Agenda of IARC meeting 84, October 25th, 2021

In attendance: Ayelet Peres, Gur Yaari, Andrew Collins, Martin Corcoran, Corey Watson, William Lees, Mats Ohlin, James Heather (guest)

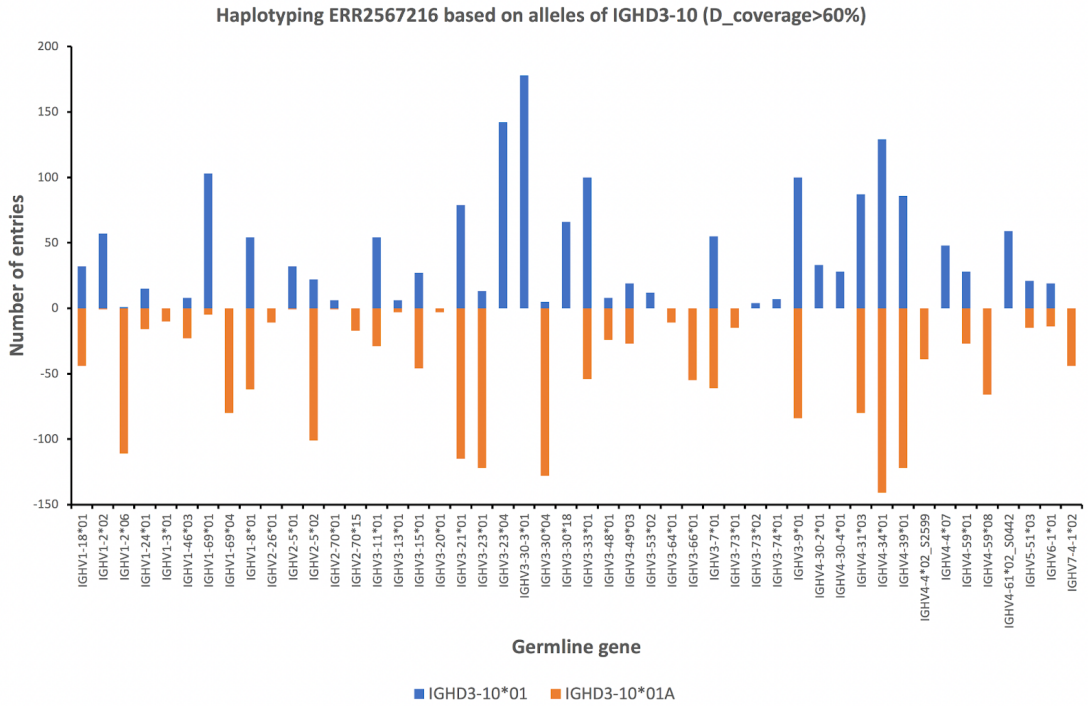
1. **Approval of minutes of meeting 83**
   Approved

2. **Inference of IGHV4-39*07_C288A (IGHV4-39*i02)**
   IGHV4-39*07_C288A was inferred in subject S29 (VDJ-base: P1_I29_S1). This inference has previously been pre-assessed at IARC meeting 63 (https://www.antibodysociety.org/wordpress/wp-content/uploads/2020/12/Meeting-63-24_11_20-minutes.pdf). The genotype also carried IGHV4-39*01. The inference was supported by many sequences (1304) and unmutated sequences (1107) a balanced allelic frequency (52%), a high overall frequency in the unmutated population (5.7%) and many unique CDR3s (1086) in the unmutated sequence set. Haplotyping based on allelic diversity in IGHJ6 was possible and the alleles distributed well between haplotypes (IGHV4-39*07_C288A: 99:1; IGHV4-39*01: 0:100). IARC infers the sequence at level 1 up to and including base 319 in agreement with past practice. It is acknowledged that the allele most likely carries one additional base, typically A at base position 320. The allele is given the name IGHV4-39*i02.

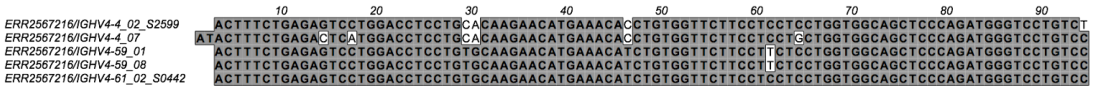
   >IGHV4-39*07_C288A
   CAGCTGCAGCTGCAGGAGTCGGGCCCAGGACTGGTGAAGCCTTCGGAG
   ACCCTGTCCCTCACCTGCACTGTCTCTGGTGGCTCCATCAGCAGTAGTAG
   TTACTACTGGGGCTGGATCCGCCAGCCCCCAGGGAAGGGGCTGGAGTG
   GATTGGGAGTATCTATTATAGTGGGAGCACCTACTACAACCCGTCCCTCAA
   GAGTCGAGTCACCATATCAGTAGACACGTCCAAGAACCAGTTCTCCCTGA
   AGCTGAGCTCTGTGACCGCAGCGGACACGGCCGTGTATTACTGTGCGAG
   AG.

3. **Inference of IGHV4-4*02_A106G (IGHV4-4*i03)**
   IGHV4-4*02_A106G was inferred in subject S39 (P1_I42_S1; ERR2567216). This inference has previously been pre-assessed at IARC meeting 63 (https://www.antibodysociety.org/wordpress/wp-content/uploads/2020/12/Meeting-63-24_11_20-minutes.pdf). The inference was supported by many sequences (294) and unmutated sequences (253) a high allelic frequency (64%), a high overall frequency in the unmutated population (1.5%) and many

unique CDR3s (243) in the unmutated sequence set. Haplotyping based on alleles of IGHJ6 was not possible. Haplotyping based on different variant sequences of IGHD3-10*01 (DOI: 10.3389/fimmu.2019.00987), one of which is not recorded in the IMGT database, was possible (but was not part of the OGRDB submission as the variant IGHD allele is not present in the database used for inference). Separate analysis following IgDiscover-based inference suggested complete separation of IGHV4-4*02_A106G (IGHV4-4*02_S2599) relative to the other allele of IGHV4-4. Overall, the genotype also carried IGHV4-4*07, IGHV4-59*01, IGHV4-59*08, and IGHV4-61*02_A234G among this set of similar genes. The upstream regions of all these alleles have been inferred in this subject in the past (DOI: 10.3389/fimmu.2021.730105) and the upstream region of IGHV4-4*02_A106G is different from those of the other alleles of this set of genes.



Upstream sequences of alleles of the IGHV4-4/IGHV4-59/IGHV4-61 set of genes present in this data set

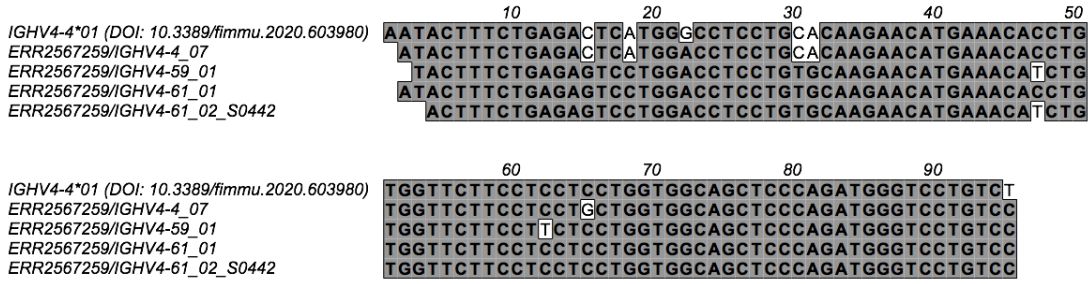


IARC now affirms, based on the extensive validation, the sequence at level 1 up to and including base 319 in agreement with past practice. It is acknowledged that the allele most likely carries one additional base, typically A at base position 320. We recognize that alleles of IGHV4-4/IGHV4-59/IGHV4-61 may residue in gene locations different from that associated to the most similar allele in the IMGT database. Although there is nothing in the data, including haplotype assessment, that suggest that

the allele does not reside in gene IGHV4-4, it must be recognized that IARC gene naming does not reflect a position on the precise gene location of the allele to a specific gene. The allele is given the name IGHV4-4*i03.

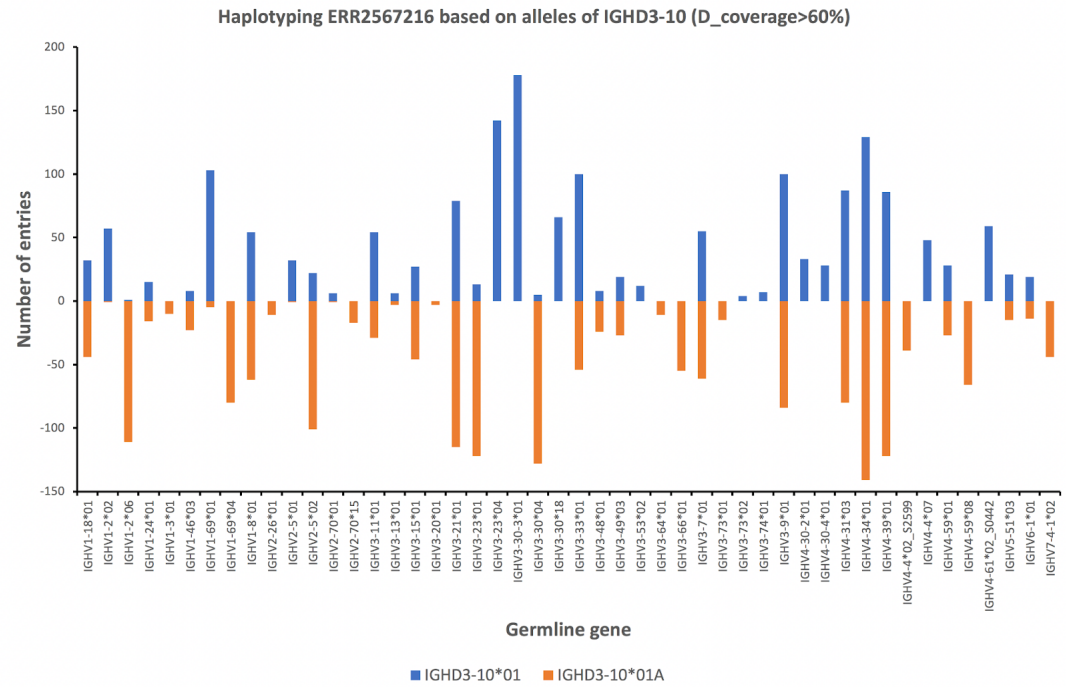
>IGHV4-4*i03
CAGGTGCAGCTGCAGGAGTCGGGCCCAGGACTGGTGAAGCCTTCGGG
GACCCTGTCCCTCACCTGCGCTGTCTCTGGTGGCTCCATCAGCAGTGGT
AACTGGTGGAGTTGGGTCCGCCAGCCCCCAGGGAAGGGGCTGGAGTG
GATTGGGGAAATCTATCATAGTGGGAGCACCAACTACAACCCGTCCCTCA
AGAGTCGAGTCACCATATCAGTAGACAAGTCCAAGAACCAGTTCTCCCTG
AAGCTGAGCTCTGTGACCGCCGCGGACACGGCCGTGTATTACTGTGCGA
GAG.

4. **Inference of IGHV4-61*02_A234G (IGHV4-61*i03)**
IGHV4-61*02_A234G was inferred in subject S82 (P1_I86_S1; ERR2567259). This inference has previously been pre-assessed at IARC meeting 63 (https://www.antibodysociety.org/wordpress/wp-content/uploads/2020/12/Meeting-63-24_11_20-minutes.pdf). The inference was supported by many sequences (613) and unmutated sequences (530) a high allelic frequency (84%), a high overall frequency in the unmutated population (2.2%) and many unique CDR3s (528) in the unmutated sequence set. The related gene set IGHV4-4, IGHV4-59 and IGHV4-61 were represented by the following inferred alleles IGHV4-4*07 (haplotyping ratio: 2:98), IGHV4-59*01 (haplotyping ratio: 36:64), IGHV4-61*01 (haplotyping ratio 100:0), and IGHV4-61*02_A234G (haplotyping ratio: 0:100). In a separate study it has been proposed that the IGHJ6*02-related haplotype in this subject was occupied by poorly expressed allele IGHV4-4*01 (DOI: 10.3389/fimmu.2020.603980). IGHV4-61*02_A234G was found to be expressed at higher frequency than the poorly expressed allele IGHV4-61*01 (528 vs 107 unique CDR3s in the unmutated sequence set). Overall expression levels, diversity and haplotyping strongly supports the inference. In addition, the upstream region of IGHV4-61*02_A234G (DOI: 10.3389/fimmu.2021.730105) in this subject is different from that of the other alleles of these three genes that are present in this subject (including the previously published upstream region of IGHV4-4*01).



>IGHV4-61*i03
CAGGTGCAGCTGCAGGAGTCGGGCCCAGGACTGGTGAAGCCTTCACAG
ACCCTGTCCCTCACCTGCACTGTCTCTGGTGGCTCCATCAGCAGTGGTA
GTTACTACTGGAGCTGGATCCGGCAGCCCGCCGGGAAGGGACTGGAGT
GGATTGGGCGTATCTATACCAGTGGGAGCACCAACTACAACCCCTCCCTC
AAGAGTCGAGTCACCATGTCAGTAGACACGTCCAAGAACCAGTTCTCCCT
GAAGCTGAGCTCTGTGACCGCCGCAGACACGGCCGTGTATTACTGTGCG
AGAG.

It has been noted that IGHV4-61*02_A234G was also found in subject S39 (P1_I42_S1; ERR2567216) used to assess the inference of IGHV4-4*i03 (although this data set was not submitted for assessment of IGHV4-61*02_A234G by IARC). In this sample, IGHV4-61*02_A234G was supported by many sequences (476) and unmutated sequences (395), a high overall frequency in the unmutated population (2.4%) and many unique CDR3s (390) in the unmutated sequence set. No other allele named as IGHV4-61 had been inferred in this subject but IGHV4-59*08, an allele suspected to residue in gene IGHV4-61 (DOI: 10.1038/ncomms14946), was found. The allelic ratio of IGHV4-61*02_A234G in this context would be 61%. Haplotyping of this subject was possible based on an allele of IGHD3-10 (10.3389/fimmu.2019.00987) not documented by the IMGT database. In this context, the two alleles of the genotype believed to be associated to IGHV4-61, IGHV4-61*02_A234G and IGHV4-59*08, showed excellent separation (see also documentation associated to the affirmation of IGHV4-4*i03).



IARC, taking all of the collective evidence under consideration, affirms the sequence at level 1 up to and including base 319 in agreement with past practice. It is acknowledged that the allele most likely carries one additional base, typically A at base position 320. Although there is nothing in the data, including haplotype assessment, that suggest that the allele does not reside in gene IGHV4-61, it must be recognized, in particular in view of the similarity of alleles linked to IGHV4-4/IGHV4-59/IGHV4-61, that IARC gene naming does not reflect an absolute position on the precise location of the allele to a specific gene. The allele is given the name IGHV4-61*i03.

5. **Followup: The problem of identical sequences associated to different/duplicated genes in germline gene databases, the impact of multiple assignments in sequence analysis and VDJbase's representation of genotypes/haplotypes, and the concept of germline gene subgroups even in well-studied organisms**
   AP presents new analysis of similarity and grouping of human IGHV alleles. Discussion.

6. **Next meeting**
   November 1st, 2021 at 10.00 UTC