V6RulesForOutputdir

Rules for outputDir

Mandatory folders:
`qc`
Optional folders:
`amplified-region-inference, asv, primer-identification, sequence-categorisation, taxonomy-summary`

In `qc`:
Can have maximum `5` files:

- `${run_id}.fastp.json`
- `${run_id}.merged.fastq.gz` **if PE** OR `${run_id}.fastp.fastq.gz` **if SE**
- `${run_id}_seqfu.tsv`
- `${run_id}_suffix_header_err.json`
- `${run_id}_multiqc_report.html`

Only required `qc` file will be `${run_id}_seqfu.tsv`, in case the first check (which is SeqFu) fails. If `${run_id}.merged.fastq.gz`/`${run_id}.fastp.fastq.gz` is empty, that means it will have failed at the fastp step because of a zero_read error

Next folder you can have after `qc` is `sequence-categorisation`
You should have `3` files minimum here:

- `${run_id}_${gene}.fasta` (depending on if the gene was SSU/LSU/ITS)
- `${run_id}.tblout.deoverlapped`
- `${run_id}_${gene}_rRNA_${domain}.${domain_id}.fa` (this will again depend on whether the domain was bacteria/archaea/eukarya). Example file name for this is: `ERR4334351_SSU_rRNA_bacteria.RF00177.fa`.
  You can have all **three domains** easily in one run so the amount of files can be higher than this, but this should be the minimum

After that is `amplified-region-inference`
You should have minimum `1` file:

- `${run_id}.tsv`

However, if that's the only file you have, then it means it didn't pass the amplified region inference thresholds (and therefore you **can't have ASV results with this run**)
If it does pass, you will have minimum `2` files:

- `${run_id}.tsv` same as before
- `${run_id}.${var_region}.txt` The var region will vary, example file could be `ERR4334351.16S.V3-V4.txt`

However, you can have a maximum of `2` `${run_id}.${var_region}.txt` files. So this means the maximum number of files should be `3` here

Next is `primer-identification`
You should have a minimum `1` file:

- `${run_id}.cutadapt.json`
  However, if you only have this one file, it has to be empty (it means no primers were found)

If not `1` file it has to be `3` files:

- `${run_id}.cutadapt.json`
- `${run_id}_primers.fasta`
- `${run_id}_primer_validation.tsv`

However, all three files can be empty (this should mean there were primers found but they failed primer validation). So if you have all three files they either have to all be empty, or all be not-empty

Then we have `asv`
The minimum is `5` files and `1` directory

- `${run_id}_dada2_stats.tsv`
- `${run_id}_DADA2-SILVA_asv_tax.tsv`
- `${run_id}_DADA2-PR2_asv_tax.tsv`
- `${run_id}_asv_seqs.fasta`
- `/${var_region}`
- `/${var_region}/${run_id}_${var_region}_asv_read_counts.tsv`

However, if you have more than one `${var_region}` in `amplified-region-inference`, then you **must** have `3` directories instead of `1`

- `${run_id}_dada2_stats.tsv`
- `${run_id}_DADA2-SILVA_asv_tax.tsv`
- `${run_id}_DADA2-PR2_asv_tax.tsv`
- `${run_id}_asv_seqs.fasta`
- `/${var_region1}` (+ `read_counts.tsv`)
- `/${var_region2}` (+ `read_counts.tsv`)
- `/concat` (+ `read_counts.tsv`)

The `/concat` dir will contain ASV read counts for both variable regions concatenated into one file. This file would have the same fileshape but insted of a `${var_region}` it will have the string `concat`, for example `${run_id}_concat_asv_read_counts.tsv`

Last directory is `taxonomy-summary`
This directory will contain maximum `7` directories, though the most common will be `6`.

- `SILVA-SSU`
- `PR2`
- `UNITE`
- `ITSoneDB`
- `DADA2-SILVA`
- `DADA2-PR2`

I haven't tested this, but if `taxonomy-summary` exists, it should contain minimum `2`directories depending on if there was SSU or ITS found:

- `SILVA-SSU`
- `PR2`
  OR
- `UNITE`
- `ITSoneDB`\

If you don't have ASV results because of a fail at `amplified_region_inference` then you **should not have** the directories `DADA2-SILVA` and `DADA2-PR2`. These **both should exist** if you do have ASV results, along with at least `SILVA-SSU+PR2`, so:

- `SILVA-SSU`
- `PR2`
- `DADA2-SILVA`
- `DADA2-PR2`

You can definitely have all `7` directories if you also have `SILVA-LSU`, though I expect it would be very rare:

- `SILVA-SSU`
- `SILVA-LSU`
- `PR2`
- `UNITE`
- `ITSoneDB`
- `DADA2-SILVA`
- `DADA2-PR2`

Then, these directories will have two different rules, the first rule applies to `SILVA-SSU/SILVA-LSU/PR2/UNITE/ITSoneDB`
These should have `4` files

- `${run_id}`.html
- `${run_id}_{db_label}.mseq`
- `${run_id}_{db_label}.tsv`
- `${run_id}_${db_label}.txt`

The other rule applies to `DADA2-SILVA`/`DADA2-PR2` and they should have minimum `3` files:

- `${run_id}_${var_region}_{db_label}_asv_krona_counts.txt`
- `${run_id}_${var_region}.html`
- `${run_id}_${db_label}.mseq`

However, if you have more than one `${var_region}`, you will have `7` files

- `${run_id}_${var_region1}_{db_label}_asv_krona_counts.txt`
- `${run_id}_${var_region1}.html`
- `${run_id}_${var_region2}_{db_label}_asv_krona_counts.txt`
- `${run_id}_${var_region2}.html`
- `${run_id}_concat_{db_label}_asv_krona_counts.txt`
- `${run_id}_concat.html`
- `${run_id}_${db_label}.mseq`