A Novel Neoantigen Discovery Approach based on Chromatin High Order Conformation: Mapping the Neoantigen to 3D Genome

Xueying Shang[+1], Mingxuan Zhang[+2], Luming Meng[+3], Xianbin Su[+1], Qingjiao Li[+4,5], Mengna Lin[1], Xin Zou[1], Qing Luo[1], Yangyang Zhai[1], Kunyan He[1], Lan Wang[1], Lintai Da*, Tom Weidong Cai[6]*, Ze-guang Han[1]*, Yi Shi[+*1]

1. *Key Laboratory of Systems Biomedicine, Ministry of Education. Shanghai Center for Systems Biomedicine, Shanghai Jiaotong University, Shanghai 200240, China.*
2. *University of California at San Diego, California, 9500 Gilman Dr. La Jolla, CA 92093, USA.*
3. *College of Biophotonics, South China Normal University, Guangzhou 510631, China.*
4. *The Eighth Affiliated Hospital, Sun Yat-Sen University, Shenzhen 518033, China.*
5. *University of California at Los Angeles, Los Angeles, CA 90095-1732, USA.*
6. *University of Sydney, NSW 2006, Australia.*

Abstract

The high-throughput sequencing technology has yielded reliable and ultra-fast sequencing for DNA and RNA. For tumor cells of cancer patients, when combining the results of DNA and RNA sequencing, one can identify potential neoantigens that stimulate immune response of the T cells. However, when the somatic mutations are abundant, it is computationally challenging to efficiently prioritize the identified neoantigen candidates according to their ability of activating the T cell immuno-response. Numerous prioritization or prediction approaches have been proposed to address this issue but none of them considers the original DNA loci of the neoantigens from the perspective of 3D genome. Here we retrospect the DNA origins of the immune-positive and non-negative neoantigens in the context of 3D genome and discovered that 1) DNA loci of the immuno- positive neoantigens tend to cluster genome-wise. 2) DNA loci of the immuno- positive neoantigens tend to belong to active chromosomal compartment (compartment A) in some chromosomes. 3). DNA loci of the immuno- positive neoantigens tend to locate at specific regions in the 3D genome. We believe that the 3D genome information will help more precise neoantigen prioritization and discovery and eventually benefit precision and personalized medicine in cancer immunotherapy.

[+]: These authors contributed equally as co-first authors.

*: Corresponding authors. yishi@sjtu.edu.cn, darlt@sjtu.edu.cn, tom.cai@sydney.edu.au, hanzg@sjtu.edu.cn.

Introduction

In a variety of human malignancies, immunotherapies via boosting the endogenous T cell ability in destroying cancer cells have demonstrated therapeutic efficacy[1]. Based on clinical practices in a substantial fraction of patients, the inference of endogenous T cell with mounted cancer-killing ability is that the T cell receptor (TCR) is able to recognize peptide epitopes that are displayed on major histocompatibility complexes (MHCs) on the surface of the tumor cells. These cancer rejection epitopes may be derived from two origins: the first origin of potential cancer rejection antigens is formed by non-mutated proteins to which T cell tolerance is incomplete for instance, because of their restricted tissue expression pattern; the second origin of potential cancer rejection antigens is formed by peptides that cannot be found from the normal human genome, so-called neoantigens[1]. With the development of genome sequencing, it has been revealed that during cancer initiation and progression, tens to thousands of different somatic mutations are generated. Most of these mutations are passenger mutations, meaning no obvious growth advantage, and are often caused by genomic instability within the tumor cells. A limited number of cancer mutations are driver mutations which interfere with normal cell regulation and help to drive cancer growth and resistance to targeted therapies[2]. Both passenger mutations and driver mutations can be nonsynonymous that alter protein coding sequences, causing tumor to express abnormal proteins that cannot be found in normal cells. When cell metabolize, the proteins possessing abnormal sequences are cut into short peptides, namely epitopes, and are presented on the cell surface by the major histocompatibility complex (MHC, or human leukocyte antigen (HLA) in humans) which have a chance to be recognizable by T cells as foreign antigens[2].

According to the discoveries mentioned above, in theory therefore, if the potential neoantigens can be identified via sequencing technology, one can synthesize epitope peptides in vitro and validate their efficacy in vivo (cancer cell-line or in mouse model) before clinical practice[1,2]. Indeed, cancers with a single dominant mutation

can often be treated effectively by targeting the dominant driver mutation[2,3]. However, when the somatic mutations are abundant, which is the case in most cancer types, it is computationally challenging to efficiently prioritize the identified neoantigen candidates according to their ability of activating the T cell immuno-response[4]. Over the last decades, numerous neoantigen prediction approaches have been proposed to address this issue[5-7]. These approaches can be classified into two major categories: the protein 3D structure-based approaches which considers the pMHC and TCR 3D conformation, and the protein sequence-based approaches which consider the amino acid sequence of protein antigens. For the protein 3D structure-based approaches, in some specific cases when high quality pMHC 3D structures are available, molecular dynamic (MD) methods are used to explore the contact affinity of pMHC-TCR complex[8-10], in most cases however, the modelling or simulation by protein docking and threading has to be used due to the lack of high quality pMHC 3D structures. Most approaches belong to the sequence-based category as there are much larger data sets for training and validation[11,12] and they are usually very efficient to set up[4,13].

Early sequence-based methods relied on position-specific scoring matrices (PSSMs), such as BIMAS[14] and SYFPEITHI[15], in which the PSSMs are defined from experimentally confirmed peptide binders of a particular MHC allele[4]. Later, more advanced methods based on machine-learning techniques have been developed to capture and utilize the nonlinear nature of the pMHC-TCR interaction which indeed demonstrated better performance than the PSSM-based methods. Consensus methods that combine multiple tools to obtain more reliable predictions were also developed, such as CONSENSUS[16] and NetMHCcons[17], which demonstrated better performances; these methods. However, the performance gain is determined by the weighting scheme which cost increased computational power. When considering peptide binding, the large majority of HLA alleles has not been investigated, therefore, there are pan-specific methods, such as NetMHCpan[6,7], are developed which allow the HLA type independent prioritization of neoantigen.

As one of the widely adopted practices in neoantigen prioritization, NetMHCpan first train a neural network is based on multiple public datasets, and the affinity of a given peptide-MHC considering the polymorphic HLA types HLA-A, HLA-B or HLA-C is then computed according to the trained neural network. NetMHCpan[7] and NetMHCIIpan[18] performance remarkably, even compared to allele-specific approaches[4,19]. However, although several assessments and criteria were proposed in the past aiming at a more fair and effective comparison[19-21], there are no recent independent benchmark studies that can be used to recommend specific tools up to now. More importantly, to the best of our knowledge however, none of the neoantigen prediction methods mentioned above consider the mutation DNA loci of the neoantigens in the perspective of 3D genome, which carries much richer information comparing to the amino acid sequence alone[22]. In this work, we retrospect the DNA origin of the immune-positive and non-negative neoantigens in the context of 3D genome and demonstrate some interesting discoveries.

Results

Neoantigen proximity in individual chromosome (Intra-chromosome)

We generated all peptide pairs between immune-positive peptides and peptide pairs between immune-negative peptides. Then on each chromosome (intra-chromosomal), we generate each pair's contact frequency on IMR90 and hESC Hi-C data. The results are shown in Table 1 and Table 2. Jointly from these results, we found that positive peptides' corresponding DNA loci tend to be more proximate than the negative ones on chr1, chr7, chr10, and chr12, while negative peptides' corresponding DNA loci tend to be more proximate than the positive ones on chromosome chr2, chr5, chr8, chr11, and chr20.

Table 1. Average contact frequency contact of immune-positive peptide pairs and

immue-negative peptide pairs based on IMR90 Hi-C data.

Chromosome	P-Value	Mean CF of Positive Pairs	Mean CF of Negative Pairs
Chr1	0.000041	0.6307	0.5718
Chr2	0.007454	0.4701	0.5241
Chr3	0.152573	0.5846	0.6205
Chr4	0.22844	0.6488	0.603
Chr5	0.012444	0.6043	0.6654
Chr6	0.784581	0.6206	0.6111
Chr7	0.01599	0.6703	0.6013
Chr8	0.010156	0.5056	0.5824
Chr9	0.157004	0.5969	0.6418
Chr10	0.005747	0.6347	0.5588
Chr11	0.000000	0.52	0.6248
Chr12	0.047414	0.5821	0.5483
Chr13	0.191002	0.6046	0.5502
Chr14	0.451133	0.5925	0.5612
Chr15	0.653759	0.5629	0.5524
Chr16	0.309672	0.5968	0.6353
Chr17	0.014733	0.5697	0.6116
Chr18	0.017744	0.2984	0.5135
Chr19	0.351049	0.5415	0.5665
Chr20	0.047292	0.5014	0.5782
Chr21	0.268375	0.6534	0.5861
Chr22	0.051617	0.5323	0.4617
ChrX	0.755006	0.6159	0.602

Table 2. Average contact frequency contact of immune-positive peptide pairs and immue-negative peptide pairs based on hESC Hi-C data.

Chromosome	P-Value	Mean CF of Positive Pairs	Mean CF of Negative Pairs
Chr1	0.010864	0.5994	0.5703
Chr2	0.019324	0.5137	0.5534
Chr3	0.458643	0.6309	0.6163
Chr4	0.273972	0.5506	0.5846
Chr5	0.000000	0.552	0.6725
Chr6	0.003641	0.6982	0.6065
Chr7	0.01784	0.6729	0.6168
Chr8	0.003467	0.5096	0.589
Chr9	0.036334	0.5384	0.6012
Chr10	0.000145	0.6374	0.5465

Chr11	0.000464	0.5516	0.6122
Chr12	0.000016	0.6056	0.5428
Chr13	0.009135	0.6481	0.541
Chr14	0.352111	0.5679	0.5347
Chr15	0.54664	0.5363	0.5484
Chr16	0.03483	0.5704	0.6363
Chr17	0.104837	0.5769	0.6022
Chr18	0.190727	0.4432	0.5554
Chr19	0.005872	0.4957	0.5549
Chr20	0.007596	0.4936	0.5771
Chr21	0.688848	0.5817	0.5618
Chr22	0.000278	0.5876	0.4652
Chr23	0.175523	0.604	0.5702

Neoantigen proximity in the whole genome (Inter-chromosome)

For the inter-chromosomal peptide pairs, both positive and negative, we also collect their contact frequency and calculate the average values. As shown in Figure 1, on both hESC and IMR90 Hi-C data, immune-positive peptide pairs are more proximate to each other comparing to immune-negative peptide pairs. The corresponding P-values are close to zero's.

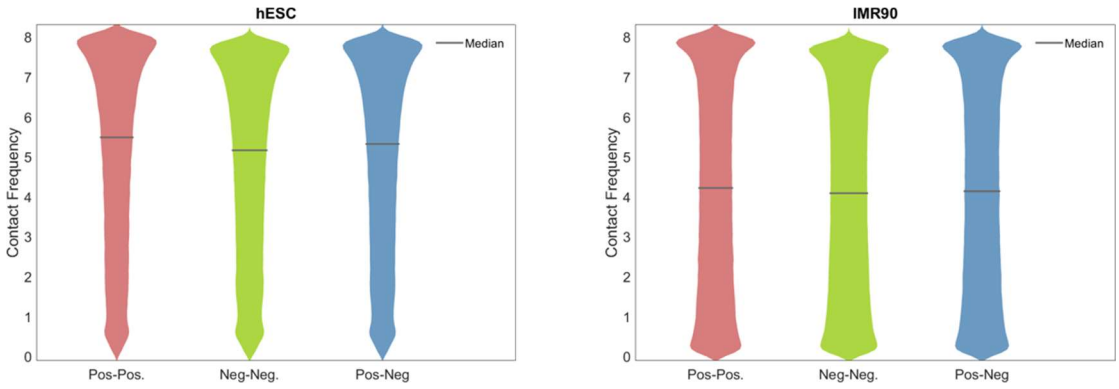

Figure 1. Contact frequency distribution comparison between immune-positive and immune-negative peptide pairs on hESC and IMR90 Hi-C data.

Neoantigen distribution on active and inactive compartment

For each chromosome, we compute the compartment type (A or B) for each chromosomal region (bin), shown in Figure 2A and 2B. Then we assign positive and negative peptides with their corresponding A/B compartment type. We found that in some chromosome, immune-positive neoantigens tend to be located on compartment A, comparing to immune-negative neoantigens, as shown in Figure 2C.

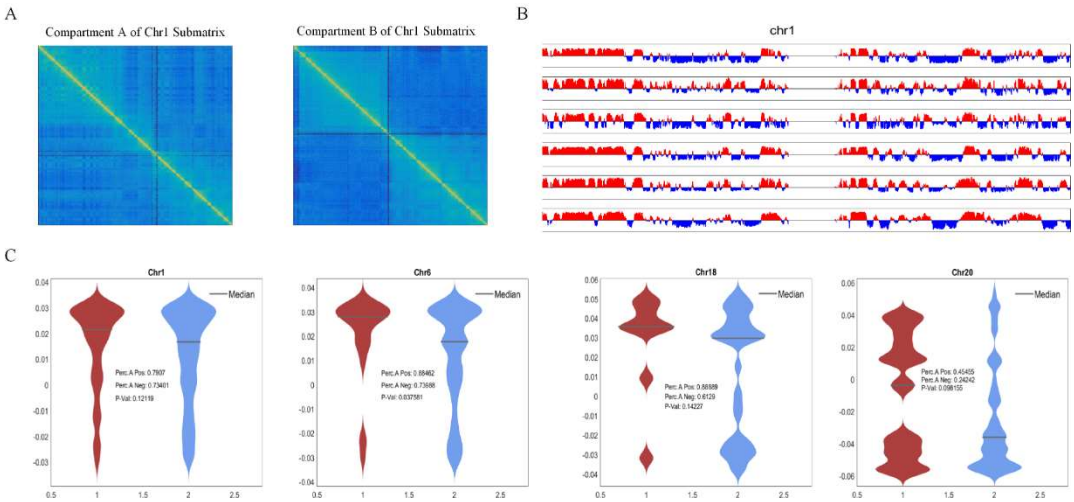

Figure 2. Neoantigen distribution on active and inactive compartment. (A) Example of compartment A and compartment B submatrix on chromosome 1. (B) Example of compartment A/B values on chromosome 1. (C) Distribution of percentage of compartment A of immune-positive and immune-negative peptides.

The radius position distribution of neoantigen

We developed a novel molecular dynamic based approach to model the 3D conformation of the human genome, on both hESC and IMR90 Hi-C data. We then map the positive and negative peptides' corresponding chromosomal loci on the constructed 3D genome and calculate their radius distance to the nucleus center, as shown in Figure 3A. We found that immune-positive peptide's corresponding loci tend to locate closer to the nuclear periphery, comparing to the negative ones, as Figure 3B demonstrates. We then used the radius position as the immunogenicity predictor and found that surprisingly, this feature along can discriminate the

immune-positive peptides from the immune-negative peptides, as shown in Figure 3C.

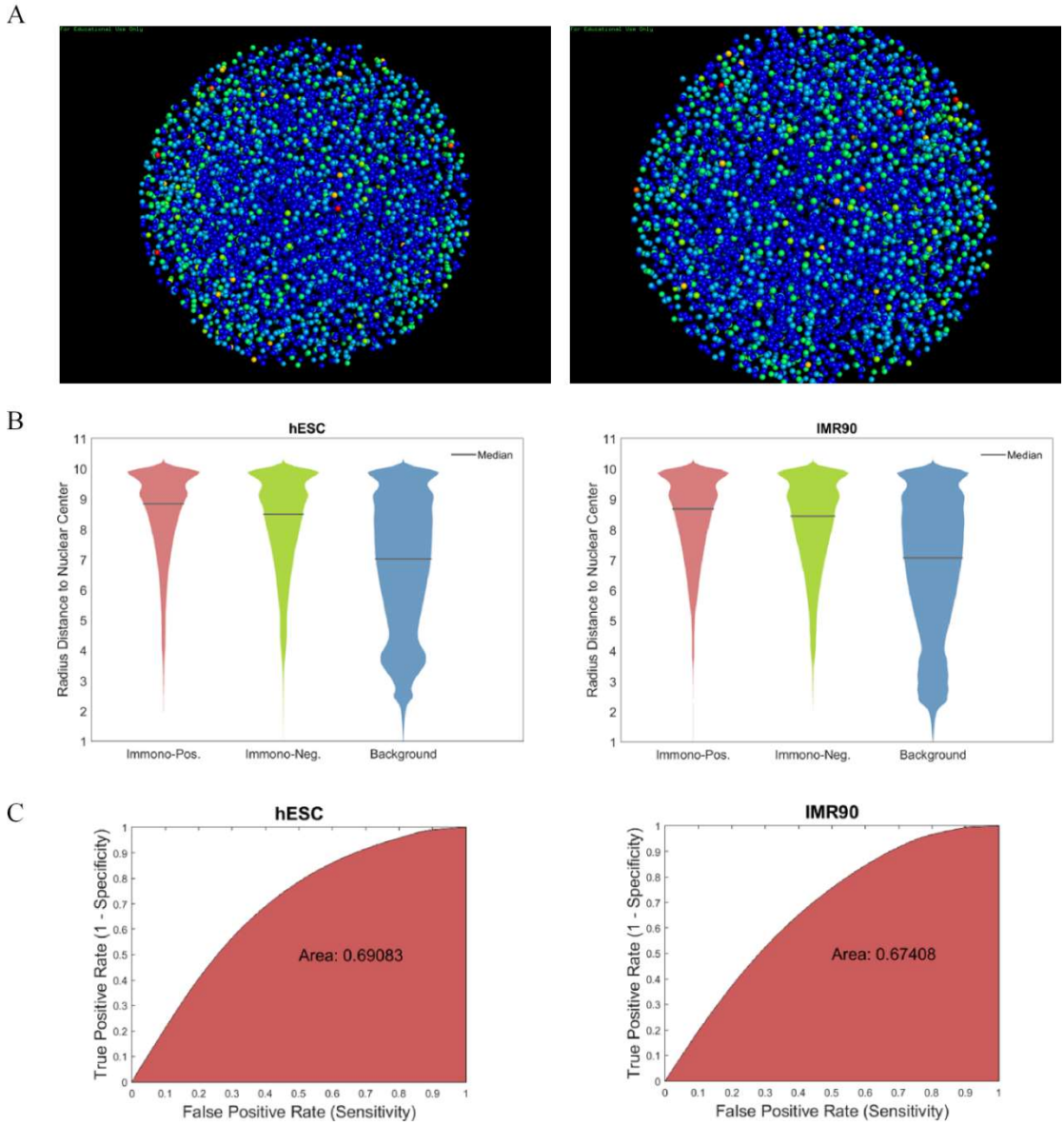

Figure 3. The radius position distribution of neoantigen. (A) Positive and negative peptides' corresponding chromosomal loci on 3D genome structure. (B) Radius position distribution of the positive peptides comparing to the negative peptides. (C) ROC curve demonstrating the discriminatal power of radius position in immunogenicity prediction.

Methods

Data Collection and curation

All the peptide sequences and their corresponding immune effectiveness were collected from IEDB (T Cell Assay)[12] on May 27th 2018; the raw dataset contains 337'248 peptide records. We then narrow down to Homo Sapiens and MHC-I subtypes and then further restrain the AA length equals to 9 with duplicated peptide merged. Finally, we obtained 3'909 qualified records with 809 immuno-positive peptides and 3'100 immuno-negative peptides that can be mapped in human hg19 reference genome. Note that for identical peptides with multiple immune experiments, we define positive rate > 0.8 as immune-positive peptides and positive rate < 0.2 as immune-negative peptides. In detail, there are 2 steps in this procedure:

Step I. Extracting the Homo Sapiens peptide sequences and cleaning up the dataset from initial dataset. We use PANDAS library to create an data frame object for processing. Then we assign the column name by passing in a name dictionary. Then we filter the dataset so that the only entries left have "homo sapiens" as their hostname, after which we transfer the dataset to a tsv file for the future procedures. Then, we further clean up the dataset by introducing two helper functions: 1. Letter_check takes in a string as the parameter and check for letters that are not in the alphabet for a proper peptide sequence. The function returns an indicator which value is set to true if the string is a legal peptide sequence. 2. Drop_legal takes a data frame and iterate through and delete the rows with illegal strings. With these helper functions, we can remove all the illegal entries in the dataset.

Step II: Counting the number of sequences that have positive qualitative measure the algorithm here takes advantage of the layout of the dataset. We keep counting until the last appearance of a certain sequence, add 1 to the counter of positivity if a positive qualitative measure is detected. For the last appearance of the sequence we can either add 1 if it's positive or we skip to the next step. The counter resets every

time we finish counting positivity for a sequence and move on to the next one. We store the counted values into a dictionary that have sequence and MHC as its key.

For the chromatin 3D conformation data, we used the Hi-C data of IMR90 and hESC cell lines generated by Bin Ren's lab[23]. The contact frequencies and the subsequent chromatin 3D modeling are based on these Hi-C data.

Mapping peptides to human genome

To map the peptides to human genome (hg19), we wrote a pipeline to query BLAST[24] web server and map the gene names to chromosome and starting positions the algorithm first divides the dataset to 711 folds where each fold has 100 sequences for BLAST server to process. To set up the BLAST search, we regulate the searching algorithm to search for Homo Sapien matches only with entrez keywords and use the PAM30 matrix to search for matches. We also adjust the gap costs to regulate gap penalty. After the setup, we call BLAST iteratively and write the result onto a tsv file. For each match, we get the first hit and save the accession and raw bit score. Now having the accessions, we upload a list of refseq id to the DAVID tool[25] to obtain the gene names composed with gene symbols. The algorithm mapping gene names to chromosome positions, we start with a dataset that records chromosome positions and gene names for numerous genes as our database. To save the iteration time, we create two dictionaries recording chromosome positions with gene names as keys, one from the dataset we produced from BLAST results and one from the database. We iterate through the dictionaries simultaneously. If we find a match for the keys, we record the chromosome positions on the result file. The final is in the form of a tuple that contains peptides, HLA subtype, chromosome number, and chromosome position.

Active and inactive compartment determination

To determine the compartment activeness (compartment A: active, compartment B: inactive) of each chromosome bin, we used individual chromosome Hi-C contact maps. We first diagonal normalized each contact map by dividing contact frequencies by their corresponding off-diagonal mean. Then we compute the Pearson correlation coefficient (PCC) matrices for each chromosome and the compartment type is jointly determined by the sign of eigenvalue of the first eigenvector of the PCC matrices and the signal of the epigenetic marker H3k4me1.

Chromatin 3D modeling

We adopted our previous chromatin 3D modeling method for individual chromosome using molecular dynamic (MD) based approach[26] with resolution 40kb (bin size) for the IMR90 and hESC Hi-C data. For the whole genome 3D modeling, we used 500kb resolution genome-wise Hi-C contact map of IMR90 and hESC and extended and re-developed our previous individual chromosomal modeling method to be able to handle the whole genome data. Once we have the genome-wise 3D structure, we calculate the radius distance of each bin to the center of the nucleus to obtain radius position of each neoantigen peptide.

Discussion

In cancer immune therapy, the neoantigen therapy is a rising and promising topic as it can be genuinely personalized and precise. However, when the somatic mutations are abundant, it is computationally hard to efficiently prioritize the identified neoantigen candidates according to their ability of activating the T cell immuno-response and numerous prioritization or prediction approaches have been proposed to address this issue. However, none of existing approaches considers the original DNA loci of the

neoantigens in the 3D genome perspective, to the best of our knowledge. Here, we retrospect the DNA origin of the immune-positive and non-negative neoantigens in the context of 3D genome and discovered that 1) DNA loci of the immuno- positive neoantigens tend to cluster genome-wise. 2) DNA loci of the immuno- positive neoantigens tend to belong to active chromosomal compartment (compartment A) in some chromosomes. 3). DNA loci of the immuno- positive neoantigens tend to locate at specific regions in the 3D genome. We believe that the 3D genome information will help more precise neoantigen prioritization and discovery and eventually benefit precision and personalized medicine in cancer immunotherapy.

References

1 Schumacher, T. N. & Schreiber, R. D. Neoantigens in cancer immunotherapy. *Science* **348**, 69-74, doi:10.1126/science.aaa4971 (2015).

2 Yarchoan, M., Johnson, B. A., 3rd, Lutz, E. R., Laheru, D. A. & Jaffee, E. M. Targeting neoantigens to augment antitumour immunity. *Nat Rev Cancer* **17**, 209-222, doi:10.1038/nrc.2016.154 (2017).

3 O'Brien, S. G. *et al.* Imatinib compared with interferon and low-dose cytarabine for newly diagnosed chronic-phase chronic myeloid leukemia. *N Engl J Med* **348**, 994-1004, doi:10.1056/NEJMoa022457 (2003).

4 Hackl, H., Charoentong, P., Finotello, F. & Trajanoski, Z. Computational genomics tools for dissecting tumour-immune cell interactions. *Nat Rev Genet* **17**, 441-458, doi:10.1038/nrg.2016.67 (2016).

5 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**, W509-512, doi:10.1093/nar/gkn202 (2008).

6 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, doi:10.1186/s13073-016-0288-x (2016).

7 Jurtz, V. *et al.* NetMHCpan-4.0: Improved Peptide-MHC Class I Interaction Predictions Integrating Eluted Ligand and Peptide Binding Affinity Data. *J Immunol* **199**, 3360-3368, doi:10.4049/jimmunol.1700893 (2017).

8 Blevins, S. J. *et al.* How structural adaptability exists alongside HLA-A2 bias in the human alpha beta TCR repertoire. *P Natl Acad Sci USA* **113**, E1276-E1285, doi:10.1073/pnas.1522069113 (2016).

9 Riley, T. P. *et al.* T cell receptor cross-reactivity expanded by dramatic peptide-MHC adaptability. *Nat Chem Biol* **14**, 934-+, doi:10.1038/s41589-018-0130-4 (2018).

10 Wang, Y. *et al.* How an alloreactive T-cell receptor achieves peptide and MHC specificity. *P*

Natl Acad Sci USA **114**, E4792-E4801, doi:10.1073/pnas.1700459114 (2017).

11 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, doi:10.1016/j.jim.2011.07.007 (2011).

12 Vita, R. *et al.* The Immune Epitope Database (IEDB): 2018 update. *Nucleic Acids Res* **47**, D339-D343, doi:10.1093/nar/gky1006 (2019).

13 Gupta, S. K. *et al.* Personalized cancer immunotherapy using Systems Medicine approaches. *Brief Bioinform* **17**, 453-467, doi:10.1093/bib/bbv046 (2016).

14 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).

15 Schuler, M. M., Nastke, M. D. & Stevanovikc, S. SYFPEITHI: database for searching and T-cell epitope prediction. *Methods Mol Biol* **409**, 75-93 (2007).

16 Moutaftsi, M. *et al.* A consensus epitope prediction approach identifies the breadth of murine TCD8+-cell responses to vaccinia virus. *Nat Biotechnol* **24**, 817-819, doi:DOI 10.1038/nbt1215 (2006).

17 Karosiene, E., Lundegaard, C., Lund, O. & Nielsen, M. NetMHCcons: a consensus method for the major histocompatibility complex class I predictions. *Immunogenetics* **64**, 177-186, doi:10.1007/s00251-011-0579-8 (2012).

18 Karosiene, E. *et al.* NetMHCIIpan-3.0, a common pan-specific MHC class II prediction method including all three human MHC class II isotypes, HLA-DR, HLA-DP and HLA-DQ. *Immunogenetics* **65**, 711-724, doi:10.1007/s00251-013-0720-y (2013).

19 Trolle, T. *et al.* Automated benchmarking of peptide-MHC class I binding predictions. *Bioinformatics* **31**, 2174-2181, doi:10.1093/bioinformatics/btv123 (2015).

20 Peters, B. *et al.* A community resource benchmarking predictions of peptide binding to MHC-I molecules. *PLoS Comput Biol* **2**, e65, doi:10.1371/journal.pcbi.0020065 (2006).

21 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, doi:10.1371/journal.pcbi.1000048 (2008).

22 Shi, Y. *et al.* Chromatin accessibility contributes to simultaneous mutations of cancer genes. *Sci Rep* **6**, 35270, doi:10.1038/srep35270 (2016).

23 Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376-380, doi:10.1038/nature11082 (2012).

24 Boratyn, G. M. *et al.* BLAST: a more efficient report with usability improvements. *Nucleic Acids Res* **41**, W29-33, doi:10.1093/nar/gkt282 (2013).

25 Huang, D. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **4**, 44-57, doi:10.1038/nprot.2008.211 (2009).

26 Xie, W. J. *et al.* Structural Modeling of Chromatin Integrates Genome Features and Reveals Chromosome Folding Principle. *Sci Rep* **7**, 2818, doi:10.1038/s41598-017-02923-6 (2017).