

Estimating immune diversity: a practical session

Mikhail Shugay,

Genomics of Adaptive Immunity Lab, CEITEC



INSTITUTO
GULBENKIAN
DE CIÊNCIA

molecul|ar
systems
biology



EMBOpress



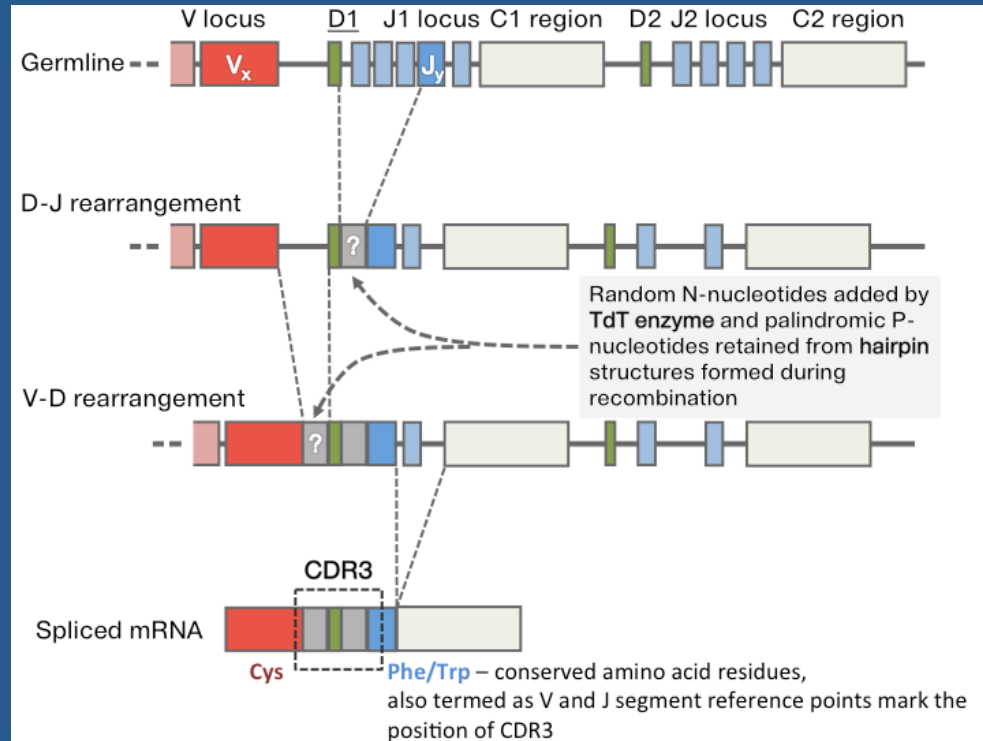
CEITEC

Central European Institute of Technology
BRNO | CZECH REPUBLIC

Challenges in estimating IR diversity

Theoretical diversity of immune receptors is extremely high

- 10^8 - 10^{10} clonotypes, single chain V-(D)-J variants



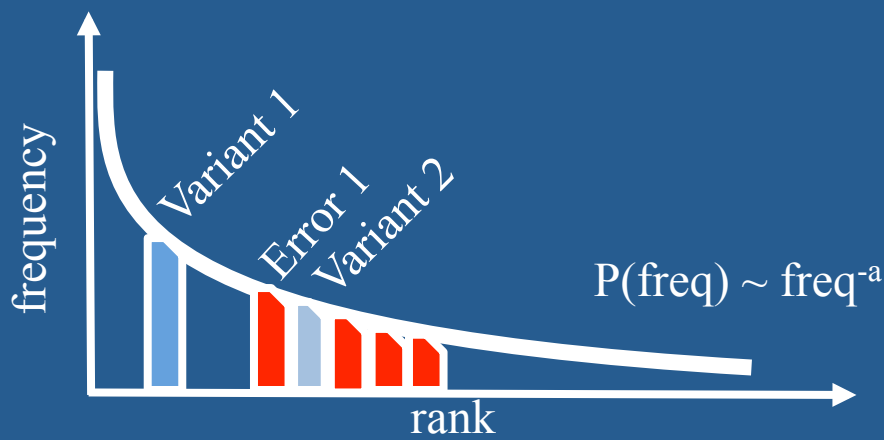
- 10^{16} - 10^{18} clones, heterodimers TRA/B, etc
- Further increased by somatic hypermutations for BCR
- Typical depth of profiling is $\sim 10^6$ cells per individual

Challenges in estimating IR diversity

- Hard to distinguish convergent recombination from errors
 - Random inserts and segment truncations

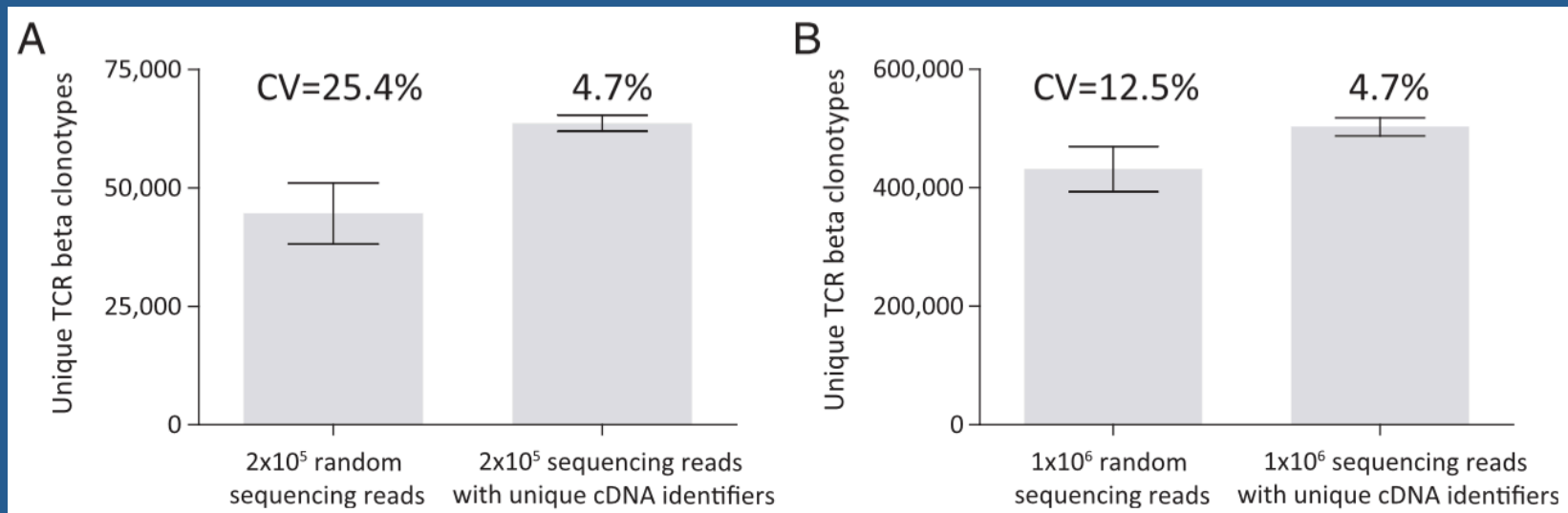
CDR3AA	V	D	J	CDR3NT
CASSLAPGATNEKLFF	TRBV7-6	TRBD2	TRBJ1-4	TGTGCCAGCAGCTTAGCGCCGGGAGCAACTAATGAAAACTGTTTTTT
CASSLAPGATNEKLFF	TRBV7-6	TRBD2	TRBJ1-4	TGTGCCAGCAGCTTAGCCCCGGGGCAACTAATGAAAACTGTTTTTT
CASSLAPGATNEKLFF	TRBV7-6	TRBD1	TRBJ1-4	TGTGCCAGCAGCTTAGCGCCTGGAGCAACTAATGAAAACTGTTTTTT

- Errors in abundant clonotypes can have frequency comparable with real clonotypes
 - Clonotype frequency is distributed according to power law



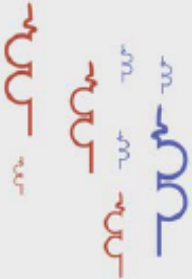
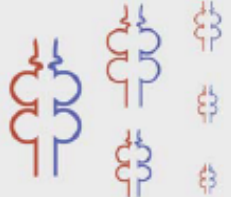


Challenges in estimating IR diversity

- Sample size is hard to define for HTS data: have we sequenced 1mln T-cells with 1x or 100k with 10x?
 - Clonotype represented by 10 reads can be either 10x-sequenced or present in 10 copies
 - Amplicon library – no read offsets, etc
- With our protocol a single UMI tag roughly corresponds to a single cDNA molecule and the protocol yield is ~ 0.5 cDNA molecules per immune receptor transcript



Practical importance

Task	Application	Conventional data analysis	Molecular barcoding approach
Repertoire diversity estimation	Emerging biomarker in cancer studies	 <p>PCR bias, stochastic sampling and artificial diversity resulting from sequencing errors</p>	 <p>Robust diversity estimates computed from corrected data</p>
TCRa- β pairing	Designing TCRs for adoptive T-cell transfer therapy	 <p>Noisy frequencies of TCR chains originating from the same clonotype</p>	 <p>Frequency-based pairing possible for normalized data</p>
TCR pattern analysis	De-novo discovery of tumor-specific TCRs	<p>CASS LVAGTV TEAFF CASS LVAGTV TEAFF CASS LVAGGV TEAFF CASS LIAGTA TEAFF CASS LIAGTG TEAFF</p> <p>Artificial TCR variants, frequency-based error correction not applicable</p>	<p>CASS LVAGTV TEAFF CASS LVAGTV TEAFF CASS LVAGTV TEAFF CASS LIAGTG TEAFF CASS LIAGTG TEAFF</p> <p>Accurate TCR pattern analysis in highly convergent populations</p>

Software for RepSeq data processing

22 software tools currently listed in RepSeq section of OMICtools

– <http://omictools.com/rep-seq-c424-p1.html>

IGBLAST



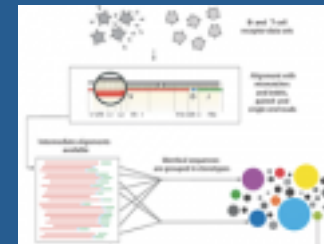
MIGEC



MITCR



MIXCR



VDJtools



Software for RepSeq data processing

IGBLAST



- V-(D)-J mapping
- “Gold standard”
- Full-length*

MITCR



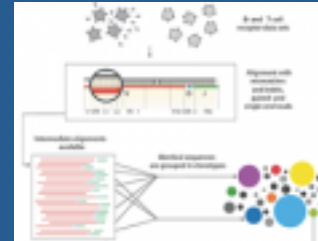
- Extremely fast
- V-(D)-J mapping
- Clonotype assembly
- Error correction (freq-based)
- TCR only

MIGEC



- Pre-processing (including UMIs)
- V-(D)-J mapping
- Clonotype assembly**
- Error correction

MIXCR



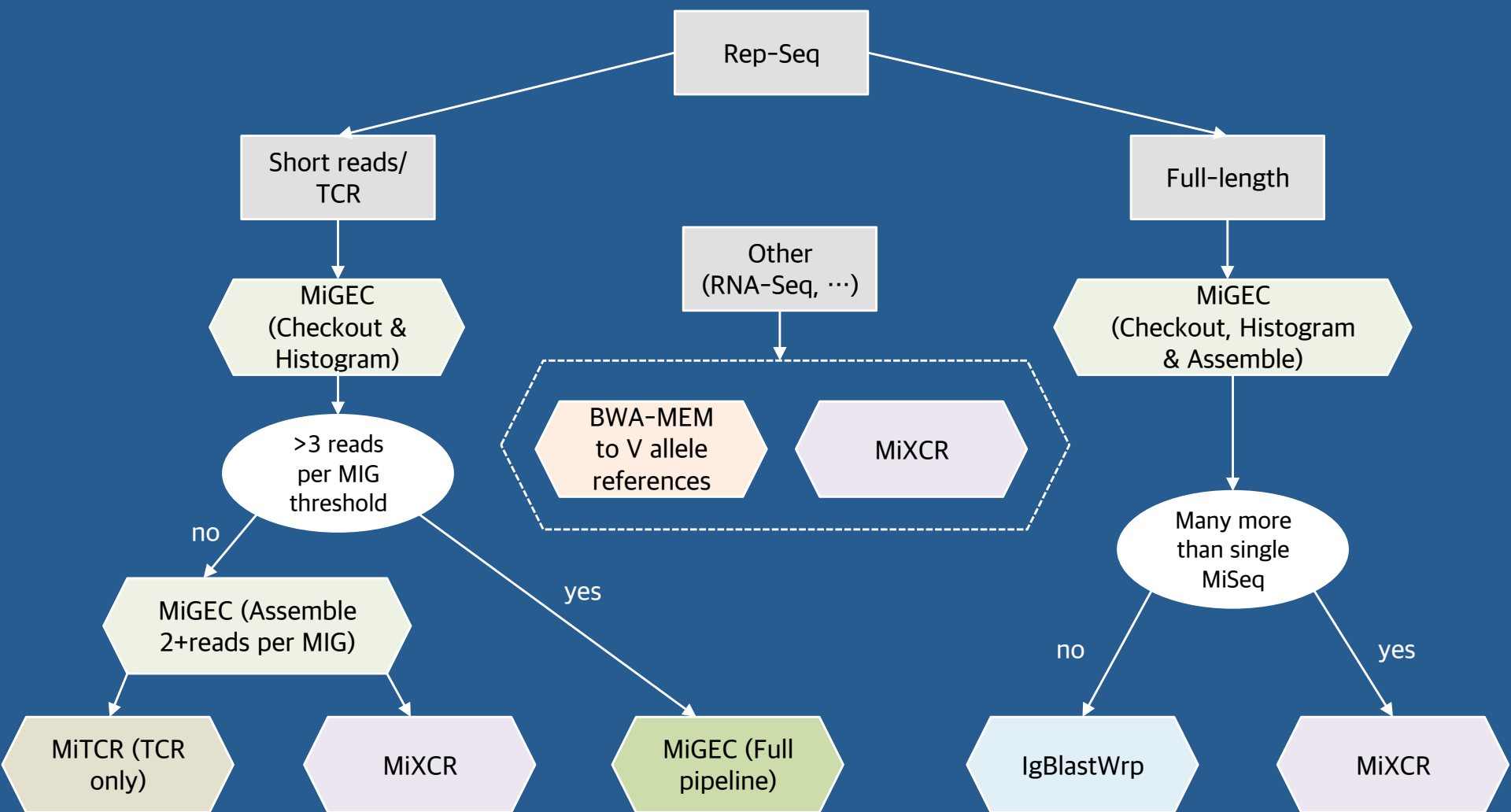
- MITCR extended for B-cells
- V-(D)-J + 5'UTR, isotype, ...
- More robust algorithms
- Powerful API
- Full-length

* Full-length mapping maps whole Variable segment and extract somatic hypermutations (SHMs)/alleles,

Most algorithms simply identify V and J gene and only extract CDR3 sequence

** Assembling V-(D)-J mapping results into a V-CDR3-J (+SHM) table, correcting errors and summarizing counts

A typical analysis pipeline



Software for RepSeq post-analysis

RepSeq processing



Raw data

MiTCR, IgBlast,
MiGEC, IMGT, ImmunoSEQ



Clonotypes

Post-analysis

VDJtools



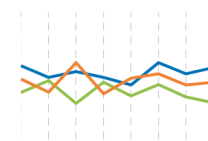
Basic statistics and segment usage



Repertoire overlap



Diversity analysis



Data joining and clonotype tracking



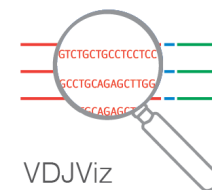
Repertoire clustering



Clonotype filtering and annotation



Flexible API



VDJViz