| 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. |
| 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. |
| calcBaseline\_L |  |  |  |

Pipeline stages:

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).
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. Adding to main db the expected and observed mutation columns.

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. Take care of clones of length 1
5. Compare to Daniel: done – same results on 2 examples.