Bioinformatics analysis of NSCLC multi-omics data

Janne Lehtiö (janne.lehtio@ki.se)
Department of Oncology and Pathology, Karolinska Institutet, Science for Life Laboratory, Solna, SE-17165, Sweden

Taner Arslan
Department of Oncology and Pathology, Karolinska Institutet, Science for Life Laboratory, Solna, SE-17165, Sweden

Ioannis Siavelis
Department of Oncology and Pathology, Karolinska Institutet, Science for Life Laboratory, Solna, SE-17165, Sweden

Yanbo Pan
Department of Oncology and Pathology, Karolinska Institutet, Science for Life Laboratory, Solna, SE-17165, Sweden

Fabio Socciarelli
Department of Oncology and Pathology, Karolinska Institutet, Science for Life Laboratory, Solna, SE-17165, Sweden

Olena Berkovska
Department of Oncology and Pathology, Karolinska Institutet, Science for Life Laboratory, Solna, SE-17165, Sweden https://orcid.org/0000-0002-8811-0591

Husen M. Umer
Department of Oncology and Pathology, Karolinska Institutet, Science for Life Laboratory, Solna, SE-17165, Sweden

Georgios Mermelekas
Department of Oncology and Pathology, Karolinska Institutet, Science for Life Laboratory, Solna, SE-17165, Sweden

Mohammad Pirmoradian
Department of Oncology and Pathology, Karolinska Institutet, Science for Life Laboratory, Solna, SE-17165, Sweden

Mats Jönsson
Division of Oncology, Department of Clinical Sciences, Lund and CREATE Health Strategic Center for Translational Cancer Research, Lund University, Lund, Sweden

Hans Brunnström
Department of Pathology, Laboratory Medicine Region Skåne, Lund, Sweden, Division of Pathology, Department of Clinical Sciences, Lund, Lund University, Lund, Sweden

Odd Terje Brustugun

Section of Oncology, Drammen Hospital, Vestre Viken Health Trust, Drammen, Norway, Department of Cancer Genetics, Institute for Cancer Research, Oslo University Hospital, Oslo, Norway

Krishna Pinganksha Purohit

University of Edinburgh Centre for Inflammation Research, Institute for Regeneration and Repair, Queen's Medical Research Institute, Edinburgh bioQuarter, 47 Little France Crescent, Edinburgh EH16 4TJ, UK, MRC Centre for Regenerative Medicine, Institute for Regeneration and Repair, University of Edinburgh, Edinburgh bioQuarter, 5 Little France Drive, Edinburgh EH16 4UU, UK

Richard Cunningham

University of Edinburgh Centre for Inflammation Research, Institute for Regeneration and Repair, Queen's Medical Research Institute, Edinburgh bioQuarter, 47 Little France Crescent, Edinburgh EH16 4TJ, UK, MRC Centre for Regenerative Medicine, Institute for Regeneration and Repair, University of Edinburgh, Edinburgh bioQuarter, 5 Little France Drive, Edinburgh EH16 4UU, UK

Hassan Foroughi Asl

Genomic Medicine Center, Karolinska University Hospital, Stockholm, Sweden. Clinical Genomics Facility, Department of Microbiology, Tumour and Cell Biology, Karolinska Institutet, Stockholm, Sweden.

Sofi Isaksson

Division of Oncology, Department of Clinical Sciences, Lund and CREATE Health Strategic Center for Translational Cancer Research, Lund University, Lund, Sweden

Elsa Arbajian

Division of Oncology, Department of Clinical Sciences, Lund and CREATE Health Strategic Center for Translational Cancer Research, Lund University, Lund, Sweden

Mattias Aine

Division of Oncology, Department of Clinical Sciences, Lund and CREATE Health Strategic Center for Translational Cancer Research, Lund University, Lund, Sweden

Anna Karlsson

Division of Oncology, Department of Clinical Sciences, Lund and CREATE Health Strategic Center for Translational Cancer Research, Lund University, Lund, Sweden

Marija Kotevska

Division of Oncology, Department of Clinical Sciences, Lund and CREATE Health Strategic Center for Translational Cancer Research, Lund University, Lund, Sweden, Department of Respiratory Medicine and Allergology, Skåne University Hospital, Lund, Sweden

Carsten Gram Hanson

University of Edinburgh Centre for Inflammation Research, Institute for Regeneration and Repair, Queen's Medical Research Institute, Edinburgh bioQuarter, 47 Little France Crescent, Edinburgh EH16 4TJ, UK, MRC Centre for Regenerative Medicine, Institute for Regeneration and Repair, University of Edinburgh, Edinburgh bioQuarter, 5 Little France Drive, Edinburgh EH16 4UU, UK

Vilde Drageset Haakensen

Department of Cancer Genetics, Institute for Cancer Research, Oslo University Hospital, Oslo, Norway, Department of Oncology, Oslo University Hospital, Oslo, Norway

Åslaug Helland

Department of Cancer Genetics, Institute for Cancer Research, Oslo University Hospital, Oslo, Norway, Department of Oncology, Oslo University Hospital, Oslo, Norway, Faculty of Medicine, University of Oslo, Norway

David Tamborero

Department of Oncology and Pathology, Karolinska Institutet, Science for Life Laboratory, Solna, SE-17165, Sweden

Henrik J. Johansson

Department of Oncology and Pathology, Karolinska Institutet, Science for Life Laboratory, Solna, SE-17165, Sweden

Rui M. Branca

Department of Oncology and Pathology, Karolinska Institutet, Science for Life Laboratory, Solna, SE-17165, Sweden

Maria Planck

Division of Oncology, Department of Clinical Sciences, Lund and CREATE Health Strategic Center for Translational Cancer Research, Lund University, Lund, Sweden, Department of Respiratory Medicine and Allergology, Skåne University Hospital, Lund, Sweden

Johan Staaf

Division of Oncology, Department of Clinical Sciences, Lund and CREATE Health Strategic Center for Translational Cancer Research, Lund University, Lund, Sweden

Lukas M. Orre

Department of Oncology and Pathology, Karolinska Institutet, Science for Life Laboratory, Solna, SE-17165, Sweden

---

Method Article

Keywords: cancer, multi-omics, proteomics, bioinformatics

Posted Date: November 22nd, 2021

DOI: https://doi.org/10.21203/rs.3.pex-1562/v1

Abstract

The associated publication reports proteogenomic analysis of non-small cell lung cancer (NSCLC), where we identified molecular subtypes with distinct immune evasion mechanisms and therapeutic targets, and validated our classification method in separate clinical cohorts. This protocol describes sections of the bioinformatics analysis of the multi-omics data, namely, data analysis and processing for panel sequencing, identification of cancer- and driver-related proteins in proteomics data, proteogenomics search, and machine learning-based classifiers for NSCLC subtyping. Specifically, a cohort classifier was built using support-vector machine-recursive feature elimination (SVM-RFE) algorithm applied to in-depth proteomics data from a cohort of 141 samples. The classifier was then validated in three external datasets. Another classifier, suitable for single-sample subtyping, was built using k-top scoring pairs (k-TSP) algorithm applied to label-free data from a cohort of 136 samples. The k-TSP-based classifier was validated in two independent cohorts and an additional external dataset.

Introduction

Reagents

Equipment

Procedure

Panel sequencing of early-stage NSCLC cohort: data analysis

BALSAMIC workflow v4.0.0[1] was used to analyze each of the FASTQ files. In summary, we first quality controlled FASTQ files using FastQC v0.11.5[2]. Adapter sequences and low-quality bases were trimmed using fastp v0.20.0[3]. Trimmed reads were mapped to the reference genome hg19 using BWA MEM v0.7.15[4]. The resulted SAM files were converted to BAM files and sorted using samtools v1.6[5,6]. Duplicated reads were marked using Picard tools MarkDuplicate v2.17.0 and promptly quality controlled using CollectHsMetrics, CollectInsertSizeMetrics, and CollectAlignmentSummaryMetrics functionalities. Results of the quality-controlled steps were summarized by MultiQC v1.7[7]. For each sample, somatic mutations were called using VarDict v2019.06.04[8] in tumor-only mode and annotated using Ensembl VEP v94.5[9]. Variants recurrently found (more than 10 cases) in the cohort and not previously described as oncogenic were manually reviewed to detect likely artifacts, which were removed from downstream analyses together with variants showing low quality calls. Variants were classified as putative functional versus passengers by using the interpretation pipeline developed by the Molecular Tumor Board Portal (accessed 2/2020), a clinical decision support tool that evaluates the functional and predictive relevance of genomic alterations[10]. Briefly, the portal classifies a variant as biologically relevant combining up-to-date results from clinical and preclinical studies, bona fide biological assumptions and bioinformatics calculations.

For tumor mutational load calculations, first all low-quality variants were removed via a hard filter of total read depth (DP) > 50 and alternative allele depth (AD) > 5. Thereafter, we followed the procedure demonstrated by Chalmers et al[11].

Downstream analysis of proteomics and proteogenomics data

Cancer- and driver-related proteins (CDRPs)

CDRPs were defined based on membership in 10 cancer-related signaling pathways as previously described[12], and/or if causally linked to cancer according to the COSMIC cancer gene census effort[13]. In total 832 CDRPs were identified and quantified in the current early-stage NSCLC cohort. CDRP annotation was performed using previously published information related to protein function as transcription factors, chromatin remodeling factor or transcription factor co-factor according to AnimalTFdb[14]; protein kinase[15]; protein phosphatase[16]; ubiquitin E3 ligase[17]; protein subcellular localization according to SubCellBarCode resource (www.subcellbarcode.org)[18]; and annotation as drug target[19].

Proteogenomics 6FT search

The IPAW proteogenomics pipeline for novel peptides was implemented as previously described[20]. Novel peptides from the 6-reading frame translation (6FT) search that passed SpectrumAI filter in the majority of TMT sets and lacked a SNPdb match were retained for outlier detection. Assuming that such peptides should be present in one or in a few samples and that the per set quantification depends on the sample composition, ratios to the reference pool were re-centered by the median and log2 transformed. Outlying peptides were determined by the same threshold used for the cancer-testis antigen analysis (*i.e.*, ratio > 3).

Peptides from the 6FT search were further annotated with ANNOVAR[21] (genes: RefSeq[22], UCSC[23], ENSEMBL[24], GENCODE[25] hg19; long non-coding RNAs: LNCipedia v.5.2[26], gencode.v34.long_noncoding_RNAs after liftOver from hg38 to hg19 coordinates , pseudogenes: gencode.v34.2wayconspseudos[27] after liftOver from hg38 to hg19 coordinates), a custom-made script for alternative open reading frame identification, and Uniprot[28] protein names (release 03/2020) for transposable elements assignment according to the blastp protein ID. Annotations were prioritized similar to ANNOVAR precedence rules with emphasis on the exon translation complexity (*AltOrf* -alternative opening reading frame) and the putative origin of the peptides (*ERV* -endogenous retro-viral elements, pseudogenes): *AltOrf, ERV, pseudogene, exonic, splicing, ncRNA_exonic, ncRNA_splicing, ncRNA_intronic, lncrna, UTR5, UTR3, UTR5;UTR3, intronic, upstream, downstream, upstream;downstream, intergenic.*

Machine-learning based classifiers for NSCLC proteomics data

Support-vector machine (SVM)-based cohort classifier

For an initial filtering to remove uninformative proteins (features) and to reduce computation time for downstream analysis, we applied DEqMS[29] as in Lehtiö et al. (associated publication) (BH adjusted p-value < 0.01 and |log2(ratio)| > 0.5, 5,872 proteins). Next, for a balanced first selection of features, for each comparison, the most upregulated and downregulated 200 (100 × 2) proteins were included, resulting in a list of 1,549 proteins after removal of redundant proteins.

Support-vector machine with linear kernel was used to build the classifier using scikit-learn library (v0.21.2) in Python (version 3)[30]. Numpy (v1.17.4) was used for data manipulation and operation.Hyperparameter C and the model was optimized using 5-fold cross-validation.

Due to data-availability constraints in this study, we used the Monte Carlo cross-validation (MCCV) method[31] to provide an unbiased performance estimation and to optimize the model. The whole process (described below) was repeated 100 times to maximize the number of samples included in training and testing. From each iteration, the testing performance (accuracy) and 200 most important features were reported.

First, we partitioned the dataset randomly into two parts: 80% for training and 20% for testing. To select the most important features in each iteration, support-vector machine-recursive feature elimination (SVM-RFE) algorithm was applied[32]. The algorithm was implemented using scikit-learn library (v0.21.2) in python (version 3)[30]. The model with the 200 most important features was then applied to the testing data to estimate the accuracy.

Finally, the overall accuracy was reported as the average accuracy from the 100 MCCV iterations, and we selected the most frequently used 200 features from the output of MCCV (100 iterations) to build the final model and deploy it.

Applying SVM classifier to external data

As the model was built on normalized proteomics data, training and testing data should be in the same scale in order to estimate the evaluation of the model robustly. Therefore, the model was built on Z-score-distributed data and the external data (GEO[33], TCGA[34], and Gillette et al.[35]) were transformed to Z-score distribution.

k-Top Scoring Pairs (k-TSP)-based single-sample classifier

The k-TSP algorithm[36], developed for solving binary classification problems, was used here for development of a diagnostic single-sample classifier intended for a clinical setting. The classifier was trained and applied on label-free data-independent acquisition mass spectrometry (DIA-MS) data generated as described in Lehtiö et al. (associated publication). To remove samples with low-quality DIA

data, sample-wise correlation (Spearman) analysis between the in-depth TMT-HiRIEF-LC-MS-DDA data (Lehtiö et al. associated publication) and the DIA-MS analysis was performed for overlapping proteins. This analysis revealed five samples with low correlation, possibly due to low amounts of available starting material for the DIA-MS analysis, and these samples were excluded from downstream analysis.

For an initial filtering to remove uninformative proteins (features) and to reduce computation time for downstream analysis, we applied DEqMS[29] as described above (BH adjusted p-value < 0.01 and |log2(FC)| > 0.5). Comparison between differentially abundant 5,872 proteins and the 6,717 proteins identified in the DIA analysis resulted in an overlap of 3,028 proteins.

Missing values in DIA data were imputed by filling baselines signals for each protein, individually. We assumed that any resulting missing value was due to the lack of protein abundance in the sample. Therefore, we imputed the missing values with baseline signals instead of inferring the missing value based on protein abundance of other samples. We sampled value from a Gaussian distribution N(μ, σ) where μ is half of the minimum MS1 peak area of the protein abundance and σ is 2 in order to replace missing values with baseline signals for each sample independently.

Protein-wise correlation (Spearman and Pearson) between TMT-HiRIEF-LC-MS-DDA and imputed DIA-MS was computed for these 3,028 proteins, and proteins with greater than 0.3 Spearman and 0.5 Pearson correlations were included, resulting in a list of 2003 proteins. Next, for each comparison, the most upregulated and downregulated 100 (50 × 2) proteins were included in subsequent analysis resulting in a list of 760 proteins.

For k-TSP classification, we modified the 'switchbox' R package (v1.24.0)[37] for multi-class classification problems. The only parameter to tune is the number of feature pairs (k) used in the k-TSP algorithm (optimized k = 13). One-versus-one classifiers were built to classify samples (in total 15 classifiers for the 6 subtypes), and for each classifier the sample was classified into either of the subtypes. Consequently, each sample was classified 15 times and the final decision was made based on a majority vote. In case of a tie in classifications, direct comparison between the subtypes with the highest equal votes determined the final classification. When there was a tie between direct subtypes comparison, the sample was labeled "unclassified" to prevent final ambiguous calls.

As for the SVM classifier we used the MCCV method[31] to provide an unbiased performance estimation and to optimize the classifier. The whole process (described below) was repeated 100 times and for each iteration the testing performance (accuracy) and 195 (15 × 13) most important feature pairs were reported.

First, we partitioned the dataset randomly into two parts: 80% for training and 20% for testing. In the training data, 15 classifiers (*Subtype 1* vs. *Subtype 2*, *Subtype 1* vs. *Subtype 3*, etc.) were built independently, while simultaneously determining the 13 feature pairs for each classifier. Next, the corresponding classifiers were applied to the testing data to estimate the classifier accuracy.

Finally, the overall accuracy was reported as the average accuracy from the 100 MCCV iterations. To build the final model and deploy it, all feature pairs from the MCCV iterations were sorted based on frequency and the top 13 most frequent pairs for each of the 15 classifiers were selected resulting in a total of 195 feature pairs (244 marker proteins).

Applying k-TSP classifier to independent validation- and late-stage cohorts datasets

The k-TSP algorithm does not require any data normalization steps. It only compares the quantitative values of the proteins in each pair and assign samples to subtypes based on rules established during training. Thus, the k-TSP algorithm was directly applied to new DIA-MS sample data from two independent NSCLC cohorts after imputation of the missing values with 1. Furthermore, the classifier was applied to an external data-dependent acquisition (DDA)-MS data from a NSCLC adenocarcinoma cohort[38].

Troubleshooting

Time Taken

Anticipated Results

References

1. Foroughi Asl, H. BALSAMIC: A bioinformatic analysis pipeline for somatic mutations in cancer [Online]. Available online at: https://github.com/Clinical-Genomics/BALSAMIC. (2019).

2. Andrews, S. A Quality Control Tool for High Throughput Sequence Data [Online] Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/. (2010).

3. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884-i890, doi:10.1093/bioinformatics/bty560 (2018).

4. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760, doi:10.1093/bioinformatics/btp324 (2009).

5. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987-2993, doi:10.1093/bioinformatics/btr509 (2011).

6. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079, doi:10.1093/bioinformatics/btp352 (2009).

7. Broad Institute. Picard toolkit. *Broad Institute, GitHub repository* (2019).

8. Lai, Z. *et al.* VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. *Nucleic Acids Res* **44**, e108, doi:10.1093/nar/gkw227 (2016).

9. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol* **17**, 122, doi:10.1186/s13059-016-0974-4 (2016).

10. Tamborero, D. *et al.* Support systems to guide clinical decision-making in precision oncology: The Cancer Core Europe Molecular Tumor Board Portal. *Nat Med*, doi:10.1038/s41591-020-0969-2 (2020).

11. Chalmers, Z. R. *et al.* Analysis of 100,000 human cancer genomes reveals the landscape of tumor mutational burden. *Genome Med* **9**, 34, doi:10.1186/s13073-017-0424-2 (2017)

12. Sanchez-Vega, F. *et al.* Oncogenic Signaling Pathways in The Cancer Genome Atlas. *Cell* **173**, 321-337 e310, doi:10.1016/j.cell.2018.03.035 (2018).

13. Futreal, P. A. *et al.* A census of human cancer genes. *Nat Rev Cancer* **4**, 177-183, doi:10.1038/nrc1299 (2004).

14. Zhang, H. M. *et al.* AnimalTFDB 2.0: a resource for expression, prediction and functional study of animal transcription factors. *Nucleic Acids Res* **43**, D76-81, doi:10.1093/nar/gku887 (2015).

15. Manning, G., Whyte, D. B., Martinez, R., Hunter, T. & Sudarsanam, S. The protein kinase complement of the human genome. *Science* **298**, 1912-1934, doi:10.1126/science.1075762 (2002).

16. Chen, M. J., Dixon, J. E. & Manning, G. Genomics and evolution of protein phosphatases. *Sci Signal* **10**, doi:10.1126/scisignal.aag1796 (2017).

17. Li, W. *et al.* Genome-wide and functional annotation of human E3 ubiquitin ligases identifies MULAN, a mitochondrial E3 that regulates the organelle's dynamics and signaling. *PLoS One* **3**, e1487, doi:10.1371/journal.pone.0001487 (2008).

18. Orre, L. M. *et al.* SubCellBarCode: Proteome-wide Mapping of Protein Localization and Relocalization. *Mol Cell* **73**, 166-182 e167, doi:10.1016/j.molcel.2018.11.035 (2019).

19. Santos, R. *et al.* A comprehensive map of molecular drug targets. *Nat Rev Drug Discov* **16**, 19-34, doi:10.1038/nrd.2016.230 (2017).

20. Zhu, Y. *et al.* Discovery of coding regions in the human genome by integrated proteogenomics analysis workflow. *Nat Commun* **9**, 903, doi:10.1038/s41467-018-03311-y (2018).

21. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* **38**, e164, doi:10.1093/nar/gkq603 (2010).

22. O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* **44**, D733-745, doi:10.1093/nar/gkv1189 (2016).

23. Kent, W. J. *et al.* The human genome browser at UCSC. *Genome Res* **12**, 996-1006, doi:10.1101/gr.229102 (2002).

24. Yates, A. D. *et al.* Ensembl 2020. *Nucleic Acids Res* **48**, D682-D688, doi:10.1093/nar/gkz966 (2020).

25. Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res* **22**, 1760-1774, doi:10.1101/gr.135350.111 (2012).

26. Volders, P. J. *et al.* LNCipedia 5: towards a reference set of human long non-coding RNAs. *Nucleic Acids Res* **47**, D135-D139, doi:10.1093/nar/gky1031 (2019).

27. Pei, B. *et al.* The GENCODE pseudogene resource. *Genome Biol* **13**, R51, doi:10.1186/gb-2012-13-9-r51 (2012).

28. UniProt, C. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res* **47**, D506-D515, doi:10.1093/nar/gky1049 (2019).

29. Zhu, Y. *et al.* DEqMS: A Method for Accurate Variance Estimation in Differential Protein Expression Analysis. *Mol Cell Proteomics* **19**, 1047-1057, doi:10.1074/mcp.TIR119.001646 (2020).

30. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* **12** (2011).

31. Xu, Q.-s. & Liang, Y.-Z. Monte Carlo cross validation. *Chemometrics and Intelligent Laboratory Systems* **56**, 1-11 (2001).

32. Guyon, I., Weston, J., Barnhill, S. & Vapnik, V. Gene Selection for Cancer Classification using Support Vector Machines. *Machine Learning* **46**, 389-422, doi:10.1023/A:1012487302797 (2002).

33. Lim, S. B., Tan, S. J., Lim, W. T. & Lim, C. T. A merged lung cancer transcriptome dataset for clinical predictive modeling. *Sci Data* **5**, 180136, doi:10.1038/sdata.2018.136 (2018).

34. Cancer Genome Atlas Research, N. Comprehensive molecular profiling of lung adenocarcinoma. *Nature* **511**, 543-550, doi:10.1038/nature13385 (2014).

35. Gillette, M. A. *et al.* Proteogenomic Characterization Reveals Therapeutic Vulnerabilities in Lung Adenocarcinoma. *Cell* **182**, 200-225 e235, doi:10.1016/j.cell.2020.06.013 (2020).

36. Tan, A. C., Naiman, D. Q., Xu, L., Winslow, R. L. & Geman, D. Simple decision rules for classifying human cancers from gene expression profiles. *Bioinformatics* **21**, 3896-3904, doi:10.1093/bioinformatics/bti631 (2005).

37. Afsari, B., Fertig, E. J., Geman, D. & Marchionni, L. switchBox: an R package for k-Top Scoring Pairs classifier development. *Bioinformatics* **31**, 273-274, doi:10.1093/bioinformatics/btu622 (2015).

38. Xu, J. Y. *et al.* Integrative Proteomic Characterization of Human Lung Adenocarcinoma. *Cell* **182**, 245-261 e217, doi:10.1016/j.cell.2020.05.043 (2020).