# May logs

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)

# June logs

## Discussion Vera and Luca 18/06/2025

Hi everyone,

I hope you’re going well.

### Meeting tomorrow:

First, a little reminder of our meeting tomorrow. Main objectives:

* Finalise data management and cleaning, share the tasks related to it.
* Determine minimal analyses and figures required for the paper + share the tasks among us.
* Identify potentially additional analytical tasks worth including in Supplementary material (for instance, comparison of different approaches for selecting the most contributing barcode IDs).

### Tasks distribution, beyond the meeting (this is a suggestion only 😊)

* For @David.alexandre@univ-rouen.fr, double check the exp281022gamme.tsv (aka the **time-course experiment**). On the most recent version of the dataset, uploaded on Google Drive,
  + Four T-75 Controls replicates were provided, however, samples 22 and 24 (CTRL1 and CTRL3) only **contain null values**. More generally, what’s the role of these Controls -> for now, I’m not considering them in the analysis:
    - CTRL1-T75\_000u\_exp281022\_run110723\_22,
    - CTRL2-T75\_000u\_exp281022\_run110723\_23
    - CTRL3-T75\_000u\_exp281022\_run110723\_24
    - CTRL4-T75\_000u\_exp281022\_run110723\_25
  + On the other hand, the **true Controls, after 9 days of incubation, were missing**, while previously available in older versions. What was their fate?
    - **9d\_Contro1\_000u\_exp281022\_run110723\_01**
    - **9d\_Contro2\_000u\_exp281022\_run110723\_02**
    - **9d\_Contro3\_000u\_exp281022\_run110723\_03**
    - **9d\_Contro4\_000u\_exp281022\_run110723\_04**.
* For @ [luca.grumolato@univ-rouen.fr](mailto:luca.grumolato@univ-rouen.fr), can you double check the **mapping** in the Attached table (file “drugs\_generic\_to\_drugbankID\_2024-06-24.xlsx”) between:
  + Original names you provided me with (“**Compound\_Luca**”)
  + **Generic\_Name** (The most common generic DrugBank used for naming the therapeutic compound)
  + This task is required for the final stage of the barcoding pipeline, proposed by @ [vera.pancaldi@inserm.fr](mailto:vera.pancaldi@inserm.fr), which relies on <https://github.com/VafaeeLab/drugSimDB>: we need to map each compound to its cognate DrugBankID (for instance, Azacitidine: DB00928).
  + In **red and bold, I highlighted the drugs requiring more specifically your investigation**, with an additional comments Bastien describing my research process.
  + If something is not correct, please detail in “comments Luca” column.
  + (**Of note, some of the drug compounds (once excluded with unknown MoAs) were only available on MedChemExpress and not on DrugBankID!!)**
  + Once this mapping validated, I will be able to switch to next step to validate **MoA** and **Pathway** in a data-driven fashion 😊.
* For @vera.pancaldi@inserm.fr:
  + Waiting for recovering my previous academic email address, may you download the most **up-to-date XML version of DrugBank**: <https://go.drugbank.com/public_users/sign_up> …
  + … and **upload it on Google Driv**e, or, if too large, on my own OneDrive account, as: ‘**DNABarcode-DrugFingerprint\data\drugbank\drugbank\_{version\_number}.xml’**: <https://1drv.ms/f/c/5147dea5b81ac1f3/En8obUg08kpInFYDawkvktsBjf8vW34lgbtEsB6WJfVVsA?e=UgzvlP>
  + Among all the Cytoscape networks and views generated (42 in total), detail exactly which one I should use? (I guess **Gcond25nov24\_only05neighx13271(1**), linking unknown MoA compound **X13271\_008u** and **\_009u** to other drug compounds)
  + Detail a bit more how you’ve generated Fig.6 of the Nina Verstraete poster, comparing similarity identified by barcode profiles with existing similarity scores generated by DrugSimDB.
    - Specifically, how could we generate those scores for drugs having only a MedChemExpress input without DrugBank ID, meaning we’ll likely have to recompute Chemical structure, Targets, Gene Ontology, Cellular component, Molecular Function and Biological process of induced pathways scores!!
* Drug bank IDs :
  + - <https://github.com/VafaeeLab/drugSimDB>? With the 6 scores Chemical structure, Targets, Gene Ontology, Cellular component, Molecular Function and Biological process of induced pathways
    - Based on the files located here, <https://figshare.com/articles/dataset/DrugSimDB/11948904>, I need 1-1 mapping between DrugBankID: DB12218, and its generic name: Capivasertib.
    - Will use this R package: https://github.com/girke-lab/drugbankR

to automate the mapping between DrugBankID, Generic Name, MoA and Pathwy

* + - Need the original XML DrugBank file, but do not have anymore academic mail and affiliation!!: <https://go.drugbank.com/public_users/sign_up>, connect here.
    - HTTP status was '403 Forbidden, without proper login
    - When “distance” colname is empty, indicator that semi-automated approaches did not work. On the other hand, if drugbankID is **null (not empty!!),** means that we could have mapped generic name, but it’s still not mapped to official DrugBankID.
    - Also add MedChemExpress ID? In that case, can’t map against DrugSimDB.
    - Which information on drug compounds to place in Tables? Pathway and MoA based on DrugBankID? MedChemExpress?
* Dvc remote to be defined!!

exp200921 has 4 controls, while and exp200921\_dose\_response does not have any, but Time-zero-controls. Which reference should I take?

• Sorafenib, only 3 replicates, for experience 5 uM, date 300821, and run date 290921.

• Run versus Experimental data, is it truly 9 days? Same remark for time-zero controls

• Run Date, from 120122\_180122 to 180122 only!! In `Run date`, when there was a range instead of a fixed Date, I kept only the end of the interval.

• 28/10/22 and Control experiences:

* time\_course, aka gamme, run.date: 110723
* 4 “normal” controls
* 4 T-75, role? Time-zero? I remove them in the end, exhibiting for 2 of them entirely null values!!
* Standard experience, aka t25 including cetuximab and drift control. 10 controls in total:
* P42 and P43 drift control
* 4 Controls for T-75. Role? Apparently for cetuxmibal control!!
* 4 standard controls, previously mentioned as cDNA/