Cellular Immunology 288 (2014) 31–38

Contents lists available at ScienceDirect

Cellular Immunology

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

Qualifying high-throughput immune repertoire sequencing

Norbert Niklas [a,*], Johannes Pröll [a], Johannes Weinberger [a], Agnes Zopf [a], Karin Wiesinger [a], Konstantin Krismer [b], Peter Bettelheim [c], Christian Gabriel [a]

[a] Red Cross Transfusion Service for Upper Austria, Krankenhausstraße 7, 4020 Linz, Austria
[b] University of Applied Sciences Upper Austria, Softwarepark 11, 4232 Hagenberg, Austria
[c] Division of Hematology/Oncology, Elisabethinen Hospital Linz, Fadingerstraße 1, 4020 Linz, Austria

ARTICLE INFO

Article history:
Received 18 June 2013
Accepted 12 February 2014
Available online 23 February 2014

Keywords:
Next-generation sequencing
Immune repertoire
Clonotyping
Immunoglobulin heavy chain
T cell receptor β
VDJ recombination
Diversity generation
Somatic hypermutation
Big data visualization

ABSTRACT

Diversity of B and T cell receptors, achieved by gene recombination and somatic hypermutation, allows the immune system for recognition and targeted reaction against various threats. Next-generation sequencing for assessment of a cell's gene composition and variation makes deep analysis of one individual's immune spectrum feasible. An easy to apply but detailed analysis and visualization strategy is necessary to process all sequences generated. We performed sequencing utilizing the 454 system for CLL and control samples, utilized the IMGT database and applied the presented analysis tools. With the applied protocol, malignant clones are found and characterized, mutational status compared to germline identity is elaborated in detail showing that the CLL mutation status is not as monoclonal as generally thought. On the other hand, this strategy is not solely applicable to the 454 sequencing system but can easily be transferred to any other next-generation sequencing platform.

1. Introduction

In immunology, the adaptive immune system's B cell and T cell receptor molecules characterize the antigen-specific humoral and cellular arsenal, where a clone is defined by its antigen receptor repertoire sequences [1,2]. This collection of sequences, the so called immune repertoire (IR), is generated by recombination and diversification mechanisms, namely V(D)J recombination (variability (V), diversity (D) and joining (J)) occurring in bone marrow (B cells) and thymus (T cells) as well as somatic hypermutation (SHM) and class switch recombination (CSR) taking place in secondary lymphoid organs for B cells [3,4].

Therefore, the IR is representing an individual's immunologic 'power' at any given time point and is mirroring age [5], vaccination [6], stem cell transplantation [7] and any other reason with direct influence on the diversity of B and T cell populations [8,9]. With the spread of next-generation sequencing (NGS) technologies into many laboratories, quantitative and qualitative analysis of B and T cell receptor sequences has become the method of choice when using elaborate primer sets and well established alignment tools in combination. Based on amplification of a single molecule in emulsion PCR or bridge PCR and the downstream clonal analysis, it can be estimated, that one sequence (read) in the data set represents a single cell's immune receptor. Nevertheless, sample preparation, the sequencing strategy itself as well as the selection of appropriate data analysis and visualization tools are the three main issues, which have to be met when the use of NGS for B cell or T cell clonotyping is intended.

Here we have determined primary simple and straightforward analysis tools for 454 pyrosequencing data on the variable-, diversity- and joining-loci of immunoglobulin heavy (IGHV) and T cell receptor β chain regions (TCRβ). These methods were applied to ten human peripheral blood donations, five healthy control samples and five chronic lymphatic leukemia (CLL) samples. Using these developed data mining strategies and high-throughput data visualization tools, we want to provide fast clonotyping and IR

Abbreviations: PCR, polymerase chain reaction; MID, multiplex identifier; TCR, T cell receptor; BCR, B cell receptor; CLL, chronic lymphatic leukemia; IMGT, international ImMunoGeneTics information system; GI, germline identity; DNA, desoxyribonucleic acid; GS, genome sequencer; PBMC, peripheral blood mononuclear cell; NGS, next-generation sequencing; IR, immune repertoire; M, mutated; UM, unmutated.

* Corresponding author. Fax: +43 732 777 000 430.
E-mail addresses: norbert.niklas@o.roteskreuz.at (N. Niklas), johannes.proell@o.roteskreuz.at (J. Pröll), johannes.weinberger@o.roteskreuz.at (J. Weinberger), agnes.zopf@o.roteskreuz.at (A. Zopf), karin.wiesinger@o.roteskreuz.at (K. Wiesinger), konstantin.krismer@students.fh-hagenberg.at (K. Krismer), peter@bettelheim.eu (P. Bettelheim), christian.gabriel@o.roteskreuz.at (C. Gabriel).

http://dx.doi.org/10.1016/j.cellimm.2014.02.001
0008-8749/

profiling of nucleic acid sequence information for high-throughput data in general. Our goal is not primarily to identify sequences that are rare in general, but those which are unique (biomarker) to a certain situation where the high-frequency clones will be dominant. Applying this comprehensive bioinformatic analysis will reveal previously unrecognized diversity score changes following vaccination-, immunotherapy- or immune disorder studies.

2. Materials and methods

2.1. Clinical setting and experimental design

Peripheral blood samples were collected from five patients with diagnosed lymphoma and from five control samples with a known inconspicuous immune status after signing a written consent. Peripheral blood mononuclear cells (PBMCs) were isolated within 24 h using gradient isolation with FicoLite-H (Linaris blue GmbH, Wertheim, Germany). Genomic DNA was extracted from PBMCs by an automated isolation system (MagnaPure Compact, Roche Diagnostics, Mannheim, Germany). Subsequent amplification for immunoglobulin heavy chain of B cells and β-chain of T cells was performed with multiplex PCR protocols and target specific primers published in [10]. VDJ regions of B cells were captured with 21 primers and 53 primers for T cells, respectively. Schematic sequencing target amplification and gene transcript structure is shown in Fig. 1. 454-Specific adapters and multiplex identifiers (MID, barcode, 10 bp long each) were attached to target specific primers to allow for bead binding and multiple sample tracing.

Briefly, after automated DNA library preparation and emulsion PCR, the created DNA library was sequenced with 200 cycles on the GS FLX + system with Titanium chemistry according to manufacturer's instructions (Roche 454 Life Sciences) without modifications. Samples were sequenced in five GS FLX runs along with other projects in separate regions (different region sizes). B cell sequencing was performed for lymphoma and healthy control samples; T cell sequencing was applied to control samples.

2.2. Data preprocessing

Data analysis took place off-instrument on a HP Proliant server (2x HP PL Intel Xeon QC E5440 with 16GB RAM) and lasted for approximately 9 h per run. After the run, sequences of samples were demultiplexed (separated) by their sequence barcode (MID) and 5′ primers were trimmed. Sequences shorter 100 bp were discarded. Thereafter, preprocessed sequences were uploaded to IMGT/HighV-QUEST database tool with default settings, but excluding individual result files from output [11,12]. IMGT/HighV-QUEST was restricted to maximum 150,000 sequences per job at the date of analysis, for T cell receptor, sequences were split in two parts for uploading and merged before further analysis. The IMGT tool assigns to both B and T cells the most applicable genotype(s) for V-, D- and J-gene. Detailed information on framework, hypervariable region (HVR1-3, also known as complementarity determining regions, CDR) along with the total junction region sequence is provided separately. Complete database output consists of 11 files, containing alignment per region (framework, CDR, junction) and scores, assigned genotypes (and partially alleles), mutational hotspots and statistics, amino acid changes and functionality of sequences (productive, unproductive).

An additional statistics module is provided by the IMGT database for a basic evaluation of the produced sequencing data, however, this module was not used for our analysis strategy or visualization.

2.3. In-depth bioinformatics analysis

Downstream analysis was performed using multiple Perl 5.10.0 scripts (The Perl Foundation, Walnut, CA, USA) together with R 2.14.2 (2012-02-29) [13] for graphics generation and statistical

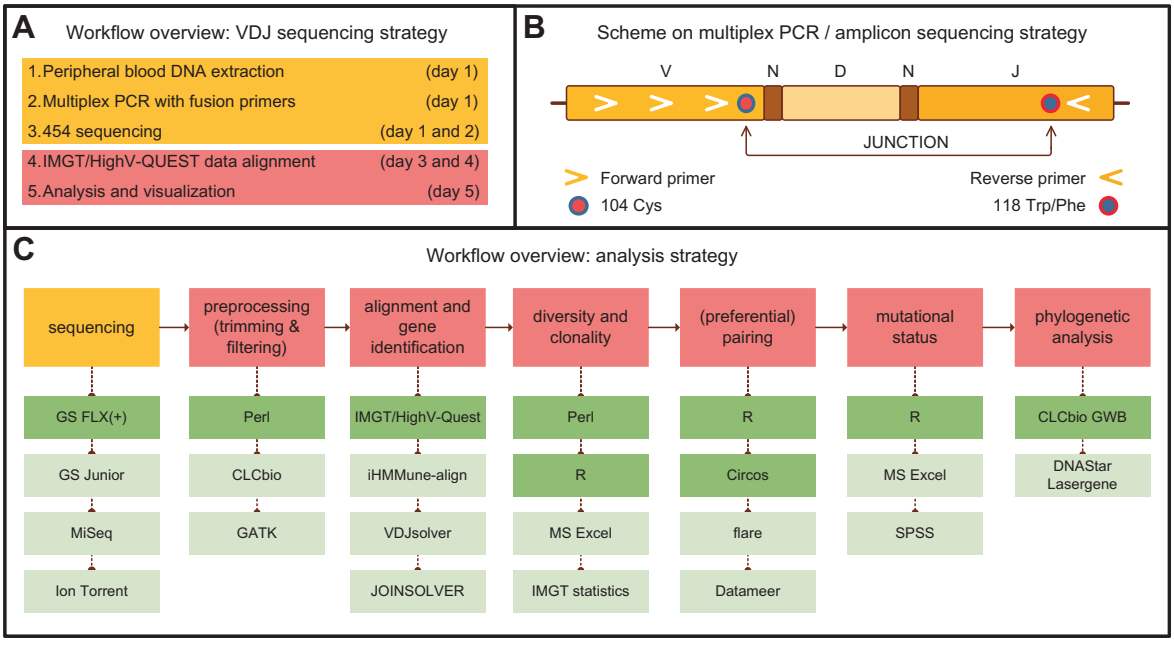

Fig. 1. Overview of workflow and gene structure. (A) Steps and duration of workflow, complete workflow is manageable in 5 days. (B) Structure of rearranged genes in B and T cells (for heavy chain respectively β chain), one V-gene is paired with one (or more) D- and one J-gene, joined with random nucleotides between the genes (N), junction region is determined by IMGT/HighV-QUEST, this is looking for 104 Cys and 118 Trp/Phe [12]. (C) Overview of bioinformatic tasks (red) starting with the sequencing process (yellow), library preparation is not included. Next to each analysis and visualisation, the tools used in this publication are listed (dark green) with a suggestion of other software products (among multiple more) that might be utilized for the same tasks (light green). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)