COMPUTATIONAL TOOLS

Computational genomics tools for dissecting tumour–immune cell interactions

Hubert Hackl, Pornpimol Charoentong*, Francesca Finotello* and Zlatko Trajanoski*

Abstract | Recent breakthroughs in cancer immunotherapy and decreasing costs of high-throughput technologies have sparked intensive research into tumour–immune cell interactions using genomic tools. The wealth of the generated data and the added complexity pose considerable challenges and require computational tools to process, to analyse and to visualize the data. Recently, various tools have been developed and used to mine tumour immunologic and genomic data effectively and to provide novel mechanistic insights. Here, we review computational genomics tools for cancer immunology and provide information on the requirements and functionality in order to assist in the selection of tools and assembly of analytical pipelines.

Precision oncology
The use of systematic assessment of cancer genomic information for personalized diagnosis and therapy.

Cancer immunotherapy
Activation of the immune system to specifically target and kill cancer cells using checkpoint blockers, therapeutic vaccines or engineered T cells.

Antigens
Short peptides that are produced from digested proteins and presented on the surface on the cell by the major histocompatibility complex or the human leukocyte antigen.

Division of Bioinformatics, Biocenter, Medical University of Innsbruck, Innrain 80–82, 6020 Innsbruck, Austria.

Correspondence to Z.T.
zlatko.trajanoski@i-med.ac.at

**These authors contributed equally to this work.*

doi:10.1038/nrg.2016.67
Published online 4 Jul 2016

In the past few decades, a large proportion of the cancer research efforts in academia and industry has been directed towards the development of targeted agents, and a number of alterations that affect protein-coding genes in tumours were identified, which subsequently led to the development of numerous anticancer drugs approved by the regulatory agencies worldwide. With the advent of next-generation sequencing (NGS) technologies, it is now possible to identify alterations at the base-pair resolution of all cancer genes in individual samples and in large cohorts (for example, the International Cancer Genome Consortium (ICGC)[1] and The Cancer Genome Atlas (TCGA), which first published data for glioblastoma in 2008)[2]. Driven by these technological advances, many centres are currently implementing precision oncology programmes that aim to use genomics approaches to inform cancer therapy. However, whereas integration of cancer genomic data into clinical oncology disease management is ongoing, drug resistance remains a major issue. For 71 drugs approved by the US Food and Drug Administration (FDA) from 2002 to 2014 for metastatic, advanced or refractory solid cancers, the median gains in progression-free and overall survival were 2.5 and 2.1 months, respectively[3]. Moreover, many cancer mutations are not druggable by the available targeted agents and therefore limit a broader applicability of precision oncology.

In contrast to other systemic therapies, cancer immunotherapy has the potential to adapt itself to changes of the tumour because specific T cells can emerge and kill tumour clones that evolved and changed surface antigens[4]. However, tumour cells can escape detection by the immune system by upregulating immune checkpoint molecules — that is, immunological 'brakes' — on the cell surface of immune cells, such as cytotoxic T-lymphocyte associated protein 4 (CTLA4) or programmed cell death 1 (PD1; also known as PDCD1)[5]. Recently, several antibodies that block immune checkpoints and thereby enhance antitumour T cell response have been introduced and show remarkable clinical effects. Analysis of long-term data of patients who received CTLA4-targeted antibodies in unresectable or metastatic melanoma[6] shows a plateau in the survival curve after 3 years, suggesting durable benefits or even curative potential. Furthermore, the efficacy of PD1-targeted antibodies has been shown not only in melanoma, but also in nine different tumour types as diverse as non-small cell lung cancer, liver cancer, kidney cancer and lymphoma[7]. We are currently witnessing a rapid pace of development of checkpoint blockers, as evident from the more than 150 clinical trials into their use as monotherapies or combination therapies[7]. However, only a fraction of the patients is responsive to monotherapies with checkpoint blockers, and the identification of the precise mode of action and predictive markers is a subject of intense research.

Following the first cancer immunotherapy that transformed cancer care — that is, the use of monoclonal antibodies — and the development of checkpoint blocker immunotherapy, as well as other immunotherapeutic strategies, including therapeutic vaccines and engineered T cells[8] (BOX 1), tumour–immune cell interactions came into focus. The dissection of these complex interactions

Immune checkpoint
An inhibitory pathway of the immune system, commonly a ligand–receptor pair, that maintains self-tolerance and modulates immune responses in peripheral tissues in order to minimize collateral tissue damage.

Checkpoint blockers
Antibodies that target immune checkpoint molecules to activate the immune system.

holds promise to lead to the identification of predictive biomarkers, to the development of novel drugs or therapeutic strategies, as well as to provide novel mechanistic insights. However, the investigation of cancer–immune cell interactions poses considerable challenges owing to the evolving and heterogeneous nature of these two multicellular ecosystems: the development of cancer, which can be seen as an evolutionary process; and the immune system, which has numerous innate and adaptive immune cell subpopulations, some of which show phenotypic plasticity and possess memory. NGS techniques and other medium- to high-throughput technologies are generating a wealth of data and require information systems to process and analyse data, to extract information to develop mechanistic theories and to support clinical decision making. Thus, cancer immunogenomics can also be seen as information science and, as such, will pave the way for the development and successful application of novel immunotherapeutic strategies.

In this Review, we first give a brief overview of the tumour–immune cell interactions and then discuss computational genomics tools for mining cancer genomic data and extracting immunological parameters. We focus on higher-level analyses of NGS data, including quantification of tumour-infiltrating lymphocytes (TILs), identification of tumour antigens and profiling of T cell receptors (TCRs), and provide information on the requirements and functionality in order to assist in

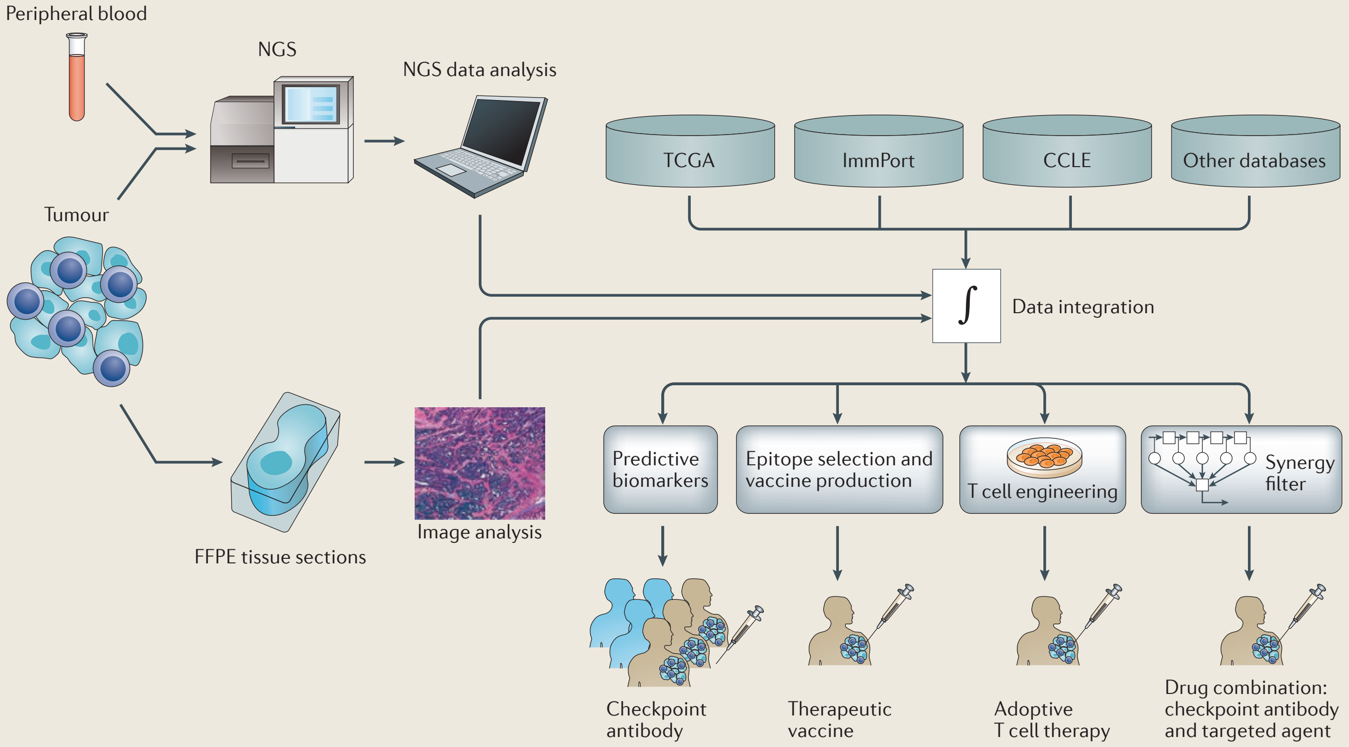
Box 1 | **Cancer immunotherapy and precision oncology**

Cancer immunotherapy is based on agents that induce or augment immune responses to cancer. Currently, besides monoclonal antibodies that target cancer cells, monotherapy is based on three strategies: use of checkpoint blockers, vaccination with neoantigens and adoptive T cell transfer (see the figure). Additionally, combinations of immune monotherapies as well as combinations of immunotherapies and targeted therapies are being investigated.

Several antibodies against the checkpoint molecules cytotoxic T-lymphocyte associated protein 4 (CTLA4), programmed cell death 1 (PD1) or PD1 ligand 1 (PDL1) have been approved or are in late-stage clinical trials and show benefit and even durable response. Still, only a fraction of the patients is responsive, and the quest for predictive markers and the investigation of the mechanisms that underlie resistance to therapy with these checkpoint blockers is an active area of research. Vaccination strategies based on molecular profiling of the tumour using next-generation sequencing (NGS) technologies, and production of a peptide- or RNA-based vaccine are considered as personalized approaches and are currently evaluated in a number of clinical trials. Adoptive cell transfer with engineered T cells for the treatment of solid cancers is another personalized approach, and it is expected that with the development of novel genome-editing technologies[121] and with the gained knowledge on the tumour–immune cell interaction, this strategy will be routinely used in the near future.

An infrastructure for precision oncology based on vaccination strategy or adoptive cell transfer with engineered T cells requires state-of-the-art molecular tools, such as NGS instruments and dedicated IT infrastructure, as depicted in the figure. Both hardware and software need to be designed and tested for their capability to handle large amounts of data. Some of the components are rather mature, such as analytical pipelines for NGS data and software for analyses of images. Other components are in their early stages of development and require extensive conceptual and technical work. For example, integrative data analyses still represent a major challenge and require novel theoretical studies. Similarly, the identification of the right combination of drugs with synergistic properties (for example, a checkpoint blocker with a targeted agent) is in its infancy and additional experimental and computational research will be necessary to develop a robust component.

CCLE, Cancer Cell Line Encyclopedia; FFPE, formalin-fixed, paraffin-embedded; TCGA, The Cancer Genome Atlas.



Therapeutic vaccines
Cancer treatment or therapeutic vaccines use specific antigens to boost the immune system's ability to recognize and to destroy cancer cells.

Engineered T cells
Genetically modified T cells (for example, by expressing a chimeric antigen receptor) that are designed to recognize particular tumour antigens as non-self and lead to tumour destruction.

Tumour-infiltrating lymphocytes
(TILs). Subpopulations of the immune system infiltrating the tumour.

T cell receptors
(TCRs). Proteins that consist of an α-chain and a β-chain on T lymphocytes (T cells), which recognize fragments of antigens bound to the major histocompatibility complex.

Neoantigens
Acquired somatic mutations in the cancer genome that lead to new antigens recognized by the immune system.

Major histocompatibility complex
(MHC). Protein complex that presents antigens on the cell surface. In humans, the MHC is encoded by the human leukocyte antigen (HLA) gene locus.

Human leukocyte antigens
(HLAs). Loci of genes that encode for proteins on the surface of cells and present antigens from inside (class I) and outside (class II) of the cell to T lymphocytes. HLA is the human form of the major histocompatibility complex.

Microsatellite unstable
Pertains to microsatellite instability, which involves changes in the number of repeats of microsatellites (that is, short, repeated DNA sequences) in certain cell types, such as cancer cells, relative to that of inherited DNA. Microsatellite instability is often produced by an impaired DNA mismatch repair system.

the selection of tools and assembly of analytical pipelines. Although the focus here is on cancer immunology, the computational approaches discussed provide the means also to study other diseases, such as autoimmune, inflammatory, infectious or graft-versus-host diseases.

Tumour–immune cell interactions

The tumour–immune cell interactions can be conceptualized as a series of events referred to as the cancer–immunity cycle[9] (FIG. 1a). The first step is the generation of neoantigens: that is, peptides that result from somatic mutations[8]. As the number of somatic mutations can range from a few dozen to several tens of thousands in an individual tumour, the resulting neoantigens render tumour cells highly heterogeneous at the molecular level. The neoantigens are presented at the surface of antigen-presenting cells (APCs) by highly variant major histocompatibility complex (MHC) alleles — which, in humans, are called human leukocyte antigens (HLAs) — adding another layer of molecular heterogeneity. The release of neoantigens following cancer cell death initiates various processes, leading to the expansion of molecularly heterogeneous T cells. These T cells recognize cancer cells by distinct TCRs through the interaction with neoantigen–MHC complexes.

In addition to the molecular heterogeneity, the intratumoural immune infiltrates are characterized by great cellular heterogeneity composed of diverse cell subpopulations related to innate and adaptive immunity (FIG. 1b), the distribution of which varies between and within tumour types. For example, tumours with high mutational rate and hence large neoantigen load, such as melanoma or microsatellite unstable colorectal cancers, are heavily infiltrated with many immune subpopulations. By contrast, microsatellite-stable tumours, which have low mutational load, show markedly reduced infiltration levels[10]. As a result of the immune infiltrates, sculpting of the tumour phenotypes during progression occurs in a process described as cancer immunoediting[11]. These molecular and cellular heterogeneities are building a complex network of tumour–immune cell interactions and pose considerable challenges to scientists aiming to disentangle the network and identify mechanisms of antitumour immunity. It is evident that comprehensive molecular characterization of the tumour–immune cell interaction requires genomics tools.

Large-scale cancer genomics projects such as TCGA and the ICGC are providing a wealth of information, particularly through charting tumour mutational landscapes, which now comprise more than 30 million mutations from more than 26,000 cancer samples. This information needs to be complemented by immunogenomics data. Although immunogenomics data from large cohorts is missing, it is clear that immunogenomics in general, and cancer immunogenomics in particular, will be one of the big data resources in the life sciences for several reasons. First, the adaptive immune system is the largest source of human genetic variation. The HLA locus contributes to more than half of the four to five million single-nucleotide polymorphisms (SNPs) in each individual genome[12]. Second, there are more than 200 immune cell types and more than 300 immune cell state transitions[13]. A database with expression profiles for the effects of more than 1,300 drugs on each cell type and each cell state will result in a resource with 0.1 petabytes of normalized RNA sequencing (RNA-seq) data. And third, the number of possible TCR–neoantigen–MHC complexes is vast. Considering the number of different TCR αβ-pairs (that is, the combinations of α- and β-subunits that make up the heterodimeric TCR, which is estimated to be around 10^{20}), the number of different antigens that can be bound by the TCR receptor (2.5×10^{10}–1.6×10^{18}), as well as the number of HLA alleles (>13,000), it becomes clear that the development of cancer immunogenomics will be considerably driven by the development of novel computational tools and our ability to effectively analyse these data.

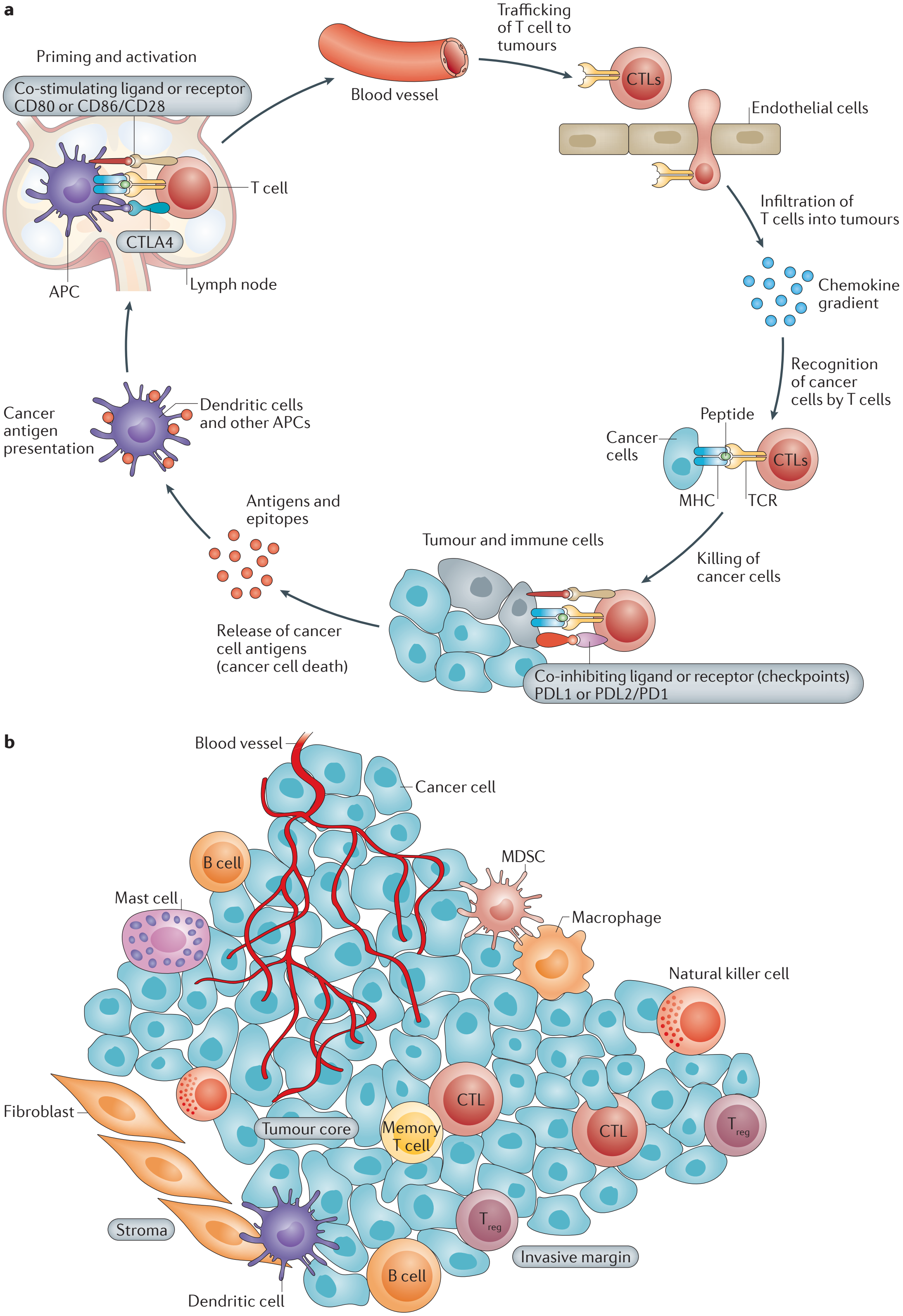
Overview of omics data analysis

The application of NGS techniques for genome, transcriptome or epigenome profiling is building the main source of data for cancer immunogenomics. Additionally, recent advances have been made in imaging technologies and associated software tools, as well as in cellular phenotyping techniques, which are enabling the generation of data types that are complementary to the genomic type (BOX 2). For most questions addressed in cancer immunogenomics, the same NGS techniques as those successfully used in cancer genomics can be applied, and they include whole-exome sequencing (WES), whole-genome sequencing (WGS), RNA-seq and bisulfite sequencing for DNA methylation (FIG. 2). However, for specific applications, such as TCR sequencing, careful consideration of the read length, the depth and the type of sequencing data (that is, WES, WGS or RNA-seq) is required. Notably, in most of the omics studies, analyses are carried out on bulk tumour samples that consist of heterogeneous cell types. Although this type of sample impurity can have confounding effects on some analyses, the molecular signatures from non-tumour cells can be leveraged to understand the interaction of the tumours with their microenvironment, including immune cells.

Alternatively, single-cell characterization based on sorting and subsequent analysis can be carried out. In the past few years, progress has been made not only in the development of sequencing technologies, but also in the barcoding and microfluidics in the area of single-cell isolation, nucleic acid amplification and transcriptome profiling[14]. This has led to novel single-cell sequencing technologies[15–18], which will have a growing role in the future, in particular for the phenotyping of immune cells[19]. For example, using these technologies, it is now possible to reconstruct pairs of TCR αβ-chains as well as transcriptional profiles from single-cell RNA-seq data[20]. Although these technologies are opening new frontiers by dissecting the contribution of individual immune cells or cancer cells, technological challenges[21], sequencing costs and practical restrictions are limiting their use. Moreover, as bulk tissue is used for diagnostic purposes in routine clinical settings, it is likely that single-cell techniques will be complementary to the analysis of bulk tissue.

Cancer immunoediting
A concept that describes the complex interactions that occur between a developing tumour and the immune system, in which immune cells not only protect the host, but also sculpt or edit the immunogenicity of the tumour.

The analysis of omics data in the context of cancer immunology can be viewed as a two-step procedure (FIG. 2). Following pre-processing of the raw data, which includes evaluation of the quality of data and removal of artefacts[22], the first step is the genomic analysis of omics data, focusing primarily on the tumour itself. A plethora of computational tools for analysing cancer genomes have been developed and are being continuously improved[23]. This step includes tools for the identification of SNPs, small insertions and deletions (indels),



Gene set enrichment analysis
(GSEA). A computational method to identify whether a predefined set of genes shows statistically significant concordant differences between two biological states (for example, phenotypes) based on gene expression profiling.

Deconvolution
A computational method to discern and quantitate individual components based on bulk measurements of a mixture (for example, gene expression measurement of a complex tumour sample).

Chemokine
A family of small secreted cytokines, the gradient of which causes immune cells with the respective receptors to migrate. This process is known as chemotaxis and is important for guiding the activated immune cells to the tumour site.

Inverse problem
A mathematical problem in which the cause is deduced based on the observed effects of a system.

copy-number variations (CNVs), structural variants and gene fusions, as well as variant annotation and interpretation[22,23]. Another set of tools in the group of genomic analyses are used to analyse the expression of genes assessed using RNA-seq, to estimate tumour heterogeneity from WES and/or SNP array data[24,25] or to analyse DNA methylation patterns[26]. The second type of analyses uses immunogenomic tools and focuses more closely on tumour–immune cell interactions. As input data, they use the output of the genomic analyses and/or raw sequencing data. The results of these immunogenomic analyses provide information on the two crucial characteristics of the tumour microenvironment: composition and functional orientation of infiltrated immune cells and origin and quantity of the tumour antigens. In the following paragraphs, we therefore focus on these two aspects: determining cellular composition of immune infiltrates in tumours and identification of tumour antigens. Additionally, as several techniques to specify T cell reactivity are emerging, we also discuss tools for TCR profiling.

Cellular characterization of immune infiltrates

As different types of TILs have different effects on tumour progression[27], the determination of the cellular composition of immune infiltrates in tumours provides not only prognostic information[28], but can also lead to the development of predictive markers and novel therapeutic strategies. Imaging and cellular phenotyping techniques are widely used and can provide partial information about the immune contexture (BOX 2). Although automated image analysis is increasingly being used, the inherent limitations of the cellular phenotyping techniques hinder the characterization of a larger number of TIL subpopulations. Thus, computational genomic tools were developed to provide a comprehensive picture of the TILs. The computational genomics tools applied for this purpose can be grouped into gene set enrichment analysis (GSEA) and deconvolution methods (FIG. 3a). It is noteworthy that both GSEA and deconvolution methods rely on a matrix of expression profiles for individual cell populations. The TIL subpopulations reconstructed with these methods include the immune subpopulations defined in the reference matrix of expression profiles.

Enrichment methods rely on gene set analysis techniques based on comparison between samples[29,30] or single-sample approaches[31]. GSEA evaluates ranked gene lists for statistical enrichment of genes involved in defined pathways and cellular processes[32]. In the comparative approach, genes are ranked based on differential expression between two biological states. Alternatively, a single-sample GSEA (ssGSEA) enrichment score can be used, which represents the degree to which genes in a particular gene set are coordinately up- or downregulated within a single sample[31]. GSEA can be used for interpreting gene expression data that has been obtained from either microarrays[29]or RNA-seq[33].

The advantage of GSEA is that it can easily be applied using existing tools, and there are no extra sample size requirements compared with classical gene expression analysis[30]. The necessary requirement of GSEA is the assembly of gene signatures related to specific immune subpopulations (FIG. 3b). In a seminal study a set of immune signatures were defined from whole-blood microarray expression data from immune and non-immune cells[34]. More recently, a gene set collection for immunological signatures (ImmuneSigDB)[35] from the Human Immunology Project was included in the Molecular Signatures Database (MSigDB)[36]. A compendium of about 5,000 well-annotated signatures was generated by analysing 389 published studies of cell states and perturbations in the mouse and human immune systems[35]. In the context of cancer immunology, a recent study analysed the TCGA data for colorectal cancer, and using GSEA provided the first high-resolution view on the immune phenotypes[10].

Deconvolution methods use expression signature matrices to infer specific cell proportions from expression data from cell mixtures and an algorithm to solve the inverse problem (FIG. 3c). Deconvolution of cell proportions from whole-blood gene expression microarray data was first introduced using a heuristic algorithm based on standard linear regression[34]. Later, an R package for deconvolution of heterogeneous tissues was developed that uses quadratic programming[37]. This package, which is called DeconRNASeq, can handle RNA-seq data, but it has been validated only on mixtures with few cell types. Several other methods have been developed that use diverse techniques for solving the ill-conditioned inverse problem (TABLE 1). Recently, a computational approach for inferring leukocyte subtypes from microarray data from bulk tumours (called CIBERSORT) was introduced[38]. CIBERSORT uses a signature expression matrix for 22 leukocyte subpopulations and implements linear support vector regression[38]. Despite the various successful applications of computational methodologies, several issues remain to be improved[30]. First, a reference matrix with gene expression profiles from sorted immune cell subpopulations from blood samples or preferably from tumour samples using RNA-seq is required. It will also be necessary to

◀ Figure 1 | **Tumour immunity at a glance.** Cancer immunity cycle and the immune contexture of the tumour. **a** | The cancer immunity cycle comprises several consecutive steps: neoantigens generated by the cancer cells are released after the cancer cell death and captured by dendritic cells. Next, dendritic cells present the captured antigens on the major histocompatibility complex (MHC) molecules to T cells, resulting in the priming and activation of effector T cell responses against the cancer-specific antigens. Guided by a chemokine gradient the activated T cells traffic to and infiltrate the tumour site. T cells specifically recognize and bind to cancer cells by the interaction between the T cell receptor (TCR) and the neoantigen–MHC complex and kill the cancer cells (cytolytic activity). Various molecular and genomics tools are available to assess each phase of these cancer immune cell interactions and their stimulating or inhibiting factors. **b** | Tumours are frequently infiltrated by immune cell types of the adaptive immune system, including B cells, cytotoxic T lymphocytes (CTLs), memory T cells, helper T cells and regulatory T cells (T_{reg}). Additionally, tumours are infiltrated by cells of the innate immune system, including macrophages, dendritic cells, mast cells, natural killer cells and myeloid-derived suppressor cells (MDSCs). APC, antigen-presenting cell; CD80, T-lymphocyte activation antigen CD80; CTLA4, cytotoxic T lymphocyte-associated protein 4; PD1, programmed cell death 1; PDL1, PD1 ligand 1. Part **a** is adapted with permission from REF. 9, Elsevier. Part **b** is from REF. 27, Nature Publishing Group.

Box 2 | Imaging technologies and cellular phenotyping

Tumour tissues are not merely a collection of malignant cells but rather a mixture of different cell types that build a complex and dynamic network[122]. Several population- and single-cell-based methods have emerged to determine the identity, the activation status and the localization of cells, including immune cell subpopulations. Commonly, formalin-fixed, paraffin-embedded (FFPE) tissue sections are processed, and thin slices are deposited on slides and prepared for haematoxylin and eosin (H&E) staining and immunohistochemistry (IHC). Tissue microarrays[123] can be used to multiplex samples onto a single slide and to study immune infiltrates in parallel in a larger number of samples. Conventional IHC or immunofluorescence methodologies and software for image analysis are frequently used to quantify several markers. For a larger number of markers, mass spectrometry can be combined with imaging techniques, as is the case with matrix-assisted laser desorption–ionization mass spectrometry imaging (MALDI MSI). Recently, imaging mass cytometry was introduced, which couples high-resolution (1 μm) laser ablation with mass spectrometry, as demonstrated on breast cancer FFPE samples for 32 proteins[124].

Flow cytometry, specifically fluorescence-activated cell sorting (FACS), which was developed 50 years ago, became a standard technology in discerning cell populations and paved the way for systems immunology. Flow cytometry can be used not only for the analyses of cell phenotypes with surface makers, but also for the functional analysis, immune responses and reactivity of the tumour-infiltrating lymphocytes (TILs). In flow cytometry, the effect of the overlapping emission spectra of the used fluorophores limits the number of possible channels. The use of stable isotopes of non-biological rare earth metals led to the introduction of mass cytometry and facilitated the extension of up to 40 channels. A recent development of this single-cell method represents cytometry time-of-flight (CyTOF) technology[125], in which metal-isotope-conjugated antibodies are used to stain the cell and are then subjected to a quadrupole time-of-flight mass spectrometer. For each cell, the mass spectrum gives quantitative information of all labelled markers of interest.

As the number of parameters analysed with these techniques is increasing, methods for the visualization of high-dimensional data are required. Several computational tools, such as SPADE[126] and viSNE[127], have been developed to reveal cell subpopulations from high-dimensional single-cell data, such as those produced by mass cytometry[127]. SPADE is complementary to existing approaches for analysing cytometry data by enabling multiple cell types to be visualized in a branched tree structure without requiring the user to define a known cellular ordering[126]. The aim of viSNE is to reduce high-parameter biological data to two dimensions. Recently, a workflow for high-dimensional mass cytometry data was suggested in which SPADE, viSNE and heatmaps are used sequentially to comprehensively characterize and compare healthy and malignant human tissue samples[128]. For identifying subpopulations in high-dimensional single-cell data, PhenoGraph[129] provides a computationally efficient graph-based method.

test whether a universal matrix can be used or whether matrices for specific tumour entities are required. Second, as deconvolution is sensitive to noise, the development and implementation of robust algorithms is necessary. And third, validation of the methods using independent methods such as fluorescence-activated cell sorting (FACS) or immunohistochemistry needs to be carried out.

Like deconvolution methods based on gene expression profiles, cell-lineage-specific DNA methylation patterns can be used to detect and to quantify leukocyte subsets[39]. For that purpose a number of methods and tools[39–41] were developed using information from a few methylated CpG loci to genome-wide loci using microarray platforms (that is, Illumina Infinium 27k and 450k DNA methylation arrays). The application of this approach to data from bisulfite sequencing technologies is straightforward. The epigenome can be highly variable across different cell types as evident from epigenome-wide association studies[42]. These cell-type-specific effects can be useful for the quantification of TILs and has been already harnessed by recent approaches[43]. Although the current availability of the reference methylation patterns from purified cell types is still limited, these methods hold promise for determining TIL composition from bulk tumour tissue.

Identification of tumour antigens

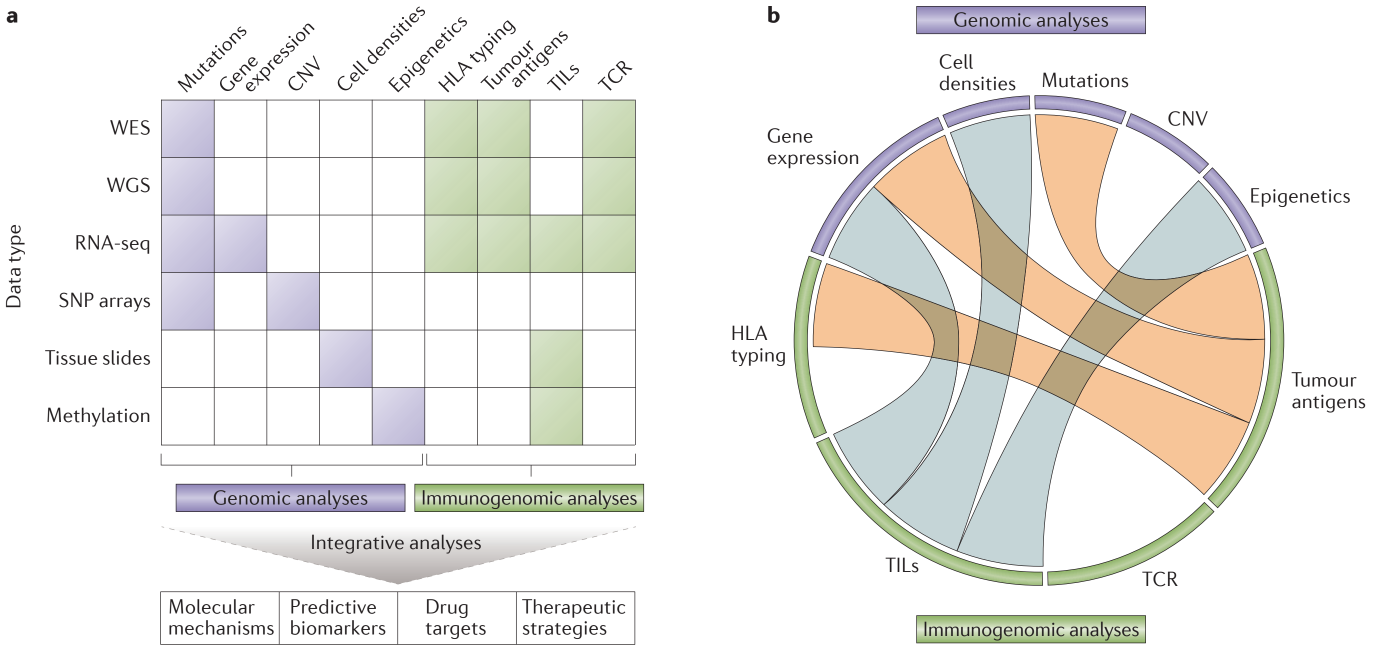
T cells are able to reject tumours on recognition of tumour-specific antigens bound to the MHC molecules of tumour cells. Antigens with high tumoural specificity — that is, displayed by tumour cells but not by normal cells — have the potential to elicit a tumour-specific immune response, minimizing the risk of adverse side effects and are therefore of great interest for cancer immunotherapies such as engineered T cells and therapeutic vaccines[8,44] (BOX 1). Three classes of antigens have high tumoural specificity: first are viral antigens, which are derived from viral genes expressed in virus-infected tumour cells; second are cancer germline antigens (CGAs), which were previously called cancer testis antigens and are proteins that are normally expressed only by trophoblasts and germline cells but that have aberrant expression in tumour cells; and third are neoantigens, which are peptides that arise from the expression of somatically mutated genes. Since the discovery of the first CGA[45], melanoma antigen 1 (MAGE1; also known as MAGEA1), a large panel of cancer germline genes expressed in several tumour types has been identified[44]. To date, the Cancer-Testis database represents the most updated and curated resource for CGAs and contains information about CGAs, their expression in tumour and normal tissues and the induced immune responses[46]. With the available list of CGAs, the extraction of the expression levels from RNA-seq data from tumour and normal samples is straightforward.

Neoantigens can be considered to be strictly tumour-specific because they originate from the expression of mutated genes that are present in malignant cells but not in the normal genome. To elicit an immune response, the mutated proteins must be proteolytically processed into short peptides and then bound to MHC molecules, to be presented to T cells (FIG. 4). When NGS data are available from matched tumour and normal samples, neoantigens can be predicted *in silico* by integrating three computational tasks (FIG. 4): the identification of mutated proteins from matched tumour–normal samples, followed by HLA typing and then prediction of neoantigen–MHC binding affinity.

Identification of mutated proteins. Mutated proteins can be estimated from NGS data of tumour tissue and matched normal sample taken from blood or from adjacent healthy tissue, through a two-step procedure. Following NGS data pre-processing consisting of quality control, read pre-processing and mapping[47], somatic DNA mutations can be identified using tools for variant detection. Then, software for variant annotation can be used to predict the affected isoform and the effect at protein level. As these tools were recently reviewed[23], we provide here only a brief overview of

High-dimensional data
Data with a few dozen to thousands of dimensions that are typically generated when each sample of an experiment or a large cohort is studied by high-throughput genomics or proteomics technologies or when many cells are studied in parallel: that is, using single-cell technologies.

Cancer germline antigens
(CGAs). Proteins that are normally expressed only by trophoblasts and germline cells but that are aberrantly expressed in cancer and recognized by the immune system. Formerly, they were often termed cancer testis antigens.



Data type	Mutations	Gene expression	CNV	Cell densities	Epigenetics	HLA typing	Tumour antigens	TILs	TCR
WES	🟪					🟩	🟩		🟩
WGS	🟪					🟩	🟩		🟩
RNA-seq	🟪	🟪				🟩	🟩	🟩	🟩
SNP arrays	🟪		🟪						
Tissue slides				🟪				🟩	
Methylation					🟪			🟩	

Figure 2 | **Computational tools for genomic and immunogenomic analyses. a** | Integrative analyses of omics data bring together genomic and immunogenomic analyses to reveal the underlying mechanisms of tumour–immune cell interactions and to identify molecular mechanisms, predictive biomarkers, drug targets and novel therapeutic strategies for cancer. The grid provides an overview of how different types of omics data (rows) are used by computational tools for individual genomic or immunogenomics analyses (columns). **b** | Immunogenomic analyses and tools are based on genomic analysis of omics data and can be divided into human leukocyte antigen (HLA) typing, quantification of tumour-infiltrating lymphocytes (TILs), identification of T cell receptors and prediction of tumour antigens. For instance, the prediction of tumour antigens requires the integration of the analysis of gene expression, mutations and HLA typing. CNV, copy-number variation; RNA-seq, RNA sequencing; SNP, single-nucleotide polymorphism; TCR, T cell receptor; WES, whole-exome sequencing; WGS, whole-genome sequencing.

selected tools. Among the tools for variant detection, the Genome Analysis Toolkit (GATK)[48] is one of the best documented and most developed pipelines and is applicable to WES, WGS and RNA-seq data. Another tool, MuTect[49], identifies SNPs with high accuracy and sensitivity by leveraging a Bayesian classifier that guarantees high specificity even in the case of variants with low allele frequency[50]. Finally, EBCall[51] uses prior knowledge about sequencing errors derived from a set of non-paired normal samples to discriminate better between true variants and sequencing errors. Despite the achieved progress, there is still room for improvement, and the next efforts should be directed towards the reduction of false positives[23]. To predict the functional impact of a genetic variant on the affected protein, annotation tools rely on publicly available repositories of gene, transcript and protein sequences, such as Ensembl[52], RefSeq[53] and Uniprot[54]. As the same variant can have different functional consequences on different transcripts, the use of different databases and strategies for prioritizing isoforms can produce largely different results. McCarthy and colleagues[55] have demonstrated that both the choice of algorithms and annotation databases have a strong impact on the final predictions. We expect that in the near future benchmark studies will be carried out and will both characterize the available tools and guide the standardization of analytical pipelines.

HLA typing. The HLA locus, which is located on chromosome 6, harbours more than two hundred genes and pseudogenes and is among the most polymorphic regions of the human genome[12]. The international immunogenetics project HLA (IMGT/HLA) database is a curated and constantly updated collection of genomic and coding DNA sequences and currently contains more than 13,000 annotated HLA alleles (release 3.22, 2015–10)[56]. The standard nomenclature of HLA alleles uses the gene name (for example, HLA-A, HLA-B or HLA-C), followed by an asterisk (*) and four sets of digits, separated by a colon (for example, HLA-A*02:01:01:05). The first set of digits defines groups of HLA alleles with similar serological specificity. The second and the third set of digits identify non-synonymous or synonymous nucleotide substitutions at the DNA level, respectively. Finally, differences in the introns or in the 3′/5′ untranslated regions are encoded in the fourth set of digits. For example, HLA alleles that are equal at two-digit resolution but not at four-digit resolution (for example, HLA-A*02:02 and HLA-A*02:01) have similar serological specificity for a peptide, but have a different protein sequence, resulting in different T cell recognition of the peptide–MHC (pMHC) complex.

Standard HLA typing is carried out using serology, polymerase chain reaction (PCR)-based methods or targeted sequencing. The increasing use of NGS in

Allele frequency
Measure of the relative frequency of an allele at a particular genetic locus in a population.

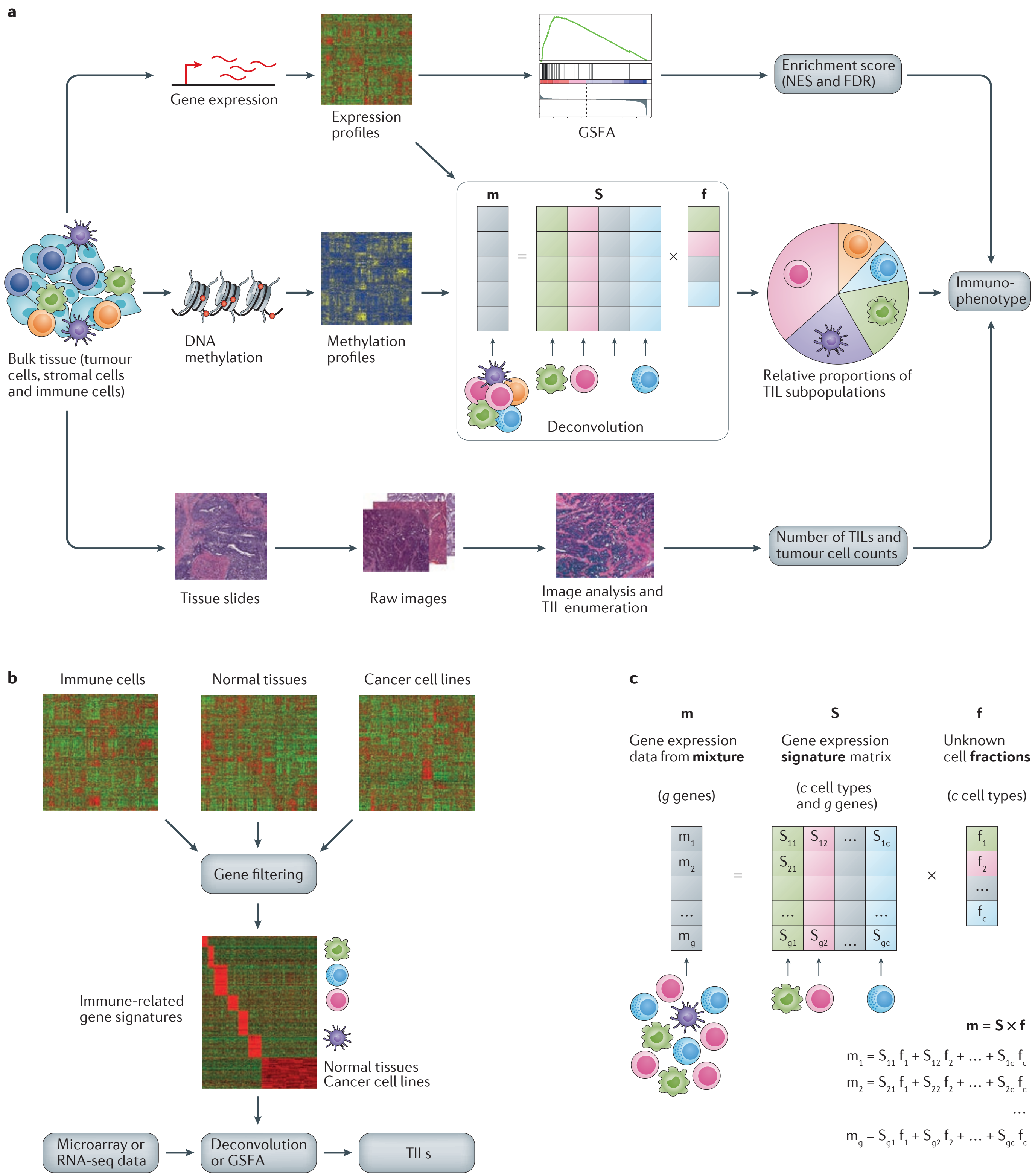

a
Gene expression
Expression profiles
GSEA
Enrichment score (NES and FDR)
Bulk tissue (tumour cells, stromal cells and immune cells)
DNA methylation
Methylation profiles
m
S
f
=
×
Deconvolution
Relative proportions of TIL subpopulations
Immuno-phenotype
Tissue slides
Raw images
Image analysis and TIL enumeration
Number of TILs and tumour cell counts
b
Immune cells
Normal tissues
Cancer cell lines
Gene filtering
Immune-related gene signatures
Normal tissues
Cancer cell lines
Microarray or RNA-seq data
Deconvolution or GSEA
TILs
c
m
S
f
Gene expression data from **mixture**
Gene expression **signature** matrix
Unknown cell **fractions**
(*g* genes)
(*c* cell types and *g* genes)
(*c* cell types)
m_1
m_2
...
m_g
=
S_{11} S_{12} ... S_{1c}
S_{21}
... ...
S_{g1} S_{g2} ... S_{gc}
×
f_1
f_2
...
f_c
$\mathbf{m = S \times f}$
$m_1 = S_{11} f_1 + S_{12} f_2 + \ldots + S_{1c} f_c$
$m_2 = S_{21} f_1 + S_{22} f_2 + \ldots + S_{2c} f_c$
...
$m_g = S_{g1} f_1 + S_{g2} f_2 + \ldots + S_{gc} f_c$

Figure 3 | **Determining cellular composition of tumour infiltrates using genomic data. a** | The immunophenotype, in particular the composition of tumour-infiltrating lymphocytes (TILs) in the tumour tissue, can be estimated based on different entities (gene expression profiles, DNA methylation profiles or immunohistochemistry) using individual or a combination of computational tools (that is, gene set enrichment analysis (GSEA), deconvolution and/or image analyses). **b** | Immune-related gene signatures are derived from expression profiles for different immune cell types, cell states and perturbations from publicly available data (including normal tissue and cancer cell lines as out-groups). **c** | Gene expression profiles in bulk tissues (m) are the result of the convolution of cell-type-specific gene expression signatures (S) with the fractions of different admixed cell types (f). Deconvolution approaches computationally trace back the unknown cell fractions that leverage on a signature expression matrix. FDR, false discovery rate; NES, normalized enrichment score; RNA-seq, RNA sequencing; TIL, tumour-infiltrating lymphocyte.

oncology now offers the possibility to identify HLA alleles directly from NGS data. In addition to the high sequence similarity of the different HLA genes and the difficulty in reconstructing phased genotypes of heterozygous alleles, HLA typing from shotgun sequencing data is further challenged by sequencing errors, read length and low coverage. With respect to standard algorithms for read mapping, methods for HLA typing must extract reads that originate from the HLA locus and assign them to the correct gene and allele. For this purpose, all algorithms rely on well-annotated reference sequences derived from the IMGT/HLA database and either implement a sensitive mapping strategy[57–63] or combine read mapping and assembly[57,64,65].

Since the publication of the first tools for HLA typing from NGS data, HLAminer[57] and Seq2HLA[58], several methods have been developed (TABLE 1), which have improved HLA typing performance in terms of both accuracy and resolution. The latest versions of the methods are optimized for four-digit resolution and consider different types of NGS data (TABLE 1). Among the available methods, the recently developed Polysolver[63] and Optitype[60] show high accuracy for HLA typing at four-digit resolution. The high sensitivity of Polysolver also facilitates the identification of somatic mutations in HLA genes, which has been shown to be in the range of a few per cent in different cancers[63]. However, as of today, an independent benchmarking study has not been carried out so far.

Prediction of neoantigen–MHC binding affinity. The MHC molecules that are directly involved in tumour rejection belong to class I (MHC-I) and are present on all nucleated cells. Viral or tumour antigens bound to class I MHC molecules are presented to CD8$^+$ T lymphocytes, which can consequently recognize and kill infected cells or tumour cells. Class II MHC molecules (MHC-II) are expressed only on specific cell types, such as dendritic cells, B lymphocytes and macrophages. MHC-II-bound antigens are presented to CD4$^+$ T cells and are involved in the activation of helper T cells. The relevance of class II MHC molecules for cancer immunotherapy has been recently emphasized by data that show that immunogenic mutations are recognized by CD4$^+$ T cells[66]. MHC-I molecules bind to peptides with a narrow length distribution, between 8 and 11 amino acids, whereas MHC-II molecules can accommodate longer peptides, up to 30 amino acids[67]. The complex that consists of a specific peptide bound to an MHC-I molecule is termed pMHC-I, and a complex with a peptide bound to a MHC-II molecule is termed pMHC-II. The binding affinity of pMHC-II is influenced by the peptide core and by the peptide flanking residues[68]. This promiscuous binding mechanism and the limited availability of training data render the task of pMHC-II binding affinity prediction more challenging than for pMHC-I, and algorithms for pMHC-II binding prediction are less accurate than pMHC-I methods[69].

Phased genotypes
Sets of alleles that are co-located on the same chromosome.

Several computational methods are now available to predict pMHC binding affinity, which can be classified into two major categories: structure-based methods, which consider protein 3D structures, and sequence-based methods, which consider the primary sequence of protein antigens. Owing to the limited number of 3D structures of pMHC complexes, modelling or simulation by protein threading and docking methods has to be used. Here, we focus on sequence-based methods, as there are large data sets for training and validation, and they are more suited for screening high-throughput data sets[70]. Early sequence-based methods, such as BIMAS[71] and SYFPEITHI[72], relied on position-specific scoring matrices (PSSMs), which were defined from experimentally confirmed peptide binders of a particular MHC allele. An allele-specific PSSM matrix encodes the binding affinity of each amino acid at every possible binding position. To model the nonlinear nature of the binding process, more advanced methods based on machine-learning techniques have been developed (TABLE 1). Nonlinear methods showed better performance than PSSM-based algorithms, owing to their ability to capture the complexity of the binding process and the interdependencies between protein residues. Higher performance was also obtained with consensus methods, such as NetMHCcons[73] and CONSENSUS[74], which combine multiple tools to obtain more reliable predictions. However, the performance gain over the single approaches is rather limited and should be weighed against the increased computational costs.

The development and validation of the methods discussed above would not have been possible without the wealth of information about MHC alleles and ligands that has been collected and made publicly available through Web-based databases, such as the Immune Epitope Database and Analysis Resource (IEDB)[75], IMGT/HLA[56] and the Dana–Farber Repository for Machine Learning in Immunology (DFRFMLI)[76] (TABLE 2). However, as the large majority of HLA alleles has not been investigated with respect to peptide binding, methods have been developed that go beyond the allele-specific approaches without requiring peptide data that are specific for each allele in question. These so-called pan-specific methods, such as NetMHCpan[77], allow the prediction of peptide binding to any MHC molecules of known protein sequence. To summarize briefly, a neural network is trained to output the affinity of a given pMHC pair that has the MHC represented by pseudo sequences constructed from HLA residues in contact with bound peptides, which are polymorphic in known functional HLA-A, HLA-B or HLA-C alleles. Pan-specific tools, such as NetMHCpan[77] and NetMHCIIpan[78], scored among the best performers, even compared to allele-specific approaches[79,80]. However, although several assessments and comparisons have been made in the past[79,81,82], there are currently no recent independent benchmark studies that can be used to recommend specific tools.

Although pMHC binding is the most selective event in the process of antigen presentation, the preceding steps of antigen processing also have a role in the MHC-I pathway: proteasomal cleavage, which is required to convert large proteins into smaller peptides; and transport of the peptide into the endoplasmic reticulum by

Table 1 | **Computational tools for cancer immunology**

Tool	Description	URL	Refs
TIL quantification			
CellMix	Framework for application of different deconvolution methods to expression data	http://web.cbio.uct.ac.za/~renaud/CRAN/web/CellMix	130
CIBERSORT	Deconvolution of 22 immune cell types from gene expression microarray data of cell mixtures, based on support vector regression and gene expression signatures	http://cibersort.stanford.edu	38
CRImage	R package for classifying cells and calculating tumour cellularity based on H&E-stained whole-section slide images and microarray expression profiles	http://www.markowetzlab.org/software/CRImage.php	131
deconf	Marker-gene-based or unsupervised deconvolution of cell fractions and gene expression signatures from microarray data of cell mixtures	http://www.biomedcentral.com/1471-2105/11/27	132
DeconRNAseq	Deconvolution of admixtured tissues profiled with RNA-seq data through quadratic programming	http://www.bioconductor.org/packages/release/bioc/html/DeconRNASeq.html	37
ImmunoRatio	Web application for automated image analysis of oestrogen receptor (ER), progesterone receptor (PR) and Ki-67 immunostained tissue sections	http://153.1.200.58:8080/immunoratio	133
MMAD	Simultaneous deconvolution of cell fractions and expression profiles from microarray data of cell mixtures using marker genes	http://sourceforge.net/projects/mmad	134
PERT	Probabilistic deconvolution of microarray data from cell mixtures, accounting for variation with respect to the cell-specific signature matrix	http://github.com/gquon/PERT	135
SPEC	Prediction of the major cellular source of gene expression microarray data of cell mixtures by computation of enrichment scores for cell-type-specific marker genes	http://clip.med.yale.edu/SPEC	136
ssGSEA	Estimation of enrichment or depletion of immune cell types from single-sample microarray or RNA-seq data using GSEA and expression markers	http://www.broadinstitute.org/cancer/software/genepattern/modules/docs/ssGSEAProjection	29,31
Mutated proteins			
ANNOVAR	Annotation of SNPs and indels detected from human and non-human genomes	http://annovar.openbioinformatics.org	137
EBCall	Detection of genomic SNPs and indels in the presence of sequencing errors and low sequencing depths	http://github.com/friend1ws/EBCall	51
GATK	SNP and indel calling and quality control, applicable to WES, WGS and RNA-seq data	http://www.broadinstitute.org/gatk	48
MuTect	High-sensitivity calling of somatic SNPs, even in cases of low allele fractions	http://www.broadinstitute.org/cancer/cga/mutect	49
Oncotator	Annotation of SNPs and indels detected from cancer data	http://www.broadinstitute.org/oncotator	138
SNPeff	Prediction of coding effects of SNPs and small indels	http://snpeff.sourceforge.net	139
SomaticSniper	Calling of somatic SNPs and indels from matched tumour–normal NGS data	http://gmt.genome.wustl.edu/packages/somatic-sniper	140
Strelka	Detection of SNPs and indels from matched tumour–normal NGS data with various degrees of tumour purity	http://sites.google.com/site/strelkasomaticvariantcaller	141
TransVar	Annotation of genetic variants at RNA and protein level and inverse annotation of isoforms to their genomic origin	http://bioinformatics.mdanderson.org/main/Transvar	142
VarScan	Calling of somatic and germline SNPs and indels from data generated with different NGS platforms	http://varscan.sourceforge.net	143
VEP	SNP consequence prediction for species annotated in the Ensembl database	http://www.ensembl.org/Tools/VEP	144

Table 1 (cont.) | **Computational tools for cancer immunology**

Tool	Description		Refs
HLA typing			
ATHLATES	Genotyping of HLA-I and HLA-II alleles, from WGS and WES Illumina data	http://www.broadinstitute.org/scientific-community/science/projects/viral-genomics/athlates	64
HLAforest	Hierarchical reconstruction of HLA-I and HLA-II alleles from RNA-seq data	http://code.google.com/p/hlaforest	59
HLAminer	Extraction of HLA-I and HLA-II types from non-targeted RNA-seq, WGS and WES data based on read mapping or *de novo* assembly	http://www.bcgsc.ca/platform/bioinfo/software/hlaminer	57
HLAreporter	WGS- and WES-based genotyping of HLA-I and HLA-II alleles at six-digit resolution	http://paed.hku.hk/genome/software.html	65
HLA-VBseq	Extraction of eight-digit resolution HLA-I and HLA-II from WGS data	http://nagasakilab.csml.org/hla	62
Optitype	High-accuracy genotyping of classical HLA-I alleles from RNA-seq, WGS and WES data	http://github.com/FRED-2/OptiType	60
PHLAT	Genotyping of HLA-I and HLA-II alleles from RNA-seq, WGS, WES and targeted sequencing for different read lengths and coverages	http://sites.google.com/site/phlatfortype	61
Polysolver	Genotyping of HLA-I alleles from WES data and calling of somatic mutations in the HLA loci	http://www.broadinstitute.org/cancer/cga/polysolver	63
Seq2HLA	Extraction of HLA-I and HLA-II types from whole-genome RNA-seq, currently optimized also for four-digit resolution	http://bitbucket.org/sebastian_boegel/seq2hla	58,145
pMHC binding			
CONSENSUS	Consensus approach for prediction of pMHC-I and pMHC-II binding affinity hosted on the IEDB website	http://tools.immuneepitope.org/mhcii	74,82
netMHC	Machine-learning-based prediction of pMHC binding affinity to human and non-human MHC-I molecules	http://www.cbs.dtu.dk/services/NetMHC	146
netMHCcons	Consensus-based prediction of pMHC-I binding affinity integrating the predictions of NetMHC, NetMHCpan and PickPocket	http://www.cbs.dtu.dk/services/NetMHCcons	73
netMHCpan	Pan-specific version of netMHC	http://www.cbs.dtu.dk/services/NetMHCpan	147
netMHCstab	Machine-learning-based prediction of the stability of binding of small peptides to HLA-A and HLA-B molecules	http://www.cbs.dtu.dk/services/NetMHCstab-1.0	83
netMHCII	Machine-learning-based prediction of binding affinity to human and mouse class-II MHC molecules	http://www.cbs.dtu.dk/services/NetMHCII	68
netMHCIIpan	Pan-specific version of netMHCII	http://www.cbs.dtu.dk/services/NetMHCIIpan	78
PickPocket	Pan-specific predictor of pMHC-I binding affinity based on PSSM of peptides and MHC pockets	http://www.cbs.dtu.dk/services/PickPocket	148
Pipelines for neoantigen prediction			
FRED 2	HLA typing and T-cell epitope prediction, selection and assembly	http://github.com/FRED-2/Fred2	112
NetCTL	Prediction of immunogenic peptides by integration of proteasomal cleavage, TAP transport and pMHC-I affinity	http://www.cbs.dtu.dk/services/NetCTL	110
NetCTLpan	Pan-specific version of NetCTLpan	http://www.cbs.dtu.dk/services/NetCTLpan	149
EpiToolKit	Web-based, flexible workbench for integration of class-I HLA typing and T-cell epitope prediction and selection	http://www.epitoolkit.de	111
NetTepi	Identification of antigenic peptides based by integrating prediction of pMHC-I binding affinity and stability and T-cell propensity	http://www.cbs.dtu.dk/services/NetTepi-1.0	90

Table 1 (cont.) | **Computational tools for cancer immunology**

Tool	Description		Refs
Pipelines for neoantigen prediction (cont.)			
pVAC-Seq	Identification and prioritization of personalized neoantigens from mutation and expression data	http://github.com/griffithlab/pVAC-Seq	113
WAPP	Integrated prediction of pMHC-I processing and presentation: proteasomal cleavage, TAP transport and pMHC-I binding affinity	http://abi.inf.uni-tuebingen.de/Services/WAPP	87
TCR profiling			
Decombinator	Estimation of V and J usage and nucleotide deletion at junctional sites of TCRs profiled with Illumina Rep-seq	http://github.com/uclinfectionimmunity/Decombinator	102
IMGT/HighV-QUEST	Webserver for reconstruction of TCR and BCR repertoires form medium-length Rep-seq reads	http://www.imgt.org/HighV-QUEST	105
IMSEQ	Extraction of CDR3 clonotypes from TCR and BCR Rep-seq data, embedding error correction	http://www.imtools.org	101
LymAnalyzer	Estimation of V(D)J usage from TCR and BCR Rep-seq data reconstruction of lineage mutation trees of BCR maturation	http://sourceforge.net/projects/lymanalyzer	97
MIGEC	Reconstruction of TCR and BCR clonotypes from Rep-seq data generated with UMI protocol for removal of experimental errors	http://github.com/mikessh/migec	99
MiTCR	Extraction of TCR clonotypes and V(D)J usage from Rep-seq and RNA-seq data with removal of PCR artefacts and sequencing errors	http://github.com/milaboratory/mitcr	96
MiXCR	Extraction of error-corrected TCR and BCR clonotypes and estimation of V(D)J segment usage from Rep-seq and RNA-seq data	http://github.com/milaboratory/mixcr	98
TCRklass	TCR profiling from short-read Rep-seq handling also TCR sequences with no full-length CDR3 sequences	http://sourceforge.net/projects/tcrklass	100

BCR, B cell receptor; CDR3, complementary determining region 3; GSEA, gene set enrichment analysis; H&E, haematoxylin and eosin stain; HLA, human leukocyte antigen; indel, small insertion or deletion; MHC, major histocompatibility complex; NGS, next-generation sequencing; PCR, polymerase chain reaction; pMHC, peptide–MHC complex; PSSM, position-specific scoring matrix; Rep-seq, repertoire sequencing; RNA-seq, RNA sequencing; SNP, single-nucleotide polymorphism; TAP, transporter associated with antigen processing; TCR, T cell receptor; TILs, tumour infiltrating lymphocytes; UMI, unique molecular identifier; V(D)J, variable, diversity and joining genomic loci; WES, whole-exome sequencing;WGS, whole-genome sequencing.

transporter associated with antigen processing (TAP), which is needed to bind the peptide to the MHC-I molecule. In addition, the stability of the pMHC complex and the CD8$^+$ T cell propensity for recognizing the pMHC complex is another factor that determines the immunogenicity of a mutated peptide. Currently, there is only one tool available for the prediction of binding stability: NetMHCstab[83]. Tools for the prediction of proteasomal cleavage (for example, netChop Cterm[84] and Pcleavage[85]), TAP transport (for example, PredTAP[86] and SVMTAP[87]) and T cell propensity (for example, POPI[88] and POPISK[89]) are available, but the value of predictions is rather limited[69]. Despite their limitations, these methods can be integrated into antigen prediction pipelines to reduce false positives and hence to minimize the experimental workload needed for epitope validation. A good example of method integration is implemented in NetTepi[90], which is a computational pipeline for the prediction of T-cell epitopes that combines binding affinity, binding stability and T-cell propensity predicted using an extension of the immunogenicity model described in REF. 91 (an implementation of this model is included on the IEDB website). Although integrative analyses, such as the one implemented in NetTepi, hold promise for future research, improvements of the methods that address all steps of antigen processing are required to unlock their full potential.

TCR profiling

The TCRs must be able to mount an immune response for a wealth of different antigens. The immense receptor repertoire diversity is the result of a process called V(D)J recombination, which consists of the somatic rearrangement of different gene segments belonging to the variable (V), diversity (D) and joining (J) genomic loci[67]. This combinatorial diversity is further increased by the addition or removal of random nucleotides at the junction sites and by combination of different α- and β-chains. Originally profiled using Sanger sequencing or spectratyping, TCR repertoires can be now analysed with NGS technologies at unprecedented resolution and throughput[92,93]. However, careful experimental design and data pre-processing[92,94,95] are needed to represent receptor diversity without bias. In particular, PCR artefacts and sequencing errors can hamper the quantification of junctional diversity. Therefore, recent tools for TCR profiling[81] implement strategies for error correction in their analytical procedures[96–101]. Alongside

T cell propensity
Measure of how much T cell receptors are prone to interact with specific major-histocompatibility-complex-binding peptides.

Epitope
Part of an antigen that is recognized by the immune system.

Spectratyping
A method to study the T cell receptor repertoire. It is based on polymerase chain reaction amplification of rearranged genes of the T cell receptor beta variable gene family. The density of heterogeneous complementarity-determining region 3 (CDR3) lengths, which are separated by electrophoresis, results in a specific spectrum that is then further analysed.

TCR repertoires
(T cell receptor repertoires). Diversity of TCRs that allows the T cells of the immune system to specifically recognize the huge number of various antigens.

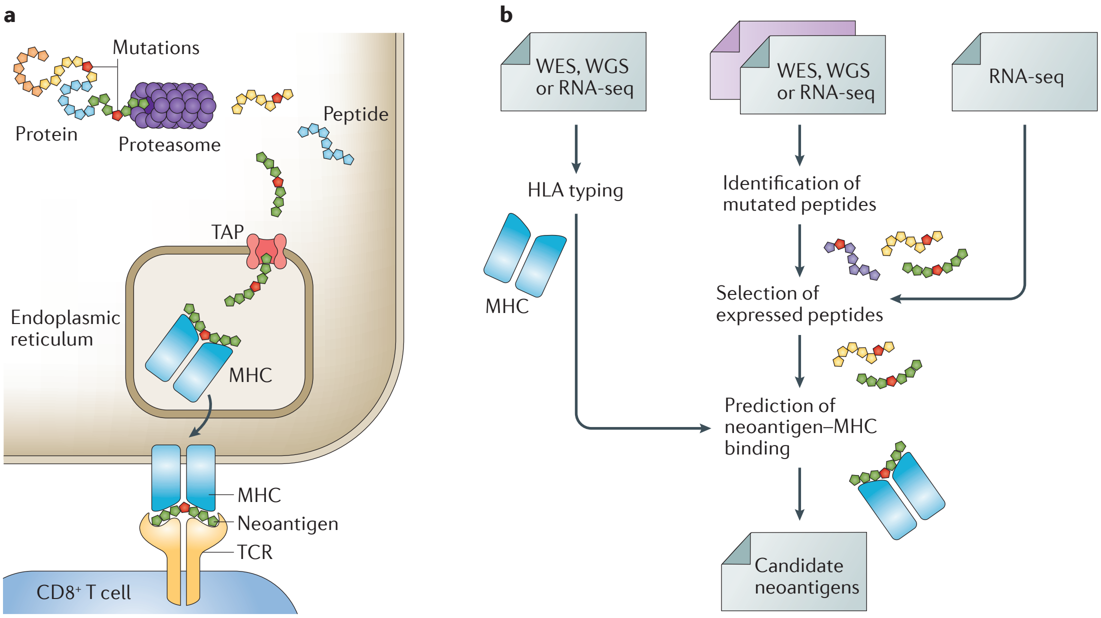


Figure 4 | **Identification of cancer neoantigens. a** | Neoantigens originate from mutated proteins expressed in cancer cells. The mutated protein is cleaved into shorter peptides by the proteasome and transported by transporter associated with antigen processing (TAP) to the endoplasmic reticulum, in which the peptides bind the major histocompatibility complex (MHC) molecule. Then, the peptide–MHC complex is displayed on the cell surface of the antigen presenting cell to be recognized by the T cell receptor (TCR) of CD8$^+$ T cells. **b** | Prediction of candidate neoantigens from next-generation sequencing (NGS) data requires the implementation of several computational tasks: prediction of mutated peptides from whole-exome sequencing (WES), whole-genome sequencing (WGS) or RNA sequencing (RNA-seq) data from matched tumour–normal samples; selection of expressed peptides by integrating RNA-seq data of the tumour sample; human leukocyte antigen (HLA) typing from WES, WGS or RNA-seq data of the tumour sample; and prediction of peptide–MHC binding for specific HLA alleles.

error-correction strategies, recent tools[96–98,102] are more suited for the analysis of large NGS data sets than are early tools[103–105] owing to the improved computational performance. All TCR profiling tools reported in TABLE 1 can be used on data from targeted repertoire sequencing (Rep-seq) of the TCR locus and, in some cases, also of the B cell receptor locus[97–99]. The latest version of MiXCR[98] directly supports the extraction of immune repertoires from whole-transcriptome RNA-seq data sets. Recently, a computational strategy based on MiTCR was used to characterize TCR repertoire using RNA-seq data from TCGA[106]. Most of the published studies focused on the complementarity-determining region 3β (CDR3β), because this region accounts for most of the variation in a person's T cell repertoire. With the advent of single-cell sequencing technologies, it is now possible to carry out paired analysis of the TCR αβ-chains[107], which contribute to the TCR specificity.

Assembling analytical pipelines

The development of computational methods for cancer immunogenomics and the availability of extensive and curated databases are providing the basis for tools to dissect tumour–immune cell interactions, thus paving the way towards applications such as cancer vaccination and personalized cancer immunotherapy. However, despite the availability of several bioinformatics tools, the lack of standardization prevents an easy assembly of analytical pipelines. Although platforms that allow integration of tools for data pre-processing and genomic analysis are available (for example, Galaxy[108]), pipelines that integrate several types of immunogenomic analyses have to be assembled *ad hoc*. In this respect, standalone solutions are better suited, as they can be more easily integrated and allow parallel analyses of high-throughput data.

In assembling a computational pipeline for cancer immunogenomics, particular care must be taken to assess whether the data allows the extraction of unbiased and meaningful information. For instance, read length and depth of coverage have a strong impact on the analysis of immune repertoires and HLA alleles from sequencing data[57,94]. In addition, the sequencing platform that is used for generating the data must also be taken into consideration to ascertain whether the selected tools are capable of distinguishing platform-specific sequencing errors from true variants. Another issue to take into consideration is batch effects, which are technical sources of variation caused by different experimental conditions. These artefacts can be identified using dimensionality-reduction tools, such as principal component analysis, and subsequently corrected using surrogate variable analysis[109].

Only recently, computational solutions for neoantigen predictions that integrate several analytical steps have been reported (TABLE 1). Besides NetTepi[90], other solutions are currently available, such as: NetCTLpan[110],

Repertoire sequencing
(Rep-seq). Targeted sequencing of the genome loci encoding the T cell (or B cell) receptor taking the complexity of different arrangements into account.

Table 2 | **Selection of databases and Web servers for cancer immunology**

Database	Synopsis	URL	Refs
Cancer molecular databases			
Cancer Cell Line Encyclopedia (CCLE)	Analysis and visualization of DNA copy number, mRNA expression and mutation data for 1,000 cancer cell lines	http://www.broadinstitute.org/ccle	150
Cancer Genome Anatomy Project (CGAP)	Gene expression profiles of normal, precancer and cancer cells. Interconnected modules provide access to data and bioinformatics analysis tools	http://cgap.nci.nih.gov	151
Cancer genomics browser at UCSC	Browser for visualization, integration and analysis of cancer genomics data and associated clinical information	http://genome-cancer.ucsc.edu	152
Catalogue of Somatic Mutation in Cancer (COSMIC)	Curated information about somatic mutations in human cancers	http://cancer.sanger.ac.uk/cosmic	153
cBio Cancer Genomics Portal	Web server for exploration, analysis and download of large-scale cancer genomics data sets	http://cbioportal.org	154
Clinical Proteomic Tumour Analysis Consortium (CPTAC)	Integrative repository of genomics and proteomics tumour data	http://proteomics.cancer.gov	155
GDAC Firehose/Firebrowse	Web browser for easy download and analysis of TCGA data	http://firebrowse.org	–
International Cancer Genome Consortium (ICGC)	Data portal that provides tools for visualizing, querying and downloading of ICGC released data	http://dcc.icgc.org	1
IntOGen	Repository of somatic mutations in thousands of tumour genomes and tools for driver gene identification, mutation mapping and visualization	http://www.intogen.org	156
The Cancer Genome Atlas (TCGA)	Data portal to search, download and analyse all data sets generated by TCGA for more than 30 cancer types	http://cancergenome.nih.gov	2
The Human Protein Atlas	Data repository of the human proteome by tissue based analysis (including cancer tissue sections)	http://proteinatlas.org	157
Immunology databases			
AntiJen	Database of quantitative binding data for MHC ligands, TCR–MHC complexes, T cell epitopes, TAP binders and B cell epitopes	http://www.ddg-pharmfac.net/antijen	158
CTdatabase	Curated information of CGA genes and protein products, expression and induced immune response	http://www.cta.lncc.br	46
GPXdb	Macrophage expression atlas	http://gpxmea.gti.ed.ac.uk	159
HaemAtlas	Gene expression profiles in differentiated human blood cells	http://t1dbase.org/page/HaemAtlasHome	160
EPIMHC	Naturally processed MHC-restricted peptide ligands and epitopes	http://imed.med.ucm.es/epimhc	161
Immune Epitope Database and Analysis Resource (IEDB)	Manually curated database of experimentally characterized immune epitopes and tools for the prediction and analysis of immune epitopes	http://www.iedb.org/	75
IMGT/HLA Database	Sequences of HLA	http://ebi.ac.uk/ipd/imgt/hla	162
ImmGen	Comprehensive resource of gene expression and its regulation in the immune system of the mouse provided by the Immunological Genome Project	http://immgen.org	13
Immport	Resource for searching and downloading shared immunological data from different studies	http://immport.org	163

Table 2 (cont.) | **Selection of databases and Web servers for cancer immunology**

Database	Synopsis	URL	Refs
Immunology databases (cont.)			
ImmuneSigDB	Collection of immunological gene signatures as part of the MSigDB, which can be used for gene set enrichment analysis	http://broadinstitute.org/gsea/msigdb	35
ImMunoGeneTics (IMGT) information system	Repository of BCR, TCR and HLA sequences and structural data of human and other vertebrate species	http://imgt.org	56
InnateDB	Genes, proteins, experimentally verified interactions and signalling pathways involved in the innate immune response of humans, mice and bovines to microbial infection	http://innatedb.com	164
SYFPEITHI	Resource including MHC ligands, peptide motifs and tools for the prediction of immune epitopes	http://syfpeithi.de	72
T cell-defined tumour antigens	Resource with MHC ligands and peptide motifs and tools for the prediction of immune epitopes	http://cancerimmunity.org/peptide	165

BCR, B cell receptor; CGA, cancer germline antigen; HLA, human leukocyte antigen; MHC, major histocompatibility complex; TAP, transporter associated with antigen processing; TCR, T cell receptor; UCSC, University of California Santa Cruz.

which is a pan-specific method for predicting proteasomal cleavage, TAP transport and pMHC binding; EpiToolkit[111], which is a Web-based platform for flexible integration of pre-selected computational modules for epitope prediction and prioritization; FRED 2 (REF. 112), which is a web resource for HLA typing, epitope prediction and selection, that also allows prototyping of customized pipelines; and pVAC-Seq[113], which is a neoantigen identification pipeline that takes into consideration mutation coverage, variant allele frequency and the expression of the mutated genes. With the increased availability of cloud computing solutions, we expect that in the near future, analytical pipelines for cancer immunogenomics will be developed that make use of this computational infrastructure.

Conclusions

Cancer immunotherapy that targets T cells to enhance immune response has elicited durable clinical responses and long-term remissions in melanoma and is showing efficacy in various cancers[7]. As several agents that target checkpoint molecules are entering the stage, it is likely that cancer immunotherapy will become a major cancer treatment regimen. However, identifying patients that are likely to benefit from immunotherapy is of utmost importance and depends on our ability to deeply mine genomics data using existing and novel computational tools for the characterization of the neoantigen landscape and the immunophenotypes of the tumours. As neoantigens are recognized as major determinants of tumour recognition by the immune system, their identification using NGS and computational genomics tools is an important prerequisite for elucidating the causes of resistance to cancer immunotherapy with checkpoint blockers and for developing cancer vaccination strategies. For example, a recent study using several tools reviewed here showed that sensitivity to PD1 and CTLA4 blockade in lung cancer and melanoma is enhanced in tumours that are enriched with neoantigens that have a clonal origin[4]. Similarly, the characterization of the immunophenotype of the tumour using TILs is of utmost importance for the identification of the tumour escape mechanisms[10] and the development of novel immunotherapeutic approaches that include depletion of immunosuppressive TILs and modulation of the tumour microenvironment[114]. The identification of the immunophenotypes and the prediction of neoantigens before checkpoint blocker treatment might also be beneficial for the stratification of patients. For instance, patients with metastatic melanoma who respond to PD1 blockade showed higher numbers of cells expressing CD8, PD1 and/or PD1 ligand 1 (PDL1; also known as CD274) at the invasive tumour margin and inside tumours and a more clonal TCR repertoire[115]. Similarly, neoantigen load and cytolytic activity were associated with CTLA4 blockade response in melanoma patients[116].

With the increasing number of ongoing trials and the use of checkpoint blockers in routine settings, the amount of generated data will pose considerable technical and scientific challenges. It is evident that the exploitation of these data sets requires not only the use and improvement of available computational genomics tools, but also the development of novel methods that address challenging issues. As in cancer genomics, it will be also necessary to improve the tools further and to carry out systematic benchmarking studies for the HLA typing methods, prediction of pMHC binding affinity and TCR profiling. Furthermore, it will be also necessary to initiate crowdsourcing-based Dialogue on Reverse-Engineering Assessment and Methods (DREAM) Challenges to solve fundamental problems, including prediction of immunogenicity of neoantigens and identification of synergistic combinations of checkpoint blockers and targeted drugs.

Previously, mutational load was suggested to be a limiting factor for responsiveness to checkpoint blocker therapy; however, a recent study showed that even tumours with low mutational load are responding[117]. Thus, the identification of neoantigens that are immunogenic might help to stratify the patients. Moreover, predicting immunogenicity of neoantigens will also dramatically improve cancer vaccination. The number of epitopes that can be included in a vaccine is currently limited to 10–20 owing to manufacturing constraints. The clinical response to epitope-based vaccines crucially depends on the number of peptides that are present on patient MHC molecules: the number of peptides that evoke an immune response strongly correlates with patient survival[118]. Thus, from the large set of neoantigens, it is crucial to pick those that have the highest likelihood of success: that is, to find an optimal design for the neoantigen-based vaccine. Hence, the identification of rules that define immunogenicity of neoantigens is essential, and novel computational and experimental approaches are required to tackle this highly challenging task. For example, a recent case report demonstrated that NGS can be used to identify CD4$^+$ T cells that are specific for a mutated antigen in a patient with metastatic cholangiocarcinoma. Infusion with an expanded population of mutation-specific T cells resulted in tumour regression[119].

As immune checkpoint therapy results in a durable response only in a fraction of patients, efforts are underway to increase the response rate by combining molecularly targeted agents and immunotherapy[120] (BOX 1). Given the large number of TIL subpopulations and their phenotypic plasticity, the investigation of the immunological impact of targeted agents will be highly challenging and is likely to require novel concepts. In this context, predicting synergistic effects of drug combinations using computational tools will be tremendously helpful and will accelerate the development of these promising strategies. All of these efforts will enable precision oncology with cancer immunotherapy to be one of the therapeutic pillars with molecular and computational tools as indispensable components.

1. International Cancer Genome Consortium. International network of cancer genome projects. *Nature* **464**, 993–998 (2010).
2. Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455**, 1061–1068 (2008).
The TCGA Research Network provides clinical information, genomic characterization data and high-level sequence analysis of the tumour genomes.
3. Fojo, T., Mailankody, S. & Lo, A. Unintended consequences of expensive cancer therapeutics—the pursuit of marginal indications and a me-too mentality that stifles innovation and creativity: the John Conley Lecture. *JAMA Otolaryngol. Head Neck Surg.* **140**, 1225–1236 (2014).
4. McGranahan, N. *et al.* Clonal neoantigens elicit T cell immunoreactivity and sensitivity to immune checkpoint blockade. *Science* **351**, 1463–1469 (2016).
5. Pardoll, D. M. The blockade of immune checkpoints in cancer immunotherapy. *Nat. Rev. Cancer* **12**, 252–264 (2012).
6. Schadendorf, D. *et al.* Pooled analysis of long-term survival data from Phase II and Phase III trials of ipilimumab in unresectable or metastatic melanoma. *J. Clin. Oncol.* **33**, 1889–1894 (2015).
7. Wolchok, J. D. PD-1 blockers. *Cell* **162**, 937 (2015).
8. Schumacher, T. N. & Schreiber, R. D. Neoantigens in cancer immunotherapy. *Science* **348**, 69–74 (2015).
9. Chen, D. S. & Mellman, I. Oncology meets immunology: the cancer–immunity cycle. *Immunity* **39**, 1–10 (2013).
10. Angelova, M. *et al.* Characterization of the immunophenotypes and antigenomes of colorectal cancers reveals distinct tumor escape mechanisms and novel targets for immunotherapy. *Genome Biol.* **16**, 64 (2015).
11. Schreiber, R. D., Old, L. J. & Smyth, M. J. Cancer immunoediting: integrating immunity's roles in cancer suppression and promotion. *Science* **331**, 1565–1570 (2011).
12. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
13. Heng, T. S. P. & Painter, M. W. & Immunological Genome Project Consortium. The Immunological Genome Project: networks of gene expression in immune cells. *Nat. Immunol.* **9**, 1091–1094 (2008).
The Immunological Genome Project provides a comprehensive compendium of gene expression and regulation networks in immune cells and cell lineages.
14. Saadatpour, A., Lai, S., Guo, G. & Yuan, G.-C. Single-cell analysis in cancer genomics. *Trends Genet.* **31**, 576–586 (2015).
15. Hashimshony, T., Wagner, F., Sher, N. & Yanai, I. CEL-seq: single-cell RNA-seq by multiplexed linear amplification. *Cell Rep.* **2**, 666–673 (2012).
16. Fan, H. C., Fu, G. K. & Fodor, S. P. A. Expression profiling. Combinatorial labeling of single cells for gene expression cytometry. *Science* **347**, 1258367 (2015).
17. Klein, A. M. *et al.* Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161**, 1187–1201 (2015).
18. Macosko, E. Z. *et al.* Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202–1214 (2015).
19. Gaublomme, J. T. *et al.* Single-cell genomics unveils critical regulators of Th17 cell pathogenicity. *Cell* **163**, 1400–1412 (2015).
20. Stubbington, M. J. T. *et al.* T cell fate and clonality inference from single-cell transcriptomes. *Nat. Methods* **13**, 329–332 (2016).
21. Gawad, C., Koh, W. & Quake, S. R. Single-cell genome sequencing: current state of the science. *Nat. Rev. Genet.* **17**, 175–188 (2016).
22. Pabinger, S. *et al.* A survey of tools for variant analysis of next-generation genome sequencing data. *Brief. Bioinform.* **15**, 256–278 (2014).
23. Ding, L., Wendl, M. C., McMichael, J. F. & Raphael, B. J. Expanding the computational toolbox for mining cancer genomes. *Nat. Rev. Genet.* **15**, 556–570 (2014).
24. Carter, S. L. *et al.* Absolute quantification of somatic DNA alterations in human cancer. *Nat. Biotechnol.* **30**, 413–421 (2012).
25. Roth, A. *et al.* PyClone: statistical inference of clonal population structure in cancer. *Nat. Methods* **11**, 396–398 (2014).
26. Bock, C. Analysing and interpreting DNA methylation data. *Nat. Rev. Genet.* **13**, 705–719 (2012).
27. Fridman, W. H., Pagès, F., Sautès-Fridman, C. & Galon, J. The immune contexture in human tumours: impact on clinical outcome. *Nat. Rev. Cancer* **12**, 298–306 (2012).
28. Galon, J. *et al.* Type, density, and location of immune cells within human colorectal tumors predict clinical outcome. *Science* **313**, 1960–1964 (2006).
29. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102**, 15545–15550 (2005).
GSEA is a widely used tool for determining whether an a priori defined set of genes shows statistically significant, concordant differences between two biological states (for example, phenotypes).
30. Shen-Orr, S. S. & Gaujoux, R. Computational deconvolution: extracting cell type-specific information from heterogeneous samples. *Curr. Opin. Immunol.* **25**, 571–578 (2013).
31. Barbie, D. A. *et al.* Systematic RNA interference reveals that oncogenic *KRAS*-driven cancers require TBK1. *Nature* **462**, 108–112 (2009).
32. Kidd, B. A., Peters, L. A., Schadt, E. E. & Dudley, J. T. Unifying immunology with informatics and multiscale biology. *Nat. Immunol.* **15**, 118–127 (2014).
33. Rahmatallah, Y., Emmert-Streib, F. & Glazko, G. Gene set analysis approaches for RNA-seq data: performance evaluation and application guideline. *Brief. Bioinform.* **17**, 393–407 (2015).
34. Abbas, A. R. *et al.* Immune response *in silico* (IRIS): immune-specific genes identified from a compendium of microarray expression data. *Genes Immun.* **6**, 319–331 (2005).
35. Godec, J. *et al.* Compendium of immune signatures identifies conserved and species-specific biology in response to inflammation. *Immunity* **44**, 194–206 (2016).
36. Liberzon, A. *et al.* Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**, 1739–1740 (2011).
37. Gong, T. & Szustakowski, J. D. DeconRNASeq: a statistical framework for deconvolution of heterogeneous tissue samples based on mRNA-seq data. *Bioinformatics* **29**, 1083–1085 (2013).
38. Newman, A. M. *et al.* Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods* **12**, 453–457 (2015).
CIBERSORT is a computational method for characterizing cell composition of complex tissues from their gene expression profiles.
39. Accomando, W. P., Wiencke, J. K., Houseman, E. A., Nelson, H. H. & Kelsey, K. T. Quantitative reconstruction of leukocyte subsets using DNA methylation. *Genome Biol.* **15**, R50 (2014).
40. Houseman, E. A. *et al.* DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics* **13**, 86 (2012).
41. Koestler, D. C. *et al.* Blood-based profiles of DNA methylation predict the underlying distribution of cell types: a validation analysis. *Epigenetics* **8**, 816–826 (2013).
42. Zou, J., Lippert, C., Heckerman, D., Aryee, M. & Listgarten, J. Epigenome-wide association studies without the need for cell-type composition. *Nat. Methods* **11**, 309–311 (2014).
43. Houseman, E. A., Kelsey, K. T., Wiencke, J. K. & Marsit, C. J. Cell-composition effects in the analysis of DNA methylation array data: a mathematical perspective. *BMC Bioinformatics* **16**, 95 (2015).
44. Coulie, P. G., Van den Eynde, B. J., van der Bruggen, P. & Boon, T. Tumour antigens recognized by T lymphocytes: at the core of cancer immunotherapy. *Nat. Rev. Cancer* **14**, 135–146 (2014).
45. van der Bruggen, P. *et al.* A gene encoding an antigen recognized by cytolytic T lymphocytes on a human melanoma. *Science* **254**, 1643–1647 (1991).

46. Almeida, L. G. *et al.* CTdatabase: a knowledge-base of high-throughput and curated data on cancer-testis antigens. *Nucleic Acids Res.* **37**, D816–D819 (2009).
This is the most updated and curated resource for cancer germline antigens and contains information about cancer germline genes, their expression in tumour and normal tissues and induced immune response.
47. Fonseca, N. A., Rung, J., Brazma, A. & Marioni, J. C. Tools for mapping high-throughput sequencing data. *Bioinformatics* **28**, 3169–3177 (2012).
48. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
49. Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* **31**, 213–219 (2013).
50. Wang, Q. *et al.* Detecting somatic point mutations in cancer genome sequencing data: a comparison of mutation callers. *Genome Med.* **5**, 91 (2013).
51. Shiraishi, Y. *et al.* An empirical Bayesian framework for somatic mutation detection from cancer genome sequencing data. *Nucleic Acids Res.* **41**, e89 (2013).
52. Flicek, P. *et al.* Ensembl 2012. *Nucleic Acids Res.* **40**, D84–D90 (2012).
53. Pruitt, K. D., Tatusova, T., Brown, G. R. & Maglott, D. R. NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Res.* **40**, D130–D135 (2012).
54. UniProt Consortium. UniProt: a hub for protein information. *Nucleic Acids Res.* **43**, D204–D212 (2015).
55. McCarthy, D. J. *et al.* Choice of transcripts and software has a large effect on variant annotation. *Genome Med.* **6**, 26 (2014).
56. Lefranc, M. -P. *et al.* IMGT®, the international ImMunoGeneTics information system® 25 years on. *Nucleic Acids Res.* **43**, D413–D422 (2015).
57. Warren, R. L. *et al.* Derivation of HLA types from shotgun sequence datasets. *Genome Med.* **4**, 95 (2012).
58. Boegel, S. *et al.* HLA typing from RNA-seq sequence reads. *Genome Med.* **4**, 102 (2012).
Seq2HLA and HLAminer, described in references 57 and 58, were the first tools that used NGS data to derive HLA alleles.
59. Kim, H. J. & Pourmand, N. HLA typing from RNA-seq data using hierarchical read weighting. *PLoS ONE* **8**, e67885 (2013).
60. Szolek, A. *et al.* OptiType: precision HLA typing from next-generation sequencing data. *Bioinformatics* **30**, 3310–3316 (2014).
61. Bai, Y., Ni, M., Cooper, B., Wei, Y. & Fury, W. Inference of high resolution HLA types using genome-wide RNA or DNA sequencing reads. *BMC Genomics* **15**, 325 (2014).
62. Nariai, N. *et al.* HLA-VBSeq: accurate HLA typing at full resolution from whole-genome sequencing data. *BMC Genomics* **16**, S7 (2015).
63. Shukla, S. A. *et al.* Comprehensive analysis of cancer-associated somatic mutations in class I HLA genes. *Nat. Biotechnol.* **33**, 1152–1158 (2015).
64. Liu, C. *et al.* ATHLATES: accurate typing of human leukocyte antigen through exome sequencing. *Nucleic Acids Res.* **41**, e142 (2013).
65. Huang, Y. *et al.* HLAreporter: a tool for HLA typing from next generation sequencing data. *Genome Med.* **7**, 25 (2015).
66. Kreiter, S. *et al.* Mutant MHC class II epitopes drive therapeutic immune responses to cancer. *Nature* **520**, 692–696 (2015).
67. Abbas, A. K., Lichtman, A. H. & Pillai, S. *Cellular and Molecular Immunology* (Elsevier, 2014).
68. Nielsen, M. & Lund, O. *NN-align*. An artificial neural network-based alignment algorithm for MHC class II peptide binding prediction. *BMC Bioinformatics* **10**, 296 (2009).
69. Backert, L. & Kohlbacher, O. Immunoinformatics and epitope prediction in the age of genomic medicine. *Genome Med.* **7**, 119 (2015).
70. Gupta, S. K. *et al.* Personalized cancer immunotherapy using systems medicine approaches. *Brief. Bioinform.* **17**, 453–467 (2016).
71. Parker, K. C., Bednarek, M. A. & Coligan, J. E. Scheme for ranking potential HLA-A2 binding peptides based on independent binding of individual peptide side-chains. *J. Immunol.* **152**, 163–175 (1994).
72. Rammensee, H., Bachmann, J., Emmerich, N. P., Bachor, O. A. & Stevanović, S. SYFPEITHI: database for MHC ligands and peptide motifs. *Immunogenetics* **50**, 213–219 (1999).
73. Karosiene, E., Lundegaard, C., Lund, O. & Nielsen, M. *NetMHCcons*: a consensus method for the major histocompatibility complex class I predictions. *Immunogenetics* **64**, 177–186 (2012).
74. Moutaftsi, M. *et al.* A consensus epitope prediction approach identifies the breadth of murine T_{CD8+}-cell responses to vaccinia virus. *Nat. Biotechnol.* **24**, 817–819 (2006).
75. Vita, R. *et al.* The immune epitope database (IEDB) 3.0. *Nucleic Acids Res.* **43**, D405–D412 (2015).
76. Zhang, G. L., Lin, H. H., Keskin, D. B., Reinherz, E. L. & Brusic, V. Dana–Farber repository for machine learning in immunology. *J. Immunol. Methods* **374**, 18–25 (2011).
77. Hoof, I. *et al.* NetMHCpan, a method for MHC class I binding prediction beyond humans. *Immunogenetics* **61**, 1–13 (2009).
NetMHCpan predicts the binding affinity of peptides to class-I MHC molecules. It provides high-accuracy predictions for both well-annotated and novel alleles.
78. Andreatta, M. *et al.* Accurate pan-specific prediction of peptide-MHC class II binding affinity with improved binding core identification. *Immunogenetics* **67**, 641–650 (2015).
79. Trolle, T. *et al.* Automated benchmarking of peptide-MHC class I binding predictions. *Bioinformatics* **31**, 2174–2181 (2015).
80. Nielsen, M., Justesen, S., Lund, O., Lundegaard, C. & Buus, S. *NetMHCIIpan-2.0* — improved pan-specific HLA-DR predictions using a novel concurrent alignment and weight optimization training procedure. *Immunome Res.* **6**, 9 (2010).
81. Peters, B. *et al.* A community resource benchmarking predictions of peptide binding to MHC-I molecules. *PLoS Comput. Biol.* **2**, e65 (2006).
82. Wang, P. *et al.* A systematic assessment of MHC class II peptide binding predictions and evaluation of a consensus approach. *PLoS Comput. Biol.* **4**, e1000048 (2008).
83. Jørgensen, K. W., Rasmussen, M., Buus, S. & Nielsen, M. NetMHCstab — predicting stability of peptide-MHC-I complexes; impacts for cytotoxic T lymphocyte epitope discovery. *Immunology* **141**, 18–26 (2014).
84. Nielsen, M., Lundegaard, C., Lund, O. & Keşmir, C. The role of the proteasome in generating cytotoxic T-cell epitopes: insights obtained from improved predictions of proteasomal cleavage. *Immunogenetics* **57**, 33–41 (2005).
85. Bhasin, M. & Raghava, G. P. S. Pcleavage: an SVM based method for prediction of constitutive proteasome and immunoproteasome cleavage sites in antigenic sequences. *Nucleic Acids Res.* **33**, W202–W207 (2005).
86. Zhang, G. L., Petrovsky, N., Kwoh, C. K., August, J. T. & Brusic, V. PRED^{TAP}: a system for prediction of peptide binding to the human transporter associated with antigen processing. *Immunome Res.* **2**, 3 (2006).
87. Dönnes, P. & Kohlbacher, O. Integrated modeling of the major events in the MHC class I antigen processing pathway. *Protein Sci.* **14**, 2132–2140 (2005).
88. Tung, C.-W. & Ho, S. -Y. POPI: predicting immunogenicity of MHC class I binding peptides by mining informative physicochemical properties. *Bioinformatics* **23**, 942–949 (2007).
89. Tung, C. -W., Ziehm, M., Kämper, A., Kohlbacher, O. & Ho, S. -Y. POPISK: T-cell reactivity prediction using support vector machines and string kernels. *BMC Bioinformatics* **12**, 446 (2011).
90. Trolle, T. & Nielsen, M. NetTepi: an integrated method for the prediction of T cell epitopes. *Immunogenetics* **66**, 449–456 (2014).
91. Calis, J. J. A. *et al.* Properties of MHC class I presented peptides that enhance immunogenicity. *PLoS Comput. Biol.* **9**, e1003266 (2013).
92. Calis, J. J. A. & Rosenberg, B. R. Characterizing immune repertoires by high throughput sequencing: strategies and applications. *Trends Immunol.* **35**, 581–590 (2014).
93. Bolotin, D. A. *et al.* Next generation sequencing for TCR repertoire profiling: platform-specific features and correction algorithms. *Eur. J. Immunol.* **42**, 3073–3083 (2012).
94. Greiff, V., Miho, E., Menzel, U. & Reddy, S. T. Bioinformatic and statistical analysis of adaptive immune repertoires. *Trends Immunol.* **36**, 738–749 (2015).
95. Yaari, G. & Kleinstein, S. H. Practical guidelines for B-cell receptor repertoire sequencing analysis. *Genome Med.* **7**, 121 (2015).
96. Bolotin, D. A. *et al.* MiTCR: software for T-cell receptor sequencing data analysis. *Nat. Methods* **10**, 813–814 (2013).
97. Yu, Y., Ceredig, R. & Seoighe, C. LymAnalyzer: a tool for comprehensive analysis of next generation sequencing data of T cell receptors and immunoglobulins. *Nucleic Acids Res.* **44**, e31 (2015).
98. Bolotin, D. A. *et al.* MiXCR: software for comprehensive adaptive immunity profiling. *Nat. Methods* **12**, 380–381 (2015).
MiXCR is a tool for B cell receptor and TCR profiling, which is applicable to data from targeted receptor sequencing and non-targeted RNA-seq.
99. Shugay, M. *et al.* Towards error-free profiling of immune repertoires. *Nat. Methods* **11**, 653–655 (2014).
100. Yang, X. *et al.* TCRklass: a new K-string-based algorithm for human and mouse TCR repertoire characterization. *J. Immunol.* **194**, 446–454 (2015).
101. Kuchenbecker, L. *et al.* IMSEQ — a fast and error aware approach to immunogenetic sequence analysis. *Bioinformatics* **31**, 2963–2971 (2015).
102. Thomas, N., Heather, J., Ndifon, W., Shawe-Taylor, J. & Chain, B. Decombinator: a tool for fast, efficient gene assignment in T-cell receptor sequences using a finite state machine. *Bioinformatics* **29**, 542–550 (2013).
103. Ye, J., Ma, N., Madden, T. L. & Ostell, J. M. IgBLAST: an immunoglobulin variable domain sequence analysis tool. *Nucleic Acids Res.* **41**, W34–W40 (2013).
104. Gaëta, B. A. *et al.* iHMMune-align: hidden Markov model-based alignment and identification of germline genes in rearranged immunoglobulin gene sequences. *Bioinformatics* **23**, 1580–1587 (2007).
105. Li, S. *et al.* IMGT/HighV QUEST paradigm for T cell receptor IMGT clonotype diversity and next generation repertoire immunoprofiling. *Nat. Commun.* **4**, 2333 (2013).
106. Brown, S. D., Raeburn, L. A. & Holt, R. A. Profiling tissue-resident T cell repertoires by RNA sequencing. *Genome Med.* **7**, 125 (2015).
107. Han, A., Glanville, J., Hansmann, L. & Davis, M. M. Linking T-cell receptor sequence to functional phenotype at the single-cell level. *Nat. Biotechnol.* **32**, 684–692 (2014).
108. Giardine, B. *et al.* Galaxy: a platform for interactive large-scale genome analysis. *Genome Res.* **15**, 1451–1455 (2005).
109. Leek, J. T. svaseq: removing batch effects and other unwanted noise from sequencing data. *Nucleic Acids Res.* **42**, e161 (2014).
110. Larsen, M. V. *et al.* Large-scale validation of methods for cytotoxic T-lymphocyte epitope prediction. *BMC Bioinformatics* **8**, 424 (2007).
111. Schubert, B., Brachvogel, H.-P., Jürges, C. & Kohlbacher, O. EpiToolKit — a web-based workbench for vaccine design. *Bioinformatics* **31**, 2211–2213 (2015).
112. Schubert, B. *et al.* FRED 2: an immunoinformatics framework for python. *Bioinformatics* http://dx.doi.org/10.1093/bioinformatics/btw113 (2016).
113. Hundal, J. *et al.* pVAC-Seq: a genome-guided *in silico* approach to identifying tumor neoantigens. *Genome Med.* **8**, 11 (2016).
114. Gajewski, T. F. *et al.* Cancer immunotherapy strategies based on overcoming barriers within the tumor microenvironment. *Curr. Opin. Immunol.* **25**, 268–276 (2013).
115. Tumeh, P. C. *et al.* PD-1 blockade induces responses by inhibiting adaptive immune resistance. *Nature* **515**, 568–571 (2014).
116. Van Allen, E. M. *et al.* Genomic correlates of response to CTLA-4 blockade in metastatic melanoma. *Science* **350**, 207–211 (2015); erratum **352**, http://dx.doi.org/10.1126/science.aaf8264 (2016).
117. Tran, E. *et al.* Immunogenicity of somatic mutations in human gastrointestinal cancers. *Science* **350**, 1387–1390 (2015).
118. Walter, S. *et al.* Multipeptide immune response to cancer vaccine IMA901 after single-dose cyclophosphamide associates with longer patient survival. *Nat. Med.* **18**, 1254–1261 (2012).
119. Tran, E. *et al.* Cancer immunotherapy based on mutation-specific CD4 + T cells in a patient with epithelial cancer. *Science* **344**, 641–645 (2014).
120. Sharma, P. & Allison, J. P. Immune checkpoint targeting in cancer therapy: toward combination strategies with curative potential. *Cell* **161**, 205–214 (2015).

121. Cox, D. B. T., Platt, R. J. & Zhang, F. Therapeutic genome editing: prospects and challenges. *Nat. Med.* **21**, 121–131 (2015).
122. Chattopadhyay, P. K., Gierahn, T. M., Roederer, M. & Love, J. C. Single-cell technologies for monitoring immune systems. *Nat. Immunol.* **15**, 128–135 (2014).
123. Kononen, J. *et al.* Tissue microarrays for high-throughput molecular profiling of tumor specimens. *Nat. Med.* **4**, 844–847 (1998).
124. Giesen, C. *et al.* Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry. *Nat. Methods* **11**, 417–422 (2014).
125. Bendall, S. C. & Nolan, G. P. From single cells to deep phenotypes in cancer. *Nat. Biotechnol.* **30**, 639–647 (2012).
126. Qiu, P. *et al.* Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE. *Nat. Biotechnol.* **29**, 886–891 (2011).
127. Amir, E. D. *et al.* viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. *Nat. Biotechnol.* **31**, 545–552 (2013).
128. Diggins, K. E., Ferrell, P. B. & Irish, J. M. Methods for discovery and characterization of cell subsets in high dimensional mass cytometry data. *Methods* **82**, 55–63 (2015).
129. Levine, J. H. *et al.* Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis. *Cell* **162**, 184–197 (2015).
130. Gaujoux, R. & Seoighe, C. CellMix: a comprehensive toolbox for gene expression deconvolution. *Bioinformatics* **29**, 2211–2212 (2013).
131. Yuan, Y. *et al.* Quantitative image analysis of cellular heterogeneity in breast tumors complements genomic profiling. *Sci. Transl. Med.* **4**, 157ra143 (2012).
132. Repsilber, D. *et al.* Biomarker discovery in heterogeneous tissue samples -taking the in-silico deconfounding approach. *BMC Bioinformatics* **11**, 27 (2010).
133. Tuominen, V. J., Ruotoistenmäki, S., Viitanen, A., Jumppanen, M. & Isola, J. ImmunoRatio: a publicly available web application for quantitative image analysis of estrogen receptor (ER), progesterone receptor (PR), and Ki-67. *Breast Cancer Res.* **12**, R56 (2010).
134. Liebner, D. A., Huang, K. & Parvin, J. D. MMAD: microarray microdissection with analysis of differences is a computational tool for deconvoluting cell type-specific contributions from tissue samples. *Bioinformatics* **30**, 682–689 (2014).
135. Qiao, W. *et al.* PERT: a method for expression deconvolution of human blood samples from varied microenvironmental and developmental conditions. *PLoS Comput. Biol.* **8**, e1002838 (2012).
136. Bolen, C. R., Uduman, M. & Kleinstein, S. H. Cell subset prediction for blood genomic studies. *BMC Bioinformatics* **12**, 258 (2011).
137. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).
138. Ramos, A. H. *et al.* Oncotator: cancer variant annotation tool. *Hum. Mutat.* **36**, E2423–E2429 (2015).
139. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w^{1118}; iso-2; iso-3. *Fly* **6**, 80–92 (2012).
140. Larson, D. E. *et al.* SomaticSniper: identification of somatic point mutations in whole genome sequencing data. *Bioinformatics* **28**, 311–317 (2012).
141. Saunders, C. T. *et al.* Strelka: accurate somatic small-variant calling from sequenced tumor–normal sample pairs. *Bioinformatics* **28**, 1811–1817 (2012).
142. Zhou, W. *et al.* TransVar: a multilevel variant annotator for precision genomics. *Nat. Methods* **12**, 1002–1003 (2015).
143. 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).
144. McLaren, W. *et al.* Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* **26**, 2069–2070 (2010).
145. Boegel, S., Löwer, M., Bukur, T., Sahin, U. & Castle, J. C. A catalog of HLA type, HLA expression, and neo-epitope candidates in human cancer cell lines. *Oncoimmunology* **3**, e954893 (2014).
146. Lundegaard, C. *et al.* NetMHC-3.0: accurate web accessible predictions of human, mouse and monkey MHC class I affinities for peptides of length 8–11. *Nucleic Acids Res.* **36**, W509–W512 (2008).
147. Nielsen, M. & Andreatta, M. NetMHCpan-3.0; improved prediction of binding to MHC class I molecules integrating information from multiple receptor and peptide length datasets. *Genome Med.* **8**, 33 (2016).
148. Zhang, H., Lund, O. & Nielsen, M. The PickPocket method for predicting binding specificities for receptors based on receptor pocket similarities: application to MHC-peptide binding. *Bioinformatics* **25**, 1293–1299 (2009).
149. Stranzl, T., Larsen, M. V., Lundegaard, C. & Nielsen, M. *NetCTLpan*: pan-specific MHC class I pathway epitope predictions. *Immunogenetics* **62**, 357–368 (2010).
150. Barretina, J. *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603–607 (2012).
151. Hess, J. L. The Cancer Genome Anatomy Project: power tools for cancer biologists. *Cancer Invest.* **21**, 325–326 (2003).
152. Goldman, M. *et al.* The UCSC Cancer Genomics Browser: update 2013. *Nucleic Acids Res.* **41**, D949–D954 (2013).
153. Forbes, S. A. *et al.* COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res.* **43**, D805–D811 (2015).
154. Cerami, E. *et al.* The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov.* **2**, 401–404 (2012).
155. Ellis, M. J. *et al.* Connecting genomic alterations to cancer biology with proteomics: the NCI Clinical Proteomic Tumor Analysis Consortium. *Cancer Discov.* **3**, 1108–1112 (2013).
156. Gonzalez-Perez, A. *et al.* IntOGen-mutations identifies cancer drivers across tumor types. *Nat. Methods* **10**, 1081–1082 (2013).
157. Uhlén, M. *et al.* Proteomics. Tissue-based map of the human proteome. *Science* **347**, 1260419 (2015).
158. Toseland, C. P. *et al.* AntiJen: a quantitative immunology database integrating functional, thermodynamic, kinetic, biophysical, and cellular data. *Immunome Res.* **1**, 4 (2005).
159. Grimes, G. R. *et al.* GPX-Macrophage Expression Atlas: a database for expression profiles of macrophages challenged with a variety of pro-inflammatory, anti-inflammatory, benign and pathogen insults. *BMC Genomics* **6**, 178 (2005).
160. Watkins, N. A. *et al.* A HaemAtlas: characterizing gene expression in differentiated human blood cells. *Blood* **113**, e1–9 (2009).
161. Reche, P. A., Zhang, H., Glutting, J.-P. & Reinherz, E. L. EPIMHC: a curated database of MHC-binding peptides for customized computational vaccinology. *Bioinformatics* **21**, 2140–2141 (2005).
162. Robinson, J. *et al.* The IMGT/HLA database. *Nucleic Acids Res.* **41**, D1222–D1227 (2013).
163. Bhattacharya, S. *et al.* ImmPort: disseminating data to the public for the future of immunology. *Immunol. Res.* **58**, 234–239 (2014).
164. Breuer, K. *et al.* InnateDB: systems biology of innate immunity and beyond—recent updates and continuing curation. *Nucleic Acids Res.* **41**, D1228–D1233 (2013).
165. Vigneron, N., Stroobant, V., Van den Eynde, B. J. & van der Bruggen, P. Database of T cell-defined human tumor antigens: the 2013 update. *Cancer Immun.* **13**, 15 (2013).

Acknowledgements

The authors apologize to fellow researchers not cited owing to restrictions. This work was supported by the Tyrolean Standortagentur (Project Bioinformatics Tyrol), Jubiläumsfonds der Österreichischen Nationalbank (Project 16534) and the Horizon2020 project No 633592 APERIM (Advanced Bioinformatics Platform for Personalized Cancer Immunotherapy).

Competing interests statement

The authors declare no competing interests.

FURTHER INFORMATION

CIBERSORT: http://cibersort.stanford.edu
CTdatabase: http://www.cta.lncc.br
DREAM Challenges: http://dreamchallenges.org
GSEA: http://broadinstitute.org/gsea
IMGT: http://imgt.org
Immunological Genome Project (Immgen): http://immgen.org
Immune Epitope Database and Analysis Resource: http://www.iedb.org
International Cancer Genome Consortium: http://icgc.org
MiXCR: http://mixcr.milaboratory.com
NetMHCcon: http://cbs.dtu.dk/services/NetMHCcons
POLYSOLVER: http://broadinstitute.org/cancer/cga/polysolver
RCSB Protein Data Bank: http://www.rcsb.org/pdb/home/home.do
The Cancer Genome Atlas: http://cancergenome.nih.gov
The Cancer Immunome Atlas: http://tcia.at

ALL LINKS ARE ACTIVE IN THE ONLINE PDF