Dear Luca and Vera,

I started working on the barcoding project again this Thursday (later than I would have wanted to, but my focus has been unwillingly captured by a last-minute GWAS analysis, actually more “gene candidate” analysis, critically lacking statistical power -> but you know, proving that an analysis is vowed to failure or irrelevant is often way more difficult than the other way round).

Nonetheless, I have some comments and remarks before pursuing the analysis:

* Except mistaken, in the zip archive link sent by Luca, the latest experiences (time course, exp070222 and exp281022) were missing -> consequently, I fetched the datasets from Google Drive, shared by Vera.
* *xlsx* format for storing raw counts might not be the most efficient in terms of data integrity and memory. *Csv* (the English one, not the French, with a comma delimiter and not a semi-colon) is better, + I may also consider switching to *Parquet* format and/or *DuckDB* (efficient fetching, true database structure, universal programming access).
* Till now, all my code to reproduce the scripts is stored under my own GitHub repository, with private visibility: <https://github.com/bastienchassagnol/DNABarcode-DrugFingerprint> (only Vera and I can review the code so far)-> **however, if feasible, I would prefer that Vera creates a repo on her own GH organisation, and add myself with “maintainer” or “write” rights, being confident in Vera’s integrity**: <https://github.com/VeraPancaldiLab> . GH username: ‘bastienchassagnol’. Email: ‘bastien\_chassagnol@laposte.net’
* For some analyses, my local computer is running out of resource memory and computation. Solutions? Use a remote cluster. **However, preferentially, I would prefer using a cluster/computational resource managed by either Luca or Vera, on which you could add me.** Alternatively, we could leverage *AbIms* French cluster, or *ComputeCanada* cluster, but I would not be able to guarantee as much as wished data integrity.
* Side note for Luca: Highlighting in red the experiences you wish to keep is unfortunately not recognised by programming languages. On the other hand, adding a logical colname, with true or false statements, is much easier to parse (**no worries, I was still able to do the mapping** 😊).

Once I will have ended these tedious but mandatory and paramount *data cleansing* analyses, I will tell you, Luca, in order to double check and validate both barcode counts content (comparing dimensions should be enough, namely number of barcode ids \* number of samples), and sample metadata labelling.

If you have questions, or if I wasn’t clear enough, please feel free to reach out,

Best, and have a nice weekend,

Bastien,