Molecular Immunology 103 (2018) 1–6

Contents lists available at ScienceDirect

Molecular Immunology

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

Critical steps for computational inference of the 3′-end of novel alleles of immunoglobulin heavy chain variable genes - illustrated by an allele of *IGHV3-7*

Linnea Thörnqvist, Mats Ohlin*

Dept. of Immunotechnology, Lund University, Lund, Sweden

ARTICLE INFO

Keywords:
Antibody
Bioinformatics
Germline gene allelic diversity
Germline gene inference
Immunoglobulin germline gene

ABSTRACT

Sequencing of immunoglobulin germline gene loci is a challenging process, *e.g.* due to their repetitiveness and complexity, hence limiting the insight in the germline gene repertoire of humans and other species. Through next generation sequencing technology, it is possible to generate immunoglobulin transcript data sets large enough to computationally infer the germline genes from which the transcripts originate. Multiple tools for such inference have been developed and they can be used for construction of individual germline gene databases, and for discovery of new immunoglobulin germline genes and alleles. However, there are challenges associated with these methods, many of them related to the biological process through which immunoglobulin coding genes are generated. The junctional diversity introduced during rearrangement of the immunoglobulin heavy chain variable (IGHV), diversity and joining genes specifically complicates the inference of the junction regions, with implications for inference of the 3′-end of IGHV genes. With the aim of coping with such diversity, an inference software package may not be able to identify novel alleles harbouring a difference in these regions compared to their closest relatives in the starting database. In this study, we were able to computationally infer one such previously uncharacterized allele, *IGHV3-7*02 A318G*. However, this was possible only if a strategy was used in which different variants of *IGHV3-7*02* were included in the inference-initiating database. Importantly, the presence of the novel allele, but not the standard *IGHV3-7*02* sequence, in the genotype was strongly supported by the actual sequences that were assigned to the allele. We thus showed that the starting database used will impact the germline gene inference process, and that difference in the 3′-end of IGHV genes may remain undetected unless specific, non-standard procedures are used to address this matter. We suggest that inferred genes/alleles should be confirmed *e.g.* by examination of the nucleotide composition of the 3′-bases of the inference-supporting sequence reads.

1. Introduction

Immunoglobulin germline gene loci are complex and repetitive, and consequently difficult to sequence with high fidelity (Watson and Breden, 2012). As a result, the immunoglobulin heavy chain variable (IGHV) locus of only one subject has been sequenced in full (Watson et al., 2013). Considering that allelic diversity, gene duplications and deletions may exist in these loci, it is inevitable that we only have a limited insight into their diverse composition in the human population. Additional genomic gene sequences have been determined, although with limited information on the genes' exact localization within the genome. For instance, the IGHV germline gene *IGHV4-59*08* was recently suggested to be an allele of *IGHV4-61*, at least in some subjects, based on the degree of sequence similarity in regions surrounding the coding sequence itself (Parks et al., 2017). This would explain why it also seems to co-exist on a single haplotype with the allelic variant *IGHV4-59*01* (Kirik et al., 2017a, 2017b). It also appears that many germline genes might have been improperly identified while other genes are missing from the officially recognized repertoire (Wang et al., 2008). Furthermore, the germline gene repertoire of many human populations has not been extensively investigated and may be poorly represented in standard databases (Scheepers et al., 2015; Wang et al., 2011). Consequently, the quality of studies of antibody repertoires and their mutational status may be compromised even when using standard

Abbreviations: CDR, complementarity determining region; D, diversity; H, heavy; IG, immunoglobulin; J, joining; N, non-templated; NGS, next generation sequencing; P, palindromic; V, variable

* Corresponding author at: Dept. of Immunotechnology, Lund University, Medicon Village building 406, S-22381 Lund, Sweden.
E-mail address: mats.ohlin@immun.lth.se (M. Ohlin).

https://doi.org/10.1016/j.molimm.2018.08.018
Received 4 July 2018; Received in revised form 10 August 2018; Accepted 18 August 2018

tools like IMGT/V-QUEST and IMGT/HighV-QUEST (Lefranc, 2011). Studies of non-human antibody repertoires are similarly complicated by a lack of comprehensive knowledge of many species' immunoglobulin germline genes, information that is only now being collected (Collins et al., 2015; Corcoran et al., 2016; Martinez-Murillo et al., 2017).

Large data sets generated through the use of next generation sequencing (NGS) technology are being used to investigate the composition of immunoglobulin repertoires (Nielsen and Boyd, 2018) in the context of *e.g.* infectious diseases, allergic and autoimmune diseases and cancer. Such technology has illustrated important features of immune repertoires, including specific features of antibodies that target unique, highly conserved epitopes on otherwise highly diversified, evolving viral envelope proteins, like those of influenza virus and human immunodeficiency virus (Jackson et al., 2014; Zhou et al., 2013). Importantly, high content datasets harbour sufficient information to computationally infer the germline genes from which they were derived. Software packages like TIgGER (Gadala-Maria et al., 2015), IgDiscover (Corcoran et al., 2016), partis (Ralph, Matsen, 2016a, 2016b), ImPre (Zhang et al., 2016) were all designed to accomplish such inference. Importantly, such technologies may be used to generate individualized databases of germline genes, thereby raising the quality of sequence and mutational analysis.

Although powerful, proper germline gene inference is complicated by biological processes and technological artefacts that may cause improper inference. Somatic hypermutation, PCR errors, PCR cross-over events, and sequencing errors are all events that may contribute to inference of germline genes that were not present in the genome of the subject under investigation. Consequently, inference of germline genes is often guarded by cut-off levels to prevent inference of germline genes that are biological and/or technological artefacts. Such settings have to be established to deliver an acceptable trade-off between inference sensitivity and specificity.

Inference of immunoglobulin germline genes is usually guided by a starting database of known germline genes. The composition of such a germline gene database may, however, influence the set of inferred germline genes (Kirik et al., 2017a, 2017b). Inference of the 3′-end of IGHV genes (as well as of the 5′- and 3′-ends of immunoglobulin heavy chain diversity (IGHD) genes, and the 5′-end of immunoglobulin heavy chain joining (IGHJ) genes) is likely to be particularly vulnerable to inference artefacts. Inference is commonly performed using transcripts of rearranged genes, and such genes have been modified by gene trimming and non-templated (N)/palindromic (P) nucleotide additions, processes that will contribute to sequence diversity at the IGHV-IGHD-IGHJ junctions. Computational inference must take these processes into account when inferring these gene ends, or must completely avoid inferences of bases that are likely to be affected by gene trimming and N/P nucleotide additions. In addition, we (Thörnqvist, Ohlin, 2018a, 2018b) and others (Ralph and Matsen, 2016a) have demonstrated that the extent of incorporation of, in particular, the terminal bases of IGHV germline genes in rearranged genes differ substantially between different IGHV germline genes. Computational inference of the 3′-end of IGHV genes is thus particularly challenging.

In efforts to study the bases of the first codons of complementarity determining region (CDR) 3, that may have their origin in the IGHV germline gene, we also identified diversity suggesting a potential allelic variant, carrying a single base difference (A318G), of *IGHV3-7*02* (Thörnqvist and Ohlin, 2018a, 2018b). The variant sequence was, however, not readily inferred by a computational process. We have now analysed a publicly available dataset that had been suggested to harbour the *IGHV3-7*02* allele (Ralph and Matsen, 2018) to define if conditions could be establish that would allow inference of the sequence variant of *IGHV3-7*02*, if it was present in the sample. We demonstrate that sequences most similar to *IGHV3-7*02*, carried G318 at a frequency of > 90%. Nevertheless, inference software, again, were unable to infer this allelic variant using standard analysis settings. Furthermore, we illustrate the impact of the composition of the germline gene database that was used to initiate the inference of the subject's germline gene repertoire on the inference output. These findings suggest that inference software developers must validate their utility's ability to properly infer sequence variants at the 3′-end of IGHV genes. Finally, we propose that inferences of the 3′-most bases of IGHV germline genes are accompanied by supporting data, such as an analysis of the nucleotide composition of these bases in reads that represent the allele in question.

2. Materials and methods

2.1. NGS data set

The data used in this study were obtained from a publicly available dataset (European Nucleotide Archive: PRJNA349143) containing immunoglobulin transcript sequences from blood samples of three individuals. The samples had been collected at ten timepoints, ranging from 8 days before to 28 days after vaccination against influenza (Laserson et al., 2014), and later re-sequenced using Illumina MiSeq. Here, the re-sequenced samples, developed after amplification using a 5′-RACE methodology that used IGHC-specific 3′-primers, of the individual IB have been analysed (European Nucleotide Archive: SRR4431764-SRR4431773).

2.2. Data pre-processing

Each of the samples of the studied individual was processed separately using pRESTO 0.5.4 (Vander Heiden et al., 2014). Low quality reads were filtered out using the FilterSeq.py script (q = 20) and reads were assembled with PairSeq.py followed by AssemblePairs.py. As only IgM sequences were desired, any reads lacking the IgM specific sequence AGGGAGTGCATCC were discarded. Finally, remaining IgM reads were pooled into one set of sequences that was used for further analysis.

2.3. Germline gene inference

The germline gene origin of all IgM sequences from the studied individual was inferred using IgDiscover 0.9 (Corcoran et al., 2016), generally applying default settings, but running three iterations. Starting databases of IGHV, IGHD and IGHJ genes were downloaded from the IMGT server (Lefranc, 2011). In total, the inference process was performed five times, each time with a slightly different version of the IGHV starting database, in relation to the *IGHV3-7*02* allele. The sequence of this allele (Winkler et al., 1992), as represented in the IMGT database, is two bases shorter than most other germline genes. This allele was, for the purpose of this inference, present in the starting database either in its original form, extended by two bases (GA, found at the 3′-end of most germline genes, including *IGHV3-7*01* and *IGHV3-7*03*), and/or modified with an A318G substitution (Table 1).

Table 1
Nucleotides downstream of position 312 of the *IGHV3-7*02* variants included in the starting databases used for germline gene inferences with IgDiscover (Corcoran et al., 2016), and the corresponding inferred bases.

	3'-end of *IGHV3-7*02* variants included in the starting database	3'-end of the inferred *IGHV3-7*02* variant
Inference A	GCGAGAGA	GCGAGAGA
Inference B	GCGAGGGA	GCGAGGGA
Inference C	GCGAGAGA and GCGAGGGA	GCGAGGGA
Inference D	GCGAGA	GCGAGA
Inference E	GCGAGA and GCGAGG	GCGAGG