nature protocols

Review article

https://doi.org/10.1038/s41596-023-00841-8

Tutorial: integrative computational analysis of bulk RNA-sequencing data to characterize tumor immunity using RIMA

Received: 7 January 2022

Accepted: 22 February 2023

Published online: 30 June 2023

Check for updates

Lin Yang[1,11], Jin Wang[1,2,11], Jennifer Altreuter[1,11], Aashna Jhaveri[1,11], Cheryl J. Wong[1,3], Li Song[1,4], Jingxin Fu[1,2,5,6], Len Taing[5,6], Sudheshna Bodapati[1], Avinash Sahu[1,4], Collin Tokheim[1,4], Yi Zhang[1,4], Zexian Zeng[1,4], Gali Bai[1], Ming Tang[1], Xintao Qiu[5,6], Henry W. Long[5,6], Franziska Michor[1,4,7,8,9,10], Yang Liu[1] ✉ & X. Shirley Liu[1,4,6] ✉

RNA-sequencing (RNA-seq) has become an increasingly cost-effective technique for molecular profiling and immune characterization of tumors. In the past decade, many computational tools have been developed to characterize tumor immunity from gene expression data. However, the analysis of large-scale RNA-seq data requires bioinformatics proficiency, large computational resources and cancer genomics and immunology knowledge. In this tutorial, we provide an overview of computational analysis of bulk RNA-seq data for immune characterization of tumors and introduce commonly used computational tools with relevance to cancer immunology and immunotherapy. These tools have diverse functions such as evaluation of expression signatures, estimation of immune infiltration, inference of the immune repertoire, prediction of immunotherapy response, neoantigen detection and microbiome quantification. We describe the RNA-seq IMmune Analysis (RIMA) pipeline integrating many of these tools to streamline RNA-seq analysis. We also developed a comprehensive and user-friendly guide in the form of a GitBook with text and video demos to assist users in analyzing bulk RNA-seq data for immune characterization at both individual sample and cohort levels by using RIMA.

In the past two decades, it has become clear that the immune system plays a significant role in tumor progression and metastasis. Cancer immunotherapies harness a patient's innate and adaptive immune system to attack cancer cells. These immunotherapies include immune checkpoint blockade (ICB) therapy targeting cytotoxic T lymphocyte–associated protein (CTLA)-4, programmed death 1 (PD-1) and programmed death-ligand 1; adoptive T cell transfer of tumor-infiltrating lymphocytes; chimeric antigen receptor T cells; and personalized cancer vaccines[1]. Cancer immunotherapy treatments have shown durable remission and clinical success in various cancer types.

[1]Department of Data Science, Dana-Farber Cancer Institute, Boston, MA, USA. [2]School of Life Science and Technology, Tongji University, Shanghai, China. [3]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA. [4]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA 02115, USA. [5]Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA, USA. [6]Center for Functional Cancer Epigenetics, Dana-Farber Cancer Institute, Boston, MA, USA. [7]The Broad Institute of MIT and Harvard, Cambridge, MA, USA. [8]Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, MA, USA. [9]Center for Cancer Evolution, Dana-Farber Cancer Institute, Boston, MA, USA. [10]The Ludwig Center at Harvard, Boston, MA, USA. [11]These authors contributed equally: Lin Yang, Jin Wang, Jennifer Altreuter, Aashna Jhaveri. ✉e-mail: yangliu@ds.dfci.harvard.edu; xsliu.res@gmail.com

However, patient outcomes are heterogeneous and vary considerably. Many studies have been conducted to identify molecular features associated with tumor immunity and immunotherapy response. These molecular features include (i) genetic markers, (ii) gene expression signatures, (iii) measures of tumor immune infiltration, (iv) immune receptor repertoires and (v) characteristics of the microbiome. First, tumor mutation burden is a well-known genetic marker of immunotherapy response, because a large tumor mutation burden is associated with better ICB patient outcomes in non-small cell lung cancer and metastatic melanoma[2–4]. In addition, strong MHC binding affinity and T cell recognition of missense mutation-derived neoantigens have also been correlated with positive survival[5]. Second, in addition to cancer genetic markers, immune-related gene expression signatures have also been shown to have prognostic and predictive value for tumor immunity and immunotherapy response. Rooney et al.[6] quantified tumor cytolytic activity from granzyme A and perforin gene expression and correlated this measure with a survival benefit and improved prognosis. Ayers et al.[7] developed a 28-gene interferon-γ (INF-γ) signature predictive of anti-PD1 response that encompasses genes related to antigen presentation, chemokine expression, cytolytic activity and adaptive immune resistance. Third, measures of tumor immune infiltration also have predictive power for tumor immunity: Gentles et al.[8] revealed that tumor-associated leukocytes and prognostic genes are associated with tumor heterogeneity and cancer outcomes, and Thorsson et al.[9] integrated The Cancer Genome Atlas (TCGA) pan-cancer tumor gene expression profiles and identified six immune subtypes discriminated by tumor microenvironment (TME) features and survival outcomes. Fourth, profiling the immune repertoires of T cell receptors (TCRs) and B cell receptors (BCRs) helps elucidate the mechanisms of T and B cell tumor immunity: Zhang et al.[10] revealed the effect of T and B cell clonal expansion in a TCGA acute myeloid leukemia dataset, showing that highly expanded IgA2 B cells were associated with overall survival; Hopkins et al.[11] showed that the clonality of the T cell receptor repertoire is associated with patient survival and outcomes in anti-CTLA4– and anti-PD1–treated pancreatic ductal adenocarcinoma; and Tumeh et al.[12] also found a broader T cell repertoire inside the tumor of metastatic melanoma patients who responded to anti-PD1 therapy than in patients who did not. Finally, the microbiome also influences the host immune system and may contribute to cancer diagnosis and prognosis. For instance, Poore et al.[13] examined microbial reads from TCGA transcriptome data and found tumor-specific microbial signatures in tissue and blood samples, providing novel insights into the potential of microbiome-based cancer diagnostics. Furthermore, Gopalakrishnan et al.[14] found that higher gut microbiome diversity is associated with an improved response to anti-PD1 immunotherapy in metastatic melanoma.

RNA-seq is a cost-effective and versatile assay for the characterization of cancer cells and the tumor microenvironment. Computational methods using transcriptomic profiles can contribute to our understanding of tumor immunity and our ability to delineate prognostic and predictive markers of immunotherapy response. These methods provide valuable insights into immune response predictors such as gene expression signatures, estimates of tumor immune cell infiltration, immune repertoire profiles and microbiome features associated with immune response. To the best of our knowledge, there is no systematic pipeline to perform integrative RNA-seq analysis focused on tumor immunity and immunotherapy. In this tutorial and the accompanying online guide (https://liulab-dfci.github.io/RIMA/), we propose efficient, accurate and user-friendly practices for analyzing RNA-seq data, highlighting analyses related to tumor immunity and immunotherapy (Fig. 1 and Table 1). The GitBook website is continuously maintained and updated with immunotherapy features discovered in the latest studies. A companion immune-focused analysis pipeline (RNA-seq IMmune Analysis, RIMA) provides the ability to characterize tumor immunity from RNA-seq data, and its usage is demonstrated in the GitBook website. The RIMA pipeline was developed for clinical ICB analysis within the Cancer Immune Monitoring and Analysis Centers-Cancer Immunologic Data Commons[15] network (https://cimac-network.org/) as part of the Cancer Moonshot initiated by the U.S. National Cancer Institute. RIMA performs data preprocessing, differential gene expression analysis and immune-focused downstream analysis, including immune infiltration estimation, immune repertoire inference, immune response prediction, human leukocyte antigen (HLA) identification, gene fusion detection and microbiome analysis. The pipeline implements analysis at both individual and cohort levels, integrating comparative and association analyses between different immune features and clinical outcomes.

Read alignment and quality control

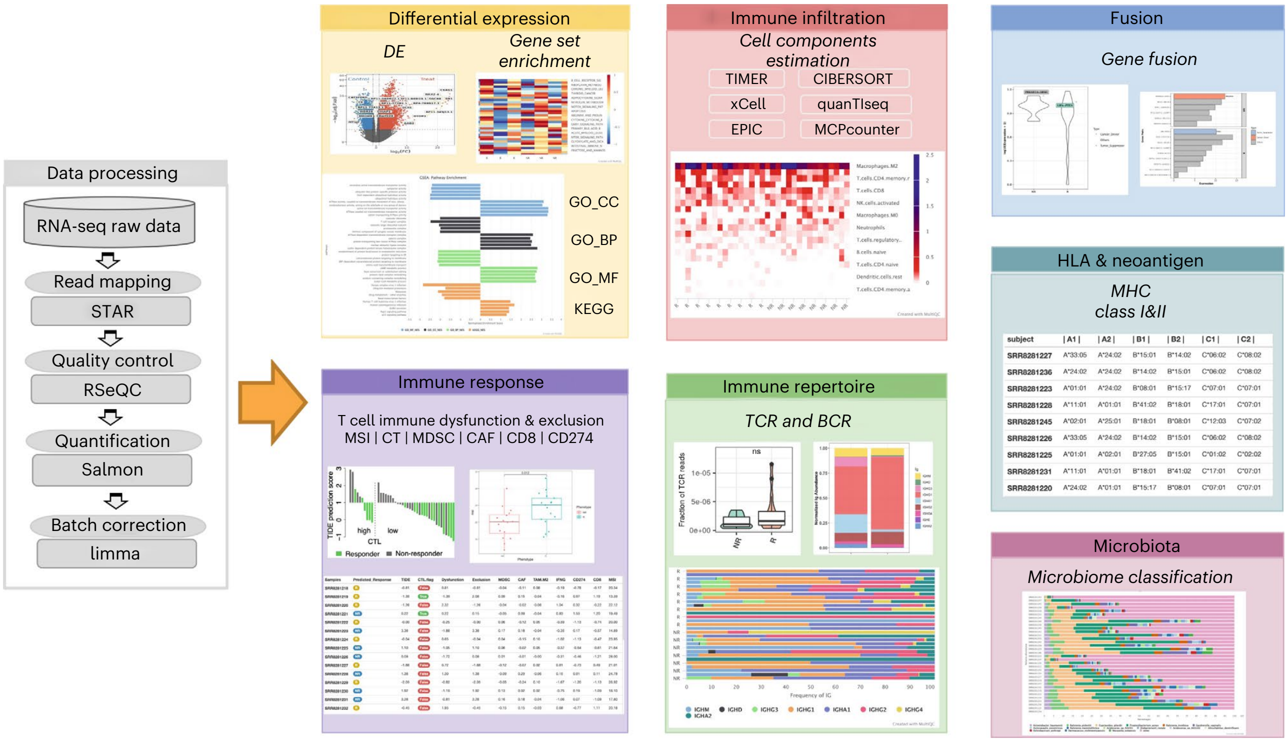
Quality control and read alignment are essential preliminary steps for RNA-seq data analysis. STAR[16] is commonly used for RNA-seq read alignment and generates results in the binary alignment map (BAM) format at the transcriptome or genome level for downstream analysis. RseQC[17] is a standard package for verifying the quality of raw RNA-seq reads and alignments with multiple built-in functions. For example, 'read_quality.py' graphically displays the quality of each base call from the 5′ end to the 3′ end of a read. 'tin.py' quantifies the RNA transcript integrity number at the transcriptome level, and a median transcript integrity number (medTIN) score for all expressed transcripts can be used to infer the overall sample integrity and quality. 'read_distribution.py' summarizes the fraction of reads aligned into different genomic regions, such as exon and intron regions. 'geneBody_coverage.py' provides RNA-seq read coverage over the gene body for all genes in the sample. 'junction_saturation.py' detects splicing junctions with different resampling percentages of reads to determine if the sequencing depth is sufficient to perform alternative splicing analysis[18–20]. In downstream analysis, users might consider removing low-quality samples with low alignment fractions (percentage of reads mapping to the genome <70%) and low integrity (medTIN < 30)[21,22].

Gene quantification

After the alignment of RNA-seq reads, HTSeq[23], RSEM[24], Kallisto[25] and Salmon[26] are widely used for gene quantification using the BAM files. HTseq counts map reads to transcripts for gene quantification, which is not normalized by the sequencing depth and gene lengths. Transcripts per million (TPM) and reads or fragments per kilobase of exon per million reads are alternative measures of normalized gene expression levels that account for both sequencing depth and gene length. RSEM performs both gene-level and transcript-level quantification, generating both TPM and reads or fragments per kilobase of exon per million reads values. Kallisto and Salmon conduct fast transcript-level quantification with less memory consumption needed to generate TPMs as compared to RSEM. Pseudo-aligners like Kallisto and Salmon speed up alignment by using a transcript-level reference rather than a genome reference. The pseudo-aligner Kallisto runs on raw fastq files, while Salmon runs on both raw fastq and alignment BAM files. Although RSEM is considered the gold standard of RNA-seq quantification[27], the pseudo-aligners Salmon and Kallisto, which can align reads to the transcripts faster without aligning reads to the genome, achieve almost as good accuracy as that achieved by RSEM, with a significant speed advantage (Fig. 2a).

Batch effect removal

Batch effects arise from systematic biases in experimental batches and can confound downstream cohort-level analyses such as the identification of differentially expressed genes. Batch effects are easily overlooked but necessary to consider when preparing the expression matrix input for bioinformatic algorithms. Principal component analysis or unsupervised clustering are generally used to identify and visualize potential batch effects within a cohort. If batch effects are present



subject	A1	A2	B1	B2	C1	C2
SRR8281227	A*33:05	A*24:02	B*15:01	B*14:02	C*06:02	C*08:02
SRR8281236	A*24:02	A*24:02	B*14:02	B*15:01	C*06:02	C*08:02
SRR8281223	A*01:01	A*24:02	B*08:01	B*15:17	C*07:01	C*07:01
SRR8281228	A*11:01	A*01:01	B*41:02	B*18:01	C*17:01	C*07:01
SRR8281245	A*02:01	A*25:01	B*18:01	B*08:01	C*12:03	C*07:02
SRR8281226	A*33:05	A*24:02	B*14:02	B*15:01	C*06:02	C*08:02
SRR8281225	A*01:01	A*02:01	B*27:05	B*15:01	C*01:02	C*02:02
SRR8281231	A*11:01	A*01:01	B*18:01	B*41:02	C*17:01	C*07:01
SRR8281220	A*24:02	A*01:01	B*15:17	B*08:01	C*07:01	C*07:01

Fig. 1 | Flowchart of immune analysis of bulk RNA-seq data using RNA-seq IMmune Analysis (RIMA). RIMA is comprised of a preprocessing data module and seven downstream modules related to the tumor immune microenvironment. The preprocessing module includes read alignment, quality control, gene quantification and batch effect removal. Downstream modules include differential expression and gene set enrichment analysis, immunotherapy response prediction using known gene expression signatures, immune cell infiltration estimation, immune repertoire profiling, gene fusion identification, human leukocyte antigen (HLA) typing and microbiome analysis. DE, differential expression; KEGG, Kyoto Encyclopedia of Genes and Genomes.

(i.e., when samples cluster by batches instead of by biological conditions), limma[28] and Combat[29] are commonly used to correct for these technical artefacts. Combat uses an empirical Bayes approach to eliminate batch effects, which is critical to avoid over-correction when the batch size is small, whereas limma applies a linear model to correct for batch effects and is more efficient for datasets of more than four batches[30]. It is essential to mitigate batch effects in gene expression profiles for informative downstream analysis.

Differential expression and gene set enrichment

Differential expression analysis facilitates the discovery of genes and pathways whose representation differs in specific biological conditions. DESeq2, EdgeR and limma-voom[31–33] are popular tools for differential expression analysis. DESeq2 fits a generalized linear model to estimate the coefficient and log fold change of genes between treatment and control conditions, whereas edgeR uses a negative binomial distribution model to achieve that same goal. Both tools apply Empirical Bayes shrinkage for estimating the expression dispersion when each condition has <20 replicates per condition[34]. limma-voom also fits a linear model and is more efficient for large sample sizes but less sensitive than DESeq2 and EdgeR. For calculating differential expression at the gene level, the 'tximport' R package[35] is used to convert transcript-level abundance (TPM) to gene-level estimated counts before performing differential expression analysis.

Gene set enrichment analysis (GSEA)[36] is usually conducted after differential expression analysis to detect gene expression patterns that affect pathways, molecular functions, cellular components and processes. Databases such as the Kyoto Encyclopedia of Genes and Genomes[37] (KEGG), Gene Ontology[38] (GO) and the Molecular Signature Database[39] (MSigDB) are generally used for enrichment analysis. Results from differential expression analysis are used to rank genes according to metrics of choice (e.g., the difference in expression means, fold change, a *P* value from a *t* test or Simpson or Shannon diversity indices). GSEA then uses this ranked list of genes to calculate enrichment scores for the pre-defined gene sets contained in databases such as KEGG, GO and MSigDB. Single-sample gene set enrichment analysis (ssGSEA)[40] is an extension of GSEA that directly calculates enrichment scores of user-specified pathways or gene signatures from the expression profile of each sample. The ssGSEA score of each sample on different pathways can then be combined and compared to identify associations with phenotypes.

Cell infiltration estimation

The TME consists of endothelial cells, immune cells, stromal cells and extracellular factors surrounding tumor cells. Estimating cellular components is essential for correctly classifying TME phenotypes[41] and untangling the mechanisms of tumor immune evasion[42]. There are two major approaches for estimating immune cell infiltration in the TME: deconvolution-based and marker-based approaches. Deconvolution-based methods, such as TIMER[43], quanTIseq[44], EPIC[45] and CIBERSORT[46], consider a given gene expression profile as a linear combination of pre-defined immune gene signatures present at different ratios. A linear regression model is often applied to estimate the coefficients of genes, which are later used to infer immune cell abundances or fractions. Marker-based approaches, such as xCell[47] and MCP-counter[48], quantify the signature enrichment score of a list of cell type–specific marker genes from gene expression profiles. ImmuneDeconv[49] is an R package that integrates the above six algorithms to estimate the extent of infiltration of immune and stromal cells. In addition, EPIC and quanTIseq also assess uncharacterized cells defined as cancer cells. CIBERSORT absolute mode, EPIC and quanTIseq support inter-comparison between sample groups and intra-comparison between

Table 1 | The most common computational tools and methods for RNA-seq immune analysis

Tasks	Approach	Tool	References	Notes
Read mapping	Aligns spliced transcripts to the genome reference by using the suffix array method	**STAR**	16	Common alignment tool
Quality control	Different methods for generating multiple quality control metrics	**RSeQC**	17	Common tool for alignment quality checks
Gene quantification	Pseudo aligner	**Salmon**, Kallisto	25,26	Salmon supports fastq and bam inputs; Kallisto supports fastq
	Iteration of expectation-maximization to estimate alignment length	RSEM	24	Relatively slow; quantifies isoforms
	Counts mapping reads over exons	HTSeq	23	Exon-level read count quantifications
	Linear model to analyze RNA data-integrated experiment	**limma**	33	Two-way ANOVA to avoid over-correction
Batch correction	Negative binomial regression model	Combat	29	Empirical Bayes to avoid over-correction; better for small batches
	Shrinkage estimation for dispersions and fold changes	DESeq2	31	–
Differential	Empirical Bayes methods	EdgeR, limma-voom	28,32	limma-voom is efficient for large-scale data
	Applies the Kolmogorov-Smirnov statistic to a gene list	**ClusterProfile**	36	Supports KEGG, GO and MSigDB enrichment
Gene set enrichment	Extension of GSEA, single-sample gene set enrichment	**ssGSEA**	40	–
	De novo assembly of candidate reads from TCR/BCR genes	**TRUST4**	66	Faster and sensitive for longer assemblies; supports both bulk and single-cell RNA-seq data
Immune repertoire	V-support vector regression	**CIBERSORT**	8	Detects 22 immune cell types
Immune infiltration	Constrained least square regression	**TIMER**, TIMER2	43	Estimates 6 immune cell types; TIMER2 allows users to choose the most-related cancer types inferred from TCGA
		quanTIseq	44	Estimates 10 immune cell types
	Estimates the immune cell abundance by using ssGSEA from curated public datasets	**xCell**	47	Estimates 64 cell types, including lymphoid, stem, myeloid and stromal cells
	A deconvolution-based method that uses constrained least square regression	**EPIC**	45	Estimates 6 immune cell types, cancer-associated fibroblasts and endothelial cells
	A marker-based method that uses curated transcriptomic markers	**MCP-counter**	48	Estimates 8 immune cell types, cancer-associated fibroblasts and endothelial cells
Immune response	Estimates T cell dysfunction and exclusion from expression data	**TIDE**	54	http://tide.dfci.harvard.edu/
MSI	Uses Pearson's chi-squared test on the distribution of repeated microsatellite sequences	**MSIsensor2**, MSIsensor	57,59	MSIsensor2 supports tumor-only and tumor-normal paired data; MSIsensor supports tumor-normal paired data
	Uses the Z-score test to calculate single-locus instability	mSINGS	58	Long running time; supports tumor-only data
HLA allele	Alignment-based approach	**arcasHLA**	85	Supports tumor/normal paired RNA sequencing data
		OptiType	87	Supports MHC class I RNA/DNA data
		POLYSOLVER	88	Supports MHC class I DNA data and detects HLA mutation
Neoantigen	Integrates neoantigen identification tools: NetMHC, SMMAlign and MHCflurry	pVACSeq	93	–
Gene fusion	Uses chimeric and discordant read alignments from STAR to predict fusion genes	**STAR-fusion**	81	Fast and efficient common fusion calling tool
	Uses BLAST to calculate the similarity of fusion genes	**pyPRADA**	82	Detects homologous genes and filters those genes from the fusion-calling results
Microbiome	Expectation-maximization algorithm to classify microbial sequences	**Centrifuge**	106	Fast alignment and sensitive for short read exact matches
	Lowest common ancestor algorithm to match to taxa in a reference database	Kraken	102	Fast, accurate and sensitive labeling of reads and quantification of species
	Aligns the non-host reads to microbial genomes by using BWA-MEM aligner	PathSeq	105	Pathogen discovery

The tools shown in bold are integrated into the RIMA pipeline.

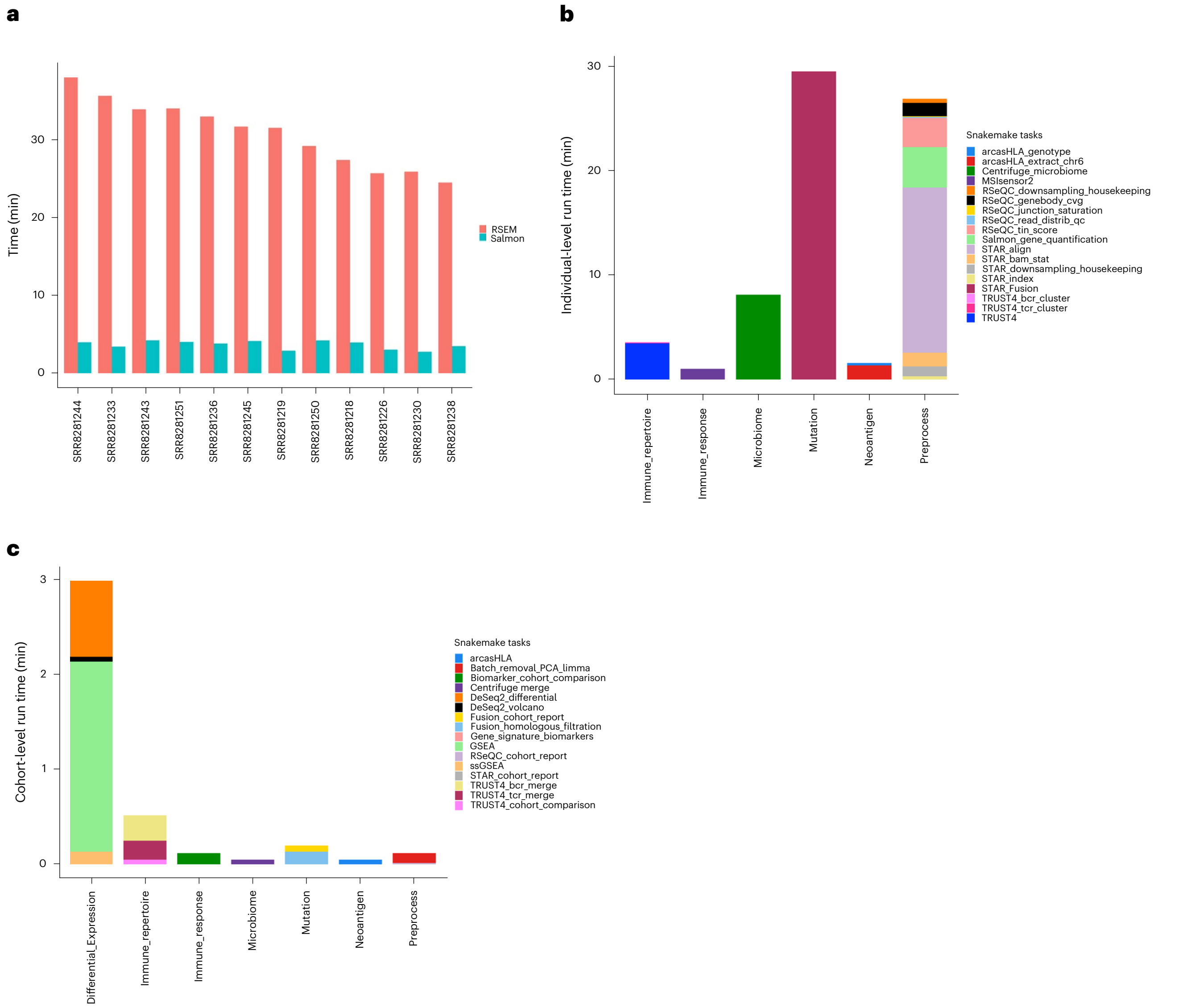


Fig. 2 | Running time of RIMA pipeline. a, Comparison of running time (minutes) when using Salmon and RSEM for gene expression quantification. **b**, Median running time (minutes) of individual-level Snakemake tasks for each module of RIMA. **c**, Running time (minutes) of cohort-level Snakemake tasks for each module. More detailed information about RIMA modules can be found at https://liulab-dfci.github.io/RIMA/.

cell types, whereas TIMER, xCell and MCP-counter support only inter-comparisons between sample groups within the same cell type. Depending on the specific gene signature used by each algorithm, different algorithms cover slightly different cell types and thus perform differently on specific immune or stromal cells[50]. A user may wish to select tools for particular immune cells and to evaluate the consistency of predictions stemming from different algorithms, the agreement of results with estimations from other modalities and/or the derivation of the marker genes used by each tool. Newer algorithms are being developed to improve predictions that use well-annotated single-cell RNA-seq datasets as reference and attempt to impute cell type–specific gene expression[51,52].

Immunotherapy response prediction

Immunotherapy, especially ICB, has made excellent progress in treating advanced-stage cancer patients[53]. An increasing number of gene expression biomarkers have been used for predicting ICB response. These biomarkers include ICB-related gene signatures (*PDCD1*, *CD274* and *CTLA4*), interferon-gamma signaling signatures and MHC class I&II antigen presentation levels. For instance, Jiang et al.[54] developed the TIDE algorithm to measure T cell dysfunction and exclusion from tumor expression data and to predict ICB outcomes. The T cell dysfunction score was derived by systematically identifying genes whose effects on survival are dependent on cytotoxic T cell infiltration levels as estimated from TCGA data. In contrast, the T cell exclusion score was calculated from the average expression of gene signatures for cancer-associated fibroblasts, myeloid-derived suppressor cells and M2 tumor-associated macrophages. The algorithm has been integrated into the TIDE website (http://tide.dfci.harvard.edu/), which contains >1,000 tumor RNA-seq profiles from samples obtained after ICB treatment across >20 cohorts to evaluate and compare different biomarkers. In addition to T cell dysfunction and exclusion, the TIDE website provides other metrics to infer the TME and T cell status (e.g., levels of cytotoxic T lymphocytes and other cell types known to restrict T cell infiltration[6], an IFN-γ–related gene signature[7] and an 18-gene T cell–inflamed expression signature developed by Merck[55]). The performance of applying these gene signatures for predicting the observed clinical outcomes is evaluated across all cohorts contained on the TIDE website.

Microsatellite instability (MSI) was identified as a predictive biomarker for cancer immunotherapy response in multiple cancer types[56].

Several tools were developed to estimate MSI scores from RNA-seq data: for instance, MSIsensor[57] predicts the MSI score from tumor-normal paired profiles, whereas mSINGS[58] accepts tumor-only inputs but is less sensitive than paired tumor-normal estimation. Recently, MSIsensor2[59] was developed to improve prediction accuracies, allowing tumor-only inputs.

Immune repertoire inference

V(D)J recombination is a genome rearrangement event affecting TCRs and BCRs during T and B cell maturation. The random joining of V, D and J genes and the indels and mutations introduced at the joining junctions produce the highly variable complementarity-determining region 3 (CDR3) on TCRs/BCRs. TCR/BCR CDR3 diversity allows T cells and B cells to recognize different external pathogens or tumor-associated antigens[60,61] for immune cell activation. Upon antigen recognition and clonal expansion, B cells also undergo somatic hypermutation to improve antigen-binding affinity and isotype class switching to elicit different downstream immune signaling pathways. Emerging sequencing techniques such as TCR-seq and BCR-seq have been developed to profile the T and B cell repertoires[62]. MiXCR[63] is a widely used algorithm to analyze TCR/BCR-seq data. It first aligns candidate reads to the reference germline sequences from the ImMunoGeneTics (IMGT), which integrates knowledge of immunoglobulins, TCRs, major histocompatibility and other immune-related proteins. MiXCR then identifies the reads spanning the CDR3 region and assembles the overlapping reads into CDR3 sequences. However, TCR/BCR-seq assays are expensive and not always feasible to perform on small biopsy samples. Because TCR and BCR transcripts are also present in RNA-seq data, computational algorithms have been designed to infer immune repertoires from RNA-seq data. TRUST is an algorithm that assembles TCR[64] and BCR[65] sequences from bulk RNA-seq. The latest version of the pipeline, TRUST4[66], significantly improves assembly accuracy and computational efficiency over those of its predecessors and supports the analysis of single-cell RNA-seq data. TRUST4 directly conducts de novo assembly of candidate reads from TCR/BCR genes and then aligns those assembled contigs to the IMGT reference database[67] to identify CDR3s. Both MiXCR and TRUST tools output the CDR3 sequence, the frequency of V/J genes and VJ gene pairs, BCR constant gene usage and the full-length VJ sequence.

TCR/BCR clonality and diversity are important TME characteristics associated with immunotherapy response[68,69], which can be calculated from the normalized Shannon entropy of individual clonotypes derived from TCRs/BCRs. For TCRs, CDR3 sequences are directly used to represent the clonotypes. For BCRs, because highly similar CDR3s from somatic hypermutations belong to the same lineage, these sequences should be clustered first to represent a single clonotype. In addition to TCR/BCR clonality and diversity, the somatic hypermutation rate and isotype quantification of BCRs are also informative for B cell activation and TME status[64,70]. The somatic hypermutation rate can be measured by the variance within clustered CDR3 sequences or the dissimilarity between the V gene and the germline reference from the IMGT database[65]. Ig isotype quantification is determined by the abundances of each isotype of the BCR heavy chain[65].

Tumor mutation detection and characterization of the mutational landscape

Genome instability is a hallmark of cancer[71], and tumor mutations continue to accumulate during tumor initiation, progression and metastasis. The tumor mutation burden has been identified as an effective marker for predicting ICB response[2,4]. The standard approach for somatic mutation detection is the analysis of whole-exome or whole-genome sequencing data from tumor-normal matched pairs. Computational algorithms are used to call mutations from the resulting data. For example, MuTect is based on a Bayesian classifier[72], MuSE uses a Bayesian Markov model[73], SomaticSniper[74] uses a Bayesian comparison of genotype likelihoods and Varscan2 is based on a heuristic algorithm classifying somatic mutation status[75]. Among these approaches, Varscan2 can be applied to RNA-seq data with good coverage[76,77], although the mutation calls are still expected to be noisier than those from DNA sequencing. After somatic mutation detection, VEP[78] can be used to annotate mutations on genes, transcripts and regulatory regions.

In addition to point mutations, gene fusions are another type of genetic alteration arising in cancer cells and can act as cancer drivers[79]. There are two major ways to form gene fusions: (i) a genome structural rearrangement at the chromosome level, such as an insertion, deletion, or translocation; and (ii) chimeric RNAs generated through read-through splicing or trans-splicing, which can be detected from RNA-seq data[80]. STAR-Fusion[81] is a popular algorithm with high speed and accuracy that leverages chimeric and discordant read alignments from STAR to predict splicing fusions from the chimeric alignment transcripts. pyPRADA[82] is another tool conducting supervised and unsupervised detection of fusion events from RNA sequences and helping to filter out homologous genes with high sequence similarity.

Neoantigen prediction

Neoantigens are highly expressed new peptides derived from tumor mutations that can bind to antibodies or T cell receptors[71]. These peptides are small polymers generated from protein degradation by proteasomes inside the cell and are then presented on the cell surface by MHCs. The HLA complex on chromosome 6 encodes most of the proteins that make up the human MHCs. Thus, the HLAs are polymorphic and encode MHCs that have different propensities to present different peptides. Two major classes of human MHC, class I and class II, are involved in antigen presentation. MHC class I molecules are generally expressed on normal cells and present internal degraded proteins, whereas MHC class II molecules are normally expressed on professional antigen-presenting cells and present processed external antigens[83,84]. Many algorithms have been developed to accurately call HLA alleles from DNA and/or RNA sequencing data. Current alignment-based HLA typing methods can predict both class I and II HLA alleles against reference sequences from the IMGT database. For example, arcasHLA[85] and seq2HLA[86] are designed to predict both MHC class I and class II alleles from RNA-seq data. OptiType[87] can be applied to RNA-seq, whole-exome and whole-genome sequencing data for the identification of MHC class I alleles. POLYSOLVER[88] is another HLA-typing tool that predicts class I alleles from whole-exome sequencing data and detects mutations in HLA genes.

Neoantigen recognition is useful for the development of cancer neoantigen vaccines and helps characterize the essential drivers of T cell activity[89]. In addition, a fraction of predicted neoantigens has been confirmed to be immunogenic[90], and the neoantigen load, which can be calculated as the total number of predicted neoantigens, is associated with ICB response[83,91]. Different HLA alleles present different neoantigens on the cell surface, which contribute to neoantigen recognition by the immune system. Many computational tools have been developed to help identify potential neoantigens by predicting the binding affinity of peptides to specific HLAs. Usually, tumor mutated peptides with stronger MHC binding affinity (half maximal inhibitory concentration < 500 nm)[92] compared to normal peptides are considered neoantigens. The pVAC-Seq[93] is a robust neoantigen prediction pipeline integrating somatic mutations, HLA alleles and common MHC binding affinity prediction tools, such as NetMHC[94], MHCflurry[95] and SMMalign[96].

Microbiota classification and quantification

Microbiota and their interaction with host tissues contribute to the human innate and adaptive immune system and influence a host's susceptibility to cancer[97,98]. For instance, modulating gut microbial components has been shown to influence ICB response in melanoma[99,100]. In addition, tumor microbial profiles are also important for virus study and are cancer type specific[13,101]. Traditional 16S rRNA sequencing is commonly used to determine microbial phylogeny and taxonomy from the hypervariable regions of 16S ribosomal RNA. High-throughput sequencing, generating microbial DNA/RNA sequences, is also widely used for microbiome classification. Kraken[102,103] is a highly used ultrafast tool for metagenomic sequence classification from DNA sequences. RNA-seq has also been used for virus detection, including human papillomavirus, hepatitis B and Epstein-Barr virus[104]. PathSeq[105] was first developed to detect microbes by using both transcriptome and whole-exome data, which aligns non-host reads to predefined microbial organisms. Centrifuge[106] significantly speeds up the classification process with comparable accuracy to Kraken and is more sensitive for short, exact matches. Thus, Centrifuge better captures signals from RNA-seq reads spanning exons within the human host and viral genomes.

RIMA pipeline implementation

We developed a comprehensive bulk RNA-seq analysis pipeline named 'RIMA' to characterize the TME on the basis of RNA-seq data (https://liulab-dfci.github.io/RIMA/)[21,107,108]. RIMA distinguishes itself from other RNA-seq analysis pipelines by integrating a basic preprocessing module with seven downstream modules focused primarily on immune-related analyses. The preprocessing module includes read alignment, quality control, gene expression quantification and an evaluation of potential batch effects. Downstream modules include differential expression analysis, immune infiltration estimation, immune repertoire profiling, fusion identification, HLA typing, immunotherapy response prediction and microbiome analysis. RIMA performs both individual- and cohort-level analyses. For individual-level analysis, each sample's raw fastq reads or alignment BAMs are used to quantify expression, profile the immune repertoire, identify gene fusions and determine the HLA alleles, MSI score and microbiome abundance. The cohort-level analysis applies customized scripts to format individual-level analysis outputs and combine them for downstream comparisons between customized clinical phenotypes. For example, the merged gene expression matrix can be used to identify differentially expressed genes, estimate immune infiltration and calculate gene expression signature scores for immune response prediction. Individual results from TRUST4 are combined for downstream comparison analysis of TCR/BCR clonality and diversity. The Python-based Snakemake pipeline management system is used to support the running flow between the different bioinformatics tools applied in the preprocessing module and downstream analysis modules of RIMA, and R Bioconductor packages are used for statistical analysis and result visualization. Our Gitbook tutorial demos the RIMA pipeline by using 12 glioblastoma samples[109] on a Google cloud n2-standard-48 machine equipped with 48 CPUs and 198 GB of memory. The running times of Snakemake tasks at both individual and cohort levels are summarized in Fig. 2b,c. The RIMA pipeline is useful for large-scale data processing and transcriptomic characterization of the TME, thus offering insights into cancer genomics and immuno-oncology.

RIMA is one uniform pipeline that integrates key immune-associated features, which allow users to study the TME from different perspectives. The differential analysis module helps identify potential prognostic biomarkers. The immune cell infiltration and the immune repertoire module uncover the extent of cancer heterogeneity. In combination with whole-exome sequencing data, the neoantigen expression module helps novel neoantigen discovery and vaccine designs. Within the Cancer Immune Monitoring and Analysis Centers-Cancer Immunologic Data Commons network as part of the National Cancer Institute cancer moonshot program, RIMA was built to support the consistent analysis of data from multiple studies and perform flexible cohort-level analysis for the comparison of different clinical phenotypes. Our pipeline has been used to study the effect of immunotherapy drugs on novel ICB targets and chimeric antigen receptor T therapy[21,107,108,110,111].

Our pipeline has some limitations: first, neoantigens derived from post-translational modifications and gene fusions are important for T cell response[112,113], but these tools are currently not included in RIMA. Several tools were developed to include neoantigens derived from these and other sources of variation[114]. In the future, we will expand RIMA to include gene fusion analysis to identify some of these neoantigens. Second, RIMA currently provides several gene signatures (T cell dysfunction and exclusion, INF-γ and tumor-infiltrating lymphocytes) that are associated with immunotherapy response. However, prognostic gene expression signatures are still being developed and are often cancer type specific. Examples include hematopoietic stem cell and granulocyte-macrophage progenitor genes found in acute myeloid leukemia[115] and genes involved in the mitogen-activated protein kinases kinase activity in triple-negative breast cancer[116]. To support customized gene signature evaluation, RIMA provides text results of expression matrices to support customized analysis depending on diverse cancer characteristics. The text expression output of RIMA will allow users to extract potential biomarker genes and explore expression differences in these genes between responders and non-responders. We will also continue to collect additional cancer-specific gene signatures and will update the pipeline and online guide as discoveries are made in the field.

Outlook

The advent and promise of successful immunotherapies has galvanized the cancer research community and has led to exciting insights into tumor immunity. However, it is difficult to disentangle the complex dynamics of cancer cells, the tumor microenvironment and the immune system. The inherent inter- and intra-tumor heterogeneity of tumors impedes the development of efficient clinical biomarkers, the discovery of novel immunotherapy targets and the development of new cancer drugs. The development of single-cell sequencing techniques, such as single-cell RNA-seq and single-cell sequencing assay for transposase-accessible chromatin, provides high-dimensional resolution data on the immune cell population and enables the detection of differences between individual cells and groups thereof[117–119]. Multi-omics approaches integrating genetics, transcriptomics, epigenetics and metagenomics data could also enhance our understanding of immune and tumor cell interactions[120–123]. RIMA is our immune-focused pipeline specifically developed for bulk RNA-seq data. It provides a consistent processing pipeline with which to extract meaning from immunotherapy RNA-seq datasets. As additional datasets, biological insights and assay modalities become available, we will continue to improve RIMA's analysis features and integrate its results with other data types. We also plan to update our online tutorial with the latest cancer immunotherapy knowledge and discoveries. We expect that our Gitbook tutorial will be a valuable resource for cancer immunology studies in the future.

Conclusion

The past decade has seen cancer immunotherapy transform cancer treatment. At the same time, RNA-seq has become a mature and cost-effective profiling technique and is increasingly used in cancer immunology and immunotherapy studies. RNA-seq is versatile, allowing scientists to investigate many aspects of tumor immunity such as identifying treatment response gene signatures, differentially expressed pathways, tumor immune cell infiltration, TCR/BCR repertoire features and microbiota abundance. Many computational methods have been developed to analyze RNA-seq data and characterize different aspects of tumor immunity. We anticipate that integrative and immune-specific computational analysis of tumor RNA-seq data will help advance cancer immunology and immuno-oncology research. Our online tutorial and companion RIMA pipeline will serve as a comprehensive resource to capture the latest progress in the field and provide insights into tumor immunity and immunotherapy response. We anticipate that ongoing research will lead to the discovery of progressively more immune markers and novel immune targets.

Data availability

The dataset used for testing the pipeline running time in Fig. 2 and in the RIMA online tutorial was obtained from Sequence Read Archive PRJNA482620 via ref. 109.

Code availability

The RIMA source code is available at https://github.com/liulab-dfci/RIMA_pipeline and as Supplementary Software 1. The online tutorial is available at https://liulab-dfci.github.io/RIMA/.

References

1. Waldman, A. D., Fritz, J. M. & Lenardo, M. J. A guide to cancer immunotherapy: from T cell basic science to clinical practice. *Nat. Rev. Immunol.* **20**, 651–668 (2020).
2. Hellmann, M. D. et al. Nivolumab plus ipilimumab in lung cancer with a high tumor mutational burden. *N. Engl. J. Med.* **378**, 2093–2104 (2018).
3. Hugo, W. et al. Genomic and transcriptomic features of response to anti-PD-1 therapy in metastatic melanoma. *Cell* **165**, 35–44 (2017).
4. Chan, T. A. et al. Development of tumor mutation burden as an immunotherapy biomarker: utility for the oncology clinic. *Ann. Oncol.* **30**, 44–56 (2019).
5. Łuksza, M. et al. A neoantigen fitness model predicts tumour response to checkpoint blockade immunotherapy. *Nature* **551**, 517–520 (2017).
6. Rooney, M. S., Shukla, S. A., Wu, C. J., Getz, G. & Hacohen, N. Molecular and genetic properties of tumors associated with local immune cytolytic activity. *Cell* **160**, 48–61 (2015).
7. Ayers, M. et al. IFN-γ–related mRNA profile predicts clinical response to PD-1 blockade. *J. Clin. Invest.* **127**, 2930–2940 (2017).
8. Gentles, A. J. et al. The prognostic landscape of genes and infiltrating immune cells across human cancers. *Nat. Med.* **21**, 938–945 (2015).
9. Thorsson, V. et al. The immune landscape of cancer. *Immunity* **48**, 812–830.e14 (2018).
10. Zhang, J. et al. Immune receptor repertoires in pediatric and adult acute myeloid leukemia. *Genome Med.* **11**, 73 (2019).
11. Hopkins, A. C. et al. T cell receptor repertoire features associated with survival in immunotherapy-treated pancreatic ductal adenocarcinoma. *JCI Insight* **3**, e122092 (2018).
12. Tumeh, P. C. et al. PD-1 blockade induces responses by inhibiting adaptive immune resistance. *Nature* **515**, 568–571 (2014).
13. Poore, G. D. et al. Microbiome analyses of blood and tissues suggest cancer diagnostic approach. *Nature* **579**, 567–574 (2020).
14. Gopalakrishnan, V. et al. Gut microbiome modulates response to anti-PD-1 immunotherapy in melanoma patients. *Science* **359**, 97–103 (2018).
15. Chen, H. X., Song, M., Maecker, H. T. & Gnjatic, S. Network for biomarker immunoprofiling for cancer immunotherapy: Cancer Immune Monitoring and Analysis Centers and Cancer Immunologic Data Commons (CIMAC-CIDC). *Clin. Cancer Res.* **27**, 5038–5048 (2021).
16. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
17. Wang, L., Wang, S. & Li, W. RSeQC: quality control of RNA-seq experiments. *Bioinformatics* **28**, 2184–2185 (2012).
18. Shen, S. et al. rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proc. Natl Acad. Sci. USA* **111**, E5593–E5601 (2014).
19. Halperin, R. F. et al. Improved methods for RNAseq-based alternative splicing analysis. *Sci. Rep.* **11**, 1–15 (2021).
20. Trincado, J. L. et al. SUPPA2: fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions. *Genome Biol.* **19**, 40 (2018).
21. Zeng, Z. et al. Cross-site concordance evaluation of tumor DNA and RNA sequencing platforms for the CIMAC-CIDC Network. *Clin. Cancer Res.* **27**, 5049–5061 (2021).
22. Conesa, A. et al. A survey of best practices for RNA-seq data analysis. *Genome Biol.* **17**, 13 (2016).
23. Anders, S., Pyl, P. T. & Huber, W. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).
24. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinforma.* **12**, 323 (2011).
25. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).
26. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* **14**, 417–419 (2017).
27. Zhang, C., Zhang, B., Lin, L.-L. & Zhao, S. Evaluation and comparison of computational tools for RNA-seq isoform quantification. *BMC Genomics* **18**, 583 (2017).
28. Ritchie, M. E. et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).
29. Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E. & Storey, J. D. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* **28**, 882–883 (2012).
30. Espín-Pérez, A. et al. Comparison of statistical methods and the use of quality control samples for batch effect correction in human transcriptome data. *PLoS One* **13**, e0202947 (2018).
31. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
32. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
33. Law, C. W., Chen, Y., Shi, W. & Smyth, G. K. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* **15**, R29 (2014).
34. Schurch, N. J. et al. How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? *RNA* **22**, 839–851 (2016).
35. Soneson, C., Love, M. I. & Robinson, M. D. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Res.* **4**, 1521 (2015).
36. Yu, G., Wang, L.-G., Han, Y. & He, Q.-Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* **16**, 284–287 (2012).
37. Wixon, J. & Kell, D. The Kyoto Encyclopedia of Genes and Genomes—KEGG. *Yeast* **17**, 48–55 (2000).
38. Gene Ontology Consortium. The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Res.* **49**, D325–D334 (2021).
39. Liberzon, A. et al. The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst.* **1**, 417–425 (2015).
40. Hänzelmann, S., Castelo, R. & Guinney, J. GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinforma.* **14**, 7 (2013).
41. Binnewies, M. et al. Understanding the tumor immune microenvironment (TIME) for effective therapy. *Nat. Med.* **24**, 541–550 (2018).
42. Lavin, Y. et al. Innate immune landscape in early lung adenocarcinoma by paired single-cell analyses. *Cell* **169**, 750–765.e17 (2017).

43. Li, T. et al. TIMER: a web server for comprehensive analysis of tumor-infiltrating immune cells. *Cancer Res.* **77**, e108–e110 (2017).
44. Finotello, F. et al. Molecular and pharmacological modulators of the tumor immune contexture revealed by deconvolution of RNA-seq data. *Genome Med.* **11**, 34 (2019).
45. Racle, J., de Jonge, K., Baumgaertner, P., Speiser, D. E. & Gfeller, D. Simultaneous enumeration of cancer and immune cell types from bulk tumor gene expression data. *Elife* **6**, e26476 (2017).
46. Newman, A. M. et al. Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods* **12**, 453–457 (2015).
47. Aran, D., Hu, Z. & Butte, A. J. xCell: digitally portraying the tissue cellular heterogeneity landscape. *Genome Biol.* **18**, 220 (2017).
48. Becht, E. et al. Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. *Genome Biol.* **17**, 218 (2016).
49. Sturm, G., Finotello, F. & List, M. Immunedeconv: an R package for unified access to computational methods for estimating immune cell fractions from bulk RNA-sequencing data. *Methods Mol. Biol.* **2120**, 223–232 (2020).
50. Sturm, G. et al. Comprehensive evaluation of transcriptome-based cell-type quantification methods for immuno-oncology. *Bioinformatics* **35**, 14 (2019).
51. Newman, A. M. et al. Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nat. Biotechnol.* **37**, 773–782 (2019).
52. Wang, K. et al. Deconvolving clinically relevant cellular immune cross-talk from bulk gene expression using CODEFACS and LIRICS stratifies patients with melanoma to Anti-PD-1 therapy. *Cancer Discov.* **12**, 1088–1105 (2022).
53. Ribas, A. & Wolchok, J. D. Cancer immunotherapy using checkpoint blockade. *Science* **359**, 1350–1355 (2018).
54. Jiang, P. et al. Signatures of T cell dysfunction and exclusion predict cancer immunotherapy response. *Nat. Med.* **24**, 1550–1558 (2018).
55. Cristescu, R. et al. Pan-tumor genomic biomarkers for PD-1 checkpoint blockade-based immunotherapy. *Science* **362**, eaar3593 (2018).
56. Chang, L., Chang, M., Chang, H. M. & Chang, F. Microsatellite instability: a predictive biomarker for cancer immunotherapy. *Appl. Immunohistochem. Mol. Morphol.* **26**, e15–e21 (2018).
57. Niu, B. et al. MSIsensor: microsatellite instability detection using paired tumor-normal sequence data. *Bioinformatics* **30**, 1015–1016 (2014).
58. Salipante, S. J., Scroggins, S. M., Hampel, H. L., Turner, E. H. & Pritchard, C. C. Microsatellite instability detection by next generation sequencing. *Clin. Chem.* **60**, 1192–1199 (2014).
59. Niu, B. et al. msisensor2: Microsatellite instability (MSI) detection for tumor only data. *Github* https://github.com/niu-lab/msisensor2 (2019).
60. Akira, S., Uematsu, S. & Takeuchi, O. Pathogen recognition and innate immunity. *Cell* **124**, 783–801 (2006).
61. Yam-Puc, J. C., Zhang, L., Zhang, Y. & Toellner, K.-M. Role of B-cell receptors for B-cell development and antigen-induced differentiation. *F1000Res.* **7**, 429 (2018).
62. Teraguchi, S. et al. Methods for sequence and structural analysis of B and T cell receptor repertoires. *Comput. Struct. Biotechnol. J.* **18**, 2000–2011 (2020).
63. Bolotin, D. A. et al. Antigen receptor repertoire profiling from RNA-seq data. *Nat. Biotechnol.* **35**, 908–911 (2017).
64. Li, B. et al. Landscape of tumor-infiltrating T cell repertoire of human cancers. *Nat. Genet.* **48**, 725–732 (2016).
65. Hu, X. et al. Landscape of B cell immunity and related immune evasion in human cancers. *Nat. Genet.* **51**, 560–567 (2019).
66. Song, L. et al. TRUST4: immune repertoire reconstruction from bulk and single-cell RNA-seq data. *Nat. Methods* **18**, 627–630 (2021).
67. Lefranc, M.-P. et al. IMGT®, the international ImMunoGeneTics information system®. *Nucleic Acids Res.* **37**, D1006–D1012 (2008).
68. Yost, K. E. et al. Clonal replacement of tumor-specific T cells following PD-1 blockade. *Nat. Med.* **25**, 1251–1259 (2019).
69. Selitsky, S. R. et al. Prognostic value of B cells in cutaneous melanoma. *Genome Med.* **11**, 36 (2019).
70. Xu-Monette, Z. Y. et al. Immunoglobulin somatic hypermutation has clinical impact in DLBCL and potential implications for immune checkpoint blockade and neoantigen-based immunotherapies. *J. Immunother. Cancer* **7**, 272 (2019).
71. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144**, 646–674 (2011).
72. Cibulskis, K. et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* **31**, 213–219 (2013).
73. Fan, Y. et al. MuSE: accounting for tumor heterogeneity using a sample-specific error model improves sensitivity and specificity in mutation calling from sequencing data. *Genome Biol.* **17**, 178 (2016).
74. Larson, D. E. et al. SomaticSniper: identification of somatic point mutations in whole genome sequencing data. *Bioinformatics* **28**, 311–317 (2012).
75. Koboldt, D. C. et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* **22**, 568–576 (2012).
76. Sun, Z., Bhagwate, A., Prodduturi, N., Yang, P. & Kocher, J.-P. A. Indel detection from RNA-seq data: tool evaluation and strategies for accurate detection of actionable mutations. *Brief. Bioinform.* **18**, 973–983 (2017).
77. Kaya, C. et al. Limitations of detecting genetic variants from the RNA sequencing data in tissue and fine-needle aspiration samples. *Thyroid* **31**, 589–595 (2021).
78. McLaren, W. et al. The ensembl variant effect predictor. *Genome Biol.* **17**, 122 (2016).
79. Gao, Q. et al. Driver fusions and their implications in the development and treatment of human cancers. *Cell Rep.* **23**, 227–238.e3 (2018).
80. Latysheva, N. S. & Babu, M. M. Discovering and understanding oncogenic gene fusions through data intensive computational approaches. *Nucleic Acids Res.* **44**, 4487–4503 (2016).
81. Haas, B. J. et al. Accuracy assessment of fusion transcript detection via read-mapping and de novo fusion transcript assembly-based methods. *Genome Biol.* **20**, 1–16 (2019)
82. Torres-García, W. et al. PRADA: pipeline for RNA sequencing data analysis. *Bioinformatics* **30**, 2224–2226 (2014).
83. Schumacher, T. N. & Schreiber, R. D. Neoantigens in cancer immunotherapy. *Science* **348**, 69–74 (2015).
84. Zhang, Z. et al. Neoantigen: a new breakthrough in tumor immunotherapy. *Front. Immunol.* **12**, 672356 (2021).
85. Orenbuch, R. et al. arcasHLA: high-resolution HLA typing from RNAseq. *Bioinformatics* **36**, 33–40 (2020).
86. Boegel, S. et al. HLA typing from RNA-Seq sequence reads. *Genome Med.* **4**, 102 (2012).
87. Szolek, A. et al. OptiType: precision HLA typing from next-generation sequencing data. *Bioinformatics* **30**, 3310–3316 (2014).
88. Shukla, S. A. et al. Comprehensive analysis of cancer-associated somatic mutations in class I HLA genes. *Nat. Biotechnol.* **33**, 1152–1158 (2015).
89. Peng, M. et al. Neoantigen vaccine: an emerging tumor immunotherapy. *Mol. Cancer* **18**, 128 (2019).

90. Lu, Y.-C. & Robbins, P. F. Cancer immunotherapy targeting neoantigens. *Semin. Immunol.* **28**, 22–27 (2016).
91. Howitt, B. E. et al. Association of polymerase e-mutated and microsatellite-instable endometrial cancers with neoantigen load, number of tumor-infiltrating lymphocytes, and expression of PD-1 and PD-L1. *JAMA Oncol.* **1**, 1319–1323 (2015).
92. Chang, K. et al. Immune profiling of premalignant lesions in patients with lynch syndrome. *JAMA Oncol.* **4**, 1085–1092 (2018).
93. Hundal, J. et al. pVAC-Seq: a genome-guided in silico approach to identifying tumor neoantigens. *Genome Med.* **8**, 11 (2016).
94. Jurtz, V. et al. NetMHCpan-4.0: improved peptide–MHC class I interaction predictions integrating eluted ligand and peptide binding affinity data. *J. Immunol.* **199**, 3360–3368 (2017).
95. O'Donnell, T. J. et al. MHCflurry: open-source class I MHC binding affinity prediction. *Cell Syst.* **7**, 129–132.e4 (2018).
96. Nielsen, M., Lundegaard, C. & Lund, O. Prediction of MHC class II binding affinity using SMM-align, a novel stabilization matrix alignment method. *BMC Bioinforma.* **8**, 238 (2007).
97. Vivarelli, S. et al. Gut microbiota and cancer: from pathogenesis to therapy. *Cancers (Basel)* **11**, 38 (2019).
98. Helmink, B. A., Khan, M. A. W., Hermann, A., Gopalakrishnan, V. & Wargo, J. A. The microbiome, cancer, and cancer therapy. *Nat. Med.* **25**, 377–388 (2019).
99. Andrews, M. C. et al. Gut microbiota signatures are associated with toxicity to combined CTLA-4 and PD-1 blockade. *Nat. Med.* **27**, 1432–1441 (2021).
100. Vétizou, M. et al. Anticancer immunotherapy by CTLA-4 blockade relies on the gut microbiota. *Science* **350**, 1079–1084 (2015).
101. Nejman, D. et al. The human tumor microbiome is composed of tumor type-specific intracellular bacteria. *Science* **368**, 973–980 (2020).
102. Wood, D. E. & Salzberg, S. L. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* **15**, R46 (2014).
103. Lu, J. et al. Metagenome analysis using the Kraken software suite. *Nat. Protoc.* **17**, 2815–2839 (2022).
104. Khoury, J. D. et al. Landscape of DNA virus associations across human malignant cancers: analysis of 3,775 cases using RNA-Seq. *J. Virol.* **87**, 8916–8926 (2013).
105. Walker, M. A. et al. GATK PathSeq: a customizable computational tool for the discovery and identification of microbial sequences in libraries from eukaryotic hosts. *Bioinformatics* **34**, 4287–4289 (2018).
106. Kim, D., Song, L., Breitwieser, F. P. & Salzberg, S. L. Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome Res.* **26**, 1721–1729 (2016).
107. Zeng, Z. et al. TISMO: syngeneic mouse tumor database to model tumor immunity and immunotherapy response. *Nucleic Acids Res.* **50**, D1391–D1397 (2022).
108. Schoenfeld, J. D. et al. Durvalumab plus tremelimumab alone or in combination with low-dose or hypofractionated radiotherapy in metastatic non-small-cell lung cancer refractory to previous PD(L)-1 therapy: an open-label, multicentre, randomised, phase 2 trial. *Lancet Oncol.* **23**, 279–291 (2022).
109. Zhao, J. et al. Immune and genomic correlates of response to anti-PD-1 immunotherapy in glioblastoma. *Nat. Med.* **25**, 462–469 (2019).
110. Penter, L. et al. Mechanisms of response and resistance to combination decitabine and ipilimumab for transplant naïve and post-transplant AML/MDS. *Blood* **140**, 10198–10199 (2022).
111. Penter, L. et al. Molecular and cellular features of CTLA-4 blockade for relapsed myeloid malignancies after transplantation. *Blood* **137**, 3212–3217 (2021).
112. Yang, W. et al. Immunogenic neoantigens derived from gene fusions stimulate T cell responses. *Nat. Med.* **25**, 767–775 (2019).
113. Hsu, J.-M., Li, C.-W., Lai, Y.-J. & Hung, M.-C. Posttranslational modifications of PD-L1 and their applications in cancer therapy. *Cancer Res.* **78**, 6349–6353 (2018).
114. Gopanenko, A. V., Kosobokova, E. N. & Kosorukov, V. S. Main strategies for the identification of neoantigens. *Cancers (Basel)* **12**, 2879 (2020).
115. van Galen, P. et al. Single-cell RNA-seq reveals AML hierarchies relevant to disease progression and immunity. *Cell* **176**, 1265–1281.e24 (2019).
116. Loi, S. et al. RAS/MAPK activation is associated with reduced tumor-infiltrating lymphocytes in triple-negative breast cancer: therapeutic cooperation between MEK and PD-1/PD-L1 immune checkpoint inhibitors. *Clin. Cancer Res.* **22**, 1499–1509 (2016).
117. Ranzoni, A. M. et al. Integrative single-cell RNA-seq and ATAC-seq analysis of human developmental hematopoiesis. *Cell Stem Cell* **28**, 472–487.e7 (2021).
118. Muto, Y. et al. Single cell transcriptional and chromatin accessibility profiling redefine cellular heterogeneity in the adult human kidney. *Nat. Commun.* **12**, 1–17 (2021).
119. Grosselin, K. et al. High-throughput single-cell ChIP-seq identifies heterogeneity of chromatin states in breast cancer. *Nat. Genet.* **51**, 1060–1066 (2019).
120. Menyhárt, O. & Győrffy, B. Multi-omics approaches in cancer research with applications in tumor subtyping, prognosis, and diagnosis. *Comput. Struct. Biotechnol. J.* **19**, 949–960 (2021).
121. Leng, D. et al. A benchmark study of deep learning-based multi-omics data fusion methods for cancer. *Genome Biol.* **23**, 171 (2022).
122. Li, B. et al. Fresh tissue multi-omics profiling reveals immune classification and suggests immunotherapy candidates for conventional chondrosarcoma. *Clin. Cancer Res.* **27**, 6543–6558 (2021).
123. Yang, Y. et al. A multi-omics-based serial deep learning approach to predict clinical outcomes of single-agent anti-PD-1/PD-L1 immunotherapy in advanced stage non-small-cell lung cancer. *Am. J. Transl. Res.* **13**, 743–756 (2021).

Acknowledgements

We acknowledge members of the Shirley Liu laboratory, the Center for Functional Cancer Epigenetics and the Cancer Immune Data Commons for their helpful suggestions and support during pipeline development. Support for this project is provided by the PACT and made possible through funding provided to the FNIH by AbbVie Inc., Amgen Inc., Boehringer-Ingelheim Pharma GmbH & Co. KG, Bristol-Myers Squibb, Celgene Corporation, Genentech Inc., Gilead, GlaxoSmithKline plc, Janssen Pharmaceutical Companies of Johnson & Johnson, Novartis Institutes for Biomedical Research, Pfizer Inc. and Sanofi.

Author contributions

L.Y., J.W., A.J., S.B., C.J.W. and Y.L. developed and optimized the RIMA pipeline. Y.L., J.W., L.Y. and J.A. drafted the manuscript. Y.L., J.A. and L.Y. drafted the online tutorial. L.S. provided suggestions for immune repertoire analysis. J.F. provided suggestions for immune response analysis. L.T., A.S., C.T., Y.Z., Z.Z., G.B., M.T., X.Q. and H.W.L. participated in conceptualization and project discussion. Y.L., F.M. and X.S.L. supervised the project. All authors read and approved the final manuscript.

Competing interests

X.S.L. conducted the work while being on the faculty at the Dana-Farber Cancer Institute and is currently a board member and CEO of GV20 Therapeutics. F.M. is a cofounder of and has equity in Harbinger Health, has equity in Zephyr AI and serves as a consultant for Harbinger Health, Zephyr AI and Red Cell Partners. F.M. declares that none of these relationships are directly or indirectly related to the content of this manuscript. All other authors do not have any conflicts.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41596-023-00841-8.

Correspondence and requests for materials should be addressed to Yang Liu or X. Shirley Liu.

Peer review information *Nature Protocols* thanks Zlatko Trajanoski, Zemin Zhang and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.