Molecular Immunology 87 (2017) 12–22

Contents lists available at ScienceDirect

Molecular Immunology

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

Parallel antibody germline gene and haplotype analyses support the validity of immunoglobulin germline gene inference and discovery

Ufuk Kirik[a], Lennart Greiff[b,c], Fredrik Levander[a], Mats Ohlin[a,*]

[a] Dept. of Immunotechnology, Lund University, Lund, Sweden
[b] Dept. of Clinical Sciences, Division of Otorhinolaryngology, Head and Neck Cancer, Lund University, Sweden
[c] Dept. of Otorhinolaryngology, Skåne University Hospital, Lund, Sweden

ARTICLE INFO

Article history:
Received 16 December 2016
Received in revised form 7 March 2017
Accepted 8 March 2017

Keywords:
Antibody
Antibody repertoire
Bioinformatics
Germline gene
Germline gene inference
Haplotype
Heavy chain variable domain

ABSTRACT

Analysis of antibody repertoire development and specific antibody responses important for e.g. autoimmune conditions, allergy, and protection against disease is supported by high throughput sequencing and associated bioinformatics pipelines that describe the diversity of the encoded antibody variable domains. Proper assignment of sequences to germline genes are important for many such processes, for instance in the analysis of somatic hypermutation. Germline gene inference from antibody-encoding transcriptomes, by using tools such as TIgGER or IgDiscover, has a potential to enhance the quality of such analyses. These tools may also be used to identify germline genes not previously known. In this study, we exploited such software for germline gene inference and define aspects of analysis settings and pre-existing knowledge of germline genes that affect the outcome of gene inference. Furthermore, we demonstrate the capacity of IGHJ and IGHD haplotype inference, whenever subjects are heterozygous with respect to such genes, to lend support to IGHV gene inference in general, and to the identification of novel alleles presently not recognized by germline gene reference directories. We propose that such haplotype analysis shall, whenever possible, be used in future best practice to support the outcome of germline gene inference. IGHJ-directed haplotype inference was also used to identify haplotypes not expressing some IGHV germline genes. In particular, we identified a haplotype that did not express several major germline genes such as IGHV1-8, IGHV3-9, IGHV3-15, IGHV1-18, IGHV3-21, and IGHV3-23. We envisage that haplotype analysis will provide an efficient approach to identify subjects for further studies of the link between the available immunoglobulin repertoire and outcomes of immune responses.

1. Introduction

Antibodies are critical components in higher organisms', including mans', defence against bacteria, viruses, and other threats. Consequently, they have been studied intensively since their discovery more than a century ago. Recent advances in cell culture technology, microdroplet technology, structural biology, and next generation sequencing (Georgiou et al., 2014) have substantially enhanced our understanding of immunoglobulins and their recognition of antigen and development following encounter with antigen. The establishment of large germline gene databases and accompanying analysis tools (Alamyar et al., 2012; Giudicelli et al., 2005; Retter et al., 2005; Gaeta et al., 2007) now allows detailed analysis of antibody sequence information. However, germline gene databases are not complete, and analysis of many species' antibody germline gene repertoire is lacking in depth. For instance, sequencing of immunoglobulin genes still identifies new alleles in several cases (Scheepers et al., 2015; Watson and Breden, 2012). The very recent identification of extensive differences in germline gene repertoires of Balb/c and C57BL/6 mouse strains further highlights the lack of completeness of germline databases (Collins et al., 2015). Furthermore, databases contain erroneous entries (Wang et al., 2008) that, depending on the research question, may affect downstream analysis. Although sequencing of the immunoglobulin gene loci of every subject of a study to allow proper germline gene assignment of identified transcripts is a possibility, it represents a substantial obstacle. Therefore, such an approach may not be a realistic option in many investigations. Inference of individual subjects' germline gene repertoires from their immunoglobulin-encoding transcripts using bioinformatics approaches is a more realistic

Abbreviations: BM, bone marrow; CDR, complementarity determining region; H, heavy; PB, peripheral blood; V, variable.

* Corresponding author at: Dept. of Immunotechnology, Lund University, Medicon Village building 406, S-223 81 Lund, Sweden.
E-mail address: mats.ohlin@immun.lth.se (M. Ohlin).

http://dx.doi.org/10.1016/j.molimm.2017.03.012

alternative. Consequently, software tools and pipelines have been developed for this type of inference and they are utilized to facilitate new discoveries in the field of antibody repertoires (Gadala-Maria et al., 2015; Elhanati et al., 2015; Boyd et al., 2010; Kidd et al., 2012; Corcoran et al., 2016; Zhang et al., 2016). The present investigation was designed to assess how analysis is affected by a variety of factors related to e.g. composition of germline gene databases, prior knowledge, and somatic hypermutation/PCR and sequencing errors, in an effort to support future development of best practice in the field of germline gene inference. We specifically exploited inferred haplotypes of IGHJ and IGHD genes to lend support to inferred IGHV genes. Finally, we used the outcome of germline gene and haplotype inference to define deletions in individual haplotypes. In particular we defined one deletion spanning many of the major IGHV genes, in an individual that proved to be heterozygous with respect to alleles of IGHJ6.

2. Materials and methods

2.1. Antibody heavy chain variable domain transcriptomes

Six allergic donors had been recruited for studies of immunoglobulin-encoding repertoires, a study that had been approved by the regional ethical review board at Lund University (Levin et al., 2017). Duplicate samples of both bone marrow (BM) and peripheral blood (PB) cells were obtained out of season of exposure to seasonal allergens (Levin et al., 2017). Transcripts encoding the heavy (H) chain variable (V) domains of different isotypes found in cells of these samples were individually amplified by PCR using primers based on the Biomed2 primer set annealing to the sequence encoding V domain framework region (FR) 1 (van Dongen et al., 2003) and those annealing to the sequence encoding the first constant (C) domain. After addition of barcodes and sequencing adaptors the PCR products were sequenced using the 2×300 bp MiSeq technology (Illumina, Inc., San Diego, CA, USA) at the National Genomics Infrastructure (SciLifeLab, Stockholm, Sweden). FASTQ sequence files (study accession number: PRJEB18926) are available from the European Nucleotide Archive.

2.2. Initial data processing

Paired end reads of IgM-encoding transcripts of one BM sample of each donor was processed using the pRESTO pipeline (Vander Heiden et al., 2014) and sequences were binned based on the isotype-specific 3′ primer as reported in Supplementary Methods and Supplementary Table EIV in Reference Levin et al. (2017). This set was used as such for gene inference using IgDiscover (Fig. 1). Furthermore, sequences not carrying an amplified gene sequence identical to a part of the sequence encoding the first constant domain of the isotype were removed. Sequences were also subsequently analysed (Levin et al., 2017) using IMGT HighV-QUEST tool (Alamyar et al., 2012) and further adapted for use in TIgGER using the Change-O pipeline (Gupta et al., 2015) (Fig. 1). IgM-encoding sequences of duplicate PB samples were pooled and used together for some studies using IgDiscover. The numbers of sequences of each sample at different steps in the analysis pipeline are summarized in Table 2 in Reference Kirik et al. (2017).

2.3. Germline gene inference using TIgGER

TIgGER (Gadala-Maria et al., 2015) uses the output of IMGT/HighV-QUEST (Alamyar et al., 2012) analysis of a set of transcripts, analyses the observed mutational pattern to compute a likely germline gene database, a database that is used to re-analyse the germline gene assignments, the output of which is used to define a likely germline gene set. TIgGER version 0.2.7 (downloaded via CRAN (https://cran.r-project.org)) was used for analysing sequences and inferring genotype information. For each donor, a two-step procedure was carried out, in compliance with the documentation of the software. IGHV gene database was retrieved from IMGT (Lefranc et al., 2015) homepage on 2016-08-16. As a first step novel alleles were searched. At this stage, the range of nucleotides to be considered by the algorithm was set to 79–312, numbering according to IMGT definitions, based on the location of the primers. All other settings were left to defaults, except the number of processors to be used for calculation. In the second step, an IGHV germline genotype inference was carried out for each donor, both with and without filtration of mutated sequences. Furthermore, the gene_cutoff value was set to 1e-3, meaning that a gene must be observed at least 1/1000 of the total allele calls to be included in the genotype.

2.4. Germline gene inference using IgDiscover

IgDiscover (Corcoran et al., 2016) infers germline genes by an iterative process that involves initial assignment of sequences using IgBlast (Ye et al., 2013) to a pre-existing germline gene database, a cluster identification process, and subsequent filtering steps to generate a new germline gene database. IgDiscover version 0.5 was downloaded via the Bioconda channel (https://bioconda.github.io) together with all the dependencies. For consistency in analysis between the TIgGER and IgDiscover, the same pre-processing pipeline was used whenever possible. Thus the input files to IgDiscover consisted of merged paired reads filtered and sorted based on isotype primer by the pRESTO tool. The same list of genes from IMGT was used as reference database, but nucleotides corresponding to amino acids 1–25 were initially removed. Following initial studies (Sections 3.1–3.2 below), nucleotides corresponding to residues 106–107 were also removed from database entries since IgDiscover does not allow for defining a range in V-gene analysis. The gene lists were processed by a Perl script to shorten the headers and remove the gaps, in compliance with IgBLAST requirements.

For testing the influence of the reference database on the inference outcomes, two smaller versions of the reference database containing a representative sequence for different clans were generated, referred to as V123 (containing IGHV1-18*01, IGHV2-5*01, and IGHV3-23*01) and V345 (containing IGHV3-15*01, IGHV4-39*07, and IGHV5-51*01), respectively.

The input data was analysed with the differences parameter (henceforth referred to as diff) set to 0, 1, and 2 (default) for both the pre-germline and germline filters. This parameter controls the number of differences allowed between a sequence and a reference gene for the sequence to be assigned to that genes cluster. In other words, setting this parameter to 0, 1, and 2 allows for separation of alleles that differ by a single, two, and three nucleotides, respectively. Additionally, the minimal number of unique CDR3s parameter of the germline filter was set to 50 in order to weed out potential PCR and sequencing artefacts.

In order to assess sequence read quality related to specific inferred genes/alleles we identified sequence read identities from the output from IMGT/HighV-QUEST analysis. The relevant sequences were to deviate from the gene/allele by not more than one base (script default: >99.4% sequence identity) and to carry a gene/allele-defining sequence motif (e.g. GCTTGAGTGGATGGGA[CT]GGATCAACCCTAACAG of IGHV1-2; bases within square brackets represent the alternatives of the allele-differentiating nucleotide, in this case nucleotide 163 of IGHV1-2). The corresponding quality-defining entities were retrieved from the FASTQ file obtained after pRESTO processing. Average read qualities of bases of the motif were calculated and plotted. The quality control analysis of the sequences was carried