KSBi-BIML 2026

Bioinformatics & Machine Learning(BIML) Workshop for Life Scientists

생명정보학 & 머신러닝 워크샵 (온라인)

Introduction to cancer-immune analysis

김상우 _ 연세대학교

KSBI KOREAN SOCIETY FOR BIOINFORMATICS | 한국생명정보학회

본 강의 자료는 한국생명정보학회가 주관하는 BIML 2026 워크샵을 목적으로
제작된 것으로 해당 목적 이외의 다른 용도로 사용할 수 없음을 분명하게 알립니다.

이를 다른 사람과 공유하거나 복제, 배포, 전송할 수 없으며 만약 이러한 사항을 위반할 경우
발생하는 **모든 법적 책임은 행위자 본인에게 있음**을 알립니다.

KSBi-BIML 2026

Bioinformatics & Machine Learning (BIML) Workshop for Life Scientists

한국생명정보학회가 주최하는 BIML-2026 동계 Bioinformatics &Machine Learning 교육 워크숍에 여러분을 초대합니다.

BIML 워크숍은 생명정보학 연구자들이 최신 AI바이오 분야의 인공지능 기반 분석 기술과 바이오 데이터 분석 기법을 이론과 실습을 통해 체계적으로 배울 수 있는 전문 교육 프로그램입니다. 2015년에 시작된 BIML 워크숍은 올해로 12년 차를 맞이하며, 국내 생명정보학 분야의 최초이자 최고 수준의 교육 프로그램으로 자리 잡았습니다. 이번 워크숍은 크게 인공지능바이오(AI바이오) 분야와 디지털바이오 분야, 두 분야로 구성됩니다.

AI바이오 분야에서는 생명정보 분석에 폭넓게 응용되고 있는 다양한 인공지능 기반 자료 모델링 기법을 다룰 예정입니다. 특히, 인공지능 심층학습을 활용한 단백질 구조 예측, 유전체 분석, 신약 개발에 대한 이론 및 실습 강의가 진행됩니다.

또한 디지털바이오 분야에서는 단일세포오믹스, 공간오믹스, 멀티오믹스, 메타오믹스에 대한 강의도 마련되어 있어, 연구자들의 분석 역량 강화에 실질적인 도움을 줄 것으로 기대됩니다.

또한 2024년부터 추가된 의료정보 자료 분석을 다루는 강의를 올해도 지속해서 운영하고자 합니다. 이는 최근 의료정보 자료 분석에 관한 연구 수요 증가를 반영한 것으로, 관련 연구를 수행하는 의과학자 및 의료정보 연구자들에게 유용한 지침을 제공할 것입니다.

또한, 올해도 생명정보학 기술의 다양화에 발맞춰 온라인 강좌를 대폭 확대했습니다. 올해는 무료 강좌 10개를 포함한 총 40개 이상의 강좌가 개설되며, 연구 주제에 맞는 강좌 추천과 강연료 할인 혜택도 제공합니다.

BIML-2026는 국내 주요 연구 중심 대학의 전임 교수 및 각 분야 최고 전문가들의 강의로 구성되어 있으며, 기초 이론부터 최신 연구 동향까지 아우르는 심도 있는 교육의 장이 될 것으로 확신합니다.

여러분의 많은 관심과 참여를 기대합니다!

2026년 2월

한국생명정보학회장 류 성 호

강의개요

Introduction to Cancer Immune Analysis

암은 인간의 면역과 밀접한 관계를 가진다. 암이 처음 생겨나는 과정에서 다양한 면역을 이겨내고 무력화시키기도 하고, 암을 치료하는 과정에서도 면역이 적극적으로 활용되기도 한다. 암이 가지는 신항원 (neoantigen) 은 암 면역치료의 핵심 타겟이 되는 한편, 암 주변의 미세환경 (microenvironment) 에 따라 그 효과가 달라지기도 한다. 이렇듯, 암의 예방과 치료에 대한 핵심전략으로 떠오르는 면역과의 상관성을 분석하는 것은 암 유전체학의 매우 중요한 부분이다.

본 강의에서는 WES, RNA-seq, Single-cell 및 Spatial Transcriptomics 를 기반으로 한 암 면역성과 미세환경을 분석하는 방법에 대한 전반적인 이론과 실습을 수행한다. 이를 통해 면역치료의 타겟, 바이오마커 발굴, 종양의 면역학적 특성을 이해할 수 있다.

강의는 다음의 내용을 포함한다:

- 암 면역성과 면역치료 전략 (이론)
- DNA-seq을 이용한 종양 내 신항원 예측 분석 (이론 및 실습)
- RNA-seq 을 이용한 종양미세환경 분석 (이론 및 실습)
- Single-cell 및 spatial transcriptomics를 이용한 종양미세환경 분석 (이론 및 실습)

* 교육생준비물: 노트북 (메모리 8GB 이상, 디스크 여유공간 30GB 이상)

* 강의 난이도: 중급

* 강의: 김상우 교수 (연세대학교 의과대학) / 홍지윤 조교

Curriculum Vitae

Speaker Name: Sangwoo Kim, Ph.D.

▶ **Personal Info**

Name	Sangwoo Kim
Title	Associate Professor
Affiliation	Yonsei University College of Medicine

▶ **Contact Information**

Address	50 Yonsei-ro, Seodaemun-gu, Seoul 03722, Korea
Email	swkim@yuhs.ac

Research Interest

Translational Genomics, Variant analysis, Cancer Genomics, Bioinformatics

Educational Experience

2002	B.S. in Computer Science, KAIST, Korea
2004	M.S. in Bioinformatics, KAIST, Korea
2010	Ph.D. in Bioinformatics, KAIST, Korea

Professional Experience

2010-2013	Post-doc Research Fellow, UC San Diego, USA
2014-2020	Assistant Professor, Yonsei University College of Medicine, Korea
2021-current	Associate Professor, Yonsei University College of Medicine, Korea

Selected Publications (5 maximum)

1. Yoo-Jin Ha, Seungseok Kang, Jisoo Kim, Jun Han Kim, Se-Young Jo, and **Sangwoo Kim***, Comprehensive benchmarking and guidelines of mosaic variant calling strategies, *Nature Methods* 2023
2. Bhumsuk Keam, Min Hee Hong, Seong Hoon Shin, Seong Gu Heo, Ji Eun Kim, Hee Kyung Ahn, Yun-Gyoo Lee, Keon-Uk Park, Tak Yun, Keun-Wook Lee, Sung-Bae Kim, Sang-Cheol Lee, Min Kyung Kim, Sang Hee Cho, So Yeon Oh, Sang-Gon Park, Shinwon Hwang, Byung-Ho Nam, **Sangwoo Kim***, Hye Ryun Kim*, Hwan-Jung Yun*, *Journal of Clinical Oncology* 2023
3. Tae-Min Kim, In Seok Yang, Byung-Joon Seung, Sejoon Lee, Dohyun Kim, Yoo-Jin Ha, Mi-kyoung Seo, Ka-Kyung Kim, Hyun Seok Kim, Jae-Ho Cheong, Jung-Hyang Sur, Hojung Nam, and **Sangwoo Kim***, Cross-species Oncogenic Signatures of Breast Cancer in Canine Mammary Tumors, *Nature Communications* 2020
4. Se-Young Jo†, Eunyoung Kim†, and **Sangwoo Kim***, Impact of mouse contamination in genomic profiling of patient-derived models and best practice for robust analysis, *Genome Biology* 2019
5. Sora Kim†, Han Sang Kim†, Eunyoung Kim, Min Goo Lee, Eui-Cheol Shin, Soonmyung Paik, and **Sangwoo Kim***, Neopepsee: accurate genome-level prediction of neoantigens by harnessing sequence and amino acid immunogenicity information, *Annals of Oncology* 2018

KSBI | 한국생명정보학회

Introduction to Cancer Immune Analysis

2024 BIML
연세대학교 김상우

강의 개론

Introduction to Cancer Immune Analysis

암은 인간의 면역과 밀접한 관계를 가진다. 암이 처음 생겨나는 과정에서 다양한 면역을 이겨내고 무력화시키기도 하고, 암을 치료하는 과정에서도 면역이 적극적으로 활용되기도 한다. 암이 가지는 신항원 (neoantigen) 은 암 면역치료의 핵심 타겟이 되는 한편, 암 주변의 미세환경 (microenvironment) 에 따라 그 효과가 달라지기도 한다. 이렇듯, 암의 예방과 치료에 대한 핵심전략으로 떠오르는 면역과의 상관성을 분석하는 것은 암 유전체학의 매우 중요한 부분이다.

본 강의에서는 WES, RNA-seq, Single-cell 및 Spatial Transcriptomics 를 기반으로 한 암 면역성과 미세환경을 분석하는 방법에 대한 전반적인 이론과 실습을 수행한다. 이를 통해 면역치료의 타겟, 바이오마커 발굴, 종양의 면역학적 특성을 이해할 수 있다.

강의는 다음의 내용을 포함한다:

- 암 면역성과 면역치료 전략 (이론)
- DNA-seq을 이용한 종양 내 신항원 예측 분석 (이론 및 실습)
- RNA-seq 을 이용한 종양미세환경 분석 (이론 및 실습)
- Single-cell 및 spatial transcriptomics를 이용한 종양미세환경 분석 (이론 및 실습)

Introduction to Cancer Immune and Immunotherapy

3

Cancer Immunotherapy:

Exploit host's immune system to treat cancer

– Generate or augment an immune response against cancer

SBi 한국생명정보학회 Korean Society for Bioinformatics

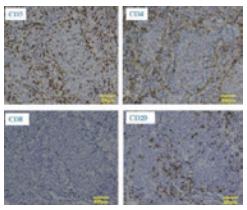
Immune and cancer

- Immunosuppressed patients have a higher risk for cancer

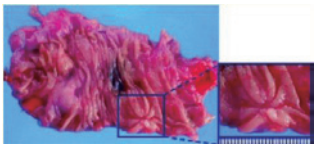
- Spontaneous regression occurs one in every 60,000 to 100,000 ca ncer cases

First visit

1 week later



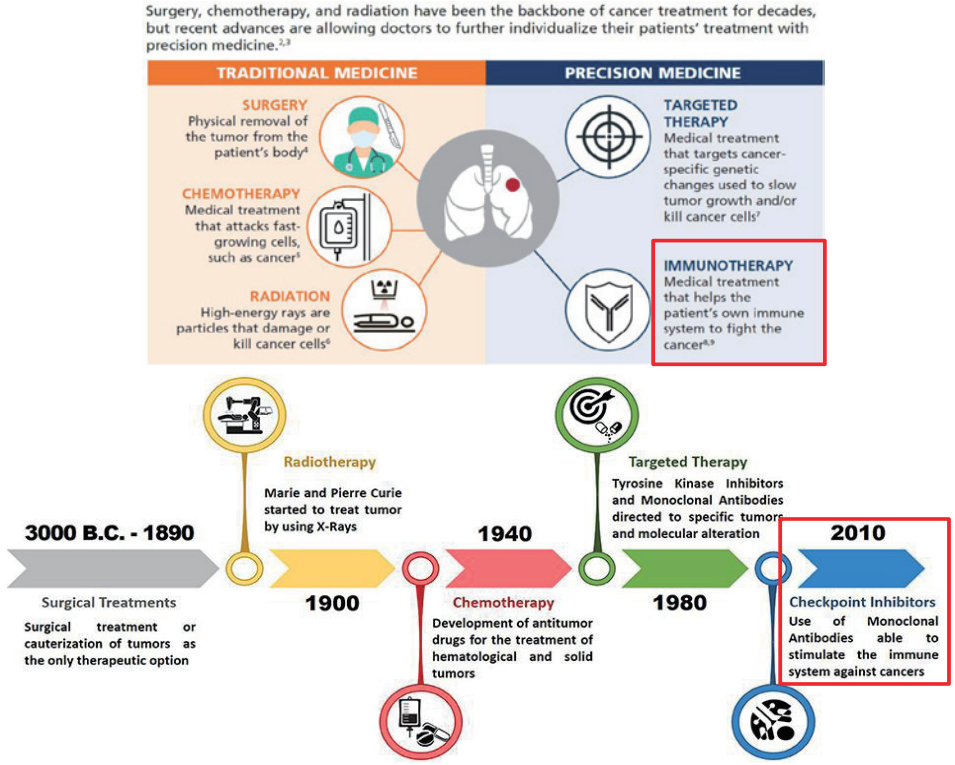
CD3+CD4+ T-cell



No cancer

Chida *et al, Surg Case Rep 2017*

Cancer Immunotherapy as a new hope



The history of immunotherapy

New York Times - July 29, 1908

ERYSIPELAS GERMS AS CURE FOR CANCER

Dr. Coley's Remedy of Mixed Toxins Makes One Disease Cast Out the Other.

MANY CASES CURED HERE

Physician Has Used the Cure for 15 Years and Treated 430 Cases— Probably 150 Sure Cures.

Following news from St. Lou's that two men have been cured of cancer in the City Hospital there by the use of a fluid discovered by Dr. William B. Coley of New York, it came out yesterday that nearly 100 cases of that supposely incurable disease have been cured in this city during the last few years, all through the use of the fluid discovered by Dr. Coley.

William Bradley Coley

erysipelas

CONTRIBUTION TO THE KNOWLEDGE OF SARCOMA.[1]

By WILLIAM B. COLEY, M.D.,

OF NEW YORK.

I. A Case of Periosteal Round-Celled Sarcoma of the Metacarpal Bone; Amputation of the Forearm; General Dissemination in Four Weeks; Death Six Weeks Later.

II. The General Course and Prognosis of Sarcoma, Based Upon an Analysis of Ninety Unpublished Cases.

III. The Treatment of Sarcoma by Inoculation With Erysipelas, With a Report of Three Recent (Original) Cases.

I. THE patient a young lady, æt. 18, had been in perfect health from earliest childhood. The family history was likewise good with the exception of a remote tubercular tendency, and the fact that an ancestor, three generations before, had died of "cancer" of the lip, presumably epithelioma.

In the early part of July, 1890, she received a slight blow upon the back of the right hand. The hand became a little swollen and somewhat painful the first night. The next few days the pain became a trifle less and the swelling subsided, but did not entirely disappear. About a week later the swelling again began to increase very slowly, and the pain became more severe. She consulted a physician at the time of the injury, but there being no evidence of anything more than an ordinary bruise the usual local applications were applied.

August 12. The pain and swelling continuing, she again sought

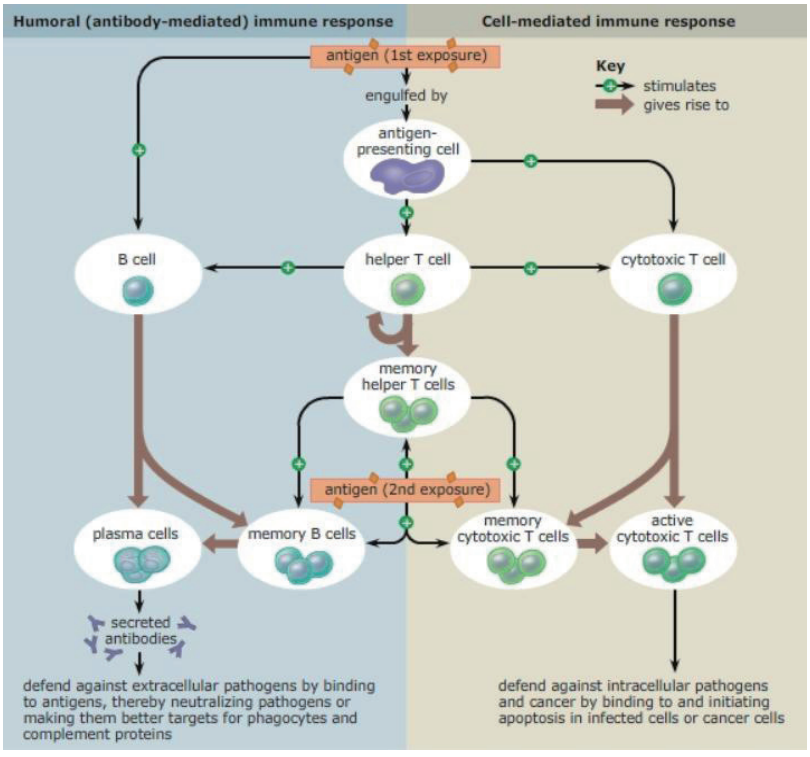
[1]Read before the Surgical Section of the New York Academy of Medicine, April 27, 1891. (With a report of three cases treated since).

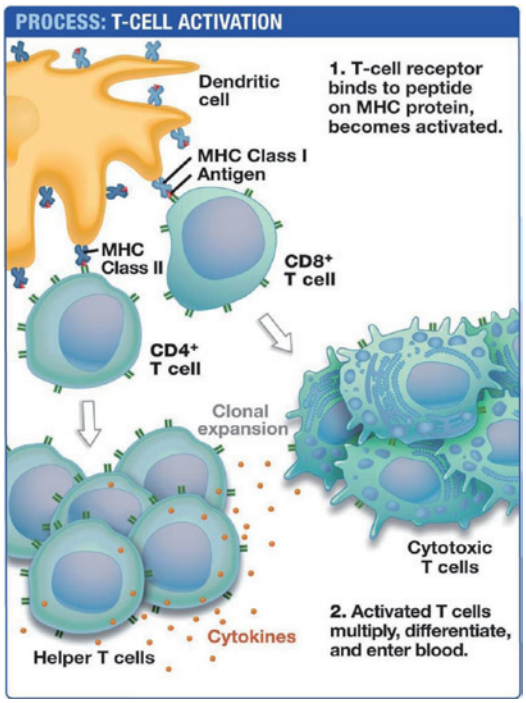
(199)

Coley, Annals of Surgery, 1981

SBi 한국생명정보학회 Korean Society for Bioinformatics

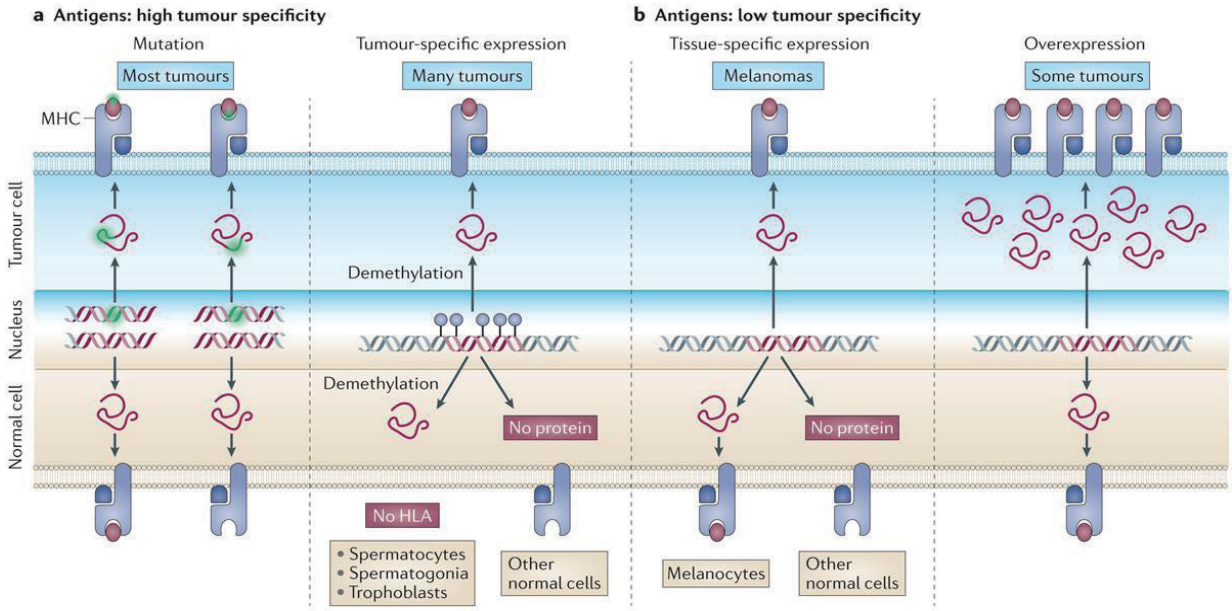
Adaptive Immunity / T-cell activation





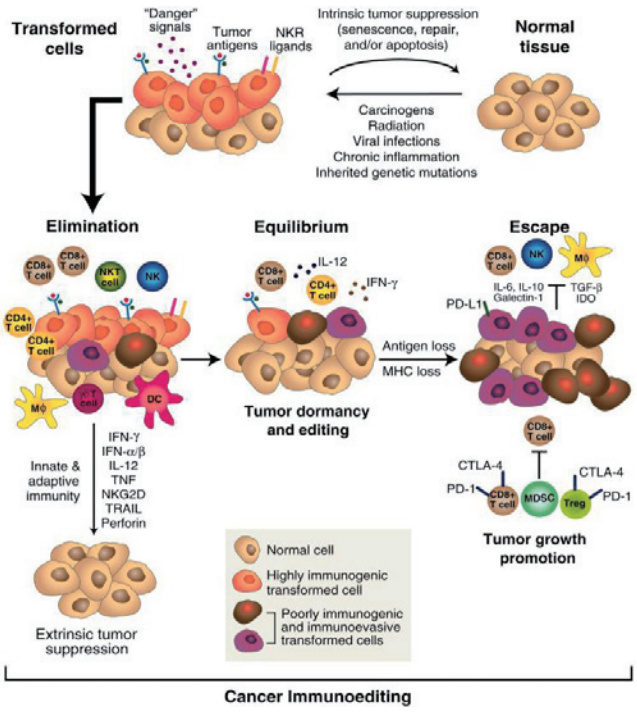
SBi 한국생명정보학회 Korean Society for Bioinformatics

Tumor Antigens



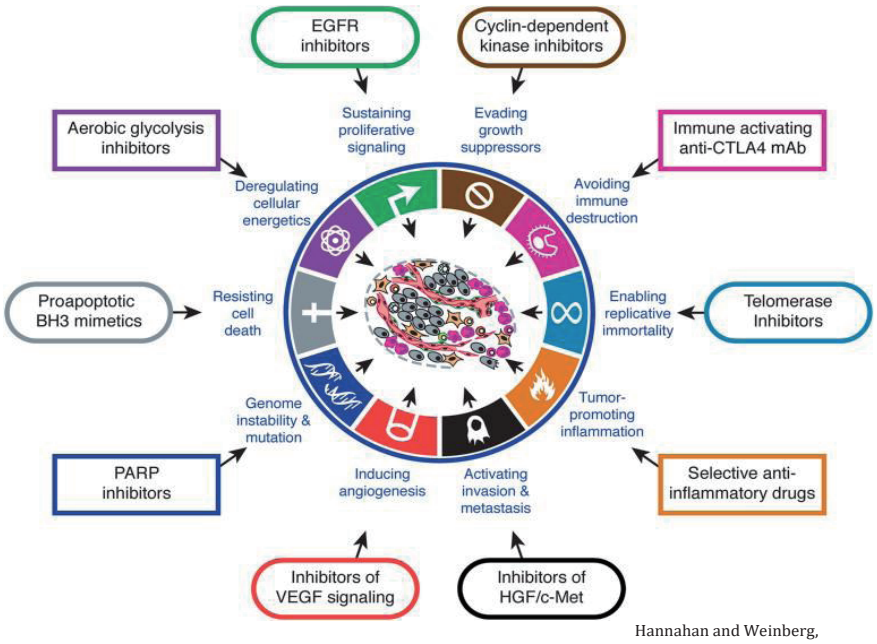
TAA (Tumor Associated Antigen): presented in tumor cells + (some normal cells)
TSA (Tumor Specific Antigen): presented only in tumor cells

Immunoediting of cancer



- **Elimination (immunosurveillance):**
  - Initial damage (possible destruction) of tumor cells by innate immune system
  - Tumor antigen presentation and attacked by CD4+, CD8+ T-cells
- **Equilibrium:**
  - Survived tumor cells do not progress and remain dormant
- **Escape:**
  - Cancer cells grow and metastasize due to the loss of control by the immune system

Immune evasion

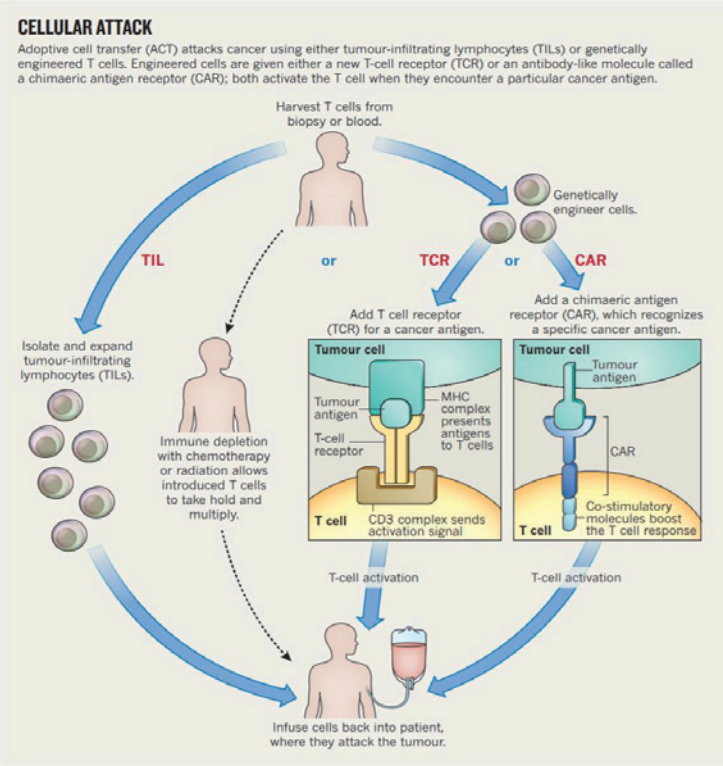


Hannahan and Weinberg, Hallmarks of cancer: The Next Generation, Cell 2011

- Paralyze CTLs and NK cells by s ecreting TGF-β or immunosup pressive factors
- Recruitment of regulatory T-ce ll (Tregs) and myeloid-derived suppressor cells (MDSCs)
- Loss of MHC class I expresssion

CURRENT APPROACHES

1. Adoptive Cell Transfer



Courtney Humpreies, Nature 504, S13-15, 2013

- TILs (tumor-infiltrating lymphocytes) - metastatic melanoma
  - tissue surrounding tumor may contain immune cells and antitumor activity
  - culture TILs and re-infuse
  - deplete endogenous immune cells
- TCR (T-cell receptor)
  - give cells new receptor
  - viral vector in patient's T-cell
  - T-cell receptor must be genetically match to the patient's immune type
- CAR (chimeric antigen receptor)
  - artificial, antibody-like protein
  - antibody (binding to cancer antigen)
  - cell activating receptor
  - stimulatory molecule

SBi 한국생명정보학회 Korean Society for Bioinformatics

Adverse effects and personalization

Table 1

Select examples of adverse events resulting from clinical application of immunotherapies targeting public antigens

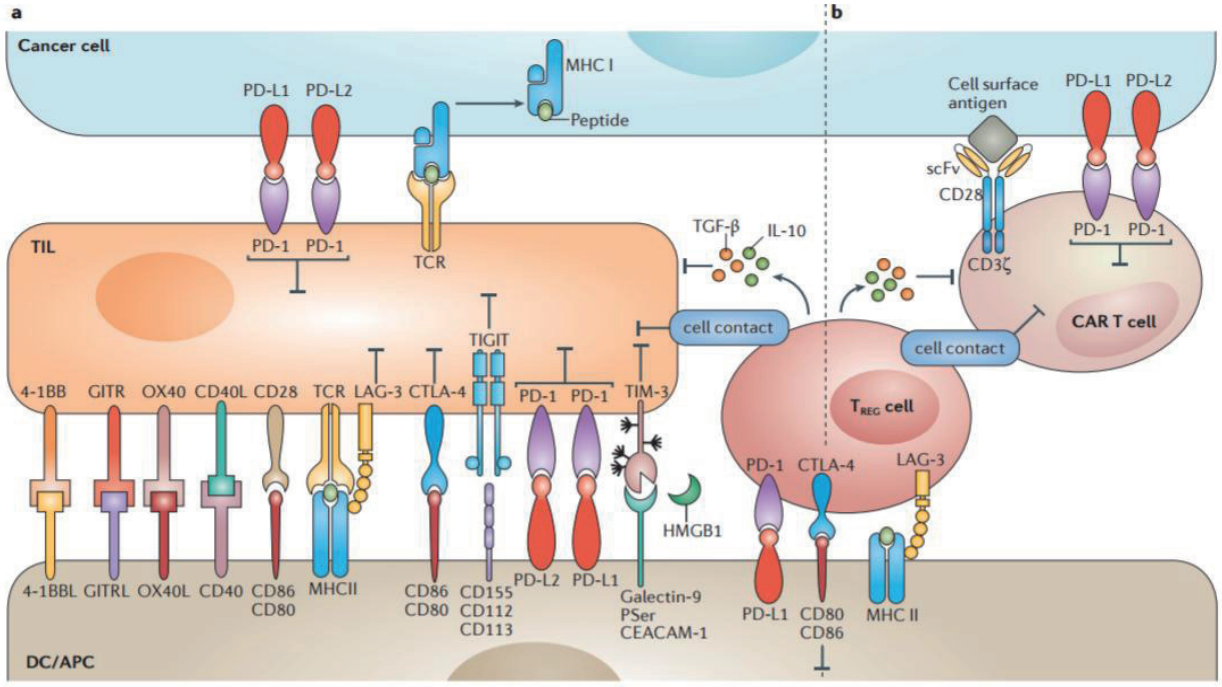
Antigen	Immunotherapy	Adverse event	Cause	Ref.
MART-1/MelanA	TCR	Fatal neural and cardiac toxicity	High levels of inflammatory cytokines alone or in combination with semi-acute heart failure and epileptic seizure	[30]
		Uveitis, Hearing loss, Loss of pigmentation	On-target activity of TCR-engineered T cells targeting normal cells expressing the cognate epitope	[24•]
	TCR + DC vaccination	Acute respiratory distress	High levels of inflammatory cytokines	[31]
NY-ESO-1	TCR (Affinity enhanced)	Skin rash with lymphocytosis, diarrheal syndrome	Autologous GVHD-like syndrome possibly due to loss of self-tolerance	[32]
MAGE-A3	TCR (Affinity enhanced)	Fatal cardiogenic shock	Cross-reactivity with an unrelated epitope from the Titin protein presented on cardiac tissue	[28]
	TCR (Affinity enhanced)	Mental status changes, comas, necrotizing leukoencephalopathy with extensive white matter defects	Reactivity to similar MAGE-A12-derived epitope presented on neural cells	[33]

- Adverse effects in ACT
  - cytokine storm
- Need to target "tumor-specific" antigen
  - Neoantigen?

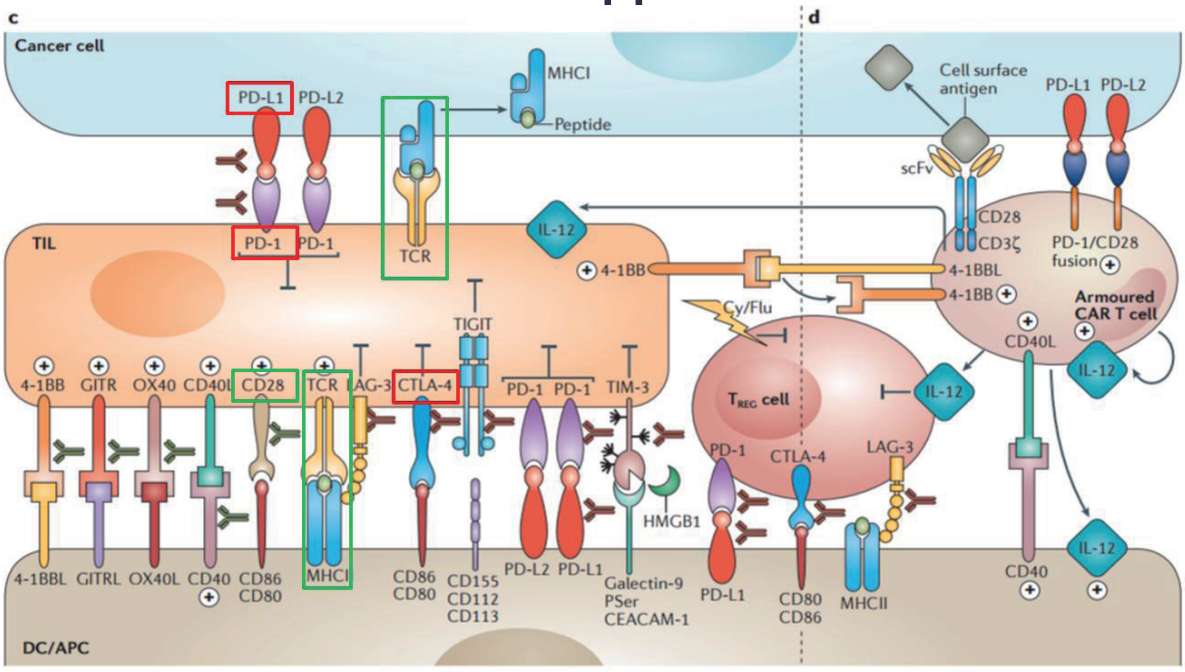
Courtney Humpreies, Nature 504, S13-15, 2013

SBi 한국생명정보학회 Korean Society for Bioinformatics

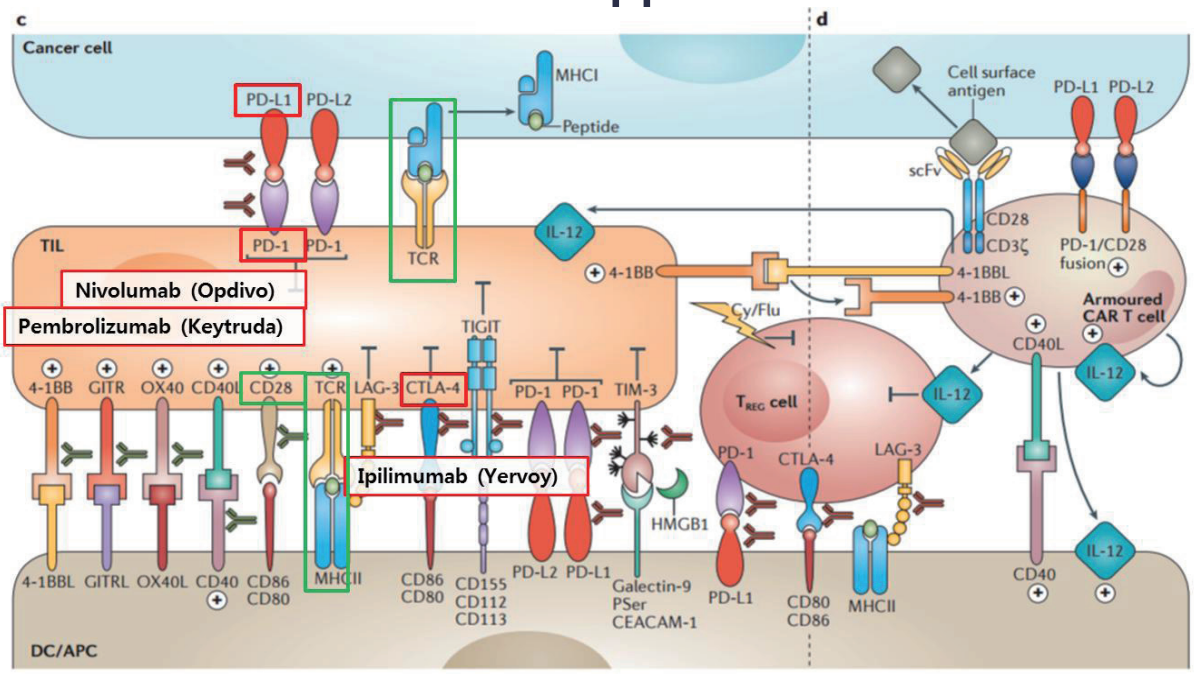
2. Checkpoint inhibitors



Immunomodulatory mABs to overcome immunosuppression

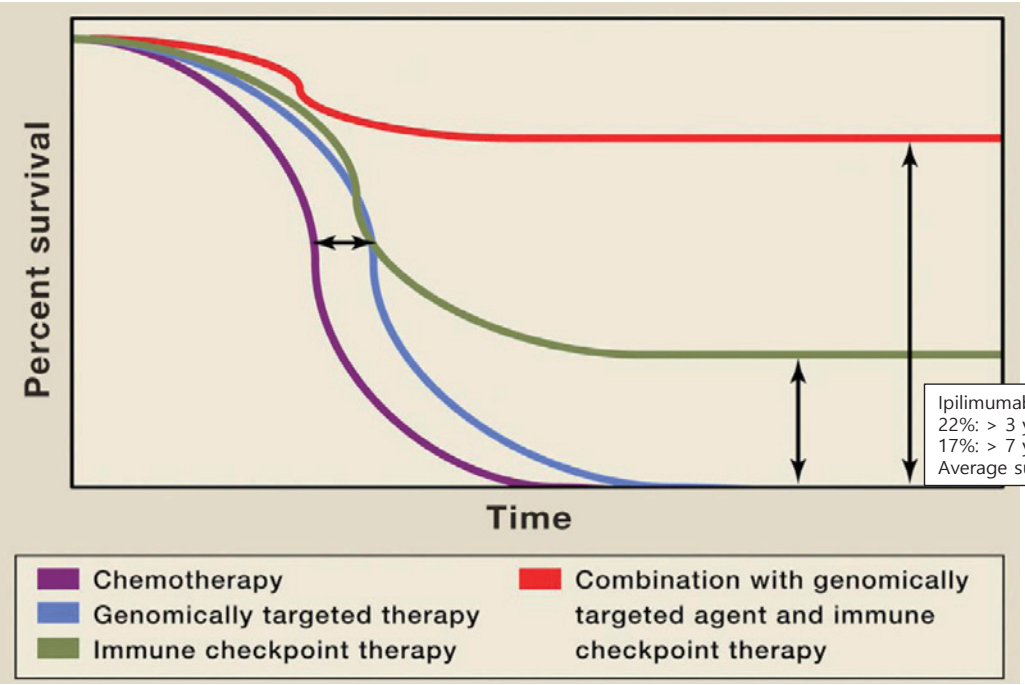


Immunomodulatory mABs to overcome immunosuppression



SBi 한국생명정보학회 Korean Society for Bioinformatics

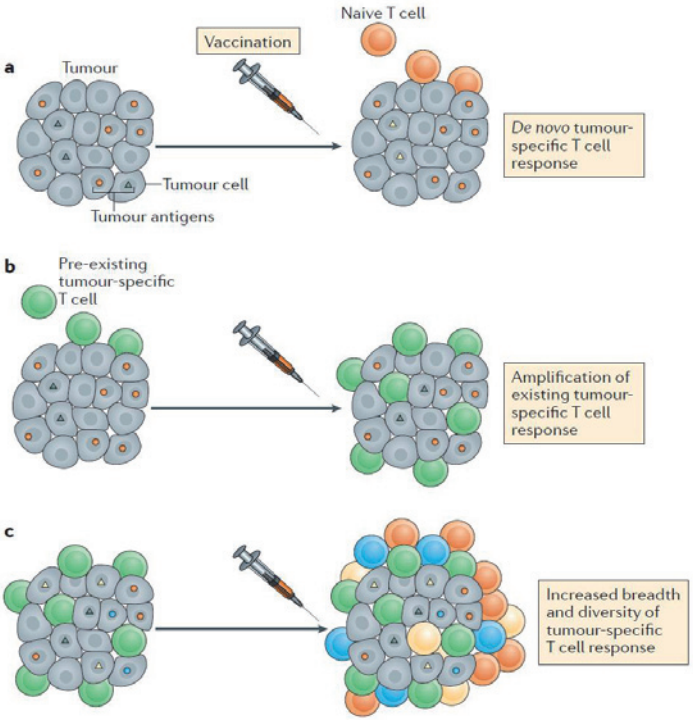
The benefits from cancer immunotherapy



Ipilimumab in melanoma (1861 patients)
22%: > 3 years
17%: > 7 years
Average survival (6-9 month to >1 yrs)

SBi 한국생명정보학회 Korean Society for Bioinformatics

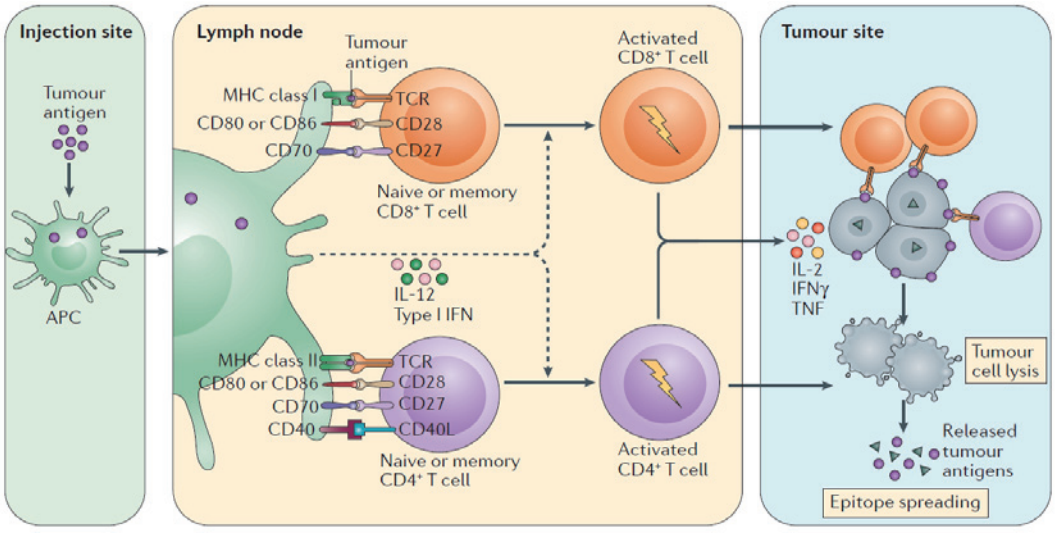
3. Cancer Vaccine



Hu *et al, Nat. Rev. Immunol 2018*

Cancer vaccines:
- Injection of tumor antigens
- generate new antigen-specific T-cell response
- amplification of existing T-cell response
- increase breadth and diversity of T-cell response

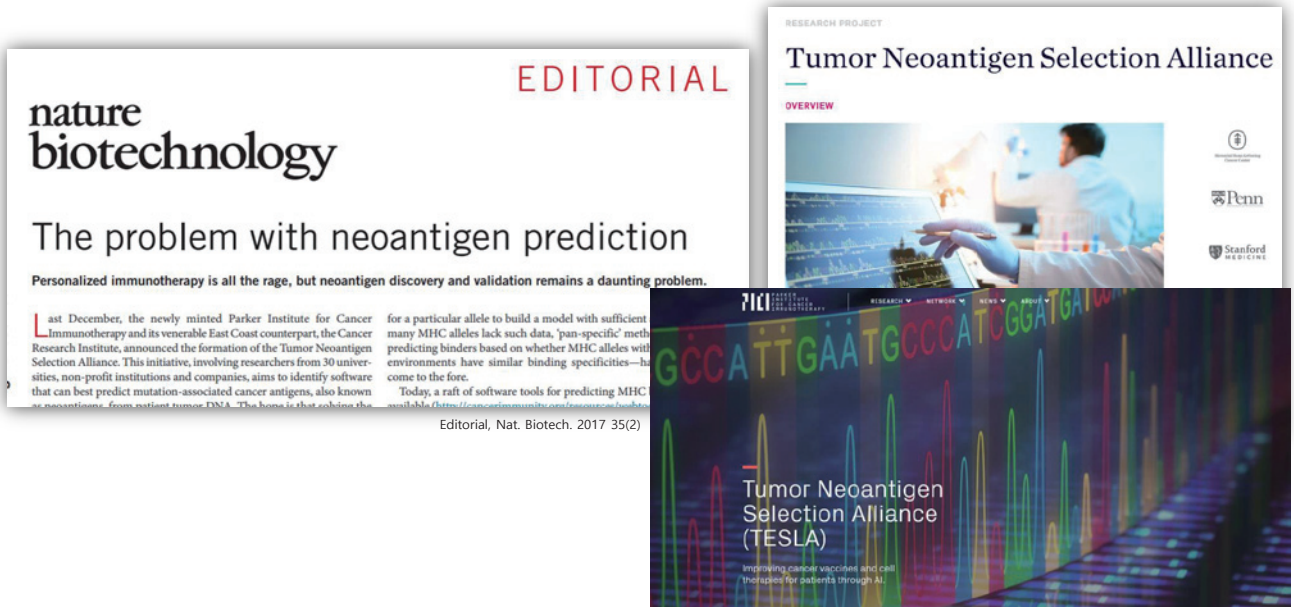
How cancer vaccine works



Hu *et al, Nat. Rev. Immunol 2018*

- Antigen injection (or DC vaccine):
- Migration of APC to present antigens to T-cells (signal 1)
- Co-stimulatory signals (signal 2)
- Migration of T-cells to tumor site
- Kill tumor cells (cytotoxicity, IFNγ, TNF..)

Neoantigen prediction is a key challenge



Editorial, Nat. Biotech. 2017 35(2)

- Neoantigen prediction for markers of checkpoint inhibitor
- Neoantigen prediction for finding tumor-specific (non-self) antigens for ACT

TUMOR MUTATION BURDEN (TMB)

Who can benefit from checkpoint inhibitor?

The NEW ENGLAND JOURNAL of MEDICINE

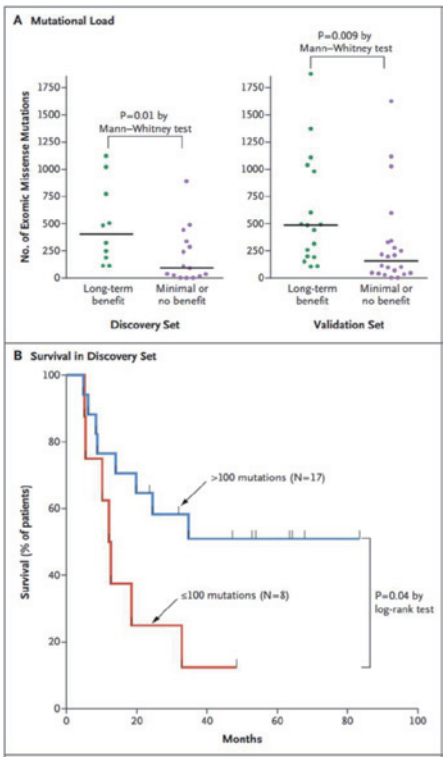
ORIGINAL ARTICLE

Genetic Basis for Clinical Response to CTLA-4 Blockade in Melanoma

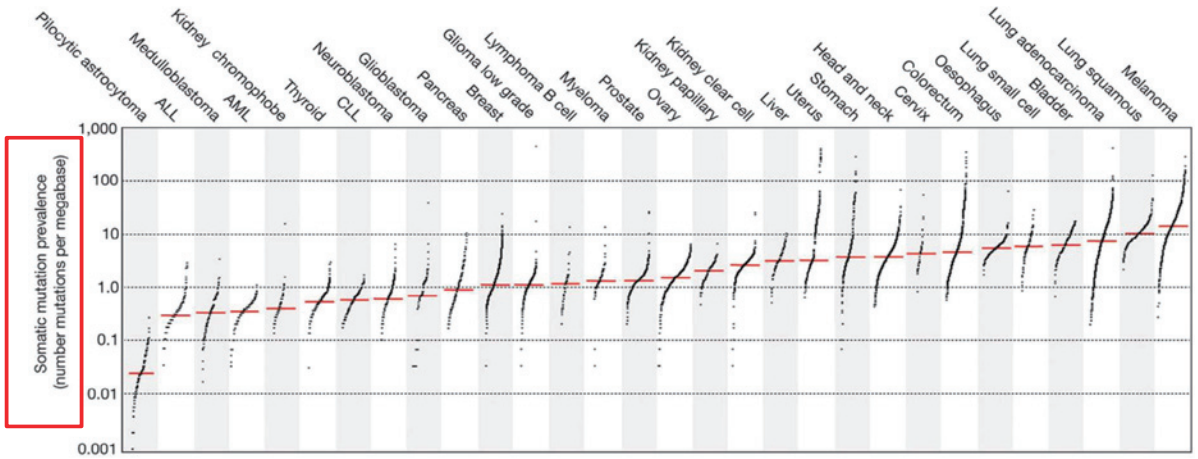
Alexandra Snyder, M.D., Vladimir Makarov, M.D., Taha Merghoub, Ph.D., Jianda Yuan, M.D., Ph.D., Jesse M. Zaretsky, B.S., Alexis Desrichard, Ph.D., Logan A. Walsh, Ph.D., Michael A. Postow, M.D., Phillip Wong, Ph.D., Teresa S. Ho, B.S., Travis J. Hollmann, M.D., Ph.D., Cameron Bruggeman, M.A., Kasthuri Kannan, Ph.D., Yanyun Li, M.D., Ph.D., Ceyhan Elipenahli, B.S., Cailian Liu, M.D., Christopher T. Harbison, Ph.D., Lisu Wang, M.D., Antoni Ribas, M.D., Ph.D., Jedd D. Wolchok, M.D., Ph.D., and Timothy A. Chan, M.D., Ph.D.

64 melanoma patients (25 discovery set, 39 validation set) treated with Ipilimumab .

Patients with high mutation burden: good survival, long-term benefit

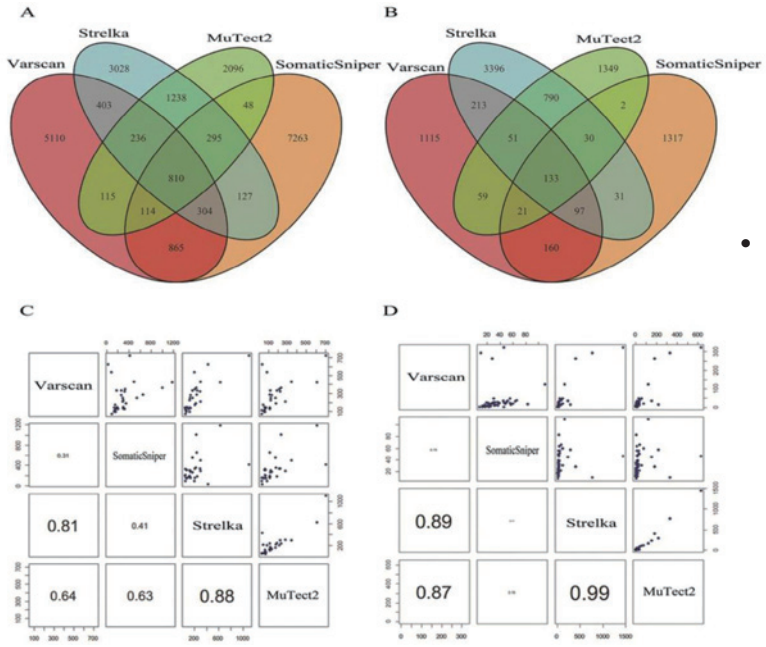


Tumor mutation burden



- Tumor Mutation Burden (TMB) $= \frac{\#total\_somatic\_mutation}{total\_targeted\_genome\_size(Mb)}$

Inconsistence of somatic mutation calls



Low depth (~50x)

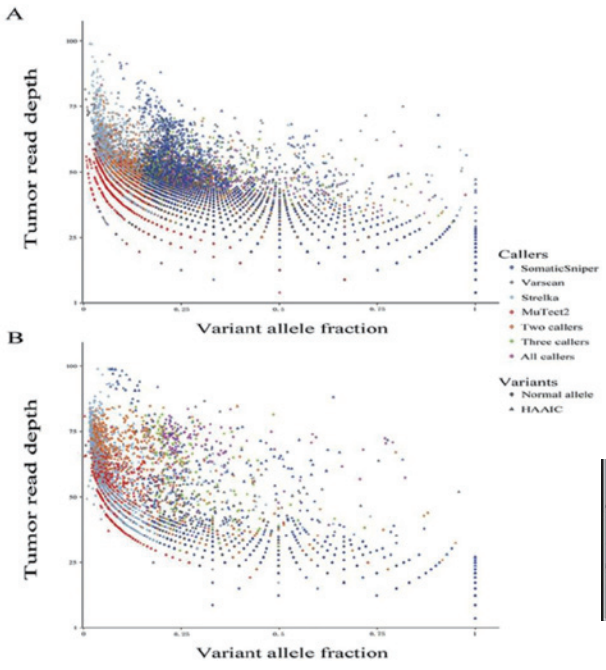
Targeted, High depth (~370x)

- The number of somatic mutations are largely dependent on the variant caller used

Cai *et al, Sci Rep. 2016*

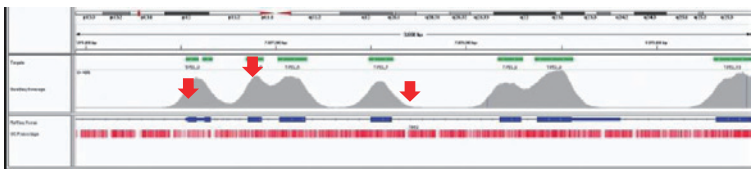
SBi 한국생명정보학회 Korean Society for Bioinformatics

Tumor mutation burden



- The number of somatic mut ations are largely dependent on the read depth

- And the read depth is simply not uniform



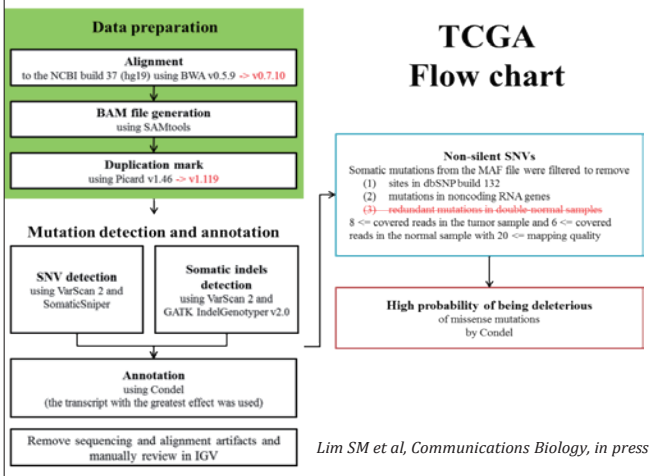
Cai *et al, Sci Rep. 2016*

SBi 한국생명정보학회 Korean Society for Bioinformatics

Fixing pipeline

mut/MB (SNV)
5.533993
5.178398
3.056166
1.459616
1.471475
1.453428
1.706356

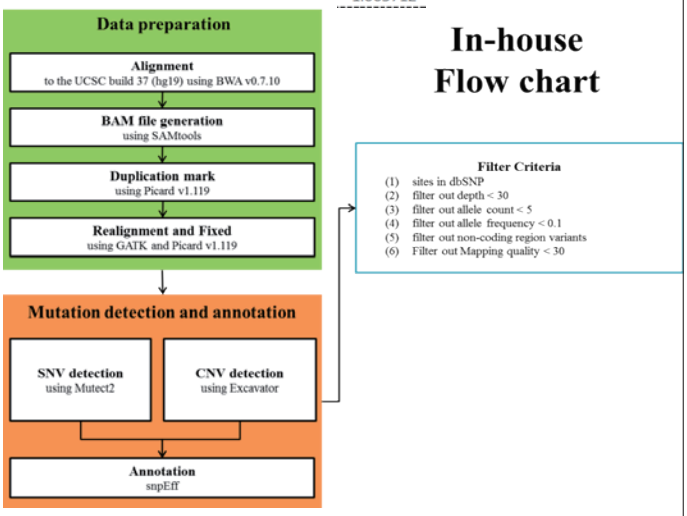
TCGA Flow chart



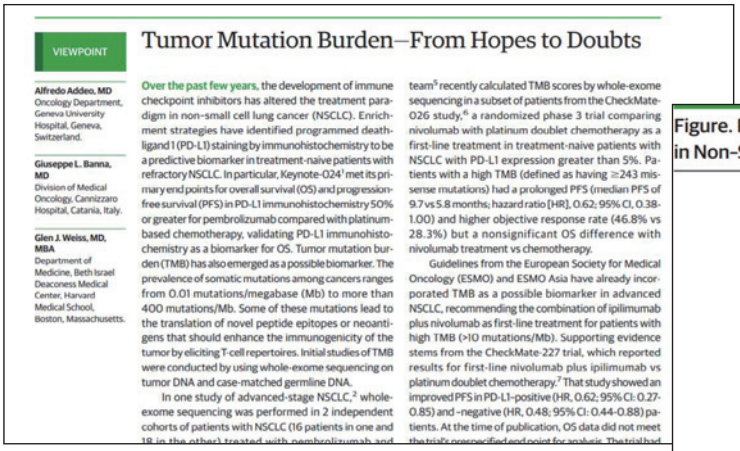
Lim SM et al, Communications Biology, in press

mut/MB
2.991871
2.4023
1.641857
1.27743
1.820113
1.108711
1.003712

In-house Flow chart



Potential pitfalls (use with care)



VIEWPOINT

Tumor Mutation Burden—From Hopes to Doubts

Alfredo Addeo, MD
Oncology Department, Geneva University Hospital, Geneva, Switzerland.

Giuseppe L. Banna, MD
Division of Medical Oncology, Cannizzaro Hospital, Catania, Italy.

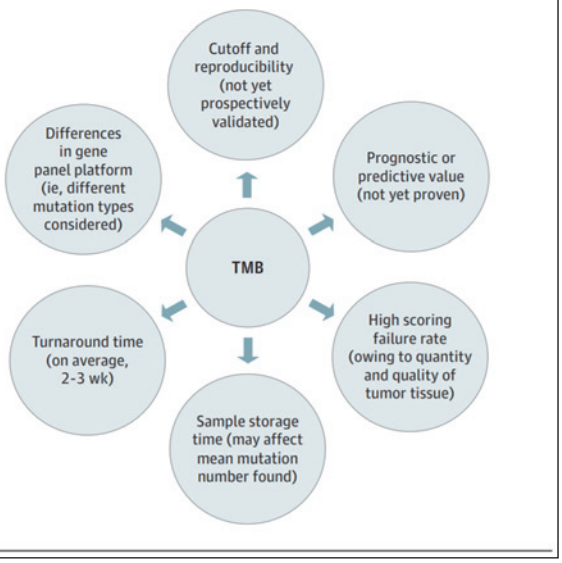
Glen J. Weiss, MD, MBA
Department of Medicine, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, Massachusetts.

Over the past few years, the development of immune checkpoint inhibitors has altered the treatment paradigm in non–small cell lung cancer (NSCLC). Enrichment strategies have identified programmed death-ligand 1 (PD-L1) staining by immunohistochemistry to be a predictive biomarker in treatment-naive patients with refractory NSCLC. In particular, Keynote-024[1] met its primary end points for overall survival (OS) and progression-free survival (PFS) in PD-L1 immunohistochemistry 50% or greater for pembrolizumab compared with platinum-based chemotherapy, validating PD-L1 immunohistochemistry as a biomarker for OS. Tumor mutation burden (TMB) has also emerged as a possible biomarker. The prevalence of somatic mutations among cancers ranges from 0.01 mutations/megabase (Mb) to more than 400 mutations/Mb. Some of these mutations lead to the translation of novel peptide epitopes or neoantigens that should enhance the immunogenicity of the tumor by eliciting T-cell repertoires. Initial studies of TMB were conducted by using whole-exome sequencing on tumor DNA and case-matched germline DNA.

In one study of advanced-stage NSCLC,[2] whole-exome sequencing was performed in 2 independent cohorts of patients with NSCLC (16 patients in one and 18 in the other) treated with pembrolizumab and team[5] recently calculated TMB scores by whole-exome sequencing in a subset of patients from the CheckMate-026 study,[6] a randomized phase 3 trial comparing nivolumab with platinum doublet chemotherapy as a first-line treatment in treatment-naive patients with NSCLC with PD-L1 expression greater than 5%. Patients with a high TMB (defined as having ≥243 missense mutations) had a prolonged PFS (median PFS of 9.7 vs 5.8 months; hazard ratio [HR], 0.62; 95% CI, 0.38-1.00) and higher objective response rate (46.8% vs 28.3%) but a nonsignificant OS difference with nivolumab treatment vs chemotherapy.

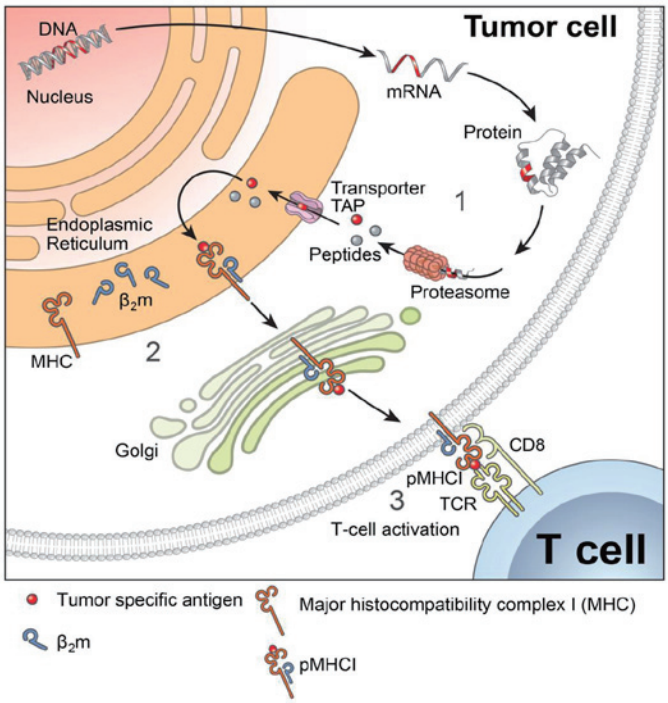
Guidelines from the European Society for Medical Oncology (ESMO) and ESMO Asia have already incorporated TMB as a possible biomarker in advanced NSCLC, recommending the combination of ipilimumab plus nivolumab as first-line treatment for patients with high TMB (>10 mutations/Mb). Supporting evidence stems from the CheckMate-227 trial, which reported results for first-line nivolumab plus ipilimumab vs platinum doublet chemotherapy.[7] That study showed an improved PFS in PD-L1–positive (HR, 0.62; 95% CI: 0.27-0.85) and –negative (HR, 0.48; 95% CI: 0.44-0.88) patients. At the time of publication, OS data did not meet the trial's prespecified endpoint for analysis. The trial had

Figure. Pitfalls of Tumor Mutation Burden (TMB) for Clinical Application in Non-Small Cell Lung Cancer



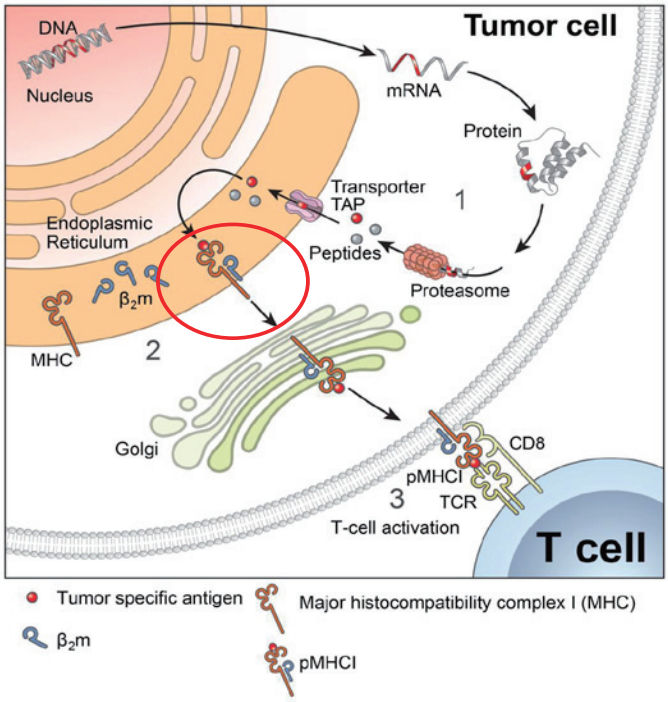
HLA TYPING IN THE ANTIGEN PROCESSING

Neoantigen processing



somatic (passenger) mutations
↓
amino acid alternation
↓
protein degradation
↓
neopeptides
↓
binding to MHC molecule
↓
peptide-MHC presentation
↓
recognition by T-cell
↓
immune response

Neoantigen processing



somatic (passenger) mutations

↓

amino acid alternation

↓

protein degradation

↓

neopeptides

↓

binding to MHC molecule

↓

peptide-MHC presentation

↓

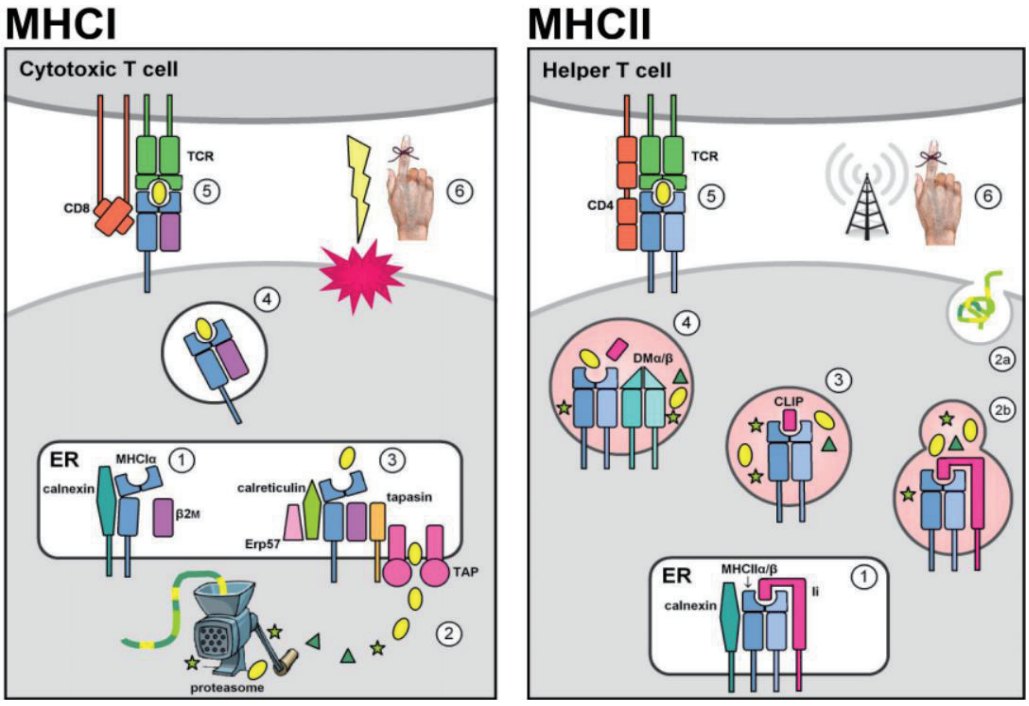
recognition by T-cell

↓

immune response

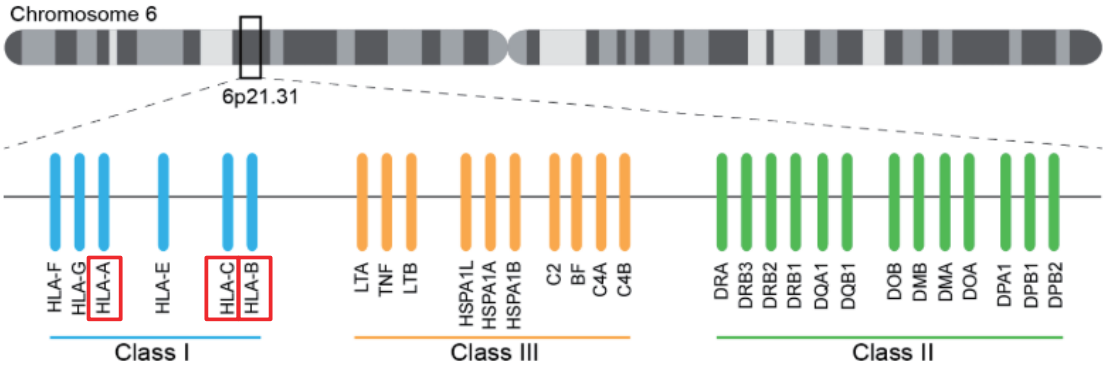
SBi 한국생명정보학회 Korean Society for Bioinformatics

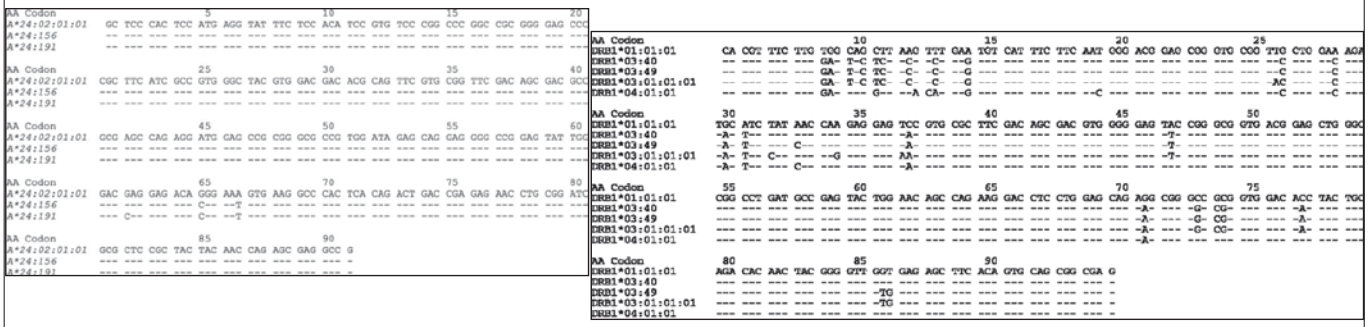
MHC (Major Histocompatibility Complex)



SBi 한국생명정보학회 Korean Society for Bioinformatics

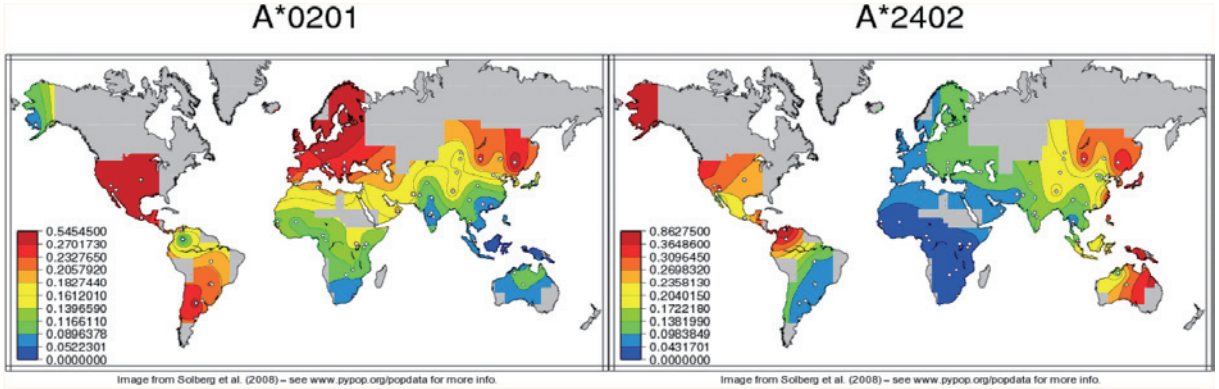
HLA (Human Leukocyte Antigen)

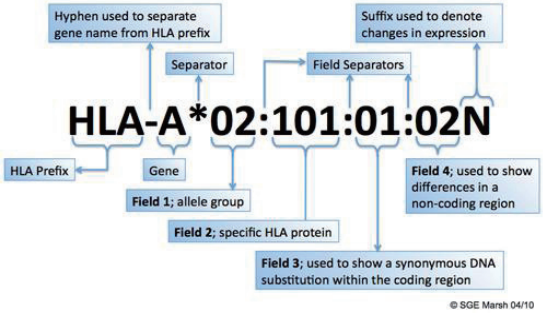




SBi 한국생명정보학회 Korean Society for Bioinformatics

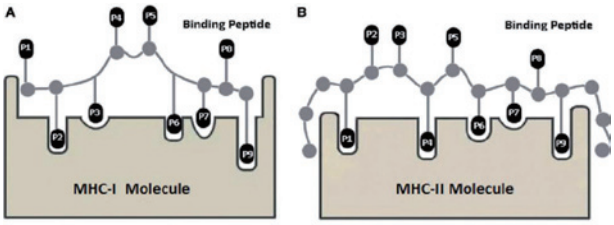
HLA alleles are ethnic specific





SBi 한국생명정보학회 Korean Society for Bioinformatics

MHC-peptide binding



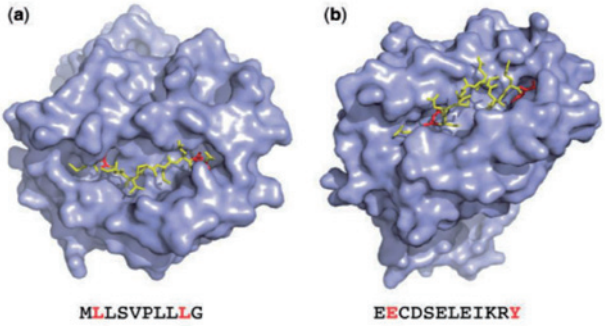
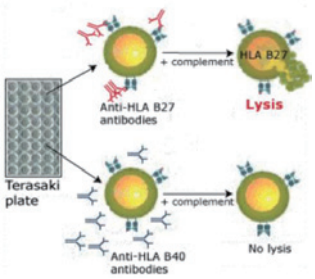


Fig. 5. 3D structures for two MHC class I molecules with bound peptides longer than 9 amino acids (PDB references 2CLR and 4JQX). (**a**) The 10mer peptide MLLSVPLLLG bound to HLA-A*02:01 extends at the C terminus with a glycine (G) amino acid. The residues at the anchor positions P2 (L) and P9 (L) are highlighted. (**b**) The 12mer EECDSELEIKRY bound to HLA-B*44:03 has anchors at its second (E) and last (Y) positions and bulges out from the middle of the MHC binding groove

But it is highly dependent on the HLA alleles
- That's why we need to know HLA allele (of the patient)

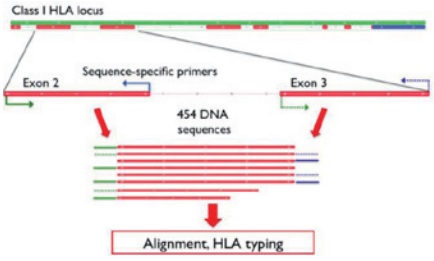
HLA typing methods

1. Serology-based typing



- Use of microcytotoxicity
  - complement mediated lysis
- Simple and low-cost
- Mostly used in HLA-A and HLA-B
- Can type allele groups an d alleles only

2. Sanger sequencing



3. Sequence-specific Oligonucleotide Hybridization (SSO)

- Amplify targeted regions with biotin-labeled primers
- Hybridized sequences emit fluorescence

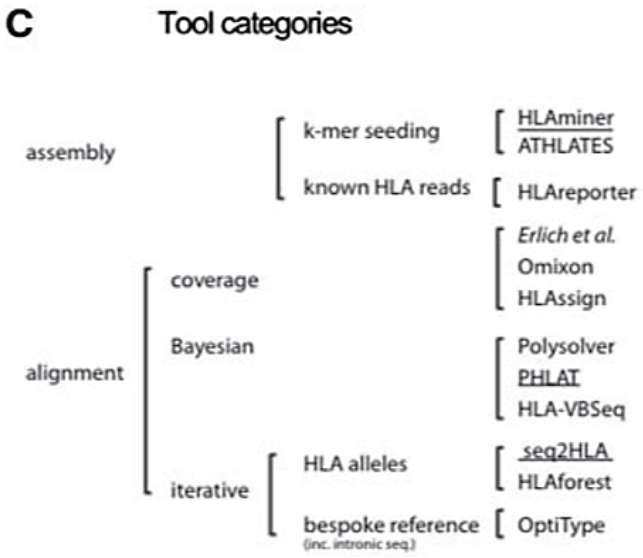
NGS-based HLA typing

- PROS
- Use of (already) produced NGS-data
- No extra-cost
- Fast

- Threat
- Short-read
- HLA genes are GC-rich: lower-sequencing coverage

SBi 한국생명정보학회 Korean Society for Bioinformatics

NGS-based HLA typing



Bauer *et al, Briefings in Bioinformatics. 2018*

SBi 한국생명정보학회 Korean Society for Bioinformatics

Assembly-based HLA typing

Warren et al. Genome Medicine 2012, 4:95
http://genomemedicine.com/content/4/12/95

Genome Medicine

METHOD Open Access

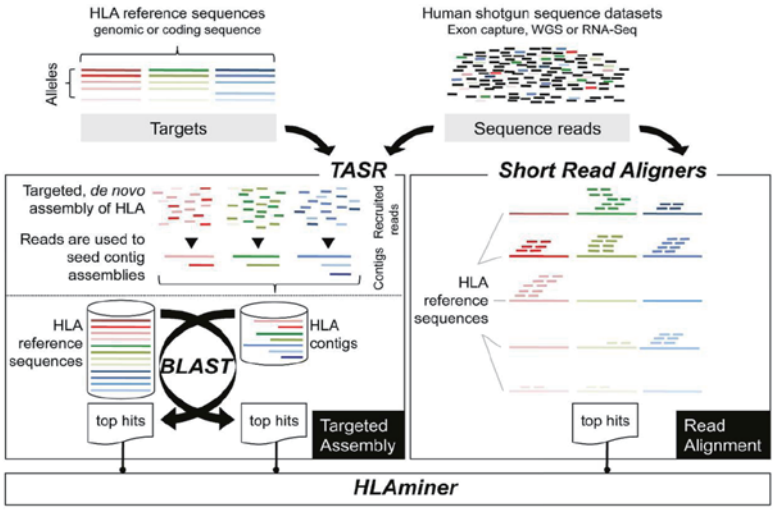
Derivation of HLA types from shotgun sequence datasets

René L Warren[1], Gina Choe[1], Douglas J Freeman[1], Mauro Castellarin[1], Sarah Munro[1], Richard Moore[1] and Robert A Holt[1,2*]

Abstract

The human leukocyte antigen (HLA) is key to many aspects of human physiology and medicine. All current sequence-based HLA typing methodologies are targeted approaches requiring the amplification of specific HLA gene segments. Whole genome, exome and transcriptome shotgun sequencing can generate prodigious data but due to the complexity of HLA loci these data have not been immediately informative regarding HLA genotype. We describe HLAminer, a computational method for identifying HLA alleles directly from shotgun sequence datasets (http://www.bcgsc.ca/platform/bioinfo/software/hlaminer). This approach circumvents the additional time and cost of generating HLA-specific data and capitalizes on the increasing accessibility and affordability of massively parallel sequencing.

HLAminer



SBi 한국생명정보학회 Korean Society for Bioinformatics

Alignment-based HLA typing

ANALYSIS

computational BIOLOGY

nature biotechnology

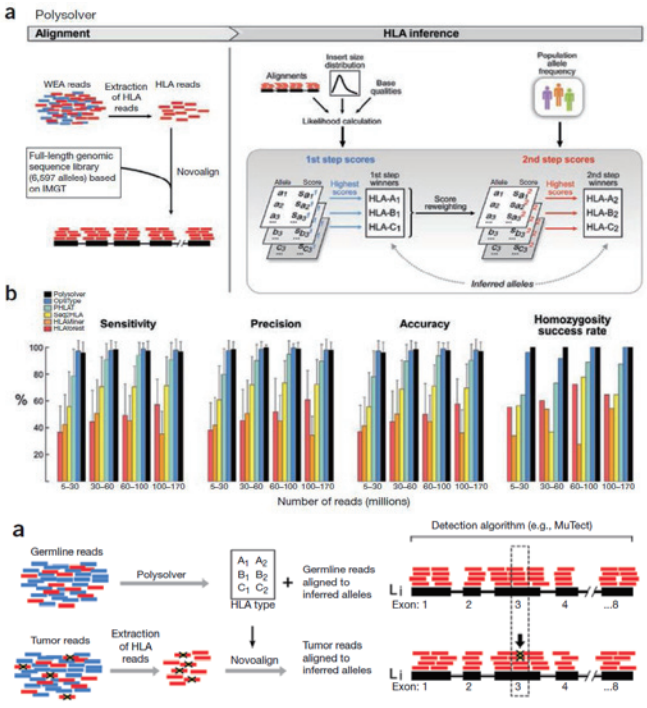
Comprehensive analysis of cancer-associated somatic mutations in class I HLA genes

Sachet A Shukla[1–3], Michael S Rooney[2,4], Mohini Rajasagi[1,5], Grace Tiao[2], Philip M Dixon[3], Michael S Lawrence[2], Jonathan Stevens[6], William J Lane[6,7], Jamie L Dellagatta[8], Scott Steelman[2], Carrie Sougnez[2], Kristian Cibulskis[2], Adam Kiezun[2], Nir Hacohen[2,8,9], Vladimir Brusic[1,3], Catherine J Wu[1,2,5,8,11] & Gad Getz[2,10,11]

Detection of somatic mutations in human leukocyte

adenocarcinoma and diffuse large B-cell lymphoma[1–3]. The HLA

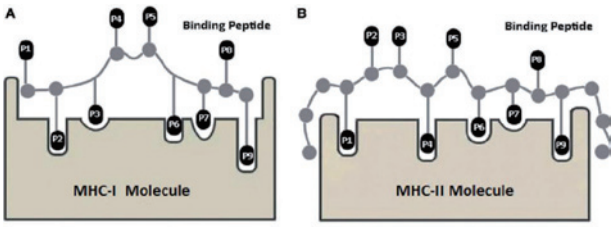
Polysolver



SBi 한국생명정보학회 Korean Society for Bioinformatics

MHC BINDING PREDICTION

MHC-peptide binding



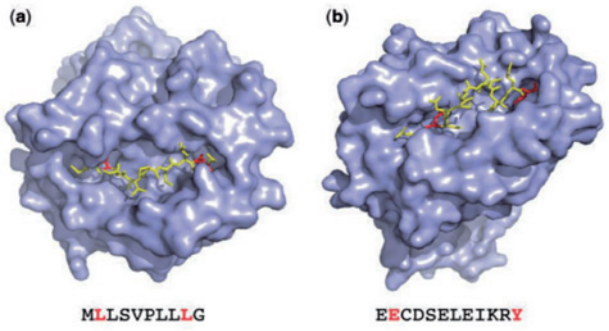


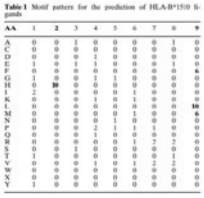
Fig. 5. 3D structures for two MHC class I molecules with bound peptides longer than 9 amino acids (PDB references 2CLR and 4JQX). (**a**) The 10mer peptide MLLSVPLLLG bound to HLA-A*02:01 extends at the C terminus with a glycine (G) amino acid. The residues at the anchor positions P2 (L) and P9 (L) are highlighted. (**b**) The 12mer EECDSELEIKRY bound to HLA-B*44:03 has anchors at its second (E) and last (Y) positions and bulges out from the middle of the MHC binding groove

Can we predict if a given peptide will bind to MHC?

Prediction algorithms

- SYFPEITHI: using PSSM


SYFPEITHI: database for MHC ligands and peptide motifs

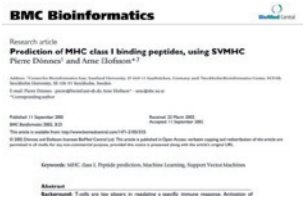


123456
ATPKAE
KKPKAA
AKPKAK
TKPKPA
AKPKT-
AKPAAK
KLPKAD
AKPKAA

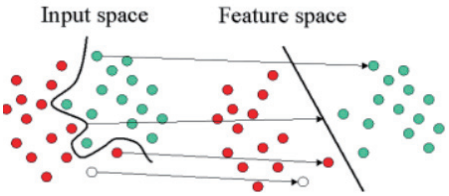
AKPKA-

Position:	1.	2.	3.	4.	5.	6.
A	0.625	0	0	1/8	6/8	3/8
D	0	0	0	0	0	1/8
E	0	0	0	0	0	1/8
K	0.25	6/8	0	7/8	0	2/8
L	0	1/8	0	0	0	0
P	0	0	1	0	1/8	0
T	1/8	1/8	0	0	1/8	0
-	0	0	0	0	0	1/8
Sum	1	1	1	1	1	1

- SVMHC: using Support Vector Machine

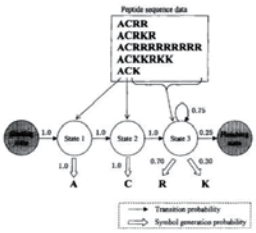

BMC Bioinformatics

Prediction of MHC class I binding peptides, using SVMHC



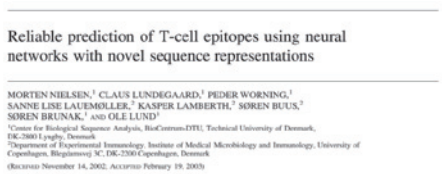
- S-HMM: using Hidden Markov Model

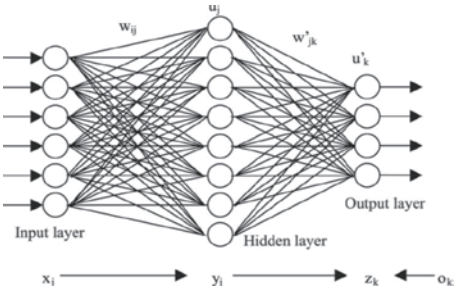

Hidden Markov Model-Based Prediction of Antigenic Peptides That Interact with MHC Class II Molecules



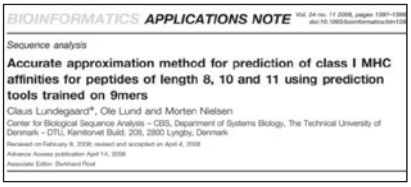
ANN based algorithms

NetMHC: Classification of MHC-I binding peptides using ANN

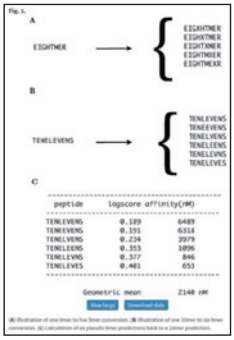

Reliable prediction of T-cell epitopes using neural networks with novel sequence representations



NetMHC-3.0

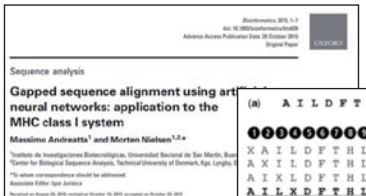

BIOINFORMATICS APPLICATIONS NOTE

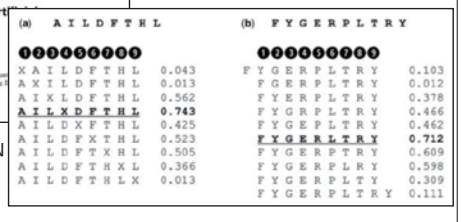
Accurate approximation method for prediction of class I MHC affinities for peptides of length 8, 10 and 11 using prediction tools trained on 9mers



Approximation of 8, 10, 11 from 9 mer model

NetMHC-4.0

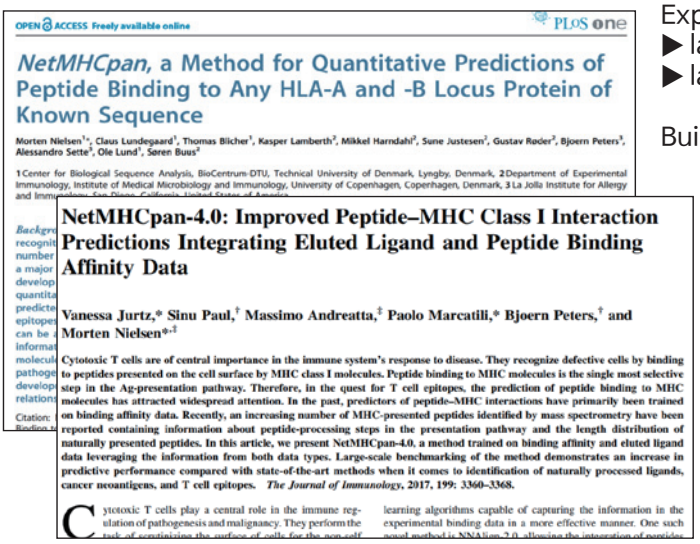

Gapped sequence alignment using artificial neural networks: application to the MHC class I system



Gapped alignment to ANN
: 9 to 8~11 mer

Regarding all HLA-types at once

NetMHCpan: Prediction on all HLA-A/B alleles, simultaneously



OPEN ACCESS Freely available online PLoS one

NetMHCpan, a Method for Quantitative Predictions of Peptide Binding to Any HLA-A and -B Locus Protein of Known Sequence

Morten Nielsen[1*], Claus Lundegaard[1], Thomas Blicher[1], Kasper Lamberth[2], Mikkel Harndahl[2], Sune Justesen[2], Gustav Røder[2], Bjoern Peters[3], Alessandro Sette[3], Ole Lund[1], Søren Buus[2]

1 Center for Biological Sequence Analysis, BioCentrum-DTU, Technical University of Denmark, Lyngby, Denmark, 2 Department of Experimental Immunology, Institute of Medical Microbiology and Immunology, University of Copenhagen, Copenhagen, Denmark, 3 La Jolla Institute for Allergy and Immunology, San Diego, California, United States of America

NetMHCpan-4.0: Improved Peptide–MHC Class I Interaction Predictions Integrating Eluted Ligand and Peptide Binding Affinity Data

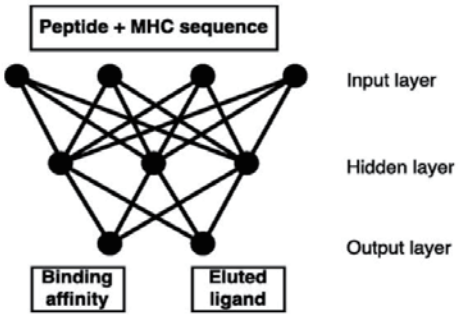
Vanessa Jurtz,* Sinu Paul,† Massimo Andreatta,‡ Paolo Marcatili,* Bjoern Peters,† and Morten Nielsen*,‡

Cytotoxic T cells are of central importance in the immune system's response to disease. They recognize defective cells by binding to peptides presented on the cell surface by MHC class I molecules. Peptide binding to MHC molecules is the single most selective step in the Ag-presentation pathway. Therefore, in the quest for T cell epitopes, the prediction of peptide binding to MHC molecules has attracted widespread attention. In the past, predictors of peptide–MHC interactions have primarily been trained on binding affinity data. Recently, an increasing number of MHC-presented peptides identified by mass spectrometry have been reported containing information about peptide-processing steps in the presentation pathway and the length distribution of naturally presented peptides. In this article, we present NetMHCpan-4.0, a method trained on binding affinity and eluted ligand data leveraging the information from both data types. Large-scale benchmarking of the method demonstrates an increase in predictive performance compared with state-of-the-art methods when it comes to identification of naturally processed ligands, cancer neoantigens, and T cell epitopes. *The Journal of Immunology*, 2017, 199: 3360–3368.

Cytotoxic T cells play a central role in the immune regulation of pathogenesis and malignancy. They perform the task of scrutinizing the surface of cells for the non-self learning algorithms capable of capturing the information in the experimental binding data in a more effective manner. One such novel method is NNAlign_2.0, allowing the integration of peptides

Experimental data are biased to major HLA alleles
▶ lack of training data in rare alleles
▶ lack of accuracy

Build a classifier that work on HLA-peptide pair



SBi 한국생명정보학회 Korean Society for Bioinformatics

Too many methods. Need a consensus

NetMHCcons: Prediction on all HLA-A/B alleles, simultaneously

Immunogenetics (2012) 64:177–186
DOI 10.1007/s00251-011-0579-8

ORIGINAL PAPER

NetMHCcons: a consensus method for the major histocompatibility complex class I predictions

Edita Karosiene · Claus Lundegaard · Ole Lund · Morten Nielsen

Received: 2 July 2011 / Accepted: 28 September 2011 / Published online: 20 October 2011
© Springer-Verlag 2011

Abstract A key role in cell-mediated immunity is dedicated to the major histocompatibility complex (MHC) molecules that bind peptides for presentation on the cell surface. Several *in silico* methods capable of predicting peptide binding to MHC class I have been developed. The accuracy of these methods depends on the data available characterizing the binding specificity of the MHC molecules. It has, moreover, methods for alleles with more remote neighbours. The final method, NetMHCcons, is publicly available at www.cbs.dtu.dk/services/NetMHCcons, and allows the user in an automatic manner to obtain the most accurate predictions for any given MHC molecule.

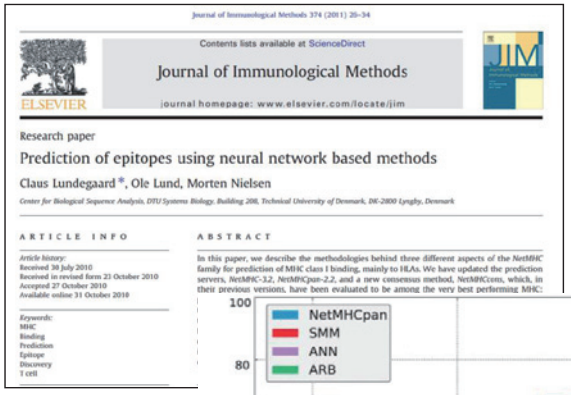
Keywords MHC class I · T cell epitope · MHC binding

$$\text{NetMHCcons} = \begin{cases} \text{NetMHCpan} & \text{for } N_p < 50 \text{ and } N_b < 10 \\ \text{NetMHC} + \text{NetMHCpan} & \text{otherwise} \end{cases}$$

We demonstrate that a simple combination of NetMHC and NetMHCpan gives the highest performance when the allele in question is included in the training and is characterized by at least 50 data points with at least ten binders. Otherwise, NetMHCpan is the best predictor.

SBi 한국생명정보학회 Korean Society for Bioinformatics

Benchmarks and competitions



Journal of Immunological Methods 374 (2011) 26–34

Contents lists available at ScienceDirect

Journal of Immunological Methods

journal homepage: www.elsevier.com/locate/jim

Research paper

Prediction of epitopes using neural network based methods

Claus Lundegaard *, Ole Lund, Morten Nielsen

Center for Biological Sequence Analysis, DTU Systems Biology, Building 208, Technical University of Denmark, DK-2800 Lyngby, Denmark

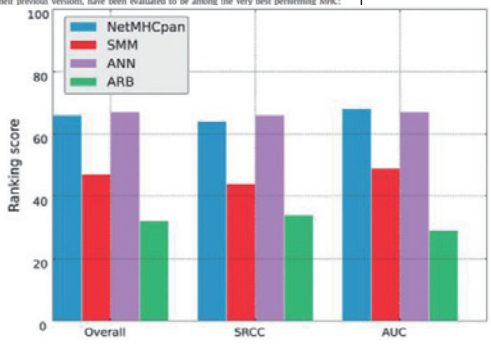
ARTICLE INFO

Article history:
Received 30 July 2010
Received in revised form 23 October 2010
Accepted 27 October 2010
Available online 31 October 2010

Keywords:
MHC
Binding
Prediction
Epitope
Discovery
T cell

ABSTRACT

In this paper, we describe the methodologies behind three different aspects of the NetMHC family for prediction of MHC class I binding, mainly to HLAs. We have updated the prediction servers, NetMHC-3.2, NetMHCpan-2.2, and a new consensus method, NetMHCcons, which, in their previous versions, have been evaluated to be among the very best performing MHC:



2nd Machine Learning Competition in Immunology 2012

Sponsors: InCoB 2012 and ICIW 2012

Prediction task:
Predict peptides naturally processed by MHC Class I pathway ("eluted peptides") for each target MHC molecule. For e
target molecule, the competitors are asked to submit a set of predicted eluted peptides from the test set.

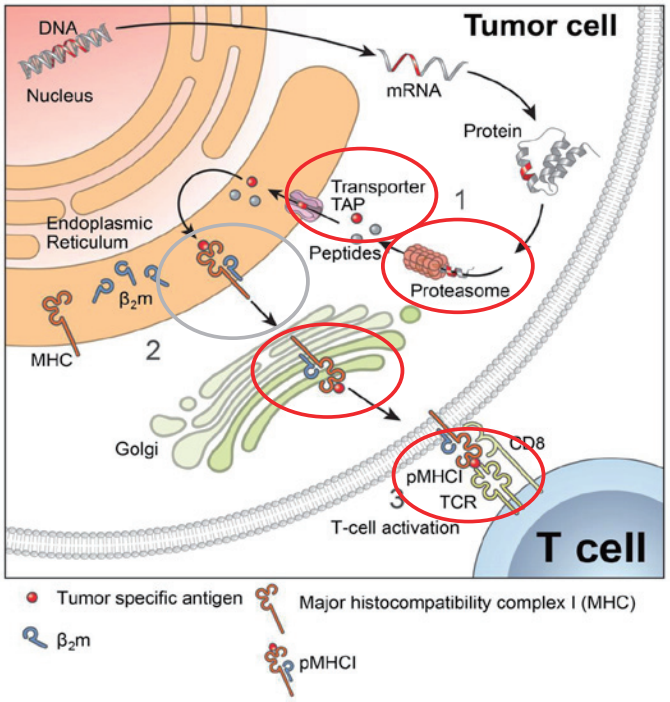
Winners MLG 2012

A total of 32 submissions were submitted for the competition. Of these, 24 submissions (Group 1) provided a set of thresholds (elution score based predictors) for each peptide and each MHC molecule. Another 8 submissions (Group 2) provided lists of peptides that were predicted as eluted from specific MHC molecules (eluted peptide list based predictors) for each of 8 studied MHC alleles. The NetMHC 3.2 server (1D-BENCH) results were used as a benchmark method.

Winning Team	Predictor No.	Prediction Method	Winning Category
Lundegaard C, Lamberth K, Harndahl M, Buus S, Lund O, Nielsen M, Technical University of Denmark	1D-BENCH	NetMHC 3.2 (Reference)	Group 1: A*0201
Giguere S, Drouin A, Lacoste A, Laval University, Canada	2F	A Bayesian model averaging method over several SVMs using the GS kernel.	Group 1: B*0702, H-2D^b, and H-2K^b
Nielsen M, et al., Technical University of Denmark	9D	A combination of NetMHC, NetMHCpan and MHCkernel predictions.	Group 1: B*3501 and B*4403
Giguere S, Drouin A, Lacoste A, Laval University, Canada	2D	A SVM classifier and a novel string kernel (GS kernel).	Group 1: B*5301
Xiang Z, He Y, University of Michigan Medical School, Ann Arbor, MI, USA.	20D	A position-specific scoring matrix (PSSM) with statistical P-value as the cutoff.	Group 1: B*5701
Yu Ting Wei, Department of Probability and Statistics, School of Mathematical Sciences, Peking University; Wen Jun Shen and Hau-San Wong, Department of Computer Science, City University of Hong Kong	14A	ConsMHC: a consensus program incorporating the results of kernelRLSpan-I, NetMHC, NetMHCpan and PickPocket by SVM	Group 2

ANTIGEN PROCESSING STEPS

Neoantigen processing revisited



somatic (passenger) mutations

↓

amino acid alternation

↓

protein degradation

↓

neopeptides

↓

binding to MHC molecule

↓

peptide-MHC presentation

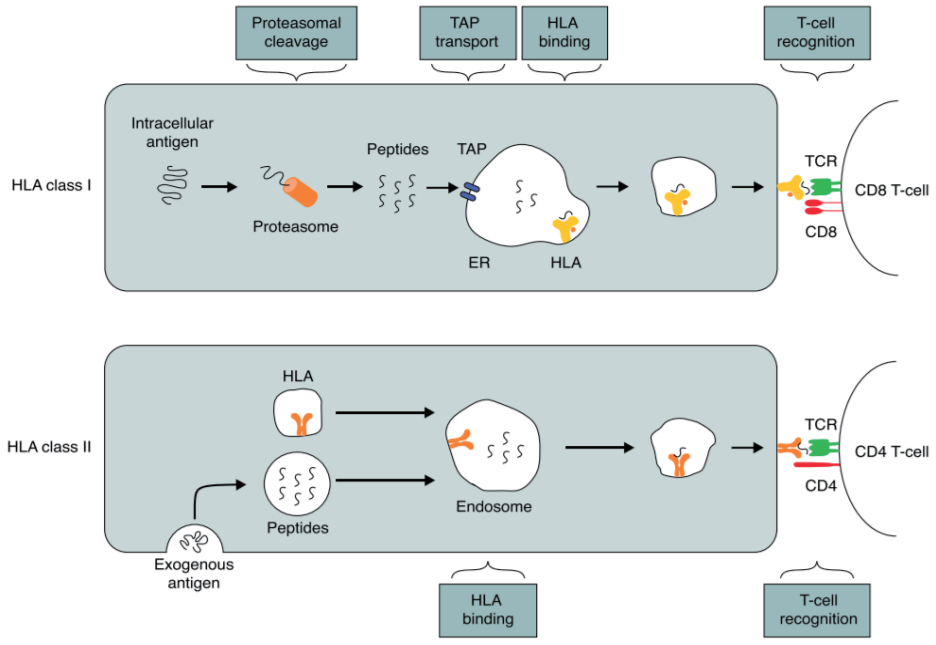
↓

recognition by T-cell

↓

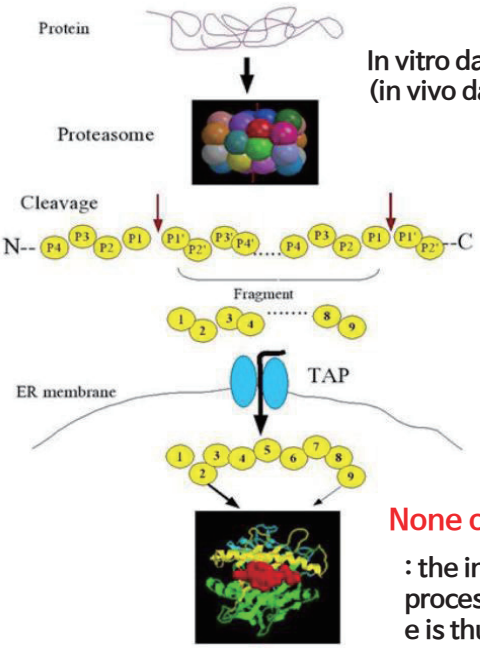
immune response

Antigen Processing Pathways for MHC class I/II



Backert and Kohlbacher, *Genome Medicine, 2015*

Proteasomal cleavage



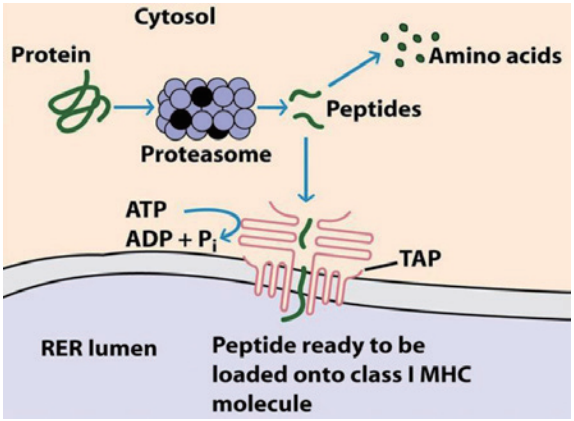
In vitro data created with purified proteasomes in the laboratory
(in vivo data are harder to collect)

C-terminus: commonly determined by proteasomal cleavage
N-terminus: can undergo further trimming by proteases located in the cytosol or ER

None of the predictors achieved an MCC above 0.3

: the in vitro data do not capture the full complexity of proteasomal processing in vivo. The value of predictions of proteasomal cleavage is thus rather limited

TAP transport prediction



- Primarily owing to the scarcity of data, there are few published methods on TAP transport prediction.
- No unbiased blind benchmarks for TAP transport methods have been published so far, and a comparative assessment of the various methods is thus currently difficult

Considering MHC-binding stability, not affinity

European Journal of Immunology

Peptide-MHC class I stability is a better predictor than peptide affinity of CTL immunogenicity

Mikkel Harndahl[1], Michael Rasmussen[1], Gustav Roder[1], Ida Dalgaard Pedersen[1], Mikael Sørensen[2], Morten Nielsen[2] and Søren Buus[1]

[1] Laboratory of Experimental Immunology, Faculty of Health Sciences, University of Copenhagen, Denmark
[2] Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, Denmark

Efficient presentation of peptide-MHC class I (pMHC-I) complexes to immune T cells should benefit from a stable peptide-MHC-I interaction. However, it has been difficult to distinguish stability from other requirements for MHC-I binding, for example, affinity. We have recently established a high-throughput assay for pMHC-I stability. Here, we have generated a large database containing stability measurements of pMHC-I complexes, and re-examined a previously reported unbiased analysis of the relative contributions of antigen processing and presentation in defining cytotoxic T lymphocyte (CTL) immunogenicity [Assarsson et al., J. Immunol. 2007. 178: 7890–7901]. Using an affinity-balanced approach, we demonstrated that immunogenic peptides tend to be more stably bound to MHC-I molecules compared with nonimmunogenic peptides. We also developed a bioinformatics method to predict pMHC-I stability, which suggested that 30% of the nonimmunogenic binders hitherto classified as "holes in the T-cell repertoire" can be explained as being unstably bound to MHC-I. Finally, we suggest that nonoptimal anchor

Binding (kinetic) stability

We also developed a bioinformatics method to predict pMHC-I stab ility, which suggested that 30% of the nonimmunogenic binders hith erto classified as "holes in the T-cell repertoire" can be explained as being unstably bound to MHC-I.

Prediction on the stability

NetMHCstab: predicting stability of pMHC-I complexes

Immunology
The Journal of cells, molecules, systems and technologies
British Society for immunology

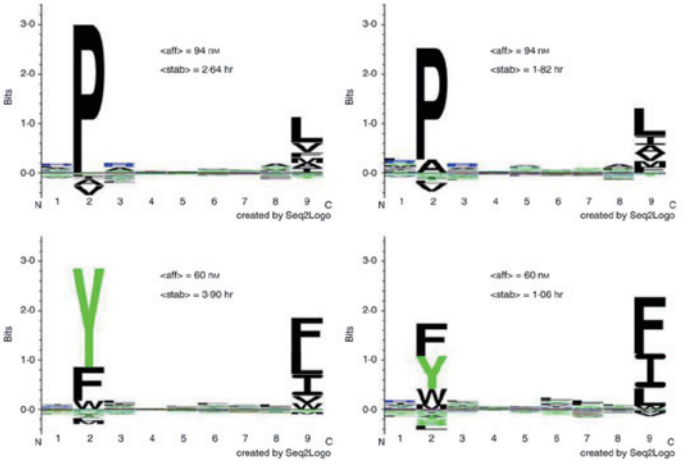
IMMUNOLOGY ORIGINAL ARTICLE

NetMHCstab – predicting stability of peptide–MHC-I complexes; impacts for cytotoxic T lymphocyte epitope discovery

Kasper W. Jørgensen,[1,*] Michael Rasmussen,[2,*] Søren Buus[2] and Morten Nielsen[1,3]
[1]*Department of Systems Biology, Centre for Biological Sequence Analysis, Technical University of Denmark, Lyngby,* [2]*Laboratory of Experimental Immunology, University of Copenhagen, Copenhagen N, Denmark, and* [3]*Instituto de Investigaciones Biotecnológicas, Universidad Nacional de San Martín, San Martín, Buenos Aires, Argentina*

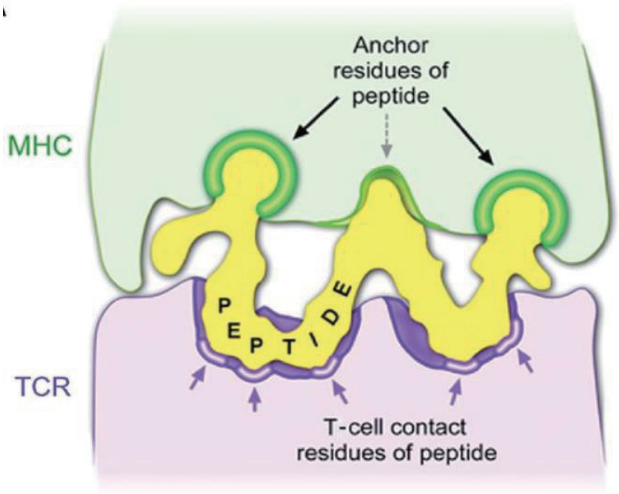
Summary

Major histocompatibility complex class I (MHC-I) molecules play an essential role in the cellular immune response, presenting peptides to cytotoxic T lymphocytes (CTLs) allowing the immune system to scrutinize ongoing intracellular production of proteins. In the early 1990s, immunogenicity and stability of the peptide–MHC-I (pMHC-I) complex were shown to be correlated. At that time, measuring stability was cumbersome and time consuming and only small data sets were analysed. Here, we investigate this fairly unexplored area on a large scale compared with earlier studies. A recent small-scale study demonstrated that pMHC-I complex stability was a better correlate of CTL immunogenicity than peptide–MHC-I affinity. We here extended this study and analysed a total of 5509 distinct peptide stability measurements covering 10 different HLA class I molecules. Artificial neural networks were used to construct stability predictors capable of predicting the half-life of the pMHC-I complex. These



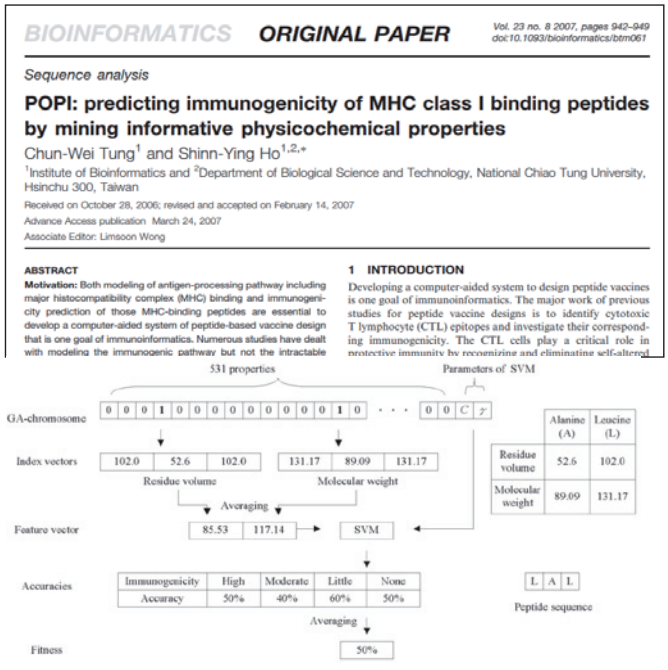
stable

Prediction on pMHC-TCR binding



Fritsch *et al, Cancer Immunology Research. 2014*

TCR immunogenicity prediction



BIOINFORMATICS ORIGINAL PAPER Vol. 23 no. 8 2007, pages 942–949 doi:10.1093/bioinformatics/btm061

Sequence analysis

POPI: predicting immunogenicity of MHC class I binding peptides by mining informative physicochemical properties

Chun-Wei Tung[1] and Shinn-Ying Ho[1,2,*]

[1]Institute of Bioinformatics and [2]Department of Biological Science and Technology, National Chiao Tung University, Hsinchu 300, Taiwan

Received on October 28, 2006; revised and accepted on February 14, 2007

Advance Access publication March 24, 2007

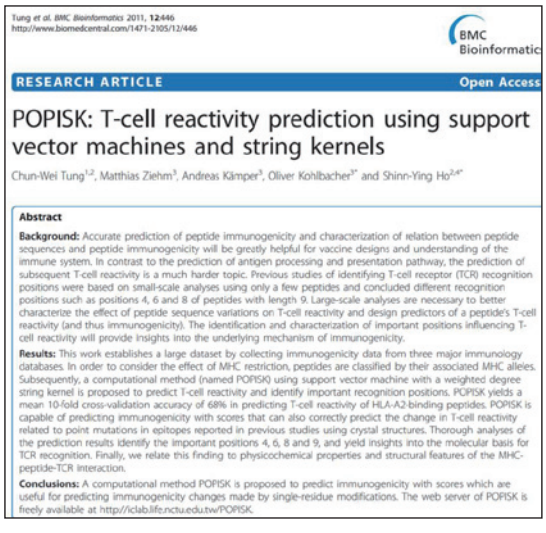
Associate Editor: Limsoon Wong

ABSTRACT

Motivation: Both modeling of antigen-processing pathway including major histocompatibility complex (MHC) binding and immunogenicity prediction of those MHC-binding peptides are essential to develop a computer-aided system of peptide-based vaccine design that is one goal of immunoinformatics. Numerous studies have dealt with modeling the immunogenic pathway but not the intractable

1 INTRODUCTION

Developing a computer-aided system to design peptide vaccines is one goal of immunoinformatics. The major work of previous studies for peptide vaccine designs is to identify cytotoxic T lymphocyte (CTL) epitopes and investigate their corresponding immunogenicity. The CTL cells play a critical role in protective immunity by recognizing and eliminating self-altered



Tung et al. BMC Bioinformatics 2011, 12:446
http://www.biomedcentral.com/1471-2105/12/446

BMC Bioinformatic

RESEARCH ARTICLE Open Access

POPISK: T-cell reactivity prediction using support vector machines and string kernels

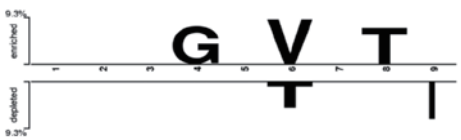
Chun-Wei Tung[1,2], Matthias Ziehm[3], Andreas Kämper[3], Oliver Kohlbacher[3*] and Shinn-Ying Ho[2,4*]

Abstract

Background: Accurate prediction of peptide immunogenicity and characterization of relation between peptide sequences and peptide immunogenicity will be greatly helpful for vaccine designs and understanding of the immune system. In contrast to the prediction of antigen processing and presentation pathway, the prediction of subsequent T-cell reactivity is a much harder topic. Previous studies of identifying T-cell receptor (TCR) recognition positions were based on small-scale analyses using only a few peptides and concluded different recognition positions such as positions 4, 6 and 8 of peptides with length 9. Large-scale analyses are necessary to better characterize the effect of peptide sequence variations on T-cell reactivity and design predictors of a peptide's T-cell reactivity (and thus immunogenicity). The identification and characterization of important positions influencing T-cell reactivity will provide insights into the underlying mechanism of immunogenicity.

Results: This work establishes a large dataset by collecting immunogenicity data from three major immunology databases. In order to consider the effect of MHC restriction, peptides are classified by their associated MHC alleles. Subsequently, a computational method (named POPISK) using support vector machine with a weighted degree string kernel is proposed to predict T-cell reactivity and identify important recognition positions. POPISK yields a mean 10-fold cross-validation accuracy of 68% in predicting T-cell reactivity of HLA-A2-binding peptides. POPISK is capable of predicting immunogenicity with scores that can also correctly predict the change in T-cell reactivity related to point mutations in epitopes reported in previous studies using crystal structures. Thorough analyses of the prediction results identify the important positions 4, 6, 8 and 9, and yield insights into the molecular basis for TCR recognition. Finally, we relate this finding to physicochemical properties and structural features of the MHC-peptide-TCR interaction.

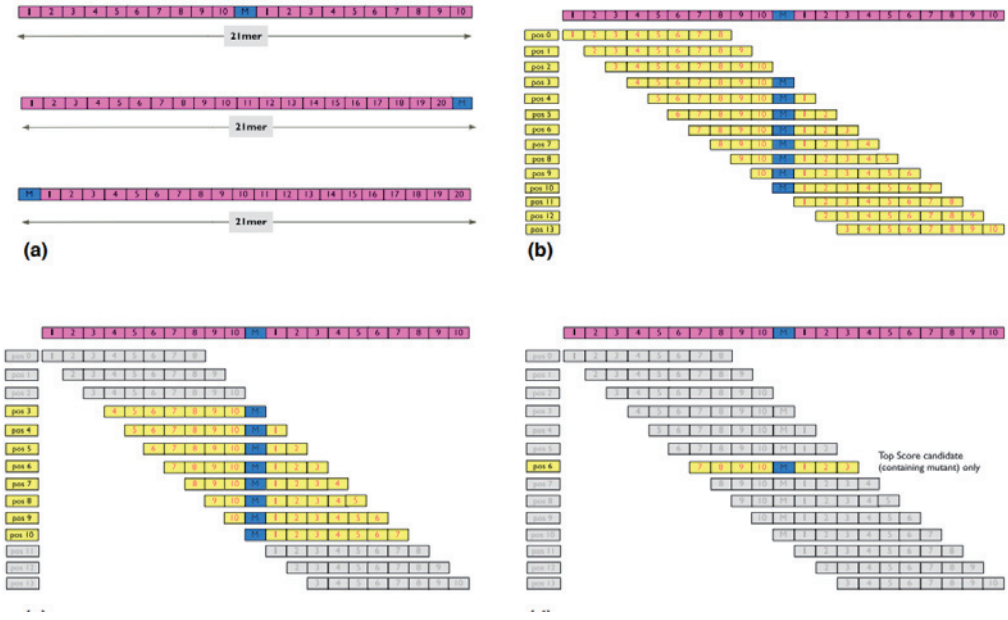
Conclusions: A computational method POPISK is proposed to predict immunogenicity with scores which are useful for predicting immunogenicity changes made by single-residue modifications. The web server of POPISK is freely available at http://iclab.life.nctu.edu.tw/POPISK.



The current performance of immunogenicity predictors is certainly not satisfying.
The amount and reliability of experimental data on T-cell reactivity is certainly one reason for this. But clearly our lack of understanding of the details of the processes leading to central and peripheral tolerance hamper the development of more predictive methods too (Toussant *et al*, BCB11, 2011)

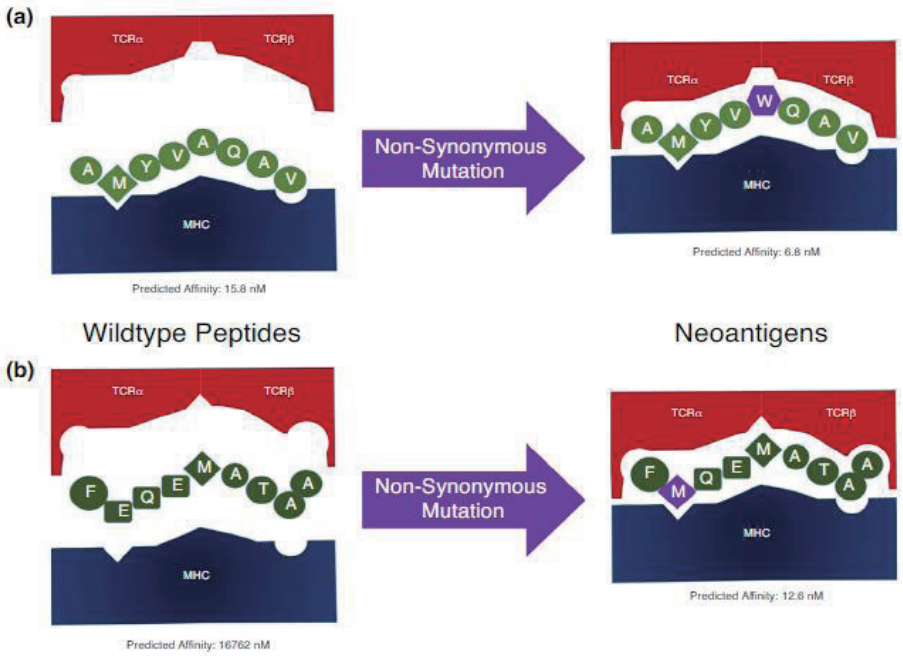
NEOANTIGEN ANALYSIS & INTEGRATED PIPELINES

Somatic mutation derived neopeptide



SBi 한국생명정보학회 Korean Society for Bioinformatics

And Neoantigens



Oiseth et al, *J Cancer Metastasis and Treatment, 2017*

SBi 한국생명정보학회 Korean Society for Bioinformatics

Overall Pipeline

Hundal et al. Genome Medicine (2016) 8:11
DOI 10.1186/s13073-016-0264-5

Genome Medi

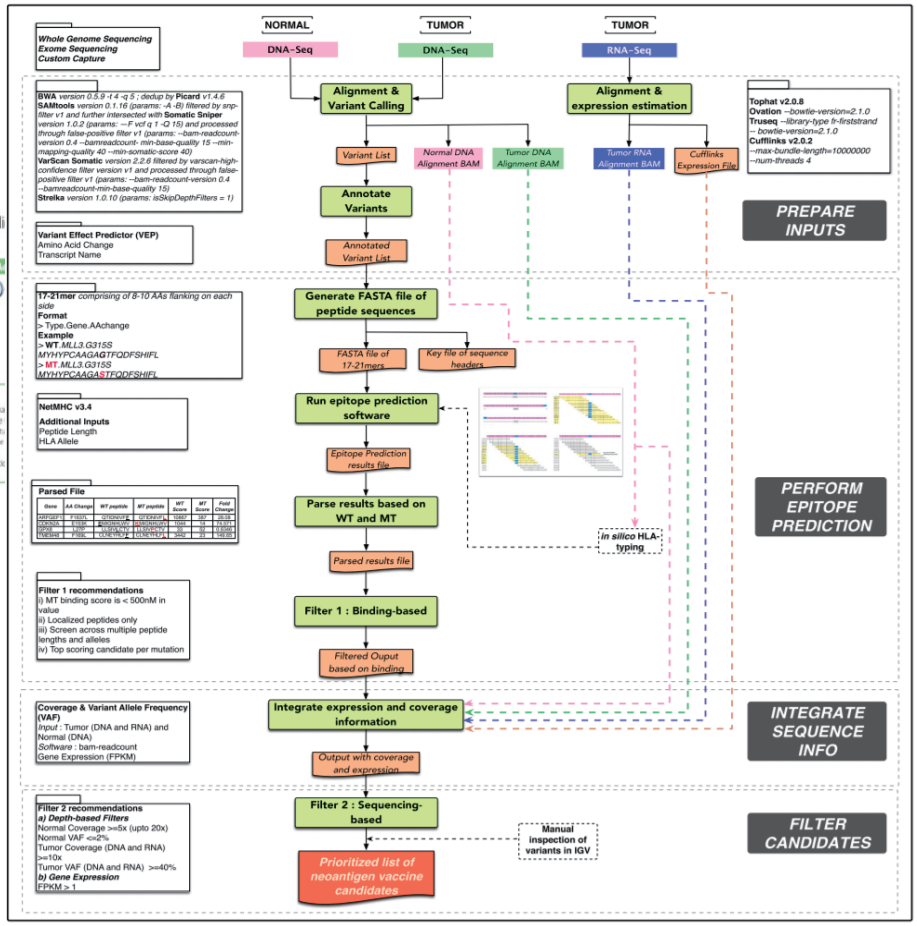
METHOD

pVAC-Seq: A genome-guided *in silico* approach to identifying tumor neoantigens

Jasreet Hundal[1], Beatriz M. Carreno[2], Allegra A. Petti[3], Gerald P. Linette[2], Obi L. Griffith[1,2,4,5], Elaine R. Mardis[1,3,4,5,6*] and Malachi Griffith[1,4,5]

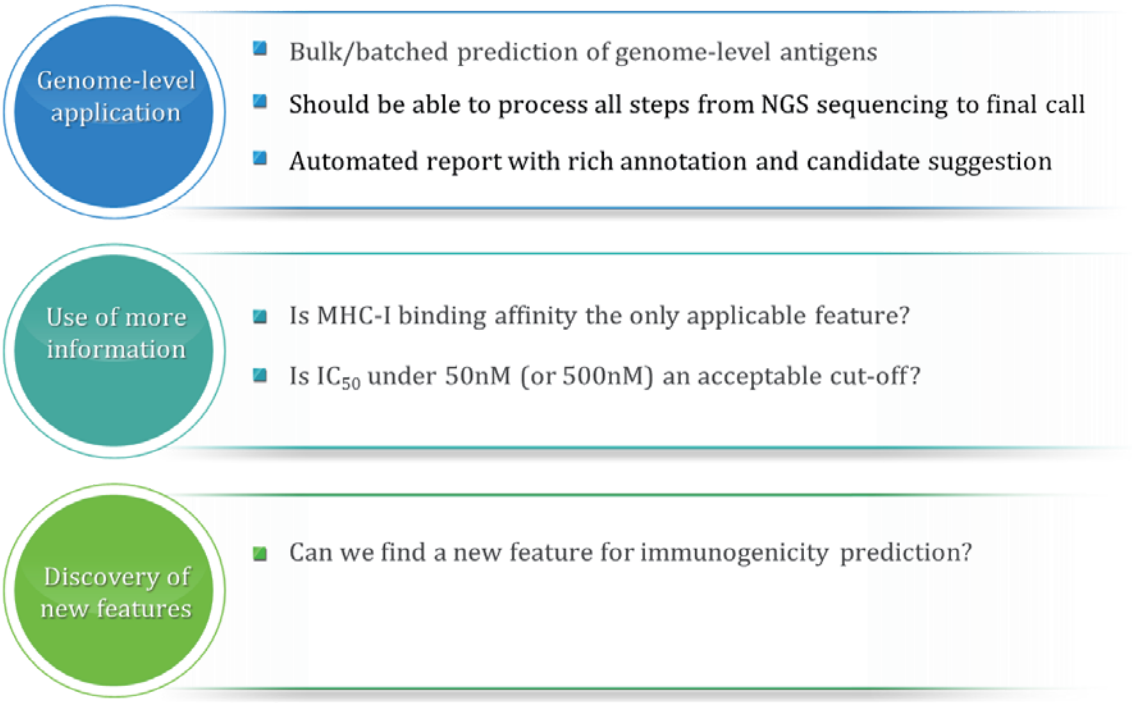
Abstract

Cancer immunotherapy has gained significant momentum from recent clinical successes of checkpoint blocka inhibition. Massively parallel sequence analysis suggests a connection between mutational load and response this class of therapy. Methods to identify which tumor-specific mutant peptides (neoantigens) can elicit anti-t T cell immunity are needed to improve predictions of checkpoint therapy response and to identify targets fo vaccines and adoptive T cell therapies. Here, we present a flexible, streamlined computational workflow for identification of personalized *Variant Antigens by Cancer Sequencing* (pVAC-Seq) that integrates tumor mutati and expression data (DNA- and RNA-Seq). pVAC-Seq is available at https://github.com/griffithlab/pVAC-Seq.



SBi 한국생명정보학회 Korean Society for Bioinformatics

Things need to be resolved for practical application



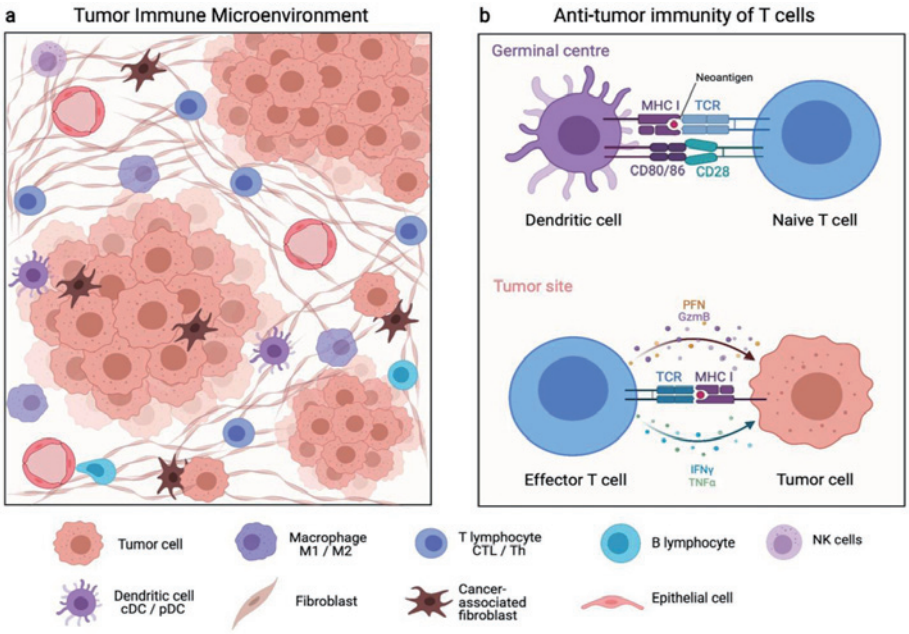
SBi 한국생명정보학회 Korean Society for Bioinformatics

IDENTIFYING TUMOR IMMUNE MICROENVIRONMENT

SBi 한국생명정보학회 Korean Society for Bioinformatics

62

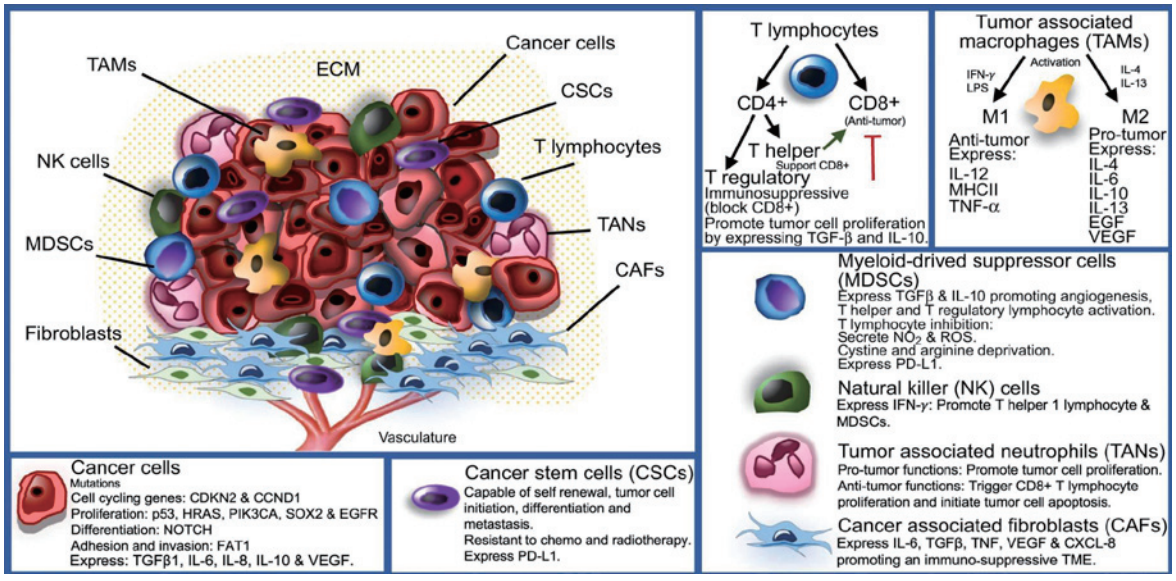
Tumor Immune Microenvironment



- A complex, organ-like structure (tumor cells, immune cells, fibroblasts, va scular endothelial cells, and other stromal cells)
- Immune cells + secreted factors (cytokines, chemokines, growth factors)

SBi 한국생명정보학회 Korean Society for Bioinformatics

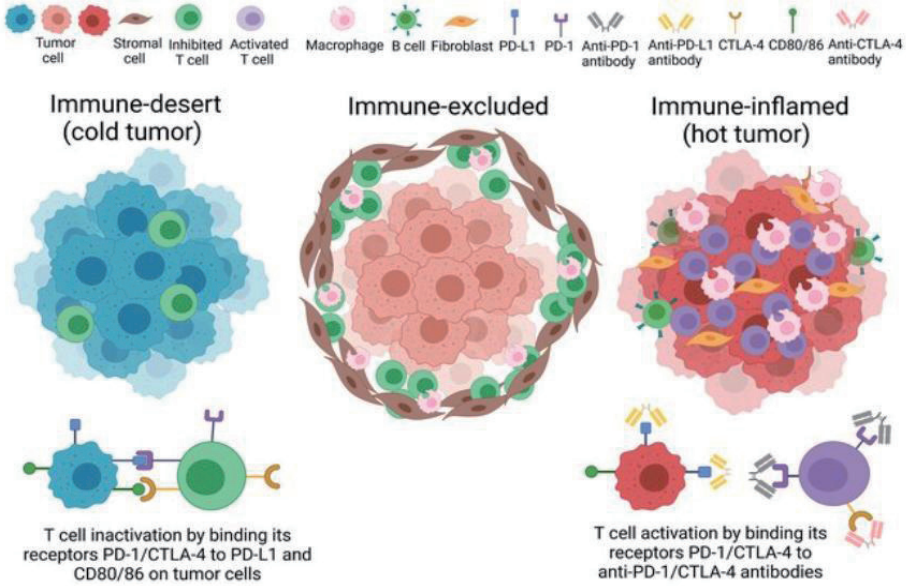
Tumor Immune Microenvironment



- TME components often inhibit or promote anti-tumor immunity
- But their roles are not definitive, and can be context-specific

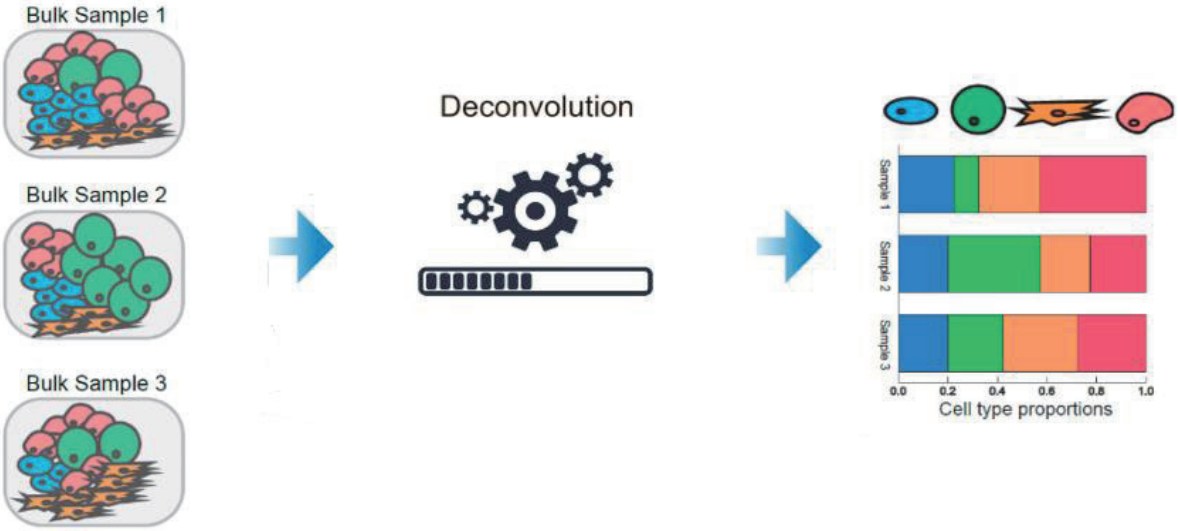
SBi 한국생명정보학회 Korean Society for Bioinformatics

Tumor Immune Phenotype

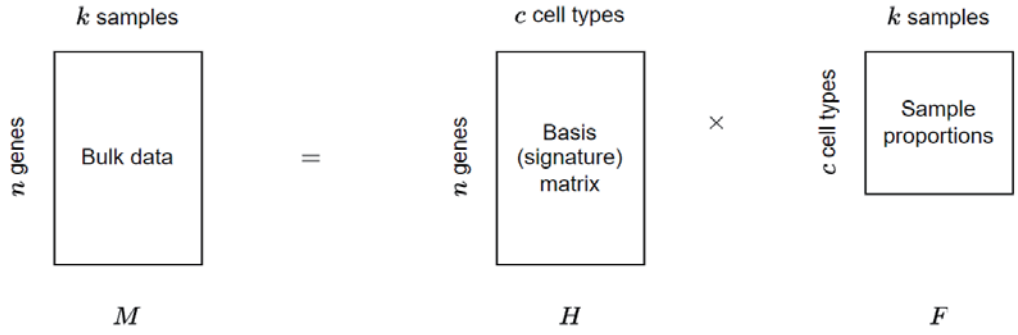


- Immune inflamed: immune cells infiltrated the tumor
- Immune excluded: immune cells are restricted to the stroma
- immune desert: T cells are not recruited

Identifying TME by Cell-type decomposition

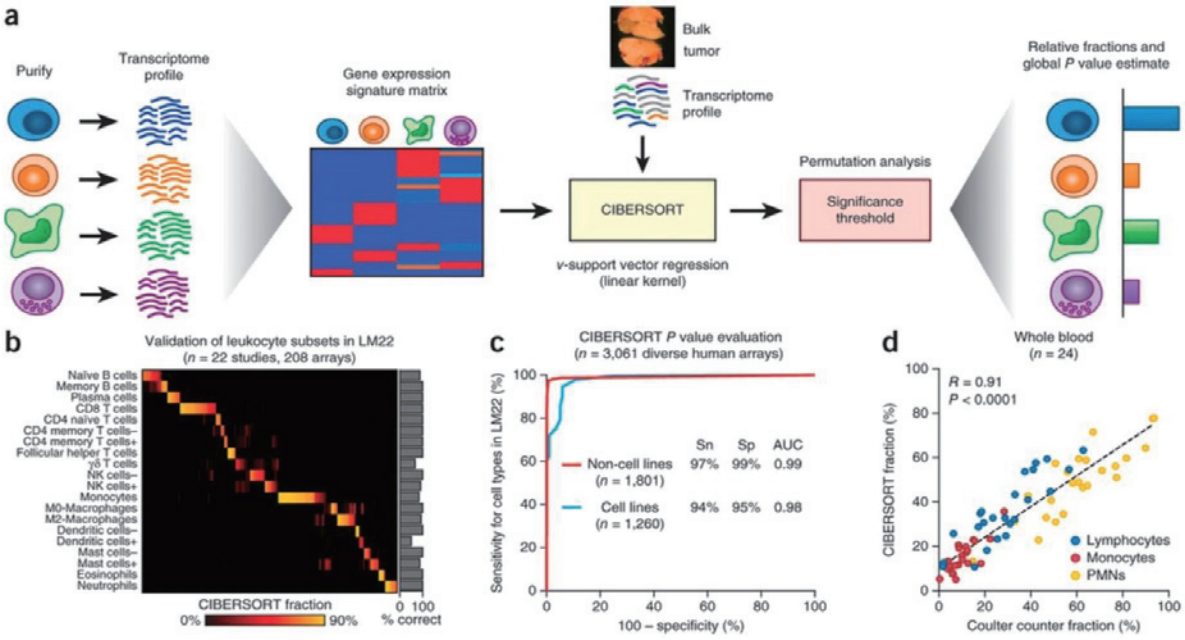


Identifying TME by Cell-type decomposition



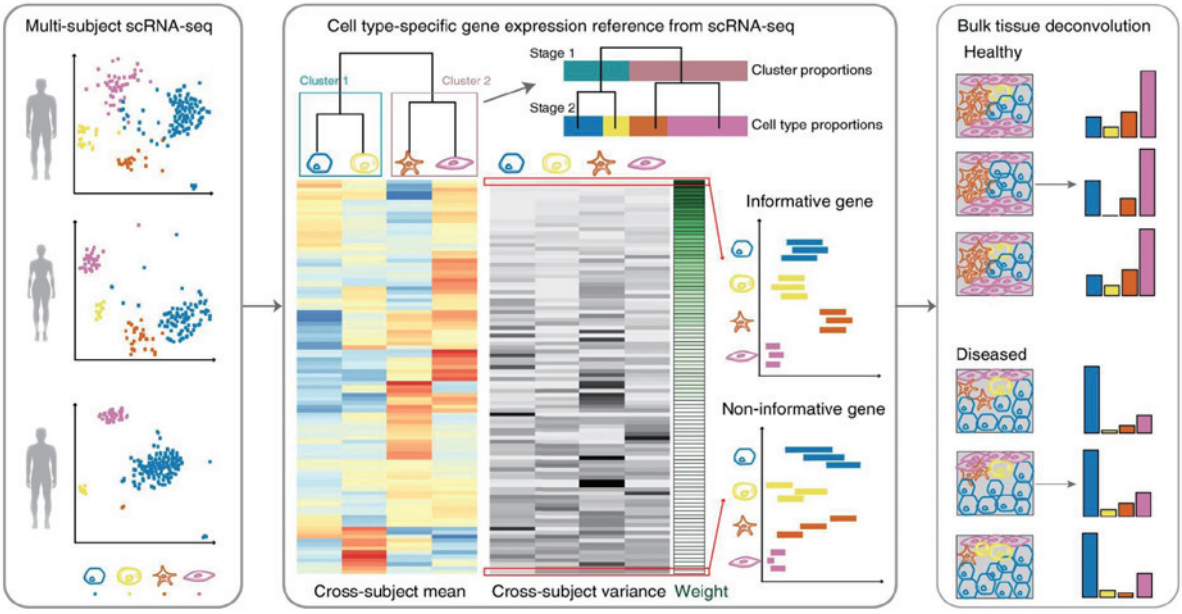
Problem	Given	Estimate	Requires
Estimate cell type proportions from bulk profile and signature matrix	M, H	F	$n > c$
Generate signature matrix from bulk profile and known cell type proportions	M, F	H	$k > c$
Estimate bulk profile from signature matrix and cell type proportions	H, F	M	none

CIBERSORT



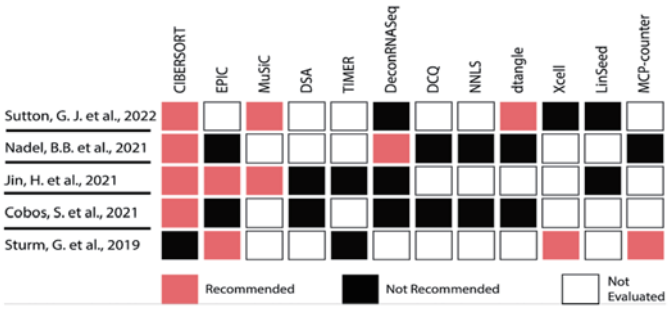
- Given a validated leukocyte gene signature matrix (LM22), deconvolute a n input bulk gene expression profile to generate cell-type fractions
- Support vector regression

MuSiC



- Utilize scRNA-seq from multiple subjects, identifying reference cell type-specific gene expression
- Extract genes that are informative (low cross-subject variance)

ImmuneDeconvR



Home > Bioinformatics for Cancer Immunotherapy > Protocol

Immunedeconv: An R Package for Unified Access to Computational Methods for Estimating Immune Cell Fractions from Bulk RNA-Sequencing Data

Gregor Sturm, Francesca Finotello & Markus List

Protocol | First Online: 03 March 2020

5035 Accesses | 88 Citations | 1 Altmetric

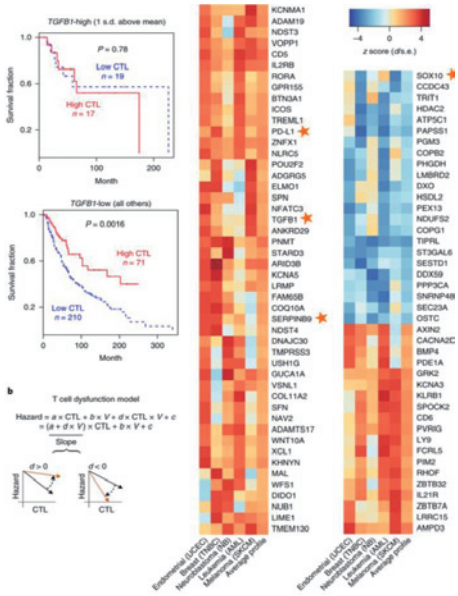
Part of the Methods in Molecular Biology book series (MIMB,volume 2120)

Abstract

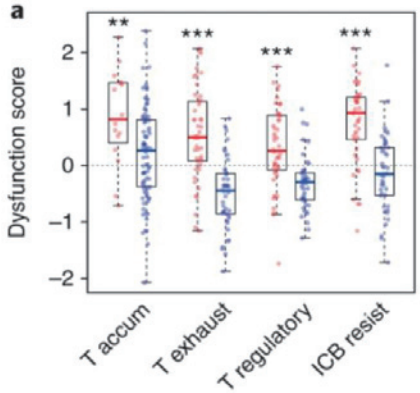
Since the performance of in silico approaches for estimating immune-cell fractions from bulk RNA-seq data can vary, it is often advisable to compare results of several methods. Given numerous dependencies and differences in input and output format of the various computational methods, comparative analyses can become quite complex. This motivated us to develop *immunedeconv*, an R package providing uniform and user-friendly access to seven state-of-the-art computational methods for deconvolution of cell-type fractions from bulk RNA-seq data. Here, we show how *immunedeconv* can be installed and applied to a typical dataset. First, we give an example for obtaining cell-type fractions using quanTIseq. Second, we show how dimensionless scores produced by MCP-counter can be used for cross-sample comparisons. For each of these examples, we provide R code illustrating how *immunedeconv* results can be summarized graphically.

- Each tool has its own pros and cons, and do not agree each other
- Immunedeconv provides a unified access to immune decomposition tools, so users can see different results and finally find a consensus

Predicting of T-cell evasion mechanisms (TIDE)



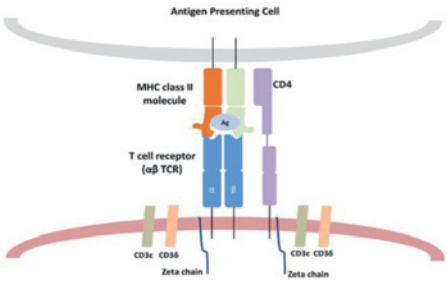
TGFB is interacting with CTL infiltration because:
- Survival of High vs. Low CTL-infiltrated patients are discriminated only when TGFB is highly expressed

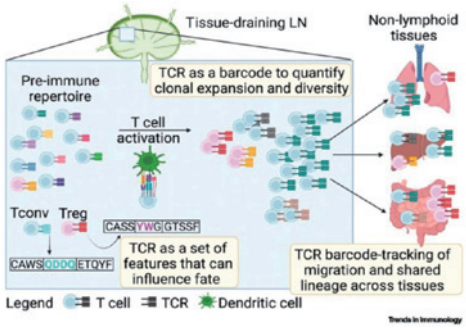


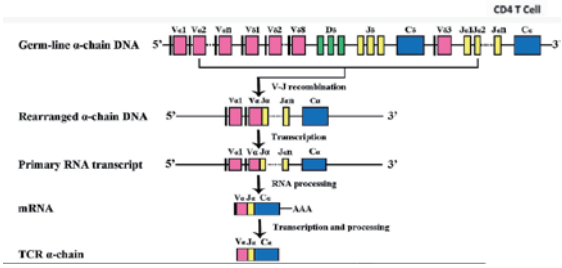
- Predicting T-cell dysfunction model: high infiltration but dysfunctional T-cells or excluded T-cells
- Extract T-cell dysfunction genes from interaction test in treatment naïve data
- Calculate T-cell dysfunction score

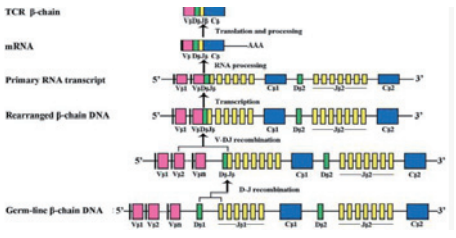
SBi 한국생명정보학회 Korean Society for Bioinformatics

TCR repertoire





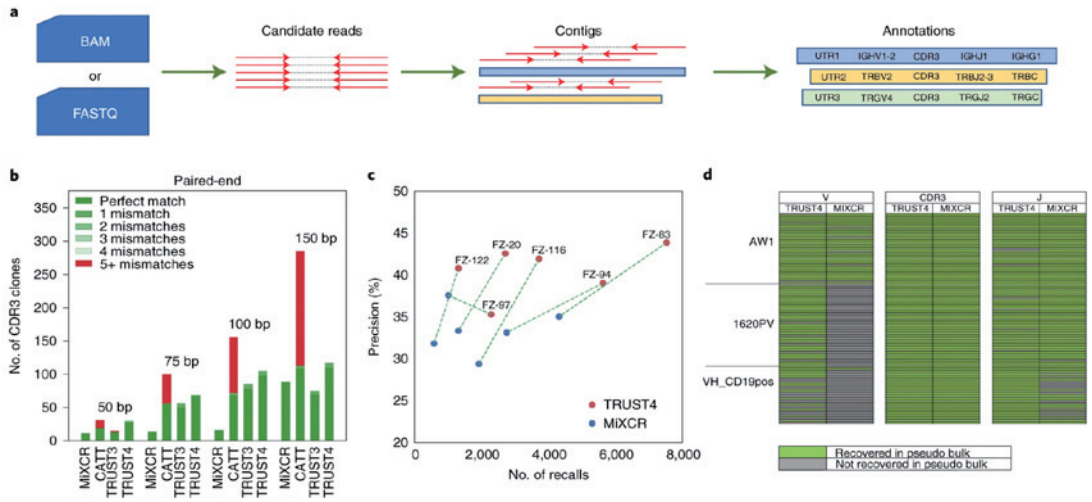




- T-cell diversity and clonality is the overall resultant response to the compl ex T-cell immune environment
- Diversity is inversely related to clonality
- High clonality is generally a marker for good response

SBi 한국생명정보학회 Korean Society for Bioinformatics

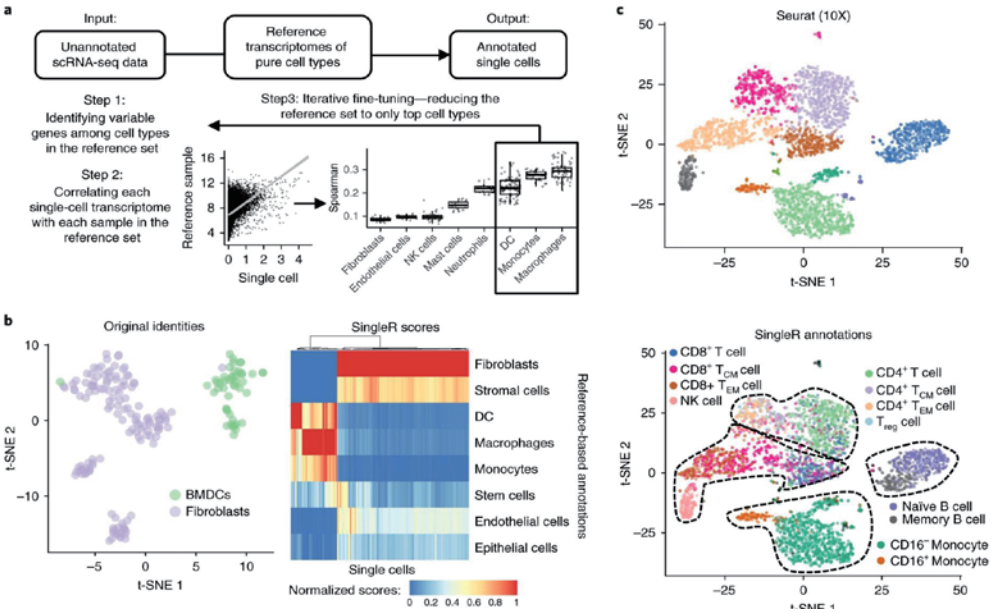
TCR repertoire reconstruction



- Generally, TCR or BCR sequencing is employed for repertoire reconstruction
- Conventional bulk RNA-seq can be also used using specialized tools, such as TRUST
- TRUST 1) extract TCR/BCR candidate reads, 2) assembles to form contigs, 3) identify somatic hypermutations, 4) reconstruct repertoire

SBi 한국생명정보학회 Korean Society for Bioinformatics

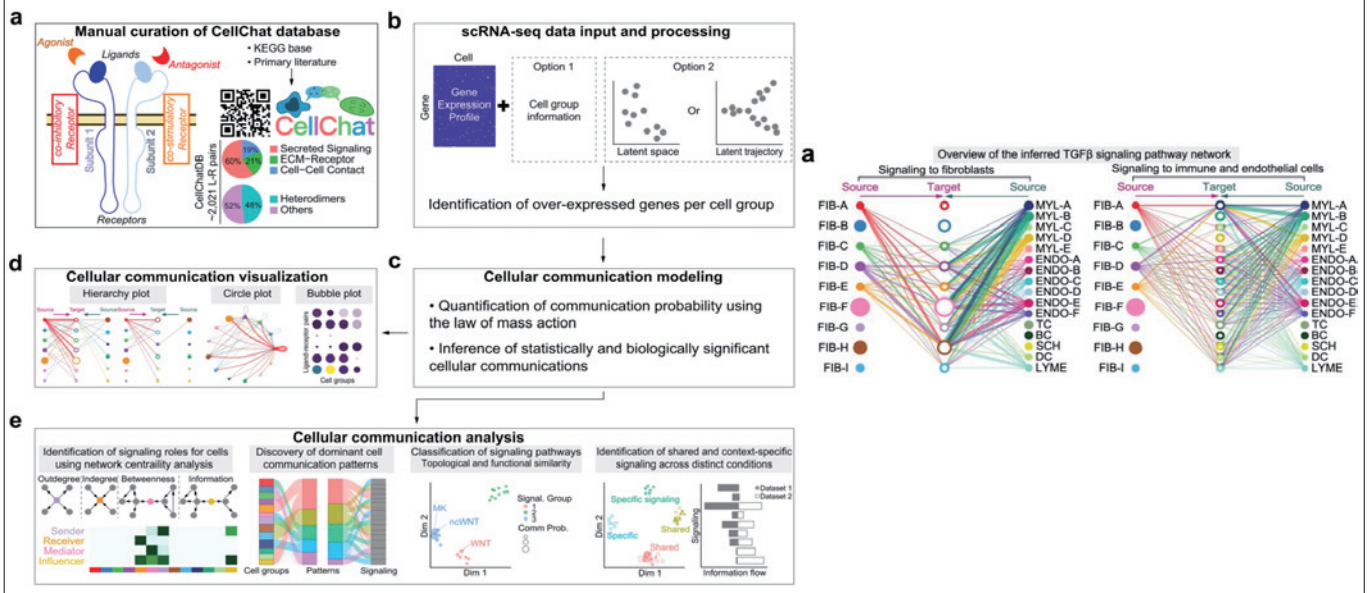
Use of single- and spatial transcriptomics



- In single cell sequencing, a complex decomposition is not necessary once t he single cells are well clustered.
- Clusters should be annotated using reference gene expression

SBi 한국생명정보학회 Korean Society for Bioinformatics

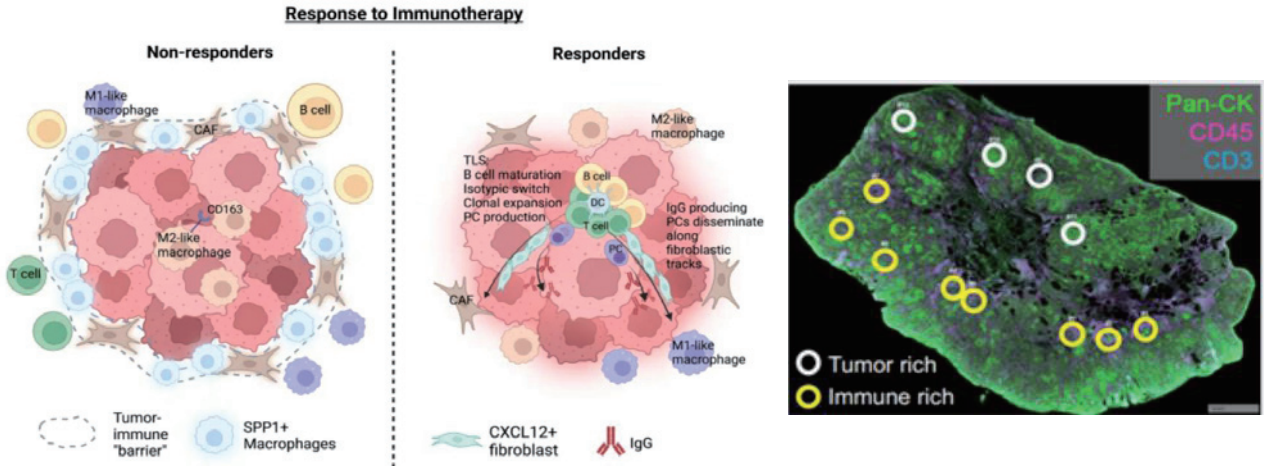
Use of single- and spatial transcriptomics



- Cell type-level gene expression with predefined ligand-receptor interactio ns, cellular communications can be inferred, wherein which cell type influe nced others through effector molecules

SBi 한국생명정보학회 Korean Society for Bioinformatics

Use of single- and spatial transcriptomics



- Using spatial transcriptomic, we can profile gene expression at the selected regio n of interest (ROI).
- Not only the abundance, but also the localization of immune cells direct the tumo r immune microenvironment
- Similar bulk cell sequencing analysis techniques can be also applied to the spatia l transcriptomics data

SBi 한국생명정보학회 Korean Society for Bioinformatics

Conclusion

- 다양한 cancer immunotherapy 의 발전으로 자신의 면역 시스템을 이용한 치료가 각광받고 있음
- 더 큰 효과와 적은 부작용을 위하여 환자, 종양 특이적 antigen 발굴이 필요함
- HLA type, MHC binding, Antigen processing 등 다양한 step 단계를 예측할 수 있는 computational algorithm 이 존재하며, 발전하고 있음
- Bulk, single, spatial transcriptomics 를 이용하여, 종양 주변의 면역환경인 Tumor immune microenvironment를 알아내고, 종양의 면역치료에 대한 환경에 따라 최적의 치료를 할 수 있음
- 결과적으로, NGS 에 기반하여 면역항암치료의 반응을 예측하고, 환자 특이적 치료를 할 수 있는 분석을 진행할 수 있음

Thank you

Your success is our success. We've prescription for your business.
We are professional communication group.

KSBI | 한국생명정보학회

KSBi-BIML 2024

Introduction to cancer-immune analysis

실습용 도구 및 환경 안내

Linux

특징

- 운영체제
- 무료 오픈소스
- 높은 통용성
- 높은 안정성
- 서버 환경으로 자주 사용됨

CLI (Command-Line Interface)

특징

- 명령어 입력 방식 (아이콘 사용 X)
- 보다 가벼움
- 보다 안정적
- 자동화 용이

실습용 도구 및 환경 안내

R

특징

- 생물정보학 분석 필수 프로그램
- 무료 오픈소스

R studio

특징

- R 사용 보조
- 변수 관리, 명령어 입력 및 기록, figure 생성 등을 위한 통합 환경 제공
- 무료 오픈소스로 사용 가능

Bioconductor

특징

- 생물정보학 분석용 패키지 모음
- 무료 오픈소스 프로그램 사용 보조

실습 진행 순서

1. **DNA-seq을 이용한 neoantigen prediction**
2. **Bulk RNA-seq을 이용한 tumor immune microenvironemnts 분석**
3. **Single cell RNA-seq을 이용한 cell-to-cell interaction prediction**
4. **Spatial RNA-seq을 이용한 TME 분석**

실습용 데이터 안내

DNA-seq을 이용한 neoantigen prediction

Prerequisites

Raw bam file (GRCh38)

- ACC_T_01.recaled.bam
- ACC_T_01.recaled.bai

Processed vcf file – Mutect2

- ACC_T_01.PASS.somatic.vcf

Processed data

Processed fastq

- ACC_T_01.chr6_1.fastq
- ACC_T_01.chr6_2.fastq

HLA typing (MHC class I) – OptiType

- ACC_T_01.MHC.I.processed.tsv
- ACC_T_01.MHC.I.list.txt

HLA typing (MHC class II) – HLA-HD

- ACC_T_01.MHC.II.processed.tsv
- ACC_T_01.MHC.II.list.txt

pVACseq (NetMHCpan, NetMHCIIpan)

- ACC_T_01.filtered.tsv (MHC Class I)
- ACC_T_01.filtered.tsv (MHC Class II)

실습 데이터: /home/jyhong906/BIML_2024/Bulk_WES/Data

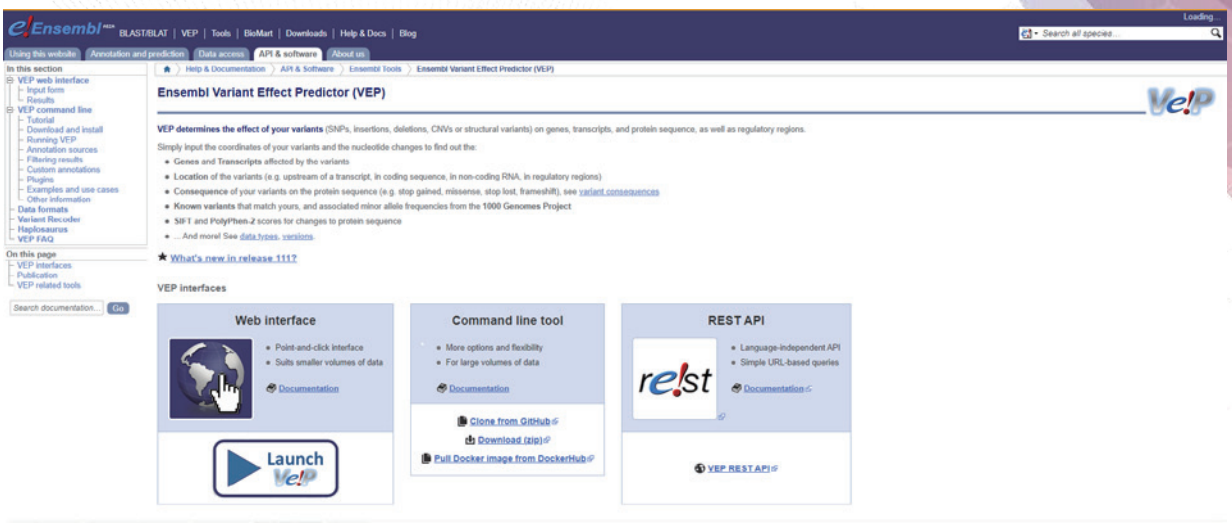
실습 스크립트: /home/jyhong906/BIML_2024/Bulk_WES/Script

환경 변수 설정

```
#!/usr/bin/env bash   # shebang
#$ -cwd    # 현재 디렉토리 내 실행

# PATH #
HLA_PATH=/home/jyhong906/BIML_2024/Bulk_WES/Data          # Input data, 결과 저장 디렉토리
optitype_PATH=${HLA_PATH}/OptiType      # MHC class I 관련 HLA typing 결과 저장 디렉토리
hlahd_PATH=${HLA_PATH}/HLA-HD           # MHC class II 관련 HLA typing 결과 저장 디렉토리

# MAKE FOLDER #
Path_list=(${HLA_PATH} ${optitype_PATH} ${hlahd_PATH})
for path in ${path_list[@]}; do
    mkdir -p ${path}  # 상위 디렉토리 모두 생성
Done

# FILE #
Ref=/home/jyhong906/Project/Reference/Ref/hg38/genome.fa    # Reference genome
IEDB_MHCI=/opt/Yonsei/IEDB-MHC_I        # 사전 설치 필요
IEDB_MHCII=/opt/Yonsei/IEDB-MHC_II      # 사전 설치 필요
hlahd_freq=/opt/Yonsei/HLA-HD/hlahd.1.7.0/freq_data
hlahd_split=/opt/Yonsei/HLA-HD/hlahd.1.7.0/HLA_gene.split.3.50.0.txt
hlahd_dict=/opt/Yonsei/HLA-HD/hlahd.1.7.0/dictionary

# EXECUTE #
vep_run=/opt/Yonsei/ensembl-vep/104.3/vep
optitype_run=/opt/Yonsei/OptiType/1.3.4/OptiTypePipeline.py
hlahd_run=hlahd.sh
pvacseq_run=/opt/Yonsei/python/3.8.1/bin/pvacseq

# SAMPLE #
patient_id=ACC_T_01

# FORMAT #
bam_format=.recaled.bam
chr6_bam_format=.sorted.chr6.bam
chr6_fastq1_format=.chr6_1.fastq
chr6_fastq2_format=.chr6_2.fastq
vcf_format=.PASS.somatic.vcf
ann_format=.vep.PASS.somatic.vcf
```

IEDB I, II installation
https://pvactools.readthedocs.io/en/latest/install.html#iedb-install

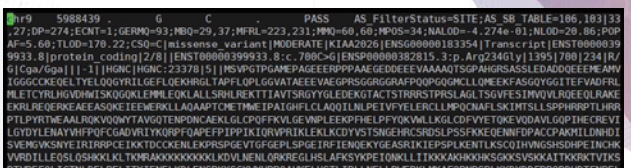
VEP (Variant Effect Predictor)



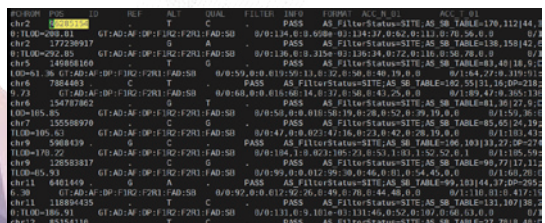
- 변이의 종류에 따른 영향(예: missense, nonsense, frameshift 등)을 포함한 상세한 annotation을 제공함.
- 유전체 데이터에서 발견된 변이의 기능적 영향을 분석하고, 이 정보를 바탕으로 생물학적 해석을 가능하게 함.
- VEP를 통한 변이 annotation은 변이가 단백질에 미치는 영향을 이해함으로써, potent neoantigen을 예측할 수 있음.

VEP annotation

```
${vep_run} \
-i ${HLA_PATH}/${patient_id}${vcf_format} \
-o ${HLA_PATH}/${patient_id}${ann_format} \
--vcf \
--symbol \
--terms SO \
--tsl \
--hgvs \
--fasta ${ref} \
--force_overwrite \
--assembly GRCh38 \
--plugin Wildtype \
--plugin Frameshift \
--offline \
--cache \
--dir_cache /data/public/VEP/104 \
--dir_plugins /data/public/VEP/104/Plugins \
--pick \
--transcript_version \
```



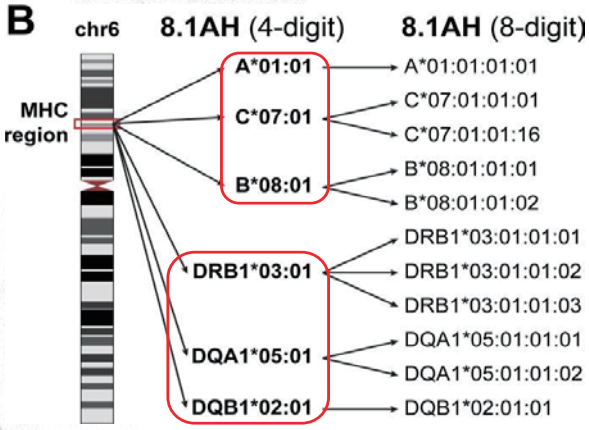
ACC_T_01.PASS.somatic.vcf



ACC_T_01.vep.PASS.somatic.vcf

BAM to chr6 fastq

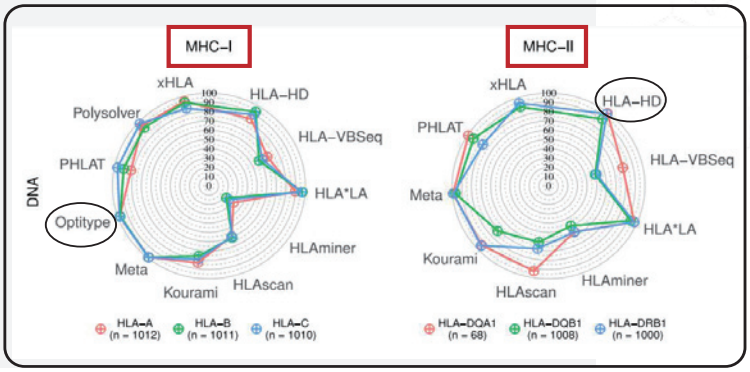
```
samtools view -h -b ${HLA_PATH}/${patient_id}${bam_format} chr6 > ${HLA_PATH}/${patient_id}${chr6_bam_format}
samtools fastq -1 ${HLA_PATH}/${patient_id}${chr6_fastq1_format} -2 ${HLA_PATH}/${patient_id}${chr6_fastq2_format} -F 4 ${HLA_PATH}/${patient_id}${chr6_bam_format}
```



Systematic genetic analysis of the MHC region reveals mechanistic underpinnings of HLA type associations with disease.

HLA typing (MHC class I, II)

```
# HLA typing - MHC class I (OptiType) #
python2 ${optitype_run} \
-i ${HLA_PATH}/${patient_id}${chr6_fastq1_format} ${HLA_PATH}/${patient_id}${chr6_fastq2_format} \
-e 4 \
--dna \
-v \
-c /opt/Yonsei/OptiType/1.3.4/config.ini \
-o ${optitype_PATH}/${patient_id} \
--prefix ${patient_id}

# convert format #
python3 source_make_MHC_list.py MHC_I ${optitype_PATH} ${patient_id}

# HLA typing - MHC class II (HLA-HD) #
${hlahd_run} \
-t 10 \
-m 50 \
-f ${hlahd_freq} \
${HLA_PATH}/${patient_id}${chr6_fastq1_format} \
${HLA_PATH}/${patient_id}${chr6_fastq2_format} \
${hlahd_split} \
${hlahd_dict} \
${patient_id} \
${hlahd_PATH}

# convert format #
python3 source_make_MHC_list.py MHC_II ${hlahd_PATH} ${patient_id}
```



Benchmark of tools for in silico prediction of MHC class I and class II genotypes from NGS data

HLA typing (MHC class I, II)

```
# HLA typing - MHC class I (OptiType) #
python2 ${optitype_run} \
-i ${HLA_PATH}/${patient_id}${chr6_fastq1_format} ${HLA_PATH}/${patient_id}${chr6_fastq2_format} \
-e 4 \
--dna \
-v \
-c /opt/Yonsei/OptiType/1.3.4/config.ini \
-o ${optitype_PATH}/${patient_id} \
--prefix ${patient_id}

# convert format #
python3 source_make_MHC_list.py MHC_I ${optitype_PATH} ${patient_id}
```

```
[jyhong906@master ACC_T_01]$ pwd
/home/jyhong906/BIML_2024/Bulk_WES/Data/OptiType/ACC_T_01
[jyhong906@master ACC_T_01]$ ll
total 1132
-rw-r--r-- 1 jyhong906 jyhong906 1154428 Feb  8 17:26 ACC_T_01_coverage_plot.pdf
-rw-r--r-- 1 jyhong906 jyhong906     337 Feb  8 17:26 ACC_T_01_result.tsv
```

HLA_PATH=/home/jyhong906/BIML_2024/Bulk_WES/Data

MHC class I – ACC_T_01_result.tsv

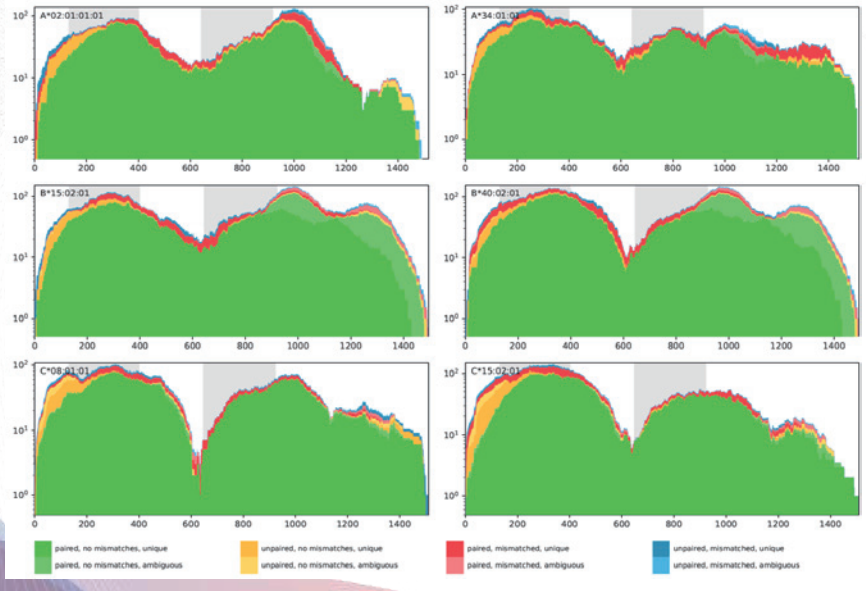
	A1	A2	B1	B2	C1	C2	Reads	Objective
0	A*02:01	A*34:01	B*40:02	B*15:02	C*15:02	C*08:01	2893.0	2762.8149999999955
1	A*02:01	A*34:01	B*15:02	B*40:06	C*15:02	C*08:01	2871.0	2741.8049999999953
2	A*02:01	A*34:05	B*40:02	B*15:02	C*15:02	C*08:01	2863.0	2734.154999999995
3	A*34:01	A*02:16	B*40:02	B*15:02	C*15:02	C*08:01	2861.0	2732.244999999996

MHC class I – ACC_T_01.MHC.I.list.txt

```
HLA-A*02:01,HLA-A*34:01,HLA-B*40:02,HLA-B*15:02,HLA-C*15:02,HLA-C*08:01
```


HLA typing (MHC class I, II)

```
# HLA typing - MHC class I (OptiType) #
python2 ${optitype_run} \
-i ${HLA_PATH}/${patient_id}${chr6_fastq1_format} ${HLA_PATH}/${patient_id}${chr6_fastq2_format} \
-e 4 \
--dna \
-v \
-c /opt/Yonsei/OptiType/1.3.4/config.ini \
-o ${optitype_PATH}/${patient_id} \
--prefix ${patient_id}

# convert format #
python3 source_make_MHC_list.py MHC_I ${optitype_PATH} ${patient_id}
```

HLA typing (MHC class I, II)

```
[jyhong906@master result]$ pwd
/home/jyhong906/DLML_2024/Bulk_WES/Data/HLA-HD/ACC_T_01/result
[jyhong906@master result]$ ll
total 1676
-rw-rw-r-- 1 jyhong906 jyhong906   4711 Feb  9 17:42 ACC_T_01_A.est.txt
-rw-rw-r-- 1 jyhong906 jyhong906 117230 Feb  9 17:42 ACC_T_01_A.read.txt
-rw-rw-r-- 1 jyhong906 jyhong906    965 Feb  9 17:42 ACC_T_01_B.est.txt
-rw-rw-r-- 1 jyhong906 jyhong906  85393 Feb  9 17:42 ACC_T_01_B.read.txt
-rw-rw-r-- 1 jyhong906 jyhong906   1506 Feb  9 17:42 ACC_T_01_C.est.txt
-rw-rw-r-- 1 jyhong906 jyhong906  85495 Feb  9 17:42 ACC_T_01_C.read.txt
-rw-rw-r-- 1 jyhong906 jyhong906    260 Feb  9 17:42 ACC_T_01_DMA.est.txt
-rw-rw-r-- 1 jyhong906 jyhong906  59235 Feb  9 17:42 ACC_T_01_DMA.read.txt
-rw-rw-r-- 1 jyhong906 jyhong906    899 Feb  9 17:42 ACC_T_01_DMB.est.txt
-rw-rw-r-- 1 jyhong906 jyhong906  54103 Feb  9 17:42 ACC_T_01_DMB.read.txt
-rw-rw-r-- 1 jyhong906 jyhong906    376 Feb  9 17:43 ACC_T_01_DOA.est.txt
-rw-rw-r-- 1 jyhong906 jyhong906  64703 Feb  9 17:43 ACC_T_01_DOA.read.txt
-rw-rw-r-- 1 jyhong906 jyhong906    619 Feb  9 17:41 ACC_T_01_DOB.est.txt
-rw-rw-r-- 1 jyhong906 jyhong906  49139 Feb  9 17:41 ACC_T_01_DOB.read.txt
-rw-rw-r-- 1 jyhong906 jyhong906   1949 Feb  9 17:42 ACC_T_01_DPA1.est.txt
-rw-rw-r-- 1 jyhong906 jyhong906  42928 Feb  9 17:42 ACC_T_01_DPA1.read.txt
-rw-rw-r-- 1 jyhong906 jyhong906    249 Feb  9 17:43 ACC_T_01_DPA2.est.txt
-rw-rw-r-- 1 jyhong906 jyhong906   3229 Feb  9 17:43 ACC_T_01_DPA2.read.txt
-rw-rw-r-- 1 jyhong906 jyhong906   2524 Feb  9 17:41 ACC_T_01_DPB1.est.txt
-rw-rw-r-- 1 jyhong906 jyhong906  58668 Feb  9 17:41 ACC_T_01_DPB1.read.txt
-rw-rw-r-- 1 jyhong906 jyhong906    748 Feb  9 17:42 ACC_T_01_DQA1.est.txt
-rw-rw-r-- 1 jyhong906 jyhong906  75414 Feb  9 17:42 ACC_T_01_DQA1.read.txt
-rw-rw-r-- 1 jyhong906 jyhong906   1568 Feb  9 17:42 ACC_T_01_DQB1.est.txt
-rw-rw-r-- 1 jyhong906 jyhong906  47342 Feb  9 17:42 ACC_T_01_DQB1.read.txt
-rw-rw-r-- 1 jyhong906 jyhong906    097 Feb  9 17:42 ACC_T_01_DRA.est.txt
-rw-rw-r-- 1 jyhong906 jyhong906 105944 Feb  9 17:42 ACC_T_01_DRA.read.txt
-rw-rw-r-- 1 jyhong906 jyhong906    433 Feb  9 17:40 ACC_T_01_DRB1.est.txt
-rw-rw-r-- 1 jyhong906 jyhong906  33356 Feb  9 17:40 ACC_T_01_DRB1.read.txt
-rw-rw-r-- 1 jyhong906 jyhong906    162 Feb  9 17:43 ACC_T_01_DRB2.est.txt
-rw-rw-r-- 1 jyhong906 jyhong906   1159 Feb  9 17:43 ACC_T_01_DRB2.read.txt
-rw-rw-r-- 1 jyhong906 jyhong906    181 Feb  9 17:43 ACC_T_01_DRB3.est.txt
-rw-rw-r-- 1 jyhong906 jyhong906  12652 Feb  9 17:43 ACC_T_01_DRB3.read.txt
-rw-rw-r-- 1 jyhong906 jyhong906     14 Feb  9 17:41 ACC_T_01_DRB4.est.txt
-rw-rw-r-- 1 jyhong906 jyhong906      0 Feb  9 17:41 ACC_T_01_DRB4.read.txt
-rw-rw-r-- 1 jyhong906 jyhong906    185 Feb  9 17:43 ACC_T_01_DRB5.est.txt
-rw-rw-r-- 1 jyhong906 jyhong906  17365 Feb  9 17:43 ACC_T_01_DRB5.read.txt
-rw-rw-r-- 1 jyhong906 jyhong906     73 Feb  9 17:43 ACC_T_01_DRB6.est.txt
-rw-rw-r-- 1 jyhong906 jyhong906   2616 Feb  9 17:43 ACC_T_01_DRB6.read.txt
-rw-rw-r-- 1 jyhong906 jyhong906     14 Feb  9 17:43 ACC_T_01_DRB7.est.txt
-rw-rw-r-- 1 jyhong906 jyhong906      0 Feb  9 17:43 ACC_T_01_DRB7.read.txt
-rw-rw-r-- 1 jyhong906 jyhong906     14 Feb  9 17:41 ACC_T_01_DRB8.est.txt
-rw-rw-r-- 1 jyhong906 jyhong906      0 Feb  9 17:41 ACC_T_01_DRB8.read.txt
-rw-rw-r-- 1 jyhong906 jyhong906     80 Feb  9 17:43 ACC_T_01_DRB9.est.txt
-rw-rw-r-- 1 jyhong906 jyhong906   7838 Feb  9 17:43 ACC_T_01_DRB9.read.txt
-rw-rw-r-- 1 jyhong906 jyhong906   1080 Feb  9 17:43 ACC_T_01_E.est.txt
-rw-rw-r-- 1 jyhong906 jyhong906  89461 Feb  9 17:43 ACC_T_01_E.read.txt
-rw-rw-r-- 1 jyhong906 jyhong906    494 Feb  9 17:42 ACC_T_01_F.est.txt
-rw-rw-r-- 1 jyhong906 jyhong906    958 Feb  9 17:44 ACC_T_01_final.result.txt
-rw-rw-r-- 1 jyhong906 jyhong906  56222 Feb  9 17:42 ACC_T_01_F.read.txt
-rw-rw-r-- 1 jyhong906 jyhong906    689 Feb  9 17:43 ACC_T_01_G.est.txt
-rw-rw-r-- 1 jyhong906 jyhong906  75325 Feb  9 17:43 ACC_T_01_G.read.txt
-rw-rw-r-- 1 jyhong906 jyhong906    219 Feb  9 17:43 ACC_T_01_H.est.txt
-rw-rw-r-- 1 jyhong906 jyhong906 100146 Feb  9 17:43 ACC_T_01_H.read.txt
-rw-rw-r-- 1 jyhong906 jyhong906    352 Feb  9 17:44 ACC_T_01_J.est.txt
-rw-rw-r-- 1 jyhong906 jyhong906  87700 Feb  9 17:44 ACC_T_01_J.read.txt
-rw-rw-r-- 1 jyhong906 jyhong906     92 Feb  9 17:42 ACC_T_01_K.est.txt
-rw-rw-r-- 1 jyhong906 jyhong906  31139 Feb  9 17:42 ACC_T_01_K.read.txt
-rw-rw-r-- 1 jyhong906 jyhong906     95 Feb  9 17:43 ACC_T_01_L.est.txt
-rw-rw-r-- 1 jyhong906 jyhong906  48231 Feb  9 17:43 ACC_T_01_L.read.txt
-rw-rw-r-- 1 jyhong906 jyhong906    232 Feb  9 17:44 ACC_T_01_T.est.txt
-rw-rw-r-- 1 jyhong906 jyhong906  11007 Feb  9 17:44 ACC_T_01_T.read.txt
-rw-rw-r-- 1 jyhong906 jyhong906    118 Feb  9 17:44 ACC_T_01_V.est.txt
-rw-rw-r-- 1 jyhong906 jyhong906  27658 Feb  9 17:44 ACC_T_01_V.read.txt
-rw-rw-r-- 1 jyhong906 jyhong906    193 Feb  9 17:42 ACC_T_01_W.est.txt
-rw-rw-r-- 1 jyhong906 jyhong906   5123 Feb  9 17:42 ACC_T_01_W.read.txt
-rw-rw-r-- 1 jyhong906 jyhong906    343 Feb  9 17:44 ACC_T_01_Y.est.txt
-rw-rw-r-- 1 jyhong906 jyhong906  71265 Feb  9 17:44 ACC_T_01_Y.read.txt
```

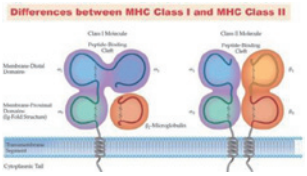
```
# HLA typing - MHC class II (HLA-HD) #
${hlahd_run} \
-t 10 \
-m 50 \
-f ${hlahd_freq} \
${HLA_PATH}/${patient_id}${chr6_fastq1_format} \
${HLA_PATH}/${patient_id}${chr6_fastq2_format} \
${hlahd_split} \
${hlahd_dict} \
${patient_id} \
${hlahd_PATH}

# convert format #
python3 source_make_MHC_list.py MHC_II ${hlahd_PATH} ${patient_id}
```

MHC class II - ACC_T_01_final.result.txt

A	HLA-A*02:01:01	HLA-A*34:01:01
B	HLA-B*40:02:01	HLA-B*15:02:01
C	HLA-C*15:02:01	HLA-C*08:01:01
DRB1	HLA-DRB1*15:02:01	HLA-DRB1*12:02:01
DQA1	HLA-DQA1*01:02:01	HLA-DQA1*06:01:01
DQB1	HLA-DQB1*03:01:01	HLA-DQB1*05:02:01
DPA1	HLA-DPA1*01:03:01	HLA-DPA1*02:02:02
DPB1	HLA-DPB1*02:01:02	HLA-DPB1*01:01:01
DMA	HLA-DMA*01:02:01	-
DMB	HLA-DMB*01:01:01	-
DOA	HLA-DOA*01:01:04	HLA-DOA*01:01:01
DOB	HLA-DOB*01:01:01	-
DRA	HLA-DRA*01:02:02	HLA-DRA*01:01:01
DRB2	HLA-DRB2*01:01	-
DRB3	HLA-DRB3*03:01:03	-
DRB4	Not typed	Not typed
DRB5	HLA-DRB5*01:01:01	-
DRB6	HLA-DRB6*02:01	-
DRB7	Not typed	Not typed
DRB8	Not typed	Not typed
DRB9	HLA-DRB9*01:02:01	-
DPA2	HLA-DPA2*01:01:02	HLA-DPA2*02:01
E	HLA-E*01:03:01	HLA-E*01:03:02
F	HLA-F*01:01:01	-
G	HLA-G*01:01:01	HLA-G*01:01:03
H	HLA-H*01:01:01	HLA-H*02:27
J	HLA-J*01:01:01	HLA-J*01:01:01
K	HLA-K*01:02	-
L	HLA-L*01:02	-
T	HLA-T*01:01:01	HLA-T*02:01:01
V	HLA-V*01:01:01	-
W	HLA-W*03:01:01	-
Y	HLA-Y*01:01	HLA-Y*03:01

MHC class II – ACC_T_01.MHC.II.list.txt

```
DRB1*15:02,DRB1*12:02,DQA1*01:02-DQB1*03:01,DQA1*06:01-DQB1*05:02
```



pVACseq (neoantigen prediction)

```
# pVACseq - MHC class I (NetMHCpan) #
allele=`cat ${optitype_PATH}/${patient_id}.MHC.I.list.txt`
${pvacseq_run} run ${HLA_PATH}/${patient_id}${ann_format} \
${patient_id} \
${allele} \
NetMHCpan \
-e1 8,9,10,11 \
--pass-only \
${HLA_PATH}/${patient_id} \
--iedb-install-directory ${IEDB_MHCI}

# pVACseq - MHC class II (NetMHCIIpan) #
allele=`cat ${hlahd_PATH}/${patient_id}.MHC.II.list.txt`
${pvacseq_run} run ${HLA_PATH}/${patient_id}${ann_format} \
${patient_id} \
${allele} \
NetMHCIIpan \
-e2 12,13,14,15,16,17,18 \
--pass-only \
${HLA_PATH}/${patient_id}
--iedb-install-directory ${IEDB_MHCII}
```

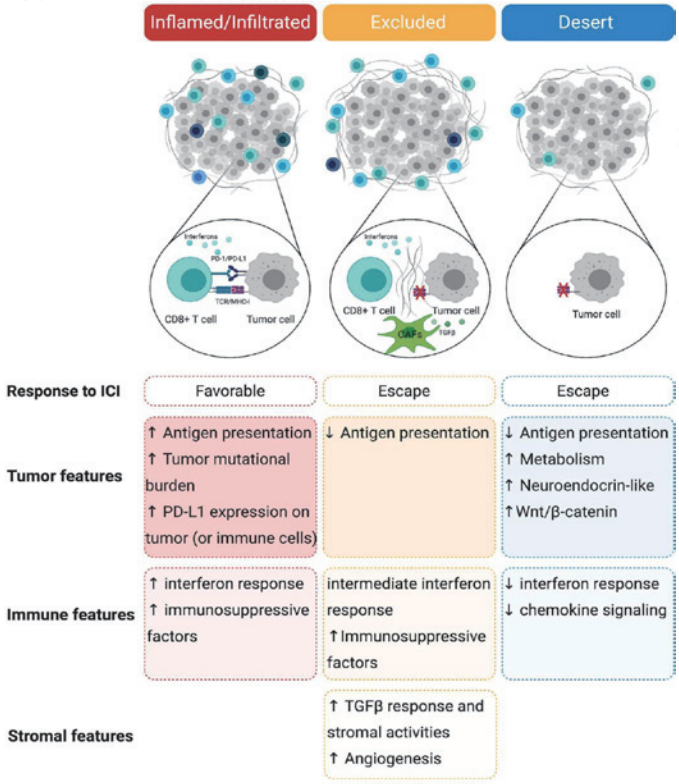
MHC class I – ACC_T_01.filtered.tsv

```
Chromosome  Start  Stop  Reference  Variant  Transcript  Transcript Support Level  Transcript Length  Biotype  Ensembl Gene ID  Variant Type  Mutation  Protein Position
  Gene Name  HGVSc  HGVSp  HLA Allele  Peptide Length  Sub-peptide Position  Mutation Position  MT Epitope Seq  WT Epitope Seq  Best MT IC50 Score  Method  Best MT IC50 Score  Corr
esponding WT IC50 Score  Corresponding Fold Change  Best MT Percentile  Method  Best MT Percentile  Corresponding WT Percentile  Tumor DNA Depth  Tumor DNA VAF  Tumor RNA Depth  Tumor RNA
VAF  Normal Depth  Normal VAF  Gene Expression  Transcript Expression  Median MT IC50 Score  Median WT IC50 Score  Median Fold Change  Median MT Percentile  Median WT Percentile  NetMHCpa
n WT IC50 Score  NetMHCpan MT IC50 Score  NetMHCpan WT Percentile  NetMHCpan MT Percentile  Index  cterm_7mer_gravy_score  max_7mer_gravy_score  difficult_n_terminal_residue  c_terminal_cysteine  c_term
inal_proline  cysteine_count  n_terminal_asparagine  asparagine_proline_bond_count
chr6  7884402  7884403  C  T  ENST00000379757.9  1  432  protein_coding  ENSG00000239264  missense  E/K  378  TXNDC5  ENST00000379757.9:c.1132G>A  ENSP00000369081.4:p.Gl
u378Lys  HLA-A*02:01  10  2  9  GLAGVKIAKV  GLAGVKIAEV  NetMHCpan  59.53  17.05  0.286  NetMHCpan  0.48  0.16  136  0.365  NA  NA  NA  NA
  NA  59.53  17.05  0.286  0.48  0.16  17.05  59.53  0.16  0.48  1.TXNDC5.ENST00000379757.9.missense.378E/K  0.9285714285714286  1.6857142857142855  False  False  False  0
  False  0
chr9  128583816  128583817  C  G  ENST00000372739.7  1  2477  protein_coding  ENSG00000197694  missense  Q/E  681  SPTAN1  ENST00000372739.7:c.2041C>G  ENSP00
000361824.4:p.Gln681Glu  HLA-B*40:02  9  3  7  REANQQQEQF  REANQQQQF  NetMHCpan  361.59  384.65  1.064  NetMHCpan  0.57  0.6  96  0.321  NA  NA  NA
  NA  NA  NA  361.59  384.65  1.064  0.57  0.6  384.65  361.59  0.6  0.57  4.SPTAN1.ENST00000372739.7.missense.681Q/E  -1.8428571428571427  -1.8428571428571427  False  Fa
lse  False  0  False  0
```

MHC class II – ACC_T_01.filtered.tsv

```
Chromosome  Start  Stop  Reference  Variant  Transcript  Transcript Support Level  Transcript Length  Biotype  Ensembl Gene ID  Variant Type  Mutation  Protein Position
  Gene Name  HGVSc  HGVSp  HLA Allele  Peptide Length  Sub-peptide Position  Mutation Position  MT Epitope Seq  WT Epitope Seq  Best MT IC50 Score  Method  Best MT IC50 Score  Corr
esponding WT IC50 Score  Corresponding Fold Change  Best MT Percentile  Method  Best MT Percentile  Corresponding WT Percentile  Tumor DNA Depth  Tumor DNA VAF  Tumor RNA Depth  Tumor RNA
VAF  Normal Depth  Normal VAF  Gene Expression  Transcript Expression  Median MT IC50 Score  Median WT IC50 Score  Median Fold Change  Median MT Percentile  Median WT Percentile  NetMHCII
pan WT IC50 Score  NetMHCIIpan MT IC50 Score  NetMHCIIpan WT Percentile  NetMHCIIpan MT Percentile  Index  cterm_7mer_gravy_score  max_7mer_gravy_score  difficult_n_terminal_residue
  c_terminal_cysteine  c_terminal_proline  cysteine_count  n_terminal_asparagine  asparagine_proline_bond_count
chr6  7884402  7884403  C  T  ENST00000379757.9  1  432  protein_coding  ENSG00000239264  missense  E/K  378  TXNDC5  ENST00000379757.9:c.1132G>A  ENSP00000369081.4:p.Gl
u378Lys  HLA-DQA1*01:02/DQB1*03:01  16  3  14  KKEFPGLAGVKIAKVD  KKEFPGLAGVKIAEVD  NetMHCIIpan  375.12  373.19  0.995  NetMHCIIpan  2.7  2.6  136  0.365  NA
  NA  NA  NA  NA  NA  375.12  373.19  0.995  2.7  2.6  373.19  375.12  2.6  2.7  1.TXNDC5.ENST00000379757.9.missense.378E/K  0.48571428571428577  1.6857142857142855
  False  False  False  0  False  0
```

Tumor immune microenvironments (TIME)



실습용 데이터 안내

Bulk RNA-seq을 이용한 tumor immune microenvironemnts 분석

Prerequisites

Gene quantification file – HTseq 등

- ~. htseq.count.txt

Raw fastq 파일

- ACC_T_01_1.fastq.gz
- ACC_T_02_1.fastq.gz

Processed data

Normalized expression matrix

- normalized_TPM.rds

Cell type decomposition

- abis.rds
- cibersort_abs.rds
- consensus_tme.rds
- epic.rds
- estimate.rds
- mcp_counter.rds
- quantiseq.rds
- timer.rds
- xcell.rds

Immune cell repertoire

- TRUST4_dat.rds

Tumor immune dysfunction and exclusion

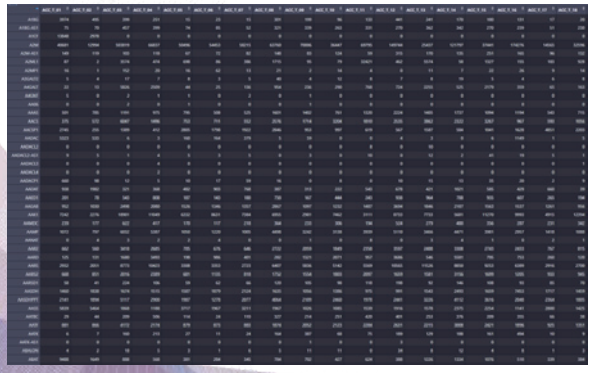
- TIDE_dat.rds

실습 데이터: /home/jyhong906/BIML_2024/Bulk_RNA/Data

실습 스크립트: /home/jyhong906/BIML_2024/Bulk_RNA/Script

TPM (Transcripts Per Million) normalization

```
##########################################
# Load expression data & TPM normalization #
##########################################
SGC_dir <- "/data/project/BIML_2024/Bulk_RNA/Sample"
SGC_files <- list.files(SGC_dir)
SGC_path <- paste0(SGC_dir, "/", SGC_files)
SGC_names <- gsub(".htseq.count.txt", "", SGC_files)

tmp_df_list <- c()
for (idx in seq(SGC_names)) {
  tmp_df <- read.table(
    SGC_path[idx],
    header = F,
    sep = "\t",
    stringsAsFactors = F,
    col.names = c("Symbol", SGC_names[idx])
  )
  tmp_df_list[[idx]] <- tmp_df
}

SGC_count_df <- Reduce(merge, tmp_df_list)[-c(1:5),]
rownames(SGC_count_df) <- SGC_count_df$Symbol; SGC_count_df <- SGC_count_df[,-1]

# Gene filter #
SGC_mat <- as.matrix(SGC_count_df[rowSums(SGC_count_df >= 1) >= ncol(SGC_count_df),])

load("/data/project/BIML_2024/Bulk_RNA/ImmuneDeconv/gene_cov.rda")
normalized_TPM <- countToTpm(SGC_mat,
                             keyType = "SYMBOL",
                             gene_cov = gene_cov)
```



	ACC_T_01	ACC_T_02	ACC_T_03	ACC_T_04	ACC_T_05	ACC_T_06
A1BG	8.818330e+01	1.198082e+01	1.740194e+00	1.715061e+00	3.228799e-01	4.781070e-01
A1BG-AS1	3.672197e+00	3.738398e+00	4.397919e+00	6.015682e+00	3.514696e+00	3.898723e+00
A2M	1.355642e+03	4.723013e+02	3.299852e+03	6.858320e+02	1.632307e+03	1.699864e+03
A2M-AS1	1.864123e+01	9.269988e+00	1.445869e+00	2.594986e+00	4.641643e+00	4.817003e+00
A2ML1	2.559659e+00	6.418735e-02	2.066730e+01	4.294261e+00	1.992093e+01	2.370282e+00
A3GALT2	1.829029e+00	8.979255e-01	6.876585e-01	4.436115e-01	1.597125e+00	5.783857e-01
A4GALT	1.379352e+00	8.898345e-01	7.179415e+01	4.883958e+01	2.676064e+00	1.668356e+00
AAAS	2.231440e+01	3.813643e+01	1.042619e+01	1.337209e+01	3.434836e+01	2.119581e+01
AACS	4.047587e+00	6.734165e+00	1.282836e+01	4.972145e+00	7.804083e+00	7.189029e+00
AACSP1	1.709853e+02	1.733529e+01	1.709513e+01	7.902435e+00	1.694886e+02	1.049168e+02
AADAT	6.549355e+01	1.535086e+02	4.477477e+00	8.032889e+00	3.316477e+01	5.906574e+01
AAGAB	4.843727e+01	5.716138e+01	2.498052e+01	3.258767e+01	7.531622e+01	6.415440e+01
AAK1	6.028251e+01	2.066464e+01	1.095231e+01	2.832064e+01	5.112873e+01	6.730273e+01
AAMDC	2.817508e+01	1.629794e+01	1.012031e+01	1.083777e+01	1.392121e+01	9.232547e+00
AAMP	7.387827e+01	5.991051e+01	8.170519e+01	1.143184e+02	7.019441e+01	7.876269e+01
AAR2	4.494583e+01	4.147078e+01	4.561102e+01	5.613361e+01	4.643140e+01	4.295489e+01
AARD	1.094734e+01	1.251700e+01	2.093402e+01	9.419067e+01	1.642113e+01	8.087361e+01
AARS2	2.682950e+01	3.728108e+01	1.591450e+01	2.954606e+01	2.653228e+01	4.270422e+01
AARSD1	3.296782e+00	2.541961e+00	2.502514e+00	1.855305e+00	3.253160e+00	3.301358e+00
AASDH	6.604800e+01	9.069327e+01	1.488431e+01	2.110809e+01	6.964260e+01	7.962919e+01
AASDHPPT	2.508909e+01	7.265482e+01	3.527334e+01	3.131916e+01	6.750113e+01	4.198809e+01
AASS	1.844746e+02	1.882919e+02	1.159956e+01	1.155745e+01	1.139153e+02	5.821585e+01
AATEC	1.252696e+00	2.073116e+00	1.774416e+00	6.730470e+00	4.276874e+00	9.711753e-01
AATF	3.050280e+01	3.270385e+01	2.839022e+01	2.312741e+01	2.952140e+01	2.811465e+01
AATK	9.388597e-02	1.194731e-01	4.920001e-01	1.026305e+00	4.098307e-01	1.612430e-01
ABAT	1.995402e+02	3.782676e+01	3.670583e+00	3.678335e+00	7.772701e+00	5.595163e+00
ABCA1	2.609129e+02	1.176262e+02	2.447538e+01	3.352699e+01	1.362989e+02	7.617583e+01
ABCA10	2.123992e+01	1.581991e+01	2.366041e-01	2.170487e+00	4.806001e+00	6.705508e+00
ABCA11P	1.501424e+01	2.589371e+01	3.363030e+00	6.891310e+00	1.909869e+01	2.357910e+01
ABCA12	5.036825e-01	5.971615e-01	4.773390e+00	2.474813e+00	1.571998e+00	7.180188e-01
ABCA13	5.218505e+01	3.363343e+02	6.717041e+01	1.339675e+02	3.532175e+01	1.342195e+02
ABCA17P	1.281042e+01	1.470647e+01	1.447764e-01	3.057119e-01	3.306667e-01	4.200203e-01
ABCA2	4.391191e+00	4.350612e+00	1.614670e+01	5.403588e+01	5.590786e+00	8.685879e+00
ABCA3	2.944068e+00	5.417196e+00	9.751472e+00	2.478848e+01	7.827581e+01	2.345458e+01
ABCA4	9.689918e-01	2.955527e-01	2.888108e-01	4.589843e+00	5.236945e-01	5.538210e-01

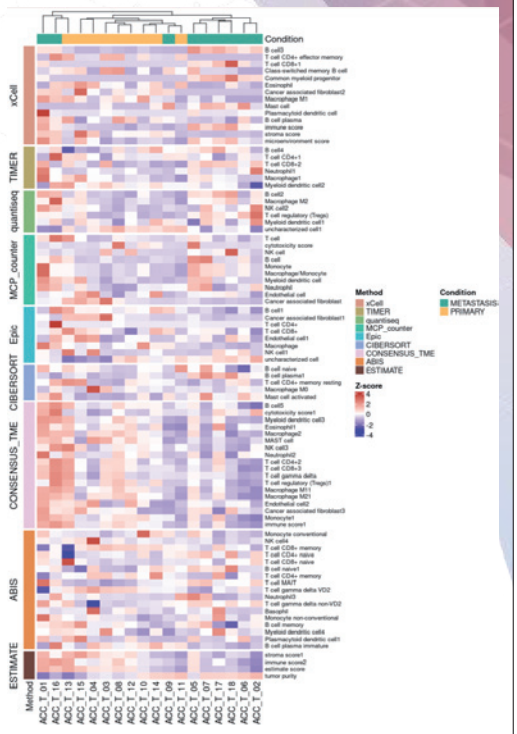
Load library & normalization

```
##################
# Load packages #
##################
library(GeoTcgaData) # normalized TPM
library(immunedeconv) # Cell type decomposition
library(ComplexHeatmap) # Visualization
library(ggplot2) # Visualization
library(ggpol) # Visualization
library(gridExtra) # Visualization
library(ggpubr) # statistics
library(circlize) # color
library(metapod) # combined p-value
library(gtools) # processing
library(dplyr) # processing

##########################################
# Load expression data & TPM normalization #
##########################################
SGC_dir <- "/data/project/BIML_2024/Bulk_RNA/Data/ImmuneDeconv"
SGC_files <- list.files(SGC_dir,pattern = "*txt")
SGC_path <- paste0(SGC_dir, "/", SGC_files)
SGC_names <- gsub(".htseq.count.txt", "", SGC_files)

tmp_df_list <- c()
for (idx in seq(SGC_names)) {
  tmp_df <- read.table(
    SGC_path[idx],
    header = F,
    sep = "\t",
    stringsAsFactors = F,
    col.names = c("Symbol", SGC_names[idx])
  )
  tmp_df_list[[idx]] <- tmp_df
}

SGC_count_df <- Reduce(merge, tmp_df_list)[-c(1:5),]
rownames(SGC_count_df) <- SGC_count_df$Symbol; SGC_count_df <- SGC_count_df[,-1]

# Gene filter #
SGC_mat <- as.matrix(SGC_count_df[rowSums(SGC_count_df >= 1) >= ncol(SGC_count_df),])

load("/data/project/BIML_2024/Bulk_RNA/Data/ImmuneDeconv/gene_cov.rda") # https://github.com
/YuLab-SMU/GeoTcgaData
normalized_TPM <- countToTpm(SGC_mat,
                             keyType = "SYMBOL",
                             gene_cov = gene_cov)

# saveRDS(normalized_TPM, file = "/data/project/BIML_2024/Bulk_RNA/Data/ImmuneDeconv
/normalized_TPM.rds")
# read_rds("/data/project/BIML_2024/Bulk_RNA/Data/ImmuneDeconv/normalized_TPM.rds")

# Primary & metastasis 정보 확인 #
SGC_groups <- read.table("/data/project/BIML_2024/Bulk_RNA/Script/SGC_groups.txt",
                         sep = "\t",
                         header = T)
P_idx <- SGC_groups$Condition == "PRIMARY"; M_idx <- SGC_groups$Condition == "METASTASIS"
```

Install.packages()

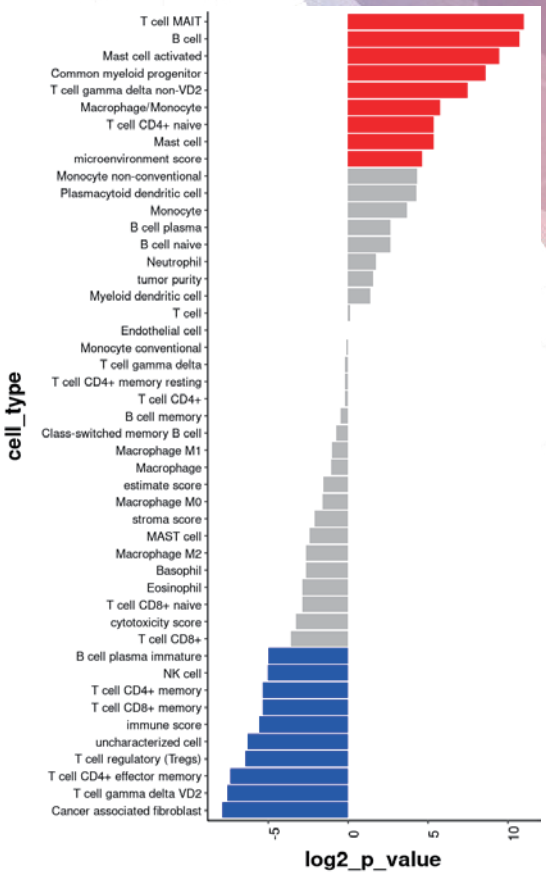
Immune cell deconvolution

```
tool_df_list <- c()
for (idx in seq(length(deconvolution_methods))){
        tool <- deconvolution_methods[idx]

        if (tool %in% "cibersort_abs") {
                set_cibersort_binary('/data/project/BIML_2024/Bulk_RNA/Script/CIBERSORT.R'
)
                set_cibersort_mat('/data/project/BIML_2024/Bulk_RNA/Script/LM22.txt')
                print(tool)
                assign(tool, as.data.frame(deconvolute(normalized_TPM, tool)))
                assign(paste0(tool, '_df'), data.frame(get(tool),
                                                       row.names = get(tool)$cell_type)[
,-1])
                # saveRDS(get(tool), file = paste0("/data/project/BIML_2024/Bulk_RNA/Data
/ImmuneDeconv/",tool,".rds"))

        }

        else if (tool %in% c("timer", "consensus_tme")) {
                print(tool)
                assign(tool, deconvolute(normalized_TPM, tool,
                                         indications=c(rep("hnsc", ncol(normalized_TPM
)))))
                assign(paste0(tool, '_df'), data.frame(get(tool),
                                                       row.names = get(tool)$cell_type)[
,-1])
                # saveRDS(get(tool), file = paste0("/data/project/BIML_2024/Bulk_RNA/Data
/ImmuneDeconv/",tool,".rds"))

        }

        else if (!tool %in% c("cibersort")) {
                print(tool)
                assign(tool, as.data.frame(deconvolute(normalized_TPM, tool)))
                assign(paste0(tool, '_df'), data.frame(get(tool),
                                                       row.names = get(tool)$cell_type)[
,-1])
                # saveRDS(get(tool), file = paste0("/data/project/BIML_2024/Bulk_RNA/Data
/ImmuneDeconv/",tool,".rds"))
        }

        tool_df_list <- append(tool_df_list, paste0(tool, '_df'))
}
# read_rds("/data/project/BIML_2024/Bulk_RNA/Data/ImmuneDeconv/normalized_TPM.rds")

for (tool_df_idx in setdiff(tool_df_list, "cibersort_df")) {
        tmp_tool_df <- get(tool_df_idx)
        tmp_tool_df <- tmp_tool_df[which(rowSums(is.na(tmp_tool_df)) != ncol(tmp_tool_df)),]
        tmp_tool_df <- tmp_tool_df[which(rowSums(tmp_tool_df == 0) < (ncol(tmp_tool_df) * 0
.2)),]
        assign(tool_df_idx, tmp_tool_df)
}
```

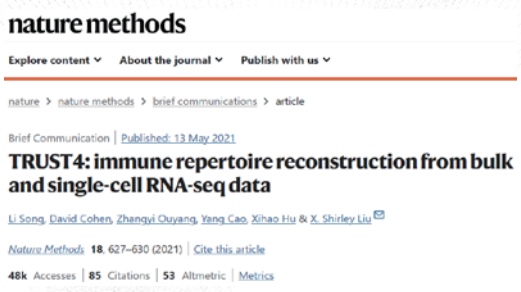


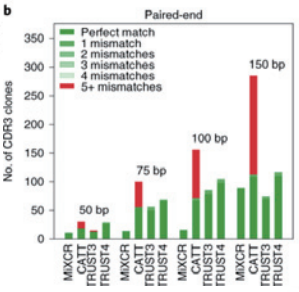
Immune cell deconvolution – cell type

```
cell_type_list <- c()
for (tool in c("mcp_counter_df", "epic_df", "quantiseq_df", "xcell_df", "cibersort_abs_df",
"timer_df","consensus_tme_df", "abis_df", "estimate_df")) {
        print(rownames(get(tool)))
        cell_type_list <- c(cell_type_list, rownames(get(tool)))
}; unique_cell_type <- unique(cell_type_list)

comb_cell_type_list <- c()
comb_p_list <- c()
for (cell_type in unique_cell_type) {
        print(cell_type)
        comb_cell_type_list <- c(comb_cell_type_list, cell_type)

        p_list <- list()
        for (tool in c("mcp_counter_df", "epic_df", "quantiseq_df", "xcell_df",
"cibersort_abs_df", "timer_df","consensus_tme_df", "abis_df","estimate_df")) {
                test_mat <- as.matrix(get(tool))

                if (sum(rownames(test_mat) == cell_type) >= 1) {
                        test <- wilcox.test(test_mat[cell_type, P_idx],
test_mat[cell_type, M_idx])
                        p_list <- c(p_list, test$p.value)
                }
        }; comb_p <- combineParallelPValues(p_list, method = "fisher"); print(comb_p$p.value
); comb_p_list <- c(comb_p_list, comb_p$p.value)
}; cor_idx <- mixedorder(comb_p_list, decreasing = F)
comb_cell_type_list <- comb_cell_type_list[cor_idx]; comb_p_list <- comb_p_list[cor_idx]
cell_type_df <- data.frame(cell_type = seq(length(comb_cell_type_list)),
                           p_value = seq(length(comb_cell_type_list)),
                           freq = length(comb_cell_type_list))
cell_type_df$cell_type <- comb_cell_type_list; cell_type_df$p_value <- comb_p_list; rownames
(cell_type_df) <- cell_type_df$cell_type

freq_list <- c()
for (cell_type in comb_cell_type_list) {
        print(cell_type)

        cell_type_line <- c()
        for (tool in c("mcp_counter_df", "epic_df", "quantiseq_df", "xcell_df",
"cibersort_abs_df",
                       "timer_df","consensus_tme_df", "abis_df","estimate_df")) {
                test_mat <- as.matrix(get(tool))

                if (sum(rownames(test_mat) == cell_type) >= 1) {
                        cell_type_line <- rbind(test_mat[cell_type,], cell_type_line)
                }
        }; mean_value <- apply(cell_type_line, 2, mean)

        if (mean(mean_value[P_idx]) > mean(mean_value[M_idx])) {
                cell_type_df[cell_type,]$freq <- "primary"
        } else {
                cell_type_df[cell_type,]$freq <- "metastasis"
        }

}; cell_type_df$log2_p_value <- -log2(cell_type_df$p_value)
cell_type_df[cell_type_df$freq == "primary",]$log2_p_value <- cell_type_df[cell_type_df$freq
== "primary",]$log2_p_value * -1
cell_type_df <- cell_type_df %>% mutate(color = ifelse(abs(log2_p_value) > -log2(0.05) & freq
== "metastasis", "red",
                                                       ifelse(abs(log2_p_value) > -log2(0.05)
& freq == "primary", "blue",
                                                              ifelse(abs(log2_p_value) < -log2
(0.05), "gray", "NA"))))
cell_type_df$cell_type <- factor(cell_type_df$cell_type, levels = cell_type_df[order
(cell_type_df$log2_p_value, decreasing = F),]$cell_type)
```



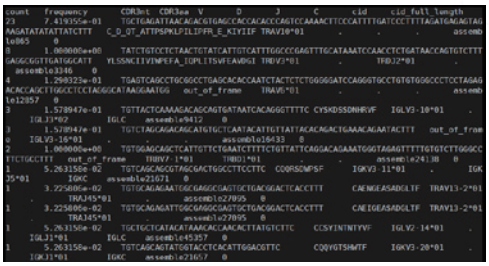
TRUST4





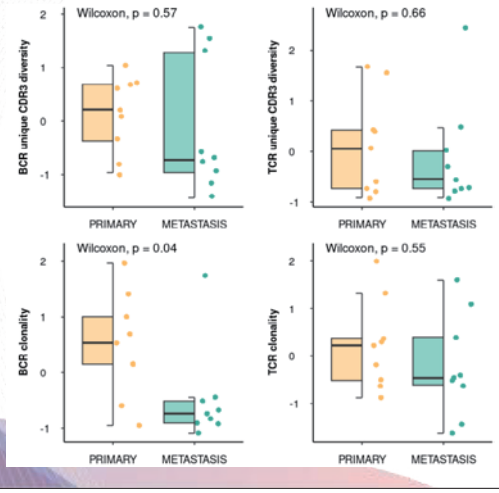

```
# TRUST4 – TCR/BCR CDR3 #
run-trust4 -1 ${FASTQ_PATH}/${patient_id}"_1.fastq.gz" -2 ${FASTQ_PATH}/${patient_id}"_2.fastq.gz" \
-t 10 \
-f ${TRUST4_PATH}/hg38_bcrtcr.fa \
-o ${TCR_PATH}/${patient_id} \
--ref ${TRUST4_PATH}/human_IMGT+C.fa

# custom processing #
python Running_process.py
```

```
-rw-r--r-- 1 jyhong906 jyhong906  101458 Feb  3 01:38 ACC_T_01_report.tsv
-rw-r--r-- 1 jyhong906 jyhong906    7870 Feb  3 01:44 ACC_T_02_report.tsv
-rw-r--r-- 1 jyhong906 jyhong906    1636 Feb  3 01:49 ACC_T_03_report.tsv
-rw-r--r-- 1 jyhong906 jyhong906    8108 Feb  3 01:38 ACC_T_04_report.tsv
-rw-r--r-- 1 jyhong906 jyhong906   83271 Feb  3 01:41 ACC_T_05_report.tsv
-rw-r--r-- 1 jyhong906 jyhong906    3720 Feb  3 01:42 ACC_T_06_report.tsv
-rw-r--r-- 1 jyhong906 jyhong906  441546 Feb  3 01:46 ACC_T_07_report.tsv
-rw-r--r-- 1 jyhong906 jyhong906  226369 Feb  3 01:34 ACC_T_08_report.tsv
-rw-r--r-- 1 jyhong906 jyhong906  242753 Feb  3 01:23 ACC_T_09_report.tsv
-rw-r--r-- 1 jyhong906 jyhong906  325457 Feb  3 01:30 ACC_T_10_report.tsv
-rw-r--r-- 1 jyhong906 jyhong906   18669 Feb  3 01:23 ACC_T_11_report.tsv
-rw-r--r-- 1 jyhong906 jyhong906  539758 Feb  3 01:37 ACC_T_12_report.tsv
-rw-r--r-- 1 jyhong906 jyhong906  151521 Feb  3 01:33 ACC_T_13_report.tsv
-rw-r--r-- 1 jyhong906 jyhong906   65924 Feb  3 01:32 ACC_T_14_report.tsv
-rw-r--r-- 1 jyhong906 jyhong906  119359 Feb  3 01:31 ACC_T_15_report.tsv
-rw-r--r-- 1 jyhong906 jyhong906 2265012 Feb  3 01:35 ACC_T_16_report.tsv
-rw-r--r-- 1 jyhong906 jyhong906    2921 Feb  3 01:36 ACC_T_17_report.tsv
-rw-r--r-- 1 jyhong906 jyhong906   14572 Feb  3 01:46 ACC_T_18_report.tsv
```



TRUST4

```
##########
# TRUST4 #
##########
TRUST4_dir <- '/data/project/BIML_2024/Bulk_RNA/Data/TRUST4'
source("/data/project/BIML_2024/Bulk_RNA/Script/trust4_metric_functions.R")

# Read the individual sample result from RIMA
Trsut4_raw_files <- list.files(path = TRUST4_dir,
                               pattern = 'report.processed.tsv')

TRUST4_list <- list()
for (idx in seq(Trsut4_raw_files)){
        TRUST4_list[[idx]] <- read.table(file = paste0(TRUST4_dir, '/', Trsut4_raw_files[idx]),
                                         sep = "\t",
                                         header = T,
                                         stringsAsFactors = FALSE
        )
        names(TRUST4_list)[[idx]] <- paste0(gsub('-','_',gsub('_T_report.processed.tsv','',Trsut4_raw_files[idx])), '_CDR3')
        sample_list <- names(TRUST4_list)
}; p_samples <- sample_list[P_idx]; m_samples <- sample_list[M_idx]

TRUST4_df <- make_TRUST4_df(TRUST4_list, sample_list, p_samples, m_samples, names(SGC_cols))
rownames(TRUST4_df) <- SGC_groups$SAMPLE
TRUST4_df[is.na(TRUST4_df)] <- 0; TRUST4_df
TRUST4_t_df <- t(TRUST4_df)
TRUST4_t_df <- TRUST4_t_df[rowSums(TRUST4_t_df) != 0,]
TRUST4_df <- as.data.frame(scale(t(TRUST4_t_df)))
int_features <- colnames(TRUST4_df)

TRUST4_df$Condition <- SGC_groups$Condition; TRUST4_dat <- melt(TRUST4_df, "Condition")
TRUST4_dat$Condition <- factor(TRUST4_dat$Condition, levels = c("PRIMARY","METASTASIS"))
# saveRDS(TRUST4_dat, file = "/data/project/BIML_2024/Bulk_RNA/Data/TRUST4/TRUST4_dat.rds")
# read_rds("/data/project/BIML_2024/Bulk_RNA/TRUST4/Data/TRUST4_dat.rds")
```



TIDE (Tumor immune dysfunction and exclusion)

nature medicine

Explore content ~ About the journal ~ Publish with us ~

nature > nature medicine > articles > article

Article | Published: 20 August 2018

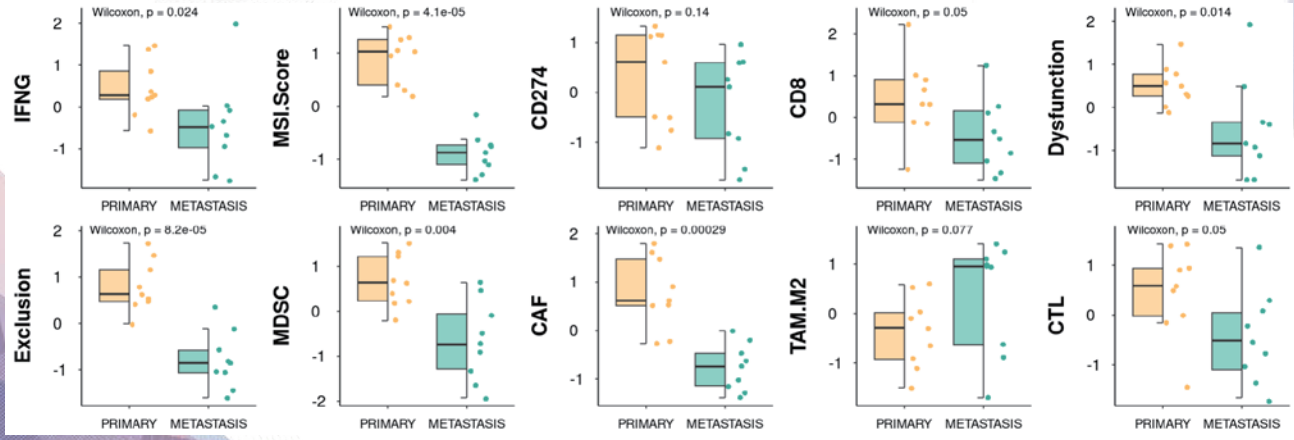
Signatures of T cell dysfunction and exclusion predict cancer immunotherapy response

Peng Jiang, Shengqing Gu, Deng Pan, Jingxin Fu, Avinash Sahu, Xihao Hu, Ziyi Li, Nicole Traugh, Xia Bu, Bo Li, Jun Liu, Gordon J. Freeman, Myles A. Brown, Kai W. Wucherpfennig & X. Shirley Liu

Nature Medicine 24, 1550–1558 (2018) | Cite this article

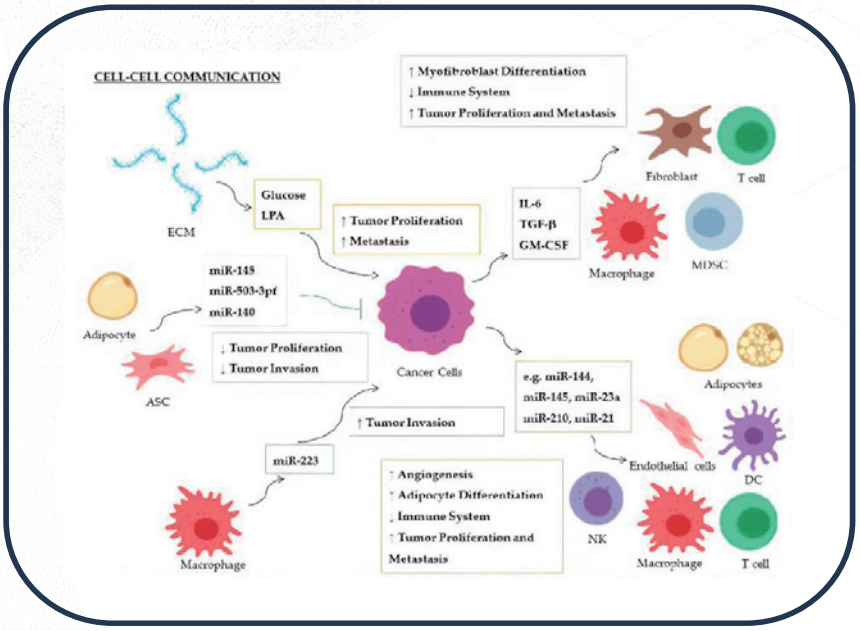
59k Accesses | 2269 Citations | 157 Altmetric | Metrics

	IFNG	MSI.Score	CD274	CD8	Dysfunction	Exclusion	MDSC	CAF	TAM.M2	CTL
ACC_T_16	2.77138452	0.30276296	0.5687156	2.2212168	1.9300687	-0.45261846	-0.051364829	0.002992495	-0.036953976	2.20361677
ACC_T_03	0.50849971	0.88954761	1.2365711	-2.2484679	-0.09121424	1.49101306	0.089043357	0.139496973	-0.006335196	-2.38583075
ACC_T_13	2.07460238	0.78585513	1.0657304	1.2062644	1.48154929	0.71345098	-0.009393384	0.127709458	-0.013916032	1.53371222
ACC_T_15	0.38777099	0.80777155	1.0323718	0.5617749	0.98075486	1.25315204	0.025434708	0.154790587	-0.012569288	0.95997048
ACC_T_08	1.93107510	0.95170196	1.0657304	1.8198446	0.60317765	0.57082723	0.039209776	0.047777594	-0.020100418?	2.25467003
ACC_T_11	-0.79437342	0.54803795	-0.7130554	-0.2614655	0.53027842	0.43202697	0.071956051	-0.019532709	-0.003782193	-0.25541541
ACC_T_04	0.34088156	0.58457538	-0.4547434	0.5755183	0.80198791	0.38928623	0.015838670	0.079258652	-0.032564151	-0.02873558
ACC_T_09	-1.36521189	0.27038910	-0.8619188	0.5130158	0.52122515	0.32742177	0.039153032	-0.016429675	0.023573842	0.42955880
ACC_T_14	0.25552726	0.81373064	-0.4547434	1.6208195	0.29709455	0.49503761	0.041898668	0.046308336	-0.024161042?	1.45824411
ACC_T_12	1.20822731	0.87802523	0.5687156	4.0100306	0.05527366	1.01367462	0.077175812	0.055399598	0.011464665	2.32943756
ACC_T_10	-0.27233313	0.61581230	-1.0279287	-0.2138869	0.32945539	0.03174874	0.013280870	-0.015384041	0.000216323?	0.80380597
ACC_T_06	-0.68389175	0.10199429	-0.7707209	-2.6987001	-1.60770152	-0.05589544	0.030368055	-0.036787790	0.020192716?	-2.23830973
ACC_T_17	-2.45351281	0.22641180	-1.6305932	-2.4062188	-1.61629434	-0.65583489	-0.027577810	-0.093492711	0.026725220?	-2.71331673
ACC_T_02	-2.36032913	0.15918562	-1.4312844	-1.9733693	-1.07253935	-0.68406541	-0.000513432	-0.114392118	0.020442665?	-1.79278383
ACC_T_18	-0.48025992	0.06943787	0.5570338	-1.4738693	-0.86554840	-0.83671020	-0.070262950	-0.068010792	0.021881858?	-1.17529126
ACC_T_07	-0.99015469	0.26026171	0.1063724	-0.9822203	-0.78611890	-0.85666243	-0.039153173	-0.106686590	0.030245230?	-0.83433846
ACC_T_01	-0.11512170	0.44700400	0.8999215	-0.5671828	-0.29825962	-1.18606179	-0.186706009	-0.052049209	-0.013716004?	0.42539174
ACC_T_05	0.03725959	0.18103876	0.2438759	0.2783981	-0.34984742	-1.31150590	-0.089348783	-0.088941825	-0.019189934?	0.07702116



TME and cell to cell interaction

Cell-cell communication within the tumor microenvironment



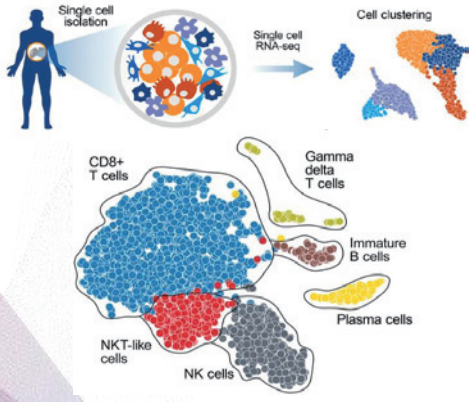
Conti I, Varano G, Simioni C, Laface I, Milani D, Rimondi E, Neri LM. miRNAs as Influencers of Cell–Cell Communication in Tumor Microenvironment. Cells. 2020 Jan 15;9(1):220. doi: 10.3390/cells9010220. PMID: 31952362; PMCID: PMC7016744.

TME and cell to cell interaction

Cell type annotation

Single R

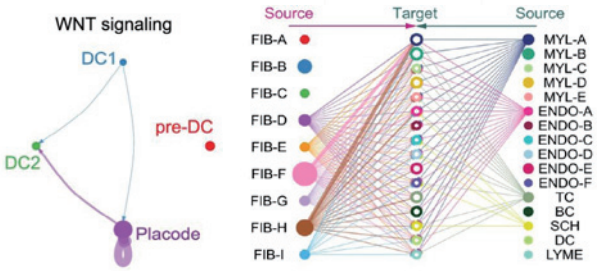
- computational method for unbiased cell type recognition of scRNA-seq
- SingleR's annotations combined with Seurat, a processing and analysis package designed for scRNA-seq



Cell to cell interaction

CellChat

- Infer cell-cell communication networks
- easy-to-use tool for extracting and visualizing



Ianevski, A., Giri, A.K. & Aittokallio, T. Fully-automated and ultra-fast cell-type identification using specific marker combinations from single-cell transcriptomic data. *Nat Commun* **13**, 1246 (2022).

실습용 데이터 안내

Single cell RNA-seq을 이용한 cell to cell interaction prediction

Prerequisites

Raw single cell data

- Barcodes.tsv
- Features.tsv
- matrix.mtx

Human single cell reference

- monaco.ref.rda
- hpca.ref.rda
- dice.ref.rda

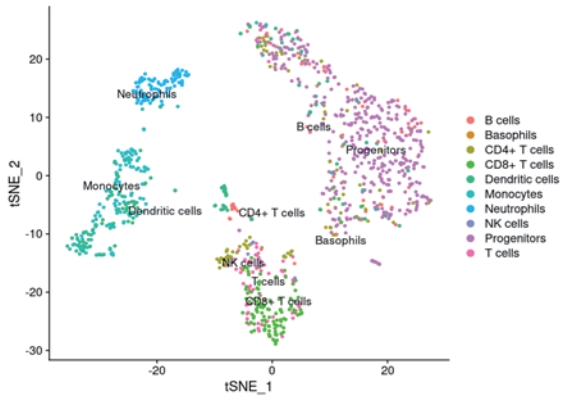
Processed data

Seurat object

- CRC_obj.rda
- CRC_count.rda

Cell type annotation

```
library('dplyr')
library('Seurat')
library('SingleR')
library('CellChat')
library('ADImpute')

#  Cell type/state annotation #
load("/data/project/BIML_2024/scRNA/ref/monaco.ref.rda")      # Reference single cell data. celldex::MonacoImmuneData() 로 다운 가능
load('/data/project/BIML_2024/scRNA/CRC_obj.rda')             #Seurat object
load('/data/project/BIML_2024/scRNA/CRC_count.rda')           # Single cell expression count file
monaco.main <- SingleR(method='single',sc_data=CRC_count, ref_data=monaco.ref@assays@data@listData$logcounts,types=monaco.ref$label.main)
CRC_obj@meta.data$monaco.main <- monaco.main$labels1
CRC_obj_monaco.main <- SetIdent(CRC_obj, value = "monaco.main")
DimPlot(CRC_obj_monaco.main, reduction = "tsne", label = TRUE, repel = TRUE, group.by = 'monaco.main')
```

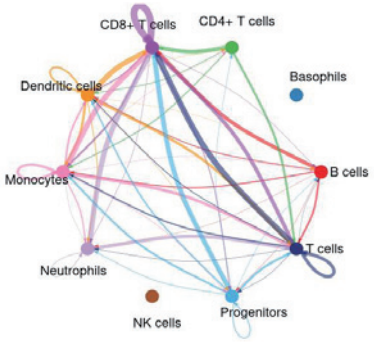


CellChat

```
#  CellChat object  #
CellChatDB <- CellChatDB.human
cellchat <- createCellChat(object = CRC_obj_monaco.main, group.by = "monaco.main", assay = "RNA")
cellchat@DB <- CellChatDB
cellchat <- subsetData(cellchat)
cellchat <- identifyOverExpressedGenes(cellchat)
cellchat <- identifyOverExpressedInteractions(cellchat)
cellchat <- computeCommunProb(cellchat)
cellchat <- filterCommunication(cellchat, min.cells = 10)
cellchat <- computeCommunProbPathway(cellchat)
cellchat <- aggregateNet(cellchat)
cellchat <- netAnalysis_computeCentrality(cellchat, slot.name = "netP")
```

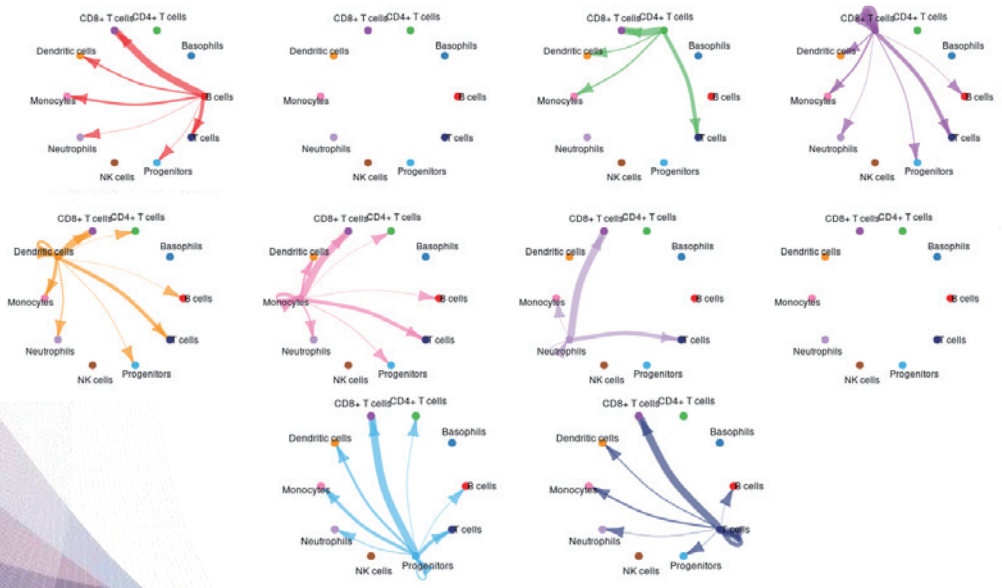
Visualization

```
netVisual_circle(cellchat@net$weight, weight.scale = T, label.edge= F, title.name = "Interaction weights/strength")      #전체 세포 상호작용
```



Visualization

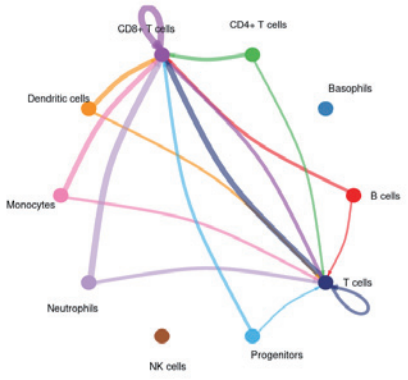
```
mat <- cellchat@net$weight
par(mfrow = c(3,4))
for (i in 1:nrow(mat)) {
        mat2 <- matrix(0, nrow = nrow(mat), ncol = ncol(mat), dimnames = dimnames(mat))
        mat2[i, ] <- mat[i, ]
        netVisual_circle(mat2, weight.scale = T,  title.name = rownames(mat)[i])
}
dev.off()
```

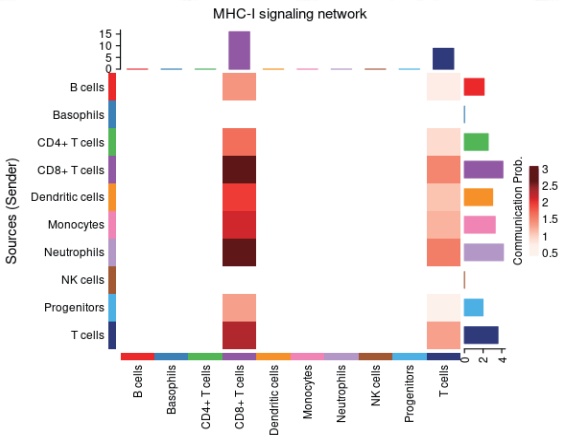


Visualization

```
pathways.show <- c("MHC-I")
netVisual_aggregate(cellchat, signaling = pathways.show, layout = "circle")
netVisual_heatmap(cellchat, signaling = pathways.show, color.heatmap = "Reds")
```

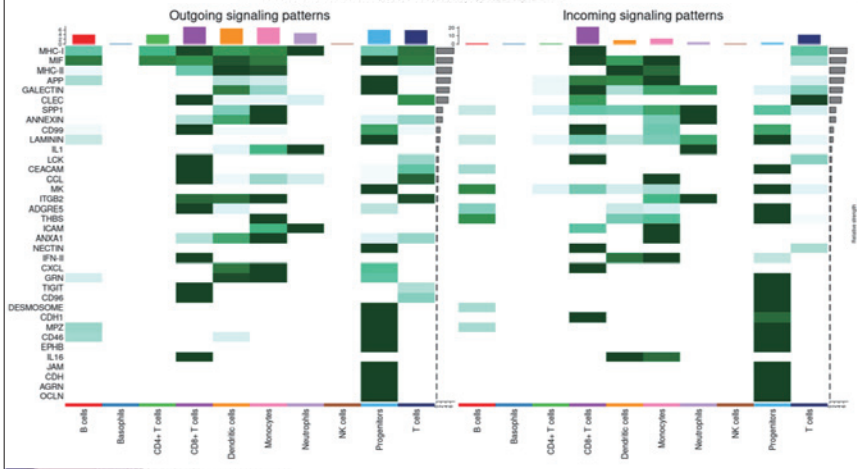
#특정 생물학적 경로 내 상호작용

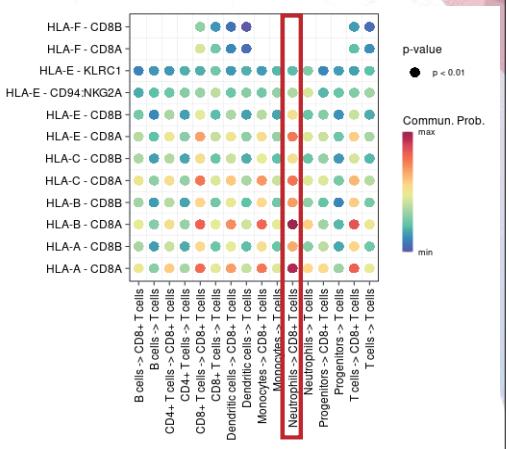




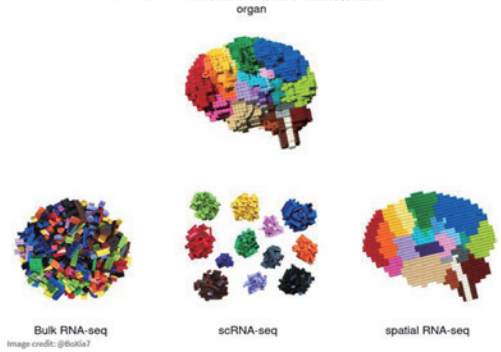
Visualization

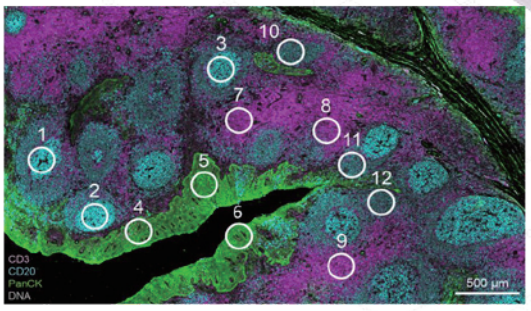
```
ht1 <- netAnalysis_signalingRole_heatmap(cellchat, pattern = "outgoing", width = 20, height = 20, font.size = 13 , font.size.title = 20) ; ht1
ht2 <- netAnalysis_signalingRole_heatmap(cellchat, pattern = "incoming", width = 20, height = 20, font.size = 13 , font.size.title = 20) ; ht2
ht1 + ht2
netVisual_bubble(cellchat, signaling=pathways.show,  remove.isolate = FALSE)          #Outgoing/Incoming signaling
```

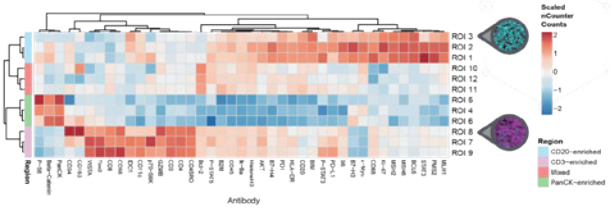




What is spatial transcriptomics?







실습용 데이터 안내

spatial RNA-seq을 이용한 DEG, GSEA 분석

Prerequisites

Processed GoeMX data

- count.rds
- anno.rds
- genemeta.txt
- msigdb_hs.RData

실습 데이터: /home/jyhong906/BIML_2024/GeoMX/Data

실습 스크립트: /home/jyhong906/BIML_2024/GeoMX/Script

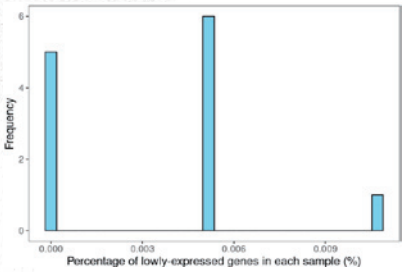
https://cumulus.readthedocs.io/en/stable/geomxngs/index.html#convert-fastq-files-into-dcc-files-by-the-nanostring-geomx-digital-spatial-ngs-pipeline

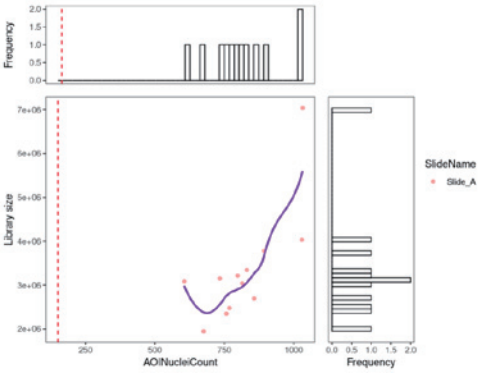
Preparation

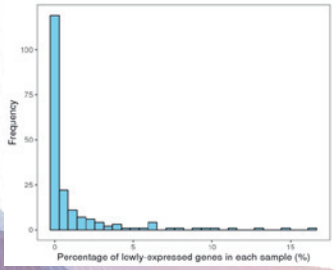
```
source("/data/project/BIML_2024/GeoMX/Script/GeoMX_function.R")

################
# Load library #
################
library(tidyverse)
library(standR)
library(SpatialExperiment)
library(edgeR)
library(limma)
library(msigdb)
library(GSEABase)
library(SpatialDecon)
library(speckle)

# Visualization #
library(ggplot2)
library(ggalluvial)
library(ggrepel)
library(DT)

#############
# Load data #
#############
countFile <- read_rds("/data/project/BIML_2024/GeoMX/Data/count.rds") %>% as.data.frame(); head(countFile)
sampleAnnoFile <- read_rds("/data/project/BIML_2024/GeoMX/Data/anno.rds") %>% as.data.frame(); head(sampleAnnoFile)
featureAnnoFile <- read_tsv("/data/project/BIML_2024/GeoMX/Data/genemeta.txt") %>% as.data.frame()
spe <- readGeoMx(countFile, sampleAnnoFile, featureAnnoFile)
```

QC

```
######
# QC #
######
# Gene level QC #
spe <- addPerROIQC(spe, rm_genes = TRUE)
plotGeneQC(spe, ordannots = "regions", col = regions, point_size = 2)

# ROI level QC #
plotROIQC(spe, x_threshold = 150, color = SlideName)
qc <- colData(spe)$AOINucleiCount > 150; spe <- spe[, qc]

# PCA #
spe <- scater::runPCA(spe)
pca_results <- reducedDim(spe, "PCA")
plotPairPCA(spe, col = SlideName, precomputed = pca_results, n_dimension = 4)
plotPairPCA(spe, col = class, precomputed = pca_results, n_dimension = 4)
```







Normalization

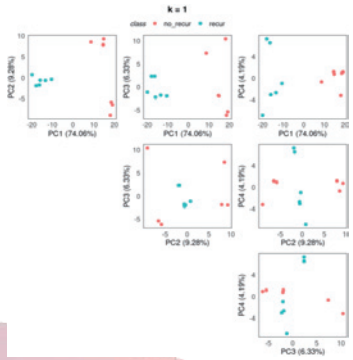
```
##################
# Normalization #
##################
spe_tmm <- geomxNorm(spe, method = "TMM")
plotRLExpr(spe_tmm, assay = 2, color = SlideName) + ggtitle("TMM")
```

TMM, RPKM, TPM, CPM

TMM: 각 샘플의 **Library size**를 이용하여 각 발현 수치를 보정하는 방법

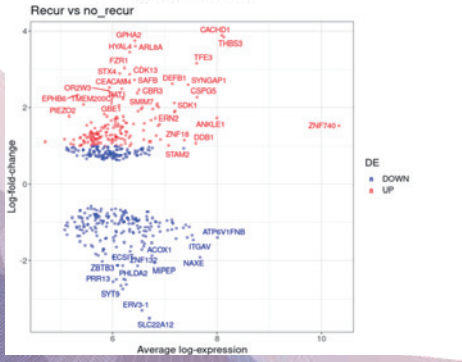
Batch correction

```
####################
# Batch correction #
####################
spe <- findNCGs(spe, batch_name = "SlideName", top_n = 300)
# for(i in seq(3)){
#           spe_ruv <- geomxBatchCorrection(spe, factors = "class",
#                                           NCGs = metadata(spe)$NCGs, k = i)
#
#           print(plotPairPCA(spe_ruv, assay = 2, n_dimension = 4, color = class, title = paste0("k = ", i)))
#
# }
# spe_ruv <- geomxBatchCorrection(spe, factors = "class",
#                                 NCGs = metadata(spe)$NCGs, k = 1)
```



DEGs

```
#######
# DEG #
#######
dge <- SE2DGEList(spe)
# design <- model.matrix(~0 + class + ruv_W1 + ruv_W2 , data = colData(spe_ruv)); colnames(design) <- gsub("^class","",colnames(design)); colnames(design) <- gsub(" ","_",colnames(design))
design <- model.matrix(~0 + class , data = colData(spe)); colnames(design) <- gsub("^class","",colnames(design)); colnames(design) <- gsub(" ","_",colnames(design))

# IM vs CT
contr.matrix <- makeContrasts(
         BvT = recur - no_recur,
         levels = colnames(design)); keep <- filterByExpr(dge, design); table(keep)
dge_all <- dge[keep, ]
v <- voom(dge_all, design, plot = TRUE)

fit <- lmFit(v)
fit_contrast <- contrasts.fit(fit, contrasts = contr.matrix)
efit <- eBayes(fit_contrast, robust = TRUE)
results_efit<- decideTests(efit, p.value = 0.05)
de_results_BvT <- topTable(efit, coef = 1, sort.by = "P", n = Inf)
de_genes_toptable_BvT <- topTable(efit, coef = 1, sort.by = "P", n = Inf, p.value = 0.05)

de_results_BvT <- de_results_BvT %>%
         mutate(DE = ifelse(logFC > 0 & adj.P.Val <0.05, "UP",
                            ifelse(logFC <0 & adj.P.Val<0.05, "DOWN", "NOT DE"))); cor <- c("DOWN" = "Blue",
                                                                                            "NOT DE" = "Grey",
                                                                                            "UP" = "Red")

DEG_vis(de_results_BvT,
        DEG_count = 500,
        title = "Recur vs no_recur",
        cor = cor)

updn_cols <- c(RColorBrewer::brewer.pal(6, 'Greens')[2], RColorBrewer::brewer.pal(6, 'Purples')[2])
DEG_table_vis(de_genes_toptable_BvT,
              DEG_count = 500,
              title = 'Recur vs. no_reucr (limma-voom)',
              cor = updn_cols)
```



Files | Plots | Packages | Help | Viewer | Presentation

Zoom | Export | Publish

Show 10 entries Search:

Recur vs. no_reucr (limma-voom)

	logFC	AveExpr	P.Value	adj.P.Val
SLC22A12	-3.508	6.705	3.699e-15	5.207e-11
ERV3-1	-3.304	6.563	8.363e-12	6.791e-9
SYT9	-2.747	6.196	3.078e-11	1.916e-8
BDKRB1	-2.656	6.160	3.521e-8	0.000004009
PIGT	-2.626	6.263	4.726e-11	2.596e-8
VPS50	-2.552	6.011	1.732e-9	3.762e-7
CD200R1	-2.501	6.246	1.774e-12	2.549e-9
PRR13	-2.490	6.074	1.958e-7	0.00001493

Showing 1 to 10 of 500 entries Previous 1 2 3 4 5 ... 50 Next

GSEA (GeneSet Enrichment Analysis)

```
########
# GSEA #
########
# msigdb_hs <- getMsigdb(version = '7.2'); save(msigdb_hs, file = "/data/project/BIML_2024/GeoMX/msigdb_hs.RData")
load("/data/project/BIML_2024/GeoMX/Data/msigdb_hs.RData")
msigdb_hs <- appendKEGG(msigdb_hs)

sc <- listSubCollections(msigdb_hs)

gsc <- c(subsetCollection(msigdb_hs, c('h')),
         subsetCollection(msigdb_hs, 'c2', sc[grepl("^CP:",sc)]),
         subsetCollection(msigdb_hs, 'c5', sc[grepl("^GO:",sc)])) %>%
          GeneSetCollection()

fry_indices <- ids2indices(lapply(gsc, geneIds), rownames(v), remove.empty = FALSE)
names(fry_indices) <- sapply(gsc, setName)
gsc_category <- sapply(gsc, function(x) bcCategory(collectionType(x)))
gsc_category <- gsc_category[sapply(fry_indices, length) > 5]

gsc_subcategory <- sapply(gsc, function(x) bcSubCategory(collectionType(x)))
gsc_subcategory <- gsc_subcategory[sapply(fry_indices, length) > 5]

fry_indices <- fry_indices[sapply(fry_indices, length) > 5]
names(gsc_category) = names(gsc_subcategory) = names(fry_indices)

fry_indices_cat <- split(fry_indices, gsc_category[names(fry_indices)])
fry_res_out <- lapply(fry_indices_cat, function (x) {
          limma::fry(v, index = x, design = design, contrast = contr.matrix[,1], robust = TRUE)
})

fry_res_sig <- post_fry_format(fry_res_out, gsc_category, gsc_subcategory) %>%
          as.data.frame() %>%
          filter(FDR < 0.25) # 0.05 ~ 0.25

# GSEA - vis
GSEA_vis(df = fry_res_sig,
         DEG_type = "Up",
         cor <- "Red",
         cnt = 10,
         title <- "up-deg")
GSEA_vis(df = fry_res_sig,
         DEG_type = "Down",
         cor <- "Blue",
         cnt = 10,
         title <- "down-deg")
```