Subject Section

DeepAntigen: A Novel Method for Neoantigen Prioritization via 3D Genome and Deep Sparse Learning

Yi Shi [1,2,3#*], Zehua Guo[4,2#], Xianbin Su[1#], Luming Meng[5*], Mingxuan Zhang[6], Jing Sun[7], Chao Wu[7], Minhua Zheng[7], Xueyin Shang[1], Xin Zou[1], Wangqiu Cheng[2,3], Yaoliang Yu[8], Yujia Cai[1], Chaoyi Zhang[9], Weidong Cai[9], Lintai Da[1*], Guang He[2,3*], Ze-Guang Han[1*]

[1]Key Laboratory of Systems Biomedicine (Ministry of Education), Shanghai Centre for Systems Biomedicine, Shanghai Jiao Tong University, Shanghai, 200240, China. [2]Bio-X Institutes, Key Laboratory for the Genetics of Developmental and Neuropsychiatric Disorders, Shanghai Jiao Tong University, 1954 Huashan Road, Shanghai, 200030, China. [3]Shanghai Key Laboratory of Psychotic Disorders, and Brain Science and Technology Research Center, Shanghai Jiao Tong University, 1954 Huashan Road, Shanghai, 200030, China. [4]Department of Instrument Science and Engineering, School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai, 200240, China. [5]College of Biophotonics, South China Normal University, Guangzhou, 510631, China. [6] Department of Mathematics, University of California San Diego, La Jolla, CA, 92093-0112, USA. [7] Department of General Surgery & Shanghai Minimally Invasive Surgery Center, Ruijin Hospital, Shanghai Jiao Tong University, Shanghai, 200025, China. [8]David R. Cheriton School of Computer Science, University of Waterloo, Waterloo, ON, N2L3G1, Canada. [9]School of Computer Science, The University of Sydney, Darlington, NSW, 2008, Australia.

[#]The authors contribute equally to this work.

*To whom correspondence should be addressed.

Abstract

Motivation: The mutations of cancers can encode the seeds of their own destruction, in the form of T cell recognizable immunogenic peptides, also known as neoantigens. It is computationally challenging however, to accurately prioritize the potential neoantigen candidates according to their ability of activating the T cell immuno-response, especially when the somatic mutations are abundant. Although a few neoantigen prioritization methods have been proposed to address this issue, advanced machine learning model that is specifically designed to tackle this problem is still lacking. Moreover, none of the existing methods considers the original DNA loci of the neoantigens in the perspective of 3D genome which may provide key information for inferring neoantigens' immunogenicity.

Results: In this study, we discovered that DNA loci of the immuno-positive and immuno-negative MHC-I neoantigens have distinct spatial distribution patterns across the genome. We therefore employed the 3D genome information along with an ensemble pMHC-I coding strategy, and developed a group feature selection based deep sparse neural network model (DNN-GFS) that is optimized for neoantigen prioritization. DNN-GFS demonstrated increased neoantigen prioritization power comparing to existing sequence-based approaches. We also developed a webserver named deepAntigen (http://yishi.sjtu.edu.cn/deepAntigen) that implements the DNN-GFS as well as other machine learning methods. We believe that this work provides a new perspective towards more accurate neoantigen prediction which eventually contribute to personalized cancer immunotherapy.

Availability: Data and implementation are available on webserver: http://yishi.sjtu.edu.cn/deepAntigen

Contact: yishi@sjtu.edu.cn, menglum@scnu.edu.cn, darlt@sjtu.edu.cn, heguang@sjtu.edu.cn, hanzg@sjtu.edu.cn

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

KEYWORDS: Deep Sparse Learning, Neoantigen, 3D genome, Hi-C, Immunotherapy, MHC-I Epitope.

1. Introduction

The approval of several immunotherapies has led to dramatic changes in cancer therapy. In a variety human malignancies, therapeutic efficacy was enhanced by immunotherapies via boosting the endogenous T cell's ability to destroy cancer cells (Schumacher and Schreiber, 2015). The 'checkpoint inhibitors' therapies work by blocking proteins that act as molecular breaks for T cells. With the breaks removed, T cells can better undertake their job to kill cancer cells. Despite the great success of checkpoint inhibitors, still many patients do not respond to the agents, and many that do temporarily respond, eventually relapse. Moreover, checkpoint inhibitors do not fully take advantage of the T cell's exquisite specificity, one of its most important characteristics (Sompayrac, 2019). This led many researchers pay more attention to the new immunotherapy strategies against tumor known as neoantigen therapies. T cells are potent at killing when they recognize 'foreign' antigens which could be some protein fragments from an invading virus or bacteria. The key ability of T cells in distinguishing foreign antigens from self prevents autoimmunity which on the contrast makes them less potent in recognizing tumor cells because they are our own but abnormal cells. The T cells overcome this dilemma in two ways. First, they tend to respond to tissue-specific antigens (TSAs) which are specific amino acid fragments produced by cells of certain types. Second, T cells respond to neoantigens which are small peptides generated in tumor cells containing high level of DNA mutations. The nonsynonymous mutations can be entirely absent from the human genome, leading the cancer cells vulnerable to T cells as they look 'foreign' (Sompayrac, 2019).

In several clinical practices, it has been demonstrated that endogenous T cells with mounted cancer-killing T cell receptor (TCR) are able to recognize epitopes which are composed of the peptides displayed on major histocompatibility complexes (MHCs) on the surface of the cancer cells (Ott, et al., 2017; Schumacher and Schreiber, 2015). With the help of DNA and RNA sequencing technology, it has been revealed that tens to thousands of different somatic mutations can be generated during cancer initiation and progression, depending on different cancer types (Castro, et al., 2019; Prior, et al., 2019; Volkov, et al., 2019). Most of these mutations are often caused by genomic instability within the tumor cells and lead to no obvious cell growth advantage; they are also known as passenger mutations. On the contrast, a small percent of these mutations are known as driver mutations which interfere with normal cell regulation and help to drive cancer growth and resistance to targeted therapies (Yarchoan, et al., 2017). Both passenger and driver mutations can cause tumor to express abnormal proteins or polypeptides that cannot be found in normal cells as they can be nonsynonymous mutations that alter protein-coding sequences. When cell metabolize, the proteins possessing abnormal sequences are cut into short peptides and are presented as epitopes on the cell surface by the MHC (also known as human leukocyte antigen, HLA, in human case) molecules, which have a chance to be recognized by T cells as foreign antigens (Yarchoan, et al., 2017). An effective neoantigen which leads to the final immunological response, is determined by many factors. For instance, Dintzis *et al.* found that size-fractionated linear polymers of acrylamide substituted with hapten can affect the immunogenicity triggering (Dintzis, et al., 1976). Other factors such as peptide degradation and transportation, peptide-MHC binding affinity and stability, and pMHC-TCR interaction should also be considered (Blaha, et al., 2019).

Based on the above knowledge, in ideal situation, after the DNA sequencing procedure, potential neoantigens can be synthesized in vitro and their efficacy can be validated in vivo via either cancer cell-line or animal model, before conducting in clinical practice (Schumacher and Schreiber, 2015; Yarchoan, et al., 2017). Indeed, the cancers with a single dominant mutation can often be effectively treated by focusing on the driver mutation (O'Brien, et al., 2003; Yarchoan, et al., 2017). Nevertheless, in many other cancer situations, the somatic mutations are usually abundant, which lead to a computationally challenging task to efficiently prioritize the potential neoantigen candidates according to their ability to activate the T cell's immuno-response (Hackl, et al., 2016). In the past decade, many prediction methods have been proposed to address the neoantigen prioritization problem (Jurtz, et al., 2017; Lundegaard, et al., 2008; Nielsen and Andreatta, 2016). These methods can be categorized into two major classes: the protein spatial conformation-based approaches which consider the pMHC and T cell receptor (TCR) 3D structures, and the protein sequence-based approaches which consider the amino acid combinatorial characters. For the protein spatial conformation-based approaches, when high quality pMHC 3D structures are available, methods such as molecular dynamic (MD) can be adopted to explore the complex interaction between TCR and pMHC (Blevins, et al., 2016; Riley, et al., 2018; Wang, et al., 2017). If high quality pMHC spatial information is lacking, by sacrificing computational complexity and spatial model accuracy, computational pMHC modelling can be adopted, followed by 3D to 1D feature transformation and machine learning approaches (Riley, et al., 2019). Most neoantigen prediction methods belong to the sequence-based class because they can usually be set up efficiently (Gupta, et al., 2016; Hackl, et al., 2016), and there are much larger data sets available for training and validation (Vita, et al., 2019; Zhang, et al., 2011).

Early sequence-based methods such as BIMAS (Parker, et al., 1994) and SYFPEITHI(Schuler, et al., 2007) utilized the position-specific scoring matrices (PSSMs), which are defined from experimentally confirmed peptide binders of a particular MHC allele (Hackl, et al., 2016). More sophisticated approaches based on machine learning techniques were later developed which were demonstrated to perform better than the PSSM-based methods; these approaches capture and utilize the nonlinear nature of the pMHC-TCR interaction. In recent years, consensus approaches such as CONSENSUS (Moutaftsi, et al., 2006) and NetMHCcons (Karosiene, et al., 2012) were exploited which combine results of multiple neoantigen prediction tools, aiming to obtain more robust and accurate outcomes, and their efficacies were supported by experimental results. Nonetheless, the performance gain of these methods is determined by the weighting scheme among different prediction components, which lead to increased computational complexity (hyper-parameter tuning). Because the peptide MHC binding can be affected by HLA allele variety, most recently, the pan-specific methods, such as NetMHCpan (Jurtz, et al., 2017; Nielsen and Andreatta, 2016), were developed which allow the HLA type independent prioritization. In NetMHCpan, a neural network is firstly trained based on multiple public datasets, then the binding affinity for a given peptide-MHC complex is predicted according to the trained neural network, with the polymorphic HLA types, e.g., HLA-A, HLA-B or HLA-C being considered. Even compared to HLA allele-specific approaches (Hackl, et al., 2016; Trolle, et al., 2015), both NetMHCpan (Jurtz, et al., 2017) and NetMHCIIpan (Karosiene, et al., 2013) could perform remarkably better. Although methods such as NetMHC or NetMHCpan were designed to predict peptide-MHC binding affinity, they were either considered as strong indicators for neoantigens' effectiveness (Harndahl, et al., 2012; Lundegaard, et al., 2011; Rasmussen, et al., 2016), or were adopted as important features in the state-of-the-art neoantigen predicting methods such as Neopepsee and pTuneos (Kim, et al., 2018; Zhou, et al., 2019). More recently, Wu *et al.* proposed a recurrent-neural-network based approach DeepHLApan which considered both pMHC binding and potential immunogenicity, yet sequence information of both peptide and HLA were still adopted as training features (Wu, et al., 2019).

For all the existing neoantigen prediction methods, although several evaluation criteria were proposed for a more fair and robust comparison (Peters, et al., 2006; Trolle, et al., 2015; Wang, et al., 2008), independent benchmark studies that can be used to recommend specific tools are still lacking. More importantly, although there are abundant previous researches indicating that somatic mutations, including point mutations, gene fusions, and copy number abnormalities do not occur at random in the perspective of genome 3D conformation (Berger, et al., 2011; Branco and Pombo, 2006; Engreitz, et al., 2012; Mani, et al., 2009; Mathas, et al., 2009; Meaburn, et al., 2007; Nikiforova, et al., 2000; Roix, et al., 2003; Wijchers and de Laat, 2011), for which we also did a thorough study and discovered the somatic co-mutation hotspot (SCH) in 3D genome (Shi, et al., 2016), none of the existing neoantigen prediction methods considers this spatial genomic information of somatic mutations, i.e., the DNA loci of these mutations in the perspective of high order genome 3D conformation. We believe that the 3D genome information could contain much richer information compared to the existing amino acid sequence based neoantigen prediction methods. Therefore, in this work, we retrospect the DNA origin of the neoantigens, both immune-positive and negative, in the context of the genome 3D conformation, and demonstrate some discoveries that worth paying attention to. We adopted the 3D genome information into an ensemble peptide feature coding scheme, and developed a group feature selection-based deep sparse neural network (DNN-GFS) model that is customized and optimized for the neoantigen prediction task. We also developed an off-the-shelf webserver that implements the DNN-GFS method along with other machine learning methods; the webserver takes sequencing result (vcf file) and produces prioritized neoantigens as well as some useful intermediate functions such as vcf annotation and candidate neoantigen enumeration, etc. The whole workflow is illustrated in Fig. 1, where the adoption of 3D genome information, ensemble feature coding, and the DNN-GFS algorithm are keys for distinguishing our neoantigen prediction from all the existing methods.

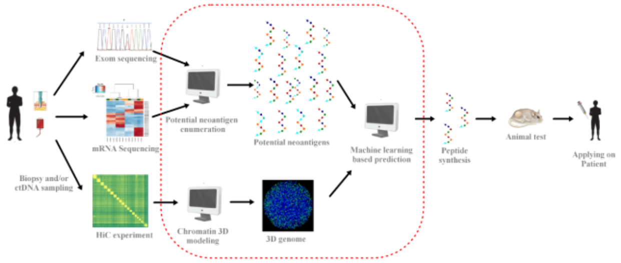

Fig. 1 **Workflow of neoantigen therapy supported by 3D genome information**. Left to right: tumor sample collection from patient; Whole-exome sequencing and mRNA sequencing for somatic mutations calling and gene expression estimation (whether the mutated DNA is expressed into mRNA and could potentially be translated into protein/peptide) respectively; Hi-C data curation to obtain 3D genome information; candidate peptides determined by NGS are generated and by combining 3D genome information immune-positive peptides are predicted machine learning methods; the top ranked peptides are screened by conducting animal experiments; the final peptide penal can be applied back to the target patient. This work aims to solve the tasks within the dashed red frame.

2. Methods

2.1 Immunogenicity data curation and reference genome mapping

The neoantigen peptide sequences and the immune responses were collected from the IEDB database under the T Cell Assay category (Vita, et al., 2019). For the cross-validation experiments, we collected training data before 2018 in IEDB; After collecting 337248 peptide records in the primary dataset, we performed filtering under *Homo Sapiens* and MHC-I subtypes and restrained the peptide length 9, as well as merging identical records and mapping to human reference genome hg19. When mapping the peptides to the reference genome, we first applied the PANDAS library to create a data frame object for subsequent processing. Then we assigned the column name by importing a name dictionary and filtered the dataset so that the only entries left have *Homo Sapiens* as their hostname. The dataset was further cleaned up by applying two functions we developed, Letter_check and Drop_legal, which checks for amino acid alphabet legitimacy. We developed a pipeline to query the BLAST (Boratyn, et al., 2013) web server and map the gene names to chromosomes and starting positions. The dataset was divided into 711 partitions where each partition contains 100 sequences. To set up BLAST queries, we restricted the search to *Homo Sapiens* using the entrez ID keywords and used the PAM30 matrix to find matches; the gap costs were adjusted to regulate gap penalty. We then queried BLAST iteratively. For each match, we adopted the accession and raw bit score for the first hit. After obtaining the accessions, we used the DAVID tool (Huang, et al., 2009) to obtain the gene names composed with gene symbols and the chromosome positions are also obtained. The final results contain a tuple of peptides, HLA subtype, chromosome number, and chromosome position. For identical peptides with multiple immune experiments, we define peptides with positive rate > 80% as immune-positive samples and with positive rate < 20% as immune-negative peptides. In the end, we obtained 3909 peptides, with 809 immuno-positive peptides and 3100 immuno-negative peptides. We also collected a standalone validation dataset from IEDB dated after 2018 and performed the same operation mentioned herein. In the end, 430 validation peptides were obtained with 125 positive samples and 305 negative samples.

2.2 Hi-C data curation & A/B compartment determination

For the chromatin 3D conformation data, we employed two well-known Hi-C data resources (Dixon, et al., 2012; Rao, et al., 2015), and obtained eight Hi-C datasets, i.e., hESC, IMR90, GM12878, HUVEC, IMR90-Rao, NHEK, K562, and KBM7. The Knight-Ruiz normalization (KR-norm) was applied on both intra-chromosomal and the inter-chromosomal (genome-wise) Hi-C contact maps. Bin sizes of 40kb, 100kb, and 500kb were adopted for intra-chromosomal contact frequency analyses, A/B compartment analyses, and inter-chromosomal contact frequency analyses and chromatin 3D modeling. 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 the contact frequencies by their corresponding off-diagonal mean. Then we computed the Pearson correlation coefficient (PCC) matrices for each chromosome, and the compartment type was jointly determined by the sign of the eigenvector corresponding to the first eigenvalue of the PCC matrices and the signal of the epigenetic marker H3k4me1.

2.3 Chromatin 3D modeling

We employed molecular dynamics (MD) and developed a human genome 3D conformation modeling approach with resolution 500kb (bin size) for all eight Hi-C datasets. The bins were coarse-grained as beads and intact genome was represented by bead-on-the-string structures consisting of 23 polymer chains. The beads' spatial positioning is affected by both chromatin connectivity that constrains linearly neighboring beads in close 3D proximity and chromatin activity that ensures active regions tend to be located closer to the nucleus center. The chromatin activity was determined according to compartment degree that can be directly calculated from Hi-C matrix as described above and also in previous work (Xie, et al., 2017). Based on compartment degree index, beads were assigned distance values with respect to the nuclear center; the conformation of chromatin was then optimized from random structures with molecular dynamics approach by applying bias potential to satisfy these distance constraints. For each cell linage, 300 feasible conformation structures were optimized from random ones to reduce possible variation for further analysis.

2.4 Deep sparse neural network methods

The deep feedforward networks, also known as multilayer perceptrons (MLPs) were employed in this work as the basic neural network architecture (Goodfellow, 2016). For a single unit, its basic form is y=*f* (x; θ), where x is the input, y is the output, and θ represents the parameters of the network that need to be optimized by adaptable methods. For a single middle layer neural network, a generic form can be given as:

$$\mathbf{y}_k = g_k\left(\mathbf{W}_k\mathbf{x}_k + \mathbf{b}_k\right) \tag{1}$$

where {$\mathbf{W}_k$, $\mathbf{b}_k$} are the optimized parameters of the layer, corresponding to θ in basic form, and $g_k(.)$ is the activation function of the layer for which we chose the widely adopted linear unit (ReLU) and the sigmoid unit in our model. Their function forms are g(z)=max {0, z} and g(z)=σ(z); $\mathbf{x}_k$ is the input and $\mathbf{y}_k$ is the output. Note that an important prerequisite in our model is $\mathbf{x}_{k+1}$=$\mathbf{y}_k$, which makes all layers form the whole network, and specially, there is no input $\mathbf{x}_k$ for input layer. In order to obtain a set of adaptable {$\mathbf{W}_k$, $\mathbf{b}_k$}, the network should be trained multiple times by minimizing the regularized objective function $\tilde{J}$:(Goodfellow, 2016)

$$\tilde{J}(\boldsymbol{\theta};\mathbf{X},\mathbf{y}) = J(\boldsymbol{\theta};\mathbf{X},\mathbf{y}) + \lambda R(\boldsymbol{\theta}) \tag{2}$$

In practice, only the weights (**W**) of θ at each layer are penalized, and to simplify the equation, θ can be replaced by **w**:(Goodfellow, 2016)

$$\tilde{J}(\mathbf{w};\mathbf{X},\mathbf{y}) = J(\mathbf{w};\mathbf{X},\mathbf{y}) + \lambda R(\mathbf{w}) \tag{3}$$

where *J*(**w;X,y**) is the standard objective function, R(**w**) is the parameter norm penalty, and $\lambda \in [0,\infty]$ is a hyper parameter that weights the two terms. Larger values of λ correspond to more regularization and setting λ=0 results in no regularization. In this work, we set *J*(.) as the cross-entropy loss. Many effective regularization strategies have been previously studied. The most common regularization strategy is the L_2 norm penalization, which is usually adopted to avoid overfitting. Its general form is:

$$R(\mathbf{w}) = \|\mathbf{w}\|_2^2 \tag{4}$$

also known as Tikhonov regularization or ridge regression. Another common practice is the L_1 regularization, which has a similar presentation:

$$R(\mathbf{w}) = \|\mathbf{w}\|_1 = \sum_i |\mathbf{w}_i| \tag{5}$$

that is the sum of absolute values of all weights. Particularly, the LASSO (least absolute shrinkage and selection operator) is a typical model that uses a L_1 penalization. The L_1 regularization can not only avoid overfitting, but also obtain a sparser solution than L_2, by making a subset of the weights to become zero (or very close to zero), suggesting that the corresponding features may safely be discarded. Due to this important property and the ability of preventing overfitting, L_1 regularization is used in feature selection scenario extensively (Goodfellow, 2016). Note that recent study has revealed that sparsity is the key to imitate human brain for the neural network (Dettmers and Zettlemoyer, 2019).

Apparently, the regularization can prevent overfitting, but its contribution is not limited to that. Scardapane *et al.* (Scardapane, et al., 2017) considered group-level sparsity, a weight grouping strategy was achieved by grouping all outgoing connections from a single neuron, which may induce the property of pruning the corresponding neuron from the network. As introduced in group lasso (Simon and Tibshirani, 2012), group sparse regularization, e.g., $L_{2,1}$ norm, can be written as:

$$R_{\ell_{21}}(\mathbf{w}) \square \sum_{\mathbf{g}\in G}\sqrt{|\mathbf{g}|}\,\|\mathbf{g}\|_2 \tag{6}$$

where |**g**| is the dimensionality of the vector **g**, vector **g** corresponds to weight matrix **W**, every **g** is one row of a matrix **W**, denoting all outgoing connections from an input neuron. *G* is the set of **g**, **g**∈*G*, which is the result of grouping **W** by row. Furthermore, sparse group Lasso (SGL) penalization was proposed by combining Lasso and group Lasso (Scardapane, et al., 2017; Simon, et al., 2013; Simon and Tibshirani, 2012)

$$R_{\mathrm{SGL}}(\mathbf{w}) \square R_{\ell_{21}}(\mathbf{w}) + R_{\ell_1}(\mathbf{w}) \tag{7}$$

which can increase the sparsity above group sparse regularization. In addition, the hyperpapermeter can be used to weight the two terms, that is (Friedman, et al., 2010):

$$R_{\mathrm{SGL}}(\mathbf{w}) \square (1-\alpha)R_{\ell_{21}}(\mathbf{w}) + \alpha R_{\ell_1}(\mathbf{w}) \tag{8}$$

where α=1 corresponds to the L_1 term and α=0 corresponds to the $L_{2,1}$ term. This form gives users more choice for their problem.

2.5 Group feature selection based DNN (DNN-GFS)

Traditional DNN and some relevant sparse DNNs have a good performance but remain to be improved in many research field (Goodfellow, 2016). When real problems are handled by deep learning, there are usually some prior knowledge neglected, leading to an unideal performance. If we only consider the datasets, it is difficult to obtain the optimal model and the corresponding parameters we expect. Moreover, the situation will get worse with decreasing sample size, especially in biology problems with more features than samples. But when the prior information is imposed on models, the model will be closer to our expectation and generalization may be improved.

For our neoantigen prioritization problem, based on the existing sparse DNN models (Friedman, et al., 2010; Simon, et al., 2013; Simon and Tibshirani, 2012), we develop a new regularization strategy that aims to tackle both feature selection and the group sparse regularization challenge, which is an extension of the L_2 and L_1 penalization. Specifically, the feature grouping nature is considered in group sparse regularization, forming a new regularization strategy. We term it Group Feature Selection (GFS) regularization. In the feature vector of our neoantigen prediction problem, some groups contain multiple features and some groups contain a single feature. In the former cases, features of the same group need to be either all selected or all rejected, simultaneously. This means that all outgoing connections from all neurons in one group should be either simultaneously all zeros, or all non-zeros (Scardapane, et al., 2017). The group feature selection regularization can be written as follows:

$$R_{GFS}(\mathbf{w}) \square \sum_{\bar{\mathbf{g}} \in G_f} |\mathbf{F}_s| \sqrt{|\bar{\mathbf{g}}|} \| \bar{\mathbf{g}} \|_2 \tag{9}$$

where vector $\bar{\mathbf{g}}$ is the average of the squares of $\mathbf{g}$ vectors of a feature group, which can efficiently reduce computational complexity. $|\bar{\mathbf{g}}|$ is the dimensionality of the vector $\bar{\mathbf{g}}$, and G_f is the result of grouping again by feature group information based on G (groups of group lasso). As Fig. 2 a illustrates, some features form new groups. $|\mathbf{F}_s|$ is the corresponding feature number matrix of G_f. Note that, when a group contains a single feature, the expression can be simplified as $L_{2,1}$. Moreover, for one-dimensional groups, it can also be reduced to the standard Lasso, while for all features in a new group, it is closer to L_2 regularization. These regularization terms other than L_2 are convex but non-smooth, since their gradient is not defined when $||\bar{\mathbf{g}}||_2=\mathbf{0}$, which is illustrated in Fig. 2c.

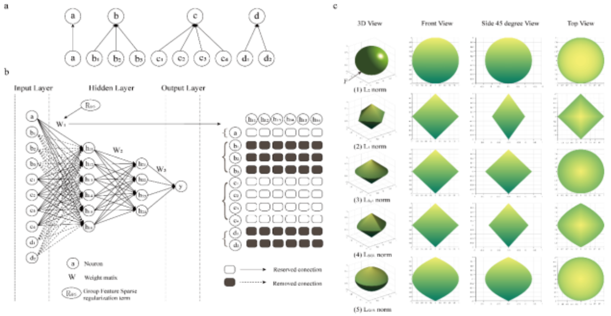

Fig. 2. The group feature selection based deep neural network method (DNN-GFS). a illustration of features belonging to groups of different sizes. All features belong to at most one group. A group can contain a single feature or multiple features. **b** Illustration of the DNN-GFS architecture and the group feature selection effect. **c** Illustration of the geometric principles of different regularization terms applied on the weighted neural network wiring and 2D projection from three representative views. F denotes Front view in c (1).

The GFS devised here is a flexible regularization strategy as G_f can be customized according to different preferences to adapt various requirements. Furthermore, $|\mathbf{F}_s|$ is also chosen skillfully in this work. i.e. $|\mathbf{F}_s|$ can be replaced or rectified by other coefficients, which is able to enlarge or narrow the differences among groups. When imposing $R_{GFS}(\mathbf{W})$ on the $\mathbf{W}$, we achieve the feature selection effect as illustrated in Fig 2b. Results that are based on other regularization strategies are showed in Fig. S5 and S6 of supplemental materials, and it is demonstrated that only GFS can achieve the group feature selection effect. The detailed comparisons including network structures (Fig. S7), sparse effects of different strategies (Fig. S8) and tuning processes (Fig. S9-S14) are also given in supplemental materials. The geometric interpretations of different approaches in 3D space and 2D projection from three representative views is shown in Fig 2c, and more details can be found in supplemental materials. Similar to L_1, GFS achieves sparsity and avoids overfitting, and moreover, the performance is improved by exploiting group information.

3. Results

3.1 The distribution of neoantigens' DNA loci in 3D genome

For all the peptides (both immuno-positive and immuno-negative) included in this study, we first generated a pool that contains all the peptide pairs. Then we classified all the peptide pairs in this pool into three categories: positive-positive pairs (Pos-Pos), negative-negative pairs (Neg-Neg), and positive-negative pairs (Pos-Neg). For each peptide pair, we computed contact frequencies for each Hi-C datasets, i.e., hESC, IMR90, GM12878, HUVEC, IMR90-Rao, NHEK, K562, and KBM7, respectively(Dixon, et al., 2012; Rao, et al., 2015). The contact frequency distribution of the three categories are shown in Fig. 3a. It is demonstrated that on all the Hi-C datasets, immune-positive peptide pairs are more proximate to each other comparing to immune-negative peptide pairs; the corresponding T-test and Wilcoxon rank sum test p-values, i.e., Pos-Pos vs. Neg-Neg, are all smaller than 10^{-99} and 10^{-18}, respectively. This indicates that the immuno-positive peptide's DNA loci tend to be more proximate in genome spatial space. We then computed the A/B compartment type (A: active; B: inactive) for each chromosomal region (bin), based on both Hi-C dataset and epigenetic markers, shown in Fig. 3b and Fig. 3c. The whole genome contact maps of the eight Hi-C datasets are shown in Fig. S2 and the A/B compartment results of each chromosome are shown in Fig. S3 of Supplementary Materials. Then we assigned the corresponding DNA loci of the positive and negative peptides with their A/B compartment type. We found that in certain chromosomes, immune-positive neoantigens tend to be located on compartment A, comparing to immuno-negative neoantigens, as shown in Fig. 3d and Fig. S4 of Supplementary Materials. This indicates that the DNA loci of the immuno-positive or negative peptides are positively correlated to chromosome compartment type, either A or B, depending on which chromosome.

We then developed a novel molecular dynamic based chromatin 3D modeling method and mapped the immuno-positive and negative peptides' corresponding chromosomal loci to the constructed 3D genome structure and calculated their radius distance to the nucleus center, as shown in Fig. 3e. We found that the immuno-positive peptide's corresponding loci tend to locate closer to the nuclear periphery (more far away from the nucleus center), compared to the immuno-negative ones, as Fig. 3f demonstrates. We found that by adopting the radius position information, the prediction power of the existing methods such as netMHCPan and netMHC can be elevated. In detail, prediction scores defined as $Y_{pred} = S_{netMHCPan} \times r^2$ or $Y_{pred} = S_{netMHC} \times r^2$ can significantly better discriminate the immune-positive peptides from the immune-negative peptides, comparing to using netMHCPan or netMHC alone. We thus believe that the DNA loci's radius positions of the immuno-positive and immuno-negative peptides are significantly differently distributed and can play an important role in predicting pMHC-I immunogenicity.

3.2 Peptide encoding and predictions

A reasonable and proper peptide encoding strategy is key to the downstream predictions as it can include and quantify more features that are plausibly related to the outcome. But by including more features into the prediction model, we also increase the risk of adding noisy (irrelevant) features into the feature pool and making the prediction prone to overfitting. To overcome this dilemma, we propose to first enumerate as many features as possible and then perform feature selection within the training process of the prediction modeling. Previous neoantigen prediction methods adopted one or more coding schemes such as amino acid (AA) composition, AA sparse coding, BLOSM, BLOMAP, etc. In this work, based on the above observation that chromatin 3D information may significantly contribute to discriminating immuno-positive peptides from immuno-negative ones, we adopted this piece of information in the peptide encoding strategy. In detail, the 3D coordinates and the radius positions of the Hi-C data based 3D modeling results, the HLA subtype encoding, the amino acid compositions, the sparse coding, BLOMAP coding, and BLOSUM coding of the peptides, the AAindex2 coding of the peptides are adopted and collected as features. At the end, we obtained a training matrix with 3909 peptides and 5459 features, shown in Fig. 4a. Note that as Fig. 4a demonstrates, there is no obvious pattern that a single feature or a group of features are correlated to the true label vector.

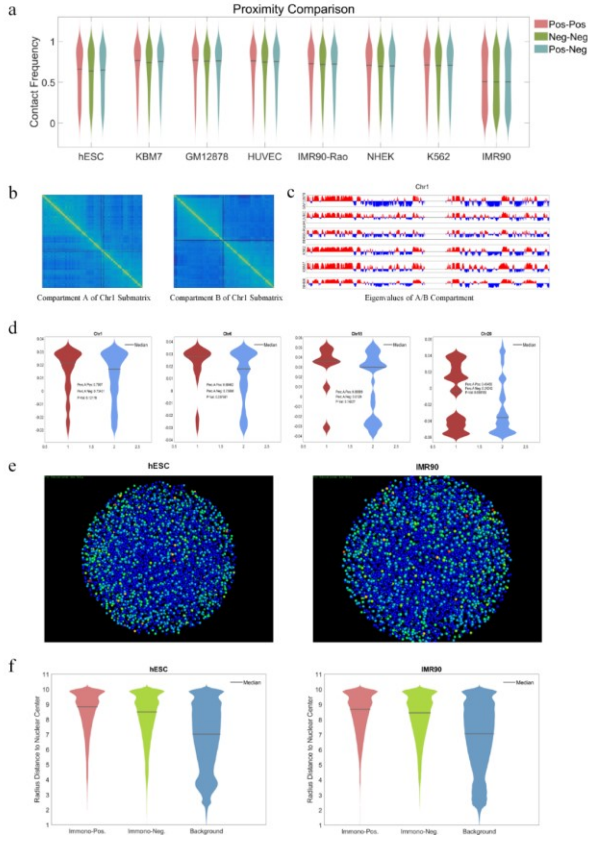

Fig. 3 The DNA loci of neoantigens in 3D genome. a Distribution of proximities between peptide pairs of different types. Immuno-positive peptide pairs tend to be more proximate to each other comparing to immuno-negative ones, while immuno-positive-negative pairs lie in between (all the P-values of the T-test comparison are smaller than 10^{-99}). **b** Illustration of Hi-C submatrices of compartment A and B on chromosome 1. **c** Illustration of eigenvalues of compartment A (red) and B (blue) on chromosome 1. **d** Comparison of percentages of immuno-positive peptide belonging to compartment A (red) and immune-negative peptide belonging to compartment A (blue). **e** The 3D genome molding results based on hESC and IMR90 Hi-C datasets and the distribution of the DNA loci of immuno-positive (yellow to red color spectrum, depending on positive occurrence on the same 500k bin) and immuno-negative peptides (green). **f** Radius position comparison of the immuno-positive and negative peptides' DNA loci 3D genome. The positive loci (red) are significantly closer to the nuclear periphery (more far away from the nucleus center), compared to the immuno-negative ones (green); they are all closer to the nuclear periphery comparing to the background distribution (blue). All T-test P-values are smaller than 10^{-99}.

In theoretical deep neural network (DNN) studies, there have been plenty of evidences pointing to the fact that the majority of weights in most deep networks are redundant and may jeopardize the prediction accuracy (Han, et al., 2015; Sainath, et al., 2013; Scardapane, et al., 2017). It is possible to learn only a small percentage of the weights, while still preserving the prediction accuracy (Han, et al., 2015). Nevertheless, studies focusing on the input feature selection based neural network is limited. Moreover, in the neoantigen prediction problem, the features that encode the peptides come in groups, e.g., the 3D coordinates <x,y,z> of a peptide's DNA loci are in one group, or the sparse coding for an amino acid is a group of 20 binary features, etc.. Therefore, when imposing feature selection on the DNN, it should be in a group fashion, i.e., features belonging to the same group should be either all selected or all rejected. The group feature selection based deep neural network (DNN-GFS) is introduced in detail in the Methods section.

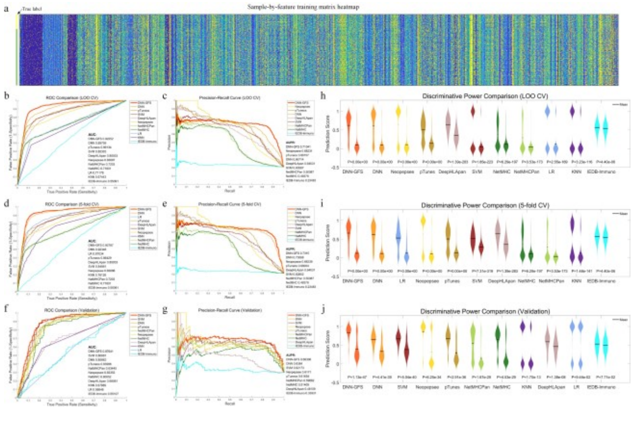

Fig. 4 Prediction results comparison. a The leftmost column vector indicates the true labels of the immuno-positive (yellow) and immuno-negative (green) for each of the 3909 peptides. The matrix heatmap indicates the column-wise normalized feature values of the 3909 peptides by 5459 features. **b** and **c** are the ROC plot comparison for DNN-GFS, DNN, SVM, LR, KNN, Neopepsee, pTuneos, DeepHLApan, NetMHCpan, NetMHC and IEDB-immuno, under 5-fold and leave one out (LOO) cross-validation respectively. **d** and **e** are the precision-recall plot comparison for different prediction methods under 5-fold and leave one out (LOO) cross-validation respectively. **f** and **g** are the prediction score (normalized) distribution comparison for immuno-positive (left violins) and immuno-negative peptides (right violins); all the P-values of the T-tests are equal to or very close to zero.

To compare the prediction efficacy, in addition to DNN-GFS, we also applied traditional L2 norm deep neural network (DNN), support vector machine (SVM), logistic regression (LR), k-nearest neighbor (KNN) classifiers on the 5459 encoded feature matrix. Moreover, we included the widely adopted methods IEDB-immunogenicity, NetMHCpan and NetMHC into the comparison, as well as the most recent popular methods Neopepsee, pTuneos and DeepHLApan. The comparison was conducted in the framework of both cross-validation (5-fold or leave-one-out) and validation alone. The ROC curves are shown in Fig. 4 b, d and f; the precision-recall curves are shown in Fig. 4 c, e, and g; the prediction score distributions for the immuno-positive and negative samples are shown in Fig. 4 h, i and j. Note that in the ten prediction methods, the KNN and LR output binary values, so for precision-recall curve comparison, we excluded them. Detailed prediction statistics are shown in Table S1-1, S1-2, and Table S2. As the comparison results demonstrate, the deep learning-based approaches DNN-GFS and DNN outperform the rest of the methods and DNN-GFS, due to its feature selection potency, is better than traditional DNN. The SVM, Neopepsee, pTuneos and DeepHLApan are also effective methods and ranked second tier among the ten methods. Although NetMHC and NetMHCpan were initially designed to predict peptide-MHC binding affinity, their capability in predicting neoantigen cannot be neglected and they are ranked third tier. The logistic regression and KNN classifiers, although performs reasonably well in cross-validation experiments, are not very stable when applied on the standalone validation set. The IEDB-immunogenicity prediction method, does not catch up with other prediction methods, possibly due to the fact that the immunogenicity scoring function is too simple to capture subtle sequence features that only advanced non-linear machine learning methods can. We also implemented other well-known sparse learning neural network models and compared their efficacy with DNN-GFS, as introduced in table S1 and S2 of Supplementary Materials, and the results indicate that DNN-GFS outperforms existing sparse neural network methods in terms of prediction statistics.

3.3 Features selected by DNN-GFS

Based on the whole training dataset, the DNN-GFS model selected 2693 features out of the 5459 features, achieving a feature sparsity ratio 49.33%. Features belonging to the same group are either all selected or all excluded. Among the selected features, all the 3D genome related features are selected, including radius position, HLA subtype, 3D coordinates of peptides' DNA loci, etc. For HLA subtype-encoding, all features are selected and cross-validation performance is improved about 3%~4% compared to dataset of not containing HLA subtype information, which illustrates their importance. For 9 AA peptides, the sparse coding of the peptide's position 1 to 5 and 7 to 8 are all selected but not position 6 and 9. BLOSUM coding features are all excluded while BLOMAP coding features for AA position 1 to 4 are selected. Except AA position 5, other side chain polarity features are all selected, and side chain charge features for position 1 to 3 are selected. For the hydropathy features, AA position 5 and 9 are selected, and for molecular weight, feature of AA position 2, 6, and 9 are selected. Other selected features are mostly AAindex2 related features. The DNN-GFS model thus suggests that the combination of these grouped features play an important role in building the prediction model and we believe that the importance of these features in neoantigen prediction is worth further investigating. Detailed feature selection and model sparsity analyses can be found in Supplementary Materials.

4. Discussion

From the association study of peptides' immunogenicity and their 3D genome information, we found that immuno-positive peptides' DNA loci tend to be more proximate to each other and locate closer to the nuclear periphery, i.e., greater radius value to the nuclear center, comparing to immuno-negative ones. This implies that if a nonsynonymous mutation happens closer to some nonsynonymous mutation that were already proven to produce immuno-positive peptides, or if it is located closer to the nuclear periphery, the mutation is more likely to generate immuno-positive neoantigens. This association can be further enhanced if the A/B compartment information of the mutation is provided. In practice, the whole genome spatial organization is more conserved across different cell types and even in mutated cancer cells. While A/B compartment characters of certain chromatin regions may flip across cell lines or in cancer cells, i.e., more transient, we only adopted the 3D coordinates and radius position of peptides' DNA loci in the prediction model, but if the A/B compartment information can also be included if the real time cancer cell's chromatin 3D experiment can be performed in the future.

To explain such intriguing relationship between 3D genome and the neoantigens' immunogenicity, factors of at least three aspects should be considered: Firstly, the non-random nature of coding sequence distribution in 3D genome: during evolution, wild type coding sequences where neoantigens of different immunogenicity characters originate are located in different regions of the nucleus (Gorkin, et al., 2019; Svozil, et al., 2008). Secondly, the gene expressions affected by 3D genome: the missense mutations need to be transcribed to generate potential neoantigens and the gene expressions are known to be affected by high order genome organization (Gorkin, et al., 2014). Thirdly, the non-random occurrences of somatic mutations in 3D genome: previous discoveries indicated that somatic mutations may not occur at random, and we systematically studied and discovered in our prior work that co-mutations may occur in a spatial clustering fashion in genome 3D space (spatial co-mutation hotspot, SCH), possibly due to abnormal chemical concentration or a systematic DNA repair protein failure at certain chromatin 3D loci (Shi, et al., 2016). This leads to a straightforward hypothesis that mutations in different chromosomal regions may carry different immunogenicity character, affected by wild type coding sequences, somatic mutation patterns, and gene transcriptions. We thus believe that it is worth considering these aspects when studying the underlying mechanism of how high order genome organization affect neoantigens' immunogenicity. For example, to explain our discovery that immuno-positive neoantigens' corresponding DNA sequences tend to locate closer to nucleus periphery (greater radius to the nuclear center), one may consider the fact that their transcribed missense mRNAs enter cytoplasm more easily (a shorter path from transcription loci to nuclear envelope). Core genes during evolution, if mutated, require stronger TCR responses, because otherwise it causes greater cancerous impact, and these genes are usually expressed across cell types and their distribution in 3D genome also worth further study. Therefore, it is also worth to investigate the relationship between neoantigen immunogenicity and gene evolutionary essentiality in the perspective of high order genome conformation.

When building the prediction models, due to the fact that most MHC-I presented peptides are of 9 amino acid long, the features we used to encode the peptides are all based on 9mer peptides, and the predictions are targeted on the 9mers as well. Nevertheless, our approach is not restricted to 9mers and can be easily extended to peptides of other length. For example, if a target peptide is longer than 9 amino acids, a sliding window of length 9 can be used to enumerate all possible 9mers, and the prediction score can be estimated by taking the maximum or average of each individual scores. In the cases where a target peptide is shorter than length 9, we only need to consider length 8 as peptides presented by MHC-I shorter than or equal to length 7 is very rare. So, for the 8mer cases, we can compensate an extra amino acid to the beginning or to the end of the sequence and enumerate all possible peptides and again take the maximum or average of each individual one's prediction score.

Most existing machine learning algorithms for the classification problem usually assume that the feature across different training examples is independent and obey the same distribution, and the links among them are usually neglected which is not reasonable for an unbalanced problem. In many real-world applications however, the small sample issue is ubiquitous and the features are usually correlated. The DNN-GFS developed here provides a new way of exploiting these links for feature selection in addition to traditional neural networks. In the machine learning area, quite a few studies have exploited introducing sparse regularization into deep neural network framework, but most of these models only focus on reducing complexity of the network as a whole, resulting in pruning edges and nodes of the network, but not specifically targeting on the input layer, i.e., the input feature vector. In this work therefore, due to the scenario that peptides are represented in an ensemble encoding which may introduce noise or redundant features into the learning process, the proposed DNN-GFS model focus on reducing the features of the input layer. Moreover, due to the nature that certain features are grouped and should be either all selected or all rejected, we considered selecting features in a grouped fashion in the model, by imposing group-specific regularization. As shown in Fig. 4, the DNN-GFS model not only exceeds the widely adopted methods NetMHCpan and NetMHC, but also exceeds other existing machine learning methods such as DNN, LR, SVM, and KNN that are performed based on the same 5459 feature encoding strategy. Moreover, DNN-GFS outperforms other sparse learning DNN models as shown in table S3-S10 of Supplementary Materials. This agrees with our conjecture that DNN-GFS is a better DNN heuristic designed specifically for the neoantigen prediction in the specific 5459 encoding scenario. Although DNN-GFS outperforms the widely adopted NetMHCpan and NetMHC methods to a large extend, due to its ability of capturing subtle nonlinear relationships of features in a grouped fashion, the prediction power can be further improved once more immunogenicity training data are provided, especially for each HLA subtypes. We also believe that DNN-GFS can also be applied in other problems where group feature selection is demanded.

To facilitate practical usage, we developed a webserver deepAntigen (Fig. S1). In the current version, if the end user only provides sequencing result vcf file, the candidate peptides will be generated by only considering nonsynonymous point mutations, i.e., 9mer peptides surrounding the mutated amino acid, while small insertions or deletions (INDEL) can also be considered as *rankPep* function is independent and user can provide their own plausible peptides for prediction. For the prediction method, we suggest to use DNN-GFS as its power of discriminating immuno-positive peptides from immuno-negative ones are most potent, but other machine learning approaches can also be considered and the consensus result maybe of more interest to an end user.

Although the mechanism of under what conditions certain specific neoantigens activate T cell immunogenicity is still under studying, this work focuses on the machine learning challenge of effectively and efficiently predict/prioritize immuno-positive neoantigens. We found that the spatial distributions of the immuno-positive and immuno-negative peptides' corresponding DNA loci follow different pattern, i.e., immuno-positive peptides' DNA loci tend to be located more proximate to the nuclear periphery and tend to be more clustered in 3D genome space, compared with immuno-negative peptides' DNA loci; the peptides' DNA loci distribution is also related to the A/B compartment of the chromatin. It is therefore salient that utilizing the 3D genome information of the peptides' corresponding DNA loci can significantly contribute to the prediction of immuno-positive neoantigens. To utilize the most of 3D genome information, we customized a group feature selection based deep neural network (DNN-GFS) model, which takes not only the 3D genome information, but also a combinatorial peptide sequence features represented by an ensemble peptide encoding strategy. The DNN-GFS selected 3D genome related features as well as some other important peptide sequence features and position specific amino acid features; the comparison studies demonstrated that DNN-GFS outperforms the widely adopted methods NetMHCpan and NetMHC, and other machine learning prediction models including DNN, SVM, LR, and KNN. DNN-GFS is implemented in the webserver deepAntigen along with other machine learning methods. To the best of our knowledge, this is the first time that the DNA origins' 3D genome perspective is considered in the neoantigen study and we hope that our work contributes novel insights to neoantigen study and eventually benefits personalized cancer immunotherapy. Although close-up studies are needed to uncover the relationship between 3D genome and neoantigen immunogenicity, in this work, we only demonstrate the contributes of 3D genome information in more accurate neoantigen prediction, as well as providing plausible explanation that it is evolution that places sequences of different immunogenicity characters in different locations in the 3D genome while different locations are prone to mutations of different causes.

References

Berger, M.F., *et al.* The genomic complexity of primary human prostate cancer. *Nature* 2011;470(7333):214-220.

Blaha, D.T., *et al.* High-Throughput Stability Screening of Neoantigen/HLA Complexes Improves Immunogenicity Predictions. *Cancer Immunol Res* 2019;7(1):50-61.

Blevins, S.J., *et al.* How structural adaptability exists alongside HLA-A2 bias in the human alpha beta TCR repertoire. *Proceedings of the National Academy of Sciences of the United States of America* 2016;113(9):E1276-E1285.

Boratyn, G.M., *et al.* BLAST: a more efficient report with usability improvements. *Nucleic Acids Res* 2013;41(Web Server issue):W29-33.

Branco, M.R. and Pombo, A. Intermingling of chromosome territories in interphase suggests role in translocations and transcription-dependent associations. *PLoS Biol* 2006;4(5):e138.

Castro, A., *et al.* Elevated neoantigen levels in tumors with somatic mutations in the HLA-A, HLA-B, HLA-C and B2M genes. *BMC Med Genomics* 2019;12(Suppl 6):107.

Dettmers, T. and Zettlemoyer, L. Sparse Networks from Scratch: Faster Training without Losing Performance. *arXiv preprint arXiv:1907.04840* 2019.

Dintzis, H.M., Dintzis, R.Z. and Vogelstein, B. Molecular determinants of immunogenicity: the immunon model of immune response. *Proc Natl Acad Sci U S A* 1976;73(10):3671-3675.

Dixon, J.R., *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* 2012;485(7398):376-380.

Engreitz, J.M., Agarwala, V. and Mirny, L.A. Three-dimensional genome architecture influences partner selection for chromosomal translocations in human disease. *PLoS One* 2012;7(9):e44196.

Friedman, J., Hastie, T. and Tibshirani, R. A note on the group lasso and a sparse group lasso. *arXiv preprint arXiv:1001.0736* 2010.

Goodfellow, I. Deep Learning/Ian Goodfellow, Yoshua Bengio, Aaron Courville. In.: MIT Press; 2016.

Gorkin, D.U., Leung, D. and Ren, B. The 3D genome in transcriptional regulation and pluripotency. *Cell Stem Cell* 2014;14(6):762-775.

Gorkin, D.U., *et al.* Common DNA sequence variation influences 3-dimensional conformation of the human genome. *Genome Biol* 2019;20(1):255.

Gupta, S.K., *et al.* Personalized cancer immunotherapy using Systems Medicine approaches. *Briefings in Bioinformatics* 2016;17(3):453-467.

Hackl, H., *et al.* Computational genomics tools for dissecting tumour-immune cell interactions. *Nat Rev Genet* 2016;17(8):441-458.

Han, S., *et al.* Learning both Weights and Connections for Efficient Neural Networks. *Adv Neur In* 2015;28.

Harndahl, M., *et al.* Peptide-MHC class I stability is a better predictor than peptide affinity of CTL immunogenicity. *Eur J Immunol* 2012;42(6):1405-1416.

Huang, D.W., Sherman, B.T. and Lempicki, R.A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols* 2009;4(1):44-57.

Jurtz, V., *et al.* NetMHCpan-4.0: Improved Peptide-MHC Class I Interaction Predictions Integrating Eluted Ligand and Peptide Binding Affinity Data. *J Immunol* 2017;199(9):3360-3368.

Karosiene, E., *et al.* NetMHCcons: a consensus method for the major histocompatibility complex class I predictions. *Immunogenetics* 2012;64(3):177-186.

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* 2013;65(10):711-724.

Kim, S., *et al.* Neopepsee: accurate genome-level prediction of neoantigens by harnessing sequence and amino acid immunogenicity information. *Ann Oncol* 2018;29(4):1030-1036.

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* 2008;36(Web Server issue):W509-512.

Lundegaard, C., Lund, O. and Nielsen, M. Prediction of epitopes using neural network based methods. *J Immunol Methods* 2011;374(1-2):26-34.

Mani, R.S., *et al.* Induced chromosomal proximity and gene fusions in prostate cancer. *Science* 2009;326(5957):1230.

Mathas, S., *et al.* Gene deregulation and spatial genome reorganization near breakpoints prior to formation of translocations in anaplastic large cell lymphoma. *Proc Natl Acad Sci U S A* 2009;106(14):5831-5836.

Meaburn, K.J., Misteli, T. and Soutoglou, E. Spatial genome organization in the formation of chromosomal translocations. *Semin Cancer Biol* 2007;17(1):80-90.

Moutaftsi, M., *et al.* A consensus epitope prediction approach identifies the breadth of murine TCD8+-cell responses to vaccinia virus. *Nature Biotechnology* 2006;24(7):817-819.

Nielsen, M. and 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* 2016;8(1):33.

Nikiforova, M.N., *et al.* Proximity of chromosomal loci that participate in radiation-induced rearrangements in human cells. *Science* 2000;290(5489):138-141.

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* 2003;348(11):994-1004.

Ott, P.A., *et al.* An immunogenic personal neoantigen vaccine for patients with melanoma. *Nature* 2017;547(7662):217-221.

Parker, K.C., Bednarek, M.A. and Coligan, J.E. Scheme for ranking potential HLA-A2 binding peptides based on independent binding of individual peptide side-chains. *J Immunol* 1994;152(1):163-175.

Peters, B., *et al.* A community resource benchmarking predictions of peptide binding to MHC-I molecules. *PLoS Comput Biol* 2006;2(6):e65.

Prior, L., *et al.* Genomic profiling of a dedifferentiated mucosal melanoma following exposure to immunotherapy. *Melanoma Res* 2019.

Rao, S.S.P., *et al.* A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping (vol 159, pg 1665, 2014). *Cell* 2015;162(3):687-688.

Rasmussen, M., *et al.* Pan-Specific Prediction of Peptide-MHC Class I Complex Stability, a Correlate of T Cell Immunogenicity. *J Immunol* 2016;197(4):1517-1524.

Riley, T.P., *et al.* T cell receptor cross-reactivity expanded by dramatic peptide-MHC adaptability. *Nature Chemical Biology* 2018;14(10):934-+.

Riley, T.P., *et al.* Structure Based Prediction of Neoantigen Immunogenicity. *Front Immunol* 2019;10:2047.

Roix, J.J., *et al.* Spatial proximity of translocation-prone gene loci in human lymphomas. *Nat Genet* 2003;34(3):287-291.

Sainath, T.N., *et al.* Low-Rank Matrix Factorization for Deep Neural Network Training with High-Dimensional Output Targets. *Int Conf Acoust Spee* 2013:6655-6659.

Scardapane, S., *et al.* Group sparse regularization for deep neural networks. *Neurocomputing* 2017;241:81-89.

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

Schumacher, T.N. and Schreiber, R.D. Neoantigens in cancer immunotherapy. *Science* 2015;348(6230):69-74.

Shi, Y., *et al.* Chromatin accessibility contributes to simultaneous mutations of cancer genes. *Sci Rep* 2016;6:35270.

Simon, N., *et al.* A sparse-group lasso. *Journal of Computational and Graphical Statistics* 2013;22(2):231-245.

Simon, N. and Tibshirani, R. Standardization and the group lasso penalty. *Statistica Sinica* 2012;22(3):983.

Sompayrac, L. How the immune system works. Hoboken, NJ: Wiley-Blackwell; 2019.

Svozil, D., *et al.* DNA conformations and their sequence preferences. *Nucleic Acids Res* 2008;36(11):3690-3706.

Trolle, T., *et al.* Automated benchmarking of peptide-MHC class I binding predictions. *Bioinformatics* 2015;31(13):2174-2181.

Vita, R., *et al.* The Immune Epitope Database (IEDB): 2018 update. *Nucleic Acids Res* 2019;47(D1):D339-D343.

Volkov, N.M., *et al.* Efficacy of immune checkpoint blockade in MUTYH-associated hereditary colorectal cancer. *Invest New Drugs* 2019.

Wang, P., *et al.* A systematic assessment of MHC class II peptide binding predictions and evaluation of a consensus approach. *PLoS Comput Biol* 2008;4(4):e1000048.

Wang, Y., *et al.* How an alloreactive T-cell receptor achieves peptide and MHC specificity. *Proceedings of the National Academy of Sciences of the United States of America* 2017;114(24):E4792-E4801.

Wijchers, P.J. and de Laat, W. Genome organization influences partner selection for chromosomal rearrangements. *Trends Genet* 2011;27(2):63-71.

Wu, J., *et al.* DeepHLApan: A Deep Learning Approach for Neoantigen Prediction Considering Both HLA-Peptide Binding and Immunogenicity. *Front Immunol* 2019;10:2559.

Xie, W.J., *et al.* Structural Modeling of Chromatin Integrates Genome Features and Reveals Chromosome Folding Principle. *Sci Rep* 2017;7(1):2818.

Yarchoan, M., *et al.* Targeting neoantigens to augment antitumour immunity. *Nat Rev Cancer* 2017;17(4):209-222.

Zhang, G.L., *et al.* Dana-Farber repository for machine learning in immunology. *J Immunol Methods* 2011;374(1-2):18-25.

Zhou, C., *et al.* pTuneos: prioritizing tumor neoantigens from next-generation sequencing data. *Genome Med* 2019;11(1):67.