KSBi-BIML 2024

Bioinformatics & Machine Learning(BIML) Workshop for Life and Medical Scientists

생명정보학 & 머신러닝 워크샵 (온라인)

Introduction to adaptive immune repertoire sequencing

박대찬 _ 아주대학교

KSBI KOREAN SOCIETY FOR BIOINFORMATICS | 한국생명정보학회

본 강의 자료는 한국생명정보학회가 주관하는 BIML 2024 워크샵 온라인 수업을 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 전적으로 불법 행위자 본인에게 있음을 경고**합니다.

KSBi-BIML 2024

Bioinformatics & Machine Learning(BIML)
Workshop for Life and Medical Scientists

안녕하십니까?

한국생명정보학회가 개최하는 동계 교육 워크샵인 BIML-2024에 여러분을 초대합니다. 생명정보학 분야의 연구자들에게 최신 동향의 데이터 분석기술을 이론과 실습을 겸비해 전달하고자 도입한 전문 교육 프로그램인 BIML 워크샵은 2015년에 시작하여 올해로 벌써 10년 차를 맞이하게 되었습니다. BIML 워크샵은 국내 생명정보학 분야의 최초이자 최고 수준의 교육프로그램으로 크게 인공지능과 생명정보분석 두 개의 분야로 구성되어 있습니다. 올해 인공지능 분야에서는 최근 생명정보 분석에서도 응용이 확대되고 있는 다양한 인공지능 기반 자료모델링 기법들에 대한 현장 강의가 진행될 예정이며, 관련하여 심층학습을 이용한 단백질구조예측, 유전체분석, 신약개발에 대한 이론과 실습 강의가 함께 제공될 예정입니다. 또한 단일세포오믹스, 공간오믹스, 메타오믹스, 그리고 롱리드염기서열 자료 분석에 대한 현장 강의는 많은 연구자의 연구 수월성 확보에 큰 도움을 줄 것으로 기대하고 있습니다.

올해 BIML의 가장 큰 변화는 최근 연구 수요가 급증하고 있는 의료정보자료 분석에 대한 현장 강의를 추가하였다는 것입니다. 특히 의료정보자료 분석을 많이 수행하시는 의과학자 및 의료정보 연구자들께서 본 강좌를 통해 많은 도움을 받으실 수 있기를 기대하고 있습니다. 또한 다양한 생명정보학 분야에 대한 온라인 강좌 프로그램도 점차 증가하고 있는 생명정보 분석기술의 다양화에 발맞추기 위해 작년과 비교해 5강좌 이상을 신규로 추가했습니다. 올해는 무료 강좌 5개를 포함하여 35개 이상의 온라인 강좌가 개설되어 제공되며, 연구 주제에 따른 연관된 강좌 추천 및 강연료 할인 프로그램도 제공되며, 온라인을 통한 Q&A 세션도 마련될 예정입니다. BIML-2024는 국내 주요 연구중심 대학의 전임 교원이자 각 분야 최고 전문가들의 강의로 구성되었기에 해당 분야의 기초부터 최신 연구 동향까지 포함하는 수준 높은 내용의 강의가 될 것이라 확신합니다.

BIML-2024을 준비하기까지 너무나 많은 수고를 해주신 운영위원회의 정성원, 우현구, 백대현, 김태민, 김준일, 김상우, 장혜식, 박종은 교수님과 KOBIC 이병욱 박사님께 커다란 감사를 드립니다. 마지막으로 부족한 시간에도 불구하고 강의 부탁을 흔쾌히 허락하시고 훌륭한 현장 강의와 온라인 강의를 준비하시는데 노고를 아끼지 않으신 모든 강사분들께 깊은 감사를 드립니다.

2024년 2월

한국생명정보학회장 이 인 석

강의개요

Introduction to adaptive immune repertoire sequencing

암 면역 환경, 감염병 반응 면역 등에 대한 관심이 증가함에 따라 적응면역의 다양성 분석에 대한 기술이 성장하고 있다. 특히, B cell receptor (=antibody, BCR)와 T cell receptor (TCR) 서열 분석은 면역항암제와 감염병 치료제 발굴 및 반응성 이해에 핵심적인 요소이다. BCR과 TCR 서열의 천문학적 다양성과 세포별 서열 특이성으로 인해 일반적인 RNA-seq, scRNA-seq, WES 등은 adaptive immune repertoire를 포괄적으로 이해하는 데 한계가 있다. 따라서, high throughput으로 BCR과 TCR full-length 서열을 얻는 방법, 단일세포 BCR의 heavy/light chain (TCR은 alpha/beta chain)을 동시에 시퀀싱하는 기법들이 개발되고 있다.

본 강의에서는 NGS를 이용한 BCR/TCR 시퀀싱 데이터 이론을 학습하고 일련의 생명정보학적 분석과정을 이해하는 것이 목표이다. 이론으로, BCR 유전자의 구조, BCR-seq library를 만드는 최신 기법을 학습한다. 기본 분석에 사용되는 IMGT와 MIXCR 활용 예시를 통해 BCR의 V gene usage와 complementarity-determining regions (CDR) 서열을 동정하는 법을 배운다. 동정된 항체 서열의 특징을 해석하는 downstream 분석 과정을 학습한다.

강의는 다음의 내용을 포함한다:

- Immune repertoire 개요
- BCR sequencing을 위한 실험 기법
- BCR sequencing 데이터 분석 및 해석

* 참고강의교재: 강의자료

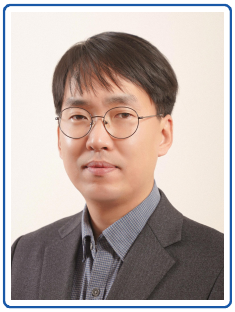
* 교육생준비물: 노트북 (메모리 8GB 이상, 디스크 여유공간 30GB 이상), RStudio 설치

* 강의 난이도: 초급

* 강의: 박대찬 교수 (아주대학교 생명과학과)

Curriculum Vitae

Speaker Name: Daechan Park, Ph.D.



▶ **Personal Info**

Name	Daechan Park
Title	Associate Professor
Affiliation	Ajou University

▶ **Contact Information**

Address	206 Worldcup-ro, Yeongtong-gu, Suwon 16499, Korea
Email	dpark@ajou.ac.kr
Phone Number	+82-31-219-2514

Research Interest

bioinformatics, immune repertoire, epigenomics, synthetic biology

Educational Experience

2014	Ph.D. in Genomics, The University of Texas at Austin, Austin, TX, USA
2008	B.S. in Biological Sciences, Seoul National University, Seoul, Korea

Professional Experience

2018-present	Assistant / Associate Professor, Ajou University
2016-2018	Senior Research Scientist, Korea Institute of Science and Technology
2014-2016	Postdoctoral Fellow, The University of Texas at Austin, Austin, TX, USA

Selected Publications (5 maximum)

1. Jae Yun Moon, Jina Seo, Jaewoo Lee[#], and **Daechan Park**[#], Assessment of attenuation of varicella-zoster virus vaccines based on genomic comparison (2023) Journal of Medical Virology 95(3), e28590 [#]co-corresponding (IF=12.7)
2. Do Young Hyeon*, Dowoon Nam*, Youngmin Han*, Duk Ki Kim*, Gibeom Kim*, Daeun Kim*, Jingi Bae Seunghoon Back, Dong-Gi Mun, Inamul Hasan Madar, Hangyeore Lee, Su-Jin Kim, Hokeun Kim, Sangyeop Hyun, Chang Rok Kim, Juhee Jeong, Suwan Jeon, Yeon Woong Choo, Kyung Bun Lee, Wooil Kwon, Seunghyuk Choi, Taewan Goo, Taesung Park, Young Ah Suh, Hongbeom Kim, Ja-Lok Ku, Min-Sik Kim, Eunok Paek, **Daechan Park**[#], Keehoon Jung[#], Sung Hee Baek[#], Jin-Young Jang[#], Daehee Hwang[#] and Sang-Won Lee[#], Proteogenomic landscape of human pancreatic ductal adenocarcinoma in an Asian population reveals tumor cell-enriched and immune-rich subtypes (2023) Nature Cancer 4(2):290-307. [#]co-corresponding (IF=22.7)
3. Se Won Park*, Jaehoon Kim*, Sungryong Oh*, Jeongyoon Lee, Joowon Cha, Hyun Sik Lee, Keun Il Kim, **Daechan Park**[#] and Sung Hee Baek[#], PHF20 is crucial for epigenetic control of starvation-induced autophagy through enhancer activation (2022) Nucleic Acids Research 50(14):7856-7872. [#]co-corresponding (IF=14.9)
4. Chuna Kim*, Sanghyun Sung*, Jong-Seo Kim*, Hyunji Lee, Yoonseok Jung, Sanghee Shin, Eunkyeong Kim, Jenny J. Seo, Jun Kim, Daeun Kim, Hiroyuki Niida, Narry V. Kim, **Daechan Park**[#] and Junho Lee[#], Telomeres reforged with non-telomeric sequences in mouse embryonic stem cells (2021) Nature Communications 12(1):1097. [#]co-corresponding (IF=16.6)
5. Sungryong Oh*, Kyungjin Boo*, Jaebeom Kim, Seon Ah Baek, Yoon Jeon, Junghyun You, Ho Lee, Hee-Jung Choi, **Daechan Park**[#], Ji Min Lee[#] and Sung Hee Baek[#], The chromatin-binding protein PHF6 functions as an E3 ubiquitin ligase of H2BK120 via H2BK12Ac recognition for activation of trophectodermal genes (2020) Nucleic Acids Research 48(16):9037-9052. [#]co-corresponding (IF=14.9)

KSBI | 한국생명정보학회

KSBi-BIML 2024

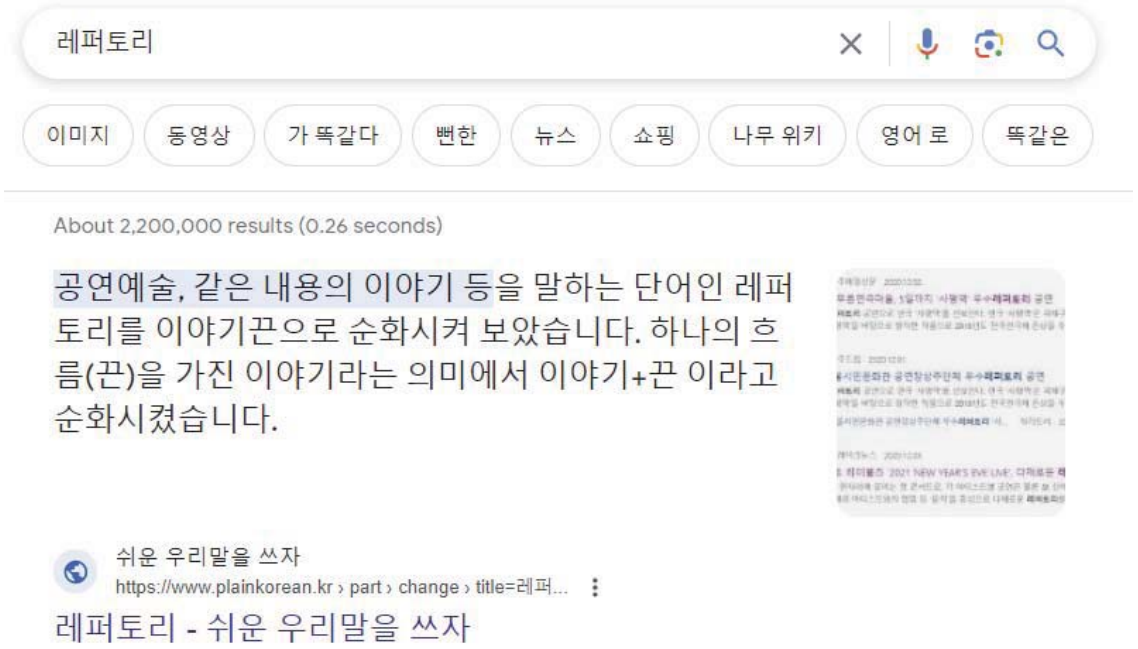
Introduction to adaptive immune repertoire sequencing

아주대학교
박대찬

강의 순서

1. Adaptive Immune Receptor Repertoire (AIRR) 의 정의와 다양성
2. AIRR 연구를 위한 오믹스 접근법
3. AIRR 생명정보 분석법
4. AIRR 연구 사례 및 분석 예시

레퍼토리 (repertoire)는 무엇인가?



우리는 다양한 레퍼토리를 기대한다.

Different People, Different Repertoire, Different B/T cells

A pool of Adaptive Immune Receptor is a Repertoire

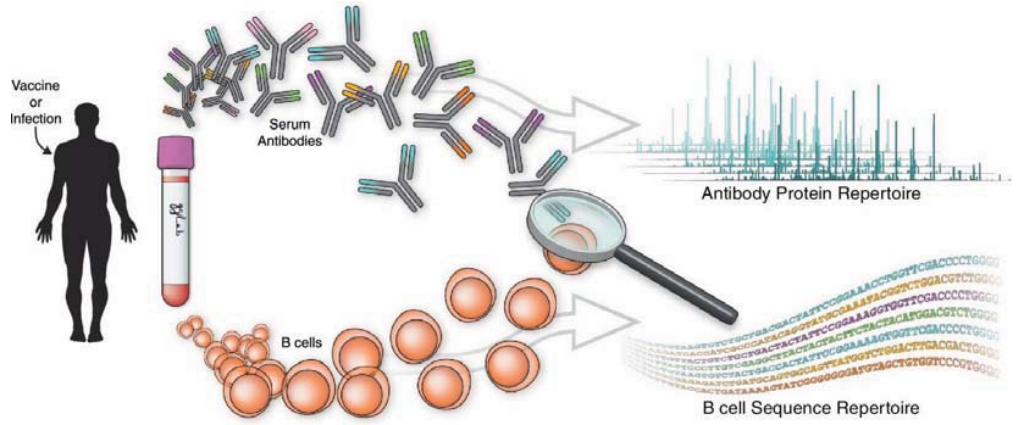
Adaptive Immune Receptor Repertoire Community recommendations for sharing immune-repertoire sequencing data

Florian Rubelt[1,21], Christian E Busse[2,21], Syed Ahmad Chan Bukhari[3,21], Jean-Philippe Bürckert[4], Encarnita Mariotti-Ferrandiz[5], Lindsay G Cowell[6], Corey T Watson[7], Nishanth Marthandan[8], William J Faison[9], Uri Hershberg[10], Uri Laserson[11], Brian D Corrie[12,13], Mark M Davis[1,14], Bjoern Peters[15], Marie-Paule Lefranc[16], Jamie K Scott[8,12,17], Felix Breden[12,13], The AIRR Community[18], Eline T Luning Prak[19,22] & Steven H Kleinstein[3,20,22]

High-throughput sequencing of B and T cell receptors is routinely being applied in studies of adaptive immunity. The Adaptive Immune Receptor Repertoire (AIRR) Community was formed in 2015 to address issues in AIRR sequencing studies, including the development of reporting standards for the sharing of data sets.

VOLUME 18 NUMBER 12 DECEMBER 2017 NATURE IMMUNOLOGY

B cell receptor (BCR) and T cell receptor (TCR) are extremely diverse and personalized due to clonal evolution



By individuals
- **MHC**
- **Genotype**
- **Age**
- **Race**

By disease
- **Cancer**
- **Infectious disease**
- **Autoimmune disease**

Lavinder *et al.*, (2015) *Current Opinion in Chemical Biology*

B cell receptor (BCR) and T cell receptor (TCR) are extremely diverse

TCR

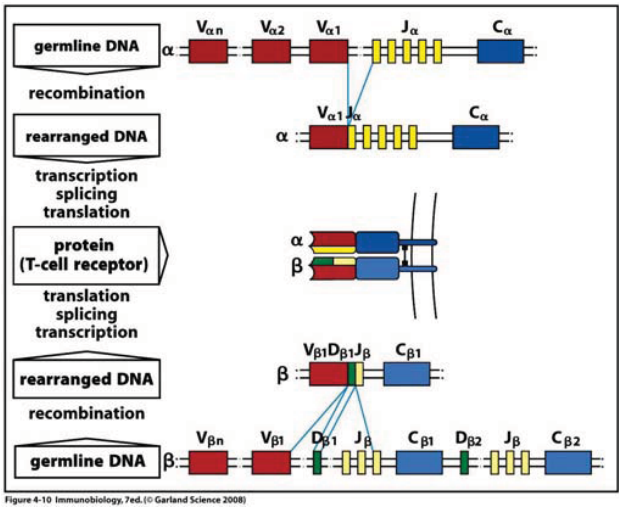


Figure 4-10 Immunobiology, 7ed. (© Garland Science 2008)

~10^{18} receptors

BCR

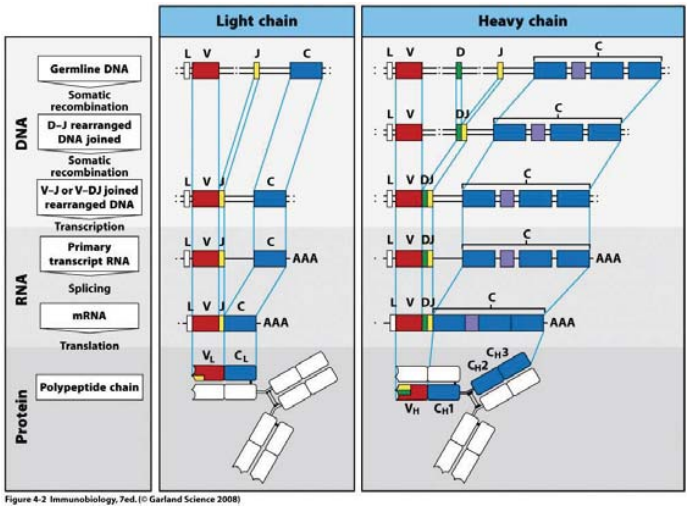
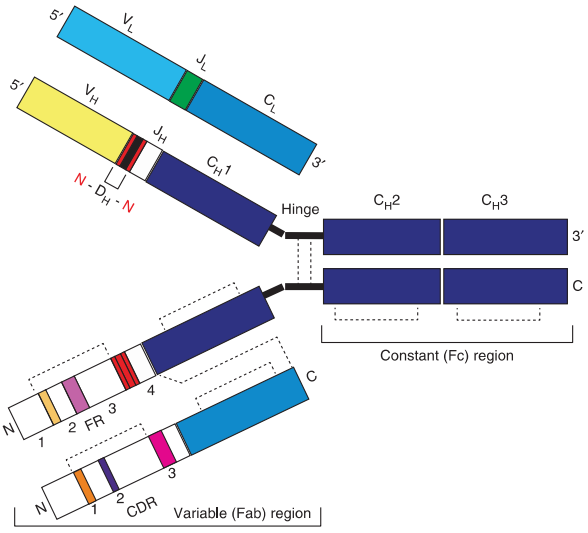


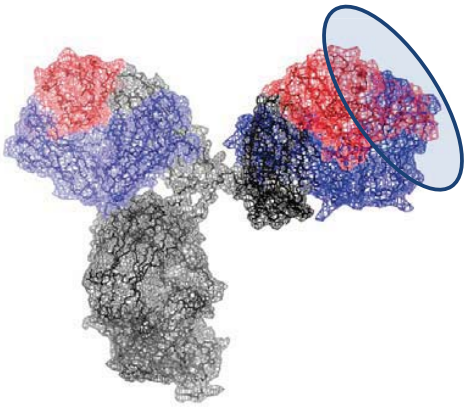
Figure 4-2 Immunobiology, 7ed. (© Garland Science 2008)

~10^{11} receptors

Complementarity-determining regions (CDRs)

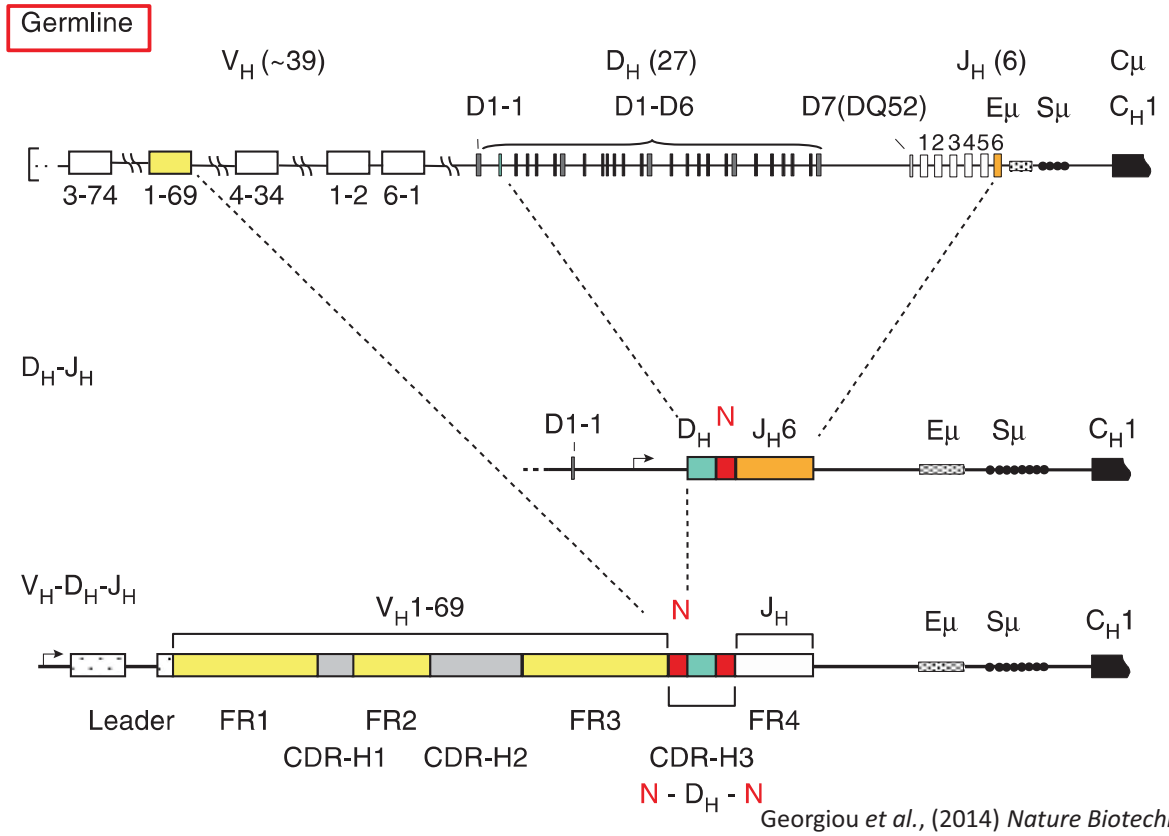


VH-VL domain: Binding to antigen



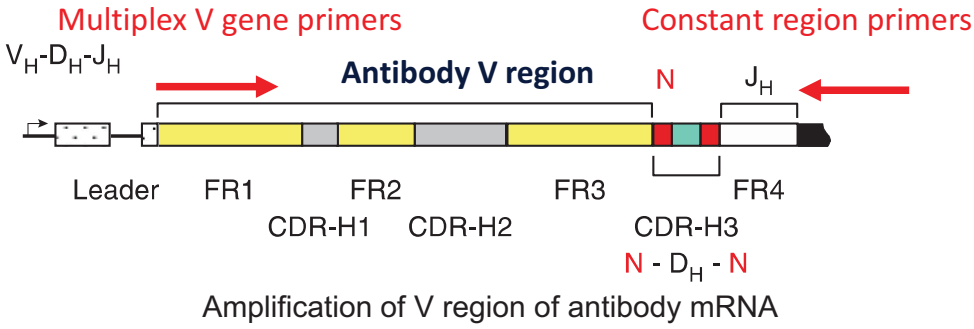
Georgiou *et al.*, (2014) *Nature Biotechnology*

Antibody transcripts are highly diverse



Georgiou *et al.*, (2014) *Nature Biotechnology*

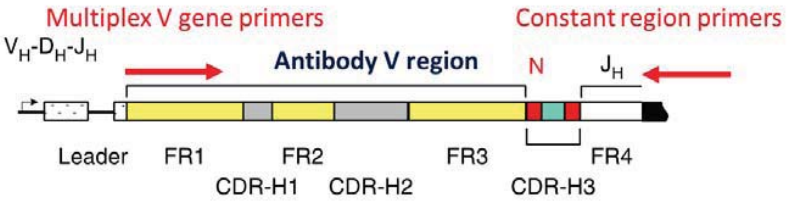
B cell Receptor sequencing (BCR-seq)



Amplification of V region of antibody mRNA

Next Generation Sequencing
MiSeq 2X300 Platform

B cell Receptor sequencing (BCR-seq)



Important points

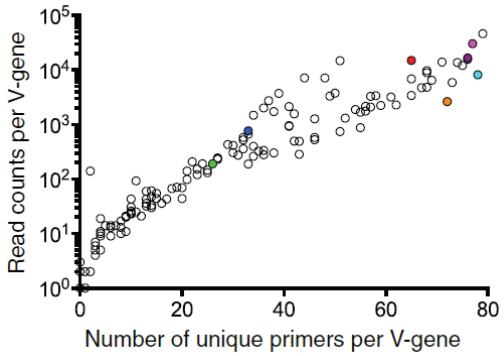
- Clone수를 많이 보는 BCR-seq은 주로 amplicon 으로 시퀀싱 한다.
- PCR을 할 때 forward primer는 주로 multiplexing primer가 사용된다.
- V region을 다 cover하려면 NGS read 길이가 길어야 한다.
- Error rate 낮은 NGS platform을 사용해야 한다 .

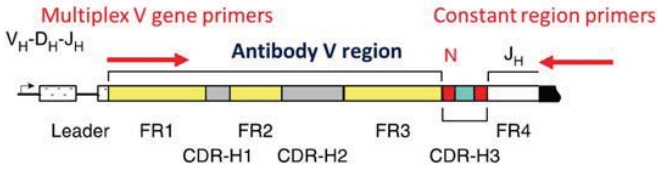
BCR-seq is erroneous

IMMUNOLOGY

Accurate and predictive antibody repertoire profiling by molecular amplification fingerprinting

Tarik A. Khan,[1] Simon Friedensohn,[1] Arthur R. Gorter de Vries,[1] Jakub Straszewski,[1,2] Hans-Joachim Ruscheweyh,[1,2,3] Sai T. Reddy[1]*

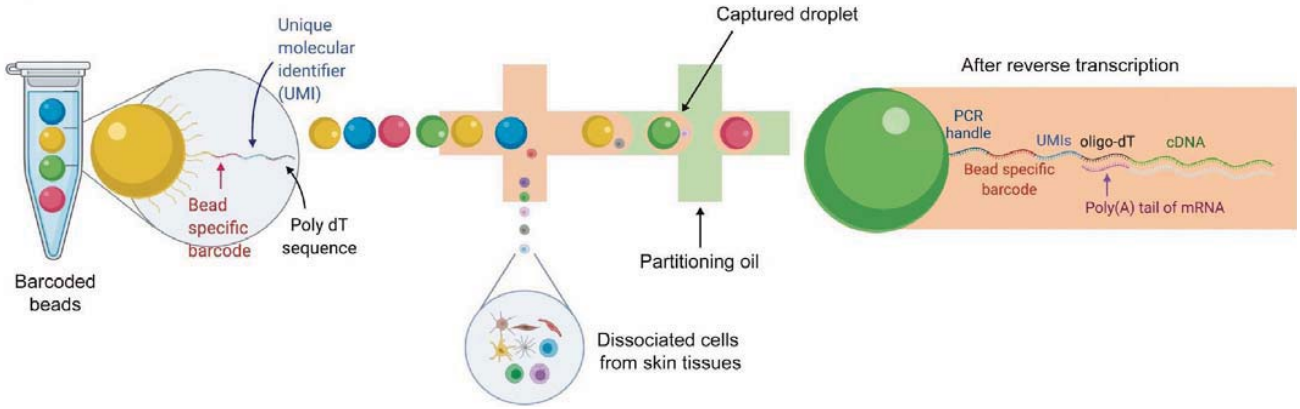




Unique Molecular Identifier를 사용하면 문제를 (일부) 해결할 수 있음

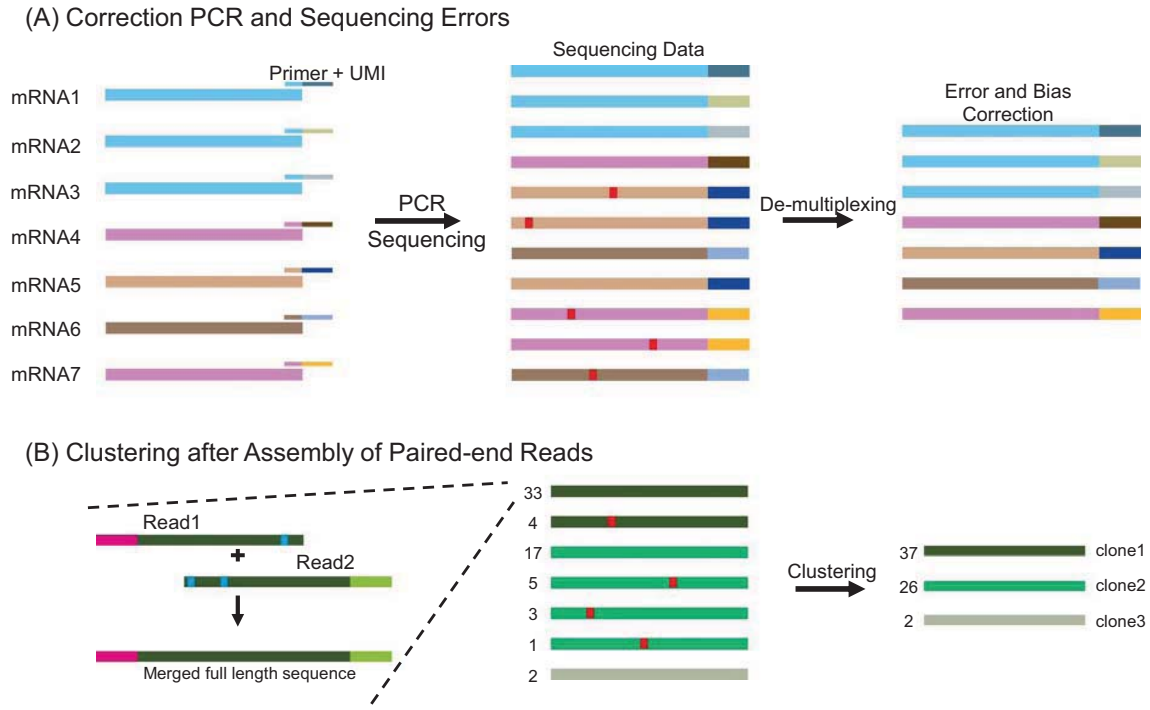
Khan *et al.*, (2016) *Science Advances*

Single cell RNA-seq의 UMI와 사용 목적이 조금 다름

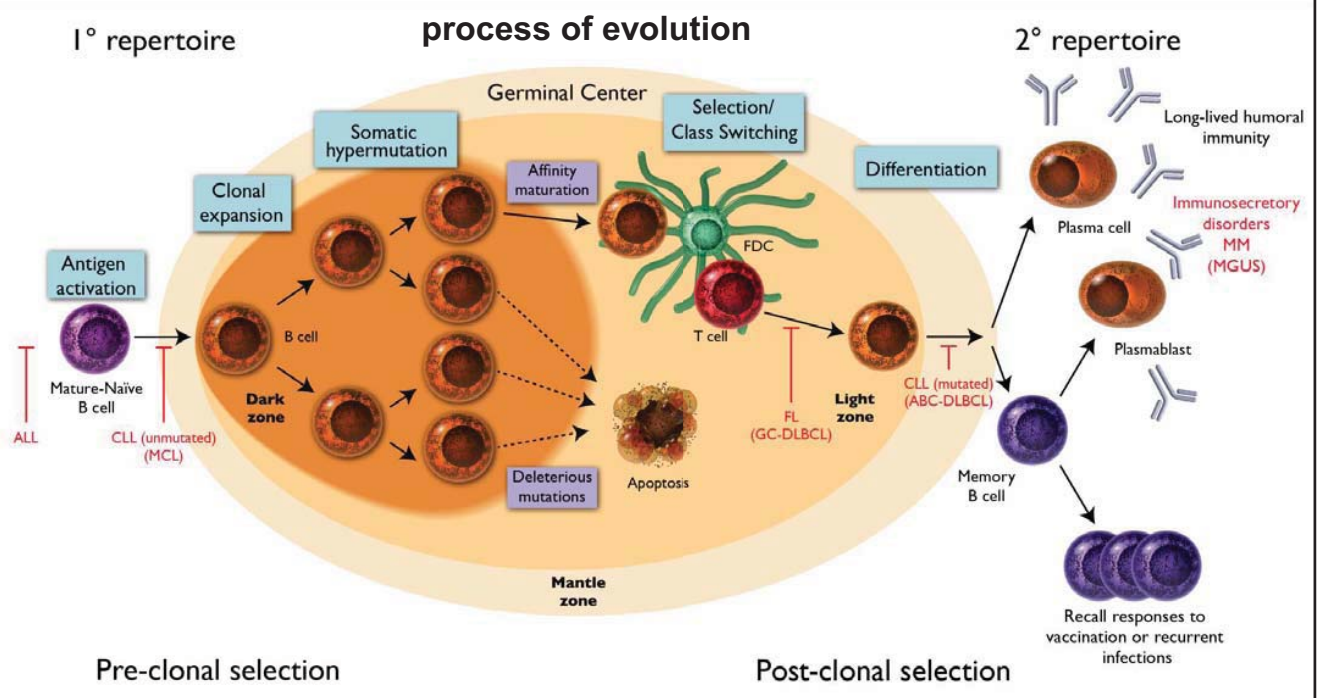


UMI counts represent gene expression levels

Error correction for detection of true diversity



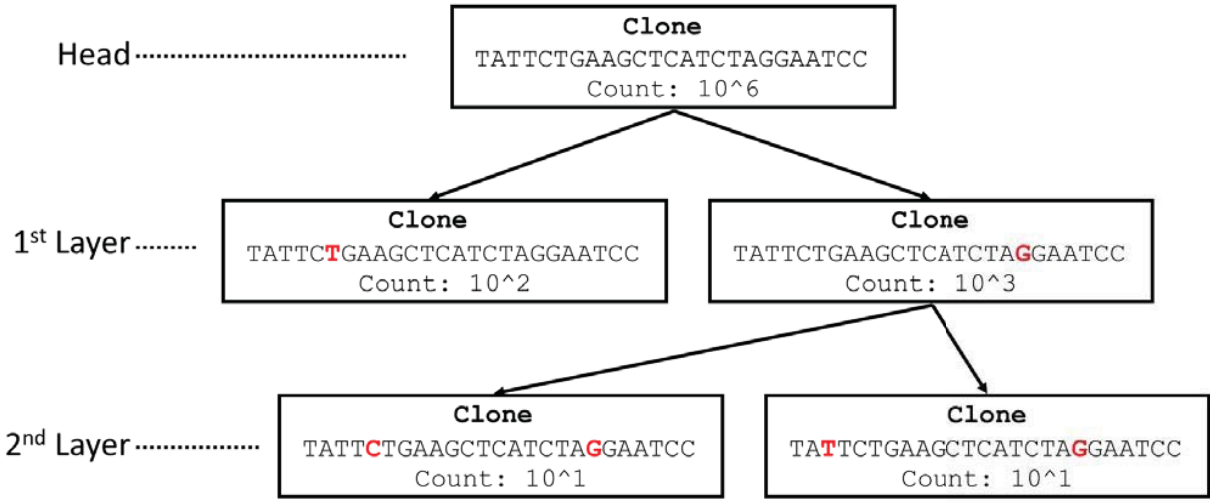
Germinal Center (GC) Reaction



ALL = Acute Lymphoblastic Leukemia
CLL = Chronic Lymphocytic Leukemia
FL = Follicular Lymphoma
DLBCL = Diffuse Large B Cell Lymphoma
MM = Multiple Myeloma

Georgiou et al., *Nature Biotechnology*

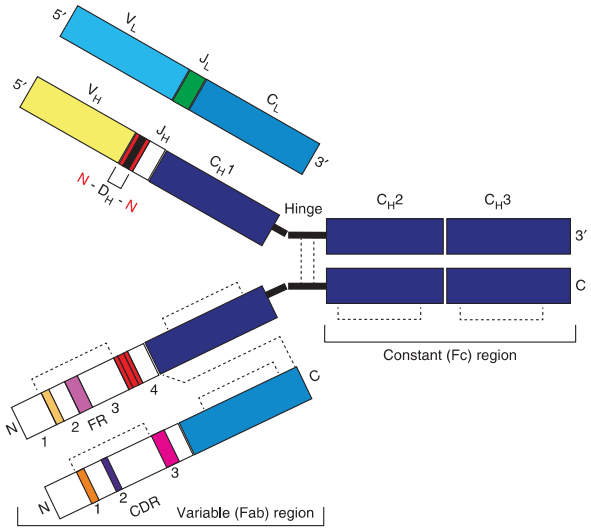
Neighbor joining algorithm to define clones



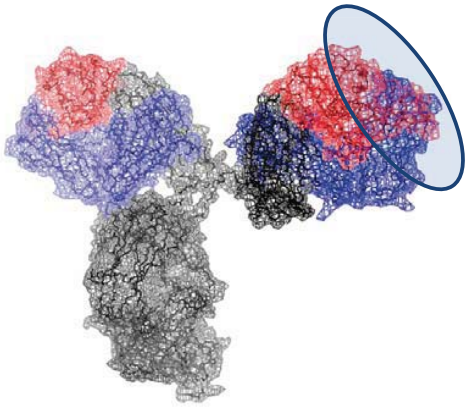
AIRR 연구를 위한 오믹스 접근법

- BCR-seq 데이터 분석에서 UMI의 사용과 DNA 서열 clustering 이 중요도
- proteomics + BCR-seq 통합 분석
- Heavy / light chain 을 동시에 시퀀싱 하기
- Short read로 assembly하기

Ig-seq is proteomics analysis of immunoglobulin

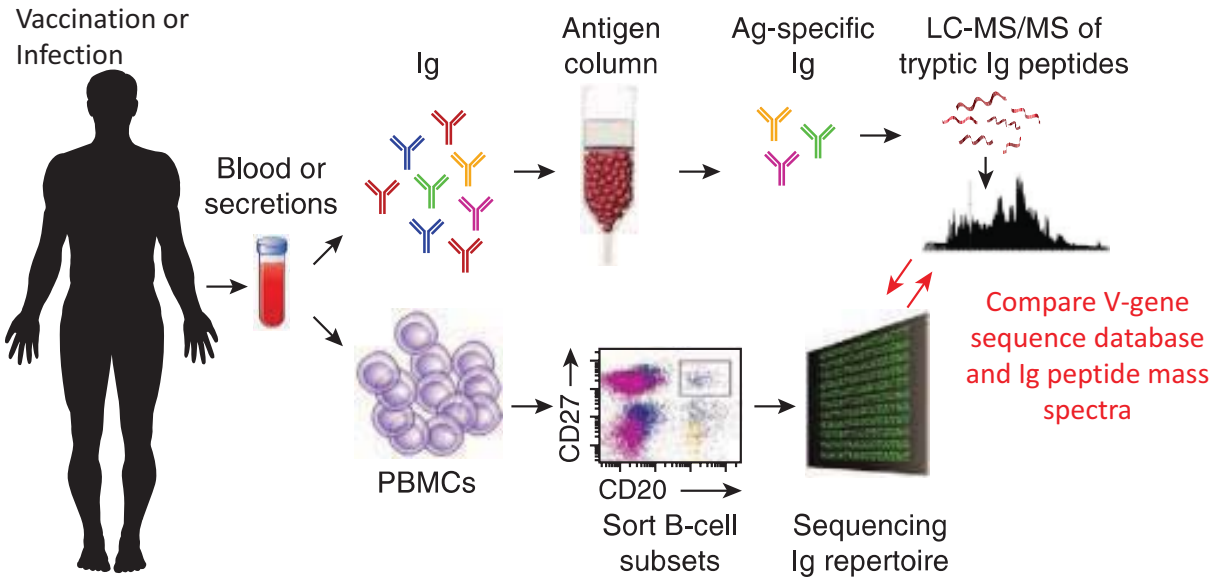


VH-VL domain: Binding to antigen



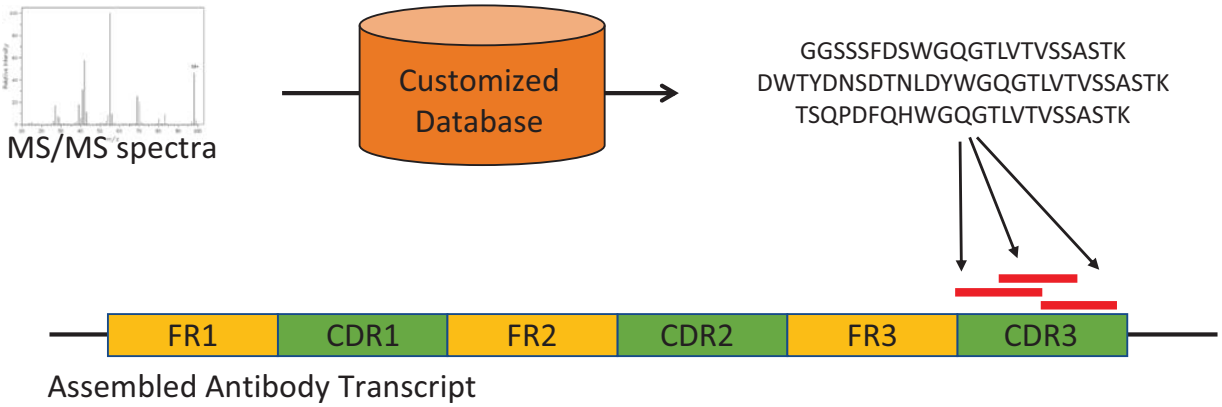
Georgiou *et al.*, (2014) *Nature Biotechnology*

Application of BCR-seq and Ig-seq



Georgiou *et al.*, (2014) *Nature Biotechnology*

Serological Antibodies Identification by CDR3 sequences in LC-MS/MS

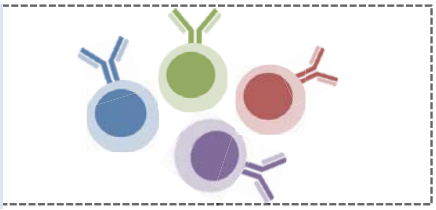


AIRR 연구를 위한 오믹스 접근법

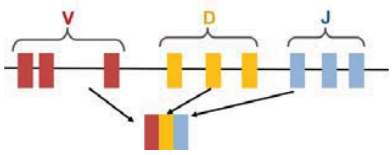
- BCR-seq 데이터 분석에서 UMI의 사용과 DNA 서열 clustering 이 중요도
- proteomics + BCR-seq 통합 분석
- Heavy / light chain 을 동시에 시퀀싱 하기
- Short read로 assembly하기

Difficulties with antibody sequencing

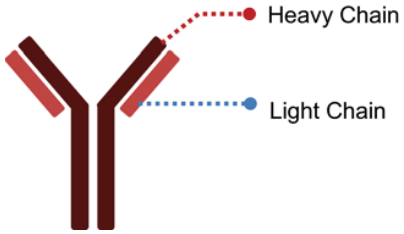
One B cell expresses one antibody

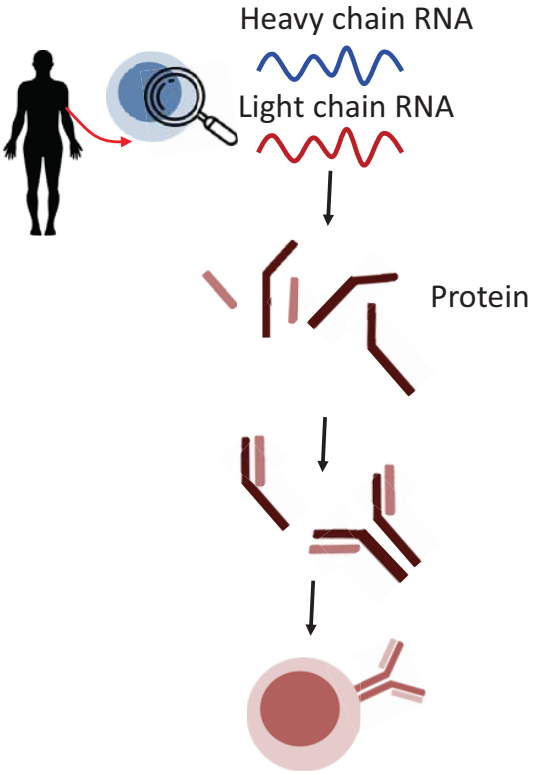


Diversity-VDJ recombination

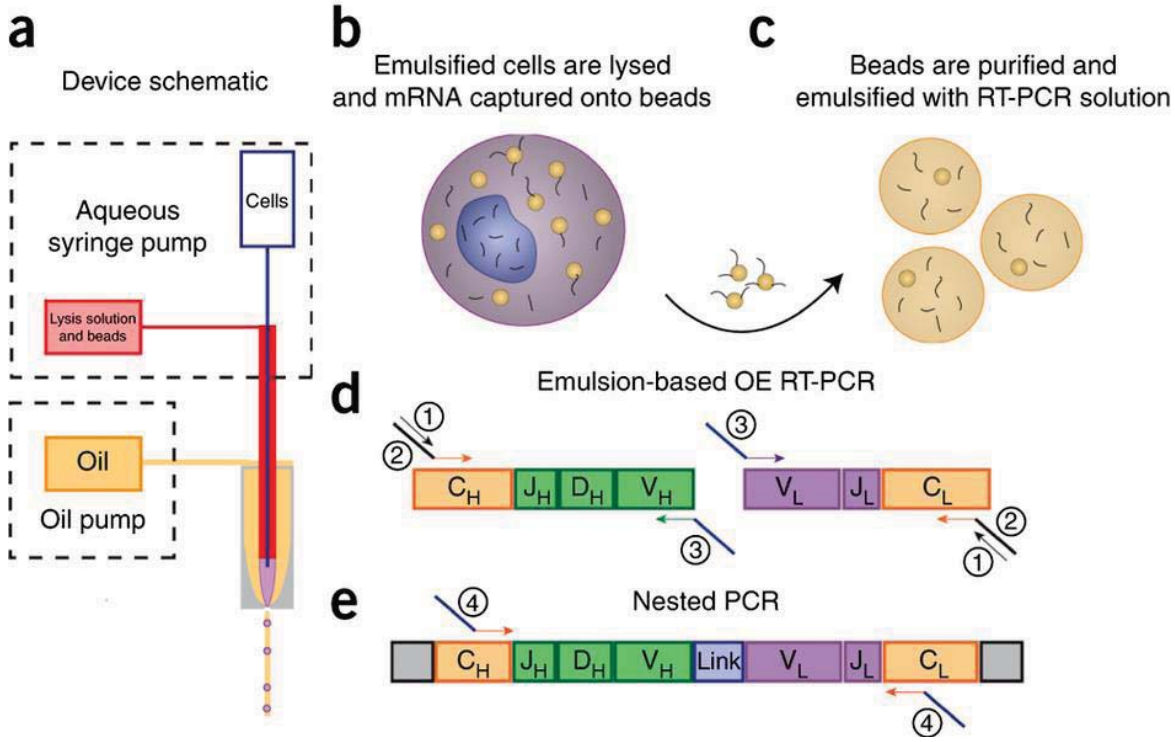


Heavy chain and Light chain are paired at protein level





Pairing sequencing technology

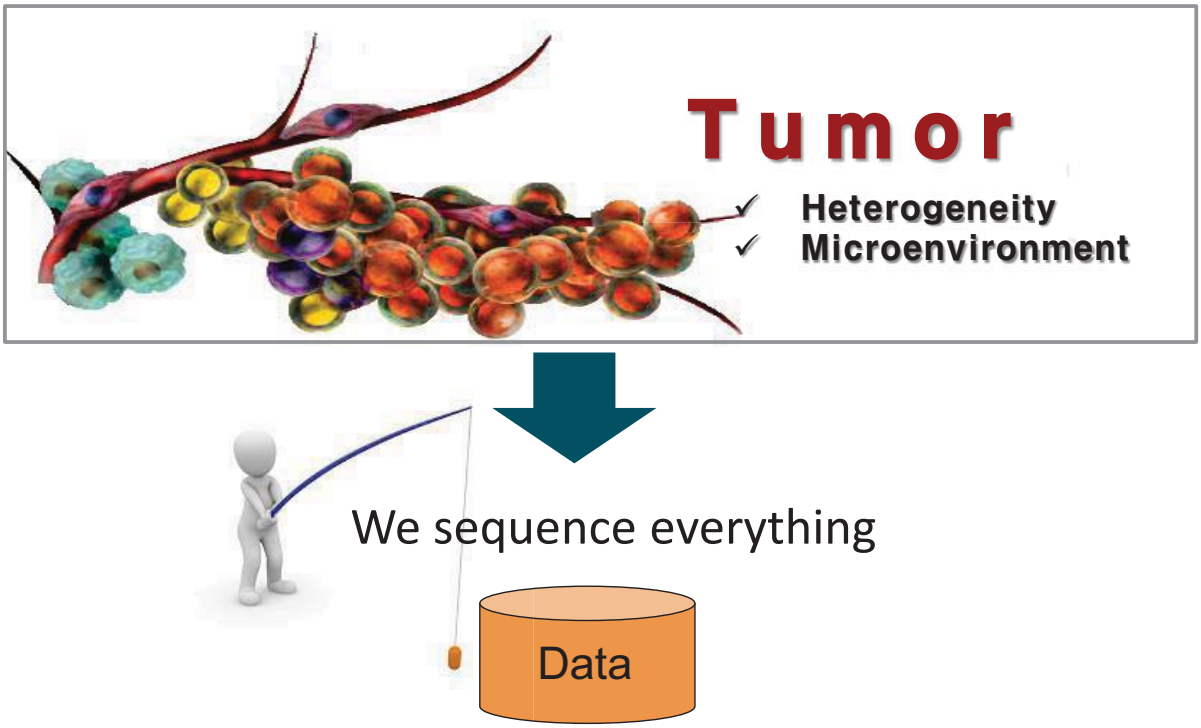


DeKosky *et al.*, (2015) *Nature Medicine*
McDaniel *et al.*, (2016) *Nature Protocols*

AIRR 연구를 위한 오믹스 접근법

- BCR-seq 데이터 분석에서 UMI의 사용과 DNA 서열 clustering 이 중요도
- proteomics + BCR-seq 통합 분석
- Heavy / light chain 을 동시에 시퀀싱 하기
- Short read로 assembly하기

Tumor infiltrating lymphocytes from mRNA-seq



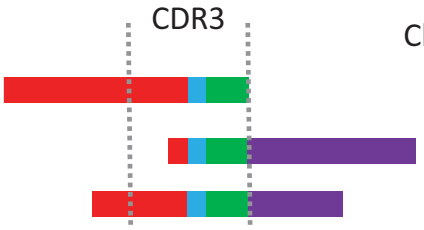
Reconstruction of Antibody from mRNA-seq

V gene
D gene
J gene
C gene

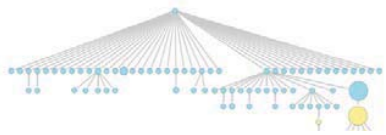
Mapping onto IMGT DB

Assembly of Reads
(de Bruijn graph)

관련 분석 프로그램
- MIXCR
- TRUST



Clonotyping



IgM
IgG
IgA

Choice for TIL repertoire: mRNA-seq and BCR-seq

	mRNA-seq	BCR-seq
	Transcriptome	**Antibody Transcripts**
Example	Gene Expression Profile	Antibody Repertoire
Platform	Mostly, short reads 2X100~150	Mostly, MiSeq 2X300
Antibody Sequences	Assembly	*de novo* after correction

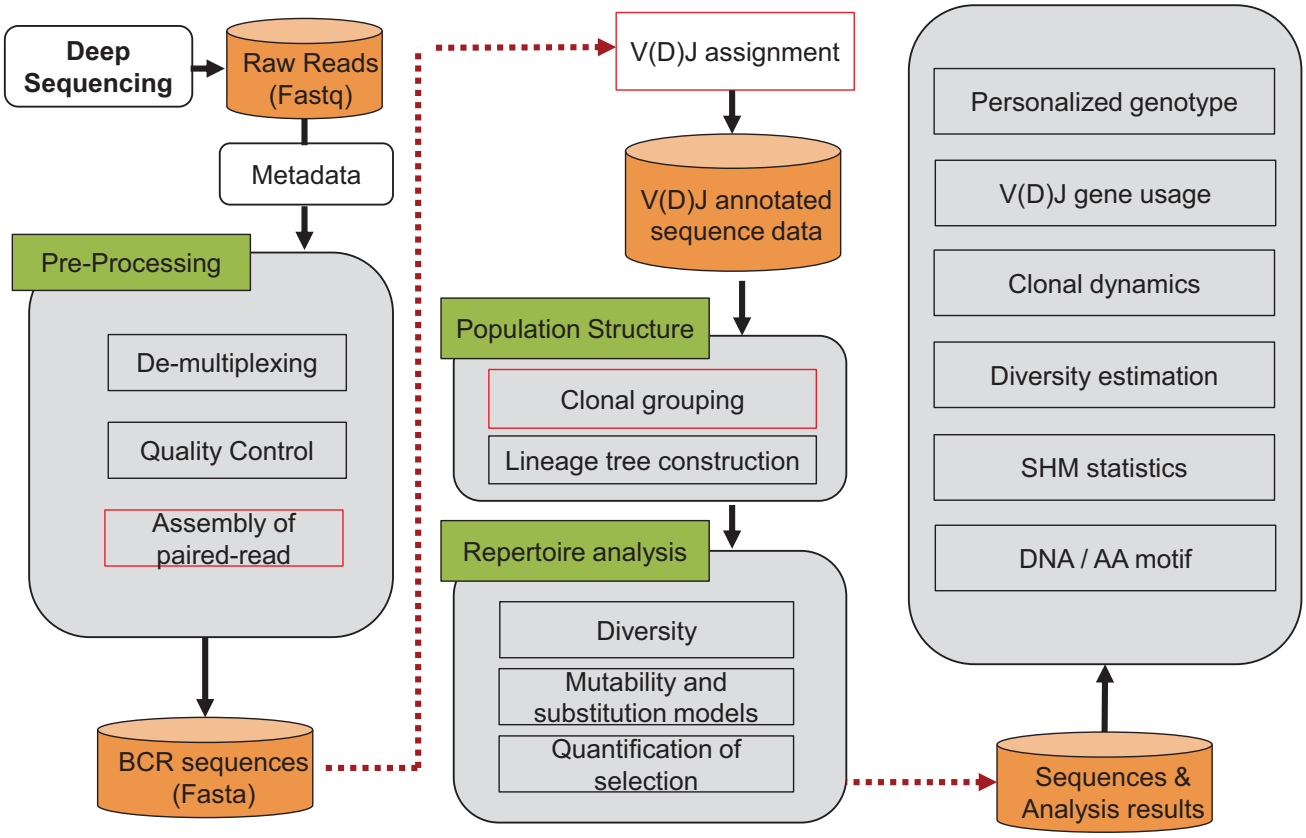
1 Adaptive Immune Receptor Repertoire (AIRR) 의 정의와 다양성

2 AIRR 연구를 위한 오믹스 접근법

3 **AIRR 생명정보 분석법**

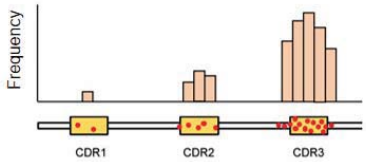
4 AIRR 연구 사례 및 분석 예시

Bioinformatics Pipeline for BCR-seq



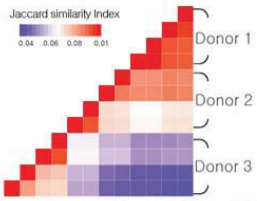
Examples of BCR-seq downstream analysis

(A) Quantification of Somatic Hyper Mutation



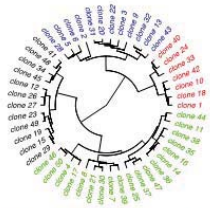
IgBlast, MIXCR, IMGT, SHazaM

(B) Pairwise Overlap Analysis



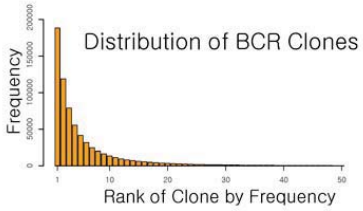
Jaccard index, Morisita Horn index

(C) Lineage Construction



PHYLIP, Alakazam, IgTree

(D) Statistics and Distribution

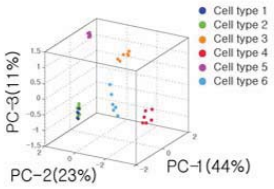


CDR3 length/charge, V gene usage

(E) Ecological Diversity

Shannon Index, Simpson Index

(F) Dimension Reduction



PCA, tSNE

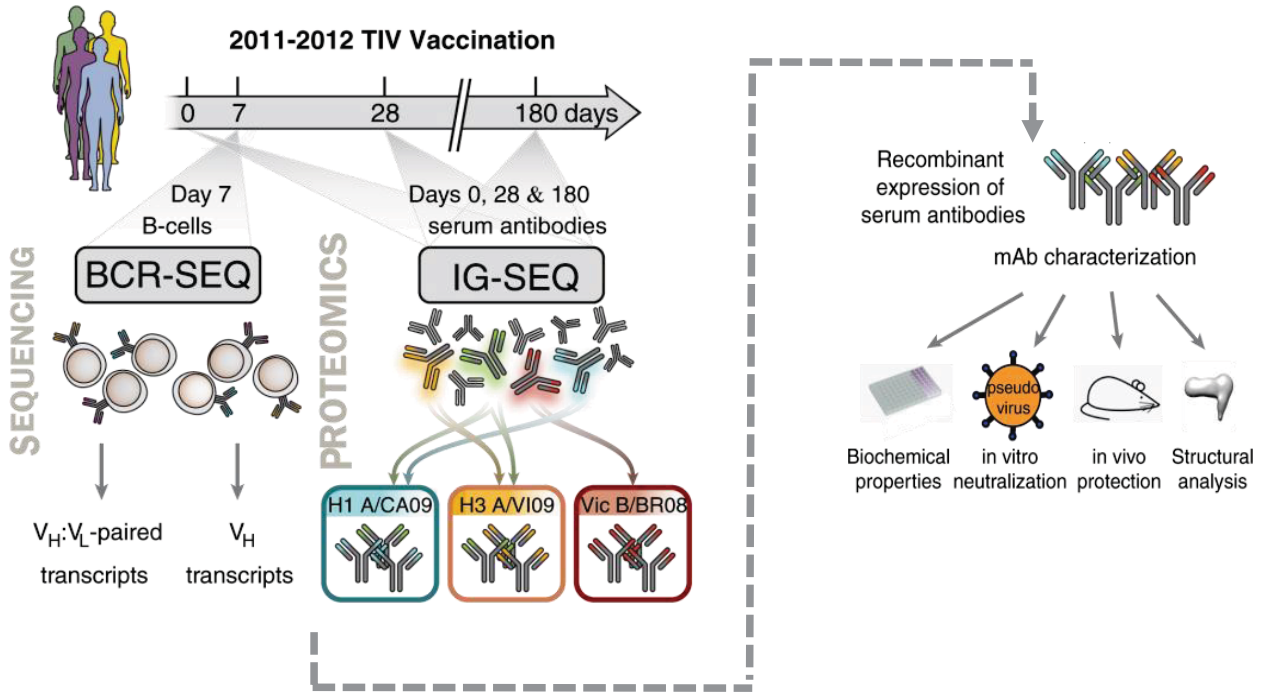
1 Adaptive Immune Receptor Repertoire (AIRR) 의 정의와 다양성

2 AIRR 연구를 위한 오믹스 접근법

3 AIRR 생명정보 분석법

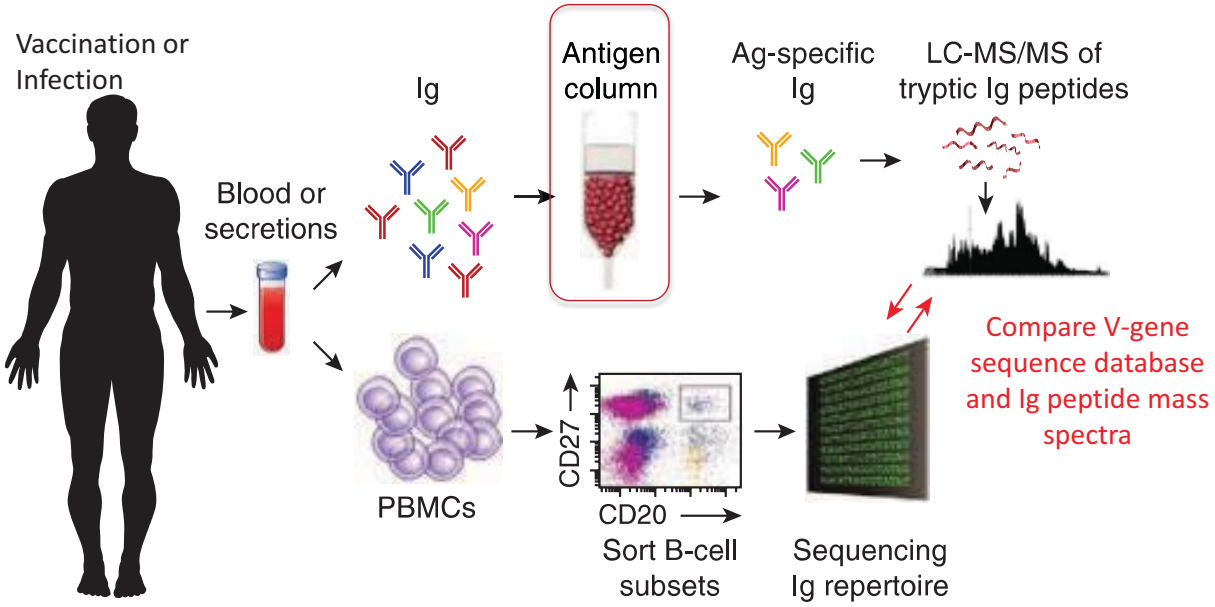
4 **AIRR 연구 사례 및 분석 예시**

Example of Immunoproteogenomics: Flu Repertoire



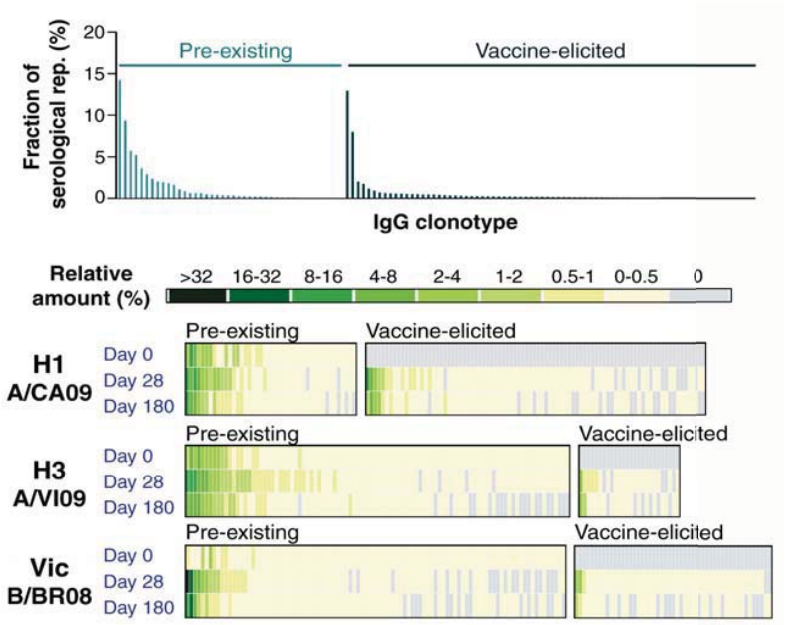
Georgiou and colleagues (2016) *Nature Medicine*

Application of BCR-seq and Ig-seq



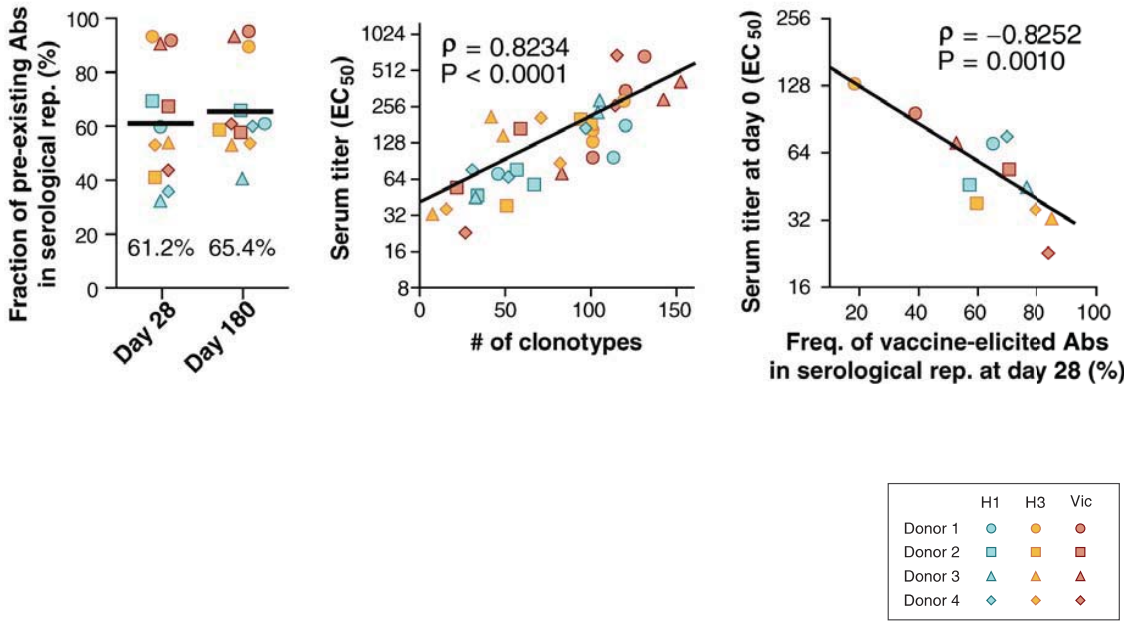
Georgiou *et al.*, (2014) *Nature Biotechnology*

Delineation of the serological repertoire



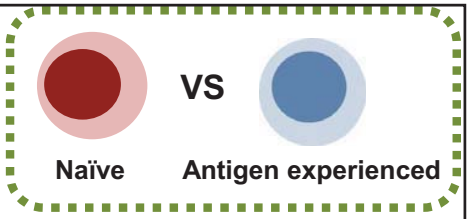
Georgiou and colleagues (2016) *Nature Medicine*

Delineation of the serological repertoire



Georgiou and colleagues (2016) *Nature Medicine*

VH:VL Pairing Experiments

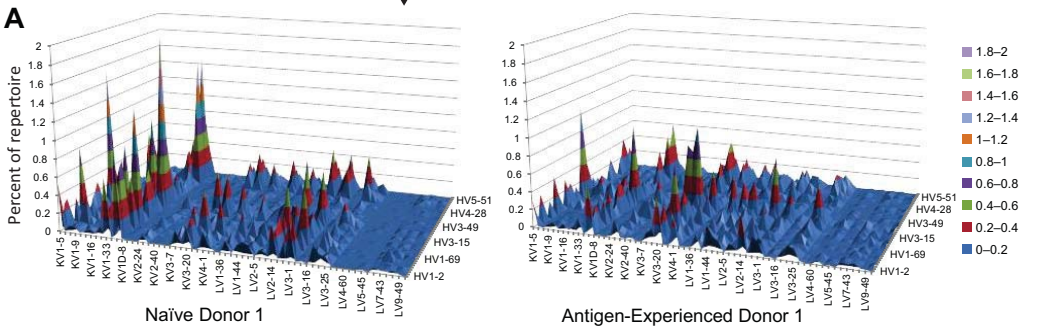


70 ml Blood Draw

BCR pairing sequencing ↓

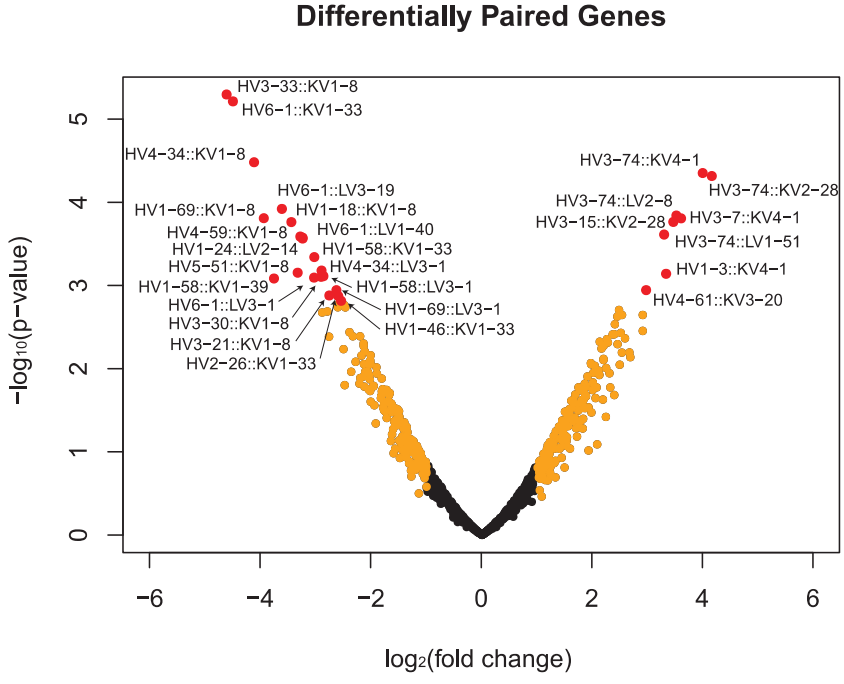
< Unique Antibody Sequences >

Donor	$CD3^-$ $CD19^+$ $CD20^+$ $CD27^-$ Naïve	$CD3^-$ $CD19^+$ $CD20^+$ $CD27^+$ Ag-Exp
1	13,780	34,692
2	26,372	89,249
3	15,203	NA
Total	55,355	123,941



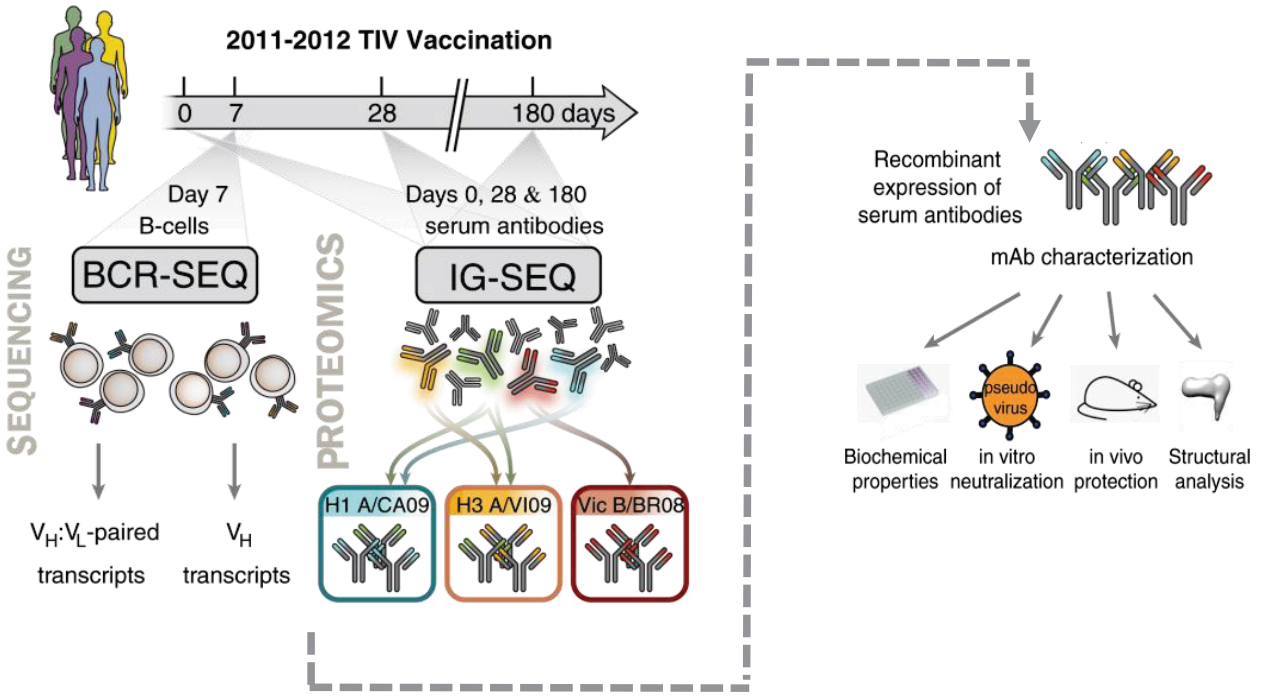
DeKosky*, Oana*, Park *et al.*, (2016) *PNAS*

Identification of differentially paired VH:VL genes



DeKosky*, Oana*, Park *et al.*, (2016) *PNAS*

Example of Immunoproteogenomics: Flu Repertoire



Georgiou and colleagues (2017) *Nature Medicine*

Summary

1 **Adaptive Immune Receptor Repertoire (AIRR) 의 정의와 다양성**

- VH의 구조

2 **AIRR 연구를 위한 오믹스 접근법**

- BCR-seq 데이터 분석에서 UMI의 사용과 DNA 서열 clustering 의 중요도
- proteomics + BCR-seq 통합 분석
- Heavy / light chain 을 동시에 시퀀싱 하기
- Short read로 assembly하기

3 **AIRR 생명정보 분석법**

4 **AIRR 연구 사례 및 분석 예시**

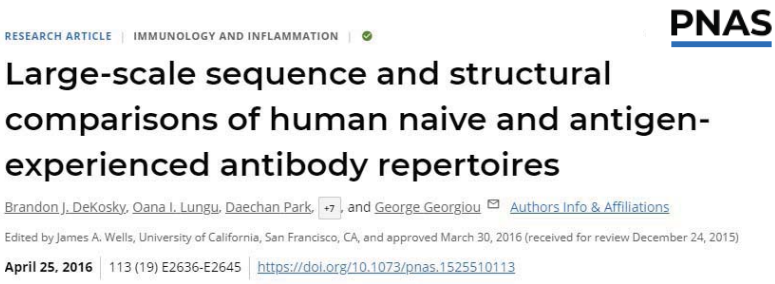
1 Adaptive Immune Receptor Repertoire (AIRR) 의 정의와 다양성

2 AIRR 연구를 위한 오믹스 접근법

3 AIRR 생명정보 분석법

4 AIRR 연구 사례 및 **분석 예시**

실습순서


RESEARCH ARTICLE | IMMUNOLOGY AND INFLAMMATION

PNAS

Large-scale sequence and structural comparisons of human naive and antigen-experienced antibody repertoires

Brandon J. DeKosky, Oana I. Lungu, Daechan Park, +7, and George Georgiou Authors Info & Affiliations

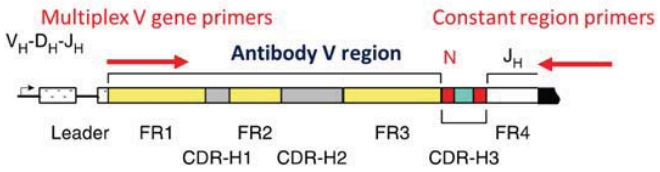
Edited by James A. Wells, University of California, San Francisco, CA, and approved March 30, 2016 (received for review December 24, 2015)

April 25, 2016 | 113 (19) E2636-E2645 | https://doi.org/10.1073/pnas.1525510113

1. 데이터 다운로드하기
   - Sratoolkit
2. 데이터 확인하기
   - 프라이머 서열 확인: 리눅스 grep & awk 명령어
   - 항체 서열인지 확인해보기: UCSC genome browser & IMGT
3. 분석하기
   - R1 and R2 이어 붙이기
   - MIXCR 프로그램 이용해서 V gene과 CDR3 서열 찾기

https://www.theserverside.com/blog/Coffee-Talk-Java-News-Stories-and-Opinions/How-do-I-install-Java-on-Ubuntu

2x250 or 2x300 technology. Forward primers targeted the antibody Framework 1 regions (3); reverse primers targeted the IgM/Igκ/Igλ constant region for $CD27^-$ NBCs, and IgM/IgG/IgA/Igκ/Igλ reverse primers were used for $CD27^+$ AEBCs. Full length VH and VL genes were generated for antigen-experienced repertoires via bioinformatic assembly of three Illumina sequencing samples (VH:VL, VH only, and VL only) as described previously (2–4). The following barcoded primers were used for VH-only amplification and sequencing (barcodes are italicized): Donor 1 Replicate 1 5'-NNNN *TGAAGG* GGCTAGCTATTCCCATCGCGG-3', Donor 1 Replicate 2 5'-NNNN *CGCGTC* GGCTAGCTATTCCCATCGCGG-3', Donor 2 Replicate 1 5'-NNNN *TAAGAA* GGCTAGCTATTCCCATCGCGG-3', Donor 2 Replicate 2 5'-NNNN *AGCGAG* GGCTAGCTATTCCCATCGCGG-3'. The following barcoded primers were used for VL-only amplification and sequencing (barcodes are italicized): Donor 1 Replicate 1 5'-NNNN *TGAAGG* GCGCCGCGATGGGAAT-3', Donor 1 Replicate 2 5'-NNNN *CGCGTC* GCGCCGCGATGGGAAT-3', Donor 2 Replicate 1 5'-NNNN *TAAGAA* GCGCCGCGATGGGAAT-3', Donor 2 Replicate 2 5'-NNNN *AGCGAG* GCGCCGCGATGGGAAT-3'.



R= A or G

	Primer Sequence (enter letters only, 5' to 3')
hVH1	tattcccatcgcggcgc CAGGTCCAGCTKGTRCAGTCTGG
hVH157	tattcccatcgcggcgc CAGGTGCAGCTGGTGSARTCTGG
hVH2	tattcccatcgcggcgc CAGRTCACCTTGAAGGAGTCTG
hVH3	tattcccatcgcggcgc GAGGTGCAGCTGKTGGAGWCY
hVH4	tattcccatcgcggcgc CAGGTGCAGCTGCAGGAGTCSG
hVH6	tattcccatcgcggcgc CAGGTACAGCTGCAGCAGTCA
hVH3N	tattcccatcgcggcgc TCAACACAACGGTTCCCAGTTA

hIgG_MiSqRev1_4N	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG NNNN ATGGGCCCTG SGATGGGCCCTTGGTGGARGC
hIgM_MiSqRev1_4N	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG NNNN ATGGGCCCTG GGTTGGGGCGGATGCACTCC
hIgA_MiSqRev1_4N	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG NNNN ATGGGCCCTG CTTGGGGCTGGTCGGGGATG