SARTORIUS

Simplifying Progress

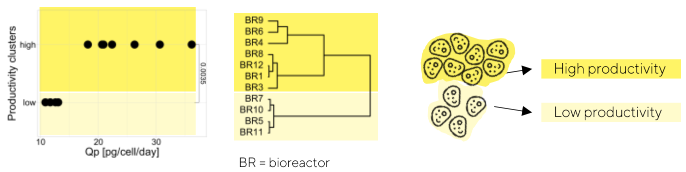
Combining data analysis strategies to identify gene targets for the optimisation of production cell lines

Merle Rattay[1*], Shanti Pijeaud[1], Athanasios Antonakoudis[2], Pär Jonsson[3], Olivier Cloarec[4], and Anne Richelle[5]

[1] Sartorius Stedim Cellca GmbH, Germany
[2] Sartorius Stedim UK Ltd., United Kingdom
[3] Sartorius Stedim Data Analytics AB, Sweden
[4] Sartorius Stedim France S.A.S, France
[5] Sartorius Stedim Belgium S.A., Belgium
[*] Corresponding author: merle.rattay@sartorius.com

Analyses of omics data can help to understand molecular mechanisms of observed phenotypes and generate hypotheses to translate this knowledge into real biological systems. In cell line development, such analyses can be used to discover gene targets for genetic engineering. By targeted knockouts or regulation of gene expression of discovered genes, cell lines can be optimised and the production of biopharmaceuticals improved. However, evaluation of omics-derived gene targets is limited by the time- and work-consuming processes of gene editing and cell culture experiments in the laboratory. To generate value through omics analyses, a well-informed choice of a limited number of most promising targets is crucial. Here, we combined classical differential expression analysis with approaches from other fields of data analytics to improve our selection of gene targets.

Using omics analyses to leverage clonal variation



Qp = specific productivity

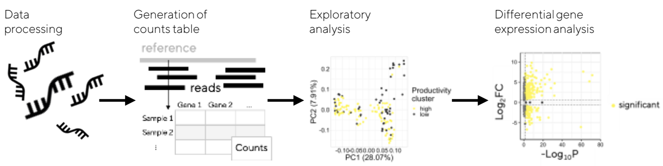
BR = bioreactor
BR1, 6 and 12 are triplicates of the same clone

Genetically similar cell line clones can display a broad range of bioprocess performance, e.g. in their productivity. This biological variation can be used to identify genes associated with a desired phenotype. 11 bioreactors of 9 clones were clustered based on specific productivity (Qp) during a 14-day bioprocess in Ambr® 250 mini bioreactors . The mean Qp was significantly different between groups defined through hierarchical clustering (t-test, p-value < 0.05). Antibody titers were measured with a Protein A-assay (Octet® R8). RNA samples were taken daily and sequenced as 150 bp PE reads (NovaSeq 6000 S4 PE150 XP). Metabolites were measured with nuclear magnetic resonance (Bruker 600 MHz AVANCE III).

Differential gene expression analysis (DGEA)

Comparison of high and low productivity clones

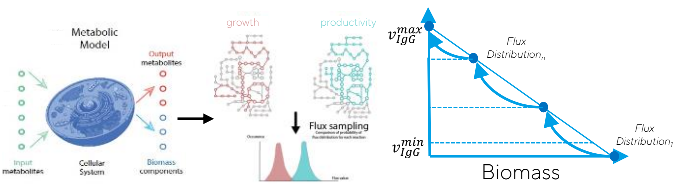
Reads were adapter- and quality-trimmed[1] and pseudo-aligned to a Chinese hamster genome assembly[2,3]. Counts were generated and normalised by variance-stabilising transformation prior to DGEA with DESeq2[4] to compare the groups defined above (design: ~productivity).



Metabolic modeling (Metmod)

Identification of reactions' flux changes related to productivity

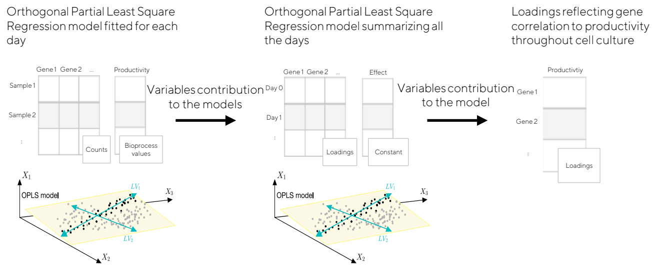
A metabolic model was constructed based on an existing genome-scale metabolic network[8], cell-specific metabolite rate data and gene expression counts. Reactions significantly influencing the acquisition of a high growth or high productivity phenotype were identified using two different FBA-based methodologies: comparison of flux sampling space with respect to phenotype objective[9] or flux scanning based on enforced objective flux[10-12].



Orthogonal least partial squares analysis (OPLS)

Correlation of gene expression with productivity

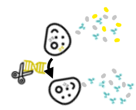
An OPLS model[5] was fitted in SIMCA® 17[6] for the gene expression and Qp for each timepoint of the dataset. The trend across timepoints was summarised by fitting OPLS-Effect Projections[7] to the scaled loadings (p(corr)) and an effect constant to yield final loadings. The models were validated by permutation tests (n=100).



Challenges of data and method integration

- Identifier mapping, depending on data providers and analysis requirements
- Comparison of qualitative and quantitative results
- Decision on (arbitrary) 'significance' cutoffs
- Combining different outcomes into an easily interpretable metric for laboratory colleagues and decision-making

Planned laboratory evaluation of gene targets

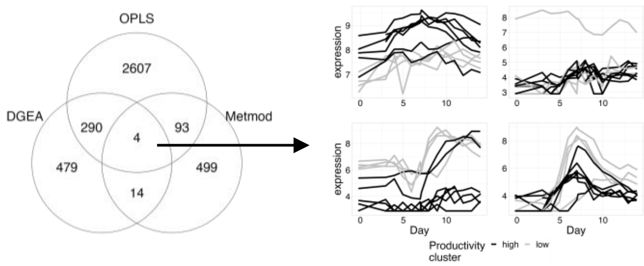


Identified genes will be evaluated through genetic engineering in the laboratory. Targeted gene knockouts will be the first step as these provide a clear readout and add information about the essentiality of the investigated gene, i.e., whether a knockout is lethal or has a negative impact on growth.

Comparison of data analysis approaches

The overlap of genes identified by each method was investigated. Genes of the DGEA were filtered for significant genes (log2FC ≥ log2(1.5) and p-adj < 0.05). OPLS results were filtered for genes with a loading (pcorr) above the 90th or below the 10th percentile. The metabolic modelling approach only reports significant pathways and genes involved therein.

787 genes were differently expressed in the DGEA, 2994 genes were present in the filtered OPLS results and 610 were reported by Metmod. Of these, 401 were identified by at least two methods. 4 genes were detected in all three. One of the four genes was overexpressed in high productivity clones (top left) while two were enriched in the low productivity clones (bottom). The fourth gene seemed to be highly expressed in one low productivity clone but had comparable expression in all other clones.



Conclusion and outlook

- > 19 000 expressed genes were analysed with three different methods to identify genes associated with productivity
  - 401 genes identified by at least two methods
  - 4 genes identified by all methods

➢ Gene targets supported by more than one method should be prioritized

Ideas for further reduction of the gene target list include

- Grouping according to biological themes (KEGG pathways and/or GO terms)
- Identifying key genes and/or transcription factors
- Modification of (arbitrary) 'significance' cutoffs
- Calculation of a single score from all three analyses for easier ranking of genes

1. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol. 2016 May;34(5):525–7.
2. CHO Genome Community. Genome assembly CriGri-PICRH-1.0 [Internet]. Available from: https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_003668045.3/
3. Trygg J, Wold S. Orthogonal projections to latent structures (O-PLS). J Chemom. 2002;16(3):119–28.
4. Sartorius. SIMCA® [Internet]. Available from: https://www.sartorius.com/en/products/process-analytical-technology/data-analytics-software/mvda-software/simca
5. Jonsson P, Wuolikainen A, Thysell E, Chorell E, Stattin P, Wikström P, et al. Constrained randomization and multivariate effect projections improve information extraction and biomarker pattern discovery in metabolomics studies involving dependent samples. Metabolomics. 2015;11(6):1667–78.
6. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014 Dec 5;15(12):550.
7. Hefzi H, Ang KS, Hanscho M, Bordbar A, Ruckerbauer D, Lakshmanan M, et al. A Consensus Genome-scale Reconstruction of Chinese Hamster Ovary Cell Metabolism. Cell Syst. 2016 Nov;3(5):434-443.e8.
8. Choi HS, Lee SY, Kim TY, Woo HM. In Silico Identification of Gene Amplification Targets for Improvement of Lycopene Production. Appl Environ Microbiol. 2010 May;76(10):3097–105.
9. Lewis NE, Hixson KK, Conrad TM, Lerman JA, Charusanti P, Polpitiya AD, et al. Omic data from evolved E. coli are consistent with computed optimal growth from genome-scale models. Mol Syst Biol. 2010 Jul 27;6:390.
10. Segrè D, Vitkup D, Church GM. Analysis of optimality in natural and perturbed metabolic networks. Proc Natl Acad Sci. 2002 Nov 12;99(23):15112–7.
11. Kaufman DE, Smith RL. Direction Choice for Accelerated Convergence in Hit-and-Run Sampling. Oper Res. 1998 Feb;46(1):84–95.
12. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014 Aug 1;30(15):2114–20.