bioRxiv preprint doi: <https://doi.org/10.1101/2020.09.02.280396>; this version posted September 3, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

¹State Key Laboratory of Organ Failure Research, National Clinical Research Center for Kidney Disease, Division of Nephrology, Nanfang Hospital, Southern Medical University, Guangzhou, 510515, China.

²Department of Bioinformatics, School of Basic Medical Sciences, Southern Medical University, Guangzhou 510515, China.

³Center for Precision Medicine, Guangdong Provincial People's Hospital, Guangdong Academy of Medical Sciences, Guangzhou 510080, China.

⁴Key Laboratory of Mental Health of the Ministry of Education, Guangdong-Hong Kong-Macao Greater Bay Area Center for Brain Science and Brain-Inspired Intelligence, Southern Medical University, Guangzhou 510515, China.

⁵Department of Nephrology, Hainan Affiliated Hospital of Hainan Medical College, Haikou 570311, China.

⁶Department of Nephrology, Hainan General Hospital, Haikou 570311, China.

⁷State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-sen University, Guangzhou 510060, China.

⁸School of public health, Sun Yat-sen University, Shenzhen 510006, China.

[†]These authors contributed equally to this work.

^{*}To whom correspondence should be addressed:

Zhenhai Zhang, zhenhaismu@163.com; zhangzhenhai@gdph.org.cn

ORCID:

^a0000-0003-1105-6491 ^b0000-0003-4036-4995 ^c0000-0003-2204-3557

^d0000-0002-5533-7263 ^e0000-0002-2248-0266 ^f0000-0002-9008-9242

29 ^g0000-0001-6759-7639 ^h0000-0002-6720-8215 ⁱ0000-0001-9043-5240

30 ^j0000-0001-8121-7786 ^k0000-0001-5030-8247 ^l0000-0002-3300-8506

31 ^m0000-0002-2000-7053 ⁿ0000-0002-4310-0525

32

33 Abstract

34 The sequence upstream of antibody variable region (Antibody Upstream Sequence, or AUS)
 35 consists of 5' untranslated region (5' UTR) and two leader regions, L-PART1 and L-PART2. The
 36 sequence variations in AUS affect the efficiency of PCR amplification, mRNA translation, and
 37 subsequent PCR-based antibody quantification as well as antibody engineering. Despite their
 38 importance, the diversity of AUSs has long been neglected. Utilizing the rapid amplification of
 39 cDNA ends (5'RACE) and high-throughput antibody repertoire sequencing (Rep-Seq) technique,
 40 we acquired full-length AUSs for human, rhesus macaque (RM), cynomolgus macaque (CM),
 41 mouse, and rat. We designed a bioinformatics pipeline and discovered 2,957 unique AUSs,
 42 corresponding to 2,786 and 1,159 unique sequences for 5' UTR and leader, respectively.
 43 Comparing with the leader records in the international ImMunoGeneTics (IMGT), while 529 were
 44 identical, 313 were with single nucleotide polymorphisms (SNPs), 280 were totally new, and 37
 45 updated the incomplete records. The diversity of AUSs' impact on related antibody biology was
 46 also probed. Taken together, our findings would facilitate Rep-Seq primer design for capturing
 47 antibodies comprehensively and efficiently as well as provide a valuable resource for antibody
 48 engineering and the studies of antibody at the molecular level.

49

50 **Keywords:** antibody upstream sequences, 5' UTR, leader sequences, Rep-Seq, antibody repertoire

51

52 Introduction

53 Antibodies represent an essential class molecule constituting adaptive immune system and
 54 can specifically bind to the invading pathogens for the subsequent eradication or clearance [1].
 55 Rep-Seq technology has enabled antibodies to be interrogated in an unprecedented coverage and
 56 depth, by which researchers can obtain millions to even billions of antibody sequences in a single
 57 experiment [2-4]. The application of Rep-Seq has led to a substantial progress in many fields, such
 58 as aging, tumor immunology, infectious diseases, immune surveillance and neutralizing antibody
 59 screening [5,6].

60 Albeit to the successes mentioned above, the potentials of Rep-Seq can be confined by a
 61 series of not fully addressed issues both experimentally and computationally [5]. One of the
 62 primary computational roadblocks is the unknown germline sequence diversity [7-9]. The
 63 uncaptured diversity will lead to germline misassignment and thus bias the downstream analyses.
 64 Besides, the germline polymorphisms of immunoglobulin loci are found to associate with
 65 expressed antibody repertoire and disease predisposition [10-13]. To capture these diversity, many
 66 tools were developed and employed in antibody repertoire studies [7-9], which then led to the
 67 findings of many novel alleles.

68 Despite the progresses mentioned above, the diversity of AUSs were less interrogated. An
 69 AUS contains two consecutive functional elements, namely 5' UTR and leader, both of which
 70 were found to implicate in mRNA transcription and translation [14-21]. Particularly, leader
 71 sequences play a vital role in antibody expression and are often engineered to improve the
 72 efficiency of monoclonal antibody production [16,21,22]. In Rep-Seq studies, the AUSs are often

73 the targets of the PCR primers for obtaining full-length antibody variable regions [23,24].
74 Furthermore, Mikocziova et al. showed that polymorphisms in AUSs can facilitate the annotation
75 of antibody variable (V) genes [25]. Owing to their importance aforementioned, we decided to
76 interrogate the possible diversity of AUSs which were long neglected in the field.

77 We therefore sequenced the antibody repertoire of both heavy and light chains of 5 species,
78 namely human, rhesus macaque (RM), cynomolgus macaque (CM), mouse and rat. Applying the
79 devised bioinformatic pipeline to the in-house dataset together with available public resources, we
80 discovered thousands of unique AUSs. We then examined their functional relevance to the
81 expression and status of downstream V genes within as well as among species. Our findings here
82 provided the first overview of antibody AUS features of multiple species and enriched the relevant
83 knowledge database which would serve as valuable resources for the community.

84 **Results**

85 **Bioinformatics pipeline for obtaining high-quality AUSs**

86 As shown in Figure 1, we obtained the candidate AUSs through the following six steps: i) V
87 allele variant detection via IgDiscover with initial species-specific databases downloaded from
88 IMGT/GENE-DB [7,26]. The newly discovered genes/alleles were merged with initial databases
89 and carried on for downstream analyses; ii) the sequences from each sample were annotated and
90 assembled via MiXCR (compared in Zhang et al. [27]) and clonotypes were consequently
91 extracted based on CDR3s and V and junctional (J) allele usages; iii) each sample were genotyped
92 via a Bayesian method adapted from TIGGER; iv) the AUS in each clonotype was extracted and
93 dimed as the initial AUS for that particular V allele; v) for each V allele, the consensus of all

94 initial AUSs were calculated and defined as the final AUS(s). Alleles not in the genotype were
 95 excluded in this step; vi) a scoring-based method was employed to retain the most confident AUSs.
 96 Further details on each of these steps can be found in [Materials and Methods](#).

Figure 1

Figure 1. The AUS identification pipeline. The AUS identification pipeline is comprised with six steps, including i) V allele variant detection, ii) sequence annotation and assembly, iii) genotyping, iv) identification of consensus sequences within clonotypes, v) identification of consensus sequences within V alleles, and vi) score-based filtration.

97 **Leaders are more conserved than 5' UTRs**

98 Utilizing the bioinformatics pipeline introduced above, we found 2,957 unique AUSs,
 99 corresponding to 2,786 and 1,159 unique sequences for 5' UTR and leader in all five species,
 100 respectively ([Table 1 and Supplementary Table 1](#)). The percentage of V genes with discovered
 101 AUSs ranged from 7.7% (52 genes for kappa chain of RM) to 56.9% (33 genes for lambda chain
 102 of human). For all chain types across five species, we discovered AUSs for up to 73.9% of the
 103 genes with complete leaders in IMGT/GENE-DB. Of these, 70% human heavy chain AUSs
 104 displayed previously unreported polymorphisms ([Supplementary figure 1 and Supplementary](#)
 105 [Table 2](#)) which may affect Rep-Seq capture efficiency and antibody production. For human V gene
 106 alleles of unknown or incomplete leader according to IMGT/GENE-DB, 43, 26, and 26 (47.3%,
 107 56.5% and 55.3%) AUSs were revealed for heavy, kappa and lambda chain V genes, respectively.

108 Besides, we found the largest number of unique novel sequences for both leader and 5' UTR
 109 in human, probably due to the largest sample size for human among studied species ([Figure 2a and](#)
 110 [Supplementary figure 2a](#)). This demonstrated high diversity of AUSs in human that should not be
 111 neglected and the power of analyzing large dataset. We observed the highest ratio of leader
 112 sequences consistent with known ones in mouse and rat but the lowest in RM. It is worth
 113 mentioning that the 5' UTR sequences curated by IMGT are either partial or contain additional
 114 lengthy intron and upstream regulatory elements [25]. Therefore, a considerable number of 5'
 115 UTRs were part of their counterparts on IMGT or *vice versa*. For CM, all 132 5' UTRs discovered
 116 in this study were novel.

117 For all species, more than half of the leaders for heavy and kappa were 57 bp in length
 118 ([Figure 2b and Supplementary figure 2b](#)). In contrast, the lambda chain exhibited a mode length of

119 60 bp. The length distribution of 5' UTRs were characterized by long tails (for example, ranging
120 from 3 to 133 bp for human heavy chain) and peaked at different lengths for different species.
121 Thus, the conservation of the leader sequence length across species might be a result of its
122 functional importance. Population wise, we found the leaders consistent with IMGT records
123 showed similar frequencies in donors with those different from IMGT, which indicated the
124 reliability of the newly identified AUSs (Figure 2c and Supplementary figure 2c). Individual genes
125 exhibited high diversity with regard to their AUSs. Only less than 5% and around 30% of human
126 V genes used single unique 5' UTRs and leaders, respectively. A single V gene may use up to 44
127 different 5' UTRs (human IGHV1-69) and 11 leader sequences (human IGHV3-53) (Figure 2d).
128 This diversity was consistent in all five species (Supplementary figure 3). The higher diversity of
129 5' UTR compared to leader was also reflected in the combinatorial frequencies, in which 54 out of
130 57 (94.7%) genes have more diverse 5' UTR than leader (Supplementary figure 4), and the more
131 consistent result of leader than 5' UTR with Mikocziova et al. [25] (Supplementary figure 5).
132 Taken together, these results indicated that leader sequences are more conservative than 5' UTRs'
133 and more critical for the functionality of antibodies.

Figure 2. Overview of the discovered AUSs for heavy chain. (a) The composition of discovered 5' UTR and leader sequences. "Consistent" indicates sequences identical to those curated by IMGT/GENE DB. "Unknown" indicates sequences whose corresponding alleles were not provided with available AUSs. "Other" includes the rest four situations, which were illustrated on top of the donut charts. The number of novel sequences is marked in the center of each donut charts. (b) The length distribution of discovered 5' UTR and leader sequences. The length and its corresponding frequency were marked for each peak. (c) The distribution of the number of donors sharing a certain sequence. Different colors in (b) and (c) mean different groups and are consistent with those in (a). (d) Percentage of V genes as a function of the number of discovered 5' UTR and leader sequences.

134

135 The AUSs coevolved with V genes

136 We further investigated the sequence-level similarities among AUSs within and across

137 species (Materials and Methods). Within the same species, we found both 5' UTR and leaders

showed clear family-specific sequence feature for both heavy and light chains (Figure 3a, b and Supplementary figure 6, 7). For human heavy chains, leader sequence sharing was only observed for genes in the same family (Figure 3c, d and Supplementary figure 8). It is also the case for 5' UTR sequences for human light chains (Supplementary figure 8). However, a single 5' UTR sequence (ACC) was shared between VH1 (IGHV1-3, IGHV1-17 and IGHV1-69) and VH3 (IGHV3-13) families (Figure 3c). This family wise sharing of 5' UTRs also exists in nonhuman species (Supplementary figure 9). Thus the AUSs, in general, are V gene family specific. Besides, these within-family interchangeable leaders would also affect the application of elevating V gene assignment accuracy by incorporating upstream sequences.

For each of the nonhuman V alleles, we identified its human counterpart by sequence similarity (Materials and Methods). Then we compared the similarities of nonhuman V gene elements to those of their human counterparts. The pairwise similarities for both 5' UTRs and leaders correlate with V allele similarities. This indicate the AUSs evolved together with V gene alleles (Figure 3e and Supplementary figure 10). Moreover, the correlations of leaders were higher than those of 5' UTRs. This result suggested that AUSs coevolved with V genes and the leaders and V genes were under more similar selective pressure. In details, we found that 5' UTR and leader sequences of RM (73.9% and 90.6% for heavy chain for 5' UTR and leader, respectively) and CM (68.4% and 87.8%) were more similar to human than mouse (44.5% and 62.2%) and rat (45.2% and 60.6%) (Figure 3f and Supplementary figure 11, Materials and Methods). This was consistent with the genomic distances reported before [28-31]. We previously classified heavy chain V genes into two groups, namely core genes and noncore genes³¹, where the former contributes to the vast majority of antibody repertoire and the latter is negligible. We then

160 classified the heavy chain V genes for four nonhuman species and calculated the sequence
 161 similarities between nonhuman and human for V genes, 5' UTRs, and leaders. Although noncore
 162 genes are more functionally inactive, they showed no significant differences in sequence similarity
 163 between human and nonhuman species, except for the V genes of mouse and rat (Figure 3f). This
 164 result indicated that core genes and noncore genes are of equal importance evolutionarily, despite
 165 their difference in contribution to antibody repertoire. Furthermore, we observed a roughly equal
 166 similarity between leaders and V genes, implicating again a coevolution between them.

Figure 3

Figure 3. 5'UTR and leader sequence similarity within human and between species. (a) 5'UTR sequence identity between V genes. (b) Leader sequence identity between V genes. (c) The number of shared 5'UTR sequences between V genes of different families. (d) The number of shared leader sequences between V genes of different families. The side color bar indicates gene families. (e) The correlation between 5'UTR or leader sequence identities and V allele sequence identities between human and four nonhuman species. The line in each scatterplot represents the fitted linear regression model. (f) V gene, 5'UTR and leader sequence identity between human and nonhuman species for antibody heavy chain. Genes were classified into core genes and noncore genes according to Yang et al., 2019. *r*, Pearson's correlation coefficient; *p*, *p* value of the linear model; **, *p*<0.01; ***, *p*<0.001.

167

168 **Leader SNPs are position-dependent and may contribute to functional reversal and**
 169 **inefficient amplification of V genes**

170 We next examined the single nucleotide polymorphisms (SNPs) in the AUSs as they were
 171 reported to affect antibody production [16] and the primer design for Rep-Seq and genetic
 172 predisposition screening [13,24]. Applying a heuristic algorithm, we investigated the ratio of

173 replacement (R) to silent (S) SNPs and positional nucleotide and amino acid diversity index (NDI
174 and ADI) of discovered leader sequences ([Materials and Methods](#)) except for RM due to the low
175 number of SNPs. For heavy chain leaders, the R/S ratio of human was the lowest (0.43), which
176 was followed by CM (0.63), mouse (2.66) and rat (4.05) ([Figure 4a](#)). Similar decreased R/S ratios
177 were also observed in light chains ([Supplementary figure 12](#)). Knowing that each leader sequence
178 contains a hydrophobic central region [16], we classified the amino acids according to their
179 polarities and looked into the amino acid conversion types for replacement SNPs. It demonstrated
180 that intra-group conversions dominated overall amino acid conversions ([Figure 4b and](#)
181 [Supplementary figure 12](#)). And the percentage of inter-group conversions followed the same order
182 as that of replacement SNPs in different species. Furthermore, both NDI and ADI were calculated
183 to evaluate the positional nucleotide and amino acid diversity. The NDI profile exhibited distinct
184 patterns ([Figure 4c](#)). To be specific, the third nucleotides in 2nd, 3rd, 9th, 12th and 19th codons in
185 leader sequences possessed clearly higher NDI than the first two nucleotides in the same codons.
186 Since the choice of the third nucleotides in a codon often do not change the encoded amino acids,
187 the higher NDI of the third nucleotides demonstrated a clearly negative selection over the
188 polymorphisms taking place in the first two nucleotides in these codons. In contrast, the 13th, 14th,
189 16th, and 17th codons were observed with higher NDI in the first two nucleotides, suggesting the
190 underlying positive selection. While for the 6th and 7th codons, the diversities were comparable
191 among the constituent nucleotides. The positional ADI profile was consistent with NDI ([Figure](#)
192 [4c](#)).

193 The SNPs in AUSs would also affect the antibody genotyping as well as Rep-Seq when these
194 polymorphic regions were targeted by PCR primer design. We selected two representative primer

195 sequences from public resources that target either the 3' end (Primer 1) or 5' end (Primer 2) of
196 leader sequences of IGHV1-24*01. We found both of them covered the SNP loci (5 for primer 1
197 and 2 for primer 2) (Figure 4d). To comprehensively evaluate the influence of novel leaders on
198 antibody sequence amplification, we compared the novel leader sequences to 6 widely applied
199 primer sets [24,32-36]. As a result, 9 to 22 genes were found with novel sequences more distant to
200 the optimal primers than their counterpart in IMGT (Supplementary Table 3). Particularly, the
201 number of mismatches can increase from 4 to 9 for IGHV1-17 in the primer set reported by Klein
202 et al. [32] (Supplementary Table 4). Indeed, when comparing amplifications of V alleles between
203 5'RACE and multiplex primers, the more mismatches in the primer, the less V allele were found
204 in the repertoire (Figure 4e, Materials and Methods). More mismatches between primers and their
205 targeted AUSs led to more compromised amplification. Therefore, the AUS set we reported here
206 would be a valuable resource for later primer designs.

207 More importantly, the SNPs in the leader region also caused functional resurrection or silence
208 to their corresponding genes. For instance, the human IGHV3-69-1*01 and IGHV3-69-1*02 were
209 both annotated as pseudo genes at IMGT because of the missing initiation codons at the starts of
210 leader sequences. However, we identified 2 (supported by 34 and 5 subjects) and 1 (3 subjects)
211 novel leader sequences that possessed normal initiation codon (*AUG*) and thus made these alleles
212 functional (Figure 4f and Supplementary figure 13a). This functional recurrence also happened in
213 IGHV1-67*01 of mouse (5 subjects) (Supplementary figure 13b). On the contrary, the SNP in the
214 leader region caused early stop codon and silenced IGHV11-2*01 in rat (2 subjects) (Figure 4g).
215 Thus, the SNPs in the AUSs are important for the activation of downstream V genes and
216 additional awareness should be paid in the future.

217

Figure 4

Figure 4. Characterization of SNPs observed in discovered leader sequences. (a) The percentage of silent (S) and replacement (R) SNPs for heavy chain. R/S ratio was marked beneath each pie plot. (b) The percentage of intra-group and inter-group amino acid conversion. Note that amino acids here were classified according to their polarities. (c) Positional nucleotide diversity profile of leader sequences. The vertical lines mean the standard errors. Boxes in red represent the third nucleotides in codons. (d) The schematic representation of mismatches of published primers with discovered leader sequences. (e) The correlation between the contribution to the overall target sequences and the number of their mismatches with primers. (f, g) Schematic representation of SNP-associated V gene functional reversals, including pseudo-to-functional (f) and functional-to-pseudo transitions (g).

218

219 **uAUGs and leader sequences do not affect antibody expression**

220 Upstream open reading frames (uORFs) or upstream AUGs (uAUGs) have been reported to
221 exist in 44%-50.5% of human annotated transcripts, implicate longer 5' UTR sequences, and
222 associate with significant reduction of downstream protein expression and modestly to markedly
223 decreased mRNA level [14,18,37]. Having obtained 5' UTR sequences from hundreds of unique
224 AUSs for human, we investigated the prevalence of uAUGs in antibody V genes and its
225 correlation with gene expression.

226 We observed uAUGs in the 5' UTR of 12.4%, 11.85% and 8.70% of AUSs for human heavy,
227 kappa and lambda chain, respectively (Table 2). Thus the uAUG frequency in antibody V genes is
228 less than that in human mRNAs and the stochastic estimation, indicating the purifying selection
229 [37,38]. Thus, this underrepresentation of uAUGs encoded in antibodies indicated that antibodies
230 tend not to be subjected to this post-transcriptional regulation. We also observed a significant
231 length difference between uAUG-containing and uAUG-absent 5' UTRs (Figure 5a). Notably,
232 these differences did not fully agree with what was reported in previous studies. The antibody
233 heavy chain 5' UTRs from human and CM demonstrated a reversed pattern, in which
234 uAUG-containing 5' UTRs were even shorter. This result implied that a longer 5' UTR is not
235 necessarily predisposed to contain uORFs and thus further address the polymorphism of
236 post-transcriptional regulation for different antibody chains and for different species.

237 The high throughput Rep-Seq approach gives us access to the measurement of gene
238 expression. We thus obtained the gene expression for each sample and correlated it with the
239 number of uAUGs present in the 5' UTR sequences of their corresponding V genes. Two and four
240 genes meeting the criteria were included in this analysis for human heavy and kappa chain,

241 respectively ([Materials and Methods](#)). Although varied mRNA level was observed in different
242 uAUG number group, statistical analysis indicated the difference was not significant ($p>0.5$,
243 unpaired Student's t-test), except for a kappa gene - IGKV1-17 ([Figure 5b and Supplementary](#)
244 [figure 14](#)). Together with the observation that the correlation was ambiguous (both seemingly
245 positive and negative correlation were obtained), we believed that the influence of uAUGs on
246 antibody expression, if any, is not critical. It is noteworthy that 21 of 23 heavy chain V genes with
247 which we observed uAUGs are core genes which are prevalent in the repertoire [39]
248 ([Supplementary figure 15](#)). As half of the heavy chain V genes are noncore genes that express very
249 lowly in the repertoire, the uAUGs in these genes might be unrevealed in this study.

250 Apart from uORFs, the leader sequence itself was also found to associate with mRNA
251 expressions [15,20]. To validate whether and to what extent leader sequences affect antibody
252 transcription, we set off to investigate the antibody expression of V genes sharing two or more
253 unique leaders. As shown in [Figure 5c and Supplementary figure 16](#), the overall pattern showed no
254 significant leader-specific gene expression was found. Thus the leader sequences involve more in
255 the regulation of translation rather than transcription. Altogether, the results above reflected the
256 polymorphism of antibody 5' UTR sequences across species and chains and the trivial functional
257 relevance to gene expression of uORFs and leader sequences.

Figure 5

Figure 5. Functional relevance of uAUGs and leader sequences to gene expression. (a) The 5' UTR length comparison between uAUG-containing group and uAUG-absent group in different species. The red asterisks marked the chains with reversed length difference between two groups. (b) The functional relevance of the number of uAUGs in 5' UTR sequences to gene expression for human heavy chain V genes. (c) The functional relevance of leader sequences to gene expression for human heavy chain V genes. L1, L2, and L3 represent unique leader sequences. The L1 and L2 in different subplots are not necessarily the same. The number in the center of or on top of each bar mean the number of samples in each group. The vertical lines mean the standard errors. n. a. not available; n. s., not significant ($p \geq 0.5$); *, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.001$.

Discussion

Rep-Seq has provided an avenue by which researchers are able to investigate antibodies with an unprecedented depth [5]. Taking advantage of Rep-Seq, we investigated the polymorphisms existing in the AUSs of both heavy chain and light chain V genes from five species. We designed a bioinformatic pipeline and captured 2,957 unique AUSs, of which 875 (29.6%) are novel (as to leader). We demonstrated that the SNPs in the AUSs affect the PCR amplification efficiency and thus the subsequent antibody quantification. Therefore, our findings are of importance for antibody gene genotyping, antibody quantifications with traditional PCR method and

267 high-throughput Rep-Seq. Moreover, the novel AUSs revealed in this study also provide a
268 resource for antibody engineering and aid antibody studies in the molecular level.

269 In-depth analyses showed that leader sequences are family-specific and are more conserved
270 than 5' UTR. The distance of the AUS between human and other nonhuman species are
271 reminiscent of that observed in genome level. And the sequence similarities between human and
272 nonhuman species also indicates a synchronous evolution between leaders and their downstream
273 V genes.

274 Notably, we found a 5' UTR sequence was shared by V genes of different families. The
275 length of the associated 5' UTR was only 3 bp. It was unlikely to be the result of RNA degradation
276 during library preparation, because the 5'cap is required for a successful template switch when
277 using 5'RACE protocol [40]. Moreover, we found the 5' UTR for IGHV1-17*02 in 6 subjects, a
278 strong evidence for its authenticity. Previous studies reported a single-nucleotide 5' UTR is
279 enough for the initiation of translation in vitro [41]. Therefore, the short 5' UTR may not mean the
280 functional abnormality. However, short 5' UTRs were found to impact the translation efficiency
281 [41]. Due to the unavailability of proteomic data, we were not able to investigate its translation
282 efficiency.

283 Informed by previous reports that uAUGs and leader sequence have an effect on gene
284 expression in other context [4,15,18,19], we also examined if there exists correlations between
285 them in antibody genes. We showed little evidence to support the link between them, which means
286 for antibodies the AUSs primarily dictate the translation process. Noteworthy is that we
287 investigated the correlations only in human, partly because we have no enough samples for
288 nonhuman species. The other reason, however, is that there are a limited number of germline

reference available [7-9]. Even worse is that some germline reference provided by IMGT/GENE-DB are provisional because their genomic locations has not been determined [26]. Hence, we proposed that the germline sequences of immunoglobulin loci should be thoroughly characterized in these species. The accurate antibody probing in these model organisms are supposed to boost the understanding of antibody repertoire in human.

In addition to the little quantitative impact on gene expression, we also observed that the SNPs in leader sequences could reverse the functional status of their downstream genes. It addressed importance of dissecting the AUSs in the evaluation of antibody functionalities, because a functional gene in a typical individual or ethnic group may be nonfunctional in another and *vice versa*. However, the prevalence of such SNP-associated functional recurrences or silences, together with their consequence for the antibody repertoire, remains to be determined. Another interesting phenomenon we observed is the higher R/S ratio in leader region for mouse and rat than human and CM. Moreover, the percentage of inter-group amino acid conversion was also higher for the two rodents. It is possibly because the immunoglobulin loci are less essential for the survival of rodents and thus they can tolerate more functional consequence caused by inter-group amino acid conversions [42].

In summary, we provided for the first time the most comprehensive knowledge database for AUSs for both human and other model organisms. Together with the characterization of these sequences, the AUS set discovered in this study will serve as valuable resources for fundamental studies and antibody engineering as well.

310 **Funding**

311 This study was supported by the National Natural Science Foundation of China (NSFC)
312 (31771479) (Z. Z.), NSFC Projects of International Cooperation and Exchanges of NSFC
313 (61661146004), and the Local Innovative and Research Teams Project of Guangdong Pearl River
314 Talents Program (2017BT01S131).

315 **Author Contributions**

316 Y. Z., X. Y., J. W., S. C., Y. C., and C. L. analyzed the data. H. T., Q. W., J. G., W. X., and M. W.
317 conducted the biological experiment. C. L. coordinated the project. Y. Z., X. Y., J. W., H. T., L. W.,
318 C. S., and Z. Z. wrote the manuscript. Z. Z. conceived the project.

319 **Competing of interests**

320 The authors declared no competing financial interests.

321 **Materials and Methods**

322 **Subjects and sample preparation**

323 In this study, we included a total number of 782 samples from five species, including human
324 (n=728), RM (*Macaca mulatta*) (n=27), CM (*Macaca fascicularis*) (n=2), mouse (*Mus musculus*)
325 (n=11), and rat (*Rattus norvegicus*) (n=14). Twenty-one and twenty-six samples are from public
326 resources for human and RM, respectively. And their accession numbers in NCBI Sequence Read
327 Archive are provided as [Supplementary Table 5](#). The in-house human samples were prepared as

described previously [39]. As for samples from nonhuman species, EDTA-treated blood specimens were collected from sixteen animals (seven rats, six mice, two CMs, and an RM). Peripheral blood mononuclear cells (PBMCs) were isolated by Ficoll-Paque density centrifugation. The PBMCs were subjected to total RNA extraction using the RNeasy Mini Kit (Qiagen, 74106), according to the manufacturer's protocol. This protocol was approved by the Ethics Committee of Southern Medical University.

Library preparation and high throughput sequencing

Total RNA was used as a template to synthesize cDNA, using a SMARTer RACE (Rapid Amplification of cDNA Ends) cDNA Amplification Kit (Clontech, 634928), according to the manufacturer's protocol. Following cDNA synthesis, 10% of the volume of cDNA was subjected to VH amplification in a 25 µl PCR reaction using the Kapa HiFi HotStart Ready Mix (KAPA Biosystems, kk2602) with universal or VH-family specific forward primers and the corresponding reverse primers ([Supplementary Table 6](#)). The thermal cycling conditions were programmed as follows: 95 °C for 3 min; 30 cycles of 98 °C for 20 s, 60 °C for 15 s, and 72 °C for 15 s; 72 °C for 5 min. PCR products were purified using the Nucleospin Gel & PCR Clean-up kit (Macherey-Nagel, 704609.25). DNA concentration was detected using the Qubit 4.0 fluorometer (ThermoFisher Scientific). Two hundred nanograms of each gel-purified product was subjected to library preparation, followed by sequencing by Illumina platform (MiSeq PE300 and NovaSeq PE250).

Extraction of AUSs and delimitation of 5' UTR and leader region

The Six steps we employed to identify qualified AUSs were briefly described in the Result section and demonstrated as [Figure 1](#). Here we provided a more detailed implementation of these

350 steps.

351 i) V allele variant detection. Considering the uncaptured polymorphisms in
 352 immunoglobulin loci in mammalian, especially for nonhuman species, we started with
 353 the identification of novel alleles for all enrolled samples and species [9,43,44]. We
 354 executed three iterations of IgDiscover (v0.12.3) based on an initial species-specific V
 355 gene database obtained from IMGT/GENE-DB (update: 06 July 2020) [7,26]. To enroll
 356 as many as possible the initial germline sequences, we included all sequences from the
 357 F+ORF+all P directory. The full set of germline reference was then subjected to a
 358 deduplication process before serving as the starting database. After the detection, all
 359 discovered novel alleles were merged with the initial ones for downstream sequence
 360 annotation.

361 ii) Sequence annotation and assembly. We then proceeded with the functionalities provided
 362 by MiXCR (v3.0.7) [45]. Taking advantage of the subcommand *align* and *assemble* and
 363 their exportation counterparts, we obtained both the sequence annotation (alignments.txt)
 364 and clustering (clones.txt) results. Sequences with the same V alleles, J alleles, and
 365 CDR3 nucleotide sequences were assembled into clonotypes. The commands we used to
 366 annotate and assemble sequences with MiXCR (v3.0.7) are as below,

367 Alignment:

368 *mixcr align --species \$species -f --library \$species.specific.library \$read1 \$read2*
 369 *alignments.vdjca*

370 Assembly:

371 *mixcr assemble -f -a -OseparateByV=true -OseparateByJ=true alignments.vdjca*

372 *clones.clna*

373 Exportation:

374 *mixcr exportAlignments -f -readIds -cloneId -vHit -vAlignment -targetSequences*

375 *clones.clna alignments.txt*

376 *mixcr exportClones -f --chains \$CHAIN clones.clna clones.txt*

377 iii) Genotyping. To guide the AUS discovery, the genotype was predetermined based on

378 allele usage for each V gene. In this study, we considered for each V gene two scenarios,

379 homozygous and heterozygous. Theoretically, if a gene is heterozygous in a typical

380 sample, each of the two variants will possess a unique AUS. Otherwise, the only variant

381 of a homozygote will probably link with two unique AUSs. We determined the genotype

382 through the Bayesian method employed in TIGGER [46], except that we considered only

383 two scenarios aforementioned. In this case, $\vec{\pi}_{H_H} = (1, 0)$ and $\vec{\pi}_{H_D} = (0.6, 0.4)$. Other

384 parameters for the proposed Bayesian model were unchanged. The allele usage,

385 measured as the number of clonotypes recombined from a typical allele, served as the

386 input of this function.

387 iv) Identification of consensus sequences within clonotypes. Each AUS is deemed to start at

388 a position immediately downstream the rGrGrGs in the 3' end of RACE primer and end

389 at the exclusively initial base of the associated V gene segment. For each clonotype, we

390 extracted a list of valid AUSs and determined the most frequent one as the consensus

391 AUS. For reliability, we discarded the consensus sequences representing no more than 50%

392 of the AUSs within clonotypes. We also required the initiation codon – “AUG” – is

393 contained in the AUS to precisely delimit 5' UTR and leader region. For AU Ss with

multiple AUGs, we selected an optimal AUG as the start position of leaders. Each optimal AUG corresponds to a leader sequence with its length nearest to a predetermined optimal leader length. The optimal leader length for each allele is same as the length of leader sequence provided by IMGTT provided it is known. Otherwise, it is same as the most frequent leader length of all other alleles corresponding to the same gene, family or all the rest alleles with known leader sequences. Noted that we extracted only leader sequences for samples from BioProject PRJNA503527 for their UTRs were found to be incomplete. The extracted leader sequences were then subjected to the same pipeline as AUSs.

v) Identification of consensus sequences within V alleles. The definition of clonotype ensured that each of consensus sequences is associated with a unique V allele. AUSs corresponding to alleles not present in the genotype list were discarded at the first place. Then we collapsed the consensus sequences belonging to the same alleles and kept the most frequent ones. Alleles with no more than 10 available AUSs were excluded. For each allele in the genotype list, we retained also the second most frequent AUSs to account for the underlying diversity in the upstream region, providing the corresponding gene is a homozygote and the most frequent one takes up less than 87.5% of total AUSs [8].

vi) Score-based filtration. For each chain of a typical species, the candidate AUSs for all samples were pooled together and then classified into two groups according to prior knowledge. The group containing known leader sequences were regarded as *bona fide* and thus serve as the positive control. The other group, comprised of novel sequences,

will be subject to a score-based filtration. The score scheme takes into account four features for each candidate sequence, namely the absolute number of supportive reads, clones and donors and the similarity to known sequences. The weight assigned to them were 20, 20, 30, and 30, respectively. The independent scores (S_{read} , S_{clone} , S_{donor} and $S_{similarity}$) and total score (S_{total}) were calculated as the formulas below,

$$S_{read} = \begin{cases} 20 \times \sqrt{\frac{n_r}{N_{rmedian}}} & (n_r < N_{rmedian}) \\ 20 & (n_r \geq N_{rmedian}) \end{cases} \quad (i)$$

$$S_{clone} = \begin{cases} 20 \times \sqrt{\frac{n_c}{N_{cmedian}}} & (n_c < N_{cmedian}) \\ 20 & (n_c \geq N_{cmedian}) \end{cases} \quad (ii)$$

$$S_{donor} = 30 \times (1 - 1/(n_d + 1)) \quad (iii)$$

$$S_{similarity} = 30 \times s \quad (iv)$$

$$S_{total} = S_{read} + S_{clone} + S_{donor} + S_{similarity} \quad (v)$$

Note, $N_{rmedian}$ and $N_{cmedian}$ mean the median number of supportive reads and clones for known AUSs that serve as positive control while n_r , n_c , and n_d mean the number of supportive reads, clones, and donors for novel AUSs. s represents the identity to known leader sequences for the leader region in each AUS. If no known leader sequence was provided by IMGT/GENE-DB for the corresponding allele, an average identity to the leaders of all other alleles corresponding to the same gene or the same family or of all the rest alleles with known leader sequences was calculated to represent the identity. Thus, the theoretical maximum score for an AUS is less than 100. After the S_{total} were calculated for each AUS, the novel sequences with a score less than the median S_{total} of

431 positive control were discarded and the rest novel sequences together with positive
432 control serve as the final AUS library. The score-based filtration step was applied to each
433 chain of each species independently. An exception was with the filtration of AUSs for the
434 kappa chain of RM. Since no known complete leader sequences were discovered, the
435 AUSs containing the known partial leader sequences were regarded as the positive
436 control for filtration.

437 **The calculation of the number of unique AUS, 5' UTR, and leader sequences**

438 The number of unique AUS, 5' UTR, and leader sequences is calculated as the number of
439 unique combinations of V allele and AUS, 5' UTR, and leader sequences. The AUSs containing
440 only leader sequences (n=4) identified from RM samples under BioProject PRJNA503527 were
441 not included in this analysis, but instead were included in all other analyses.

442 **Sequence similarity comparison**

443 Sequence similarity or identity is calculated based on the pairwise alignment result. The
444 pairwise alignment was implemented using *pairwise2* function built in Biopython (v1.70). After
445 the pairwise alignment result was obtained, we calculated the sequence similarity as the number of
446 matched nucleotides divided by the alignment length. For the measurement of leader or 5' UTR
447 sequence similarity between two genes within the same species, we performed an all-to-all
448 comparison between all leader sequences corresponding to the two compared genes and calculated
449 an average similarity for each gene pair. While for the sequence comparison between genes from
450 human and nonhuman species, we determined the homologous genes at first. The homologous
451 gene in human for each allele of nonhuman species was determined by the sequence similarity and
452 the gene corresponding to the nearest allele in human was considered as its homologous gene. In

453 this case, a typical gene in nonhuman species can have alleles with different homologous genes in
454 human. Thus a gene in nonhuman species can be homologous with multiple genes in human and
455 vice versa.

456 **Evaluation of functional relevance of uAUGs and leader sequence to gene expression**

457 The functional relevance of uAUGs and leader sequence to gene expression was investigated
458 in only human samples due to the limitation of sample size. To avoid the noise caused by
459 heterozygous 5' UTR for a gene, only genes with one unique 5' UTR sequence were considered in
460 each sample. Besides, among samples, only genes with diversity in the number of uAUGs and
461 having at least 5 samples in each group were finally included in the analysis. Applying these
462 criteria to all genes, we discovered 2 and 4 genes for heavy and light chain, respectively (Figure
463 5b and Supplementary figure 14). The gene filtration criteria for the evaluation of leader
464 sequences' influence on gene expression is similar. In each sample, only genes with one unique
465 leader sequence were retained. Besides, among samples, only genes with at least two kinds of
466 leader sequences and shared at least two leader sequences with another were finally included. The
467 gene expression in each sample is calculated as the number of antibody sequences assigned a
468 typical gene divided by the number of all assigned antibody sequences.

469 **The measurement of percentage of replacement and silent SNPs**

470 Owing to the limited number of discovered leader sequence, we did not include RM into this
471 analysis. To measure the percentage of replacement and silent SNPs, the SNP loci were firstly
472 determined by aligning leader sequences from the same genes and then identifying the
473 polymorphic positions. For heavy chain and lambda chain, only leader sequences of 57 bp were
474 considered, while for kappa chain only 60 bp. This length limitation makes the leader sequence

alignment quite straightforward and at the same time retain the overwhelming majority of leader diversity (Figure 2b). Then the SNP loci-associated codons were identified. For each SNP loci-associated codon position (e. g. from 1 to 19 for heavy chain), the frequencies of different codons were calculated. Each pair of codons will associate with a single SNP outcome, either replacement or silent. Then for a certain codon position, its contribution to silent SNPs was calculated as the sum of the product of the two frequencies for all codon pairs encoding the same amino acids. Similarly, its contribution to replacement SNPs was calculated as the sum of the product of the two frequencies for all codon pairs encoding different amino acids. Finally, the total replacement or silent SNPs were the sums of the contributions of all SNP loci-associated codon positions. The percentages of two kinds of amino acid conversion, namely intra-group conversion and inter-group conversion, were determined in the same way, except that it considered only replacement SNPs.

487 **Nucleotide and amino acid diversity index (NDI and ADI) calculation**

Only SNPs of human heavy chain leader were included in this analysis. We used NDI (or ADI) to measure the nucleotide (or amino acid) diversity of each position. The diversity index was supposed to reflect the conservation of each position in the leader. Both NDI (57 positions) and ADI (19 positions) were based on the formula calculating Shannon index as follows:

$$H' = - \sum_{i=1}^S p_i \ln p_i$$

where S represents the number of unique nucleotides or amino acids in a certain position and p_i represents the proportion of i th nucleotide or amino acid.

494 **Mismatch comparison between known and novel leaders**

495 In total 6 widely applied primer sets were included to compare the number of mismatches
496 between primers and known leaders in IMGT/GENE-DB with that between primers and novel
497 leaders found in this study. Therefore, the comparison considered only V alleles with both
498 discovered novel leaders and their complete counterparts in IMGT/GENE-DB. For each leader
499 sequence, we determined its optimal primer as well as the target region in each primer set by
500 pairwise alignment. This process also enables us to obtain the number of mismatches with its
501 optimal primer. In this way, a maximum number of mismatches with optimal primers can be
502 obtained for each V gene ([Supplementary Table 4](#)).

503 **Evaluation of primer amplification efficiency**

504 Two amplification strategies (5'RACE and multiplex PCR) were applied to the same sample
505 and the paired sequencing dataset were acquired using NovaSeq 6000 sequencing system. We
506 genotyped and extracted the AUSs together with the starting 30 bp sequence of V alleles for the
507 sequenced sample using the RACE-derived dataset. For the multiplex-derived dataset, we also
508 employed MiXCR to annotate the antibody sequences. The antibody sequences assigned to V
509 alleles absent in the genotype were excluded at first. Then the contribution of each primer to one
510 of its target V alleles was measured as this V allele's percentage of all the sequences targeted by
511 this primer. It is also straightforward to calculate the mismatches between the primer and its target
512 V alleles based on the AUSs and partial sequences of V alleles. When correlating the contributions
513 with mismatches, we found a clear negative correlation between them ([Figure 4e](#)).

514 **Software**

515 In-house scripts were written in Python (v3.7.4) employing modules including numpy

(v1.16.4), Biopython (v1.73), Levenshtein (v0.12.0) and pandas (v0.24.2). To visualize these results, we employed modules seaborn (v0.9.1) and matplotlib (v3.0.2) as well as a standalone software, BioEdit Sequence Alignment Editor.

References

1. Forthal DN. Functions of Antibodies. Microbiology spectrum 2014;2:1-17
2. Boyd SD, Marshall EL, Merker JD, et al. Measurement and clinical monitoring of human lymphocyte clonality by massively parallel VDJ pyrosequencing. SCI TRANSL MED 2009;1:12r-23r
3. Campbell PJ, Pleasance ED, Stephens PJ, et al. Subclonal phylogenetic structures in cancer revealed by ultra-deep sequencing. Proc Natl Acad Sci U S A 2008;105:13081-13086
4. Weinstein JA, Jiang N, White RR, Fisher DS, Quake SR. High-throughput sequencing of the zebrafish antibody repertoire. SCIENCE 2009;324:807-810
5. Georgiou G, Ippolito GC, Beausang J, et al. The promise and challenge of high-throughput sequencing of the antibody repertoire. NAT BIOTECHNOL 2014;32:158-168
6. Zhang Y, Xu Q, Zeng H, et al. SARS-Cov-2-, HIV-1-, Ebola-neutralizing and anti-PD1 clones are predisposed. bioRxiv 2020
7. Corcoran MM, Phad GE, Bernat NV, et al. Production of individualized V gene databases reveals high levels of immunoglobulin genetic diversity. NAT COMMUN 2016;7
8. Gadala-Maria D, Yaari G, Uduman M, Kleinstein SH. Automated analysis of high-throughput B-cell sequencing data reveals a high frequency of novel immunoglobulin V gene segment alleles.

- 537 Proceedings of the National Academy of Sciences 2015;112:E862-E870
- 538 9. Ralph DK, Matsen FA. Per-sample immunoglobulin germline inference from B cell receptor
- 539 deep sequencing data. PLOS COMPUT BIOL 2019;15:e1007133
- 540 10. Glanville J, Kuo TC, von Budingen HC, et al. Naive antibody gene-segment frequencies are
- 541 heritable and unaltered by chronic lymphocyte ablation. Proc Natl Acad Sci U S A
- 542 2011;108:20066-20071
- 543 11. Lingwood D, McTamney PM, Yassine HM, et al. Structural and genetic basis for
- 544 development of broadly neutralizing influenza antibodies. NATURE 2012;489:566-570
- 545 12. Parks T, Mirabel MM, Kado J, et al. Association between a common immunoglobulin heavy
- 546 chain allele and rheumatic heart disease risk in Oceania. NAT COMMUN 2017;8
- 547 13. Watson CT, Glanville J, Marasco WA. The Individual and Population Genetics of Antibody
- 548 Immunity. TRENDS IMMUNOL 2017;38:459-470
- 549 14. Calvo SE, Pagliarini DJ, Mootha VK. Upstream open reading frames cause widespread
- 550 reduction of protein expression and are polymorphic among humans. Proc Natl Acad Sci U S A
- 551 2009;106:7507-7512
- 552 15. Elfakess R, Dikstein R. A translation initiation element specific to mRNAs with very short
- 553 5'UTR that also regulates transcription. PLOS ONE 2008;3:e3094
- 554 16. Haryadi R, Ho S, Kok YJ, et al. Optimization of Heavy Chain and Light Chain Signal
- 555 Peptides for High Level Expression of Therapeutic Antibodies in CHO Cells. PLOS ONE
- 556 2015;10:e116878
- 557 17. Lovett PS, Rogers EJ. Ribosome regulation by the nascent peptide. Microbiol Rev
- 558 1996;60:366-385

559 18. Matsui M, Yachie N, Okada Y, Saito R, Tomita M. Bioinformatic analysis of
560 post-transcriptional regulation by uORF in human and mouse. FEBS LETT 2007;581:4184-4188

561 19. Vilela C, McCarthy JEG. Regulation of fungal gene expression via short open reading frames
562 in the mRNA 5'untranslated region. MOL MICROBIOL 2003;49:859-867

563 20. Wellensiek BP, Larsen AC, Flores J, Jacobs BL, Chaput JC. A leader sequence capable of
564 enhancing RNA expression and protein synthesis in mammalian cells. PROTEIN SCI
565 2013;22:1392-1398

566 21. Zhou Y, Liu P, Gan Y, et al. Enhancing full-length antibody production by signal peptide
567 engineering. MICROB CELL FACT 2016;15

568 22. Gibson SJ, Bond NJ, Milne S, et al. N-terminal or signal peptide sequence engineering
569 prevents truncation of human monoclonal antibody light chains. BIOTECHNOL BIOENG
570 2017;114:1970-1977

571 23. Khan TA, Friedensohn S, Gorter DVA, et al. Accurate and predictive antibody repertoire
572 profiling by molecular amplification fingerprinting. SCI ADV 2016;2:e1501371

573 24. Kreer C, Döring M, Lehnen N, et al. openPrimeR for multiplex amplification of highly
574 diverse templates. J IMMUNOL METHODS 2020;480:112752

575 25. Mikocziova I, Gidoni M, Lindeman I, et al. Polymorphisms in human immunoglobulin heavy
576 chain variable genes and their upstream regions. NUCLEIC ACIDS RES 2020;48:5499-5510

577 26. Giudicelli V. IMGT/GENE-DB: a comprehensive database for human and mouse
578 immunoglobulin and T cell receptor genes. NUCLEIC ACIDS RES 2004;33:D256-D261

579 27. Zhang Y, Yang X, Zhang Y, et al. Tools for fundamental analysis functions of TCR repertoires:
580 a systematic comparison. BRIEF BIOINFORM 2019

581 28. Ebeling M, Kung E, See A, et al. Genome-based analysis of the nonhuman primate *Macaca*
582 *fascicularis* as a model for drug safety assessment. *GENOME RES* 2011;21:1746-1756

583 29. Gibbs RA, Rogers J, Katze MG, et al. Evolutionary and biomedical insights from the rhesus
584 macaque genome. *SCIENCE* 2007;316:222-234

585 30. Gibbs RA, Weinstock GM, Metzker ML, et al. Genome sequence of the Brown Norway rat
586 yields insights into mammalian evolution. *NATURE* 2004;428:493-521

587 31. Waterston RH, Lindblad-Toh K, Birney E, et al. Initial sequencing and comparative analysis
588 of the mouse genome. *NATURE* 2002;420:520-562

589 32. Klein F, Gaebler C, Mouquet H, et al. Broad neutralization by a combination of antibodies
590 recognizing the CD4 binding site and a new conformational epitope on the HIV-1 envelope
591 protein. *The Journal of Experimental Medicine* 2012;209:1469-1479

592 33. Meng W, Zhang B, Schwartz GW, et al. An atlas of B-cell clonal distribution in the human
593 body. *NAT BIOTECHNOL* 2017;35:879-884

594 34. Scheid JF, Mouquet H, Ueberheide B, et al. Sequence and structural convergence of broad
595 and potent HIV antibodies that mimic CD4 binding. *SCIENCE* 2011;333:1633-1637

596 35. Tiller T, Meffre E, Yurasov S, et al. Efficient generation of monoclonal antibodies from single
597 human B cells by single cell RT-PCR and expression vector cloning. *J IMMUNOL METHODS*
598 2008;329:112-124

599 36. Vergani S, Korsunsky I, Mazzarello AN, et al. Novel Method for High-Throughput
600 Full-Length IGHV-D-J Sequencing of the Immune Repertoire from Bulk B-Cells with Single-Cell
601 Resolution. *FRONT IMMUNOL* 2017;8

602 37. Iacono M, Mignone F, Pesole G. uAUG and uORFs in human and rodent 5'untranslated

603 mRNAs. GENE 2005;349:97-105

604 38. Neafsey DE, Galagan JE. Dual Modes of Natural Selection on Upstream Open Reading
605 Frames. MOL BIOL EVOL 2007;24:1744-1751

606 39. Yang X, Wang M, Shi D, et al. Large-scale Analysis of 2,152 dataset reveals key features of
607 B cell biology and the antibody repertoire. bioRxiv 2019

608 40. Harbers M, Kato S, de Hoon M, et al. Comparison of RNA- or LNA-hybrid oligonucleotides
609 in template-switching reactions for high-speed sequencing library preparation. BMC GENOMICS
610 2013;14:665

611 41. Hughes MJG, Andrews DW. A single nucleotide is a sufficient 5' untranslated region for
612 translation in an eukaryotic in vitro system. FEBS LETT 1997;414:19-22

613 42. Liao B, Zhang J. Null Mutations in Human and Mouse Orthologs Frequently Result in
614 Different Phenotypes. Proceedings of the National Academy of Sciences - PNAS
615 2008;105:6987-6992

616 43. Frost SDW, Murrell B, Hossain ASMM, Silverman GJ, Pond SLK. Assigning and visualizing
617 germline genes in antibody repertoires. Philosophical Transactions of the Royal Society B:
618 Biological Sciences 2015;370:20140240

619 44. Zhang W, Li X, Wang L, et al. Identification of Variable and Joining Germline Genes and
620 Alleles for Rhesus Macaque from B Cell Receptor Repertoires. J IMMUNOL
621 2019;202:1612-1622

622 45. Bolotin DA, Poslavsky S, Mitrophanov I, et al. MiXCR: software for comprehensive
623 adaptive immunity profiling. NAT METHODS 2015;12:380-381

624 46. Gadala-Maria D, Gidoni M, Marquez S, et al. Identification of Subject-Specific

625 Immunoglobulin Alleles From Expressed Repertoire Sequencing Data. FRONT IMMUNOL

626 2019;10

627

628 **Figure legend**

629 **Figure 1. The AUS identification pipeline.** The AUS identification pipeline is comprised with six
630 steps, including **i)** V allele variant detection, **ii)** sequence annotation and assembly, **iii)** genotyping,
631 **iv)** identification of consensus sequences within clonotypes, **v)** identification of consensus
632 sequences within V alleles, and **vi)** score-based filtration.

633

634 **Figure 2. Overview of the discovered AUSs for heavy chain.** (a) The composition of discovered
635 5' UTR and leader sequences. "Consistent" indicates sequences identical to those curated by
636 IMGT/GENE DB. "Unknown" indicates sequences whose corresponding alleles were not
637 provided with available AUSs. "Other" includes the rest four situations, which were illustrated on
638 top of the donut charts. The number of novel sequences is marked in the center of each donut chart.
639 (b) The length distribution of discovered 5' UTR and leader sequences. The length and its
640 corresponding frequency were marked for each peak. (c) The distribution of the number of donors
641 sharing a certain sequence. Different colors in (b) and (c) mean different groups and are consistent
642 with those in (a). (d) Percentage of V genes as a function of the number of discovered 5' UTR and
643 leader sequences.

644

645 **Figure 3. 5'UTR and leader sequence similarity within human and between species.** (a)

646 5'UTR sequence identity between V genes. **(b)** Leader sequence identity between V genes. **(c)** The
 647 number of shared 5'UTR sequences between V genes of different families. **(d)** The number of
 648 shared leader sequences between V genes of different families. The side color bar indicates gene
 649 families. **(e)** The correlation between 5'UTR or leader sequence identities and V allele sequence
 650 identities between human and four nonhuman species. The line in each scatterplot represents the
 651 fitted linear regression model. **(f)** V gene, 5'UTR and leader sequence identity between human and
 652 nonhuman species for antibody heavy chain. Genes were classified into core genes and noncore
 653 genes according to Yang et al., 2019. *r*, Pearson's correlation coefficient; *p*, *p* value of the linear
 654 model; **, $p < 0.01$; ***, $p < 0.001$.

655
 656 **Figure 4. Characterization of SNPs observed in discovered leader sequences.** **(a)** The
 657 percentage of silent (S) and replacement (R) SNPs for heavy chain. R/S ratio was marked beneath
 658 each pie plot. **(b)** The percentage of intra-group and inter-group amino acid conversion. Note that
 659 amino acids here were classified according to their polarities. **(c)** Positional nucleotide diversity
 660 profile of leader sequences. The vertical lines mean the standard errors. Boxes in red represent the
 661 third nucleotides in codons. **(d)** The schematic representation of mismatches of published primers
 662 with discovered leader sequences. **(e)** The correlation between the contribution to the overall
 663 target sequences and the number of their mismatches with primers. **(f, g)** Schematic representation
 664 of SNP-associated V gene functional reversals, including pseudo-to-functional (f) and
 665 functional-to-pseudo transitions (g).

666
 667 **Figure 5. Functional relevance of uAUGs and leader sequences to gene expression.** **(a)** The 5'

668 UTR length comparison between uAUG-containing group and uAUG-absent group in different
669 species. The red asterisks marked the chains with reversed length difference between two groups.
670 **(b)** The functional relevance of the number of uAUGs in 5' UTR sequences to gene expression for
671 human heavy chain V genes. **(c)** The functional relevance of leader sequences to gene expression
672 for human heavy chain V genes. L1, L2, and L3 represent unique leader sequences. L1 and L2 in
673 different subplots are not necessarily the same. The number in the center of or on top of each bar
674 mean the number of samples in each group. The vertical lines mean the standard errors. n. a. not
675 available; n. s., not significant ($p \geq 0.5$); *, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.001$.

676

Table 1. Overview of discovered AUSs.

Species	Chain	# unique AUSs	# unique 5' UTRs	# unique leaders	# alleles	# genes	% genes
Human	IGH	1111	1023	346	205	61	47.3
	IGk	498	458	189	138	43	44.3
	IGL	161	152	76	65	33	56.9
RM	IGH	5	5	2	2	2	3.8
	IGK	50	48	31	24	24	20.3
CM	IGH	136	132	69	52	52	35.1
Mouse	IGH	239	237	118	110	106	30.8
	IGK	256	256	96	93	91	54.2
Rat	IGH	399	379	162	126	126	32.3
	IGK	102	96	70	60	60	36.8
Total	-	2957	2786	1159	-	-	-

Table 2. Number and percentage of genes with identified uAUGs.

Species	Chain	# uAUG-containing AUSs	% uAUG-containing AUSs	# uAUGs-containing genes	% uAUGs-containing genes
Human	IGH	138	12.4	23	37.7
	IGK	59	11.9	12	27.9
	IGL	14	8.7	4	12.1
RM	IGH	0	0.0	0	0.0
	IGK	32	64.0	15	62.5
CM	IGH	33	24.3	21	40.4
Mouse	IGH	75	31.4	40	37.7
	IGK	69	27.0	36	39.6
Rat	IGH	65	16.3	27	21.4
	IGK	28	27.5	18	30.0