**Pipeline Ab germline**

1. merge reads with pandaseq, trim random nucleotides/primers at beginning and end of read, collapes unique reads, keep only if more than 10 members

/data/AbX/germline/GermAb/1\_merge\_trim\_collapse.sh /data/MiSeq/MiSeqOutput/XXX/Data/Intensities/BaseCalls/

Input: \_R1.fastq

\_R2.fastq

Script: 1\_merge\_trim\_collapse.sh

contains primer\_trim.py

Output: \_panda.fasta

\_trimmed.fasta

\_uniq.fasta

2. align reads to IMGT reference

Input: \_uniq.fasta

Script: 2\_align.sh

Output: \_aligned.sam

\_aligned.txt

IMGT reference was modified as follows:

IGHV3-23D was deleted (identical to 3-23) -> analysis of 3-23 is that of 3-23 AND 3-23D

IGHV1-69D was deleted (identical to 1-69) -> analysis of 1-69 is that of 1-69 AND 1-69D

IGHV2-70D\*04 was deleted (identical to 2-70\*04), IGHV2-70D\*14 was renamed to IGHV2-70\*14 -> analysis of 2-70 is that of 2-70 AND 2-70D, IGHV2-70D\*04v renamed into IGHV2-70\*04v

3. filter functional Ab seqs, combine identical seqs with 0, 1 and 2 mutations from reference using sam cigar

#the following deletes reads with mutation at position 229 (or 226, depends on primers) (wt: CCAAGAACCAGTT, mut: CCAAGACCCAGTT)

filter(position != 230 | !grepl("IGHV4", allele) | !grepl("A", nt)) %>%

filter(position != 227 | !grepl("IGHV4", allele) | !grepl("A", nt))

-> run R on server (takes too long otherwise) -> delete „Volumes“ in path files for this

Input: \_aligned.txt

Script: 3\_functional\_combine\_identical.R

Output: \_alleles\_comb.txt

4. determine alleles

Input: \_alleles\_comb.txt

Script: 4\_determine\_alleles.sh

contains freq\_drop.py

Output: \_alleles\_final.txt (list of readcount and assigned alleles)

all\_results.txt (list for all patients: number of alleles per gene and patient)

\_final\_results.txt (list of alleles and number of mutations to allele)

**Analysis**

Exclude non-neutralizing patients:

* 16198 (=SB126, score=2) -> not sequenced
* 17420 (score=2)
* 18826 (score=9) -> not sequenced
* 18928 (AK170, score=11), labelled Ak170 in first run
* 26500 (score=0)
* 26586 (score=0)
* 31822 (score=10)
* 31933 (score=11) -> not sequenced
* 34545 (score=12)
* 41895 (ART)
* 42080 (score=10) -> not sequenced
* 42335 (score=10)
* 42335 (score=12)

Exclude patients with <10`000 reads -> repeat in next run

* 17811
* 18322
* 15504
* 18357
* 15224
* 18669
* 18311
* 18418
* 19138
* 13853
* 25478
* 17241
* 31396 (run 3)

Exclude controls :

recombination controls in 3rd run

Exclude read 46179\_S1 (patient was sequenced twice)

*(exclusions are done by filter(!grepl("46179\_S1\_|Hy|HD|AK170|41895|17420|26500|26586|34545|42335", patient\_ID)) on first run samples and*

*filter(!grepl("17811|18322|15504|18357|15224|18669|18311|18418|19138|13853|25478|17241|41895|31822", patient\_ID)) on second run samples, done in* combine\_n\_alleles\_pat\_characteristics.R*)*

filter(!grepl("4-59-|4-28-|mix|31396", patient\_ID)) on third run samples

**R scripts**

**Run-related parametes (Read numbers etc ):**

* reads\_per\_patient.R

analyzes reads per patient for all patients

* reads\_per\_family\_gene.R

analyzes reads per gene and family, also contains same analysis only with samples >10000 reads

* missing\_genes\_vs\_total\_reads.R

plots total reads per sample vs number of missing genes

**Reformat data, write output tables with patient characteristics and germline information**

* combine\_n\_alleles\_pat\_characteristics.R

combines “all\_results.txt” (number of alleles per gene and patient) with patient ethnicity and neut status, removes samples with <10`000 reads, removes wrongly included samples (ART, didn`t make it into top 105 etc)

-> writes table: “patients\_n\_alleles\_ethn\_neut\_subtype.txt” (contains patient, gene, n\_alleles, run, ethnicity, subtype, bnAb activity)

-> from this, check all samples with alleles > 4 using “alleles\_final” files and correct if necessary (file to view alleles with readcounts: multiple\_alleles\_raw.txt, corrections are recorded in multiple\_alleles\_corr.txt), save as patients\_n\_alleles\_corr.txt”

**Analyse**

**for now exclude**

* **4-28**
* **4-30-2**
* **4-30-4**
* **4-38-2**
* **4-39**
* **4-4**
* **4-61**
* **2-70**