Dear all,

This email to summarise the exchanges with the stakeholders of the project, along with and next scheduled tasks:

## Useful links

* **GH code** : <https://github.com/VeraPancaldiLab/DNABarcode-DrugFingerprint>
* Web Quarto reporting: <https://verapancaldilab.github.io/DNABarcode-DrugFingerprint/>
* Personal One Drive repository**, read-only** and personalised link: <https://1drv.ms/f/c/5147dea5b81ac1f3/EvLw_kCHo1pOlRkh0hB7l2sBH4KPc8MyP6DMbltB0oDMWg?e=qmxGJC> -> only method so far to reproduce locally the results, as the original barcode files are too large for GH (+ maybe not the best idea to store them there), waiting for migration to IFB Cluster (if relevant)

## Metadata cleaning and mismatch

* @luca and @david, provide on the Google Drive folder the most comprehensive barcode counts profiles (namely all replicates AND all barcode IDs), at the finer granularity level (**no need for prior aggregating runs by Galaxy** -> this can be done afterwards with R easily -> what matters here, beyond having all the individual barcode IDs, is the **Run Date information, as the variable used for potential batch effect correction)**
  + Original .TABULAR format is perfectly fine for me (actually, better than Excel, as this software tends to clip barcode IDs when the file is too large, or changing colnames) + when using CSV, check you’ve configured Excel to **the standard international format, namely using “,” comma as delimiter, and not “;”:** <https://support.microsoft.com/en-us/office/import-or-export-text-txt-or-csv-files-5250ac4c-663c-47ce-937b-339e391393ba>
  + Add the **Time-Zero controls barcode controls**
* Check general phenotype metadata information:
  + @bastien, @luca and @david: **check that my Tab ‘Experimental Design’**, notably reported Concentrations (add a ‘Concentrations Unit’ colname, specifying whether it should be read as mole or mg/kgs, or any other unit), Experiment date and Run date (by the way, @luca, can you confirm that Run Date corresponds to my “technical batch effect covariate”, when the sequencing was performed, and “Experiment date” when the culture of cell lines had been started? **) matches the metadata table that has just been forwarded by Luca**
    - Can you confirm that the only modified things are the Pathways and MoAs annotations + badly reported concentrations in yellow? 😊.
    - @bastien: Check how the Luca’s annotated MoAs and Pathways recover well automatic annotations from Drugbank and DGIdb, suggested drug databases by @vera
    - @bastien: add, whenever feasible, colname ‘DrugBankID’, for downstream Drug Network similarity analysis.
  + Discarded experiences (as discussed, we keep everything at first, except if the individual replicates were proven of bad quality 😊, for instance due to limited drug concentration) -> it’s straightforward to filter out and subset relevant experiments in subsequent downstream analyses.

## Methods

### Quality control (for @bastien):

Basically follow the bartools tutorial: <https://danevass.github.io/bartools/articles/bartools_quickstart.html> -> Objective: justify by motivated data-driven processes the thresholds proposed by Luca.

## Results

General strategy: first include everything for the generation of Heatmaps, then refine subsequently if we found afterwards that some replicates only brought noise/are irrelevant.

### Drug network

1. @bastien, export generated graph object to a format readable by Cytoscape + find references for performing graph clustering (such as Louvain)
2. @vera, find back code snippets for comparing inferred barcode similarity scores (with database DrugSimDB) + additional question: **how do you aggregate barcode counts information from the replicates level to the compound level** (so, across batches, concentrations, durations, and so on?) to compute the pairwise Pearson correlation between two compounds barcode profiles?
3. @marie-pier, provide bibliographic references and/or proven methods that compute pairwise drug similarity scores based on their weighted proximity in the network (for instance, using random Markov Chains) -> indeed, alone, the p-value score is not that relevant for filtering out spurious edges, acknowledging the larger number of observations / sample size quantified by the number of barcode IDs (also named the fallacy of large sample size p-values, see <https://pubmed.ncbi.nlm.nih.gov/22862286/> )

**Overarching objective:** try to find back Luca’s intuition/Vera’s original finding that unknown compound X13271 clusters well with compounds exhibiting proteasome inhibitors, Bortezomib et MG132 as MoA (@vera and @luca, feel free to dissert if I’m wrong here)

### Barcode IDs clustering

* In file ‘./figures/cell\_lines/heatmap\_cell\_lines\_by\_cell\_lines\_uncompressed\_2025-04-28.pdf’ (to be opened with AdobeAcrobat, being too large for web browsers), we notice that some barcode IDs cluster well with each other -> intrinsic heterogeneity of the PC9 cell line and/or induced mutations (unlikely) or epigenetic transformations by drugs? **@vera, could you textually elaborate further on this topic, with your idea of UMAP projection paired with correlations of correlations, being uncertain of your idea**? (alternatively, provide code snippets used for implementing this idea 😊)
* @bastien, find references of bi-clustering approaches (simultaneously cluster barcode IDs and samples)

## Perspectives

There are other elements to be tested, such as:

* Try more stringent selection of barcode IDs, as 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
* Run targeted analytical design experiences for Control + Control Time Zero experiences, including P42 and P43 drifts, to evaluate the underlying intrinsic heterogeneity of the P9 cell line (and generally barcode drifts over time)

As mentioned, I will be rather busy the next two weeks, being selected as reviewer of 5!! Papers for ISMB/ECCB conference + preparing my modified talk to GT Bioss, but feel free to elaborate on the mentioned items and/or ask for clarifications if unclear,