6th Educational Workshop on Immunoglobulin Gene Analysis in Chronic Lymphocytic Leukemia

Technical considerations for NGS analysis of immunoglobulin gene repertoires

Uppsala, Sep 2016
IG gene repertoire analysis: let the past go?

• Flow cytometry
  (fast identification of clonal B populations)

• CDR3 spectratyping

• Cloning-based Sanger sequencing
  (molecular characterization of the BcR IG)
Sanger-based immunogenetics: the tip of the iceberg
Next-generation sequencing

- Sequence mixtures of millions of DNA molecules simultaneously

- High-throughput, fast, easy bench work
Next-generation sequencing platforms for IG repertoire profiling

454/Roche
- Pyrosequencing technology
- 500bp (GS Junior)-700bp (GS FLX)
- 150,000 (GS Junior)- 3 million (GS FLX) reads per run

Illumina
- Cyclic reversible termination sequencing
- 100-300bp (paired-end sequencing)
- 15 million (MiSeq)- 6 billion (HiSeq) reads per run

Ion torrent/Life Technologies
- Pyrosequencing technology
- 200bp
- 1 billion reads per run
Next-generation sequencing platforms for IG repertoire profiling

- **454/Roche**
  - Longest reads
  - Highest rate of frameshift errors

- **Illumina**
  - High throughput

- **Ion torrent/Life Technologies**
  - Highest rate of sequencing errors
An efficient IG repertoire analysis requires:

1| **Methods** that describe the repertoire diversity at different levels, for affordable cost, from a little amount of material

2| **Analysis strategies** that reconstitute the best snapshot of immune diversity
High-throughput sequencing of the IG repertoire: 
considerations

1| Experimental design

2| Management of DNA sequence errors

3| Data mining and visualization tools to make biological sense of big data
High-throughput sequencing of the IG repertoire: considerations

1 | Experimental design

2 | Management of DNA sequence errors

3 | Data mining and visualization tools to make biological sense of big data
Source of B cells?

~1-2 x 10^{11} B cells in the human body

- 2% in peripheral blood
- 28% in lymph nodes
- 23% in spleen and mucosal tissues
- 17% in bone marrow
Amount of B cells?

Sequencing depth **MUST** be greater than the number of B cells in the sample

...don’t aim for a deep dive in a shallow pool
gDNA or mRNA?

**gDNA:**
- **Longer sequence** (full IGHV-D-J template: ~550bp)
- Libraries containing both productive and unproductive V(D)J rearrangements
  - decrease of throughput
- Better for estimation of clonal frequencies
  - number of reads proportional to gDNA template molecules

**mRNA (cDNA):**
- **Shorter sequence** (full IGHV-D-J template: ~450bp)
- IG gene transcription varies up to 100-fold between naïve and plasma cells
  - estimation of clonal frequencies will require sorted B cell populations
High-throughput sequencing of the IG repertoire: 

considerations

1| Experimental design

2| Management of DNA sequence errors

3| Data mining and visualization tools to make biological sense of big data
Sources of sequence error

1| Sample preparation

2| Sequencing error
Sources of sequence error

1. Sample preparation

- Base misincorporation during RT or PCR
- Multiplex PCR: differential efficiency of primers?
The 5′ RACE approach
The 5’ RACE approach – con’s

- Confined to mRNA studies
- Lengthy template (600-640bp)
- Cannot exclude differential amplification of some templates over others
Sources of sequence error

2| Sequencing error

• 454 and Ion Torrent: indels vs Illumina: substitution errors

• Qscore/base: quality of base calling
  - the longer you read, the lower the quality
Error correction: can it be done?

No reference sequence for CDR3 and SHM: discrimination of error vs true biologic variation?
Error correction algorithms... 
...never perfect

Exhaustive T-cell repertoire sequencing of human peripheral blood samples reveals signatures of antigen selection and a directly measured repertoire size of at least 1 million clonotypes


Genome Res. 2011 21: 790-797 originally published online February 24, 2011

Filter-out low-abundance clonotypes:
loss of true diversity without elimination of artificial diversity
Error correction algorithms... 
...never perfect

Identification of errors introduced during high throughput sequencing of the T cell receptor repertoire

Phuong Nguyen, Jing Ma, Deqing Pei, Caroline Obert, Cheng Cheng, Terrence L Geiger

Rely on high-quality reads:
- PCR error may accumulate in high-quality reads
Error correction approaches:
molecular barcoding

High-quality full-length immunoglobulin profiling with unique molecular barcoding

M A Turchaninova1–4, A Davydov2,4, O V Britanova1,3,4, M Shugay1,3,4, V Bikos2,4, E S Egorov1–3, V I Kirgizova1, E M Merzlyak1, D B Staroverov1,3, D A Bolotin1,3, I Z Mamedov1,2, M Izraelson1–3, M D Logacheva3, O Kladova3, K Plevova2, S Pospisilova2 & D M Chudakov1–3
First strand cDNA synthesis with template switch and molecular barcoding.

5' adaptor

Template-free added cytosines

IG heavy-chain mRNA

Multiplexed primers for cDNA synthesis: IgM, IgD, IgG, IgA, IgE isotypes

Purify cDNA, using Qiagen MinElute column kit

First PCR, semi-nested

Nested primers, all isotypes

Purify first PCR product, Qiagen column kit

Second PCR, nested/step-out, M1S + Z primers, with sample barcodes

Purify second PCR product, Qiagen column kit

Ligation of Illumina adaptors and amplification
400+100-nt paired-end sequencing using Illumina 500- or 600-cycle kit

- 5' UMI 3'
- 3' UMI 5'

400 nt 100 nt

Low quality

Umi-based assembly and error correction

- 5' UMI 3'
- 3' UMI 5'

High-quality sequence of the cDNA 5' end

- 3' UMI 5'
- 5' UMI

High-quality sequence of the cDNA 3' end

Overlap assembled sequences carrying the same UMI

- 5' UMI 3'
- 3' UMI 5'

High-quality sequence of original cDNA molecule, up to 750-nt length
Error correction approaches: molecular barcoding

Pro’s
• More sophisticated error correction
• Traceable PCR and sequencing error (except for the first-strand cDNA synthesis step)

Con’s
• mRNA-based
• Requires oversequencing of the sample, to obtain multiple reads of the same molecule
• Half of all reads (100nts) do not contribute to final sequence = loss of throughput
High-throughput sequencing of the IG repertoire: considerations

1| Experimental design

2| Management of DNA sequence errors

3| Data mining and visualization tools to make biological sense of big data
Platforms for the analysis of high-throughput IG and TR sequences

• IMGT/HighV-Quest

• ARResT/Interrogate

• ImmunoSEQ

• iRepertoire
**CMV reactivation drives posttransplant T-cell reconstitution and results in defects in the underlying TCRβ repertoire**

Yvonne Suessmuth,¹ Rithun Mukherjee,²,³ Benjamin Watkins,⁴ Divya T. Koura,⁵ Knut Finstermeier,⁶ Cindy Desmarais,⁶ Linda Stempera,¹ John T. Horan,⁴ Amelia Langston,⁵ Muna Qayed,⁴ Hanna J. Khoury,⁵ Audrey Grizzle,⁴ Jennifer A. Cheeseman,¹ Jason A. Conger,¹ Jennifer Robertson,¹ Aneesah Garrett,⁴ Allan D. Kirk,¹ Edmund K. Waller,⁵ Bruce R. Blazar,⁷ Anees K. Mehta,¹ Harlan S. Robins,² and Leslie S. Kean²-⁴,⁸

<table>
<thead>
<tr>
<th>UPN</th>
<th>Rank Order</th>
<th>Clone Frequency (%) Tem</th>
<th>Tem TCR V Gene Sequence</th>
<th>Tem TCR J Gene Sequence</th>
<th>Clone Frequency (%) Tetramer+</th>
</tr>
</thead>
<tbody>
<tr>
<td>001-001</td>
<td>1</td>
<td>11.673</td>
<td>TGTGCCAGCAGCTTAG</td>
<td>TTT</td>
<td>53.272</td>
</tr>
<tr>
<td>001-001</td>
<td>2</td>
<td>11.277</td>
<td>TGTGCCAGCAAG</td>
<td>ACAATGAGCAGTCTTC</td>
<td>41.844</td>
</tr>
<tr>
<td>001-001</td>
<td>3</td>
<td>6.844</td>
<td>TGTGCCAGCAATG</td>
<td>AAAAAACAGCTACTTC</td>
<td>12.006</td>
</tr>
<tr>
<td>001-001</td>
<td>4</td>
<td>5.114</td>
<td>TGTGCCAGCAAG</td>
<td>ACAATGAGCAGTCTTC</td>
<td>0.278</td>
</tr>
<tr>
<td>001-001</td>
<td>5</td>
<td>4.184</td>
<td>TGTGCCAGCAAG</td>
<td>TCAAGCTACTTC</td>
<td>0.159</td>
</tr>
<tr>
<td>001-001</td>
<td>6</td>
<td>3.437</td>
<td>TGTGCCAGCAAG</td>
<td>CGAGCAGTACTTC</td>
<td>0.133</td>
</tr>
<tr>
<td>001-001</td>
<td>7</td>
<td>3.291</td>
<td>TGTGCCAGCAAG</td>
<td>CTCTACAATGAGCAGTCTTC</td>
<td>0.118</td>
</tr>
</tbody>
</table>

Reporting clonotypes without V annotation, N1+N2+D sequence!
Biomathematical analysis of repertoire diversity data

- Principal component analysis
- Discriminant analysis
- Hierarchical clustering
- Specific statistics
High-throughput immunoprofiling needs multidisciplinary approaches

- Immunogenetics
- Bioinformatics
- Statistics

...and standardization

- in experimental methodology
- in analysis and interpretation approaches
IG sequencing:

Is the past gone and the future not here yet?