IRF(V1.0, immune repertoire format) specification

### IRF specification

IRF is a text-based format for storing TCR and BCR repertoire. In the format, each line typically represents the annotated information of unique nucleotide sequence. The annotated information include the alignment records and structure of clone such as CDR (Complementarity-determining region) and FR (framework region) region. If paired two chains, such as TCR α and β or heavy chain and light chain, are sequenced at the same time, the paired sequences are deposited at the adjacent two lines, which can be identified by the same ID. Each line has 27 mandatory fields. These fields always appear in the same order and must be present, but their values can be used “na” instead if the corresponding information is unavailable.

### An example



### Specific format

1. **ID**: sequence id, the id is unique in the sample and the paired sequence derived from paired chains (such as T cell α and β chains, or B cell heavy and light chains) use the same id.
2. **locus**: represents the single chain, such as TRA, TRB, TRD, TRG, IGH, IGK, IGL.
3. **functional**: the sequence is productive or unproductive. The productive sequence is marked as “in-frame” and the unproductive sequences are marked by “out-of-frame(CDR3 length)”, “out-of-frame(stop codon)” and “pseudogene”. “out-of-frame(CDR3 length)”: the length of CDR3 nucleotide sequence is not the multiple of 3. “out-of-frame(stop codon)”: the sequence contains a stop codon. “pseudogene”: the sequence is mapped to one of pseudo- V, D or J gene segment.
4. **vGene**: identified V(variable) gene segment for the sequence. The reference genes are derived from IMGT database.
5. **dGene**: identified D(diversity) gene segment for the sequence. The reference genes are derived from IMGT database.
6. **jGene**: identified V(joining) gene segment for the sequence. The reference genes are derived from IMGT database.
7. **cGene**: identified C(constant region) gene segment for the sequence. The reference genes are derived from IMGT database.
8. **ntCDR3**: CDR3 nucleotide sequence
9. **aaCDR3**: CDR3 amino acid sequence
10. **ntFragments**: the nucleotide sequence for each fragments, the displayed order like this: FR1;CDR1;FR2;CDR2;FR3;FR4:C. If there is no sequence for a specific fragment, such as FR1, “na” is used to instead. If there are partial of sequence for a specific fragment, ellipsis “…” is added the end of sequence, such as “…AACG” or “CGGG…”.
11. **aaFragments**: the amino acid sequence for each fragments, the displayed order like this: FR1;CDR1;FR2;CDR2;FR3;FR4:C. If there is no sequence for a specific fragment, such as FR1, “na” is used to instead. If there are partial of sequence for a specific fragment, ellipsis “…” is added the end of sequence, such as “…LYI” or “TVKP…”.
12. **CDR3Pos**: the start and end of CDR3 positions in both nucleotide and amino acid sequences. The start and end of positions are linked by “-”, and the positions of nucleotide and amino acid are separated by “;”. The format like this: start(nt)-end(nt);start(aa)-end(aa), for example, “214-257;91-105”.
13. **ntFragmentsPos**: the start and end of positions for each fragment in nucleotide sequence. The start and end of positions are linked by “-”, and the positions of fragments are separated by “;”. The format like this: start(FR1)-end(FR1);start(CDR1)-end(CDR1);start(FR2)-end(FR2);start(CDR2)-end(CDR2);start(FR3)-end(FR3);start(FR4)-end(FR4);start(C)-end(C). If there is no sequence for a specific fragment, such as FR1, “na” is used to instead. If there are partial of sequence for a specific fragment, “<” or “>” is added the existed position, such as “<1-15” or “402->450”.
14. **aaFragmentsPos**: the start and end of positions for each fragment in amino acid sequence. The start and end of positions are linked by “-”, and the positions of fragments are separated by “;”. The format like this: start(FR1)-end(FR1);start(CDR1)-end(CDR1);start(FR2)-end(FR2);start(CDR2)-end(CDR2);start(FR3)-end(FR3);start(FR4)-end(FR4);start(C)-end(C). If there is no sequence for a specific fragment, such as FR1, “na” is used to instead. If there are partial of sequence for a specific fragment, “<” or “>” is added the existed position, such as “<1-15” or “102->120”.
15. **v3Deletion**: the deleted nucleotide bases of 3’ end of V gene segment. If the deleted bases cannot be identified (such as because of no V gene assignment), “na” is used to instead. However, if no base is deleted during the VDJ rearrangement, “0” is used here.
16. **d5Deletion**: the deleted nucleotide bases of 5’ end of D gene segment. If the deleted bases cannot be identified (such as because of no D gene assignment), “na” is used to instead. However, if no base is deleted during the VDJ rearrangement, “0” is used here.
17. **d3Deletion**: the deleted nucleotide bases of 3’ end of D gene segment. If the deleted bases cannot be identified (such as because of no D gene assignment), “na” is used to instead. However, if no base is deleted during the VDJ rearrangement, “0” is used here.
18. **j5Deletion** : the deleted nucleotide bases of 3’ end of J gene segment. If the deleted bases cannot be identified (such as because of no J gene assignment), “na” is used to instead. However, if no base is deleted during the VDJ rearrangement, “0” is used here.
19. **vdInsertion**: the inserted nucleotide bases between the junctional region of V and D genes. If the inserted bases cannot be identified (such as because V or D gene is not assigned), “na” is used to instead. However, if no base is inserted during the VDJ rearrangement, “0” is used here.
20. **djInsertion**: the inserted nucleotide bases between the junctional region of D and J genes. If the inserted bases cannot be identified (such as because D or J gene is not assigned), “na” is used to instead. However, if no base is inserted during the VDJ rearrangement, “0” is used here.
21. **vjInsertion**: the inserted nucleotide bases between the junctional region of V and J genes. This field is used for TCR α and γ chains, and BCR light chain. If the inserted bases cannot be identified (such as because V or J gene is not assigned), “na” is used to instead. However, if no base is inserted during the VDJ rearrangement, “0” is used here.
22. **originalSource**: the sequence is derived from plus or minus chain of chromosome
23. paired: “1” represents the paired chain (TCR α and β chains, TCR γ and δ chains, BCR heavy and light chains) is also sequenced and deposited in adjacent line. “0” represents there is no information of paired chain.
24. **seqCount**: abundance of nucleotide sequence
25. **ntSequence**: nucleotide sequence
26. **aaSequence**: amino acid sequence
27. **mapInformation**: mapped information for V, D and J assignment for the sequence. The format like this: vIdentity=;vLen=;vMismatch=;vGap=;vSeqPos=;vRefPos=;dIdentity=;dLen=;dMismatch=;dGap=;dSeqPos=;dRefPos=; jIdentity=;jLen=;jMismatch=;jGap=;jSeqPos=;jRefPos=; for example, vIdentity=98.17;vLen=109;vMismatch=2;vGap=0;vSeqPos1-109=;vRefPos=183-291;dIdentity=100;dLen=7;dMismatch=0;dGap=0;dSeqPos=113-119;dRefPos=3-9; jIdentity=88.64;jLen=44;jMismatch=5;jGap=0;jSeqPos=124-167;jRefPos=4-47;