Inferring the MOA of new drugs through the analysis of heterogeneous response to treatment of different subpopulations of cancer cells

Bastien CHASSAGNOL and Vera PANCALDI

2025-02-19

Contents

# 1. Analyses

## 1.1 Data management

* [Original Google Drive repository](https://drive.google.com/drive/folders/1pMSX4M4kHcDGxcsCzsMYHHM6VRvBHHUO?usp=drive_link).
* **File structure organisation**:
  + data stores all datasets, and is further split between data-raw (the original and untouched datasets), and data-derived (stored datasets compliant with tidy standards, with proper mapping between phenotype aka experimental metadata and expression matrices).

#### 1.1.0.1 Additional comments

* exp130921\_in\_vivo: cells injected in mice.
* exp281022\_time course: time course osim and pemetrexed, n=3 for osim 3 months as we lost replicate 4.
* exp281022: control, protac, cetuximab and gefitinib were tested in T25 flasks. \*\*P42 and P43 correspond to cells grown without treatment for 42 or 43 passages, evaluating barcode drift.

### 1.1.1 Barcode expression matrix

ToDO, but already spotted 3 painful elements:

* Format used is **French csv**, instead of international CSV. As the name indicates, CSV means for ‘comma-separated value’, so prefer this format (instead of ‘;’ separation), by changing your [Excel settings](https://support.microsoft.com/en-us/office/import-or-export-text-txt-or-csv-files-5250ac4c-663c-47ce-937b-339e391393ba).
* First column name, corresponding to barcode identification, is either left ‘empty’, or associated with key name -> whatever the final convention, **be consistent**!!
* Some barcode expression matrices, such as exp220322.csv and exp070222.csv showcase a final line paired with index value. What’s this? Should I expect strange things like that in other expression profiles, and add a custom filter checking the original row key is indeed a ‘ATCG’ sequence tag? Even stranger: full empty count values for some indexes, such as in exp200921.csv, where last line is associated with GATCA key only (against a tag of around 20 nucleotides normally).

#### 1.1.1.1 Processing per batch

**Code colour**: Luca’s protocol, Vera’s code and Bastien’s comments

1. **Noise filtering**:

* Eliminate barcodes for which the combined counts of the 4 controls per barcode are below 5, and below 4 for experiments 010821 and 040821
* No dichotomy between experiences in Vera, always remove barcodes below 4.
* Density plots to evaluate the relevance of this threshold, HTSFilter, comparison with existing filtering approaches, … But removing noise is indeed a great point!!, especially with the original high number of barcodes. However, why only performing this operation on control cases?

1. **Normalisation**

* Normalize barcodes so that the total number of counts per sample is 100 000
* Same
* Close to two existing normalisation methods: Counts Per Million (CPM) which additionally scales raw counts by total library size and multiplies by 1,000,000 and Total Count Scaling (TCS): Scales raw counts by the total number of reads (or mapped reads) in each sample, then multiplies by a fixed number (e.g., 100,000). Would compare other normalisation approaches + ignore biological or technical biases + generate MA plots for verifying the mean-variance correction trend + not suitable for DEG analyses.

1. **Derivation of basal barcoding expression**

* Calculate mean of the 4 controls
* Same, use rowMeans
* No discussion of potential batch effect correction that would possibly allow pairing of all control samples, or all treatments at the same concentration -> indeed, would best require balanced designs.

1. **Differential Represented Barcode Analysis**(DRB)

* Calculate the fold change (FC) treatment vs control for each drug replicate, set threshold to 3 for keeping barcodes. Applies discretisation: set value to 1 if Fold change is above 3, 0 otherwise.
* Computed by Vera in function create\_fil, stored in variable efcfil, but finally not used in the subsequent analysis.
* As Vera emphasised it out, why keeping only positive values? Negative are also interesting, otherwise, we will bias towards drugs having a positive fitness. Why not computing the fold change at the drug level, averaging all replicate values? Why not pairing p-values and fold-change (considering indeed really small sample sizes)? All these operations can be done in one step using lm, and a model as such, Expr∼0+Gene+Batch+Drug. Indeed, note that your model implicitly assumes Gaussian-distributed expression profiles, and a contr.treatment contrast, in other terms, one fit per each combination of variables: interpretable but definitely the least powerful[[1]](#footnote-23)

1. **Derivation of drug fingerprints**:

* Calculate the sum for the four replicates per drug
* Same, yet, adds an extra-filtering step by removing probes with null expression -> by applying the FC selection stage at step 4, this step should not have been required. See also next bullet point why vera will not return integer values.
* Wondered first why using the sum instead of the mean. Makes sense in Luca’s protocol with discretisation, not in Vera’s protocol without effectively enforcing discretisation. Replace threshold of 4 by number of replicates + easier to simply use a logical AND. In any case, we need sensitivity analyses to evaluate the impact of such stringent thresholds, 3 seems really hard.
* Final output: both barcode expressions at the replicate level, stored in efc (… replicates), and at the perturbagen/drug level, stored in comb (121 treatment configurations, excluding controls?), with additional Batch\_ID (14 experiences), having removed background noise and non-DRBs.

#### 1.1.1.2 Final Concatenation

1. **Global perturbagen profile derivation**:

* Merge all barcode expression profiles keeping only the common Differentially expressed ones (or in 2/3 of the samples?).
* Simply merge common barcodes, yet, without the DRB selection as finally not performed.
* Have to check extension of unbalanced Wassertein distance, or relatives, for distinct input and output dimensions, otherwise, concatenating everything seems legit.

1. **Drug correlation**:
   1. Select barcodes displaying a sum of at least 1 returns 4106 barcodes. Same in vera’s codeCondition already checked by the pre-processing itself, in the selection of DRBs that impose in practice a total expression of at least 6.
   2. Selected barcodes displaying a sum of at least 10 returns 657 barcodesThis one makes more sense actually. Yet, we have to try other correlation methods on the continuous space, or consider other metrics if working on the discrete space.
   3. Actual computation of the correlation matrix using stats::cor.
   4. Plot weighted undirected graphs with igraph::graph\_from\_adjacency\_matrix
2. **Barcode correlation**:

* To select barcodes highly correlated within each other, must correlate, with CC>0.8 with at least 4 others (or 5, do you remove the barcode itself?).
* Not done.

1. Really wants to try the Helmert contrast to evaluate drug dose response, or time-course studies. [↑](#footnote-ref-23)