Supplementary Note 1. Extracting more but extracting truth

RNA-seq data contain only a tiny fraction of TCR/Ig reads covering CDR3. This fraction varies from sample to sample depending on the degree of immune cell infiltration, ranging from 10^{-5} to 10^{-7} for TCRs and from 10^{-4} to 10^{-7} for Igs. Samples having no target reads are also common. Additionally, due to the limited length of sequencing (typically paired-end 50-100 bp), successful detection of target V(D)J junctions implies alignment with very short fragments of germline V and J genes (12-15 bp). The main challenge in the analysis of such short sequences is the high probability of false-positive alignments. Thus, the primary objective of a robust CDR3 extraction procedure is to extract as many true CDR3 sequences as possible with nearly zero amount of CDR3-like false-positives.

To meet these challenges, we have developed and implemented a set of new algorithms (**Supplementary Fig. 1**):

(i) **Sensitive and highly selective aligner**. Since there is no prior knowledge of true TCR/Ig sequences in real data and because some sources of false-positive sequences are not random (**Online Methods**), the implementation of a fully automated procedure for optimization of the alignment algorithm is substantially complicated. Manual analysis of the structure of false- and true-positive alignments on real and artificial datasets has helped us to establish a semi-automated pipeline for the optimization of our built-in MiXCR aligner. This has resulted in high-efficiency extraction of target V(D)J rearrangements from bulk RNA-seq data with a zero detected false-positive rate. Optimized aligner efficiently filters out even the fragments of mRNA sequences that are homologous to antibody or TCR hypervariable regions, which comprise the most prominent source of reproducible false-positives in RNA-seq data analysis (**Online Methods**). We further enhanced our aligner by targeting ambiguous cases, which recurrently arise in short reads that have either V or J segment alignment but for which the fast alignment algorithm has failed to detect a J or V gene segment, respectively. In such cases MiXCR switches to a more sensitive algorithm—a modified version of the Smith-Waterman/Needleman-Wunsch algorithm— in order to reconstruct the full V(D)J junction if possible. In this way, we achieve high alignment sensitivity even on short reads, maintaining high overall performance (**Supplementary Table 1**).

(ii) **Partial alignments assembler**. For short-read data, there is little to no chance that a long CDR3 will be fully covered by a single read. To extract such sequences, we introduced an additional analysis step that assembles contigs from several initial alignments, originating from different paired- or single-end reads. The partial assembler merges left-half reads (LR), defined as reads which cover only the left boundary of CDR3 (the conserved Cys in the V gene) while not covering the right boundary (the conserved Phe/Trp in the J gene), with right-half reads (RR), which do not cover the left CDR3 boundary and cover part of the J gene. To protect the algorithm from artificial diversity generation, LR and RR are merged if and only if the following criteria are satisfied:

1. Minimal length of the overlap region is 12 nt.
2. The overlap covers at least 7 non-germline-derived letters (N region). Boundaries of N regions are determined after the realignment of fused contigs against V, D and J gene sequences, and P-segments are also aligned.
3. Sequences are 100% identical inside the overlap.

These default thresholds showed the best extraction efficiency while keeping negligible rate of observed false-positive overlaps (as verified with *in silico* generated data, **Online Methods**, doi.org/10.5281/zenodo.804326). All parameters can be adjusted by the user.

(iii) **CDR3 extension**. To safely utilize even those reads that partially cover CDR3 and were not fully reconstructed using previous step, we added an optional CDR3 extension step for TCRs (but not Igs, due to possible presence of hypermutations). This step fills in the edges of the CDR3 based on known information on the relevant germline gene segments. A substantial fraction of extracted TCR reads fully cover the N-D-N region and are assigned with definite V and J genes (often supported by companion paired-end reads), but at the same time do not cover the CDR3 sequence end-to-end, missing several germline nucleotides. Since TCRs do not undergo hypermutations and their germline sequences are quite well-conserved, it is reasonable to artificially extend the CDR3 for such junctions with existing data from reference germline genes. This allows us to make use of these sequences for clonotype assembly and further comparative analysis. Using *in silico* generated data, we estimated a false extension rate of ~10^-5 for this procedure (**Online Methods**).

Importantly, the resulting RNA-seq analysis pipeline employs the same MiXCR modules, the same error-correction algorithms, and has the same output format as for targeted TCR or Ig profiling. This allows unified processing and comparison of immune repertoires obtained from different types of raw sequencing data.

To verify the efficiency and specificity of TCR CDR3 repertoires extraction from RNA-seq data, we performed both deep targeted profiling of TCR alpha (TRA) and beta (TRB) chains repertoires (TCR-seq) as described previously[1] and 100+100 paired-end RNA-seq analysis for the same split RNA samples, which were obtained from surgically resected melanoma specimens from two patients, SPX6730 (ileocecal lymph node metastasis) and SPX8151 (small intestine resection).

We further used the deep TRA and TRB CDR3 repertoires extracted from TCR-seq data using the standard MiXCR analysis pipeline[2] (**Online Methods**) as control data. Analysis of RNA-seq data was performed with MiXCR or TRUST, a recently published software tool for extracting TCR repertoires from RNA-seq[3, 4].

For MiXCR, we compared the RNA-seq CDR3 sequences with the TCR-seq control to directly assess the equivalence of the identified CDR3 nucleotide sequences. Since TRUST does not group clonal CDR3 sequences, we used unique CDR3s from the TRUST output for comparison. Furthermore, the majority of CDR3s reported by TRUST are truncated in a nondeterministic way, such that strict equality gives almost no

matches between the TRUST results and the control data. Since it was still possible that TRUST-reported CDR3s may represent immunologically useful information, we allowed subsequence matching between the control data and TRUST-reported CDR3s (**Online Methods**).

We assessed the dependence of the number of confirmed RNA-seq clonotypes on their abundance estimated from TCR-seq data (**Fig. 1a** and **Supplementary Fig. 2**). MiXCR was able to extract all relatively abundant TRB CDR3 clonotypes (frequency in repertoire > 0.15%) from the SPX6730 sample RNA-seq, even with the short paired-end reads (50+50-bp, trimmed *in silico* from the 100+100-bp paired-end data). In contrast, TRUST failed to extract a considerable proportion of high-frequency clonotypes. As expected, performance of RNA-seq analysis degraded at shorter read lengths. Our data also indicate that paired-end RNA-seq >100-bp would be beneficial for immune repertoires profiling.

Next, we divided extracted clonotypes into several categories: clonotypes that are also present in the TCR-seq control (verified CDR3s, considered true positives), clonotypes that are absent in the TCR-seq control but have canonical amino acid sequences (potentially true low-frequency clonotypes, **Online Methods**), and clonotypes that are absent in the TCR-seq control and have non-canonical amino acid sequences (probable false positives). **Fig. 1b** shows the dependence of the number of found CDR3 clonotypes on the read length in paired-end analysis for these various categories. Most MiXCR-reported clonotypes were confirmed by the control data; ~20% of all clonotypes were unique to RNA-seq samples and had canonical amino acid sequences, and there was only a tiny fraction of unconfirmed clonotypes with non-canonical CDR3 sequences, which was less than the fraction of non-canonical CDR3s observed in control TCR-seq data. It should be noted that we did not apply a filter for CDR3 canonical sequences in the MiXCR pipeline. On the other hand, nearly none of CDR3s reported by TRUST could be confirmed by the TCR-seq control; ~20% of CDR3s were partially confirmed on the basis of subsequence matches, with truncations of up to 6 nucleotides allowed, and >50% of unconfirmed CDR3s did not match canonical pattern. Only ~16% of the CDR3s reported by TRUST had both V and J genes annotated, while in the case of MiXCR, all clones had both V and J segments assigned.

Software testing with *in silico*-generated data confirmed the high extraction efficiency of MiXCR, with zero false-positive clones observed. In contrast, TRUST efficiency was an order of magnitude lower, and the software generated a substantial number of false clonotypes, including those of non-TCR origin (**Online Methods**, **Supplementary Fig. 3**).

The frequencies of clonotypes in the TCR repertoires extracted by MiXCR from the SPX6730 sample correlated between the TCR-seq and RNA-seq data (**Fig. 1c**). This demonstrates that RNA-seq-based TCR profiling can be relatively quantitative for the abundant clonotypes that occupy >0.1% of the overall T-cell repertoire for those samples that harbor a substantial number of T-cells. It should be noted that the SPX6730 sample was an ileocecal lymph node metastasis that was enriched with T-cells. The second sample, SPX8151, was a small intestine resection that contained lower proportion of T-cells and correspondingly yielded a lower number of TCR CDR3 reads (**Fig. 1d**), resulting in poor quantification of observed clonotypes (**Supplementary**

Fig. 2). Roughly, the number of TRB CDR3-containing sequencing reads extractable from an RNA-seq dataset was proportional to TRBC coverage, and was estimated as approximately 46 TRB CDR3 reads per 1,000 TRBC reads for 50+50-bp, and 128 TRB CDR3 reads per 1,000 TRBC reads for 100+100-bp paired-end sequencing (**Online Methods**, **Supplementary Table 2**).

We also compared MiXCR performance with the recently reported V'DJer software[5], which was designed for the extraction of Ig repertoires from RNA-seq data. MiXCR successfully extracted repertoires for all immune receptor types from both melanoma samples (**Fig. 1d**), while for these large RNA-Seq datasets V'DJer failed to extract IGH and IGK repertoires within four days using 8 threads on a Xeon E5-2683 CPU with 50 GB of RAM.

Additionally, we used several representative samples analyzed in refs. 3, 4 from TCGA and SRA databases in order to compare MiXCR performance for TCR and Ig repertoires relative to the TRUST and V'DJer packages, respectively. In all comparisons, MiXCR demonstrated superior sensitivity (**Fig. 1e**). Both alternative software packages require substantially more hands-on time and implementation of third-party alignment tools with particular versions of epy human genome and particular analysis settings, which are not clearly defined in the documentation and require laborious optimization. Additionally, the output from both tools lacks useful biological information; some of this information may be recovered by additional post-processing, but other important information is irretrievably lost during analysis (**Supplementary Table 1**).

In single T-cell transcriptome analysis, MiXCR outperformed TraCeR[6] in efficiency of TRA and TRB chains detection (**Supplementary Table 3**).

References:

1. Britanova, O.V. et al. *J Immunol* **196**, 5005-5013 (2016).
2. Bolotin, D.A. et al. *Nat Methods* **12**, 380-381 (2015).
3. Li, B. et al. *Nature genetics* **48**, 725-732 (2016).
4. Li, B. et al. *Nature genetics* **49**, 482-483 (2017).
5. Mose, L.E. et al. *Bioinformatics* **32**, 3729-3734 (2016).
6. Stubbington, M.J. et al. *Nat Methods* **13**, 329-332 (2016).

Supplementary Note 2. Intratumoral Ig repertoire derived from RNA-seq

We employed MiXCR to extract immune repertoires from the TCGA 48+48-bp paired-end RNA-seq data for 458 patients with cutaneous melanoma (SKCM). In terms of functional CDR3 clonotypes/CDR3-covering reads per sample, MiXCR yielded an average of 52/69 for TRA, 54/86 for TRB, 2.4/3 for TCR gamma (TRG), 0.15/0.2 for TCR delta (TRD), 395/3924 for IGH, 620/7595 for IGK, and 414/4939 for IGL (see doi:10.6084/m9.figshare.4620739 for clonesets). Notably, the extracted Ig repertoires were an order of magnitude larger than for the TCRs, indicating the presence of intratumoral Ig-producing plasma cells.

Furthermore, we noted that high intratumoral IGH expression levels as well as high levels of IGH clonality (calculated according to ref. 1) were associated with longer survival (**Fig. 2a,b**, **Supplementary Fig. 4a**), and the two parameters had strong cumulative value for patient stratification (**Fig. 2c**).

In many patients, a single dominant intratumoral Ig clonotype occupied 30–80% of all Ig CDR3 sequences in both heavy and light chains repertoires. Hypermutating IGH CDR3 variants could be observed even in primary tumor samples (**Supplementary Table 4**, **Supplementary Fig. 4b**), which could reflect the presence of intratumoral germinal centers[2, 3], or the co-infiltration of extra-tumorally hypermutated B cells encoding homologous Igs.

Higher TRB expression levels, reflecting greater tumor infiltration by T-cells, were also associated with longer survival (**Supplementary Fig. 4a**). TCR CDR3 repertoire clonality was not significantly associated with survival, but this could be attributable to insufficient information on TCR repertoires extracted from tumor RNA-seq data.

Analysis of isotype composition revealed that the IgG1 isotype was dominant among intratumorally-produced Igs (**Supplementary Fig. 4c**). A high proportion of IgG1 among IGH was associated with longer survival, while a high IgA/IGH proportion was associated with a negative prognosis (**Fig. 2d**). Proportions of IgD and IgE also tend to correlate negatively with survival, while proportions of IgG2, IgG3, IgG4 and IgM of IGH had no clear association with prognosis (**Supplementary Fig. 4d**).

To validate the observed effects, we analyzed 1,000 subsets of the original SKCM cohort by randomly sampling 50% of the total number of samples. For each iteration, we evaluated log-rank test χ^2 statistics for groups split by the given metric's median. This analysis demonstrated that the positive correlations of IGH clonality, high IGH expression, and high IgG1/IGH ratio with survival remains significant even in these smaller subsets of the SKCM cohort (**Supplementary Fig. 4e**).

It should be noted that there was a significant difference in the number of TCR and Ig CDR3 reads extracted from the samples from regional lymph nodes versus other tissue sites (**Supplementary Fig. 4f**). However, the association of high IGH clonality, high IGH expression, and high IgG1/IGH ratio with survival remained significant for the separately analyzed samples with or without lymph node tissue (**Supplementary Fig. 4g**).

To exclude the influence of the disease stage, we separately analyzed melanoma stage III samples, which are the most abundant stage in the TCGA data. Again, all effects remained significant (**Supplementary Fig. 4h**). The proportional hazard model suggested comparable regression coefficients for all three covariates (**Supplementary Fig. 4i**). Collectively, these results indicate that intratumorally produced Igs represent a critical component of efficient anti-tumor response.

References:

1. Tumeh, P.C. et al. Nature **515**, 568-571 (2014).
2. Gottlin, E.B. et al. Journal of thoracic oncology : official publication of the International Association for the Study of Lung Cancer **6**, 1687-1690 (2011).
3. Willis, S.N. et al. J Immunol **182**, 3310-3317 (2009).

Supplementary Note 3. Functional characterization of TCR repertoires based on RNA-seq of sorted T-cells

RNA-seq data from tissues that contain relatively low T-cell counts, as is the case with many tumor samples, allow to extract information only for the most abundant TCR clonotypes. However, melanoma samples with high T-cell infiltration allowed us to extract relatively rich TCR repertoires, comparable to the shallow targeted TCR profiling (**Fig. 1c**). Therefore, we predicted that for the RNA-seq data obtained from pure T-cell samples, such as sorted T-cells, efficient extraction and comparative analysis of TCR repertoires should be possible.

To test the feasibility of such an approach, we performed 50+50-bp paired-end RNA-seq analysis for the sorted effector (T_{eff}) and regulatory (T_{reg}) CD4 T-cells from the spleen and central nervous system (CNS) of six individual *Foxp3*^{yfpcre} mice[1] with induced experimental autoimmune encephalomyelitis (EAE). The near-100% abundance of T-cells in these samples allowed MiXCR to extract high-quality TCR repertoires comprising at average 1330/3295 TRA and 1489/3933 TRB unique functional CDR3 clonotypes/CDR3 sequencing reads per sample (**Supplementary Table 5,** see doi:10.6084/m9.figshare.4620739 for clonesets).

We determined an efficiency of about 90–100 TRB CDR3 reads per million unique reads. For small samples of 500 sorted T-cells covered by 3x10^7 reads, we could identify about 350–450 distinct clonotypes, indicating almost complete repertoire extraction (taking clonality into consideration), similar to the single cell RNA-seq. For large samples, the total number of reads remains the limiting factor. 50 million reads yield approximately 5,000 TRB CDR3 reads, which means up to 5,000 TRB clonotypes in theory, although in reality the number is usually lower due to the natural clonality of the repertoire. Roughly speaking, 50 million unique 50+50-bp paired-end RNA-seq reads could reveal the TRB CDR3 repertoire for approximately 5,000 T-cells randomly chosen from a large T-cell sample.

Extracted repertoires were suitable for the routine post-analysis using the VDJtools software[2]. First, we compared the diversity of repertoires, a challenging task in TCR profiling that preferably requires unique molecular barcoding for normalization of multiple samples[3, 4]. However, in paired-end RNA-seq, each sequencing read usually covers a unique starting RNA molecule, characterized with the unique starting nucleotide positions. PCR and optical duplicates are relatively rare and were excluded from the raw data. This allowed us to normalize samples for the accurate comparison of diversity metrics by extracting 500 random unique CDR3-containing reads, representing unique fragments of template RNA molecules from each sample. The diversity correlated well between the TRA and TRB repertoires (R>0.95, **Fig. 2e**), was similar between the effector and regulatory CD4 T-cell subsets, and was significantly lower in the CNS compared to spleen samples (**Fig. 2f**), reflecting the narrowed TCR repertoire in the CNS.

Next, we analyzed CDR3 characteristics for the full extracted repertoires, weighted for clonotypes frequency. T_{reg} cells were characterized by shorter TRB CDR3 lengths (**Fig. 2g**). The functional characteristics of the amino acids comprising the middle portion of CDR3 differed between T_{reg} and T_{eff} cells TRB repertoires (**Fig.**

2h). The higher interaction "strength"[5] of T_{reg} CDR3s is in keeping with the previously observed higher TCR affinity of T_{reg}s for self-peptide:MHC complexes, which may enable thymic T_{reg} precursors to compete more efficiently for the limited amount of antigens found on thymic antigen-presenting cells[6-8]. Analysis of amino acid TRB CDR3 repertoire overlaps revealed separate clustering of T_{eff} and T_{reg} cells, indicating functional similarity of subset repertoires across mice (**Fig. 2i**). Thus, we conclude that detailed and highly informative insights into the structure of TCR repertoires can be obtained by using RNA-seq data from sorted T-cell subsets.

References:

1. Rubtsov, Y.P. et al. Immunity 28, 546-558 (2008).
2. Shugay, M. et al. PLoS computational biology 11, e1004503 (2015).
3. Best, K., Oakes, T., Heather, J.M., Shawe-Taylor, J. & Chain, B. Scientific reports 5, 14629 (2015).
4. Britanova, O.V. et al. J Immunol 196, 5005-5013 (2016).
5. Kosmrlj, A., Jha, A.K., Huseby, E.S., Kardar, M. & Chakraborty, A.K. Proc Natl Acad Sci U S A 105, 16671-16676 (2008).
6. Hsieh, C.S., Zheng, Y., Liang, Y., Fontenot, J.D. & Rudensky, A.Y. Nature immunology 7, 401-410 (2006).
7. Jordan, M.S. et al. Nature immunology 2, 301-306 (2001).
8. Feng, Y. et al. Nature 528, 132-136 (2015).

Supplementary Table 1. Comparison of the key software characteristics.

	TRUST	V'DJer	MiXCR
Analysis of T-cell receptors	✔	✘	✔
Analysis of B-cell receptors	✘	✔	✔
Paired-end analysis	✔	✔	✔
Single-end analysis	✔	✘	✔
Species	Human	Human	Human, Mouse, Rat
Analysis of TCR-seq and/or IG-seq[1)]	✘	✘	✔
Assemble CDR3 clonotypes[2)]	✘	✘	✔
Reports full CDR3 sequence	✘[3)]	✔	✔
Reports clonal abundances[2)]	✘	✘	✔
Builds full-length sequences[4)]	✘	✔	✘
Annotates V/J genes	✘/✔[5)]	✘	✔
Annotates D gene	✘	✘	✔
Annotates C gene / antibody isotype	✘	✘	✔
Annotates V/D/J gene positions[6)]	✘	✘	✔
Failed on some samples	Yes[7)]/No	Yes[8)]	No
Median % of V/J annotated CDR3's	16%	N/A	100%
Median % of canonical CDR3's[9)]	18%	100%	95%
RAM requirement	~2Gb[10)]	~30+ Gb[11)]	~2 Gb
Average time of analysis of 10^8 reads, h	20h	28h (9 hours per chain)[12)]	4h
Average sample analysis cost, USD[12)]	0.46$	1.51$ (0.5$ per chain)	0.09$
Total analysis cost (all 140 samples, including truncated), USD[13)]	64$	106$ (35$ per chain)	12$
Operating system	Cross-platform	Linux	Cross-platform
Depends on external software	TopHat [14)]	STAR [14)]	No
Source-code available	✘	✔	✔

1) MiXCR allows for homogeneous analysis of TCR-seq, IG-seq, and RNA-seq datasets producing output in the same format, making possible further comparative analysis of outputs from both sources of information. Neither TRUST nor V'DJer support analysis of TCR-seq and IG-Seq data.

2) Assembled clonotypes and information about their abundances is the standard representation of TCR/Ig repertoire by MiXCR. Clonotype abundances are highly important for further data interpretation.

3) CDR3s reported by TRUST are truncated in nondeterministic way.

4) From the beginning of Framework 1 (FR1) till the end of Framework 4 (FR4)

5) In majority of cases TRUST has annotated either V or J gene but not both simultaneously

6) Positions of boundaries of V/D/J genes. This information allows to calculate number of 5'/3' truncated nucleotides of V/D/J genes and number of inserted N nucleotides. Such information is crucial for statistical inference of clonotype assembly probability and estimation of statistical significance of co-occurrence of same clonotype in several datasets.

7) On some samples (e.g. analysis of TRA chain for paired-end SPX6730 with >80bp length) TRUST reported zero number of CDR3s which is seemed to be a bug.

8) On some samples V'DJer failed to finish execution within 4 days running in 8 computer threads on Intel(R) Xeon(R) CPU E5-2683 v3 @ 2.00GHz and occupying more than 50 Gb of RAM.

9) For TCR analysis canonical amino acid CDR3 is considered as matching regex ^C[^_]*(?:[FW]|[FW]G.G)$. Since V'DJer provides no amino acid translation, for Ig analysis canonical nucleotide CDR3 is considered as matching ^TG[TC].*(?:TT[TC]|TGG)$ pattern, not matching (?:...)*(?:TAA|TAG|TGA)(?:...)* pattern and with the sequence length multiple of 3.

10) TopHat aligner is required to align raw sequencing reads before TRUST can assemble V-J contigs. TRUST itself consumes less than 1Gb RAM, while TopHat consumed more memory in all cases. Thus amount of RAM consumed by TopHat was used as an overall memory consumption value.

11) STAR aligner is required to align raw sequencing reads before V'DJer can assemble V-J contigs. STAR aligner consumes at least 30 Gb of RAM (depending on the reference human genome used). RAM requirement for V'DJer itself highly varies from sample to sample (e.g. analysis of SPX6730 required 50Gb of RAM).

12) To analyze all IGH, IGK and IGL chains, one has to run V'DJer three times (independently for each chain).

13) Bulk analysis of a large set of data samples requires substantial computation power. To analyze all 140 samples by three software tools in our benchmark we rented computer instances at Amazon Web Services (EC2). To minimize total analysis cost we used only spot EC2 instances during weekends, when price is the lowest. Instances with minimal required amount of RAM were used for each software. In case of V'DJer and STAR which require a huge amount of RAM we used r3.2xlarge instances ($0.083/hour). For running TopHat, TRUST and MiXCR c3.xlarge instances were used ($0.035/hour). We believe that this estimation reflects the real costs of organization hosting required computational power (spent on hardware, electricity, administration, etc).

14) Both TRUST and V'DJer require reads aligned to human genome in BAM format as input. TRUST requires raw reads to be aligned with TopHat aligner, while V'DJer requires STAR aligner. We found that when using TopHat alignments as input for V'DJer its performance degrades. When using STAR alignments as input for TRUST, it produces almost no results.

Supplementary Table 2. Estimating the number of extractable TRB CDR3 reads.

Sequencing length	TRB CDR3-containing reads per 1,000 TRBC sequencing reads, CI 95%
50 bp	20.2 ± 0.3
75 bp	50.0 ± 1.0
100 bp	81.4 ± 3.8
50+50 bp	46.2 ± 0.6
75+75 bp	80.7 ± 21.1
100+100 bp	128.2 ± 6.4

Sequencing length, bp	100+100 bp		100 bp		50+50 bp		50 bp	
Software tool	MiXCR	TraCeR	MiXCR	TraCeR	MiXCR	TraCeR	MiXCR	TraCeR
Functional TRA	231(85%)	223(82%)	227(83%)	220(81%)	226(83%)	215(79%)	215(79%)	206(76%)
Functional TRB	252(93%)	247(91%)	250(92%)	245(90%)	249(91%)	242(89%)	239(88%)	235(86%)
Functional pairs	221(81%)	209(77%)	215(79%)	207(76%)	214(79%)	198(73%)	196(72%)	186(68%)

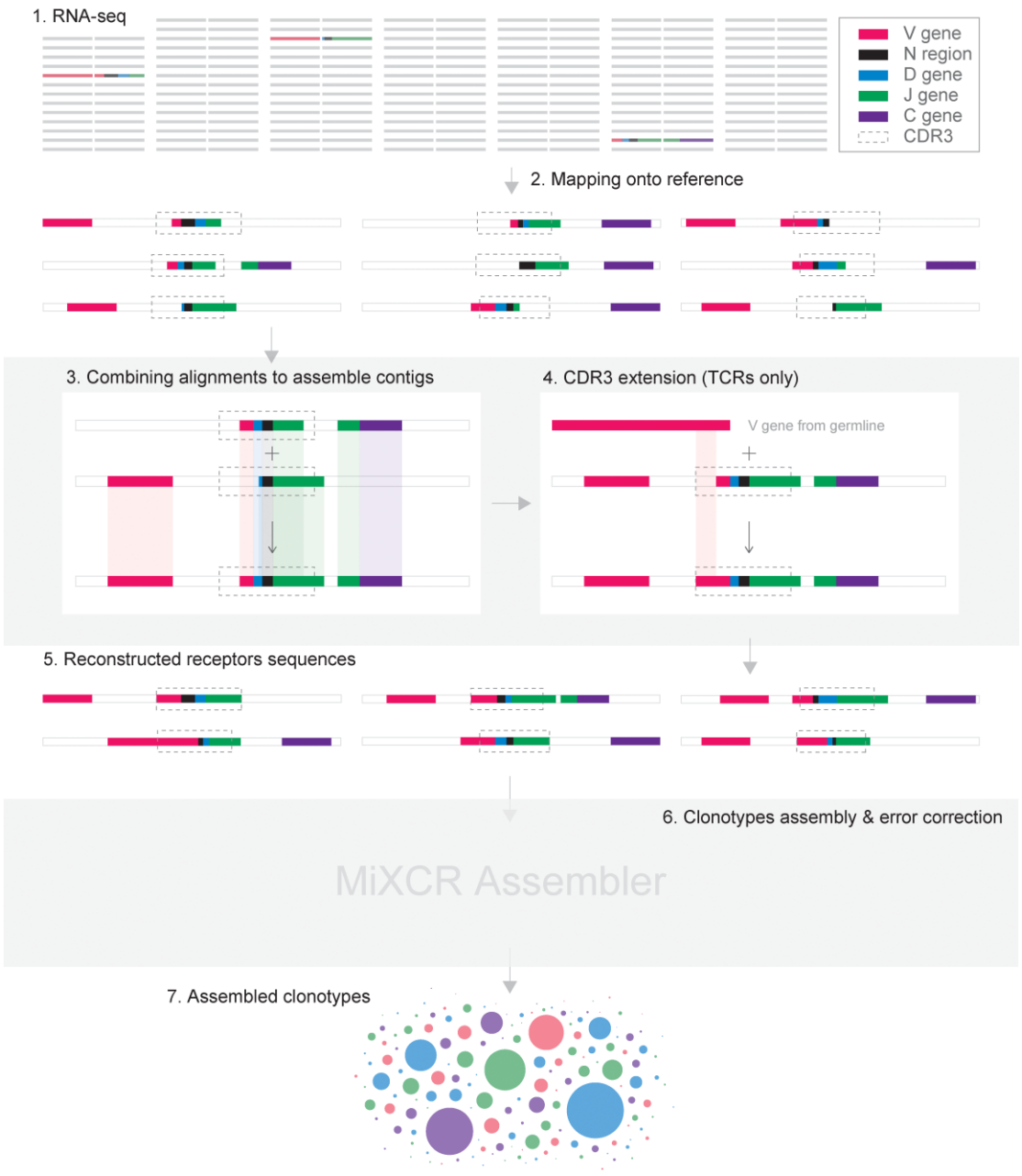
Supplementary Table 3. TRA and TRB CDR3 extraction from 272 single cells RNA-seq data. Single-cell transcriptome analysis is a rapidly developing methodology for characterization of cellular phenotypes and heterogeneity[1-3]. As MiXCR showed high efficiency for conventional bulk RNA-seq data, we tested its performance for previously published single-cell RNA-seq data from 272 $CD4^+$ T-cells, and compared the efficiency of MiXCR versus the TraCeR software that was originally used for analysis[4]. MiXCR showed higher efficiency for both functional chains, detecting at least one TRA and TRB chain in 85% and 93% of individual cells, respectively, and detecting productive pairs in 81% of cells (versus 77% for TraCeR). MiXCR also outperformed TraCeR with 50-bp single- and paired-end and 100-bp single-end reads that were *in silico*-generated from original 100-bp paired-end data.

References:

1. Macosko, E.Z. et al. Cell **161**, 1202-1214 (2015).
2. Fan, H.C., Fu, G.K. & Fodor, S.P. Science **347**, 1258367 (2015).
3. Klein, A.M. et al. Cell **161**, 1187-1201 (2015).
4. Stubbington, M.J. et al. Nat Methods **13**, 329-332 (2016).

Supplementary Table 5. **TCR repertoires extraction form sorted mice T cell RNA-seq.**

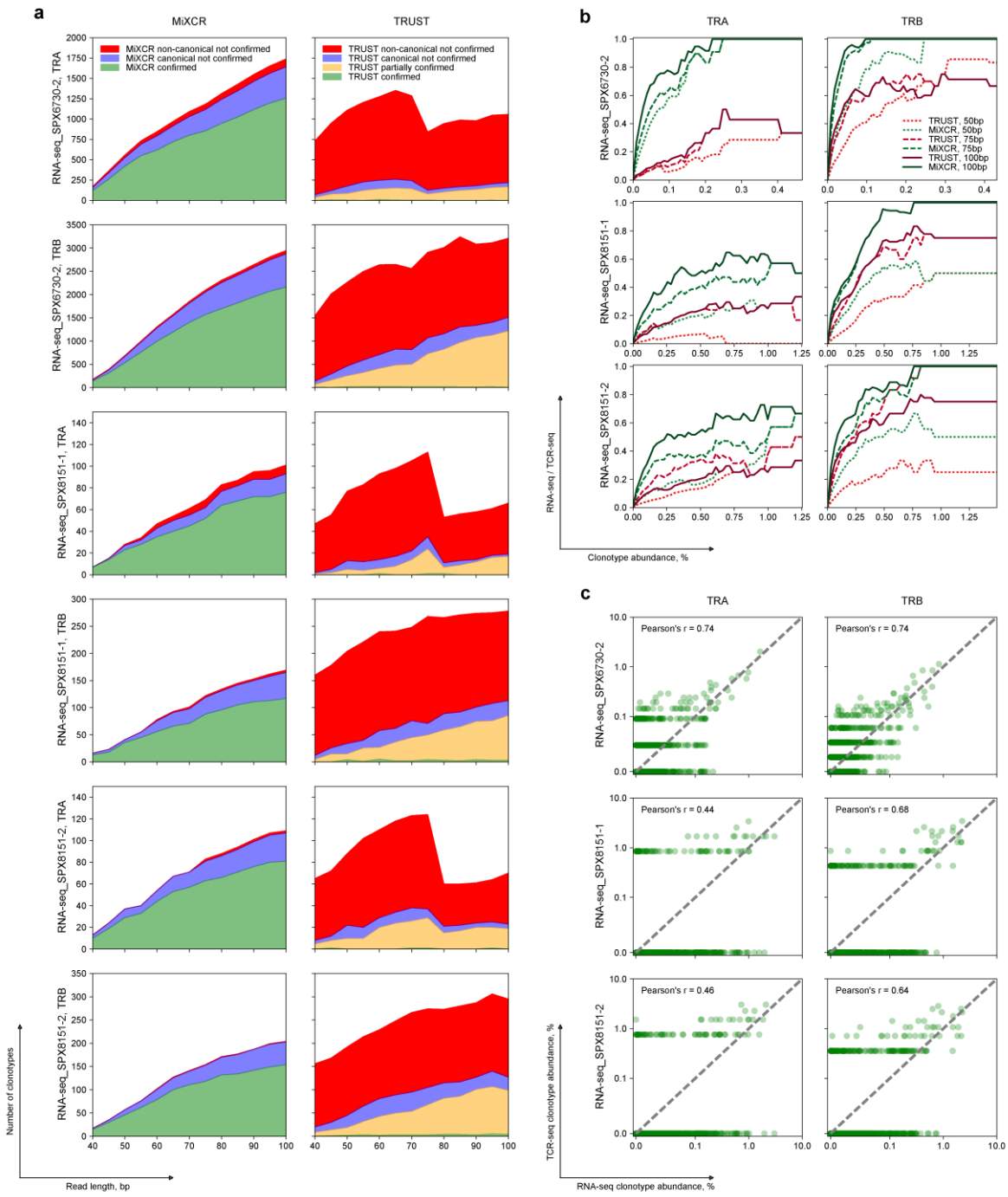
Sample ID	Approximate number of sorted T cells	RNA amount used for RNA-seq, ng	Total number of unique paired-end sequencing reads	Number of reads containing functional (in frame) TCR alpha CDR3	Number of reads containing functional TCR beta CDR3	Number of functional TCR alpha CDR3 reads per million reads	Number of functional TCR beta CDR3 reads per million reads	Number of extracted functional TCR alpha clonotypes	Number of extracted functional TCR beta clonotypes
CNS_Teff_1	8900	2	35587184	2543	2666	71	75	720	703
CNS_Teff_2	4100	all (1.3)	33431485	2789	3164	83	95	864	728
CNS_Teff_3	24000	2	32804755	2547	2589	78	79	700	611
CNS_Teff_4	6900	2	34756703	2826	3016	81	87	997	936
CNS_Teff_5	4000	all (1.0)	30220582	2396	2750	79	91	695	658
CNS_Teff_6	4000	all (1.0)	27260264	2464	2710	90	99	639	591
CNS_Treg_1	2000	all (1.0)	30444580	1767	1895	58	62	642	617
CNS_Treg_2	1300	all (0.9)	46112231	4476	4376	97	95	666	582
CNS_Treg_3	2400	all (1.3)	27457454	1793	1801	65	66	793	736
CNS_Treg_4	1900	all (0.8)	33091302	2418	2803	73	85	819	787
CNS_Treg_5	500	all (0.3)	35420782	2839	2942	80	83	462	408
CNS_Treg_6	500	all (0.3)	30011308	2622	2730	87	91	395	367
SP_Teff_1	100000	10	51204555	3439	4117	67	80	1571	1707
SP_Teff_2	100000	10	50886232	3605	4445	71	87	1970	2250
SP_Teff_3	100000	all (3.1)	52677819	3659	4648	69	88	1506	1684
SP_Teff_4	100000	10	57358961	3789	5147	66	90	1781	2006
SP_Teff_5	100000	10	48550774	2916	4127	60	85	1762	2088
SP_Teff_6	100000	10	57360066	5175	7092	90	124	2603	3352
SP_Treg_1	100000	10	49361408	3187	3549	65	72	1582	1726
SP_Treg_2	100000	10	55310370	4360	5600	79	101	2397	2977
SP_Treg_3	100000	10	54781194	4966	6156	91	112	2038	2390
SP_Treg_4	100000	10	51283048	4194	5749	82	112	1972	2398
SP_Treg_5	100000	10	53589946	4694	5596	88	104	2066	2565
SP_Treg_6	100000	10	51560460	3639	4736	71	92	2289	2881



Supplementary Figure 1

Pipeline overview.

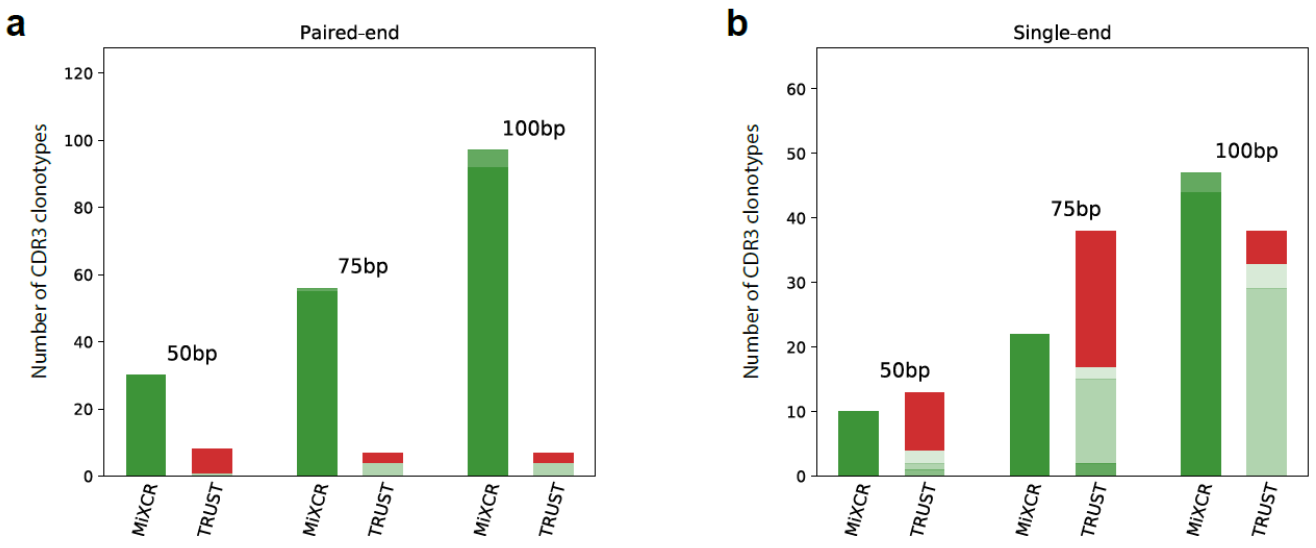
(1) MiXCR accepts raw paired- or single-end sequencing data as input. RNA-seq data contain only a tiny fraction of target Ig/TCR reads (one per 10^5-10^7 reads). (2) Alignment of raw sequencing reads to genomic sequences of V, D, J and C genes. Non-aligned reads are filtered out. (3) Reads containing only fragments of CDR3 are assembled into contigs with full or near-full coverage of CDR3. Assembly is performed only for reads with large overlap involving a substantial part of the hypervariable N region. (4) TCR sequences that contain defined V and J genes but do not fully cover the ends of the CDR3 are extended using germline sequence. This procedure is not performed for Igs because of possible hypermutations in the extended sequence. (5-6) The resulting sequences are clustered into clonotypes based on their CDR3 sequence. This step includes correction of artificial diversity from PCR and sequencing errors. (7) The primary output is a list of clonotypes with comprehensive information on their abundance, V/D/J genes composition, antibody isotype, CDR3 sequence topology, etc.



Supplementary Figure 2

Sensitivity and specificity of TCR and Ig repertoire extraction from tumor RNA-seq data, SPX6730-2, SPX8151-1 and SPX8151-2 tumor samples.

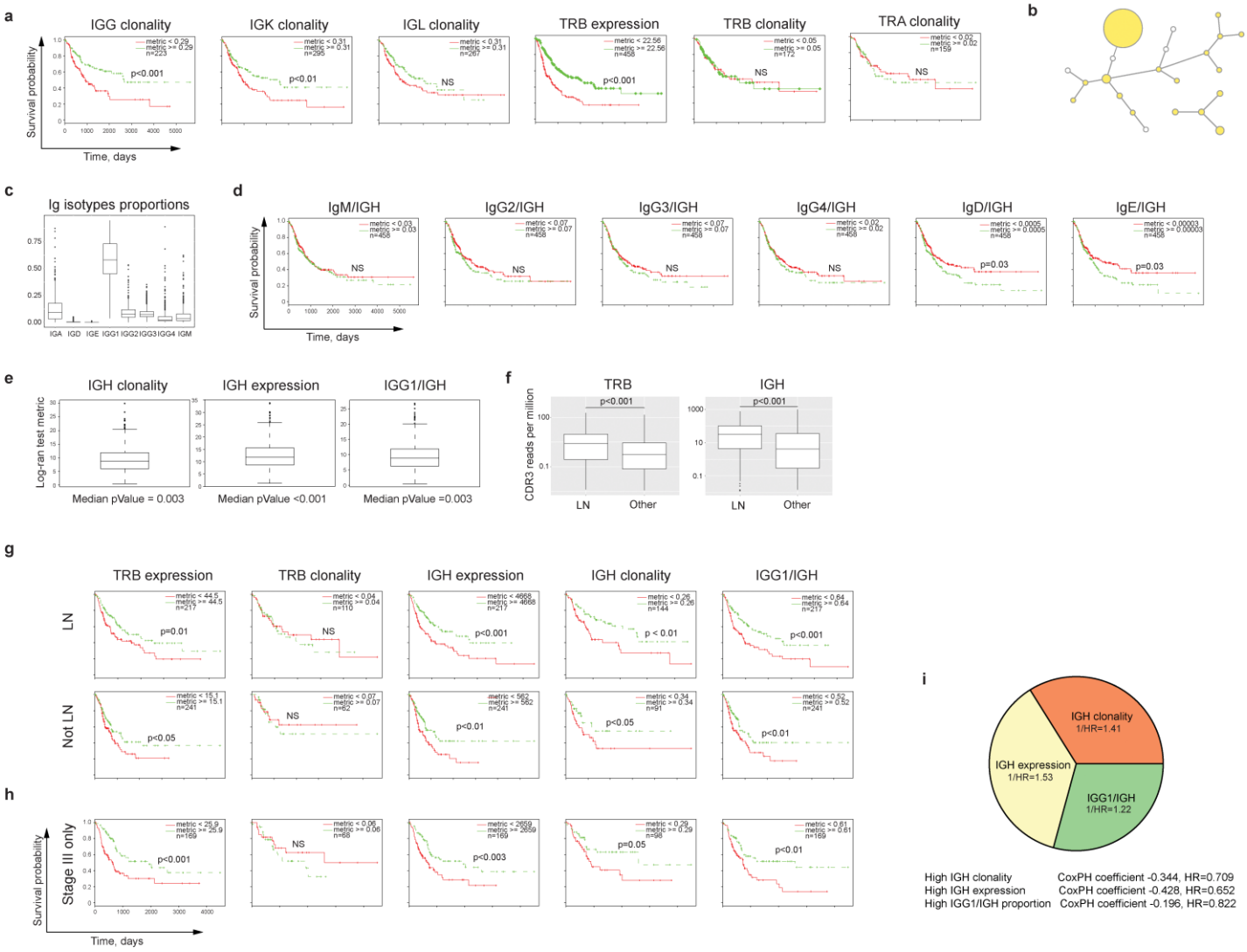
a. Dependence of total number of TCR beta CDR3 clonotypes, number of TCR-seq-confirmed clonotypes (green), number of canonical unconfirmed clonotypes (blue) and number of non-canonical unconfirmed clonotypes (red) on the paired-end sequencing reads length. In the case of TRUST green band states for definitely confirmed while orange for partially confirmed (allowing 6 nucleotides to be truncated) clonotypes. **b**. Dependence of the share of TCR-seq-confirmed TRA and TRB clonotypes extracted from RNA-seq data on the clonotype abundance (estimated by TCR-seq). The x-axis corresponds to clonotype frequency A in TCR-seq data. The y-axis corresponds to the fraction of identified clonotypes in the total number of control clonotypes with the frequency greater than A. **c**. Correlation of TCR clonotypes frequency in MiXCR-extracted repertoires from TCR-seq and RNA-seq data.



Supplementary Figure 3

MiXCR and TRUST performance comparison on *in silico* generated data.

a. Comparison on paired-end data. **b**. Single-end data. Dark green color corresponds to fully matched CDR3s (without any mismatches or indels), lighter shades of green denotes CDR3s matched with mismatches or indels (up to 3 mutations); red color denotes false-positive CDR3s (missing in the original set of synthetic clones).



Supplementary Figure 4

Validation of IGH clonality, expression and isotype ratio effects for TCGA SKCM samples.

a,d. Kaplan–Meier plots depicting the survival probability over time for high (>=cutoff) and low (<cutoff) metrics groups for IGG, IGK, IGL (for samples with >500 CDR3 reads of corresponding gene loci) and TRA, TRB (for samples with >50 CDR3 reads) repertoire clonality and TRB expressions (a) and isotype coverage ratios (d). Cutoff values were determined in terms of median values. *p* values of log-rank test for survival difference between low and high metrics groups are shown. n, number of patients. **b**. Exemplary lineage trees of hypermutating IGH CDR3 variants, sample SKCM193 (primary tumor). Adjacent nodes differ by exactly one nucleotide mismatch. The nodes are colored according to antibody isotype: IgG1 (yellow) or undetermined (white). Size of nodes is proportional to the clonotypes frequencies. **c**. Plot depicts the proportions of Ig isotypes among IGH. **e**. Analysis of 1,000 subsets of the original SKCM cohort, each representing a randomly chosen 50% of samples. For each iteration, we evaluated log-rank test χ^2 statistics for groups split by a given metric's median. Box plots report the distribution of the log-rank test score for the sampled subsets. *p* value is for the median log-rank test score over 1,000 simulations. **f**. TRB and IGH CDR3 reads extraction efficiency from lymph-node containing and non-lymph node samples. **g,h**. Kaplan–Meier survival plots for high (>=cutoff) and low (<cutoff) metrics groups for TRB expression, TRB clonality (for samples with >50 TRB CDR3 reads), IGH expression, IGH clonality (for samples with >500 IGH CDR3 reads), and IgG1/IGH proportion for lymph node-containing and non-lymph node-containing samples (f), and melanoma Stage III samples only (g). **i**. Inverse hazard ratios of the major covariates (IGH expression, IGH clonality and IgG1/IGH proportion) for all 458 samples. The hazard ratios are calculated by the proportional hazards regression model. Hazard ratios were inverted to reflect a favorable prognostic function of the covariates in the chart.