Dear all,

I mostly finished implementing in R Luca’s pipeline, along with replicating most of the barcode visualisations, that you can find on the GitHub repository, or even simpler, on the Quarto Book URL project: <https://verapancaldilab.github.io/DNABarcode-DrugFingerprint/> (note that you also download a docx version of the website). The main sections include:

## Metadata cleaning and mismatch

Most importantly, @luca and @david, can you confirm that I have the most comprehensive barcode counts profiles (namely all replicates AND all barcode IDs)?

* With R, I’m not suffering from the intrinsic memory-limitations of Excel.
* I emphasised again on this point, as the provided original `exp200921\_whole exp.TABULAR` file included much more barcode IDs with respect to originally provided CSV or Excel files.
* Besides, there’s a strong discrepancy in terms of surveyed barcode IDs from one batch to another, going from only 93,132 unique barcode IDs in batch exp07022, up to 1,067,031 unique barcode IDs in the 200921 TABULAR file.

To conclude with this section (and other spotted metadata inconsistencies), it would be great to schedule a meeting only with the 3 of us, Luca, David and myself next week, to ensure that I have the most complete barcode counts and no errors on the provided metadata per batch (Compound, Concentration, Number of replicates, …) next week (free, except on Tuesday) -> to best interpret biologically main outputs, generate the most insightful representations, and customise at best the pipeline, this data management and cleaning is critical!

Waiting for the meeting, you can check that the general metadata information provided in attached file ` barcode\_metadata\_overview\_2025-04-28.xlsx` is correct.

## Methods

* Reproduce Luca’s pipeline (background noise filtering, Fold-change of 3 with respect to Controls, binarization)
* Save resulting normalised and filtered barcode counts as standard SummarizedExperiment objects …
* … at 2 stages of the pipeline, and 3 levels of granularity.
  + After background filtering, merging of shared barcode IDs across batches and normalisation (stored under folder .\results\compounds,as se\_normalised\_\* files)
  + After binarization (assign a 1 to a given replicate barcode ID if and only if its FC is above 3 with respect to the average barcode count of Control replicates.(stored under folder .\results\compounds, as se\_discretised\_\* files)
  + At the replicate level, compound by batch level (averaging over replicates for normalised, and considering an all operator for binarized values), and compound level.
* Once done, compute the correlation matrix between replicates (or compounds), AND the correlation matrix between barcode IDs.
  + I came up with the same threshold of 10 (only by visualisation) for identifying well-separated clusters of cell barcode IDs.

### Results

* Generate ComplexHeatmap for the different levels of resolution and pipeline stages identified in point 2)
* Compute an igraph object of the drug pairs, highlighting a cluster of EGFR-driven compounds.

## Perspectives

There are definitely other elements to be tested, discussed with Marie-Pier on Thursday, such as:

* Quality controls (histograms of barcode IDs presence, PCAs, …), reported under bartools vignette: <https://danevass.github.io/bartools/articles/bartools_quickstart.html> (specifically, density plots of control counts, as we would not expect large count values for them, not being subjected to positive fitness selection by drugs)
* Improve Heatmap objects, testing different distance metrics (overlap, jaccard distance), clustering algorithms, unsupervised reordering <https://jokergoo.github.io/ComplexHeatmap-reference/book/a-single-heatmap.html#reorder-dendrograms>, and supervised splits.
* Try more stringent selection of barcode IDs, as already generated Heatmaps suggest higher discrimination performance by keeping the most informative barcode IDs only.
* Run targeted analytical design experiences for dose response and time course experiences, for which we have a unique batch, clearly defined replicates and concentrations….

… but most importantly, I want to ensure that I have the most comprehensive barcode datasets and corrected metadata, best achieved through a three-actor meeting series with Luca and David.

In the provided ` ` file, it turned out that all drug samples related to the Drug Response batch are present, while some experiences in the more standard 200921 experience have been discarded. Report to replicates\_discarded\_200921.txt and `replicates\_preserved\_200921.txt attached for details (in total, 4 replicates times 5 compounds (Palboc, Olapar, LGK9743, Chloro, Bafilo)=20 replicates overall).

* If I have well understood your mail, should I completely discard out all replicates paired with `T25` pass in `exp281022` (4 replicates times 4 compounds= 16 replicates overall) + `p42` and `p43`!? Subsequently, what does this `T25` refer to? At least from the global Correlation Drug Heatmap matrix.
* Same remark for the `exp130921 in vivo` experiment. Not included so far to compute drug similarities in the resulting network.
* Finally, other spotted inconsistencies between provided general metadata `Table of compounds.xlsx` file and actual barcode replicate IDs are listed under <https://github.com/VeraPancaldiLab/DNABarcode-DrugFingerprint/issues>/2 .

## Metadata inconsistencies

* I do not know exactly what to do with paired drug effect Osimertinb+sorafenib: 0,1uM+5uM, as the only experience trying to evaluate the synergistic effect of two drugs.
* I spotted some replicates’ names associated with the `PEM` prefix in ` exp281022\_time course` experiment. Can you confirm that it stands for `Compound: Pemetrexed`? (Indeed, for the other experiences, it’s usually prefix `Pemetr` which has been used!!).
* Finally, other inconsistencies are reported under <https://github.com/VeraPancaldiLab/DNABarcode-DrugFingerprint/issues/1> .