Computational inference, validation, and analysis of 5'UTR-leader sequences of alleles of immunoglobulin heavy chain variable genes

Yixun Huang[†], Linnea Thörnqvist[†], and Mats Ohlin*

Dept. of Immunotechnology, Lund University, Medicon Village building 406, S-223 81 Lund, Sweden

[†] These authors have contributed equally to this work and share first authorship

* To whom correspondence should be addressed. Tel: +46-46-2224322; Email: mats.ohlin@immun.lth.se

Keywords: adaptive immune receptor repertoire (AIRR), germline gene inference, immunoglobulin germline gene, immunoglobulin heavy chain variable domain, leader sequence, 5'-untranslated region

ABSTRACT

Upstream and downstream sequences of immunoglobulin genes may affect the expression of such genes. However, these sequences are rarely studied or characterized in most studies of immunoglobulin repertoires. Inference from large, rearranged immunoglobulin transcriptome data sets offers an opportunity to define the upstream regions (5'-untranslated regions and leader sequences). We have now established a new data pre-processing procedure to eliminate artifacts caused by a 5'-RACE library generation process, reanalyzed a previously studied data set defining human immunoglobulin heavy chain genes, and identified novel upstream regions, as well as previously identified upstream regions that may have been identified in error. Upstream sequences were also identified for a set of previously uncharacterized germline gene alleles. Several novel upstream region variants were validated, for instance by their segregation to a single haplotype in heterozygotic subjects. SNPs representing several sequence variants were identified from population data. Finally, based on the outcomes of the analysis, we define a set of testable hypotheses with respect to the placement of particular alleles in complex IGHV locus haplotypes, and discuss the evolutionary relatedness of particular heavy chain variable genes based on sequences of their upstream regions.

1 INTRODUCTION

Immunoglobulins play a vital role in recognition of pathogens, thereby enabling their removal or modification of their activities or functions. The typical antibody consists of two identical heavy (H) chains and two identical light chains, of which the H chain often plays a dominant role in determination of specificity (1). The diversity of antibody H chains is established by somatic recombination of immunoglobulin variable (IGHV), diversity (IGHD) and joining (IGHJ) genes, along with junction diversity and somatic hypermutation. Thanks to the development of next-generation sequencing (NGS), it has been possible to describe the nature of the adaptive immune receptor repertoire (AIRR), both in general terms and in relation to e.g. infectious disease, autoimmunity and allergy. Furthermore, it has been possible to approach features of AIRR at a personalized germline

gene level as a key factor in the nature of developing immune responses (2). The importance of the personal germline gene repertoire for the development of specific antibodies may indeed be substantial, in particular in view of the importance of stereotyped (public) immune responses against a number of antigens (3).

The germline gene repertoire that encodes final, processed, complete antibody variable domains is extensively described and addressable by bioinformatic tools (4,5). The IMGT (the international ImMunoGeneTics information system) database (6) has developed into a recognized collection of germline genes for analysis of T and B cell AIRR. Despite the development of techniques visibly expanding our knowledge of germline gene variants, the reference database of such genes still cannot be considered to be complete and accurate (7). Importantly, however, long-read sequencing (8) and other NGS technologies and bioinformatics approaches (9-12) now allow us to generate extended, and personalized databases that in the future will enable better, high-quality analysis of AIRR as they develop in health and disease.

Features used to generate antibody repertoires, other than the nucleotide sequence of the product-encoding part of germline genes, are less well defined, studied and understood. Yet, they may play a role in gene expression and generation of a functional antibody repertoire. These include the 5'-untranslated region (5'UTR), the leader sequence encoding the signal peptide that play a vital role in protein transport (13-15), introns of immunoglobulin genes, 3'-non-coding regions including the recombination signal sequence, and more distant regulatory elements (16). Bioinformatic tools developed for studies of large transcriptomic repertoire data sets, such as IgDiscover (9) and IMGT/HighV-QUEST (17), are already able to capture parts of the 5'UTRs and the signal peptide-encoding part of the genes in many existing NGS data sets. Recent studies, however, have suggested that the diversities of 5'UTR and leader sequences are not well represented in the IMGT database (14,18), strongly arguing that such information ought to be updated to enable analysis of the role of these regions in gene expression and functionality. Heterozygosity in 5'UTR and leader sequences may also be used in sequence haplotyping efforts to assess gene expression from individual chromosomes (19,20) even in cases when their associated IGHV genes are identical, thereby allowing further development of our understanding of these genes in a broader context.

NGS-derived AIRR data generated from B cell lineage transcriptomes are now made available for analysis at a large scale. Many such data sets have been generated using 5'-RACE (rapid amplification of cDNA ends) technology and thus incorporate part of the 5'UTR and the entire leader sequence (but not its intron). Proper data pre-processing is important to accurately identify the 5'UTR and leader sequence of the genes. In this study, we developed a tool to remove 5'-barcodes and the homopolymeric tail introduced during the sequencing library generation process, to enable pre-processing of an NGS data set of antibody repertoires. Immunoglobulin germline gene repertoires were inferred by IgDiscover (9) and 5'UTR-leader sequences were subsequently used to infer consensus 5'UTR-leader sequences of each IGHV gene/allele. We explored haplotype analysis, third complementarity-determining region (CDR3) length distribution patterns, and population genome data to validate inferred sequences. Several novel examples of diversity in these upstream regions are described and discussed. Our findings extend the reference database of validated 5'UTR-leader

sequences, and the study provides a new pipeline to infer and analyze the upstream sequences of IGHV-encoding genes, a pipeline that likely can be adapted to assess AIRR diversity in other transcriptome data sets.

2 MATERIALS AND METHODS

2.1 Data set

A publicly available NGS data set of antibody repertoires, first published in a study by Gidoni et.al (19), was analyzed in this study. The data set was obtained from the European Nucleotide Archive (ENA) under the accession number PRJEB26509. It contains reads of antibody heavy chain transcript data generated from naïve B-cells of 100 individuals in Norway. The 100 subjects comprise 52 patients with celiac disease and 48 healthy controls. Sequencing had been performed by a 300*2 paired-end kit by Illumina MiSeq. As in the study of Mikocziova et al. (18), data of two subjects (ERR2567273 & ERR2567275) were here excluded due to low sequencing depth.

2.2 Data pre-processing

The sequencing library of the used data set had been generated by 5'-RACE technology, where terminal deoxynucleotidyl transferase (TdT) typically is used to extend cDNA with a homopolymeric tail at its 3'-end, a tail that subsequently can be used to add an adaptor sequence. Details, such as what adaptor sequences that were used, were not stated in the original paper (19), but through inspection of a set of random raw reads, we identified that the sequences, upstream of the antibody genes, incorporated a random barcode sequence followed by a TAC-G_n adaptor. Using an in-house developed tool, we trimmed the 5' end of the forward sequences, up to and including the homopolymeric tail (Figure 1). In addition, forward reads were removed entirely if the sequence TAC-G_3 was not found within their first 40 bases. The PairSeq.py tool of the pRESTO 0.6.0 software (21) was subsequently used to combine remaining forward reads with reverse reads, into full-length sequences.

2.3 Germline gene inference and data filtering

The pre-processed sequences were analyzed by IgDiscover 0.12 (9), using default parameters, in order to infer personalized germline gene repertoires. Reference databases of human IGHV, IGHD and IGHJ genes were obtained from the IMGT database (release 202011-3) (6) (Supplementary Data 1). Thereafter, we filtered the IgDiscover processed reads by removing entries with V_errors > 0, as these either have been subject to sequencing errors or somatic hypermutations, or assigned to the wrong IGHV germline gene allele (Figure 1). Additionally, any germline gene allele with low diversity (defined as fewer than 75 unique CDR3s, all entries considered) or low number of assigned reads (less than 20, only entries with V_errors > 0 considered) were excluded, for each individual separately.

2.4 Extraction of 5’UTR-leader sequences

113 After data filtering, we extracted 5'UTR-leader reads, grouped them according to inferred IGHV
114 germline gene allele, and inferred 5'UTR-leader sequences for each analyzed individual and allele
115 (Figure 1). Sequences were built one position at a time, starting at the 3' end, by extracting any
116 nucleotide present in at least 30% of the reads. Whenever two nucleotides met this threshold, the
117 reads were split accordingly and analyzed separately, starting over from the 3' end. Length of inferred
118 5'UTR-leader sequences were set so that at least 50% of the underlying data covered the 5' most
119 inferred base.
120 We subsequently summarized 5'UTR-leader sequences of all analyzed individuals, each IGHV
121 allele separately, and counted their frequencies (Figure 1). For sequences that were identical in
122 different individuals, except with respect to how far in the 5' direction they stretched, the length was
123 set so that at least 80% of the inferred sequences would be of the same length or longer than the
124 consensus sequence. Six upstream region sequences were removed from the output data. The
125 majority of these expressed one extra base in a homopolymeric stretch, a type of region where
126 sequencing insertion errors are not uncommonly seen (22), in either the leader region (three
127 sequences), thus resulting in frame shift, or in the 5'UTR region (one sequence). One of the other two
128 removed sequences showed remains of an adaptor sequence that had not been removed by the pre-
129 processing trimming step, and the other showed low CDR3 diversity.
130 The upstream regions sequences are numbered in a way that assigns the last base of the leader
131 sequence as base -1. Most upstream regions encode a 19 amino acids long signal peptide.
132 Consequently, the initiation ATG codon will (with the exception of IGHV3-64*01 and IGHV6-1*01) be
133 represented by bases -57 – -55, and the 5'UTR will extend beyond base -57.

134 **2.5 Validation of alleles by haplotype inference and CDR3-length distribution analysis**

135 Haplotype analysis was performed for all 35 of the 98 subjects that are heterozygous in the IGHJ6
136 gene, by calculating the frequency of 5'UTR-leader sequences found in transcripts derived from each
137 allele of the IGHJ gene (Figure 1). The haplotype inference was conducted for alleles that have
138 heterozygous 5'UTR-leader-IGHV allele sequences for the subjects mentioned above, as well as for
139 some additional genes with novel inferred alleles. Clonal diversity of each inferred 5'UTR-leader
140 sequence was examined, by plotting the distribution pattern of amino acid lengths of associated
141 CDR3's, using filtered IgDiscover data (Figure 1).

142 **2.6 Base -93 of upstream regions of genes of the IGHV4 subgroup**

143 To assess the ability of raw, assembled reads to correctly infer base -93 of upstream regions of genes
144 of the IGHV4 subgroup we collected all reads of 8 subjects assembled by PEAR as part of an
145 IgDiscover process (23). These reads were subjected to an IMGT HighV-QUEST analysis process
146 (IMGT/V-QUEST program version: 3.5.18; IMGT/V-QUEST reference directory release: 202011-3).
147 The reads in these data sets that perfectly matched bases -1 – -92 of the inferred upstream region(s),
148 and that were unequivocally assigned by IMGT/HighV-QUEST to one allele of the gene in question,
149 were collected. The occurrences of base T and G (that typically defined variant upstream regions in a
150 recent study (18) of base -93) of reads associated to each haplotype (as defined by alleles of IGJ6)

151 were counted and the ratio of these reads were calculated. For comparison, reads assigned to
152 IGHV1-46*01/*03 were analyzed in the same manner as an example of variability in sets of reads that
153 typically extend well beyond base -93.

154 **2.7 Poorly expressed alleles of IGHV2-70**

155 Assembled reads representing the naïve B cell repertoire of subjects that could be haplotyped based
156 on heterozygocity of IGHJ6 had in the past been generated and subjected to IMGT/HighV-QUEST
157 analysis (23). Rare reads of IGHV2-70 of alleles not directly inferred by IgDiscover were identified, a
158 subset of which was associated to an IGHJ6-defined haplotype that did not express another, more
159 highly expressed allele of this gene. A consensus sequence, also including the 5'UTR-leader
160 sequence, of such reads was identified (23).

161 **2.8 Identification of upstream region of IGHV4-59*12 in data sets SRR5471283 and SRR5471284**

162 Raw read files SRR5471283 and SRR5471284, containing IgM library of donor LP08248, created by
163 5'-RACE technology and sequenced by 454 technology (24), were downloaded from ENA.
164 Sequences of the two sets were merged, and subsequently converted using FASTQ Groomer
165 (version 1.0.4) (25). Leading and trailing bases with a quality below 25 were discarded using
166 Trimmomatic (version 0.32.3) (26). Reads were filtered by quality (quality cut-off value: 25; percent of
167 bases ≥ quality cut-off: >95%), and reads were converted to FASTA files. The resulting reads were
168 used to infer a germline gene repertoire using IgDiscover 0.12 (9). 89 bases of the upstream region of
169 IGHV4-59*12 were identified from this output.

170 **2.9 Comparison to upstream regions of IGHV genes of Rhesus macaques**

171 Upstream regions of functional germline genes of the IGHV1 and IGHV3 subgroups of Rhesus
172 monkey (*Macaca mulatta*) were retrieved from the assembled IMGT000064 entry
173 (http://www.imgt.org/ligmdb/view?id=IMGT000064). These sequences were aligned to inferred
174 upstream regions of human IGHV genes/alleles of the same subgroups. The similarity of IGHV gene
175 coding regions of Rhesus monkeys to human IGHV germline genes was also assessed using IMGT
176 V-QUEST (IMGT/V-QUEST program version: 3.5.24; IMGT/V-QUEST reference directory release:
177 202113-2).

178 **2.10 SNPs and population data**

179 VCF files describing single nucleotide polymorphisms (SNPs) in human population data of the 1000
180 Genomes project (27) were retrieved from the International Genome Sample Resource (Phase 3
181 release, https://www.internationalgenome.org), and data for any variants with global minor allele
182 frequency (MAF) >1% within the analyzed regions were extracted. For genes not defined in the
183 GRCh37 reference genome, but in the GRCh38 reference genome, data were obtained from the
184 Ensembl Genome Browser (releases 102-103; http://www.ensembl.org) (28).

185 **2.11 Linkage disequilibrium**

186 Linkage disequilibrium of alleles of IGHV1-2, IGHV1-3, IGHV4-4, and IGHV7-4-1 has been studied in
187 the past (23). The 5'UTR-leader sequences of several of the alleles that occupy these genes were
188 determined in the present study. Some of the alleles of these genes, however, are very poorly
189 expressed and thus cannot be inferred. The conventional haplotype inference was thus extended by
190 past observations of rare transcripts in these transcriptomes, transcripts that suggest the presence of
191 these poorly alleles in haplotypes that lack expression of other, highly expressed alleles (23). In order
192 to extend the previous analysis of linkage disequilibrium of alleles of IGHV1-2, IGHV1-3, IGHV4-4,
193 and IGHV74-4-1, the expected frequency of each haplotypic combination of these 4 genes was
194 calculated, assuming random association, and compared with the observed frequency of the same
195 combinations. Only haplotypic combinations observed in at least 2 of 70 haplotypes were considered.
196 Calculation of expected frequencies was based on the separate occurrence frequency of each 5'UTR-
197 leader-allelesequencew ithin the studied haplotypes.

198

199 **3 RESULTS**

200 **3.1 Inference and validation of 166 5'UTR-leader sequences by a novel analysis pipeline**

201 In order to address the incomplete representation of 5'UTR and leader sequences of antibody genes
202 in the IMGT database (6), we have examined such sequences in a publicly available antibody
203 transcript data set of 98 individuals (19), also analyzed for the same purpose in the study by
204 Mikocziova et al. (18). Using a strict pre-processing and filtering pipeline followed by extraction of
205 consensus 5'UTR-leader sequences (Figure 1), we identified 166 sequences, found in frequencies
206 ranging from 1 individual to 98 individuals (Figure 2; Supplementary Table 1; Supplementary Data 2).
207 A 5'UTR-leader sequence detected by an inference tool as defined in the present study should
208 feature particular characteristics to be considered valid. Firstly, one would expect that these
209 sequences should be present in a number of different rearrangements, for instance as evidenced by
210 their association to a diversity of lengths of the third complementarity determining region (CDR3).
211 Thus, for each 5'UTR-leader sequence we generated a plot of the number of unmutated reads vs. the
212 length of CDR3 (Figure 3; Supplementary Figure 1), demonstrating that each inferred 5'UTR-leader
213 sequence was associated to a diversity of rearrangements. Secondly, haplotyping offers an important
214 tool to assess the outcome of an inference process (20); the inferred 5'UTR-leader sequences should
215 typically be associated with a single haplotype in subjects that are heterozygous or hemizygous for a
216 given 5'UTR-leader-IGHV gene combination. As illustrated for 5'UTR-leader sequence variants
217 associated to IGHV4-4*02 and IGHV4-4*07 (Table 1), as well as for other 5'UTR-leader IGHV genes
218 that were found in IGHJ6 heterozygous subjects (Supplementary Table 2), this proved to be the case.
219 Thirdly, diversified positions in the 5'UTR-leader sequence of an IGHV gene could also be expected
220 to be represented in genomic data. Population data as described in the Ensembl database
221 (https://www.ensembl.org) has typically been generated by short read sequencing and thereby suffer
222 from important technical caveats that may compromise the correct assembly of complex loci like
223 those representing immunoglobulin germline genes (29). Nevertheless, such data may provide

complementary information to other methods, like sequence inference. Analysis of population data of the 1000 Genome Project (27) confirmed that many of the variants seen in the inferred 5'UTR-leader sequences also were represented in the genomic data (Supplementary Table 1). Altogether these findings support the validity of the inferred 5'UTR-leader sequences.

3.2 Novel IGHV alleles

Several novel IGHV alleles have been inferred from the present data set in the past and validated by sequencing of amplified genomic clones (18). These are now featured in more recent releases of the IMGT human IGHV database. Other alleles, some of which had also been identified in the past study but have not yet been entered into the IMGT database, were also identified in the present study. Some of these have independently been reviewed and provisionally accepted by the Inferred Allele Review Committee (https://www.antibodysociety.org/the-airr-community/airr-subcomittees/inferred-allele-review-committee-iarc/), while other alleles have not been identified in the past. The not yet reviewed inferences (IGHV2-70*04_S5392 [A14G], IGHV3-13*01_S3164 [G290A T300A], IGHV3-30*02_S4989 [G49A], IGHV3-30*04_S7005 [C201T G317A], IGHV3-43D*04_S5432 [G4A], IGHV3-53*02_S9017 [C259T], IGHV3-66*02_S8911 [G303A], and IGHV4-30-2*01_S6723 [G70A]) were validated by haplotyping (when possible), CDR3 length distribution, and frequency of unmutated reads based on VDJBase (30) and IgDiscover (9) analyses (Supplementary Table 3). Their upstream regions are now reported (Figure 2).

3.3 Conserved 5'UTR-leader sequences of multiple IGHV genes

For multiple genes, the inferred 5'UTR-leader sequences were highly conserved among the alleles of the respective gene. Some genes (such as IGHV3-64, IGHV3-72, IGHV3-74, IGHV4-34, and IGHV6-1) were represented by only one allele that all featured one and the same 5'UTR-leader sequence. Furthermore, all alleles of IGHV1-2, IGHV1-46, IGHV1-8, IGHV2-5, IGHV3-11, IGHV3-13, IGHV3-15, IGHV3-20, IGHV3-23, IGHV3-43D, IGHV3-48, IGHV3-49, IGHV3-66, IGHV3-73, IGHV3-9, IGHV4-31, IGHV4-38-2, and IGHV7-4-1 were associated to one, identical 5'UTR-leder sequence/gene in this cohort.

Assessment of population data (excluding IGHV3-30-3, IGHV3-43D, IGHV3-64D, IGHV4-30-2, IGHV4-30-4, and IGHV4-38-2, as these genes are not featured in any of the reference genomes GRCh37 or GRCh38) confirmed that IGHV1-2, IGHV1-46, IGHV1-8, IGHV2-5, IGHV3-13, IGHV3-48, IGHV3-49, IGHV3-66, IGHV3-72, IGHV4-34, IGHV6-1, and IGHV7-4-1 had no diverse residues (with an overall population MAF>1%) within the sequenced part of the 5'UTR-leader (*i.e.* excluding the leader sequence intron). IGHV3-11, IGHV3-15, IGHV3-20, IGHV3-23, IGHV3-73, and IGHV3-74 all had SNPs that carried variability at high frequency in some populations, although not in European populations (Supplementary Table 1). IGHV3-9 and IGHV3-64 however, expressed variants (-60 [A/G], -88 [A/G], -101 [G/C], and -127 [G/A]; and -56 [C/T], respectively) with MAF>1% also in European population, indicating that the 5'UTR-leader sequences of these genes may contain diversity not captured by our study. However, these genomic variants could potentially also be technical artefact resulting from incorrect assembly of the complex IGHV loci, which sometimes

262 accompany short read sequencing (29). Base -56 of IGHV 5'UTR-leader sequence generally holds
263 the T of the initiation ATG codon, but is represented by an C in the herein inferred 5'UTR-leader
264 sequence of IGHV3-64 (as this gene's ATG codon is located in position -60 – -58). Thus, incorrect
265 mapping of reads derived from other IGHV genes, including the duplicate gene IGHV3-64D, to the
266 IGHV3-64 region would indeed result in a technical artifact presented as a -56T variant. Likewise, the
267 upstream region of IGHV3-9 is highly similar to e.g. those of IGHV3-20, IGHV3-43 and IGHV3-43D,
268 the latter of which is not even present in the reference genome. It is certainly conceivable that
269 improper assembly of short reads derived from these other genes to the upstream region of IGHV3-9
270 (Supplementary Figure 2) may contribute to precisely those sequence variants that were defined in
271 Ensembl. Nevertheless, the population-based studies, despite their shortcomings (Watson et al.,
272 2017), generally agreed with the observation of low diversity of these 5'UTR-leader sequences of the
273 herein studied cohort. This analysis, furthermore, also suggested that differences may exist between
274 populations with respect to diversity of the studied upstream region.

275 **3.4 Highly diversified 5'UTR-leader sequences of multiple IGHV genes**

276 The 5'UTR-leader sequences of several genes were diverse even after the stricter pre-processing
277 procedure performed prior to the present analysis. Alleles of many genes (like IGHV1-18, IGHV1-24,
278 IGHV1-3, IGHV1-58, IGHV1-69, IGHV2-26, IGHV2-70, IGHV3-21, IGHV3-7, IGHV3-30, IGHV3-43,
279 IGHV3-53, IGHV3-64D, IGHV4-30-2, IGHV4-30-4, IGHV4-39, IGHV4-4, IGHV4-61, IGHV5-10-1, and
280 IGHV5-51) were diverse in the population of this data set. Population-based studies addressing
281 diversity in the 5'UTR-leader sequence was used to examine these variants further. For a majority of
282 these genes (IGHV1-3, IGHV1-58, IGHV1-69, IGHV2-26, IGHV2-70, IGHV3-21, IGHV3-43, IGHV3-
283 53, IGHV3-7, IGHV4-4, IGHV4-61, and IGHV5-51), all identified SNPs could also be observed in the
284 population data (Supplementary Table 1). Many of these variants could also be further validated with
285 haplotype analysis. Haplotyping could also be performed for two of the 5'UTR-leader sequence
286 variants that could not be identified in population data. IGHV1-24 featured three different 5'UTR-
287 leader sequences with diversity in two positions (-70 [A/G] and -71 [C/T]), with only the latter observed
288 in analyzed population data. Yet, haplotyping of one individual, expressing both IGHV1-24*01-A (-
289 70A, -71C) and IGHV1-24*01-C (-70G, -71C) showed appropriate segregation of these upstream
290 regions between the two haplotypes, supporting the inferences (Supplementary Table 2). Similarly,
291 the diversity of base -30 (G/C) in IGHV4-39 associated 5'UTR-leader sequences could not be
292 confirmed by population studies but is supported by haplotype analysis of one individual expressing
293 IGHV4-39*01-A and IGHV4-39*01-B on different haplotypes.
294 Diversification of 5'UTR-leader sequences can be limited to a single base (*e.g.* for IGHV1-18 and
295 IGHV3-7) or include variability in multiple bases. One of the genes expressing the most diversified
296 5'UTR-leader sequences within the analyzed population is IGHV4-4, which is dominated by two quite
297 different alleles, IGHV4-4*02 and IGHV4-4*07. These alleles together carry diversity located to seven
298 positions (bases -1 [C/T], -31 [C/G], -65 [A/G], -66 [C/T], -74 [A/G], -78 [A/C], and -81 [C/G]), four of
299 which are diverse in both alleles as defined by the present study. This diversity corresponded well to
300 diversity seen in the 5'UTR-leader sequence of the gene as investigated in population studies

301 (Supplementary Table 1). Additionally, haplotype analysis provides further evidence for most of the
302 identified 5'UTR-leader sequences (Supplementary Table 2). Despite the substantial divergence of
303 the two alleles' coding regions, the 5'UTR-leader sequences are similar and several of their diversified
304 5'UTR-leader sequence residues carry similar type of diversification. In all, six different 5'UTR-leader
305 sequences were found associated to each of these alleles of IGHV4-4, several of which were not
306 identified in a previous study of the present data set (Supplementary Figure 3).
307 Some germline genes may, due to their high similarities, be hard to distinguish between in *e.g.*
308 germline gene inferences and population-based studies. One example of such very similar germline
309 genes is IGHV3-30, IGHV3-30-3, IGHV3-30-5, and IGHV3-33. We identified five different 5'UTR-
310 leader sequences among the alleles of these genes with variability in bases -80 (G/T), -103 (G/C), -
311 111 (G/A), and -124 (G/C). The 5'UTR-leader sequences of alleles like IGHV3-30*02, IGHV3-30*18,
312 and IGHV3-33*01 share common sequence features while alleles like IGHV3-30*01, IGHV3-30*04,
313 and IGHV3-30-3*01 shared another set of related sequence features (Supplementary Figure 4).
314 Population-based studies using short read sequencing technology is complicated and error-prone
315 (29), in particular in relation to sets of very similar genes, like these. In any case, analysis of data of
316 these three genes from the 1000 Genome Project provides further evidence for two of the identified
317 variable positions (-80 and -103) of the 5'UTR-leader sequence of IGHV3-30 (Supplementary Table
318 1). One additional SNP (-40 [G/T]) could be identified in the population data of IGHV3-30, but had a
319 low MAF (<1%) in the European population. Another set of highly similar genes is IGHV4-30-2,
320 IGHV4-30-4, and IGHV4-31. The 5'UTR-leader sequences of alleles of these genes are mostly
321 identical. Only 2 and 1 rare sequence variants of these upstream regions were identified in IGHV4-30-
322 2 and IGHV4-30-4, respectively. In contrast to other variants seen in this study two of these
323 sequences represented base deletions, in both cases Δ-69C. Haplotyping of such upstream regions
324 of IGHV4-30-2 was possible using one data set, in which case the haplotype with or without base -69
325 separated onto different haplotypes (Supplementary Table 2), supporting the validity of the inference.
326 The frequency of these transcripts in the data sets suggested that they were expressed at similar
327 levels as those alleles that had not deleted this particular base in the 5'UTR (0.39%±0.05% [n=3] and
328 0.45±0.16% [n=4], respectively). Population-based studies provided further validation of this deletion,
329 as one such variant was identified for IGHV4-31 (Supplementary Table 1).

330 **3.5 5'-terminal Gs in inferred 5'UTR-leader sequences**

331 In contrast to the study by Mikocziova et al. (18), 5'UTR variants with a 5'-terminal G were largely
332 eliminated in our analysis, a direct result of the strict 5' trimming process used. As a consequence we,
333 in several instances, inferred a 5'UTR that was shorter than that identified by Mikocziova et al. (18).
334 For instance, in the case of alleles of IGHV2-5, only one common upstream sequence was identified
335 in the present study, while Mikocziova et al. (18) identified two common, longer upstream sequences
336 for each allele, with only a T/G difference in the 5'-most base (position -75) (Supplementary Figure 5).
337 Population data suggest that base -75 is virtually invariant (T) in human populations (highest
338 population minor allele frequency (MAF)<0.01%), suggesting that an inferred G variant may be a
339 technical artifact. Similarly, in our hands and using a strict pre-processing protocol, 5'UTR-leader

340 sequences with a length of 92 bases were typically inferred for many genes belonging to subgroup
341 IGHV4, while alternative 5'UTR-leader sequence variants that carry either a G or a T at base -93 had
342 been identified for many such genes in the past (18). Again, many of the alleles that had previously
343 been suggested to carry a variant with a 5'-terminal G showed no evidence of such common SNPs in
344 population studies (Supplementary Table 4).
345 To further study the matter of diversity in the 5'-most base of inferred 5'UTRs, we assessed the
346 nature of the raw data generated in the sequencing process and its relation to a possible outcome of
347 the inference process. Assessment through haplotyping of unprocessed reads associated to genes of
348 subgroup IGHV4 frequently demonstrated that sequences carrying both bases at position -93 were in
349 general associated to both haplotypes of each subject (Supplementary Table 5). Such observations
350 indicate that the haplotyped individuals can only be heterozygous in position -93 if these genes are
351 duplicated on both haplotypes, a requirement that is at odds with our current understanding of the
352 locus. Altogether, these investigations suggest that further studies (such as long read sequencing) are
353 required to provide evidence of the existence of many variants of 5'UTR-leader sequences with 5'-
354 terminal Gs.

355 **3.6 Uncommon 5'UTR-leader sequences in the IMGT germline database**

356 It has previously been reported (14,18) that several 5'UTR-leader sequences associated to IGHV
357 germline genes do not correspond to the sequence of the primary entry found in the IMGT database.
358 We confirm this in several cases, such as for IGHV2-5*01, IGHV3-23*01, and IGHV5-51*01
359 (Supplementary Figure 5, Supplementary Table 6). Population data support that the sequences
360 reported by us and others represent the real upstream sequences while the primary entries of the
361 IMGT database are incorrect or represent vary rare sequence variants (MAF<0.01%) not
362 representative of many populations. Interestingly, the common leader sequence of genes like IGHV2-
363 5*01 and IGHV3-23*01 is represented in the IMGT database as secondary sequence entries. Such
364 more representative 5'UTR-leader sequences are however not readily retrieved as one download
365 upstream regions from the database.

366 **3.7 5'UTR-leader sequences as a resource for defining genotype organization**

367 Alleles of IGHV genes are commonly given a name associated to the closest known sequence even
368 when the precise genomic location of these alleles might not be known. Some genes might thus be
369 associated by name to a gene where it does not reside. Upstream regions might provide indications of
370 gene relatedness beyond the sequence of the final product. 5'UTR-leader sequences of all identified
371 alleles of the IGHV4 subgroup identified in the present study were consequently aligned to each
372 other. The sequence of some alleles of IGHV4-4 are very similar to alleles of other genes
373 (Supplementary Figure 6). One of the upstream regions, IGHV4-59*12-A identified in data set
374 ERR2567237, was shown to be most similar to some of those of IGHV4-4. In fact, it was identical to
375 IGHV4-4*02-F and IGHV4-4*07-D (Supplementary Figure 6). IGHV4-59*12 (https://ogrdb.airr-
376 community.org/genotype/32) was originally identified in a data set (24) different from those assessed
377 here. The 5'UTR-leader sequence of IGHV4-59*12 found is this genotype (donor LP08248) differed

378 by one base from IGHV4-59*12-A, and it was identical to that of IGHV4-4*07-E (Supplementary
379 Figure 6). Haplotyping of this genotype suggested that IGHV4-59*12 resided on a haplotype that
380 apparently lacked an allele of IGHV4-4 but had alleles of IGHV4-59 and IGHV4-61. There are thus
381 two instances of IGHV4-59*12 with leader sequences more similar to those of IGHV4-4 than to those
382 of IGHV4-59, and circumstantial evidence through haplotyping that suggests that IGHV4-59*12 might
383 very well be located in IGHV4-4.
384 5'UTR-leader sequence can also provide valuable information that can aid in the understanding
385 how an individual's IGHV loci are composed. For example, one genotype (defined by data set
386 ERR2567264) carries IGHV1-69*02 and IGHV1-69*06, that through haplotyping were shown to
387 segregate onto different haplotypes. Allele IGHV1-69*06 was, however, associated to two different
388 upstream regions (Supplementary Table 2). This finding suggests that allele IGHV1-69*06 may
389 occupy both gene location IGHV1-69 and IGHV1-69D. The inference of the germline gene repertoire
390 of the data set ERR2567237 also demonstrated unusual features, in this case of IGHV4-30-2 and
391 IGHV4-30-4. One allele of IGHV4-30-2 was inferred, but it was associated to three different 5'UTR-
392 leader sequences, while three different alleles of IGHV4-30-4 were inferred. This suggests that both
393 genes are duplicated, either both genes on one haplotype, or one gene on each haplotype.
394 Alternatively, one of the alleles might be located at the site of another gene. Analysis of 5'UTR-leader
395 sequences can thus provide additional evidence of genotype organization, in this case related to
396 duplicated genes, not assessable by analysis of the coding region alone.
397 The part of the locus spanning from IGHV1-69 to IGHV2-70 is highly complex as it commonly
398 harbors a large duplication (Watson et al, 2013) and numerous allelic variants of these genes. The
399 present analysis inferred 8 alleles of IGHV1-69(D) and only 3 alleles of IGHV2-70(D) in 35 subjects in
400 which the IGHV locus could be haplotyped based on heterozygocity of IGHJ6 (20). Assessment of the
401 haplotypes identified in the present investigation identified four main types of expressed gene
402 combinations in this part of the locus, as defined by the coding regions and their upstream sequences
403 (Supplementary Figure 7A, C). IGHV2-70*15 was linked to two different 5'UTRs (Supplementary
404 Figure 7D), that associated to different genomic contexts, with and without the duplication involving
405 IGHV1-69D, IGHV1-69-2, and IGHV2-70D (Supplementary Figure 7A). Genomic sequencing has in
406 the past identified haplotypes resembling some of the differences in upstream regions of genes in this
407 part of the IGHV locus (Supplementary Figure 7B). Future descriptions of haplotypes of different
408 populations will likely be required to understand the diversity of this complex part of the IGHV locus.
409 IGHV1-69-2 and IGHV2-70/70D are commonly expressed at relatively low levels (Gidoni et al,
410 2019), and may thus escape inference in samples with fewer reads and limited sequence complexity.
411 In haplotypes expressing only a single copy of IGHV1-69/69D with upstream region sequence
412 featuring -88A -100G, it was common not to infer an occurrence of IGHV2-70. Detailed analysis of
413 IMGT/HighV-QUEST output of reads of IGHV2-70 nevertheless identified poorly expressed variants of
414 IGHV2-70 in some cases, even alleles that are not currently defined or incomplete in the IMGT
415 database (Supplementary Figure 7E). One of these alleles also carries a variant upstream region
416 sequence (Supplementary Figure 7D) not seen in the other, more highly expressed alleles of this
417 gene. Genomic sequencing has in the past identified a similar allelic variant of IGHV2-70 (GenBank

418 accession number AC242528), an allele that is not yet featured in the IMGT database, that also
419 encoded multiple unusual sequence modifications, including for instance an unusual cysteine in
420 framework 3 (Supplementary Figure 7). Altogether, it is highly likely that at least some subjects carry
421 an IGHV2-70 allele in their genotype that could not be efficiently detected by transcriptome-based
422 sequencing and germline gene inference technology. Future identification and confirmation of such
423 alleles and studies of their functionality will be required to allow us to understand their contribution to
424 human functional antibody repertoires.

425 **3.8 Role of the IGHV4-4*01 5'UTR-leader sequence in the poor expression of this allele**

426 Allele IGHV4-4*01 has recently been identified as being very poorly expressed (23), and
427 consequently difficult to infer using tools like IgDiscover. As a consequence of these technical
428 aspects, it was not detected in the present study. The allele's unusual protein sequence was
429 proposed as the cause of its poor expression. Its 5'UTR-leader sequence (23) differs from the
430 corresponding upstream regions of prototype highly expressed alleles IGHV4-4*02 and IGHV4-4*07
431 as defined in the IMGT database. With the present collection of novel 5'UTR-leader sequences of
432 highly expressed alleles of IGHV4-4, it was possible to further assess the extent whereby these
433 regions might also explain the poor expression of IGHV4-4*01. Indeed, the upstream region of
434 IGHV4-4*01 (23) is identical to IGHV4-4*02-A, an upstream region identified in 7 subjects in the
435 herein investigated data set. This upstream region, in combination with IGHV4-4*02 (0.92%±0.37%)
436 expressed similarly with the other allele of IGHV4-4 (0.90%±0.13%) (n=7) in the same subject,
437 suggesting that this upstream sequence is not responsible for the poor expression of IGHV4-4*01.
438 Furthermore, while transcripts derived from IGHV4-4*01 are largely non-productive (23), transcripts
439 derived from IGHV4-4*02 in combination with the IGHV4-4*02-A upstream region typically encoded
440 an in-frame product. There is thus no evidence to suggest that the herein assessable upstream region
441 of IGHV4-4*01 is responsible for the poor expression of this allele.

442 **3.9 Length differences in the inferable part of the 5'UTR**

443 Insertions and deletions (indels) may serve as markers to assess the evolution of genes (31).
444 Inspection of the 5'UTR of genes belonging to the IGHV3 subgroup suggests that they have evolved
445 by indels resulting in length differences in this region (Supplementary Figure 8A). For instance, alleles
446 of gene IGHV3-43D (but not alleles of the related gene IGHV3-43) all lack bases -65 and -66 of the
447 5'UTR-leader sequence of other alleles. Similarly, alleles of IGHV3-23, IGHV3-30, IGHV3-30-3,
448 IGHV3-53, and IGHV3-66 all lack base -121 of other 5'UTR-leader sequences, while alleles of
449 IGHV3-7, IGHV3-21, and IGHV3-48 lack base -109 present in other 5'UTR-leader sequences. In
450 contrast, all these bases are present in IGHV3-9, IGHV3-11, IGHV3-13, IGHV3-15, IGHV3-20,
451 IGHV3-43, IGHV3-49, IGHV3-64, IGHV3-64D, IGHV3-72, IGHV3-73 and IGHV3-74. It is conceivable
452 that these groups of genes have a common evolutionary history. We also compared the upstream
453 regions of inferred human IGHV genes with the small set of functional genes of Rhesus macaques as
454 defined in the IMGT database entry IMGT000064. We identified a number of length differences in the
455 5'UTR-leader sequences of such functional genes, including such identical to those found in human

genes (Supplementary Figure 9A). Length differences were also observed in the 5'UTR of genes belonging to the IGHV1 subgroup (Supplementary Figure 8B) affecting for instance base -76. In similarity to the case of IGHV3, a similar indel event was observed in the upstream region of Rhesus macaque IGHV1 subgroup genes (Supplementary Figure 9B). The upstream region of macaque allele IGHV1-111*01 carried an indel event identical to that of the upstream region of human gene IGHV1-69-2*01. In addition, the human germline gene most similar to the V domain coding sequence of IGHV1-111*01 was IGHV1-69-2*01. In this case the close similarity of human and macaque IGHV germline genes in terms of indels in their upstream region sequences, was associated to a similarity of the coding sequences of these genes as well.

3.10 Linkage disequilibrium

We have previously identified a possible linkage disequilibrium in the IGHV locus that associates IGHV1-2*05 to IGHV4-4*01 (23). We now extended this finding by assessing the association of alleles and diverse upstream regions of these genes to each other and to alleles of IGHV1-3 and IGHV7-4-1, genes that are located close to each other on chromosome 14. This was made possible by analysis of 35 haplotypable data sets of the herein analyzed set of data. All cases of IGHV1-3*01 with upstream region C, and all cases of IGHV7-4-1*02 were associated to either all cases of poorly expressed alleles IGHV1-2*05 and IGHV4-4*01, or with all cases of IGHV1-2*06 and 6/7 cases of IGHV4-4*02 with upstream region D. These two gene combinations were found at a frequency >300-fold above those expected from the frequencies of these individual alleles/upstream regions, alone (Supplementary Figure 10). Similarly, IGHV1-2*04 and IGHV1-3*01 with upstream region D were mostly associated to IGHV4-4*02 with upstream region C or F, and poorly expressed allele IGHV7-4-1*01, at frequencies >10-fold higher than those excepted from random associations of the same alleles. Finally, IGHV1-2*02 was in most cases (>10 times more often than expected) linked to poorly expressed allele IGHV1-3*02, and IGHV4-4*07 with upstream region E or D while there was no evidence of expression of IGHV7-4-1 in these haplotypes. These conserved combinations of alleles were, however, found to be associated to a diverse set of alleles of more distal genes in the locus. For instance, the linked combination IGHV1-2*06 – IGHV1-3*01 (upstream region C) – IGHV4-4*02 (upstream region D) – IGHV7-4-1*02 was seen in haplotypes that carried IGHJ6*02 and IGHV1-69*02, or IGHV1-69*03 and IGHV1-69*02, or IGHJ6*03 and IGHV1-69*04, or IGHJ6*02 and IGHV1-69*10, or IGHJ6*02 and IGHV1-69*12). This strongly suggests that the observed linkage disequilibrium was not primarily an artifact caused by a close familiar relationship between several study subjects in the cohort, but rather that the gene combination exists in subjects with otherwise highly different IGHV loci. Altogether, although multiple alleles and upstream regions exist in IGHV1-2, IGHV1-3, IGHV4-4, and IGHV7-4-1, these are largely found only in a limited set of combinations in the herein investigated population (Supplementary Figure 10).

4 DISCUSSION

493 The present investigation has, inspired by recent studies (14,18), further investigated 5'UTR-leader
494 sequences of IGHV genes. By exploring a strict 5'-trimming pre-processing procedure we eliminated
495 strings of 5'-terminal Gs introduced during the sequencing library generation process, as these may
496 result in technical inference artefacts. We also provide extensive validation of many inferred
497 sequences in terms of haplotyping, association to rearrangement with a variety of CDR3 length, and
498 genomic evidence. Several variants of the 5'UTR-sequences identified by Mikocziova et al. (18) are
499 confirmed. We also report additional upstream sequences not identified in that study. Importantly,
500 these studies ([18]; this study) collectively indicated that some primary sequences in the IMGT
501 database do not represent common upstream regions of these genes. It is consequently suggested
502 that this database is updated to better represent typical 5'UTR-leader sequences.
503 Past investigations in several cases suggested that alternative 5'UTR sequence variants with a 5'-
504 G were proposed to be common in the investigated population (18). These variants could not be
505 confirmed in the present study. Genomic data, generated largely by short read sequencing, further
506 confirmed that these variants are at most rare in human populations. It is, however, certainly difficult
507 to apply such sequencing technology on a highly repetitive locus like that encoding antibody H chain
508 variable domains (29). This is in particular the case in a sequence discovery setting. Examples of
509 particularly complicated cases, such as the upstream region of IGHV3-9, were identified, highlighting
510 the need for caution when interpreting genomic population data. However, such data may also
511 provide independent, supportive information for validation of more common sequence variants
512 (SNPs), as we have demonstrated in this study. Genomic data indeed support many of the variants
513 we have identified and indicate that additional upstream sequence variants, not identified in the
514 present study of data sets collected in northern Europe, may exist in other populations.
515 Inference analysis cannot provide positional information on inferred sequences. However, inferred
516 sequences may stimulate development of hypotheses that later has to be proven by alternative
517 technologies, such as long-read genomic sequencing (32). In the present study, analysis of upstream
518 regions identifies several genotypes that may represent unusual or previously not well-characterized
519 structures of the IGHV locus. We identified one genotype (data set ERR2567237) that carried three
520 copies of IGHV4-30-2 (as defined by different upstream regions) and three copies of IGHV4-30-4 (as
521 defined by allelic differences in the coding region), suggesting that these closely linked genes may be
522 present in two copies on one haplotype. Furthermore, we defined that allele IGHV1-69*06 may be
523 present (data set ERR2567264) in two copies (with different upstream regions) within a single
524 haplotype. This suggests that this allele, which frequently occurs in combination with IGHV1-
525 69*01/IGHV1-69D*01, tentatively may occupy both the IGHV1-69 and the IGHV1-69D gene.
526 Inference technology also allowed us to identify tentative linkage disequilibrium between alleles and
527 their upstream regions from IGHV1-2 to IGHV7-4-1, genes that are located in close proximity to each
528 other in the IGHV locus. Whenever particular genes/alleles are associated through such
529 disequilibrium and are linked to particular (stereotyped) immune responses, these characteristics may
530 thus be co-inherited. Finally, we found evidence (in data sets ERR2567237, and
531 SRR5471283+SRR5471284) in its 5'UTR-leader sequence that IGHV4-59*12 may reside in a gene
532 different from that suggested by its name, tentatively IGHV4-4. Indeed, haplotyping of data generated

533 from a subject different from those primarily studied here suggest that IGHV4-59*12 is present on a
534 haplotype that also carries IGHV4-59*01 but no allele of IGHV4-4 (https://ogrdb.airr-
535 community.org/genotype/32). Similarly, IGHV4-4*09 has been discovered in a context in which it
536 exists on the same haplotype as IGHV4-4*03 but in the perceived lack of an allele of IGHV4-61
537 (https://ogrdb.airr-community.org/genotype/51). These cases mimic the situation of IGHV4-59*08 that
538 typically is present on the same haplotype as IGHV4-59*01. Such haplotypes commonly lack an allele
539 of IGHV4-61. It has furthermore been proposed that IGHV4-59*08 is associated to non-coding regions
540 more similar to those of alleles of IGHV4-61 than to those of other alleles of IGHV4-59 (33).
541 Altogether these studies suggest that IGHV4-59*08 might be located to the IGHV4-61 gene location.
542 Although not proof of these alleles' location in the locus, findings through inference certainly stimulate
543 debate on the organization of the locus, and the principles of allele naming that are currently in use
544 (10,34,35).
545 Upstream regions of alleles/genes belonging to the same IGHV subgroup tend to be similar,
546 suggesting a common origin. Different genes are though different in terms of diversity of the alleles
547 and its upstream regions. For instance, IGHV1-2 of the present data set features a number of slightly
548 different, highly expressed alleles (IGHV1-2*02, IGHV1-2*04, IGHV1-2*06, and IGHV1-2*07), and one
549 poorly expressed allele (IGHV1-2*05) (23), but they are all associated to the very same upstream
550 region. In contrast, IGHV4-4 is dominated by two quite different alleles (IGHV4-4*02 and IGHV4-4*07;
551 only 92% base identity, and difference in the length of CDRH1) that show more similarity to alleles of
552 other genes of the IGHV4 subgroup than to each other. These two very different alleles of IGHV4-4
553 are associated to six upstream regions each with similar sequence features, one of which are even
554 shared between them (Supplementary Figure 3). Is the upstream region of some genes like IGHV1-2
555 less amendable to diversification than the upstream region of IGHV4-4? Was IGHV4-4 populated by
556 independent duplications of other genes, or have processes like gene conversion contributed to the
557 present diversity of this gene and its alleles, and the similarity of their associated upstream regions?
558 Future phylogenetic and experimental studies are required to address these matters properly.
559 Although small differences in the expression of alleles of an IGHV gene have been identified,
560 many alleles of a single gene tend to be expressed at similar levels (19,36). Indeed, such similarity is
561 frequently used as a gatekeeper by germline gene inference tools to eliminate inference of sequence
562 variants that are artifacts of PCR and sequencing errors, or somatic hypermutation (37). However, we
563 recently described a number of very poorly expressed alleles (23). These alleles all encoded residues
564 within the variable domain not found in other germline genes. We hypothesized that these alleles
565 were poorly expressed as their encoded product in general would be non-functional. B cells encoding
566 such antibodies would rarely be selected as their products would not be able to participate in a
567 positive selection process. IGHV4-4*01 was one such poorly expressed allele. Assessment of its
568 upstream region as detected in the few transcripts that were present in IGHV-encoding transcriptome
569 suggested that it differed from upstream regions of other alleles of IGHV4-4, as defined in the IMGT
570 database. It was thus plausible that this region associated to IGHV4-4*01 is responsible for the
571 allele's low level of expression. However, we herein demonstrated that the upstream region of IGHV4-
572 4*01 assessable by analysis of IGHV transcripts amplified by 5'-RACE methodology is identical to that

573 of a subset of the well-expressed allele IGHV4-4*02. Through this analysis approach we were able to
574 extend the support for the hypothesis (23) that IGHV4-4*01 is poorly expressed, not as a
575 consequence of the upstream region' sequence, but as a consequence of a compromised ability of its
576 encoded protein product to form a folded protein.
577 Interestingly, although similar within subgroups, the 5'UTR-leader sequences show some
578 differences not only in sequence but also in terms of sequence length, differences that may relate to
579 insertion and deletion events. For instance, the upstream regions of IGHV3-43 and its duplicated
580 variant IGHV3-43D, differ by the absence of two bases in the 5'UTR of the latter gene. This difference
581 has previously been used to support the naming of previously undefined, inferred allele IGHV3-
582 43D*04, as an allele of IGHV3-43D and not of IGHV3-43 (38). This allele has independently been
583 demonstrated to reside at IGHV3-43D through sequencing of a fosmid clone
584 (http://www.imgt.org/ligmdb/view?id=AC242184). Upstream regions of some alleles of IGHV genes
585 have thus been proven to contain information that can be used to build valid hypothesis about their
586 location in the genome. Other genes of the IGHV3 subgroup differ in the length of their 5'UTR.
587 Indeed, three other major sets of upstream regions that differ by the presence of perceived
588 insertion/deletion events have been identified. Some of the genes grouped together based on the
589 similarity of indels in their upstream regions are quite similar in their coding region while others are
590 quite different in this respect (e.g. IGHV3-21 and IGHV3-48 vs. IGHV3-7, genes that all lack base -
591 109 of the 5'UTR-leader sequence). We propose that the presence of these indels events may
592 identify genes with a common evolutionary history. Intriguingly, identical insertion/deletion differences
593 as those found in the upstream regions of human IGHV1 and IGHV3 genes were identified in a limited
594 IMGT-database-defined set of functional germline genes of *Macaca mulatta* (Rhesus macaque).
595 These findings suggest that either these positions are particularly sensitive to indel events, or that
596 such events might have occurred prior to separation of linages (39) resulting in humans and Rhesus
597 macaques, respectively. As IGHV germline gene repertoires of additional species become available, it
598 might be possible to identify a line of events through which the human IGHV genes and their
599 upstream regions have evolved.
600 In conclusion, we have generated a collection of validated 5'UTR-leader sequences associated to
601 human IGHV genes in a European human population, a set that may be used for future studies of
602 human IGHV genes. Through this effort we also identified SNPs that indicate diversity in these
603 regions that may exist at high frequency in other populations. We also defined upstream region
604 sequences that may have been identified in error in the past. We describe the extent of diversity of
605 such regions in human germline genes, ranging from the invariable upstream regions of alleles of
606 IGHV1-2 to the highly diversified upstream regions of IGHV4-4. Data on upstream regions were used
607 to build hypotheses regarding for instance allele placement in the IGHV locus, in order to promote
608 further studies of the locus' structure. Finally, we used length differences in upstream regions of IGHV
609 genes to postulate a model of the gene's phylogenetic relatedness.

610

611 **AVAILABILITY**

612 Raw sequence data files of IgM-encoding transcriptomes are available from the European Nucleotide
613 Archive as project PRJEB26509. Raw sequence files that represent the transcriptome of subject
614 LP08248 are available from the European Nucleotide Archive with accession numbers SRR5471283
615 and SRR5471284. Code developed in this study is available at https://github.com/yixun-h/5-UTR-
616 leader_Infer.

617

618 **ACKNOWLEDGEMENT**

619 The computations and data storage were enabled by resources provided by the Swedish National
620 Infrastructure for Computing (SNIC) at LUNARC and Swestore, partially funded by the Swedish
621 Research Council through grant agreement no. 2018-05973. Part of the study was conducted,
622 presented, and defended by Yixun Huang as a MSc thesis project entitled "Computational inference
623 and analysis of 5'UTR-leader sequence of alleles from immunoglobulin H chain genes".

624

625 **FUNDING**

626 This study was supported by a grant from The Swedish Research Council [grant number: 2019-
627 01042].

628

629 **AUTHOR CONTRIBUTIONS**

630 Conception of study: LT, MO. Coding: YH, LT; Analysis: YH, LT, MO. Manuscript preparation and final
631 approval: YH, LT, MO.

632

633 **CONFLICT OF INTEREST**

634 The authors declare that they have no conflicts of interest in relation to the present study.

635

636 **ACKNOWLEDGEMENTS**

637 This manuscript has been released as a pre-print at bioRxiv (https://www.biorxiv.org/) (40)

638

639 **REFERENCES**

640 1. Xu, J.L. and Davis, M.M. (2000) Diversity in the CDR3 Region of VH Is Sufficient for Most
641 Antibody Specificities. *Immunity.*, 13, 37-45. doi: 10.1016/s1074-7613(00)00006-6

642 2. Avnir, Y., Watson, C.T., Glanville, J., Peterson, E.C., Tallarico, A.S., Bennett, A.S. et al. (2016)
643 IGHV1-69 polymorphism modulates anti-influenza antibody repertoires, correlates with IGHV
644 utilization shifts and varies by ethnicity. *Sci Rep.*, 6, 20842. doi: 10.1038/srep20842
645 3. Sangesland, M., Yousif, A.S., Ronsard, L., Kazer, S.W., Zhu, A.L., Gatter, G.J. et al. (2020) A
646 Single Human V(H)-gene Allows for a Broad-Spectrum Antibody Response Targeting Bacterial
647 Lipopolysaccharides in the Blood. *Cell Rep*., 32, 108065. doi: 10.1016/j.celrep.2020.108065
648 4. Benichou, J., Ben-Hamo, R., Louzoun, Y. and Efroni, S. (2012) Rep-Seq: uncovering the
649 immunological repertoire through next-generation sequencing. *Immunology*, 135, 183-191.
650 doi: 10.1111/j.1365-2567.2011.03527.x
651 5. Yaari, G. and Kleinstein, S.H. (2015) Practical guidelines for B-cell receptor repertoire
652 sequencing analysis. *Genome Med.*, 7, 121. doi: 10.1186/s13073-015-0243-2
653 6. Giudicelli, V., Chaume, D. and Lefranc, M.-P. (2005) IMGT/GENE-DB: a comprehensive
654 database for human and mouse immunoglobulin and T cell receptor genes. *Nucleic Acids Res.*,
655 33, D256-D261. doi: 10.1093/nar/gki010
656 7. Wang, Y., Jackson, K.J.L., Sewell, W.A. and Collins, A.M. (2008) Many human immunoglobulin
657 heavy-chain IGHV gene polymorphisms have been reported in error. *Immunol Cell Biol.*, 86,
658 111-115. doi: 10.1038/sj.icb.7100144
659 8. Rodriguez, O.L., Gibson, W.S., Parks, T., Emery, M., Powell, J., Strahl, M. et al. (2020) A Novel
660 Framework for Characterizing Genomic Haplotype Diversity in the Human Immunoglobulin
661 Heavy Chain Locus. *Front Immunol.*, 11, 2136. doi: 10.3389/fimmu.2020.02136
662 9. Corcoran, M.M., Phad, G.E., Bernat, N.V., Stahl-Hennig, C., Sumida, N., Persson, M.A.A. et al.
663 (2016) Production of individualized V gene databases reveals high levels of immunoglobulin
664 genetic diversity. *Nat Commun.*, 7, 13642. doi: 10.1038/ncomms13642
665 10. Ohlin, M., Scheepers, C., Corcoran, M., Lees, W.D., Busse, C.E., Bagnara, D. et al. (2019)
666 Inferred Allelic Variants of Immunoglobulin Receptor Genes: A System for Their Evaluation,
667 Documentation, and Naming. *Front Immunol.*, 10, 435. doi: 10.3389/fimmu.2019.00435
668 11. Ralph, D.K. and Matsen, F.A.I.V. (2019) Per-sample immunoglobulin germline inference from
669 B cell receptor deep sequencing data. *PLoS Comput Biol.*, 15, e1007133. doi:
670 10.1371/journal.pcbi.1007133
671 12. Gadala-Maria, D., Gidoni, M., Marquez, S., Vander Heiden, J.A., Kos, J.T., Watson, C.T. et al.
672 (2019) Identification of Subject-Specific Immunoglobulin Alleles From Expressed Repertoire
673 Sequencing Data. *Front Immunol.*, 10, 129. doi: 10.3389/fimmu.2019.00129
674 13. Lovett, P.S. and Rogers, E.J. (1996) Ribosome regulation by the nascent peptide.
675 *Microbiological Rev.*, 60, 366-385. doi: 10.1128/mr.60.2.366-385.1996
676 14. Zhu, Y., Yang, X., Wu, J., Tang, H., Wang, Q., Guan, J. et al. (2020) Antibody Upstream Sequence
677 Diversity and Its Biological Implications Revealed by Repertoire Sequencing. *bioRxiv*, doi:
678 https://doi.org/10.1101/2020.09.02.280396, 3 September 2020, pre-print: not peer-reviewed.
679 15. Wellensiek, B.P., Larsen, A.C., Flores, J., Jacobs, B.L. and Chaput, J.C. (2013) A leader sequence
680 capable of enhancing RNA expression and protein synthesis in mammalian cells. *Protein Sci.*,
681 22, 1392-1398. doi: 10.1002/pro.2325
682 16. Saintamand, A., Vincent-Fabert, C., Marquet, M., Ghazzaui, N., Magnone, V., Pinaud, E. et al.
683 (2017) E_{μ} and 3'RR IgH enhancers show hierarchic unilateral dependence in mature B-cells.
684 *Scientific Rep.*, 7, 442. doi: 10.1038/s41598-017-00575-0
685 17. Alamyar, E., Duroux, P., Lefranc, M-P. and Giudicelli, V. (2012) IMGT(®) tools for the nucleotide
686 analysis of immunoglobulin (IG) and T cell receptor (TR) V-(D)-J repertoires, polymorphisms,
687 and IG mutations: IMGT/V-QUEST and IMGT/HighV-QUEST for NGS. *Methods Mol Biol.*, 882,
688 569-604. doi: 10.1007/978-1-61779-842-9_32
689 18. Mikocziova, I., Gidoni, M., Lindeman, I., Peres, A., Snir, O., Yaari, G. and Sollid, L.M. (2020)
690 Polymorphisms in human immunoglobulin heavy chain variable genes and their upstream
691 regions. *Nucleic Acids Res.*, 48, 5499-5510. doi: 10.1093/nar/gkaa310

19. Gidoni, M., Snir, O., Peres, A., Polak, P., Lindeman, I., Mikocziova, I. et al. (2019) Mosaic deletion patterns of the human antibody heavy chain gene locus shown by Bayesian haplotyping. *Nat Commun.*, 10, 628. doi: 10.1038/s41467-019-08489-3
20. Kirik, U., Greiff, L., Levander, F. and Ohlin, M. (2017) Parallel antibody germline gene and haplotype analyses support the validity of immunoglobulin germline gene inference and discovery. *Mol Immunol.*, 87, 12-22. doi: 10.1016/j.molimm.2017.03.012
21. Vander Heiden, J.A., Yaari, G., Uduman, M., Stern, J.N., O'Connor, K.C., Hafler, D.A. et al. (2014) pRESTO: a toolkit for processing high-throughput sequencing raw reads of lymphocyte receptor repertoires. *Bioinformatics*., 30, 1930-1932. doi: 10.1093/bioinformatics/btu138
22. Minoche, A.E., Dohm, J.C. and Himmelbauer, H. (2011) Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and genome analyzer systems. *Genome Biol*., 12, R112. doi: 10.1186/gb-2011-12-11-r112
23. Ohlin, M. (2021) Poorly Expressed Alleles of Several Human Immunoglobulin Heavy Chain Variable Genes are Common in the Human Population. *Front Immunol*., 11, 603980. doi: 10.3389/fimmu.2020.603980
24. Sheng, Z., Schramm, C.A., Kong, R., N.C.S.P., Mullikin, J.C., Mascola, J.R. et al. (2017) Gene-Specific Substitution Profiles Describe the Types and Frequencies of Amino Acid Changes during Antibody Somatic Hypermutation. *Front Immunol*., 8, 537. doi: 10.3389/fimmu.2017.00537
25. Blankenberg, D., Gordon, A., Von Kuster, G., Coraor, N., Taylor, J., Nekrutenko, A. and Galaxy, T. (2010) Manipulation of FASTQ data with Galaxy. *Bioinformatics*., 26, 1783-1785. doi: 10.1093/bioinformatics/btq281
26. Lohse, M., Bolger, A.M., Nagel, A., Fernie, A.R., Lunn, J.E., Stitt, M. and Usadel, B. (2012) RobiNA: a user-friendly, integrated software solution for RNA-Seq-based transcriptomics. *Nucleic Acids Res*., 40, W622-627. doi: 10.1093/nar/gks540
27. Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O. et al. (2015) A global reference for human genetic variation. *Nature*., 526, 68-74. doi: 10.1038/nature15393
28. Yates, A.D., Achuthan, P., Akanni, W., Allen, J., Allen, J., Alvarez-Jarreta, J. et al. (2020) Ensembl 2020. *Nucleic Acids Res*., 48, D682-D688. doi: 10.1093/nar/gkz966
29. Watson, C.T., Matsen, F.A., Jackson, K.J.L., Bashir, A., Smith, M.L., Glanville, J. et al. (2017) Comment on "A Database of Human Immune Receptor Alleles Recovered from Population Sequencing Data". *J Immunol*., 198, 3371-3373. doi: 10.4049/jimmunol.1700306
30. Omer, A., Shemesh, O., Peres, A., Polak, P., Shepherd, A.J., Watson, C.T. et al. (2020) VDJbase: an adaptive immune receptor genotype and haplotype database. *Nucleic Acids Res*., 48, D1051-D1056. doi: 10.1093/nar/gkz872
31. Simmons, M.P., Ochoterena, H. and Carr, T.G. (2001) Incorporation, Relative Homoplasy, and Effect of Gap Characters in Sequence-Based Phylogenetic Analyses. *Syst Biol*., 50, 454-462.
32. Ford, M., Haghshenas, E., Watson, C.T. and Sahinalp, S.C. (2020) Genotyping and Copy Number Analysis of Immunoglobulin Heavy Chain Variable Genes Using Long Reads. *iScience*., 23, 100883. doi: 10.1016/j.isci.2020.100883
33. Parks, T., Mirabel, M.M., Kado, J., Auckland, K., Nowak, J., Rautanen, A. et al. (2017) Association between a common immunoglobulin heavy chain allele and rheumatic heart disease risk in Oceania. *Nat Commun*., 8, 14946. doi: 10.1038/ncomms14946
34. Busse, C.E., Jackson, K.J.L., Watson, C.T. and Collins, A.M. (2019) A Proposed New Nomenclature for the Immunoglobulin Genes of Mus musculus. *Front Immunol*., 10, 2961. doi: 10.3389/fimmu.2019.02961
35. Allele. IMGT®, the international ImMunoGeneTics information system®, http://www.imgt.org/IMGTindex/allele.php [Accessed June 8, 2021].
36. Boyd, S.D., Gaëta, B.A., Jackson, K.J., Fire, A.Z., Marshall, E.L., Merker, J.D. et al. (2010) Individual Variation in the Germline Ig Gene Repertoire Inferred from Variable Region Gene Rearrangements. *J Immunol*., 184, 6986-6992. doi: 10.4049/jimmunol.1000445

743 37. Gadala-Maria, D., Yaari, G., Uduman, M. and Kleinstein, S.H. (2015) Automated analysis of
744 high-throughput B-cell sequencing data reveals a high frequency of novel immunoglobulin V
745 gene segment alleles. *Proc Natl Acad Sci U S A*., 112, E862-E870. doi:
746 10.1073/pnas.1417683112
747 38. Thörnqvist, L. and Ohlin, M. (2018) Critical steps for computational inference of the 3′ -end
748 of novel alleles of immunoglobulin heavy chain variable genes - illustrated by an allele of
749 IGHV3-7. *Mol Immunol*., 103, 1-6. doi: 10.1016/j.molimm.2018.08.018
750 39. Gibbs, R.A., Rogers, J., Katze, M.G., Bumgarner, R., Weinstock, G.M., Mardis, E.R. et al. (2007)
751 Evolutionary and Biomedical Insights from the Rhesus Macaque Genome. *Science*., 316, 222-
752 234. doi: 10.1126/science.1139247
753 40. Huang, Y., Thörnqvist, L. and Ohlin, M. (2021) Computational inference, validation, and
754 analysis of 5'UTR-leader sequences of alleles of immunoglobulin heavy chain variable genes.
755 *bioRxiv*, 2021.06.10.447679; doi: 10.1101/2021.06.10.447679

756

757 **FIGURES LEGENDS**

758

759 **Figure 1.** Schematic illustration of the pre-processing of Illumina MiSeq paired-end reads and of the
760 pipeline of 5'UTR-leader sequences inference and validation process.

761 **Figure 2.** Overarching 5'UTR-leader sequence germline data set inferred in the present study. In
762 addition, upstream regions of IGHV1-3*02 and IGHV4-4*01 have been identified in a separate study
763 (23).

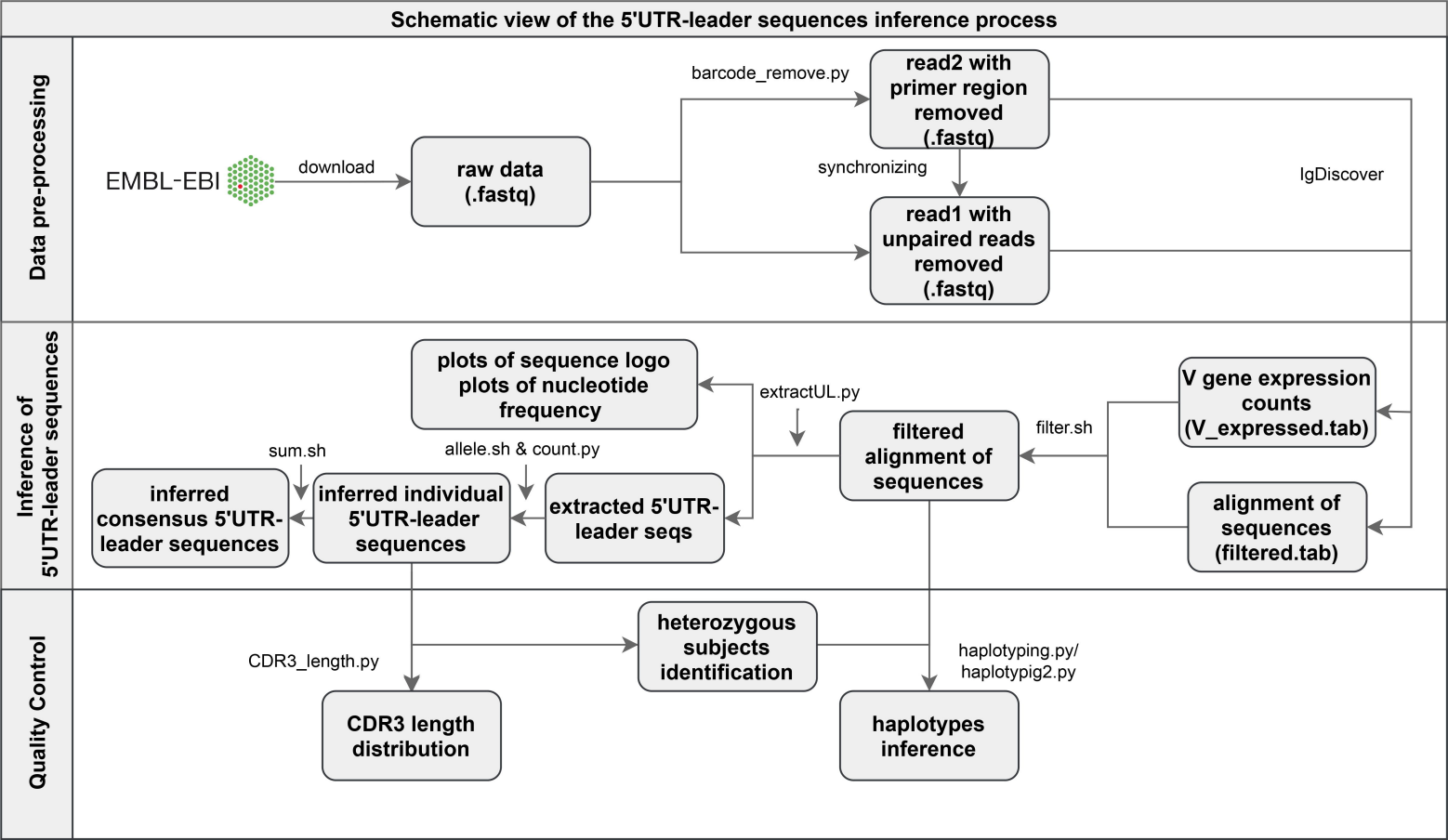
764 **Figure 3.** Distribution patterns of CDR3 length encoded by transcripts associated to 5'UTR-leader
765 sequences of (A) IGHV4-4*02, (B) IGHV4-4*07. For each 5'UTR-leader sequence of a specific allele,
766 the number of filtered reads in each length of CDR3 was counted to create the plots. Every line in the
767 plots represents the 5'UTR-leader sequence from one subject (at maximum 8 subjects were included
768 in each plot). Distribution patterns of CDR3 length for 5'UTR-leader sequences of other alleles are
769 displayed in Supplementary Figure 1.

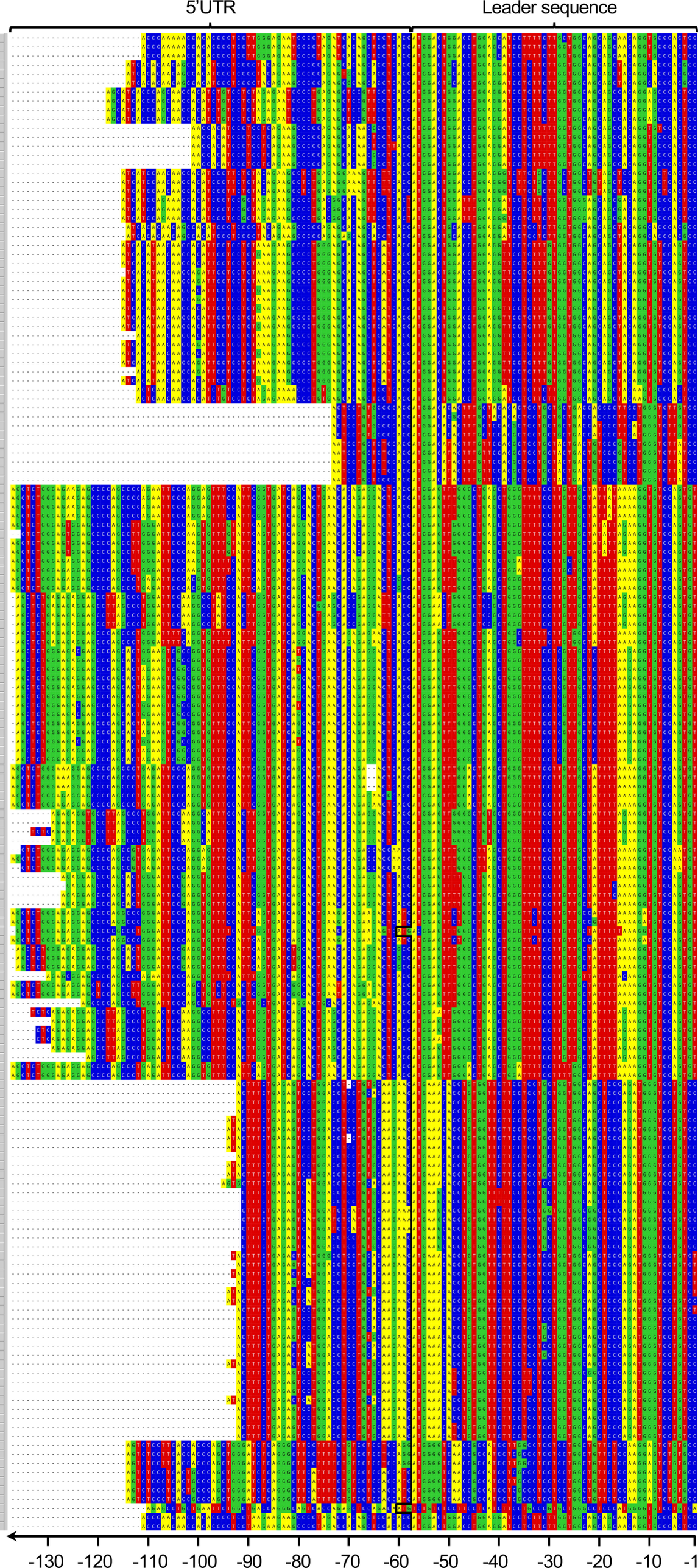
770

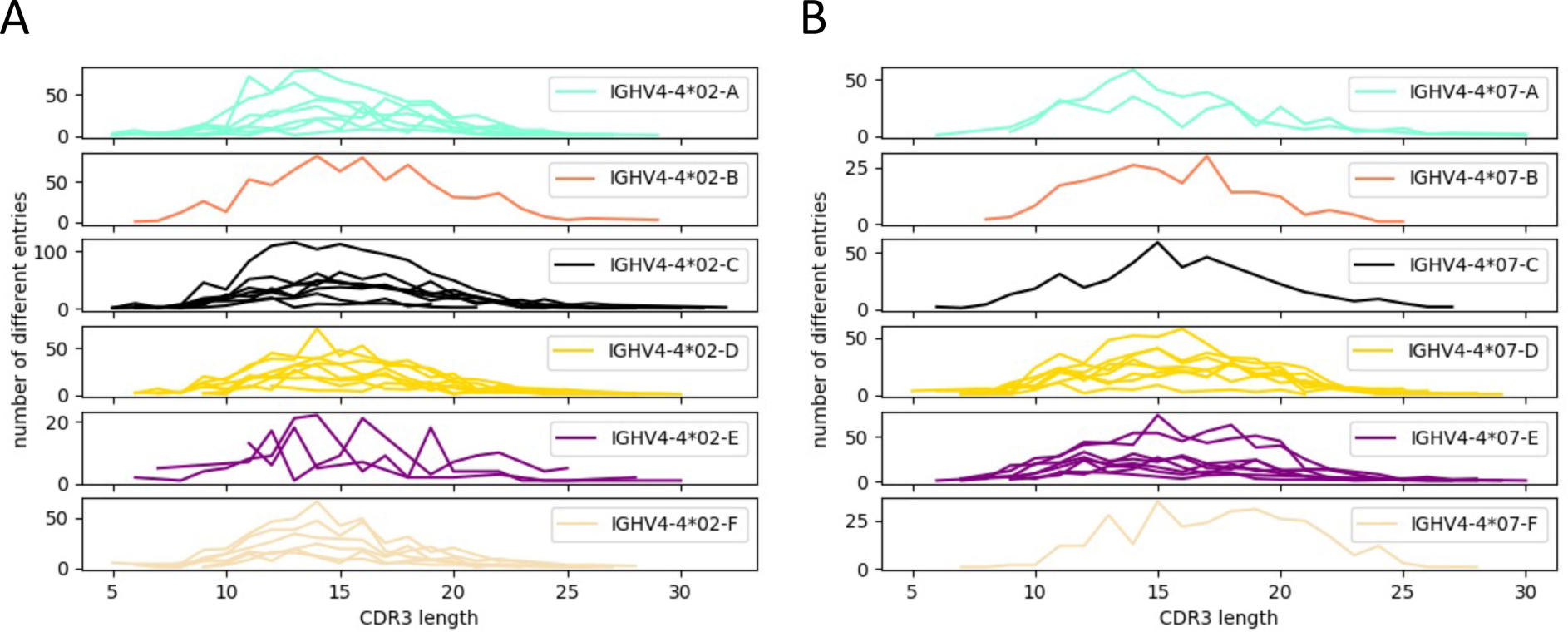
771 **Table 1.** Haplotyping to support the validity of diverse 5'UTR-leader sequence of allele IGHV4-4*02
772 and IGHV4-4*07. The sequence counts of 5'UTR-leader sequences of alleles of IGHV4-4 associated
773 to different alleles of IGHJ6 in rearranged sequences. Haplotyping data for other 5'UTR-leader
774 sequences are available in Supplementary Table 2.

Data set	IGHV gene and upstream sequence	IGHJ6 read distribution	
		IGHJ6*02	IGHJ6*03
ERR2567266	IGHV4-4*02-C	58	0
	IGHV4-4*07-A	1	107
ERR2567189	IGHV4-4*02-F	0	24
	IGHV4-4*07-E	37	0
ERR2567200	IGHV4-4*02-C	0	46
	IGHV4-4*07-B	48	0
ERR2567230	IGHV4-4*02-A	0	38
	IGHV4-4*07-D	72	0
ERR2567192	IGHV4-4*02-C	16	0
	IGHV4-4*02-A	0	17
ERR2567204	IGHV4-4*02-C	74	0
	IGHV4-4*02-D	0	75
ERR2567246	IGHV4-4*02-F	0	65
	IGHV4-4*02-C	36	0
ERR2567254	IGHV4-4*02-C	55	0
	IGHV4-4*02-F	0	42
ERR2567261	IGHV4-4*02-C	99	0
	IGHV4-4*02-F	0	78
ERR2567271	IGHV4-4*02-D	51	0
	IGHV4-4*02-F	0	5
ERR2567274	IGHV4-4*02-E	24	0
	IGHV4-4*02-C	0	21
ERR2567187	IGHV4-4*02-C	65	0
	IGHV4-4*01	-	-
ERR2567201	IGHV4-4*07-E	27	0
	IGHV4-4*07-D	0	35
ERR2567263	IGHV4-4*07-F	0	94
	IGHV4-4*07-C	84	1

775


Schematic view of the 5'UTR-leader sequences inference process
Data pre-processing
EMBL-EBI
download
raw data (.fastq)
barcode_remove.py
read2 with primer region removed (.fastq)
synchronizing
read1 with unpaired reads removed (.fastq)
IgDiscover
Inference of 5'UTR-leader sequences
plots of sequence logo
plots of nucleotide frequency
extractUL.py
filtered alignment of sequences
filter.sh
V gene expression counts (V_expressed.tab)
alignment of sequences (filtered.tab)
sum.sh
allele.sh & count.py
inferred consensus 5'UTR-leader sequences
inferred individual 5'UTR-leader sequences
extracted 5'UTR-leader seqs
Quality Control
heterozygous subjects identification
CDR3_length.py
haplotyping.py/ haplotypig2.py
CDR3 length distribution
haplotypes inference


5'UTR
Leader sequence
IGHV1-18*01-AI2
IGHV1-18*01-BI94
IGHV1-18*04-AI31
IGHV1-24*01-AI2
IGHV1-24*01-BI30
IGHV1-24*01-CI91
IGHV1-2*02-AI68
IGHV1-2*04-AI64
IGHV1-2*05-AI1
IGHV1-2*06-AI19
IGHV1-3*01-AI1
IGHV1-3*01-BI9
IGHV1-3*01-CI26
IGHV1-3*01-DI57
IGHV1-3*01_S6816-AI2
IGHV1-46*01-AI90
IGHV1-46*03-AI49
IGHV1-46*04-AI2
IGHV1-58*01-AI61
IGHV1-58*02-AI23
IGHV1-58*02-BI16
IGHV1-69-2*01-AI1
IGHV1-69-2*01-BI29
IGHV1-69*01-AI76
IGHV1-69*01-BI1
IGHV1-69*01_S8909-AI1
IGHV1-69*02-AI45
IGHV1-69*04-AI24
IGHV1-69*04-BI2
IGHV1-69*04_S3852-AI2
IGHV1-69*06-AI40
IGHV1-69*06-BI2
IGHV1-69*06_S0471-AI1
IGHV1-69*09-AI2
IGHV1-69*09-BI1
IGHV1-69*09-CI5
IGHV1-69*10-AI2
IGHV1-69*12-AI3
IGHV1-69*17-AI2
IGHV1-8*01-AI77
IGHV1-8*03-AI2
IGHV2-26*01-AI98
IGHV2-26*02_S3803-AI1
IGHV2-5*01-AI57
IGHV2-5*02-AI82
IGHV2-70*01-AI63
IGHV2-70*04-AI26
IGHV2-70*04_S5392-AI1
IGHV2-70*15-AI6
IGHV2-70*15-BI36
IGHV3-11*01-AI76
IGHV3-11*04-AI2
IGHV3-11*05-AI22
IGHV3-11*06-AI47
IGHV3-13*01-AI61
IGHV3-13*01_S3164-AI8
IGHV3-13*04-AI19
IGHV3-13*05-AI22
IGHV3-15*01-AI96
IGHV3-15*07-AI25
IGHV3-20*01-AI53
IGHV3-20*04-AI30
IGHV3-21*01-AI95
IGHV3-21*01-BI15
IGHV3-21*01_S4935-AI1
IGHV3-21*01_S5913-AI1
IGHV3-23*01-AI97
IGHV3-23*04-AI20
IGHV3-30-3*01-AI69
IGHV3-30*01-AI11
IGHV3-30*02-AI3
IGHV3-30*02-BI6
IGHV3-30*02_S4989-AI1
IGHV3-30*03-AI5
IGHV3-30*04-AI6
IGHV3-30*04-BI12
IGHV3-30*04_S7005-AI1
IGHV3-30*18-AI92
IGHV3-30*19_S5956-AI1
IGHV3-33*01-AI96
IGHV3-33*01_S3418-AI1
IGHV3-43D*03-AI15
IGHV3-43D*04-AI13
IGHV3-43D*04_S5432-AI1
IGHV3-43*01-AI73
IGHV3-43*02-AI11
IGHV3-48*01-AI50
IGHV3-48*02-AI65
IGHV3-48*03-AI34
IGHV3-48*04-AI10
IGHV3-49*03-AI59
IGHV3-49*04-AI37
IGHV3-49*05-AI60
IGHV3-53*01-AI68
IGHV3-53*02-AI49
IGHV3-53*02_S9017-AI1
IGHV3-53*04-AI29
IGHV3-64D*06-AI52
IGHV3-64D*06_S5429-AI12
IGHV3-64*01-AI46
IGHV3-64*05_S2482-AI10
IGHV3-66*01-AI42
IGHV3-66*02-AI29
IGHV3-66*02_S8911-AI1
IGHV3-72*01-AI47
IGHV3-73*01-AI51
IGHV3-73*02-AI64
IGHV3-74*01-AI98
IGHV3-7*01-AI73
IGHV3-7*01-BI2
IGHV3-7*03-AI44
IGHV3-7*03-BI1
IGHV3-7*03_S9833-AI11
IGHV3-7*04-AI21
IGHV3-9*01-AI77
IGHV3-9*03-AI2
IGHV4-30-2*01-AI3
IGHV4-30-2*01-BI1
IGHV4-30-2*01-CI69
IGHV4-30-2*01_S6723-AI1
IGHV4-30-4*01-AI62
IGHV4-30-4*01_S5061-AI2
IGHV4-30-4*07-AI2
IGHV4-30-4*07-BI1
IGHV4-30-4*08-AI1
IGHV4-31*01-AI9
IGHV4-31*03-AI89
IGHV4-34*01-AI98
IGHV4-38-2*01-AI23
IGHV4-38-2*02-AI19
IGHV4-39*01-AI2
IGHV4-39*01-BI92
IGHV4-39*01_S7498-AI1
IGHV4-39*07-AI13
IGHV4-39*07_S0654-AI2
IGHV4-4*02-AI7
IGHV4-4*02-BI1
IGHV4-4*02-CI57
IGHV4-4*02-DI22
IGHV4-4*02-EI3
IGHV4-4*02-FI6
IGHV4-4*02_S2599-AI1
IGHV4-4*07-AI2
IGHV4-4*07-BI1
IGHV4-4*07-CI1
IGHV4-4*07-DI8
IGHV4-4*07-EI55
IGHV4-4*07-FI1
IGHV4-59*01-AI1
IGHV4-59*01-BI98
IGHV4-59*08-AI42
IGHV4-59*12-AI1
IGHV4-61*01-AI63
IGHV4-61*01_S9413-AI1
IGHV4-61*02-AI30
IGHV4-61*02_S0442-AI6
IGHV5-10-1*01-AI22
IGHV5-10-1*03-AI43
IGHV5-10-1*03-BI14
IGHV5-51*01-AI87
IGHV5-51*01-BI12
IGHV5-51*03-AI40
IGHV5-51*03-BI1
IGHV6-1*01-AI95
IGHV7-4-1*01-AI1
IGHV7-4-1*02-AI26
-130
-120
-110
-100
-90
-80
-70
-60
-50
-40
-30
-20
-10
-1


A
B
IGHV4-4*02-A
IGHV4-4*02-B
IGHV4-4*02-C
IGHV4-4*02-D
IGHV4-4*02-E
IGHV4-4*02-F
IGHV4-4*07-A
IGHV4-4*07-B
IGHV4-4*07-C
IGHV4-4*07-D
IGHV4-4*07-E
IGHV4-4*07-F
number of different entries
CDR3 length