% of more abundant clonotypes (6 p.p.m. or higher) using either error distribution. Longer reads improve sensitivity, with assembled 42 and 50 nt reads identifying 82.0% and 94.7% of rare 1 p.p.m. clonotypes, respectively. Our approach illustrates the feasibility of complete profiling of the TCR repertoire using new massively parallel short read sequencing technology.

Availability: ftp://ftp.bcgsc.ca/supplementary/iSSAKE

Contact: rwarren@bcgsc.ca

Supplementary information: Supplementary methods and data are available at *Bioinformatics* online.

1 INTRODUCTION

Recognition of MHC (major histocompatibility complex)-presented antigen by the T-cell receptor (TCR) is a pivotal process in cell-mediated adaptive immunity. A vast TCR repertoire is required to recognize the enormous diversity of potential antigens in the environment. TCRs are heterodimers that consist predominantly (90–99%) of an α and a β subunit (reviewed in Lefranc and Lefranc, 2001), the remainder consisting of γ–δ heterodimers. Each chain (a TCR subunit is typically referred to as a chain) originates from the genetic rearrangement of a variable (V), joining (J) and constant (C) gene segment (Gascoigne *et al.*, 1984; Hedrick *et al.*, 1984). Rearranged TCRβ DNA also includes a short (12–16 nt) diversity (D) gene segment between the V and J gene (Fig. 1; Kavaler *et al.*, 1984). At the molecular level, two main mechanisms contribute to generate the immense TCR sequence repertoire. Akin to immunoglobulins, the combinatorial diversity of TCR arises from the genetic rearrangement of V, D and J gene segments (Sakano *et al.*, 1979) and yields $\sim 5.8 \times 10^6$ possible TCRαβ gene combinations (Janeway *et al.*, 2001). Further diversity is generated during this rearrangement by an additional mechanism of base addition and deletion at the junction of V, (D) and J segments, and is known as the N-diversity (Huck *et al.*, 1988). Addition of nucleotides by terminal deoxynucleotidyl transferases at the V–J (α) or V–D–J (β) junction (Landau *et al.*, 1984) occurs at random and is frequently preceded by base deletion at the 3′ end of V, the 5′ end of J and at both ends of D. This junctional diversity alone can generate $\sim 2 \times 10^{11}$ distinct molecules, bringing the number of theoretically possible TCRαβ to $\sim 10^{18}$ (Janeway *et al.*, 2001). The actual number of unique T-cell clonotypes in human blood is at least $\sim 10^7$ (10^6 β-chains; Arstila *et al.*, 1999). The amino acids encoded at the V–(D)–J junction, a region known as the third complementarity determining region (CDR3), are principally responsible for antigen recognition and define unique TCR clonotypes (Gorski *et al.*, 1994). Together, these somatic genome alterations create a diverse T-cell metagenome in every individual.

Profiling the cellular immune response to immune challenge by, for example, vaccination, transplantation, infection or cancer provides valuable insights into immune system integrity and function and the efficacy of prophylactic or therapeutic interventions. Unfortunately, the TCR diversity is such that complete characterization of repertoires still represents an enormous challenge. Current profiling methods, developed 15 years ago, analyze TCR β-chain repertoire complexity based on the CDR3 length diversity within Vβ gene families (Gorski *et al.*, 1994; Pannetier *et al.*, 1993; Penitente *et al.*, 2008). Although they provide a global picture of the repertoire, these low-resolution PCR-based spectratypes do not allow specific identification and quantification of individual T-cell clonotypes. DNA sequencing achieves higher resolution, but large-scale sequence profiling has been infeasible previously due to cost. For instance, sampling 1 million clonotypes (i.e. 10-fold coverage of 1 million 150 nt target CDR3 sequences in a single individual) with traditional Sanger sequencing would cost ~$1.5 M. Newer sequencing technologies

*To whom correspondence should be addressed.

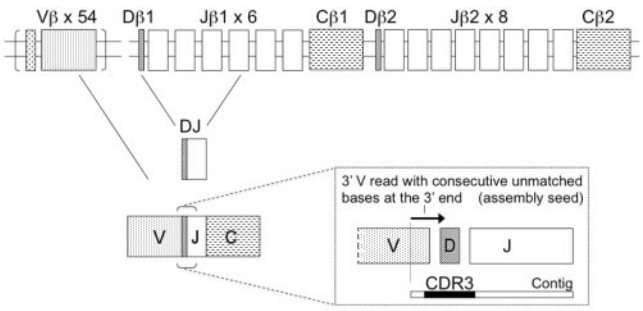

Fig. 1. Schematic diagram of the ~1 Mb human TCRβ locus on chromosome 7q34, showing the combinatorial gene rearrangement that takes place and the iSSAKE strategy for assembling CDR3 (inset). The TCRβ locus comprises a cluster of 54 predicted V genes located distantly from two separate clusters each with one D and C gene, interspersed with 6 or 8 J genes. At the DNA level, one of the D genes recombines with one of the J segments, creating partially rearranged DJ genes. Second, one of the V genes joins DJ and the intermediary DNA is deleted. During the gene rearrangements, the random base addition at the junction of V, D and J and the frequent base deletion at the 3′ end of V and 5′ end of the J gene yield the CDR3, a region with unique immune specificities. Read assembly is preceded by the segregation of assembly seeds (arrow); reads that align to the 3′ end of V with eight or more consecutive unmatched 3′ bases. A possible contiguous sequence (contig) resulting from that strategy is shown, with the CDR3-encoding region highlighted in black.

capable of producing a large amount of short reads at much lower cost have emerged in recent years (Bennett, 2004; Holt and Jones, 2008; Margulies *et al*., 2005) and make affordable TCR sequence profiling a likely prospect. Currently, sampling a million TCR clonotypes with the Illumina GAII Analyzer would cost ~1000-fold less compared to Sanger sequencing. On the flip side, next-generation sequencing technologies have much shorter read lengths and show appreciable base error (Holt and Jones, 2008). Combined, these limitations pose a computational challenge for the accurate and complete sequence reconstruction of specificity-determining regions.

Using randomly generated error-prone short reads from simulated TCRβ (sTCRβ) sequences, we have developed a strategy for profiling T-cell metagenomes. The method uses iSSAKE, a modified version of our previously published short read assembler SSAKE (Warren *et al*., 2007), and relies on annotated Vβ gene predictions to segregate partial 3′ alignments and sort corresponding seed sequences prior to assembly (Fig. 1). In this proof-of-principle study, we show that the method is over 63% sensitive for rare 1 p.p.m. clonotypes and over 91% sensitive for clonotypes as low as 2 p.p.m., when using 36 nt reads with 1% randomly distributed errors. When applying a modeled error distribution to simulated reads, we show that the sensitivity of the method is reduced to 51% for the rarest (1 p.p.m.) clonotypes, but is equally sensitive when clonotype frequencies are above 5 p.p.m. The assembly of longer read length impacts positively on the sensitivity of the method. For instance, the method is nearly 12% and 25% more sensitive in recovering 1 p.p.m. sTCRβ, when using 42 nt and 50 nt long reads compared to 36 nt. Together with high base accuracy of over 99%, we show that the majority of CDR3 sequences can be reconstructed accurately and thus characterized using error-rich short read data.

Table 1. Frequency (f) of base deletion and addition at the CDR3 between publicly available mRNA sequences and simulated TCRβ

f deleted 3′ V bases			f deleted 5′ J bases			f added CDR3 bases[a]		
Bases	Observed (N = 356)	Simulated (N = 220 000)	Bases	Observed (N = 1151)	Simulated (N = 220 000)	Bases	Observed (N = 174)	Simulated (N = 220 000)
0	0.194	0.200	0	0.209	0.212	1	0.006	0.007
1	0.160	0.158	1	0.123	0.123	3	0.006	0.006
2	0.098	0.098	2	0.122	0.122	4	0.029	0.031
3	0.118	0.113	3	0.104	0.105	5	0.017	0.019
4	0.160	0.155	4	0.117	0.117	6	0.017	0.019
5	0.118	0.119	5	0.123	0.119	7	0.052	0.056
6	0.070	0.073	6	0.086	0.085	8	0.063	0.067
7	0.045	0.047	7	0.056	0.057	9	0.069	0.073
8	0.022	0.023	8	0.031	0.031	10	0.080	0.084
9	0.008	0.009	9	0.019	0.019	11	0.057	0.059
10	0.006	0.006	10	0.010	0.010	12	0.075	0.076
						13	0.080	0.081
						14	0.086	0.085
						15	0.098	0.096
						16	0.046	0.046
						17	0.052	0.049
						18	0.029	0.028
						19	0.034	0.033
						20	0.023	0.021
						21	0.023	0.021
						22	0.011	0.010
						23	0.006	0.005
						25	0.011	0.010
						26	0.011	0.005
						27	0.006	0.005
						28	0.011	0.010

[a]We did not observe addition of 2 or 24 nt within the CDR3.

2 METHODS

2.1 Modeling TCRβ rearrangements

From the alignments of predicted V and J genes to Genbank TCRβ mRNA, four independent models of the N-diversity mechanisms were constructed, representing the frequencies of (i) random base addition at the V–D–J junction; (ii) base composition at each position where bases were added; (iii) 3′ V base deletion; and (iv) 5′ J base deletion. The models are intended as a guide for simulating distinct sequence clonotypes within the CDR3 region by estimating, using empirical data, expected frequencies of base addition, deletion and composition (Table 1). These models were used to construct sTCRβ sequences as described in Supplementary Material. Briefly, this involved, (i) randomly selecting the V and J gene sequences; (ii) deleting 3′ V bases; (iii) deleting 5′ J bases; and (iv) joining V–D–J with addition of junction bases.

For the simulations, we used only the ~150 nt of simulated sequence matching ~40 nt upstream of the 3′ end of V, spanning CDR3 and J and ending ~50 nt downstream of the 5′ end of C. The process of building sTCRβ was repeated to generate a library of 1 000 000 total sequences containing 220 000 unique sTCRβ sequences at frequencies ranging from 1 to 20 p.p.m. (Table 2).

2.2 TCRβ–CDR3 reconstruction strategy

From the above 1M sTCRβ sequences, we randomly generated, in three independent replicate experiments, 20 million 36 nt reads having 1.0, 1.5 or 2.0% randomly distributed errors and aligned them to known TCRβ gene

Table 2. Generating sTCRβ clonotypes

Clonotype frequency	p.p.m.	Number of unique sTCRβ	Number of total sTCRβ	Fold coverage ca.[a]
1:1 000 000	1	110 000	110 000	5
1:500 000	2	10 000	20 000	10
1:333 333	3	10 000	30 000	15
1:250 000	4	10 000	40 000	20
1:200 000	5	10 000	50 000	25
1:166 667	6	10 000	60 000	30
1:142 857	7	10 000	70 000	35
1:125 000	8	10 000	80 000	40
1:111 111	9	10 000	90 000	45
1:100 000	10	10 000	100 000	50
1:66 667	15	10 000	150 000	75
1:50 000	20	10 000	200 000	100

[a]Approximate coverage calculated using 150 nt sTCRβ template size, 20M 36 nt reads.

segments. In addition, we also generated 20M 42 or 50 nt 1% error reads to assess the effect of read length on assembly and tested the effect of random (Dohm *et al*., 2007) versus modeled (Using MAQ simutrain and simulate on real phiX174 Illumina sequences; Heng *et al*., 2008) 1% read error distributions. Simulated reads were aligned against Ensembl (Flicek *et al*., 2008) TCRβ gene predictions using exonerate (Slater and Birney, 2005; Software parameters used: –bestn 1 –score 1 –percent 0). Reads aligning best to V genes, at the 3′ end and having eight or more consecutive unmatched bases in 3′ were put aside as seeds for a *de novo* iSSAKE assembly. The reverse, complemented sequences of reads aligning on the reverse strand of V predictions were also considered as seeds. The assembly read pool consisted of unaligned reads, those aligning to J segments best and assembly seeds. Reads aligning best to V, C or any possible JC junction sequence combinations were discarded (Fig. 1).

2.3 Targeted seeded assemblies with iSSAKE

To create the iSSAKE assembler, modifications were made to the SSAKE v3.2.1 code base (Warren *et al*., 2007; http://www.bcgsc.ca/platform/bioinfo/software/ssake). Notably, the depth of the prefix tree was increased, augmenting the number of nodes to 15. This modification was essential to help speed the assembly process at the cost of increased memory requirements. It does so by reducing the search space when considering reads for extension. This modification was compulsory since there is great sequence conservation between the various sTCRβ CDR3 sequences, sometimes differing by only one or a few bases.

Since SSAKE release v2.0 (October 2007), we have implemented the approach for handling error-rich sequencing data described in Jeck *et al*. (2007). In essence, all overhanging bases of reads aligning perfectly to a seed sequence are considered for extension, using a majority rule approach for building consensus sequences of the overhanging bases.

To support the assembly of longer contigs with complete CDR3 without depleting the read pool, sequences used for extension are re-used. The assembly terminates only when all seeds have been maximally extended. This is easily parallelizable on a cluster of computers and permits the assembly of discrete nontruncated TCRβ CDR3 sequences ending in the same J segment. Finally, only the 3′ extension of seeds was permitted, the assembly progressing through V, D and J in this order. For each read set, we ran 100 parallel iSSAKE jobs on two dozen 2.66 GHz Quad-Core 64 bit Intel® Xeon® processors with 15 GB RAM (iSSAKE -m 15 -o 1 -r 0.6).

3 RESULTS

A library of 1 million sTCRβ was generated *in silico* using models of TCRβ diversity derived from publicly available mRNA sequences.

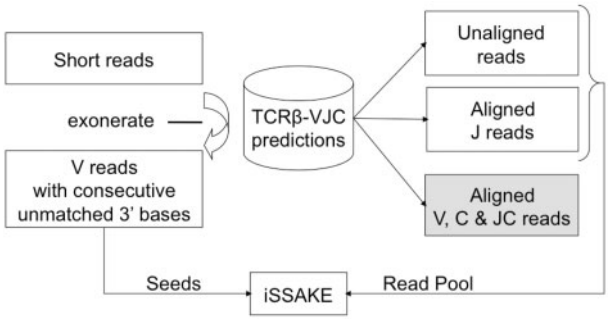

Fig. 2. TCRβ–CDR3 reconstruction strategy. Reads are aligned against Ensembl TCRβ gene predictions using exonerate or other short read aligners. Reads aligning best to V genes at the 3′ end and having user defined *n* consecutive unmatched bases in 3′ are set aside as seeds for a *de novo* iSSAKE assembly. The reverse complements of reads aligning on the reverse strand are also considered as seeds. The read pool consists of unaligned reads, those aligning J genes best and assembly seeds. Reads aligning best to V, C or any possible JC junction sequence combinations are discarded to reduce sequence space.

The library consisted of 220 000 distinct sequences present at frequencies ranging from 1 to 20 p.p.m. (110 000 unique 1 p.p.m. sequences and 10 000 unique 2 to 20 p.p.m., Table 2). To confirm that sTCRβ reflect real sequences, we kept track of the simulated N-diversity changes applied to each sequence and verified that the frequencies of V and J base deletion as well as CDR3 base addition in the sTCRβ library were consistent with the frequencies derived from experimental mRNA sequences (Table 1). Twenty million reads having 1.0, 1.5 or 2.0% random or 1.0% modeled error and 36, 42 or 50 nt in length were randomly sampled from the sTCRβ library and assembled in separate experiments. The approximate read coverage for each 1–20 p.p.m. clonotype ranged from 5- to 96-fold, respectively (Table 2).

Since we expect the redundancy of short reads derived from V, J and C to be extremely high and the redundancy over CDR3 to be very low, a classic *de novo* assembly where all sequence reads are used in turn to seed a contig assembly is not suitable. However, the sequences of human TCRβ genes (V, J and C) are known and well annotated, which makes feasible a streamlined strategy of seeded assembly. The assembly seeds we use are sequences that align to the 3′ end of V with eight or more consecutive unmatched 3′ bases in the highly diverse CDR3 region (Fig. 1 inset and Fig. 2). Using exonerate, averages of 1.755, 1.718, 1.679 and 1.599 million seeds were identified from sets of reads with random 1.0, 1.5, 2.0 and 1.0% modeled error, respectively (Table 3). The decrease in number of seeds identified at higher error rates or between random and modeled error distribution (4.3% and 8.9%, respectively) is due to an increased number of mismatched bases that prevent read alignment to the 3′ end of the V gene. Selecting seeds before assembly reduces the sequence space by ∼90% and segregates about half of the 20M input read set for contig assembly. This approach considerably increases the assembly speed and yields only contigs that represent the CDR3.

At 1% randomly distributed error, 84.2 ± 0.16% of the seeds, on average, yielded contigs that comprised complete CDR3 sequences, including D segment bases and unambiguous J segment junctions. These unambiguous contigs which are defined as having clearly

Table 3. sTCRβ contig stats from assemblies of 20M randomly generated 36, 42 and 50 nt reads

Bases	Error (%)[a]	Mean seeds	Number of iSSAKE contigs from triplicates (36 nt) or duplicates (42 and 50 nt) experiments				
			Short ambiguous[b]	Long ambiguous[c]	Unambiguous[d]	Unambiguous, but sub-optimal[e]	
						A	B
36	1.0[f]	1 599 140 ±961	498 583 ±822	36 352 ±261	1 064 399 ±400	1098 ±185	2847 ±89
36	1.0	1 755 437 ±404	265 618 ±245	11 520 ±500	1 478 299 ±236	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	1158 ±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 five or more mismatched bases.
[a]Random error distribution generated using simulators from Dohm *et al*. (2007), unless otherwise specified.
[b]Too short to unambiguously decipher J and thus, CDR3.
[c]Contigs ≥ 45 nt, sufficiently long to contain the first 15 bases of J, but base errors/polymorphisms prevent proper identification of the J segment.
[d]Captured CDR3 and first 15 bases of J unambiguously.
[e]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 five or more mismatches bases with the closest sTCRβ are built with erroneous reads that often yield misassembled contigs.
[f]Error distribution modeled using phiX174 Illumina sequence data as the training set (Heng *et al*., 2008).

demarcated V and J boundaries were subsequently trimmed to keep the last 15 V bases, the CDR3 and the first 15 bases of identifiable J sequence. The reason for trimming was to facilitate assessment by removing bases that were uninformative for characterization of CDR3. As expected, seeds from 36 nt read sets having higher error rates or errors modeled on Illumina data yield fewer unambiguous contigs, with average proportions of 78.1 ± 0.25% and 72.4 ± 0.10% for the 1.5% and 2% random error sets and 66.6 ± 0.03% for the modeled error read set, respectively (Table 3). This is because random base errors in the seed sequences cause premature termination of contig extension by iSSAKE, unless one or more reads in the pool has matching erroneous bases by chance. The latter case can lead to (i) misassembled contigs, especially if different sTCRβ have a very similar CDR3 makeup and (ii) long ambiguous contigs where J segments are undecipherable.

Misassemblies are identified as contigs that do not match sTCRβ in the source library. These were rarely observed (0.01–0.02% of unambiguous contigs), although more prevalent in the 2% error read set (Table 3). The effect of error on contig misassemblies is more pronounced when error is modeled from real Illumina data (Table 3), where error rates tend to increase toward the end of the read. However, even with a modeled error distribution, misassemblies still represent a minor proportion of all unambiguous reconstructions (0.3%).

We define long ambiguous contigs as those large enough to contain J segment bases, but because of base errors, there was not a precisely matching J segment. An increase from 1% to 2% in the error rate nearly doubles the number of long ambiguous contigs, increasing their proportion from 0.7% to 1.3%. These contigs, while ambiguous, were still useful for assessing the sensitivity of our method. Their abundance is not negligible and the contigs still produce valid alignments to a reference. With real data, identifying the CDR3 from these contigs without a reference sequence will prove more challenging. Short ambiguous contigs are defined here as those that are not long enough to span CDR3. Short ambiguous contigs are caused by early termination of contig extension due to base errors. These short ambiguous contigs represent a considerable portion of the total contigs (15% ± 0.01, 21% ± 0.04, 26% ± 0.05 and 31% ± 0.05 of assemblies using 1.0, 1.5, 2.0% random and 1% modeled error read sets, respectively).

For each assembly, the average base accuracy was calculated by counting the total number of matching bases over the aligned contig length. Although base accuracy of assembled contigs is lower when simulated sequence error rates are higher, it is above 99% at all clonotype frequencies and error rates simulated (Tables 4–6). Contigs representing clonotypes with the lowest frequencies were the least accurate. This is not unexpected since at lower read depths, there are fewer reads to offset the base error, especially in the highly diverse and thus relatively thinly covered CDR3 region. Effectively, inspection of the base error and coverage of assemblies as a function of base position over the region of interest reveals that base mismatch frequency peaks within the seed portion (any of the last 15 V bases and at least 8 consecutive mismatched bases downstream) and decreases through J as the base coverage increases (Fig. 3).

At clonotype frequencies as low as 3 p.p.m., over 93% of the sTCRβ CDR3 sequences could be characterized by iSSAKE contigs assembled from the 1% modeled error distribution read set (Table 4). This means that the sTCRβ sequence diversity can be almost entirely characterized at 15× coverage (Table 2). Although the scope of real T-cell diversity remains unknown, if it is close to the estimated lower limit of 10^6 β-chains, then substantial repertoire coverage should be easy to attain by massively parallel short read sequencing, even without trimming the reads.

For all contigs that capture sTCRβ sequences we find the accuracy to be very high, especially for clonotypes present at 5 p.p.m. or more. Interestingly, we find that read error has only a small impact on accuracy at these clonotype frequencies. Seeds with errors will rarely find a sequence match in iSSAKE, causing premature contig extension or leading to an increased number of singlets, depleting the pool of unambiguous contigs. Thus, early rejection of these reads has a much more significant impact on the sensitivity than it does on the accuracy.

Table 4. Method sensitivity and accuracy as a function of a 1% read error distribution (random or modeled)

p.p.m.	Number of sTCRβ–CDR3 characterized by iSSAKE contigs (sensitivity)		Accuracy (%)		Average number of contigs characterizing each sTCRβ	
	Error (%)		Error (%)		Error (%)	
	1.0 Random	1.0 Modeled	1.0 Random	1.0 Modeled	1.0 Random	1.0 Modeled
1	70 240	56 564	99.68	99.01	2.0	1.8
2	9111	8100	99.90	99.34	3.7	2.5
3	9747	9295	99.96	99.64	4.6	3.4
4	9883	9721	99.98	99.80	6.0	4.4
5	9932	9874	99.99	99.90	7.6	5.5
6	9936	9913	99.99	99.94	9.1	6.6
7	9935	9936	99.99	99.96	10.6	7.7
8	9939	9948	99.99	99.97	12.2	8.9
9	9948	9955	99.98	99.97	13.7	10.0
10	9956	9958	99.99	99.98	15.2	11.2
15	9972	9975	99.99	99.98	23.0	16.8
20	9955	9958	99.98	99.98	30.7	22.1

Unambiguous and long ambiguous contigs were used for this analysis. Reported values are the mean of triplicate simulations. Variation among simulations was minimal (Supplementary Table 1).

Table 5. Method sensitivity and accuracy as a function of randomly distributed error rates

p.p.m.	Number of sTCRβ–CDR3 characterized by iSSAKE contigs (sensitivity)		Accuracy (%)		Average number of contigs characterizing each sTCRβ	
	Error (%)		Error (%)		Error (%)	
	1.5	2.0	1.5	2.0	1.5	2.0
1	64 862	59 259	99.48	99.23	1.9	1.8
2	8779	8423	99.80	99.68	2.9	2.6
3	9656	9520	99.92	99.85	4.1	3.7
4	9869	9831	99.96	99.93	5.4	4.9
5	9929	9920	99.98	99.97	6.9	6.2
6	9936	9928	99.98	99.98	8.2	7.5
7	9934	9924	99.98	99.97	9.7	8.7
8	9940	9929	99.98	99.98	11.0	10.1
9	9952	9944	99.98	99.98	12.5	11.3
10	9953	9940	99.99	99.98	13.9	12.6
15	9971	9966	99.99	99.98	21.1	19.3
20	9961	9944	99.98	99.98	28.2	26.0

Unambiguous and long ambiguous contigs were used for this analysis. Reported values are the mean of triplicate simulations. Variation among simulations was minimal (Supplementary Table 1).

Table 6. Method sensitivity and accuracy as a function of read length

p.p.m.	Number of sTCRβ–CDR3 characterized by iSSAKE contigs (sensitivity)		Accuracy (%)		Average number of contigs characterizing each sTCRβ	
	Read length (nt)		Read length (nt)		Read length (nt)	
	42	50	42	50	42	50
1	90 210	104 155	99.74	99.75	2.4	2.9
2	9737	9914	99.93	99.95	4.3	5.4
3	9911	9959	99.98	99.98	6.4	8.0
4	9937	9963	99.99	99.98	8.5	10.7
5	9948	9966	99.99	99.98	10.7	13.2
6	9944	9966	99.98	99.98	12.8	15.8
7	9941	9974	99.98	99.97	14.8	18.3
8	9948	9975	99.98	99.97	17.0	20.8
9	9954	9979	99.98	99.98	19.1	23.2
10	9960	9983	99.98	99.98	21.1	25.7
15	9973	9985	99.99	99.98	31.1	37.2
20	9958	9976	99.98	99.98	40.0	47.4

Unambiguous and long ambiguous contigs were used for this analysis. Reported values are the mean of duplicate (42 and 50 nt reads) simulations at 1.0% randomly distributed errors. Variation among simulations was minimal (Supplementary Table 1).

Again, it is important that in real sequence data, errors tend to accumulate toward the 3′ ends of reads, rather than being equally distributed along the length of the read. Using error distribution modeled on real data, we see fewer contigs reconstructed at each p.p.m., due to fewer seeds being initially identified and more frequent seed extension failures (Table 4). However, at least for clonotype frequencies >5 p.p.m., reconstruction success rate is high and largely unaffected by error distribution. At lower clonotype frequencies, the reconstruction rate is lower; for instance, 63.8% versus 51.5% of 1 p.p.m. sTCRβ can be identified when using a random versus a modeled error distribution, respectively (Table 4).

A change of only 1% in the read base error (from 1% to 2%) yields a 66% increase in contigs too short to characterize sTCRβ unambiguously, usually because the J segment is incomplete and/or its position cannot be identified with certainty. This translates into a decreased sensitivity of the method of over 10% at 1 p.p.m.,

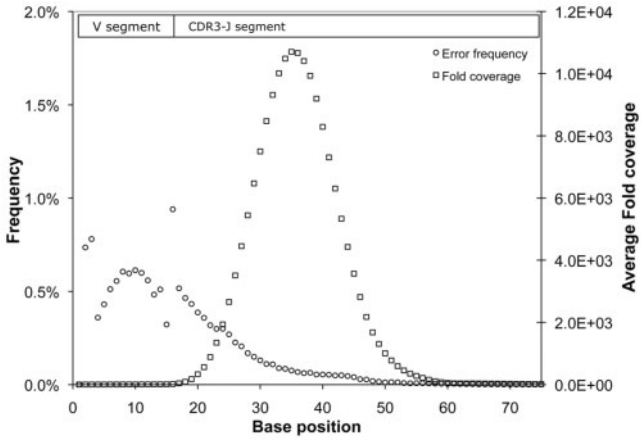

Fig. 3. Average mismatched base frequency and mean contig base coverage per position on trimmed, normalized, unambiguous contigs from triplicate 36 nt 1% random error read assemblies. Assembly base mismatch frequency (i.e. assembly errors) reaches a maximum within the seed portion (any of the last 15 V bases and at least eight consecutive mismatched bases downstream) and decreases through J as the base coverage increases. The sudden increase in average fold coverage beginning at approximately base position 30 is explained by over sampling of a limited number (14) of J segments, and the re-use of reads by the iSSAKE assembly algorithm. The position of the V segment and approximate position of the CDR3 and J gene segments on the contigs is shown on top of the graph and is depicted by the rectangles. Every contig is comprised of the last and first 15 nt of the V and J gene segment, respectively. The CDR3 and J gene segment boundaries are approximate because the length of the CDR3 varies.

7% at 2 p.p.m. and 2% at 3 p.p.m. At higher clonotype frequencies (>4 p.p.m.), the effect of base error on yielding short contigs is offset by the larger read depth (Table 5).

The robustness of assemblies of higher frequency clonotypes is further enhanced because more seeds are available at the start of assembly. When assembled, these seeds should, in theory, lead to contigs that characterize the same TCRβ. Keeping track of the average number of contigs that characterize each sTCRβ generated allows one to estimate the frequency of any given sTCRβ in the sample. Consistently, at the error rates tested, there is an almost perfect PEARSON correlation (0.9998, 0.9997 and 0.9994 at 1, 1.5 and 2% random error, respectively, and 0.9995 for the 1% modeled error set) between the average number of sTCRβ-capturing contigs and the frequency of that sTCRβ. Since the number of seeds (and thus, contigs) identified per TCRβ varies linearly in function of read coverage, as opposed to having a 1:1 relationship with the clonotype frequency, the number of contigs identified cannot be expected to reflect the exact clonality of each TCRβ in the sample. Instead, the contig count may be used to estimate relative TCRβ abundance.

To explore the effect of read length, we also simulated sets of 42 and 50 nt × 20M reads at 1% random error. These read lengths and error rates should be achievable by massively parallel short read platforms, if not currently, then in the near future. Increasing the read length has a drastic effect on the sensitivity of the method. Detection of sTCRβ increases by 18% at 1 p.p.m. when 42 nt 1% error reads are assembled compared to shorter 36 nt reads (Table 6). Using 50 nt reads for assembly recovers over 94.7% of clonotypes, an increase of >30% in detection compared to the usage of 36 nt reads with the same error rate. At 2 p.p.m., the sensitivity of the recovery increases from 91.1% to 97.4% to 99.1%, using 1% error 36, 42 and 50 nt reads, respectively. Increased sensitivity is a direct consequence of obtaining more seeds. With 42 and 50 nt reads, 40% and 77% more seed sequences could be identified from our set of 20M simulated reads (Table 3).

4 DISCUSSION AND CONCLUSION

Technological advances in sequencing (Holt and Jones, 2008) put large-scale high-resolution TCR profiling within the realm of possibility. However, shortcomings of these new sequencing technologies, namely the appreciable sequencing errors and short read lengths, require computational solutions to help make sense of the data. We have explored the feasibility of using short 36, 42 and 50 nt error-prone sequences to characterize up to 1 million sTCRβ sequences. Our strategy for reconstructing sTCRβ relies on two bioinformatics pillars: short read sequence alignment and seeded *de novo* assembly. Unidirectional *de novo* assemblies of short seeds targeting the V–D–J junction is made possible using a modified version of SSAKE (Warren *et al*., 2007) that handles sequencing errors, re-use reads and processes *k*-mers more rapidly than earlier versions (http://www.bcgsc.ca/platform/bioinfo/software/ssake). This strategy is tailored for very short reads, such as those produced by the Illumina Ltd. sequencing instrument, and constitute the main theoretical advance presented in this article.

Sequence characterization of TCRs and more specifically the variable portion encoding amino acids that directly interact with antigenic peptide permits the identification of disease-associated T-cells. Current TCR sequence profiling can at best decipher hundreds of TCRs (Ozawa *et al*., 2008; Zhou *et al*., 2006), a small number in comparison with the 10^7 TCR diversity estimated in an individual (Arstila *et al*., 1999). Larger-scale profiling techniques that examine CDR3 length heterogeneity provide a global snapshot of TCR repertoires, but do not resolve individual clonotypes at the macromolecular level (Gorski *et al*., 1994; Pannetier *et al*., 1993; Penitente *et al*., 2008). Due to the low throughput, high cost and labor requirements of traditional Sanger sequencing, sequence-profiling TCR on that same scale has not yet been explored, thereby providing the impetus for our study.

The success of TCR sequence reconstruction using short sequences relies on the very region that makes profiling the TCR repertoire challenging; the uniqueness and specificity of the CDR3 (Davis *et al*., 1998). Selection of seed sequences that comprise bases encoding a portion of the variable region ensures that a streamlined, unidirectional assembly proceeding through the junction will help characterize unique clones. This is especially true if the sequence coverage is 10-fold or above, or the frequency of the TCR is high, since higher frequencies result in higher sequence coverage of discrete TCRs. At low frequencies, base error has a strong negative impact on TCR reconstruction rates that is due to sequence coverage insufficient to offset base error in less redundant CDR3-encoding regions. iSSAKE will not extend a seed or contig with a base error in a minimum set overlap region, unless that base can be found at the same position in an overlapping *k*-mer. This impacts favorably on contig accuracy at the expense of reconstruction rates, especially at low 1 p.p.m. frequencies and 2% error.

We examined instances of failure to detect CDR3 sequences known to exist in our sTCRβ library. The majority (99.4% using

36 nt read sets) of irresolvable low-frequency sTCRβ are attributable to 3′ base errors. The problem is exacerbated for low-frequency clones because these by definition have lower coverage and therefore less chance for an error to be mitigated by read redundancy. At all clonotype frequencies, but more noticeably at higher sequence coverage, 0.3% of irresolvable sTCRβ are due to high sequence identity between modeled TCRβ, sometimes differing only by a few 5′ V bases and thereby preventing unambiguous characterization of their sequence. We see additional failure modes that are very rare (∼0.2–0.3% of irresolvable sTCRβ, 0.004% of total sTCRβ) and associated with shorter (36 and 42 nt) reads and higher frequency clonotypes (>5 p.p.m.). For example, reads originating from CDR3 may by chance align well to V segments, C segments or JC junctions, and will as a consequence be removed early in the assembly process (Fig. 2) such that CDR3's containing these sequences cannot be assembled despite high read coverage. This failure mode was not observed with longer 50 nt seeds and is explained by the increased ability of seeds (which are never discarded) to span the entire CDR3 and capture a portion of J in a single read.

In time, accurate sequence length from next generation sequencing platforms will exceed the length of CDR3. However, sequence assembly will remain advantageous because it mitigates the effect of sequence errors present in individual reads. In the present study, fewer CDR3 sequences could be identified unambiguously with the unassembled 50 nt read set compared to the assembled one (533 868 versus 2 428 086 unambiguous contigs, Supplementary Table 1). Already, a typical Illumina Sequence Analyzer run will output more bases than we generated in this study, at equal or lower error rates. This suggests that the overall strategy, as outlined, will work with sequence data from biological samples. We expect it will also be applicable to similar metagenomics projects including sequence-characterization of the immunoglobulin repertoire.

ACKNOWLEDGEMENTS

We thank Dr John Webb for advice. R.A.H. is a Michael Smith Foundation for Health Research Scholar.

Funding: Genome Canada and Genome British Columbia.

Conflict of Interest: none declared.

REFERENCES

Arstila,P.T. *et al*. (1999) A direct estimate of the human αβ T cell receptor diversity. *Science*, **286**, 958–961.

Bennett,S. (2004) Solexa Ltd. *Pharmacogenomics*, **5**, 433–438.

Davis,M.M. *et al*. (1998) Ligand recognition by alpha beta T cell receptors. *Annu. Rev. Immunol.*, **16**, 523–544.

Dohm,J.C. *et al*. (2007) SHARCGS, a fast and highly accurate short-read assembly algorithm for de novo genomic sequencing. *Genome Res.,* **17**, 1697–16706.

Flicek,P. *et al*. (2008) Ensembl 2008. *Nucleic Acids Res*., **36**, D707–D714.

Gascoigne,N.R. *et al*. (1984) Genomic organization and sequence of T-cell receptor beta-chain constant- and joining-region genes. *Nature,* **310**, 387–391.

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.

Hedrick,S. *et al*. (1984) Isolation of cDNA clones encoding T cell-specific membrane-associated proteins. *Nature*, **308**, 149–153.

Heng,L. *et al*. (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.*, **18**, 1851–1858.

Holt,R.A. and Jones,S.J. (2008) The new paradigm of flow cell sequencing. *Genome Res.*, **18**, 839–846.

Huck,S. *et al*. (1988) Variable region genes in the human T-cell rearranging gamma (TRG) locus: V-J junction and homology with the mouse genes. *EMBO J*., **7**, 719–726.

Janeway,C.A. Jr *et al*. (2001) *Immunobiology*. 6th edn. Garland Science, London, UK.

Jeck,W.R. *et al*. (2007) Extending assembly of short DNA sequences to handle error. *Bioinformatics,* **23**, 2942–2944.

Kavaler,J. *et al*. (1984) Localization of a T-cell receptor diversity-region element. *Nature,* **310**, 421–423.

Landau,N.R. *et al*. (1984) Cloning of terminal transferase cDNA by antibody screening. *Proc. Natl Acad. Sci. USA,* **81**, 5836–5840.

Lefranc,M.-P. and Lefranc,G. (2001) *The T cell Receptor Facts-Book*. Academic Press, London, UK.

Margulies,M. *et al*. (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**, 376–380.

Ozawa,T. *et al*. (2008) Comprehensive analysis of the functional TCR repertoire at the single-cell level. *Biochem. Biophys. Res. Commun.*, **367**, 820–825.

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.

Penitente,R. *et al*. (2008) Administration of PLP139-151 primes T cells distinct from those spontaneously responsive in vitro to this antigen. *J. Immunol*., **180**, 6611–6622.

Sakano,H. *et al*. (1979) Sequences at the somatic recombination sites of immunoglobulin light-chain genes. *Nature*, **280**, 288–294.

Slater,G.S. and Birney,E. (2005) Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*, **6**, 31.

Warren,R.L. *et al*. (2007) Assembling millions of short DNA sequences using SSAKE. *Bioinformatics*, **23**, 500–501.

Zhou,D. *et al*. (2006) High throughput analysis of TCR-b rearrangement and gene expression in single T cells. *Lab. Invest.,* **86**, 314–321.