CORRESPONDENCE OPEN

Check for updates

Commentary on Population matched (pm) germline allelic variants of immunoglobulin (IG) loci: relevance in infectious diseases and vaccination studies in human populations

Genes & Immunity (2021) 22:335–338; https://doi.org/10.1038/s41435-021-00152-6

Dear Editor,

In their recent publication, Khatri et al. [1] describe an immunoglobulin germline gene database inferred from short-read genomic sequence data derived from five superpopulations. The development of methods for the compilation of more complete and accurate germline gene databases would be an important achievement, but we do not believe that this has been achieved. Existing databases are clearly incomplete. Germline sequences differ substantially between subjects, and alleles found in some populations may be absent in others. Existing databases are likely biased towards alleles found in European populations and may lack many sequences found in understudied populations [2]. Improved, properly designed and curated germline gene databases are therefore needed for analysis of antibody repertoires.

Extensive efforts are under way to better document germline genes, and the study by Khatri et al. represents one such effort. Yet, their use of short-read sequencing data from the 1000 Genomes Project involves special challenges that we believe have not been met. The pitfalls of using short read data for genetic analysis of the IG loci have been discussed in detail previously [3–5]. Here, we focus on specific issues that call into question the completeness and accuracy of the pmIG database, and highlight shortcomings in the methodology which may explain them. We are motivated by concerns that less-informed users might be drawn towards the use of pmIG because it appears to contain a larger number of identified alleles than other databases. We believe that despite the breadth of the database, its problems compromise its utility. For brevity, the discussion is restricted here to the heavy chain IGHV genes, but the comments are equally relevant to the other immunoglobulin gene loci.

We first draw attention to the omission of many genes and alleles from the pmIG database. One reason for this is that the use of the GRCh37 assembly in analyses conducted to construct the pmIG database necessarily restricts it to those genes that are annotated in the assembly [3]. This excludes "duplicated" genes as noted by the authors, but also those present in other structurally variant haplotypes. In total, 15 known open reading frame IGHV genes are missing from GRCh37, and therefore from pmIG. These amount to approximately one fourth of all functional/ORF IGHV genes, and account for 43 alleles found in the IMGT and OGRDB databases. Importantly, the functional, missing genes have been commonly observed in many studies [5–12]. For example, the genes IGHV7-4-1, IGHV3-64D, IGHV5-10-1, IGHV4-30-2, and IGHV4-30-4, are identified in 46%, 65%, 62%, 75%, and 66%, respectively, of the transcribed repertoires for 421 human subjects curated by www.vdjbase.org [13].

Many common alleles of genes that are present in GRCh37 are also missing from pmIG, such as IGHV2-70*01 (found in 73% of subjects at VDJbase), IGHV3-11*06 (64%), and IGHV3-66*01 (53%), all of which are supported by genomic sequencing of unrearranged elements [8, 14, 15]. In the case of IGHV1-69, the absence of many well-documented alleles from pmIG can be traced to the absence of a single SNP in the pmIG results, as depicted in Khatri et al. Supplementary Fig. 2C [1]. These absences likely stem from the mis-mapping of reads to incorrect positions within the IGH locus, or their complete exclusion, with effects that carry through to variant calls found within the VCF files. Such cases can occur in duplicated and repetitive loci where it can be difficult to confidently assign short reads to a single position. In addition, reads from structural variants that differ from the reference assembly can also be mis-mapped to the closest-matching position within the reference, producing erroneous variant calls.

Certain genes and alleles may be particularly susceptible to erroneous read mapping. The germline gene IGHV4-4 is often represented by the IGHV4-4*07 and IGHV4-4*02 alleles, but the variant IGHV4-4*01 is also common. IGHV4-4*07 is the allele present in the GRCh37 reference genome. All IGHV4-4 alleles in pmIG are close variants of IGHV4-4*07, strongly suggesting that reads derived from the IGHV4-4*01 and IGHV4-4*02 alleles are systematically misassigned, creating chimeric sequences that contribute either to the apparent diversity of the IGHV4-4*07-like variants, or to "novel" alleles of similar genes of the IGHV4 subgroup. Similarly, a short sequence string seen in half of the currently curated alleles of IGHV3-11 is not found in any of the IGHV3-11 alleles in pmIG, but rather appears in a similar sequence context in four pmIG alleles of IGHV3-48. Evidence for this mis-mapping can be observed in sample data from the 1000 Genomes Project (Fig. 1).

Identical mapping errors can be expected to occur in multiple datasets from the 1000 Genomes Project, as many individuals will share variant haplotypes. Misidentification of "novel" alleles can also be expected in these datasets, and multiple observations of a particular variant may point to systematic errors. We therefore do not accept the assertion of Khatri et al. [1] that there can be confidence in "novel alleles" that were identified in at least 7 haplotypes. The omission of numerous common alleles supported by various studies and methods, and in particular those identified by the amplification and sequencing of unrearranged genomic DNA, strongly suggests the presence of unexplained systematic errors. The authors also claim confidence in their methods based upon analysis of a single sample for which long-read sequencing was available [5]. They report the correct identification of 61 of 66 IGHV sequences in this sample but do not report whether

Received: 26 July 2021 Revised: 29 September 2021 Accepted: 5 October 2021
Published online: 19 October 2021

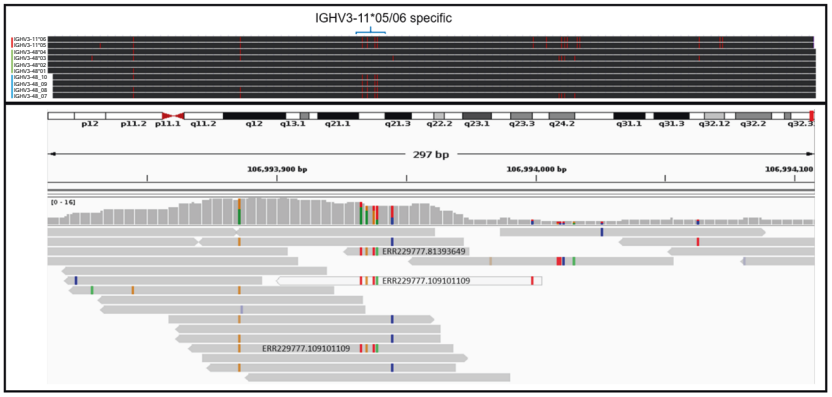

Fig. 1 Problems caused by short read mis-assignments. Upper panel: Alignment of known and novel candidate germline sequences to GRCh37 reference assembly using the BLAT genome browser. Alleles IGHV3-11*05 and IGHV3-11*06 are indicated with the vertical red bar, IGHV3-48*01, IGHV3-48*02, IGHV3-48*03 and IGHV3-48*04 with the green bar and pmIG candidate alleles IGHV3-48_7, IGHV3-48_8, IGHV3-48_9 and IGHV3-48_10 with the blue bar. Position of the IGHV3-11 specific cluster of SNP variants is shown with the blue bracket. Lower panel: Assignment of low coverage sequences from 1000 genomes case HG00105, from the (HGO0105.mapped.ILLUMINA.bwa.GBR.low_-coverage.20130415.bam) BAM file of GRCh37 assigned sequences visualized using the Broad Institute IVG viewer. Three short read sequences are incorrectly assigned to IGHV3-48 locus, ERR229777.81393649, ERR229777.109101109 and ERR229777.100808222 as shown by the presence of an IGHV3-11*05/06 specific segment containing four SNP variations, rs199879022 T (A in IGHV3-48), rs200437959 G (A in IGHV3-48) rs200973953 T (G in IGHV3-48) rs199815306 A (T in IGHV3-48). Four candidate IGHV3-48 germline sequences, IGHV3-48_7, IGHV3-48_8, IGHV3-48_9 and IGHV3-48_10, appear to be chimeric in origin, erroneously containing the IGHV3-11*05/06 segment.

false-positive calls were made. Khatri et al. explain their failure to find 5 of the 66 IGHV sequences as resulting from "differences in the human reference genome assembly used for mapping and calling alleles". Although IGHV4-4*02, IGHV3-66*01, IGHV1-69*04, IGHV2-70*01 and IGHV2-70*15 are absent from GRCh37, this should not prevent their identification in a GRCh37-based analysis, as alleles of each of these genes are present in the reference assembly.

There are other anomalies in the database. For example, there is an extra base in framework region 3 in 4 of 5 alleles of IGHV3-49 and in 18 of 19 alleles of IGHV3-53. In addition, many (but not all) genes and alleles lack a number of bases at their 3'-end, an error that may compromise gene annotation and the analysis of the generation of diversity in the third hypervariable loop of the receptor. This anomaly is particularly surprising given that genomic sequences were used to generate the database. A third issue is that pmIG assigns novel allele names to sequences that only vary in their leader sequences. Multiple sequences may therefore contain V-exons that match a named allele in the IMGT database, creating challenges for any attempt to use a combined database for AIRR-seq analysis.

Without additional work to explain and rectify errors and omissions of genes and alleles, the pmIG database is unsuited to many applications. Specifically, for AIRR-seq analysis, errors and omissions will result in erroneous germline gene and allele assignments, and ultimately impact the accuracy of other analyses including clonal inference and estimates of somatic hypermutation. The consequences will be most serious where absent genes and alleles are highly divergent from all sequences in the database. For example, if an AIRR-seq dataset from a donor carrying the allele IGHV4-30-2*01 (the most common allele of this gene) was analyzed using pmIG, reads derived from this gene would not be assigned to IGHV4-30-2, but instead to the closest germline sequence in pmIG, IGHV4-31_7, which differs by 13 nucleotides. To demonstrate the extent of these effects, we analyzed 98 naïve B cell repertoires. The cohort's repertoires were aligned with the IMGT and pmIG reference sets, and mutation levels were compared across individuals and for individual genes. The calculated median mutation level across individuals was more than twice as high when sequences were aligned with pmIG compared to IMGT (Fig. 2A). This is a consequence of the many common genes and alleles present in this Norwegian cohort that are missing from pmIG (Fig. 2B). Comparison of the identified alleles using the two reference sets shows that 93 alleles were shared between the references, 88 were only observed in IMGT and 146 were observed only in pmIG (Fig. 2C).

The need to better document variation in the human immunoglobulin loci makes it important to exploit all available data sources where possible. At present however, in our view, the errors and omissions we have found suggest fundamental flaws in gene calling from short read data. We recognize that some and perhaps many alleles identified by Khatri and colleagues [1] are genuine, but we are presently unable to determine the reliability of any of the novel calls. If these calls were publicly linked to individual data sets, a more nuanced assessment of specific calls could be made, and inferences could be checked experimentally against their corresponding samples. Until such checks have been made, and until we better understand the challenges and limitations of short read data for the compilation of germline databases, we strongly advise against implementation of the current pmIG database, or similarly derived databases, in any AIRR-seq analysis.

When clear evidence is available that points to the reliability of adaptive immune receptor gene discovery from short-read sequencing projects, it will be important for processes to be established to review, name and document well-supported sequences. We believe this important task should be pursued through community-wide processes, in conjunction with naming authorities of the International Union of Immunological Societies. Such processes have been established for the evaluation of germline genes inferred from AIRR-Seq data [16, 17]. The substantial expansion of documented alleles that should result from short-read genome assemblies will then require more careful approaches than ever for the analysis of AIRR-Seq data. In particular, if alignments of mutated V(D)J sequences against the germline repertoire are to produce unequivocal alignments, the determination of individual genotypes prior to the reanalysis of AIRR-Seq datasets will need be recognized as an essential step in studies of the antibody repertoire.

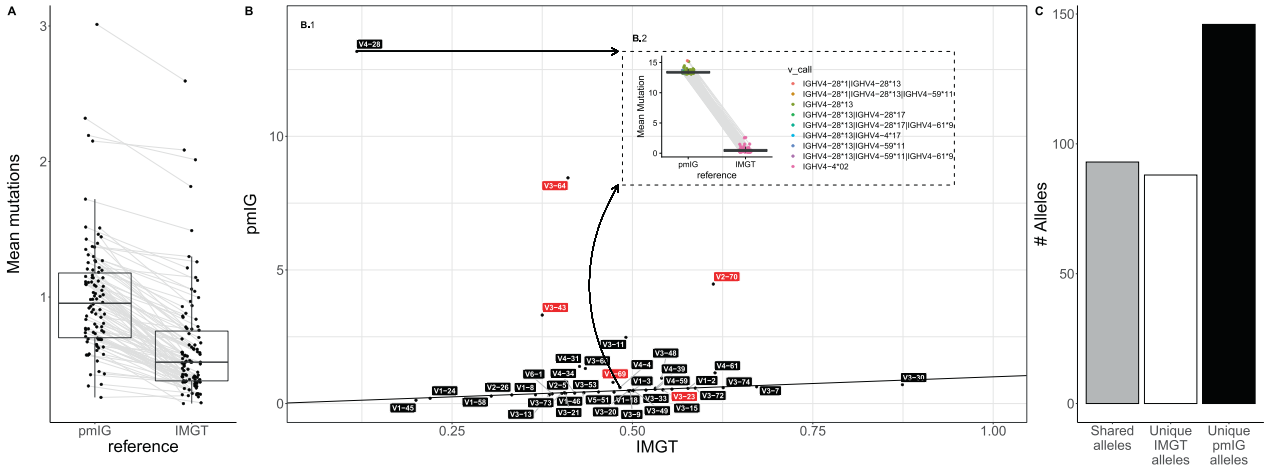

Fig. 2 pmIG reference database introduces erroneous mutations. Repertoire analysis was performed on a naïve B-cell cohort (not expected to carry somatic mutations) of 98 individuals reported by Gidoni et al. (SRA: PRJEB26509) [6]. The repertoires were sequenced using the 5′RACE protocol and pre-processing was done as described in Gidoni et al. [6]. For the downstream analysis, repertoires were initially aligned (IgBLAST version 1.16.0) with the IMGT reference (March 29, 2021). Non-functional sequences, and functional sequences that were not full-length, were not assigned to a single V-gene unambiguously, or were assigned to a V-gene not present in the pmIG database were removed. The remaining 70% of functional sequences were then aligned with the pmIG reference database (downloaded from https://pmtrig.lumc.nl/ on July 15th, 2021) and the repertoires were compared. **A** For each repertoire, the mean mutation count was calculated using each reference database. Each dot represents the mean mutation count and each boxplot represents the variation within the cohort for each of the reference databases. **B** B.1 Each dot is the median of the mean individual mutation frequency per gene. The *X* axis is based on the IMGT reference and the *Y* axis is based on the pmIG reference. Red labels represent the genes with a duplicated copy in the chromosome (e.g., IGHV1-69/IGHV1-69D). B.2 Each dot represents an individual with IGHV4-4*02 in their IMGT data. These calls were matched via sequence IDs to calls in the matching pmIG datasets, and different gene annotations are shown with different colors. Where multiple allele calls were made, these calls are separated in the legend by vertical bars. The *X* axis shows the annotations to the two datasets. The *Y* axis shows the mean mutation numbers for the sequences assigned to the IGHV4-4*02 allele and to their matching calls in the pmIG dataset. **C** The count of alleles that are represented in the cohort for each of the reference databases.

Andrew M. Collins[1], Ayelet Peres[2,3], Martin M. Corcoran[4], Corey T. Watson[5], Gur Yaari[2,3], William D. Lees[6] and Mats Ohlin[7]✉

[1]School of Biotechnology and Biomolecular Sciences, University of New South Wales, Sydney, NSW, Australia. [2]Bioengineering, Faculty of Engineering, Bar Ilan University, Ramat Gan, Israel. [3]Bar Ilan Institute of Nanotechnologies and Advanced Materials, Bar Ilan University, Ramat Gan, Israel. [4]Department of Microbiology, Tumor and Cell Biology, Karolinska Institute, Stockholm, Sweden. [5]Department of Biochemistry and Molecular Genetics, University of Louisville School of Medicine, Louisville, KY, USA. [6]Institute of Structural and Molecular Biology, Birkbeck College, University of London, London, UK. [7]Department of Immunotechnology, Lund University, Lund, Sweden. ✉email: mats.ohlin@immun.lth.se

REFERENCES

1. Khatri I, Berkowska MA, van den Akker EB, Teodosio C, Reinders MJT, van Dongen JJM. Population matched (PM) germline allelic variants of immunoglobulin (IG) loci: relevance in infectious diseases and vaccination studies in human populations. Genes Immun. 2021;22:172–86. https://doi.org/10.1038/s41435-021-00143-7.
2. Peng K, Safonova Y, Shugay M, Popejoy AB, Rodriguez OL, Breden F. et al. Diversity in immunogenomics: the value and the challenge. Nat Methods. 2021;18:588–91. https://doi.org/10.1038/s41592-021-01169-5.
3. Watson CT, Matsen FA, 4th, Jackson KJL, Bashir A, Smith ML, Glanville J. et al. Comment on "A Database of Human Immune Receptor Alleles Recovered from Population Sequencing Data". J Immunol. 2017;198:3371–3. https://doi.org/10.4049/jimmunol.1700306.
4. Luo S, Yu JA, Li H, Song YS. Worldwide genetic variation of the IGHV and TRBV immune receptor gene families in humans. Life Sci Alliance. 2019;2:e201800221. https://doi.org/10.26508/lsa.201800221.
5. Rodriguez OL, Gibson WS, Parks T, Emery M, Powell J, Strahl M. et al. A novel framework for characterizing genomic haplotype diversity in the human immunoglobulin heavy chain locus. Front Immunol. 2020;11:2136. https://doi.org/10.3389/fimmu.2020.02136.
6. Gidoni M, Snir O, Peres A, Polak P, Lindeman I, Mikocziova I. et al. Mosaic deletion patterns of the human antibody heavy chain gene locus shown by Bayesian haplotyping. Nat Commun. 2019;10:628. https://doi.org/10.1038/s41467-019-08489-3.
7. Zhu Y, Yang X, Ma C, Tang H, Wang Q, Guan J. et al. Antibody upstream sequence diversity and its biological implications revealed by repertoire sequencing, J Genet Genomics. 2021 (in press) https://doi.org/10.1016/j.jgg.2021.06.016.
8. Watson CT, Steinberg KM, Huddleston J, Warren RL, Malig M, Schein J. et al. Complete haplotype sequence of the human immunoglobulin heavy-chain variable, diversity, and joining genes and characterization of allelic and copy-number variation. Am J Hum Genet. 2013;92:530–46. https://doi.org/10.1016/j.ajhg.2013.03.004.
9. Kirik U, Greiff L, Levander F, Ohlin M. Parallel antibody germline gene and haplotype analyses support the validity of immunoglobulin germline gene inference and discovery. Mol Immunol. 2017;87:12–22. https://doi.org/10.1016/j.molimm.2017.03.012.
10. Soto C, Bombardi RG, Branchizio A, Kose N, Matta P, Sevy AM. et al. High frequency of shared clonotypes in human B cell receptor repertoires. Nature. 2019;566:398–402. https://doi.org/10.1038/s41586-019-0934-8.
11. Goldstein LD, Chen Y-JJ, Wu J, Chaudhuri S, Hsiao Y-C, Schneider K. et al. Massively parallel single-cell B-cell receptor sequencing enables rapid discovery of diverse antigen-reactive antibodies. Commun Biol. 2019;2:304. https://doi.org/10.1038/s42003-019-0551-y.
12. Gadala-Maria D, Gidoni M, Marquez S, Vander Heiden JA, Kos JT, Watson CT. et al. Identification of subject-specific immunoglobulin alleles from expressed repertoire sequencing data. Front Immunol. 2019;10:129. https://doi.org/10.3389/fimmu.2019.00129.
13. Omer A, Shemesh O, Peres A, Polak P, Shepherd AJ, Watson CT. et al. VDJbase: an adaptive immune receptor genotype and haplotype database. Nucleic Acids Res. 2020;48:D1051–6. https://doi.org/10.1093/nar/gkz872.
14. Shin EK, Matsuda F, Nagaoka H, Fukita Y, Imai T, Yokoyama K. et al. Physical map of the 3′ region of the human immunoglobulin heavy chain locus: clustering of autoantibody-related variable segments in one haplotype. EMBO J. 1991;10:3641–5. https://doi.org/10.1002/j.1460-2075.1991.tb04930.x.
15. Berman JE, Mellis SJ, Pollock R, Smith CL, Suh H, Heinke B. et al. Content and organization of the human Ig VH locus: definition of three new VH families and

linkage to the Ig CH locus. EMBO J. 1988;7:727–38. https://doi.org/10.1002/j.1460-2075.1988.tb02869.x.
16. Lees W, Busse CE, Corcoran M, Ohlin M, Scheepers C, Matsen FA, IV. et al. ORGDB: a reference database of inferred immune receptor genes. Nucleic Acids Res. 2020;48:D964–70. https://doi.org/10.1093/nar/gkz822.
17. Ohlin M, Scheepers C, Corcoran M, Lees WD, Busse CE, Bagnara D. et al. Inferred allelic variants of immunoglobulin receptor genes: a system for their evaluation, documentation, and naming. Front Immunol. 2019;10:435. https://doi.org/10.3389/fimmu.2019.00435.

AUTHOR CONTRIBUTIONS

The authors together conceived the commentary, wrote the paper and approved the final version of the paper. MC and AP carried out the analyses specifically reported in Figs. 1 and 2, respectively.

FUNDING

Parts of this study were supported by a grant from the Swedish Research Council (grant number 2019-01042) (MO).

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Correspondence and requests for materials should be addressed to Mats Ohlin.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.