EMBO Practical Course

Estimating immune diversity: a practical session

Mikhail Shugay,

Genomics of Adaptive Immunity Lab, CEITEC

INSTITUTO GULBENKIAN DE CIÊNCIA

molecular systems biology

EMBOpress

CEITEC

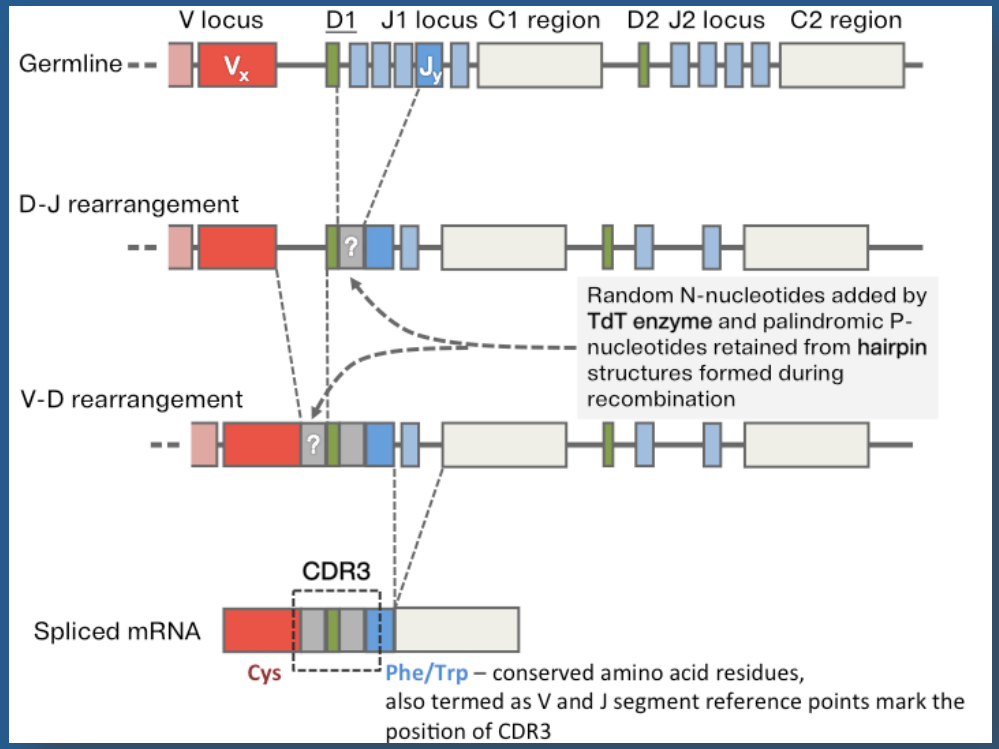
Central European Institute of Technology

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



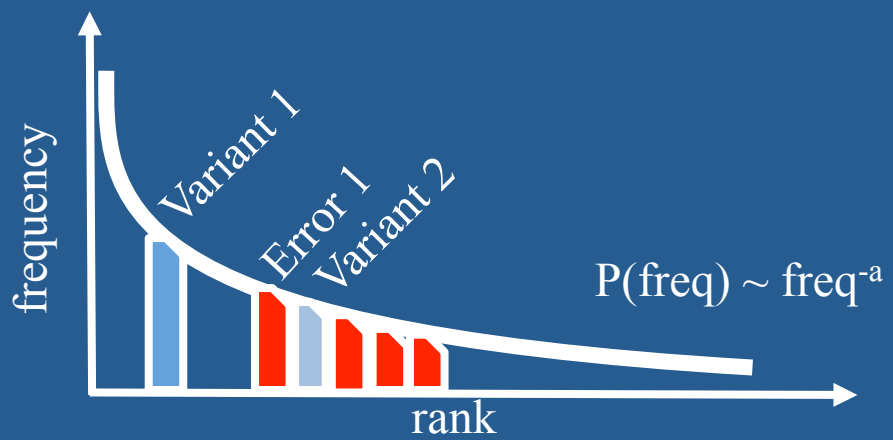
- 10^{16}-10^{18} clones, heterodimers TRA/B, etc
- Further increased by somatic hypermutations for BCR
- Typical depth of profiling is ~10^6 cells per individual

Challenges in estimating IR diversity

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

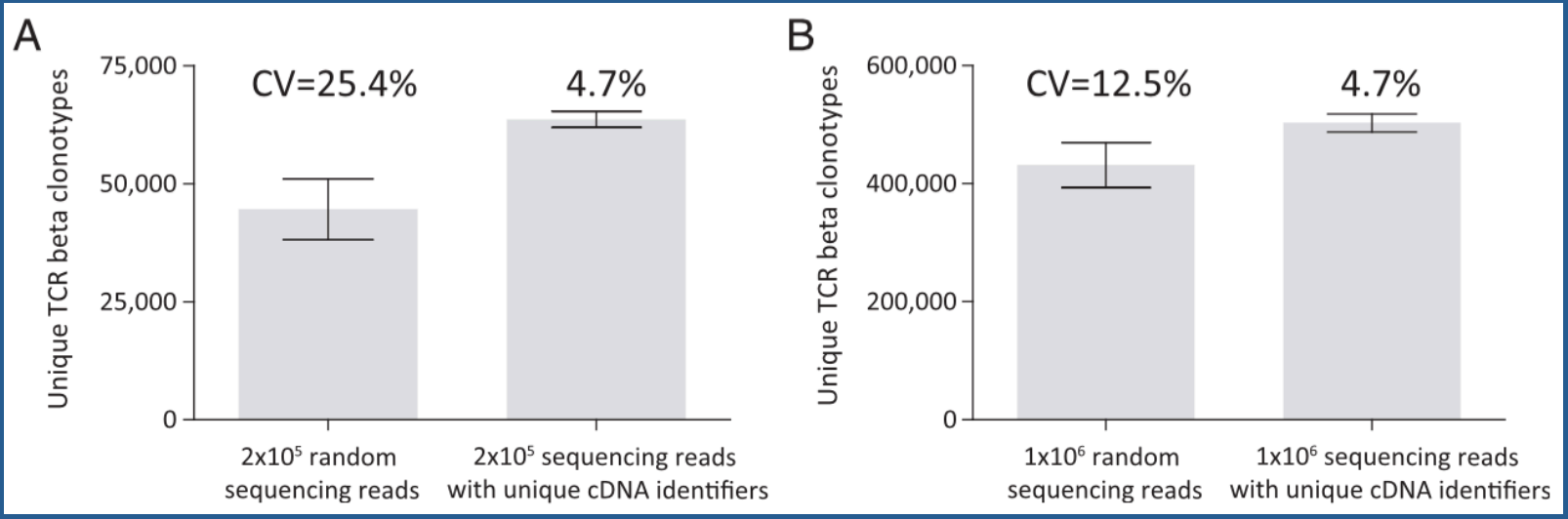
CDR3AA	V	D	J	CDR3NT
CASSLAPGATNEKLFF	TRBV7-6	TRBD2	TRBJ1-4	TGTGCCAGCAGCTTAGCGCCGGGAGCAACTAATGAAAAACTGTTTTTT
CASSLAPGATNEKLFF	TRBV7-6	TRBD2	TRBJ1-4	TGTGCCAGCAGCTTAGCCCCCGGGGCAACTAATGAAAAACTGTTTTTT
CASSLAPGATNEKLFF	TRBV7-6	TRBD1	TRBJ1-4	TGTGCCAGCAGCTTAGCGCCTGGAGCAACTAATGAAAAACTGTTTTTT

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



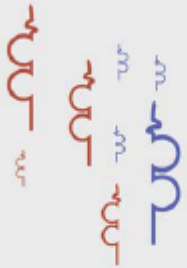
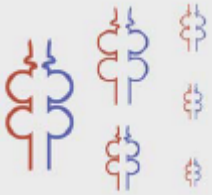
Challenges in estimating IR diversity

- Sample size is hard to define for HTS data: have we sequenced 1mln T-cells with 1x or 100k with 10x?
  - Clonotype repesented 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



Britanova et al. J Immunol 2014

Practical importance

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

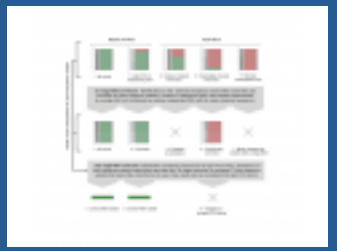
Software for RepSeq data processing

22 software tools currently listed in RepSeq section of OMICtools

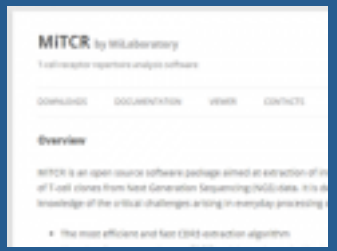
– http://omictools.com/rep-seq-c424-p1.html

IGBLAST

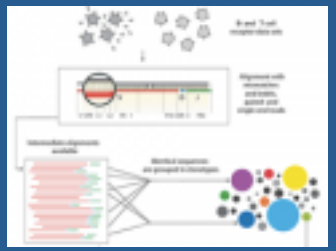
MIGEC



MITCR



MIXCR



VDJtools

OMIC tools

Software for RepSeq data processing

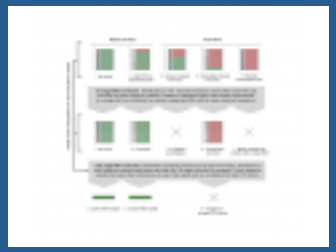
IGBLAST

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

MITCR

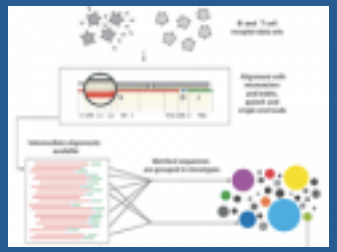
- Extremely fast
- V-(D)-J mapping
- Clonotype assembly
- Error correction (freq-based)
- TCR only

MIGEC



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

MIXCR

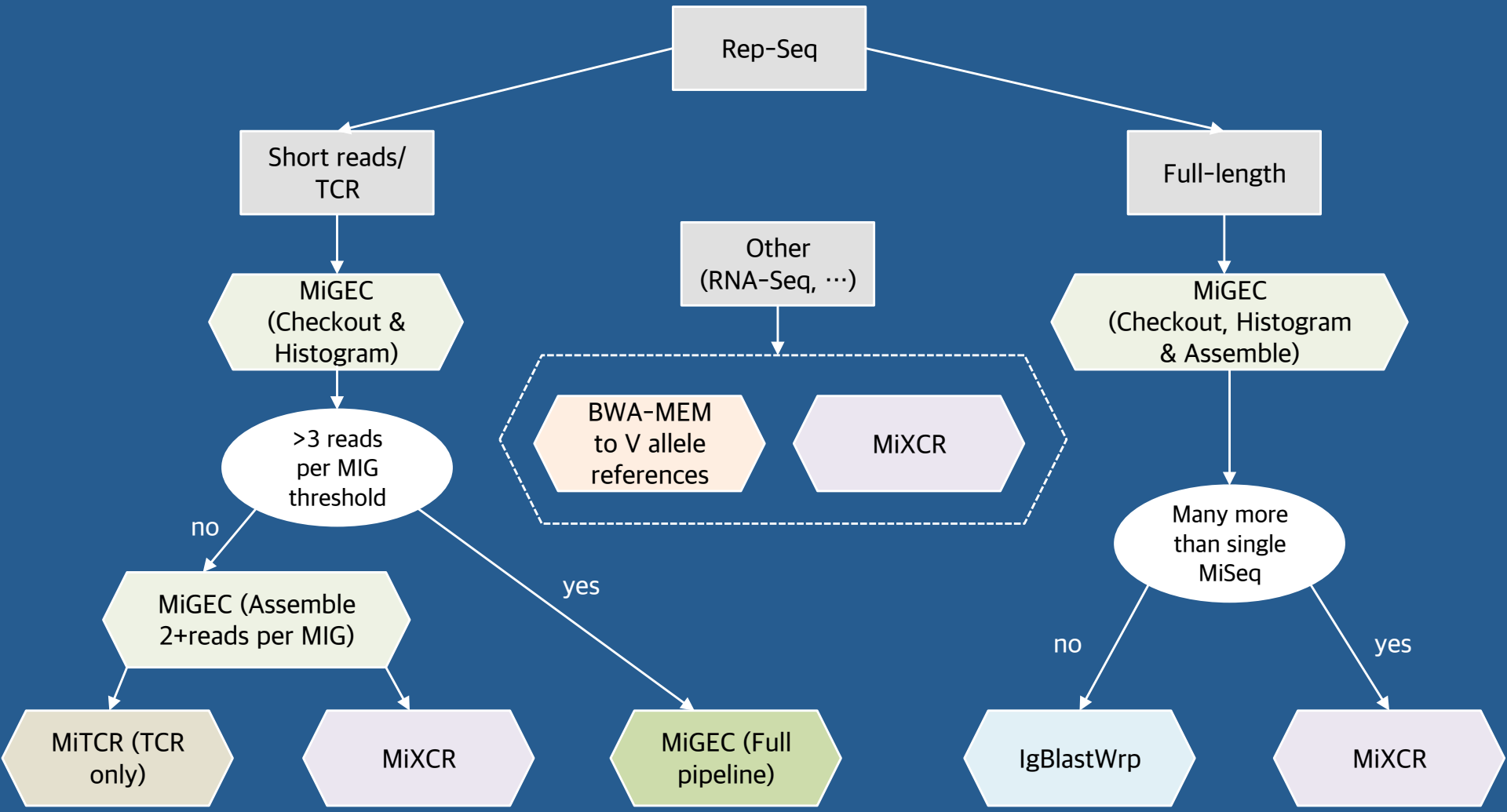


- MITCR extended for B-cells
- V-(D)-J + 5'UTR,isotype,…
- More robust algorithms
- Powerful API
- Full-length

* 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

Typical pipeline



Software for RepSeq post-analysis