Modeling Repertoire Behavior within the Energy-Based Framework: A Study on Cell Proliferation and Somatic Hypermutation during the Adaptive Immune Response

Yexing Chen[1,2*], Haiwen Ni[3], Jin Ma[1,4], Yongjie Li[1], Chen Huang[3], Sixian Yang[5],Xiangfei Xie[6], Haitao Lv[4], Peng Cao[1,2*]

[1]State Key Laboratory on Technologies for Chinese Medicine Pharmaceutical Process Control and Intelligent Manufacture, Nanjing University of Chinese Medicine, Nanjing, China.
[2]Jiangsu Provincial Medical Innovation Center, Affiliated Hospital of Integrated Traditional Chinese and Western Medicine, Nanjing University of Chinese Medicine, Nanjing, China.
[3]Department of Hematology, Affiliated Hospital of Nanjing University of Chinese Medicine, The First Clinical College of Nanjing University of Chinese Medicine, Nanjing, China.
[4]Children's Hospital of Soochow University, Suzhou, China.
[5]Key Laboratory of Quantum Information, University of Science and Technology of China, Hefei, China.
[6]Department of Computer Science, College of Engineering and Applied Science, University of Cincinnati, Cincinnati, OH, USA.

*Correspondence: xycyx@njucm.edu.cn(Y.C.);cao_peng@njucm.edu.cn(P.C.)

Monitoring and describing the adaptive immune repertoire(IR) is essential in healthcare research. Several indicators have been invented to quantify the static property of IR, while its dynamic collective behavior remains not fully studied. Here, we present a mathematical and physical model to interpret the cell proliferation and somatic hypermutation(SHM) process in IR, difficulties to generate different clones are computed and IR distance is calculated as the minimum effort required to transform one repertoire distribution to another.

IR diversity are shown to be intensive properties and remain constant across datasets with different sample size. The differences in IR observed in mice and clinical samples from various health conditions and immune stages confirm the features we have depicted. Our work dynamically characterize and quantify IR process, enabling a macroscopic immunoevaluation by sensitive immune fluctuation detection from minute samples.

Antigen recognition in adaptive immune system is achieved through B cells and T cells by their respective receptors, the B cell receptor(BCR) and T cell receptor(TCR). These receptors exhibit a high degree of diversity to ensure compatibility with the diverse range of antigens (*1,2*). Immune repertoire refers to the concept of the functional assemblies of BCR and TCR receptor, which shifts due to the selection of high-affinity cells during the adaptive immune process. Central challenges in quantitative immunology include the dynamic description of repertoire generation (*3*), summarizing the diversity of the entire repertoire (*4*), the computation of the repertoire shift (*5*), as well as the correlation between disease and repertoire behavior (*6*). Several approaches have been developed to depict the repertoire from high throughput sequencing data, such as the diversity measurements (*4*), similarity aggregating methods (*7*), and Artificial Intelligence (*8*), some methods describe the origination of diversity during cell differentiation, while some other indicators tend to describe a transient state of repertoire. However, indicators such as Shannon Index requires data normalization due to its extensive properties related to sample size, which can introduce unnecessary systematics errors(Fig. S1) (*9*). Moreover, these existing indicators are unable to demonstrate the direction of repertoire development. Thus, quantitative methods are still insufficient in repertoire research.

As the complex repertoire evolves dynamically, it is difficult to quantify the diversity or assign a scalar value to characterize the overall repertoire status, but it is easier to determine the direction and magnitude of shifts in the immune system compared to a healthy resting state. So

we first constructed a coordinate space to present the status of repertoire based on the measurable variables in HTS sequencing, then the diffculty of each point in the space for the immune system to occupy were stimulated and calculated, finally we quantified and calculated the repertoire shift as the the minimum effort required to transform a repertoire distribution to another. After blind verification with clinical and mouse data, we believe that this method can indeed describe the macroscopic state of the entire immune system and reflect its diverse activation patterns in response to different diseases.

Optimizing the presentation of repertoires to show its holistic characteristics We first built a 3D structural demonstration of repertoire, with mutation(X axis), clone kind(Y axis) and clone proliferation(Z axis) as base substrates(Fig. S2), where clones were defined as cells with the same BCR heavy chain or TCR β chain nucleotide sequence. The scale of Y and Z axis are normalized to 1. Each cell in the repertoire get its unique point occupation in the structure and the cell distribution can be intuitively embodied. Previous discoveries about the repertoire can be systematically reflected in this structure: The variations in the average and median mutation number within a repertoire offer a vivid reflection of the processes of SHM and antibody affinity maturation (*10*), mutations that occur at a frequency exceeding the random mutational background are designated as hotspot motifs (*11*), the observation that the clone exhibiting the highest frequency does not necessarily harbor the greatest number of mutations is indicative of an epitope spreading (*12*). Occupancy of various points within the 3D space engenders a spectrum of distinctive immune signatures. Therefore, we opted for this form of representation to articulate the immune repertoire.

The distinctive signatures of the immune repertoire associated with each pathology can be delineated and, following rigorous summarization and validation, can generate a multitude of bioinformatics indices. However, these indicators vary across different diseases and represent

empirical findings, and do not permit the establishment of an exclusive correspondence with particular diseases. To investigate the ubiquitous principles governing immune system activation, we chose to study the systematic difference between two different repertoire distributions rather than summarizing empirical outcomes. The key question of the study is to determine the quantitative measure of effort necessitated for the immune system to assume presence at specific points within the 3D space, and how to quantify the differences between two 3D distributions.

Calculating the difficulty to generate a clone by a scale-free preferential attachment framework To calculate the weight of each point in the space, we decided to employ computer simulation methods to study the immune activation process. We began by transforming the uneven cell proliferation process into a preferential selection model, highly proliferative cell clones have higher probabilities to add new cell members(Fig. 1A). The emergence of preference is based on the different amplification rates caused by BCR or TCR binding capacity which is determined by VDJ recombination (*13*) and SHM. Six parameters(four for TCR) are set in the model to describe the process(Fig. S3), an immune repertoire distribution is uniquely determined by these parameters (*14*). (i) the distribution of initial activating signal stimulated by antigens. (ii) the initial number of B or T cells involved in a immune response, (iii) the effective or unlethal SHM rate (*15*) in BCR which change stimulating signal randomly, (iv) the amplitude of change in stimulating signal caused by SHM, (v) the formulation to calculate the proliferation speed from stimulating signal. (vi) the overall increase in the number of cells and clones through cell division and mutation. A starting repertoire distribution is first initialized then the selection process is duplicated cell by cell to mimic the progression of the immune response, the system's attributes are extracted and analyzed throughout the process.

Real world data were collected to help validate the model(Fig. 1B). Spleen data were collected from C57BL/6 mice immunized with Candida Albicans Water Soluble fraction(CAWS)

or Virus-like particle(VLP) (*16*), meanwhile clinical data from PBMC or bone marrow blood were obtained from individuals with healthy control, infection, Kawasaki Disease, and Multiple Myeloma patients(Table. S1). BCR heavy chain and TCRβ gene were sequenced.

The rank-frequency curve of the stimulated repertoire exhibits a fat-tail power law distribution (*17*)(Fig. 1, C-D), which is consistent with the B-A preferential model that produces a scale-free network (*18*), this power law distribution is also observed in real data(Fig. 1, E-G) and previous research results (*19*). The scale-free pattern can be considered as the dimension reduction of the 3D structure on Y-Z plane after sorting(Fig. S4). The distribution of clones with varying mutation numbers, which is also the projection of the 3D structure on the X-Y plane, distributes an unimodal curve(Fig. S4), and both the average and median number of mutations per clone increase with the total number of cells(Fig. 1H), these trends are observed in different immune stages(Fig. 1, I-J) and health conditions(Fig. 1, K-M) with significant consistency. This consistency strongly implies a functional relationship between number of mutations and clone kind, suggesting that the analysis of the immune repertoire should not only be sequence-based but may also need consideration at the clonal aspects, where the clone size could be considered as a parameter of the clone in calculating the mutation rate, proportion, and biological function weight. The correlation between total cell number N_{total} and average mutation number per cell N_{Mut} in the model is almost linear(Fig. 1N).

From the perspective of mechanics, if we consider the total cell number after proliferation N_{total} in immune process as an indicator of force exerted by the immune stimulation, and the mutation number per cell N_{Mut} as a displacement of the BCR gene and whole immune system, then the effort E_{SHM} required to generate a cell with mutation number N_{Mut} is approximately proportional to N_{Mut}^2, similar to the relationship between tension and deformation of a spring according to Hooke's law. TCR as well as IgM antibodies in the early stages of the immune response do not occur the mutation process, so we expand the scope of application of the model

to cells without mutation by adding 0.1 in E_{SHM} as ground energy to avoid the occurrence of zero values as in Equation (1), ground energy pertains to the essential effort required for the maintenance of normal cellular life processes and is preliminarily and arbitrarily set to 0.1.

$$E_{\mathrm{SHM}} = \int N_{\mathrm{Mut}} * N_{\mathrm{total}} dN_{\mathrm{total}} \propto N_{\mathrm{Mut}}^2 \sim N_{\mathrm{Mut}}^2 + 0.1 \tag{1}$$

The X axis in the 3D space defines the weight of the clone, points with the same x-coordinate possess identical weight levels. The spatial distributions exhibits the direction and extent of how immune system react to different immunogenic stimuli, and can serve as an indicator for disease diagnosis.

Repertoire shift are defined and calculated as distribution distance After defining the weight of different clones in the repertoire, the repertoire shift now can be computationally quantified as the minimum effort required to transform a 3D repertoire distribution to another. Theoretically, the difference between two 3D distributions can be calculated by the 3D Wasserstein Distance. Due to the difficulty and complexity of computation process, in actual computation, we reduced the 3D distributions to a 2-dimensional Rank-Frequency curve then calculated the 2D Wasserstein distance (*20*). clone weights are set as $N_{\mathrm{Mut}}^2 + 0.1$, as calculated in Equation (1).

The diverse functions of antibodies are implemented through distinct antibody subtypes, which can be regarded as different facets in the immune response. Therefore, we modeled total repertoire in a vector space, the different subtypes of antibody genes and TCR were set as different directional components of the space, total distance between two individuals are the vector sum of repertoire distance in TCR and all antibody subtypes. Because of the class switch process, these vectors are not orthogonal to each other, at the same time, data of same subtype

are difficult to measure in clinical samples like IgE, so the distance between all accessable gene types are calculated and preliminarily use Chebyshev distance (*21*) to integrate them and evaluate the possible distance between two individuals.

Blind clustering of clinical and mouse data validates the accuracy of the model A distance matrix was built by measuring the distance between pairwise combinations of all samples, then the matrix was transformed to cluster results by t-SNE and quantified by Silhouette score(SS) (*22*) and Davies-bouldin index(DBI) (*23*). The method is unsupervised and clinical information are not used in clustering and are treated as a blind factor, the visualized clusters were labeled based on the clinical grouping for correctness validation in the last step(Fig. 2A, S5). These samples were naturally divided into several clusters, data within the group exhibited strong consistency and was consistent with their clinical grouping. The findings confirmed that the activation of the immune system displays distinctive patterns in response to different diseases.

Significance difference can be ovserved in mice with different antigen immunization like CAWS and VLP(Fig. 2B), indicating that different types of antigens can cause different activation model and immune stimulation, which lead to different immune response outcomes. Mice after VLP immunization for 3, 7, 14 and 28 days have distinct adaptive immune repertoire clusters(Fig. 2C), which proves the precision of the model in dynamic immune reaction detection during the whole immune process. With enough time point collected, it is feasible to draw the trajectory of repertoire change in the whole course of disease. The potential secondary complications and disease progression can then be detected and prevented by the deviation of the immune repertoire. These results can serve as a reference for immunological mechanism research and an identifier for the cause of disease. The VLPs capsid protein are vectors for Virus-like particle vaccines (*24*), so this method may give evaluations to vaccine effectiveness.

Clinical evidence also proved the effectiveness of this method. Differences in immune re-

sponces can be observed among primary diagnosed Multiple Myeloma(MM) and healthy controls(Fig. 2D). The MM cluster showed significant differences with healthy groups, so we believe this can be used as an early detection indicator for Multiple Myeloma. Misdiagnosis may be alleviated for some diseases in clinical practice, such as mild cases of Kawasaki disease(KD) that are easily misdiagnosed as infection(Fig. 2E), the clustering result showed the distinct difference between two diseases, and may help aiding clinicians in the rapid identification of the cause of a disease. The immune development over ontogenesis was observed by distinct clusters of healthy child and adult groups(Fig. 2F), thus the effect to immune system by age or other factors can be observed and researched, subhealthy status may then be distinguished and quantified by evaluating the background immune activation.

The model not only enables comparisons between pairs of different health conditions but also supports the simultaneous differentiation among a large number of various health status samples. The different clusters of varying health conditions separated significantly to each other(Fig. 2G), so it is feasible to draw a map containing many kind of diseases and benefits clinical diagnosis unless diseases triggers similar immune activation. Since the model is constructed based on the process analysis of the normal immune system without using any specific traits from particular disease or intervention, this approach may be extensively generalizable to diagnostic and therapeutic methods across a diverse array of pathologies, including cancers and infectious diseases.

We did not develop a machine learning classification model based on this method, as conventional classification models are only capable of identifying disease types included in the training set and lack the ability to generalize or give quantified results. The method of using false positive rate to determine accuracy is also inapplicable to this model, as this model did not employ any empirical rules as assumptions.

Speculation on the individual's overall immune status from small sample size The model holds practical value only when the results can reflect an individual's overall immune status, so we expect the clustering result to be independent of sample size. Multiple sub-datasets with varying size were resampled from sequencing data and the sample distance between them were calculated, we found the sub-datasets and the overall population holds the same distribution and the distance between them diminishes at a power-law rate with the increase of the sample size, a tenfold increase in data volume corresponds to a single-digit enhancement in the precision of estimating large samples from small samples(Fig. 3, A-C), indicating that a small data could represent the overall repertoire status of an individual. After calculating the average distance between MM patients and HC group, we found the minimum amount of sample size to tell whether an individual is in MM group or HC group to be at 10^4 level. Multiple Myeloma group have smaller distance with Kawasaki Disease group than HC group, so calculating distances from samples with $1*10^4$ sequences are not enough to distinguish Kawasaki Disease group or Multiple Myeloma group. Clustering results with all samples' data size reduced to $1*10^4$ confirmed our computational results(Fig 3D). Data size normalization is not necessary for repertoire analysis in this method, datasets could compare to each other directly regardless of sample size as they all can represent the overall macroscopic status of the individual. Differences calculated from datasets with varying sample size ranging from 20k to 120k were stable with the normalized situation and clustering results still aligned with clinical groupings(Fig 2, B-G). This gives the feasibility to estimate the overall health status of an individual from small amounts of peripheral blood, even in the level of droplets.

Discussion In this article, we provide a method to demonstrate repertoire and quantify repertoire shift. By using a scale-free network model, we modeled the cell proliferation and SHM process into a preferential attachment model with mutations. This model gives a dynamic de-

scription of immune repertoire which is accurate enough to tell the difference between various health conditions as well as dynamic changes during immune process. Rather than depict a static status of repertoire with indicators or summarize empirical rules from sequencing data and then verify their accuracy, we focus on mathematical descriptions and provide comprehensive, quantitative indicators, this ensures its effectiveness across various health states of species and enables the representation of an individual's overall health condition based on small sample sizes.

Considering the statistics property of immune system, another model using the state of ideal gas can also get identical result with that from the scale-free model (*14*). Memory cells may be seen as gas molecules with smaller mass, as it have higher activation rate when giving same amount of stimulation. Many approximations were made in both mechanical and thermodynamic model to fit the biological process and made the model not so self-consistent, and the processes of B cell class switch, the generation of immune memory, and the calming process were not included in the model, which may cause inaccuracy in modeling the whole immune process. The model still showed significant accuracy in depicting and predicting the repertoire process in several diseases dynamically over time. So we believe the model shows accuracy at a certain degree and address such models based on phase transformation, grand canonical ensemble or dissipative process can further describe the process and will amendment the residual error.

We exclusively employed cost-effective bulk HTS sequencing, excluding the need for expensive single cell techniques in the analysis, thereby guaranteeing its extensive accessibility and adaptability in clinical conditions. No specific other method was coupled with the quantitative method including wetwork library preparation or sequencing method. Any BCR heavy chain or TCR β chain data displaying a scale-free characteristic are suitable for the quantitative method.

In conclusion, we depicted the collective behavior of repertoire including B and T cells in activated immune response by modeling basic immunological cognition with mathematical distributions, and defined repertoire difference as the minimum efforts needed to transform a repertoire distribution to another. The immune system itself can be used as an indicator to detect fluctuations in changes in health status, the model may help solve the dilemma of early disease screening, misdiagnosis, disease stages identifications, potential secondary complications, and disease progression detection. The evidence-based evaluation of treatments and vaccines to regulate immune system in sick and healthy individuals can also be evaluated by monitoring the different repertoire pattern triggered. The consequent research will focus on the repertoire change in specific diseases, as the model's availability to each kind of situation should be proved independently.

References

1. J. A. Pai, A. T. Satpathy, *Nature Methods* **18**, 881 (2021).

2. E. Miho, R. Rŏskar, V. Greiff, S. T. Reddy, *Nature Communications* **10** (2019).

3. G. Altan-Bonnet, T. Mora, A. M. Walczak, *Physics Reports* **849**, 1 (2020). Quantitative immunology for physicists.

4. M. Roswell, J. Dushoff, R. Winfree, *Oikos* **130**, 321 (2021).

5. M. Brown, M. A. Schumacher, G. D. Wiens, R. G. Brennan, M. B. Rittenberg, *The Journal of Experimental Medicine* **191**, 2101 (2000).

6. F. Finotello, D. Rieder, H. Hackl, Z. Trajanoski, *Nature Reviews Genetics* **20**, 724 (2019).

7. A. M. Collins, K. J. L. Jackson, *Immunogenetics* **70**, 143 (2017).

8. J.-W. Sidhom, H. B. Larman, D. M. Pardoll, A. S. Baras, *Nature Communications* **12** (2021).

9. T. C. Hsieh, K. H. Ma, A. Chao, *Methods in Ecology and Evolution* **7** (2016).

10. L. Mesin, J. Ersching, G. Victora, *Immunity* **45**, 471 (2016).

11. M. C. Vieira, D. Zinder, S. Cobey, *Molecular Biology and Evolution* **35**, 1135 (2018).

12. C. Vanderlugt, S. Miller, *Nature Reviews Immunology* **2**, 85 (2002).

13. G. Teng, D. G. Schatz, *Molecular Mechanisms that Orchestrate the Assembly of Antigen Receptor Loci*, C. Murre, ed. (Academic Press, 2015), vol. 128 of *Advances in Immunology*, pp. 1–39.

14. Materials and methods are available as supplementary materials.

15. S. H. Kleinstein, Y. Louzoun, M. J. Shlomchik, *The Journal of Immunology* **171**, 4639 (2003).

16. T. M. Kozlovska, *et al.*, *Gene* **137 1**, 133 (1993).

17. J. Desponds, T. Mora, A. M. Walczak, *Proceedings of the National Academy of Sciences* **113**, 274 (2016).

18. A.-L. Barabási, R. Albert, *Science* **286**, 509 (1999).

19. T. Mora, A. M. Walczak, W. Bialek, C. G. Callan, *Proceedings of the National Academy of Sciences* **107**, 5405 (2010).

20. V. M. Panaretos, Y. Zemel, *Annual Review of Statistics and Its Application* (2018).

21. C. D. Cantrell, *Modern mathematical methods for physicists and engineers* (Cambridge University Press, 2000).

22. P. J. Rousseeuw, *Journal of Computational and Applied Mathematics* **20**, 53 (1987).

23. D. L. Davies, D. W. Bouldin, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **PAMI-1**, 224 (1979).

24. M. O. Mohsen, M. F. Bachmann, *Cellular and Molecular Immunology* **19**, 993 (2022).

Acknowledgments

The use of AI

Some phrases are translated by GPT, and some sentences are refined by GPT.

Funding

This paper was supported by the National Key R&D Program of China (2023YFC2308200), National Science Foundation for Distinguished Young Scholars (82125037) and the National Natural Science Foundation of China(82230120).

Author contributions

Conceptualization, Data curation, Formal analysis, Software, Validation, Visualization, Writing - original draft: YC

Funding acquisition: PC

Investigation: YC, YL, JM, CH

Methodology: YC, SY, XX

Project administration and Supervision: PC, YC

Resources: PC, HN, HL, YC

Writing - review & editing: PC, YC, XX

Competing interests

Peng Cao and Yexing Chen declare patents for the repertoire quantification, fluctuation detection and comparison algorithm in the article.

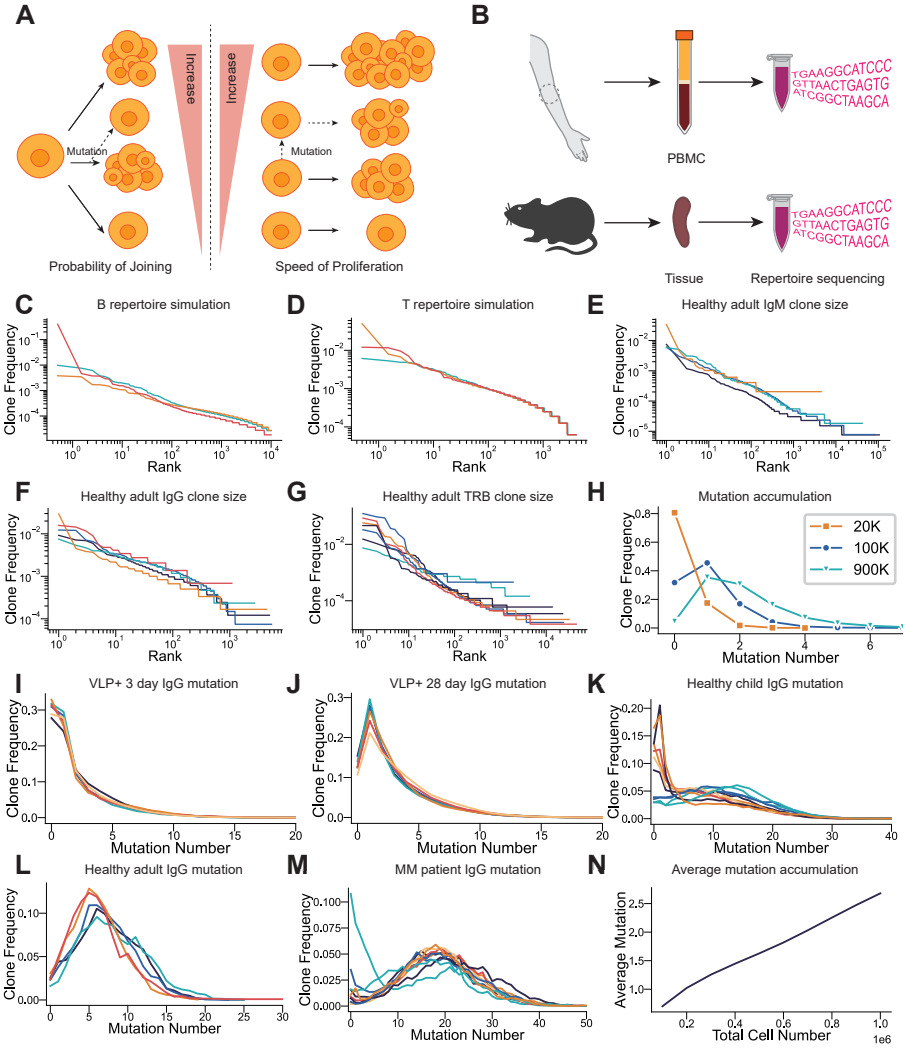
A
Mutation
Increase
Increase
Mutation
Probability of Joining
Speed of Proliferation
B
PBMC
TGAAGGCATCCC
GTTAACTGAGTG
ATCGGCTAAGCA
Tissue
Repertoire sequencing
C
B repertoire simulation
D
T repertoire simulation
E
Healthy adult IgM clone size
F
Healthy adult IgG clone size
G
Healthy adult TRB clone size
H
Mutation accumulation
20K
100K
900K
I
VLP+ 3 day IgG mutation
J
VLP+ 28 day IgG mutation
K
Healthy child IgG mutation
L
Healthy adult IgG mutation
M
MM patient IgG mutation
N
Average mutation accumulation
Clone Frequency
Rank
Mutation Number
Average Mutation
Total Cell Number
1e6

Figure 1: **The preferential attachment model simulate the rank-frequency and mutation distribution of repertoire.** (**A**) Clones with high proliferation speed equals to have high probability to attract new cells to attach to them, New clones are generated when mutation occurs. (**B**) Samples are collected from human PBMC and mouse spleen, bulk HTS sequencing are used to get repertoire data directly from PCR librarys without FACS sorting or single cell techniques. (**C, D**) BCR&TCR repertoire generated from the model show the scale free characteristics. (**E, F, G**) the scale free characteristics of IgM, IgG, and TRB repertoire data from healthy adults. (**H**) The simulated distribution of mutation-clone shifts with different repertoire size, the immune process involving more proliferation has a higher average mutation distribution. (**I, J**) C57BL/6 mice are immunized with Virus Like Particle, the mutation-clone curve in early(3 days) and late(28 days) stages shows similar distribution with the simulated repertoire. (**K, L, M**) The mutation-clone curve in clinical samples. The unexperienced(healthy child), experienced(healthy adult) and well activated(cancer patients) immune system also show the similar unimodal and shift pattern as in the simulated repertoire. (**N**) In the simulated repertoire, mutation accumulates linearly with the proliferation process.

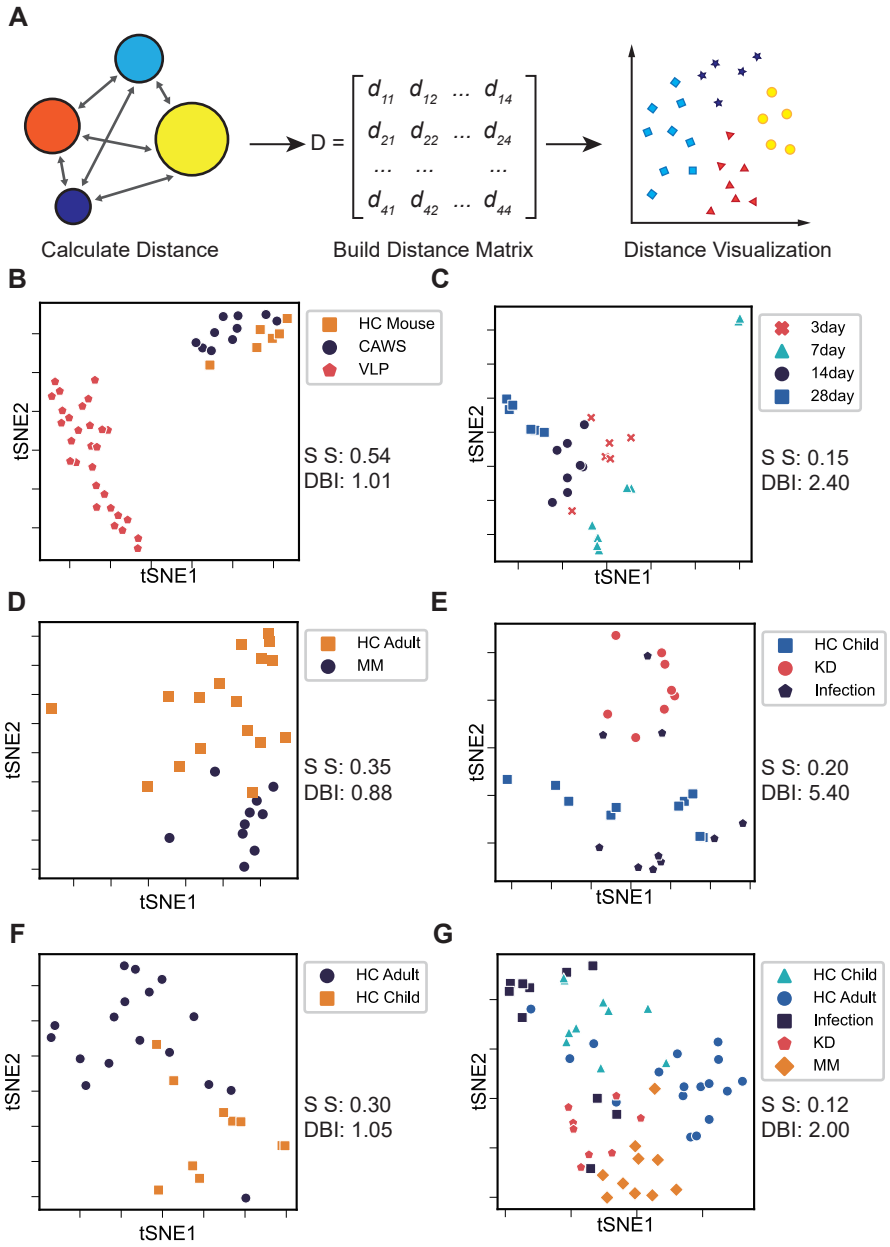
A
D =
Calculate Distance
Build Distance Matrix
Distance Visualization
B
HC Mouse
CAWS
VLP
S S: 0.54
DBI: 1.01
C
3day
7day
14day
28day
S S: 0.15
DBI: 2.40
D
HC Adult
MM
S S: 0.35
DBI: 0.88
E
HC Child
KD
Infection
S S: 0.20
DBI: 5.40
F
HC Adult
HC Child
S S: 0.30
DBI: 1.05
G
HC Child
HC Adult
Infection
KD
MM
S S: 0.12
DBI: 2.00
tSNE1
tSNE2

Figure 2: **Repertoire shift calculation and sample clustering based on repertoire distance.** (**A**) Sample distances are calculated as minimum efforts required to transform one repertoire distribution to another between any two samples, distance matrix are calculated and then blind clustered and visualized by t-SNE. (**B**) Distinct clusters are shown between C57BL/6 mice samples immunized with CAWS and VLP. (**C**) The clustering method is capable of distinguishing different immune stages in mouse VLP immunization progress. (**D**) The primary diagnosed MM shows an unique cluster result, indicating a potential detection method for cancer early diagnosis. (**E**) The different clustering of KD and infection may help solve the dilemma of misdiagnosis and stop disease progression in early stages. (**F**) The differences in healthy child and healthy adult clustering shows that the method is also capable of depicting the dynamic change in immune system development without antigen stimulation. (**G**) Clustering of clinical samples with all healthy conditions included in the study, each condition has unique immune pattern and clusters separately.

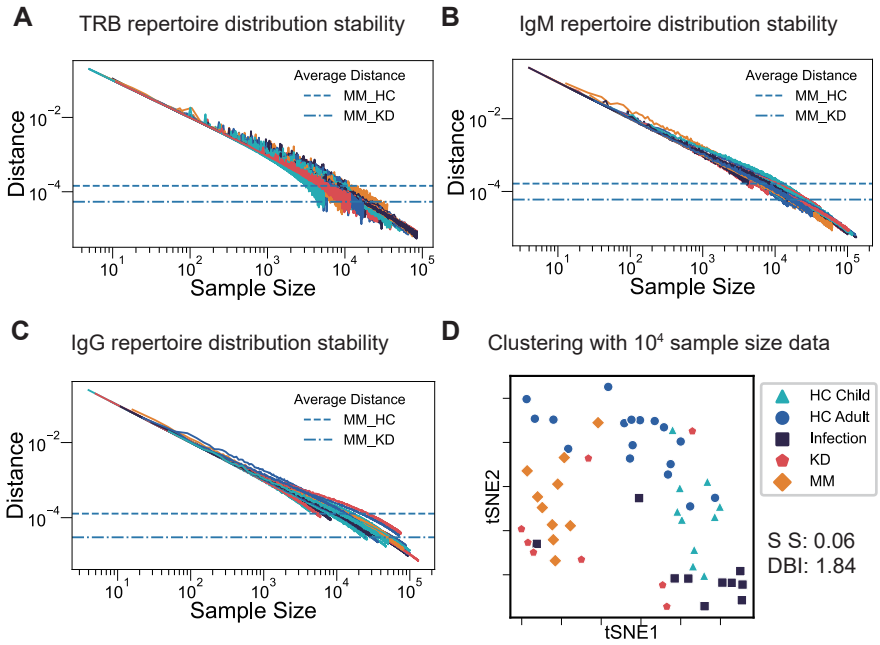

Figure 3: **Structural characteristics of repertoire.** (**A, B, C**) The distance between the overall sequenced dataset of MM, KD, and HC group and a small data subsets resampled from it decreases as the sample size increase. Dashed lines are the average distance between healthy adults and MM patients, as well as the average distance between KD patients and MM patients. The discrepancy between the data subset and the overall population decreases at a power-law rate with the increase of the sample size in IgM, IgG and TRB. (**D**) data clustering of clinical samples. data size are limited to $1*10^4$, most clusters except KD group are consistent with the clustering result with full sample size.

Table S1: Clinical information

Clinical Diagnosis	Number(M/F)	Age(median)	Sample Type
Multiple Myeloma	4/6	46-82(67)	Bone marrow
Kawasaki Disease	5/3	1-10(2)	Peripheral blood
Infection	7/3	1-9(2)	Peripheral blood
Healthy children	1/9	2-4(3)	Peripheral blood
Healthy adults	5/12	23-32(25)	Peripheral blood

Table S2: Mouse immunization information

Antigen	Number	days after immunization	Sample Type
Healthy control	6	/	spleen
CAWS	10	14	spleen
VLP	6	3	spleen
VLP	8	7	spleen
VLP	8	14	spleen
VLP	6	28	spleen

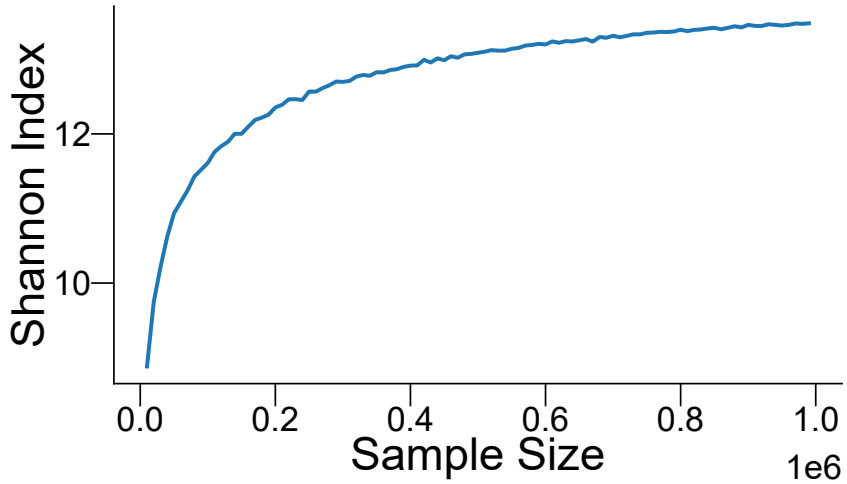

Figure S1: **Correlation between Shannon Index and sample size.** The Shannon Index value is not a constant; it increases with the augmentation of the sample size.

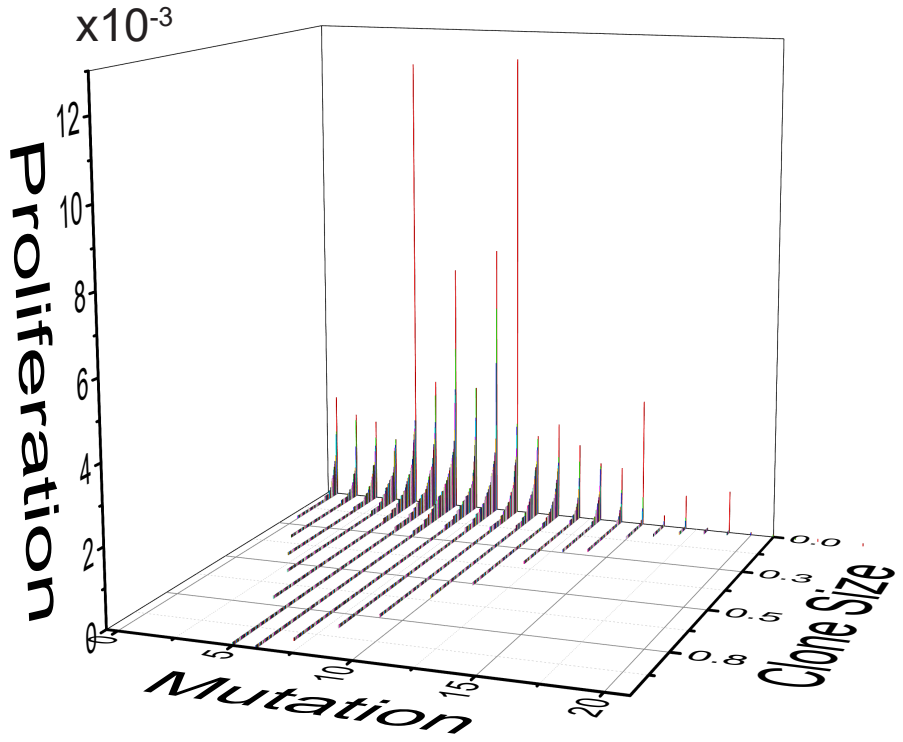

Figure S2: **A 3D structural demonstration of repertoire.** The three basis of the vector space are mutation, proliferation and clone number, each cell in repertoire occupies a point in the structure based on its BCR or TCR sequence. The 3D map is uniquely determined by a repertoire.

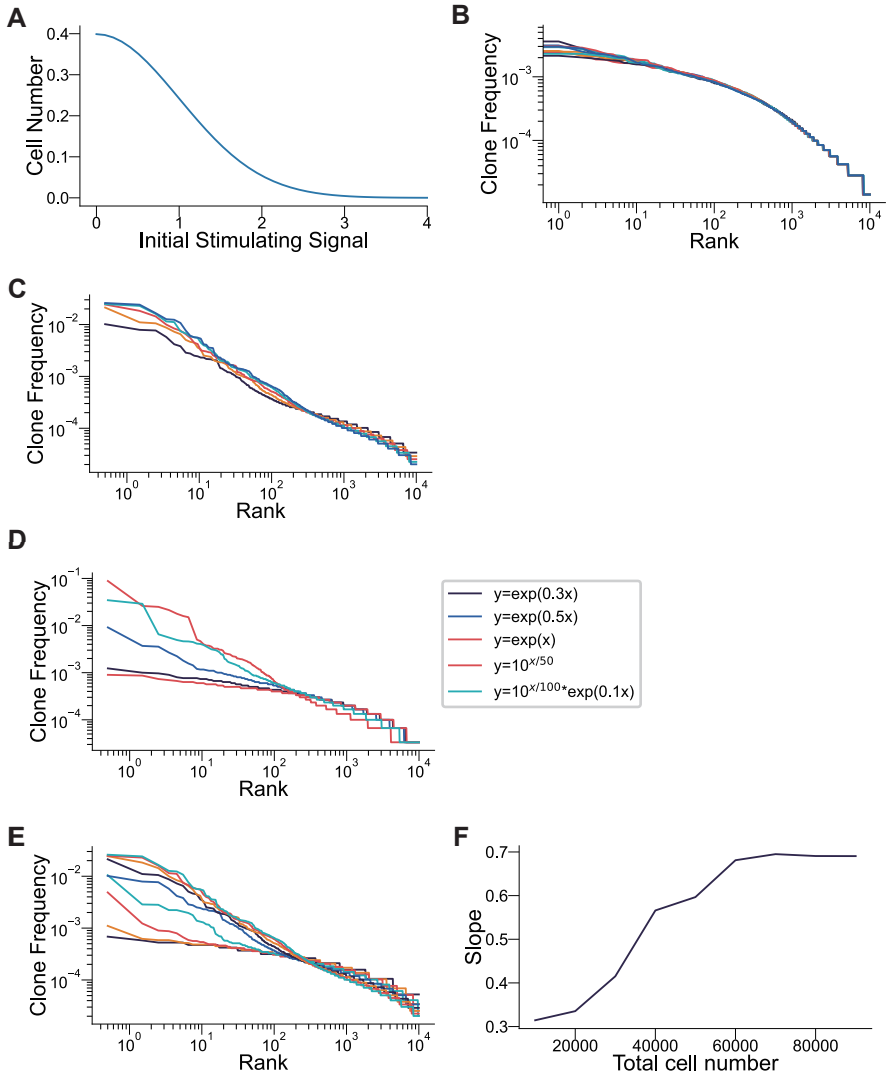

Figure S3: **Parameters affecting the scale-free distribution of simulated repertoire.** (**A**) The number of naive B or T cells with varying binding capacity and generate stimulating signal of different strength to a random antigen should follow a Gaussian distribution. (**B, C**) Stimulation starts with 1k-5k cells and proliferate to 200k cells will generate a concave distribution, while power-law distribution emerges when starting with 50-90k cells. (**D**) Different formulations calculating the proliferation speed from stimulating signal can all produce a power law distribution, the greater the derivative of the function, the steeper the corresponding curve is. (**E, F**) The scale-free distribution generated from different final cell numbers range from 10k-100k. As the number of cells increases, the slope of the curve also increases until it reaches a plateau.

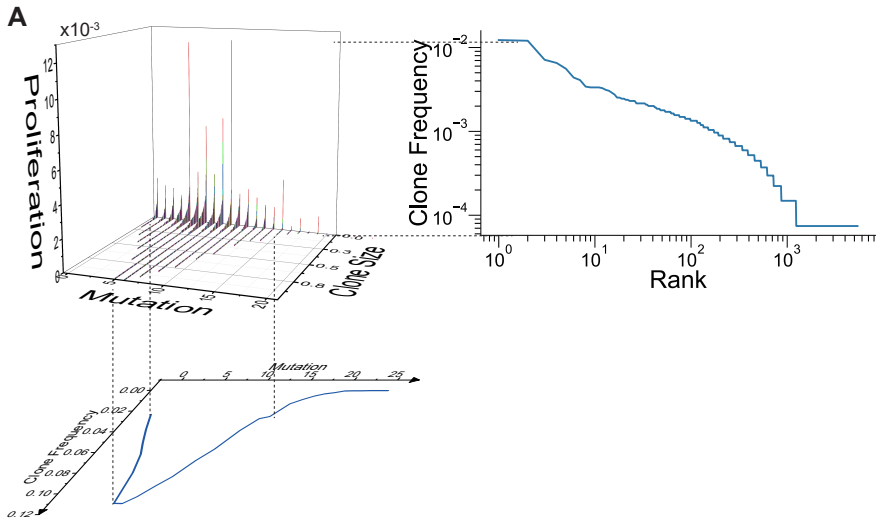

Figure S4: **Correlation between 3D structures and various distributions.** The dimensionality reduction results of the 3D structure at different angles exhibit the unimodal distribution and the scale-free distribution.

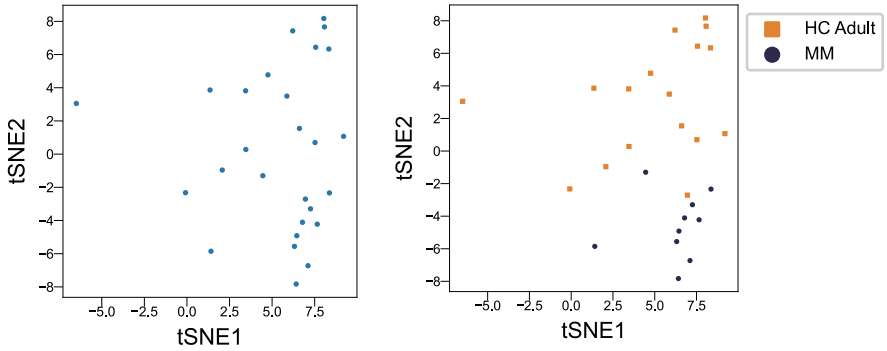

Figure S5: **Blind clustering of repertoire data.** The distances between any two samples were calculated and visualized to produce clustering results. The clinical information of HC and MM are treated as a blind factor and are subsequently incorporated to verify the accuracy of the cluster.