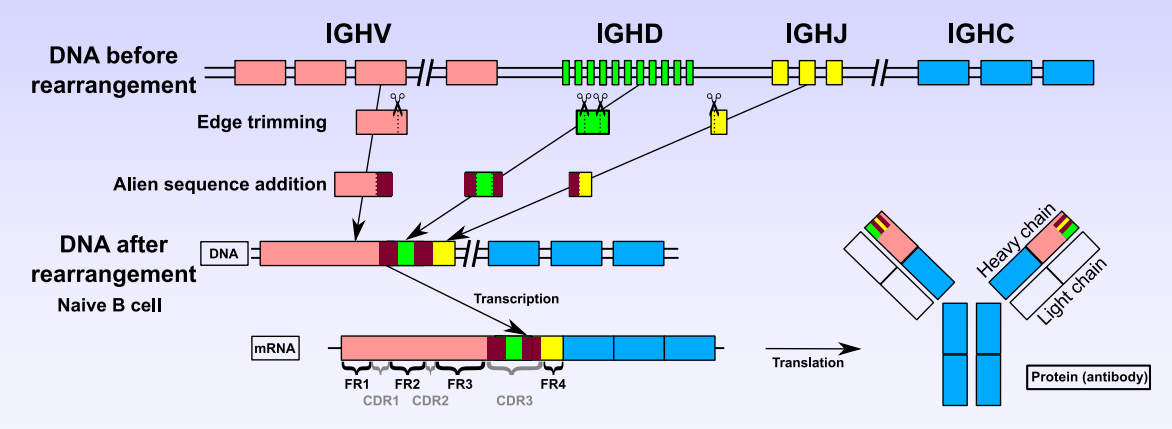
**V(D)J recombination - combinatorial and junctional diversity:**



**Good presentation for BLAST algorithm:**

<https://www.slideshare.net/Fardin6600/blast-algorithm>

**IMGT sequences:**

V-Genes:

<http://www.imgt.org/IMGTrepertoire/Proteins/alleles/list_alleles.php?species=Homo%20sapiens&group=IGHV>

D-Genes:

<http://www.imgt.org/IMGTrepertoire/Proteins/alleles/list_alleles.php?species=Homo%20sapiens&group=IGHD>

J-Genes:

<http://www.imgt.org/IMGTrepertoire/Proteins/alleles/list_alleles.php?species=Homo%20sapiens&group=IGHJ>

**Alignment explanation through Example:**

Example below is taken from HCV\_B data, file CI10.tab, 5th sequence (SEQUENCE\_ID CTAAGAGCGGGAGCGCA).

Grey background format words are the headers of the table above.

SEQUENCE\_INPUT – this is the sequence input that is aligned to IMGT sequences database.

GGCTTTCTGAGAGTCATGGATCTCTTGTGCAAGAAGATGAAGACCTGTGGTTCTTCCTCCTGCTGGTGGCGGCTCCCAGATGGGTCCTGTCCCAGTTGCAGCTGCAGGAGTCGGGCCCAGGGGTGGTGAAGCCTTCGGAGACCCTGTCCCTCNCCTGCAATGTCTCTGGTGGCTCCATCANCAGTTCTGGCTTCTACTGGGCCTGGATCCGCCAGCCCCCAGGGAAGGGGCTGGAGTGGATTGGGACTATCCACCACAGTGGGAAAACCAACCAGAACCCGTCCCTCAAGAGTCGAGTCATCATCTCCGTAGACACGTCCAAAAATCAGTTCTCCCTGAAGATGAGATCCATGACCGCCGCAGACACGGCTGTCTATTACTGTGCCAGACAGGAGACAACCTATGGCTCGGGGACCTATTATTTCAGGTACTACTTCTACGGTCTGGACGTCTGGGGCCAAGGGGCCACGGTCACCGTCTCCTCAGCCTCCACCAAGGGCC

The bellow are the V,D,J genes that were identified (by the Igblast) and are in the database.

V\_CALL: IGHV4-39\*01

D\_CALL: IGHD3-10\*01

J\_CALL: IGHJ6\*02

Their sequences are as follows from IMGT sequences sites above.

IGHV4-39\*01:

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24

cag ctg cag ctg cag gag tcg ggc cca ... gga ctg gtg aag cct tcg gag acc ctg tcc ctc acc tgc act

25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48

gtc tct ggt ggc tcc atc agc ... ... agt agt agt tac tac tgg ggc tgg atc cgc cag ccc cca ggg aag

49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 ggg ctg gag tgg att ggg agt atc tat tat agt ... ... ... ggg agc acc tac tac aac ccg tcc ctc aag

73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96

... agt cga gtc acc ata tcc gta gac acg tcc aag aac cag ttc tcc ctg aag ctg agc tct gtg acc gcc

97 98 99 100 101 102 103 104 105 106

gca gac acg gct gtg tat tac tgt gcg aga ca

IGHD3-10\*01:

1 2 3 4 5 6 7 8 9 10

gta tta cta tgg ttc ggg gag tta tta taa c

IGHJ6\*02:

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20

AT TAC TAC TAC TAC TAC GGT ATG GAC GTC TGG GGC CAA GGG ACC ACG GTC ACC GTC TCC TCA G

V\_SEQ\_START: Position of first V nucleotide in SEQUENCE\_INPUT. In our example - the value is 93.

V\_SEQ\_LENGTH: Number of V nucleotides in SEQUENCE\_INPUT. In our example - the value is 299.

N1\_LENGTH: the length of the N1 segment, which is the segment between V and D segments. In our example – the value is 9.

D\_SEQ\_START: Position of first D nucleotide in SEQUENCE\_INPUT. In example above - the value is 401.

D\_SEQ\_LENGTH: Number of D nucleotides in SEQUENCE\_INPUT. In example above - the value is 14.

N2\_LENGTH: The length of the N2 segment, which is the segment between D and J segments. In our example – the value is 14.

J\_SEQ\_START: Position of first J nucleotide in SEQUENCE\_INPUT. In example above - the value is 429.

J\_SEQ\_LENGTH: Number of J nucleotides in SEQUENCE\_INPUT. In example above - the value is 57.

Below is the SEQUENCE\_INPUT above, with all its segments:

GGCTTTCTGAGAGTCATGGATCTCTTGTGCAAGAAGATGAAGACCTGTGGTTCTTCCTCCTGCTGGTGGCGGCTCCCAGATGGGTCCTGTCCCAGTTGCAGCTGCAGGAGTCGGGCCCAGGGGTGGTGAAGCCTTCGGAGACCCTGTCCCTCNCCTGCAATGTCTCTGGTGGCTCCATCANCAGTTCTGGCTTCTACTGGGCCTGGATCCGCCAGCCCCCAGGGAAGGGGCTGGAGTGGATTGGGACTATCCACCACAGTGGGAAAACCAACCAGAACCCGTCCCTCAAGAGTCGAGTCATCATCTCCGTAGACACGTCCAAAAATCAGTTCTCCCTGAAGATGAGATCCATGACCGCCGCAGACACGGCTGTCTATTACTGTGCCAGACAGGAGACAACCTATGGCTCGGGGACCTATTATTTCAGGTACTACTTCTACGGTCTGGACGTCTGGGGCCAAGGGGCCACGGTCACCGTCTCCTCAGCCTCCACCAAGGGCC

|  |
| --- |
| V-seq N1-seq D-seq N2-seq J-seq |

SEQUENCE\_VDJ: This is the sequence of only the V-N1-D-N2-J segments above.

CAGTTGCAGCTGCAGGAGTCGGGCCCAGGGGTGGTGAAGCCTTCGGAGACCCTGTCCCTCNCCTGCAATGTCTCTGGTGGCTCCATCANCAGTTCTGGCTTCTACTGGGCCTGGATCCGCCAGCCCCCAGGGAAGGGGCTGGAGTGGATTGGGACTATCCACCACAGTGGGAAAACCAACCAGAACCCGTCCCTCAAGAGTCGAGTCATCATCTCCGTAGACACGTCCAAAAATCAGTTCTCCCTGAAGATGAGATCCATGACCGCCGCAGACACGGCTGTCTATTACTGTGCCAGACAGGAGACAACCTATGGCTCGGGGACCTATTATTTCAGGTACTACTTCTACGGTCTGGACGTCTGGGGCCAAGGGGCCACGGTCACCGTCTCCTCA

SEQUENCE\_IMGT: w/o the dots – it is exact same as SEQUENCE\_VDJ.

CAGTTGCAGCTGCAGGAGTCGGGCCCA...GGGGTGGTGAAGCCTTCGGAGACCCTGTCCCTCNCCTGCAATGTCTCTGGTGGCTCCATCANC......AGTTCTGGCTTCTACTGGGCCTGGATCCGCCAGCCCCCAGGGAAGGGGCTGGAGTGGATTGGGACTATCCACCACAGT.........GGGAAAACCAACCAGAACCCGTCCCTCAAG...AGTCGAGTCATCATCTCCGTAGACACGTCCAAAAATCAGTTCTCCCTGAAGATGAGATCCATGACCGCCGCAGACACGGCTGTCTATTACTGTGCCAGACAGGAGACAACCTATGGCTCGGGGACCTATTATTTCAGGTACTACTTCTACGGTCTGGACGTCTGGGGCCAAGGGGCCACGGTCACCGTCTCCTCA

V\_GERM\_START\_IMGT: Position of first V-gene nucleotide from SEQUENCE\_INPUT in SEQUENCE\_IMGT. (Nucleotide C underlined above). Is equal to 1 in our example.

V\_GERM\_LENGTH\_IMGT: Position of last V-gene nucleotide from SEQUENCE\_INPUT in SEQUENCE\_IMGT. (Nucleotide A double underlined above). Is equal to 320 in our example.

D\_GERM\_START: Position of first nucleotide of SEQUENCE\_INPUT D Gene, in the D-Gene of the germline (Since the D-gene is cut from both sides SEQUENCE\_IMGT). In our example it is 7 (underlined nucleotide C below):

This is the Germline D-gene: gtattactatggttcggggagttattataac

This is the SEQUENCE\_INPUT D-gene: CTATGGCTCGGGGA

D\_GERM\_LENGTH: Length of the Germline D alignment (usually equal to the length of the D-sequence part in SEQUENCE\_INPUT). In our example it is 14.

J\_GERM\_START: Position of first nucleotide of SEQUENCE\_INPUT J Gene, in the J-Gene of the germline (Since the J-gene is cut from left). In our example it is 6 (underlined nucleotide C below):

This is the Germline J-gene: ATTACTACTACTACTACGGTATGGACGTCTGGGGCCAAGGGACCACGGTCACCGTCTCCTCAG

This is the SEQUENCE\_INPUT J-gene: TACTACTTCTACGGTCTGGACGTCTGGGGCCAAGGGGCCACGGTCACCGTCTCCTCA

J\_GERM\_LENGTH: Length of the Germline J alignment (usually equal to the length of the J-sequence part in SEQUENCE\_INPUT). In our example it is 57.

JUNCTION\_LENGTH: Number of junction nucleotides in SEQUENCE\_VDJ. The junction (that includes the CDR3 region) - includes a portion of the end of V-Gene + N1 + D-Gene + N2 + a portion of the beginning of J-Gene. In our example the Junction length is 75. The region of the junction can be seen in IMGT sequences above.

JUNCTION:Junction region nucleotide sequence. Below is the SEQUENCE\_VDJ with the JUNCTION region **bolded**. The CDR3 is 2 codons less than the junction itself, as the left most codon of the junction belongs to FWR3, and the right most codon of the junction belongs to FWR4. In example below- the CDR3 region is underlined.

CAGTTGCAGCTGCAGGAGTCGGGCCCAGGGGTGGTGAAGCCTTCGGAGACCCTGTCCCTCNCCTGCAATGTCTCTGGTGGCTCCATCANCAGTTCTGGCTTCTACTGGGCCTGGATCCGCCAGCCCCCAGGGAAGGGGCTGGAGTGGATTGGGACTATCCACCACAGTGGGAAAACCAACCAGAACCCGTCCCTCAAGAGTCGAGTCATCATCTCCGTAGACACGTCCAAAAATCAGTTCTCCCTGAAGATGAGATCCATGACCGCCGCAGACACGGCTGTCTATTAC**TGTGCCAGACAGGAGACAACCTATGGCTCGGGGACCTATTATTTCAGGTACTACTTCTACGGTCTGGACGTCTGG**GGCCAAGGGGCCACGGTCACCGTCTCCTCA

CONSCOUNT: Number of reads contributing to the UMI consensus sequence. Meaning – number of sequences with same UMI (originating from same source). In our example this value is 7.

DUPCOUNT: Number of exact same sequences but different UMI. In our example it is 1.

**Following fields are calculated:**

CLONE: An integer that is same for all sequences with same SEQUENCE\_IMGT, and is different between sequences with different SEQUENCE\_IMGT\_D\_MASK. All sequences in one clone have same V,D,J alleles, and same junction length.

**Region definitions (CDR and FWR):**

| Region | Range in nucleotides in SEQUENCE\_IMGT | Range in Amino acids in SEQUENCE\_IMGT | Range for number of Amino Acids (depending on the V/J gene and allele) | Comments |
| --- | --- | --- | --- | --- |
| FWR1 | 1 - 78 | 1 - 26 | 25-26 |  |
| CDR1 | 79 - 114 | 27 - 38 | 5-12 |  |
| FWR2 | 115 - 165 | 39 - 55 | 16-17 |  |
| CDR2 | 166 - 195 | 56 - 65 | 0-10 |  |
| FWR3 | 196 - 312 | 66 - 104 | 36-39 |  |
| CDR3 | 313 - 351 | 105 - 117 | 2-13 | Can be much larger than 13 Amino Acids in none germline genes (after V(D)J recombination). |
| FWR4 | 352 - 387 | 118 - 129 | 10-12 |  |

The above is based on:

<http://www.imgt.org/IMGTScientificChart/Nomenclature/IMGT-FRCDRdefinition.html>

and

<http://www.imgt.org/IMGTScientificChart/Numbering/IMGTIGVLsuperfamily.html>

**Codon table:**

