| Function | Major input | Call functions | output |
| --- | --- | --- | --- |
| make\_clones\_freq\_df | Any data frame that have a clone clumn | --- | A 2 column data frame: CLONE, CLONE\_FREQ (that includes frequency of each clone in input data frame) |
| make\_clones\_list | ChangeO db | --- | Returns a (unique) list of all clones in db. |
| make\_chaneoClone\_cur\_clone | ChangeO db,  Clone\_num | makeChangeoClone | An obejcts with 2 slots:   * CahngeoClone object of the specific clone\_num * Clone size (after collapsing relevant lines) |
| make\_graph\_cur\_clone | ChangeoClone object | buildPhylipLineage | Igraph object of the current clone lineage tree. |
| make\_graph\_df | ChangeoClone object,  Igraph object | summarizeSubtrees | A db containing merged data from clones changeoClone object and its igrpah object. Specifically adding PARENT and PARENT\_SEQUENCE columns, adding raws for inferred sequences and renaming sequences ID to unique easier ones (related to clone). |
| draw\_clone\_lineage\_tree | ChangeO db,  Clone\_cum | make\_chaneoClone\_cur\_clone  make\_graph\_cur\_clone | Plot of a lineage tree of specific clone\_num |
| make\_region | JUNCTION\_LENGTH  SEQUENCE\_IMGT  reg\_type |  | An object of type RegionDefinition that includes all FWR and CDR regions definitions per IMGT definitions. According to “reg\_type” input – can give either 7 regions (CDR1/2/3,FWR1/2/3/4) or 2 regions (CDR/FWR). But in both cases – the full sequence is covered. |
| get\_clone\_region | ChangeO db  Clone\_num | make\_region | An object of type RegionDefinition for the specific clone\_num in the db. |
| make\_regions\_dict | ChaneO db | get\_clone\_region | A list. Element names are clone numbers. Each list element includes 2 sub elements of type RegionDefinition that are related to specific clone (one of each type described in make\_region function). |
| observedMutations\_L | ChangeO db | make\_region  observedMutations | ChangeO db with observed mutations columns per region, where the reference sequence is not the germline sequence, but the parent sequence. |
| expectedMutations\_L | ChangeO db | make\_region  expectedMutations | ChangeO db with expected mutations columns per region, where the reference sequence is not the germline sequence, but the parent sequence. |
| collapse\_one\_clone | ChangeO db  Clone\_num | get\_clone\_region  collapseClones | A db for the specific chosen clone - with added column – for the consensus sequence of this clone. |
| collapseClones\_L | ChangeO db | make\_clones\_list  collapse\_one\_clone | Same as original collapseClones function, except that the region definition will be set per clone in the database, in order to include CDR3 and FWR4 in the region. |
| calcBaseline\_one\_clone | ChangeO db  Clone\_num | get\_clone\_region  calcBaseline | Computes calcBaseline on the specific clone in db. |
| calcBaseline\_L | ChangeO db | make\_clones\_ calcBaseline\_one\_clone | Very similar to calcBaseline, but it enables using a different region definition for each clone (per its junction length), including CDR3 and FWR4 as part of the region definition. |
| groupBaseline\_L | Baseline\_L object | Basleine | Same as groupBaseline, but works on Baseline\_L object |

**Pipeline stages:**

Data preparation for Analysis:

1. Load data to R data frame.

Note that the data frame should be of type ChangeO, specifically following columns must be there:

* + SEQUENCE\_ID
  + SEQUENCE\_INPUT
  + SEQUENCE\_IMGT
  + JUNCTION\_LENGTH
  + JUNCTION
  + GERMLINE\_IMGT
  + CLONE

1. Removing from db sequences that their IMGT sequence is shorter than 312 nucleotides. The reason for this is that running Expected mutations on them does not run properly, as they are missing full range of FWR1/CDR1/FWR2/CDR2/FWR3 which is 312 nucleotides.
2. Create a list of clones to loop on.
3. Loop on all clones in db to generate an extended db that includes lineage information (parent sequence id, parent sequence). While doing so – clonesof size 1 – are put aside ona separate db (to be taken care of by later stages).
4. Taking care of clones of size 1: preparing a db for them.
5. Adding to clones of size 1 db - an additional line for their germline info.
6. Adding db of size 1 clones to main db (all\_clones\_merged\_df)
7. Data Analysis 1 (based on CI10.tab from HCV\_B data sets)– comparing PDFs of 2 methods of grouping:
8. PDFs using consensus clones:

* For each clone – set a consensus clone (using collapseClones\_L)
* Calculate baseline (using calcBaseline\_L) for each consensus clone. The reference sequence for each consensus clone is its germline sequence.
* Group all PDFs of all clones consensus above (using groupBaseline\_L)

1. PDFs using parents sequences:

* Calculate baseline (using calcBaseline\_L) for each read. The reference sequence for each read is its parent sequence.
* Group PDFs of all reads above (using groupBaseline\_L).

For all the above (a and b) – 2 region definitions are used: “ALL” (which are all 7 regions CDR1/2/3 FWR1/2/3/4), and “CDR\_FWR” which are 2 regions that cover CDR1/2/3 and FWR1/2/3/4. In both – the region definition is per clone (as each has its specific junction length).

Results are here:

A close up of a map

Description automatically generated

1. Data Analysis 2 – same as Data Analysis 1, except that db will exclude singletone reads. The reason for that is to get a stronger difference between 2 methods, since singletone reads show equal results between two methods.

Results are here (based on CI10.tab from HCV\_B data sets):

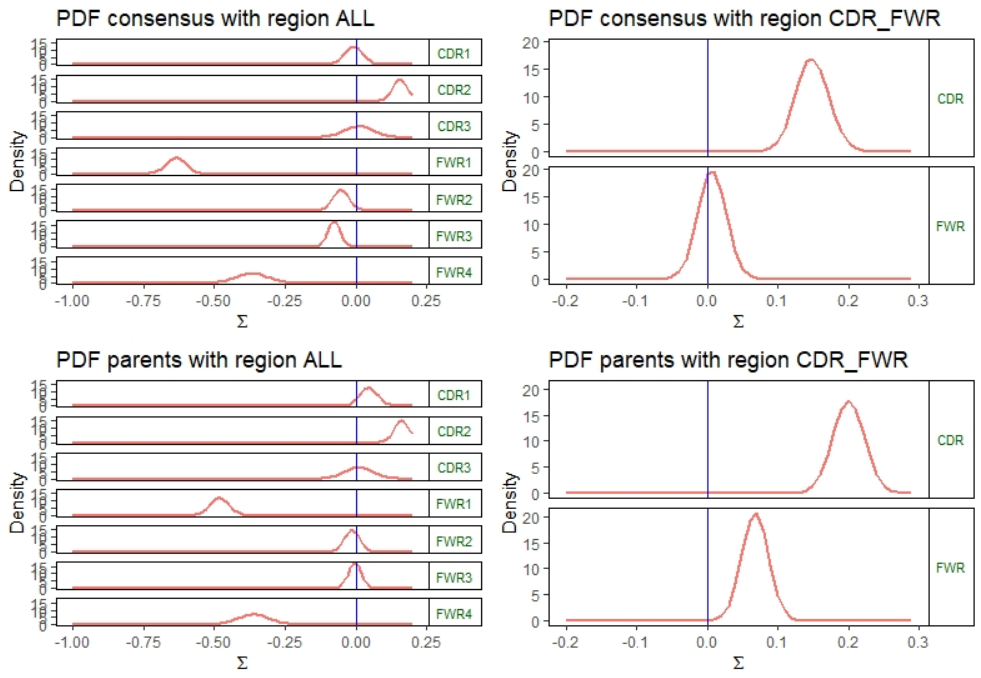
A close up of a map

Description automatically generated

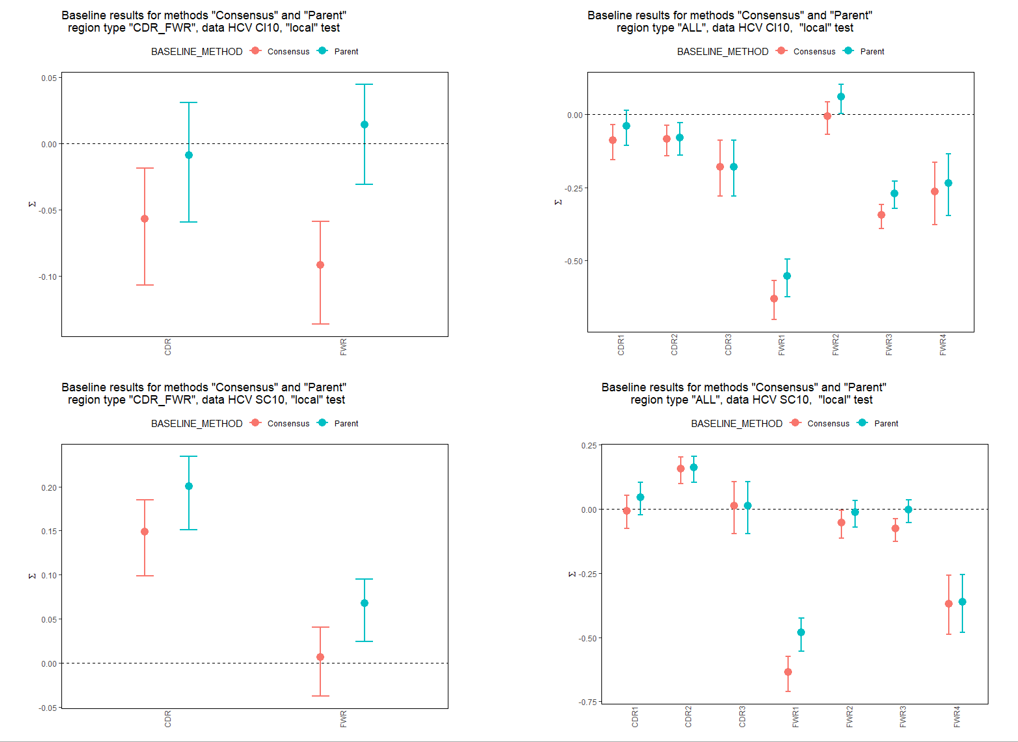
1. Data Analysis 3 - Same as Data Analysis 1, but on clones with up to 10 mutations. Results are below:
2. Data Analysis 4 – Same as Data analysis 1, but on different DB: SC10.tab from HCV\_B data sets.

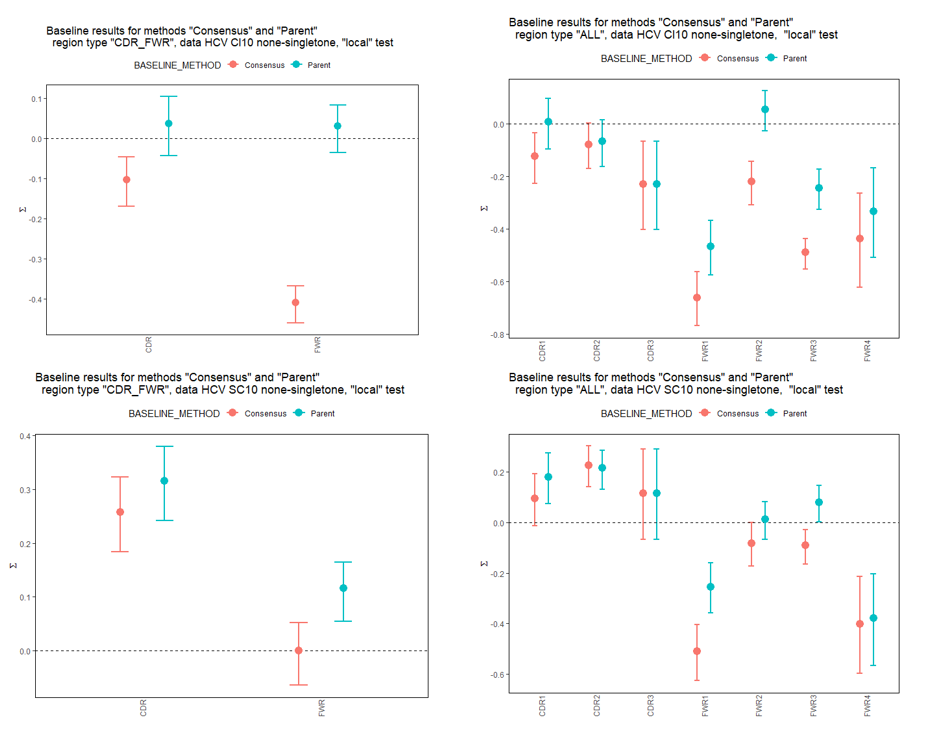
(I had to remove on read from database: SEQUENCE\_ID GAGCATGCCAAAATCAT. Not sure why, but it caused error in calcBaseline\_L function on collapsed clones. This read is a singletone).

Results are here:



1. Here are other plots related to Data Analysis above:





1. Sanity check 1:

This step is to be used as a sanity check (to check calcBaseline\_L and groupBaseline\_L functions):

Comparing results using region definition of IMGT\_V\_BY\_REGIONS (including CDR1/2 FWR1/2/3) to results using region definition of type "ALL" (including the whole sequence - CDR1/2/3 and FWR1/2/3/4).

The PDFs of CDR1/2 and FWR1/2/3 look similar, but they are not identical.

When looking deeper - the conclusion is that they cannot be identical, since the expected mutation in IMGT\_V\_BY\_REGIONS is calculated based on sequence length of 312, while a region definition of type "ALL" is longer than that. So the relative expected mutations for example of FWR1 in "ALL" region will be smaller than the expected mutations for FWR1 in the IMGT\_V\_BY\_REGIONS.

But:

exp\_cdr1/(exp\_cdr1+exp\_cdr2+exp\_fwr1+exp\_fwr2+exp\_fwr3) in region "ALL" is the same as exp\_cdr1 in region IMGT\_V\_BY\_REGIONS.

And same for cdr2, fwr1/2/3.

1. Sanity check 2:

Compare only singletons (expected to get same PDFs in parent vs consensus methods)

1. PDFs using consensus clones.
2. PDFs using parents sequences.

Results are that PDFs are the same.

Also - comparing different objects from 1 and 2 above - are equal (when using all.equal() function)

TODO:

1. Add an option that lineage tree is already done on change db. What is the format of this input: igraph? Data frame with parent column?
2. Other functions beside obserevdMutations and ExpectedMutations that need to enlarge like above? CalcBaseline, GroupBaseline.
3. Data analysis? Take any database as a test case and compare the following:
   1. pdf of mutation strength (using CalcBaseline, GroupBaseline) in old way: taking one “consensus” sequence from each clone and running the pdf against the germline.
   2. Pdf of mutation strength (using CalcBaseline, GroupBaseline) in new way: running it in each clone vs the parent sequence, and convolving all sequences in a clone to one pdf.
4. Compare the following:
   1. P=0.6, x=3, N=5. Calculate pdf
   2. groupBaelines of 5 times : P=0.6,x=1,N=5
5. Improve run time by changing loop to sapply.
6. Take care of clones of length 1
7. Compare to Daniel: done – same results on 2 examples.
8. Sanity check with:
   1. old functions on FW1-FWR3:
   2. Compare only singletons – as sanity check (expected to get same parent vs consensus)
   3. Remove singletons
   4. Another db
   5. Same with only v3\* and only v2\*…
   6. Clones with only up to 10 mutations on all reads.
   7. All patients, and using barplots
9. Add region definition for codons, and check PDFs of each codon using different tests (local, imbalanced, etc)