Erasmus MC - Cancer Computational Biology Center (CCBC)

Available Internship projects. BSc. & MSc.

Contact:

Dr. Ir. Harmen van de Werken
Managing Director Cancer Computational Biology Center
h.vandewerken@erasmusmc.nl

Or

Ing. Job van Riet
PhD-Student Bioinformatics
j.vanriet@erasmusmc.nl

Cancer Computational Biology Center
Erasmus Medical Center
The Netherlands

December 8, 2017

Contents

List of Figures

1 | Who are we?

The Cancer Computational Biology Center (CCBC) is a bioinformatics and computational biology core facility within the Erasmus Medical Center located in Rotterdam. It is the largest of the eight university medical centers in The Netherlands and ranks as 21st in the Ranking of Best Global Universities for Clinical Medicine 2016 and has world-leading scores on scientific output. [1]

The CCBC aims to innovate, facilitate and stimulate omics-based **cancer** research, including genomics, transcriptomics and proteomics, within the Erasmus Medical Center. We are a relatively young department as we have only been founded since 1st of September 2014 by Prof. Dr. Ir. Guido Jenster, Prof. Dr. Ir. John Martens, Prof. Dr. Leendert Looijenga and managing director Dr. Harmen van de Werken.

We closely collaborate (and share our knowledge) with other research groups within the Erasmus MC e.g. the department of Medical Oncology, Urology, Pathology and Hematology, and thus have a wide range of projects and biological interests. Amongst our resources, we have several high-performance PC's, servers and commercially-available software packages available. We primarily make use of Unix-based operating systems and open-source software and algorithms.

Several of our past and current projects can be seen on our website. If you have any interesting Bioinformatics (or IT) projects of your own and would like to pursue these further, we are also happy to perform these projects in our department (or in collaboration with other Erasmus MC departments)!

If you wish to be part of a young and growing team and help us on some **very interesting** projects, we can help you develop in the following fields:

1. Big-data analysis and storage of omics-based techniques (e.g. methylation sequencing, exome-sequencing, RNA-sequencing, proteomics, proteogenomics) and also microarrays.
2. Structured/OO programming and general software development using git. (Main languages: Python, Perl, C, R, Java, Unix/Bash, LaTeX)
3. Biological insight and (diagnostic) interpretation.
4. We are members of the MolMed school which allows us to offer you various courses for extra ECTS.

2 | Project I Proteogenomics

2.1 Abstract

Multiple breast-cancer cell-lines were investigated using in-house bottom-up proteomics and DNA/RNA-sequencing experiments. The capability to couple these two techniques was investigated by designing a pipeline which could detect and verify sample-specific DNA mutations which led to protein sequence alterations. This preliminary pipeline could accurately detect sample-specific DNA-verified SNPs,SNVs and (certain) InDels in the proteomics dataset which would not be detected by standard proteomics analysis. This provides compelling arguments to further develop these computational algorithms and establish an EMC-wide platform to allow future use of proteomics data in correspondence with other forms of -omics data to better understand underlying causal traits of cancer.

We seek a student to help in both developing a more advanced and general pipeline and assist in the biological analysis of using proteogenomics to detect cancer-specific mutations in multiple cancer types. Ideally, we would like to design a pipeline/package in the **R** language and publish this in the BioConductor suite. [2] This can possibly be accompanied alongside an application note in a respectable journal.

2.2 Introduction

Due to the increasing expertise, (clinical) use and ease of DNA/RNA-sequencing and LC-MS/MS within the ErasmusMC, coupled with a significant increase of sensitivity in current state-of-the-art in-house mass-spectrometers such as the Thermo Scientific Orbitrap Fusion Lumos Tribrid Mass Spectrometer, we wish to further assess and utilize the possibility for combining these multiple levels of genetic information.

A simple bottom-down proteomics design can be summarized as thus:

1. Isolate proteins from cell-cultures or patient biopsies (e.g. breast-cancer cell-lines or tissue slices of breast-cancer tumours).
2. Perform fractionation of proteins
   (a) Reduces the size of the protein pool to be measured as to detect more proteins-of-interest at various dynamic ranges, e.g. perform a measurement without the highly-expressed proteins.
3. Generation of protein fragments by digestion of a protease, most commonly trypsine which cleaves after every Arginine (R) or Lysine (K) except if a Proline (P) follows; [RK]!P. These fragments are called **proteolytic fragments**.
4. Add optional isobaric labelling as identifier of samples when pooling multiple samples in a single run. (TMT-6plex / TMT-10plex)
5. Calculate the m/z of each fragment by measuring the Time of Flight (TOF) from protein fragments after ionization through a vacuum tube of a constant length.
6. Perform peak picking to remove background noise and determine thresholds for fragment calling.
7. Match the peaks, called **spectra**, based on molecular weights to a database containing protein fragments which has been *in silico* generated from proteins treated with the same protease, preferably using multiple search algorithms. [3] Matches of spectra against peptides of proteins are called **peptide-spectrum matches (PSM)**.
8. Perform annotation and statistics of PSMs.

A large number of the measured spectra cannot be matched to a canonical (non-mutated) human protein fragment. The incorporation of different levels of sample-specific data (such as DNA-validated mutations) potentially increase the number of matching spectra and subsequently provide protein-level evidence for DNA mutations.

This intersection of multiple -omics techniques to assist proteomics analysis is termed proteogenomics. [4] This recent and emerging field has already successfully identified *novel* protein-coding transcripts and accompanying proteins, including some transcripts previously thought to be of non-coding properties.
Proteogenomics allows searching and matching of mass-tandem spectra against databases filled with additional (sample-specific) protein sequences derived from DNA/RNA experiments and/or generated *in silico*.
This approach can provide insight into the translational patterns of (mutated onco-)genes, identification of splice variants, reading-frame determination, post-translational modification and even provide evidence for *novel* protein-coding transcripts. [5–7]

We have already performed a pilot study within the Erasmus MC using multiple breast cancer cell-lines to identify sample-specific mutations in high-resolution proteomics data which would not have been detected with standard proteomics analysis. We successfully identified MSH6 and P53 mutations in the correct cell-lines using the following strategy (S1 for simplified flowchart):

1. Perform a LC-MS/MS experiment with a pooled sample-set of breast cancer cell-lines and perform peak picking.

2. Match spectra to protein databases containing all canonical human proteins and isoforms using search algorithms based on molecular weight of fragments.

3. Remove all spectra which matched with a high probability score to a protein fragment (PSM); keeping unidentified spectra which did not match correctly to canonical protein fragments.

4. Perform an additional search with the spectra left after filtering. This second database contains sample-specific validated DNA mutations translated into protein sequence.

5. Identify PSMs which matched to distinct and unique proteolytic fragments originating from mutant protein sequences; these are fragments with a different molecular weight than their canonical fragments. This provides evidence of the presence of the mutant protein.

2.3 Aims

Our pilot study proved to be successful in identifying several mutant proteins, however, we still have to implement and optimize many features. Thus, we would like to develop a more advanced pipeline in **R**, starting from scratch.

This new pipeline should make use existing BioConductor packages such as the MSnbase package [8] for the data-structure of storing and accessing MS-based data as described in BioConductor manuals. [9] Incorporating mutations from external cancer databases, such as the COSMIC database, could also potentially increase the sensitivity of the pipeline. Another aspect on which we could greatly improve

the workflow are the statistical methods we employ to limit the amount of false positives and false negatives.

We have breast cancer cell-line data, processed on a state-of-the-art high-resolution mass-spectrometer, with validated DNA mutations which allows us to also validate and hypothesize on the biological implications of any found protein variants.

The student is tasked with aiding the development of this pipeline and process the available biological data. The student is welcome to suggest any improvements or alternative approaches to tackling this project and investigate these further.

We are collaborating with experts in the proteomics field and can help the student gain insight into every aspect of proteomics, genomics and *proteogenomics*. This can seem as a challenging project but we will provide assistance at every step. We feel this project has great potential and would welcome any assistance!

2.4 Main activities and learning outcomes

1. Gaining insight into proteomics workflows employing bottom-up mass-spectrometry and subsequent data analysis.
2. **R** programming and package development using git.
   (Suitable for entry into the widely acclaimed BioConductor suite)
3. Software optimization for parallel and resource-efficient functionality.
   (Possibly write some C code for larger calculations if needed)
4. Visualization of big-data using novel approaches, e.g. Circos plot of cleavage sites with proteolytic peptides.
5. Finding sample-specific mutations (detected and validated on DNA/RNA level) on protein level to further characterize and gain insight into cancer.
6. Gaining experience with Linux environments and Bash commands.

2.5 Contact

You can always contact us for additional information or questions on this project. Please contact Harmen van de Werken and/or Job van Riet.

(h.vandewerken@erasmusmc.nl / j.vanriet@erasmusmc.nl).

3 | Project II Fusion-gene detection

3.1 Abstract

One of the root causes (and effects) of many cancers is genetic instability. This can be seen by the amount of abnormalities on DNA-level in driver or passenger mutations. One of the subsets of these genetic abnormalities are fusion-genes, these are genes with their own independent gene-bodies and surrounding elements (promotors, enhancers etc.) yet, after a chromosomal rearrangement, are found to be "fused" together. One such prominent example of a fusion-gene of clinical interest in prostate cancer is the fusion of TMPRSS2 with ETS transcription factor genes, e.g. TMPRSS2::ERG or TMPRSS2::ETV1. Current high-resolution molecular sequencing gives us more opportunities to detect these chromosomal rearrangements and we would thus like to pursue this opportunity, both on RNA level and to provide evidence on protein level using proteogenomics. Analysis can be be performed on both high-resolution in-house and external data-sources.

3.2 Introduction

Cancer is hallmarked by an enormous variety in chromosomal and cellular abnormalities due to the many mechanisms which drive tumor initiation and progression. However, amongst all this heterogeneity, recurring mutational patterns have been discovered and validated. One of these patterns is indicative of two or more independent gene-bodies seemingly 'fused' together. These chromosomal rearrangements can generally be classified in two groups. The first type of rearrangements are characterized by (partial) fusion of two or more gene-bodies resulting in a *novel* fusion-gene and subsequent fusion-protein of new or altered activity. The second type are characterized by chromosomal rearrangements juxtaposing promotor and/or enhancer elements from e.g. highly-expressed survival-related genes to proto-oncogenes. One such fusion product of interest is the fusion of TMPRSS2 and ETS transcription factor genes such as ERG and ETV1, often seen in prostate cancer. [10–12] Another famous finding is the BCR-ABL fusion-gene, which is now used as a highly sensitive test for chronic myelogenous leukemia (CML) and which led to the development of the drug imatinib mesylate. [13]

A striking example for chromosomal rearrangement in cancer is chromothripsis. [14] This is an extreme form of genomic instability resulting in high amounts of DNA lesions, random joining of broken DNA ends and occurrences of deletions. The molecular cause of this phenomenon is still debated, one postulation relates this to the formation of micronuclei following chromosome missegregation. [15] Yet, even with random joining of the DNA, recurrent patterns emerge out of this chaotic event; signifying *mostly unknown* biological functionality (or at least importance) on fusion products.

These fusion products can be found using RNA-sequencing as there will be reads overlapping the fusion sites (breakpoints) and differences in the overall gene-body of genes can be seen. E.g. if the second half of a gene is fused with another gene of a higher transcriptional rate, one can postulate that the amount of sequenced reads originating from the fused segment will be greater than those from the original gene-body due to the availability of more transcripts for sequencing. There are a multitude of algorithms utilizing these and other principles to detect and report fusion products; often with great differences in the final fusion candidates signifying a high rate of false positives.

3.3 Aims

We propose a strategy utilizing and combining multiple fusion-detection algorithms to establish the most significant overlapping results and to test these in external data-sources such as The Cancer Genome Atlas and protein databases to find evidence on protein level. We would like to combine this into a single pipeline for processing multiple cancer types. We will use known fusions products in prostate cancer to validate functionality and specificity.

3.4 Main activities and learning outcomes

1. Gaining detailed insight into RNA-sequencing and cancer biology e.g. important chromosomal aberrations.
2. **R**, Python/Perl/Bash programming (or language of choice) and software development using git.
3. Use of multiple fusion-detection algorithms and gaining insights into the approaches of these tools.
4. Multi-omics data integration, using both RNA and protein level data to provide evidence.
5. Visualization of big-data using novel approaches.
6. Gaining experience with Linux environments and Bash commands.

3.5 Contact

You can always contact us for additional information or questions on this project. Please contact Harmen van de Werken and/or Job van Riet.

(h.vandewerken@erasmusmc.nl / j.vanriet@erasmusmc.nl).

4 | Project III Allelic Imbalance

4.1 Abstract

Human genomes are usually of diploid nature and thus contain two identical copies of each gene (alleles). Allelic imbalance (AI) is the observable effect where one of the alleles is lost or multiplied due to the genomic aberrations instituted by defects in the cellular repair mechanisms leading to cancer development and progression. Allele-specific expression (ASE) is the observation that only one allele is expressed whilst there are still two (or more) alleles present. We propose to build (and publish) a novel pipeline to detect these events on a genome-wide basis using sample-specific DNA and RNA-sequencing.

4.2 Introduction

Under normal circumstances, human genomes are diploid and consist of 22 pairs of autosomal chromosomes and one allosome pair of sex chromosomes (XX / XY). The autosomal chromosomes (and XX) have two identically paired homologs, one received from the maternal side and one received from the paternal side during fertilization, which is subsequently carried onward in all future human daughter cells. The homologs of each chromosome provide the same genetic information such as genes and enhancers at the same loci, resulting in a ∼1:1 ratio of mRNA expression of the two alleles (copies of each gene). A loss (or gain due to copy number variation) of one of these two alleles is termed allelic imbalance.

Allele-specific expression (ASE) occurs when both alleles are still present, yet only one of these alleles is expressed due to genomic silencing. One striking example of ASE which occurs in all mammals is genomic imprinting; parental silencing of genes during embryogenesis and later life which has been linked to various diseases ranging from obesity to psychiatric disorders. [16] Of particular interest is that ASE has also been detected and linked to cancer development in various cancer types. [17–19]

ASE can be detected by RNA-sequencing due to heterozygous SNPs between the maternal and paternal homologs, e.g. one allele has a C on a certain position (5') whilst the other allele has a T on that position (5'). By counting the number of ATCG bases for each read on these heterozygous positions, it can be deduced if both alleles are equally expressed or if one allele is silenced, e.g. 100 reads containing the base for the reference allele and ∼zero reads containing alternative base (4.1). [20]

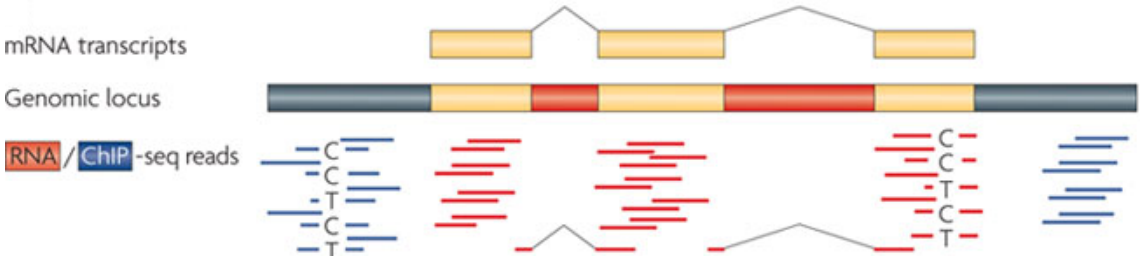


Figure 4.1: RNA-sequencing can be used to interrogate allele-specific expression when reads overlap a site that provides a high-quality heterozygote call. The ratio of reads from each allele is calculated and allelic bias at such sites is determined by deviation of the expected 50:50 ratio. *Figure: Pastinen et. al.*

This method can be used to detect local and regional silencing of genomic loci. An interesting concept is if there are two heterozygous positions flanking a gene and both show ASE, it is likely that the gene in-between is also silenced. This would be of paramount interest for flanked tumor suppressors as loss of expression of one allele of these genes is often sufficient for early-cancer development.

4.3 Aims

We previously had an intern investigate multiple methods and available algorithms/tools for discovering allelic imbalances. The main disadvantage and "room-for-improvements" are that most of these available methods work only on a smaller genomic region of interest and face complications when performed on a genome-wide basis such as computational speed and statistical complexities. We propose to develop an optimized algorithm in **R** (and/or C) and making use of additional genomic information such as sample-specific SNP arrays and whole exome sequencing (WES) data to provide an genome-wide analysis of AI/ASE in cancer and cell-lines.

4.4 Main activities and learning outcomes

1. Gaining insight into DNA assays, RNA-sequencing and cancer biology.
2. **R** programming and package development using git.
   (Suitable for entry into the widely acclaimed BioConductor suite)
3. Software optimization for parallel and resource-efficient functionality.
   (Possibly write some C code for larger calculations if needed)
4. Simulation of datasets to test statistical power and computational speed.
5. Visualization of big-data using novel approaches.
6. Gaining experience with Linux environments and Bash commands.

4.5 Contact

You can always contact us for additional information or questions on this project. Please contact Harmen van de Werken and/or Job van Riet.

(h.vandewerken@erasmusmc.nl / j.vanriet@erasmusmc.nl).

5 | Project IV Alternative Transcription and splicing in cancer

5.1 Abstract

Cancer biology is still poorly understood, even using the latest technologies at our disposal. A large amount of data has been generated from all forms of genetic information, from DNA and RNA to protein level and epigenetics. We would like to use these in-house and external datasets to investigate alternative transcription and splicing patterns in prostate cancer biology and its effect on clinical outcome. We propose to build a pipeline making use of existing algorithms.

5.2 Introduction

After the completion of the human draft genome and the exhaustive generation of omics related data, we are moving more and more from data generation to data analysis and data integration. However, the origin and progression of cancer is still poorly understood, despite all the new technologies that are made available in the last decades. Moreover, cancer-omics data analysis has been challenging, because of the genomic plasticity of the human cancer genome.

It has been known for decades that alternative promoter usage, alternative poly-adenylation and alternative splicing are profound mechanisms in cancer. Therefore, we wish to analyse our prostate RNA-seq data more deeply, and focus on alternative transcription and splicing in malignant tissue with the aim of identifying novel mechanisms and novel biomarkers in prostate cancer.

5.3 Aims

The student should make a modular RNA-sequencing pipeline (from input reads to pathways), which focuses on alternative transcripts and alternative splicing in cancer. The pipeline should be combined and complementary of our current RNA-sequencing pipeline. The results from the prostate samples can be integrated with in-house generated proteomics and clinical data and, subsequently, with publicly available datasets from The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC). The findings may contribute to better understanding of prostate cancer and improving personalized cancer treatment.

5.4 Main activities and learning outcomes

1. Gaining insight into RNA-sequencing, proteomics and cancer biology coupled with clinical data.
2. Python/Perl/Bash programming and package development using git.
3. Use of state-of-the-art algorithms in detecting alternative transcription.
4. Gaining experience with Linux environments and Bash commands.

5.5 Contact

You can always contact us for additional information or questions on this project. Please contact Harmen van de Werken and/or Job van Riet.

(h.vandewerken@erasmusmc.nl / j.vanriet@erasmusmc.nl).

6 | Project V Deciphering chemoresistancy profiles.

6.1 Introduction

Baker's yeast (*Saccharomyces cerevisiae*) is a well-known and powerful eukaryote model organism in biology. We have used yeast as a tool to obtain more insight into the mode of action of important anticancer drugs as well as the mechanisms of drug resistance.

Initially many cancers respond well to chemotherapy, however, frequently chemoresistance develops hampering further effective treatment. We initiated a semi-high throughput screen in which 4600 *S. cerevisiae* gene disruption strains (knockouts) were screened for sensitivity towards the widely used cytostatic drugs cisplatin and doxorubicin.

About 550 knockout strains have a clearly aberrant drug sensitivity profile, identifying genes whose disruption causes resistance or sensitivity. The vast majority (93%) shows increased sensitivity to cisplatin and/or doxorubicin implying that the genes involved have protective roles against these toxic substances. Only in 7% of the cases, a gene disruption results in a resistant phenotype.

In order to make sense of the genes and the biochemical pathways they belong to, we intend to use the program Cytoscape. This software will enable us to integrate our quantitative data from the screens with yeast protein interaction data, metabolite mediated interaction data and protein-DNA interaction data and thus establishing a graphic network model. Within the molecular networks, we will identify biological modules (biomodules) i.e. associations of preferred molecular partners that interact to perform a collective function. We will thus highlight biochemical pathways that govern sensitivity and resistance to commonly used anticancer drugs cisplatin and doxorubicin.

To be able to translate the findings in yeast to the human situation we will investigate if these biochemical pathways/networks are also present in humans and whether they play a determining role in drug sensitivity / resistance in cancer patients.

6.2 Approach

We are looking for a motivated and enthusiastic student with a keen interest in bioinformatics and biology to carry out this project that will result in a scientific publication. We envision the student to be a co-author on this paper. The student is expected to get acquainted with the program Cytoscape and analyse the data from the yeast screens. Once the genes and biochemical pathways involved in drug sensitivity / resistance have been elucidated, it is our intention to identify their human orthologs.

6.3 Contact

If you are interested in this project or want to know more about the project, feel free to contact Erik Wiemer or Harmen van de Werken.

(h.vandewerken@erasmusmc.nl / e.wiemer@erasmusmc.nl).

Bibliography

[1] TMF and Elsevier. *Subject Ranking 2015-2016: clinical, pre-clinical and health top 100*, 2015-2016.

[2] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015.

[3] D. Shteynberg, A. I. Nesvizhskii, R. L. Moritz, and E. W. Deutsch. Combining results of multiple search engines in proteomics. *Mol. Cell Proteomics*, 12(9):2383–2393, Sep 2013.

[4] I. Nesvizhskii. Proteogenomics: concepts, applications and computational strategies. *Nature Methods*, 11:1114âĂŞ1125, 2014.

[5] J. A. Alfaro, A. Sinha, T Kislinger, and P. C. Boutros. Onco-proteogenomics: cancer proteomics joins forces with genomics. *Nature Methods*, 11:1107–1113, 2014.

[6] B. Zhang and R. R. Townsend et al. Proteogenomic characterization of human colon and rectal cancer. *Nature*, 513(7518):382–387, Sep 2014.

[7] D. A. Bitton, D. L. Smith, Y. Connolly, P. J. Scutt, and C. J. Miller. An integrated mass-spectrometry pipeline identifies novel protein coding-regions in the human genome. *PLoS ONE*, 5(1):e8949, 2010.

[8] Gatto L and Lilley K. MSnbase - an R/Bioconductor package for isobaric tagged mass spectrometry data visualization, processing and quantitation. *Bioinformatics*, 28:288–289, 2012.

[9] Gatto L and Lilley K. *RforProteomcs: Companion package to the 'Using R and Bioconductor for proteomics data analysis' publication.*

[10] K. G. Hermans, A. A. Bressers, H. A. van der Korput, N. F. Dits, G. Jenster, and J. Trapman. Two unique novel prostate-specific and androgen-regulated fusion partners of ETV4 in prostate cancer. *Cancer Res.*, 68(9):3094–3098, May 2008.

[11] D. Gasi Tandefelt, J. Boormans, K. Hermans, and J. Trapman. ETS fusion genes in prostate cancer. *Endocr. Relat. Cancer*, 21(3):R143–152, Jun 2014.

[12] S. A. Tomlins, D. R. Rhodes, S. Perner, S. M. Dhanasekaran, R. Mehra, X. W. Sun, S. Varambally, X. Cao, J. Tchinda, R. Kuefer, C. Lee, J. E. Montie, R. B. Shah, K. J. Pienta, M. A. Rubin, and A. M. Chinnaiyan. Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science*, 310(5748):644–648, Oct 2005.

[13] G. Marcucci, D. Perrotti, and M. A. Caligiuri. Understanding the molecular basis of imatinib mesylate therapy in chronic myelogenous leukemia and the related mechanisms of resistance. Commentary re: A. N. Mohamed et al., The effect of imatinib mesylate on patients with Philadelphia chromosome-positive chronic myeloid leukemia with secondary chromosomal aberrations. Clin. Cancer Res., 9: 1333-1337, 2003. *Clin. Cancer Res.*, 9(4):1248–1252, Apr 2003.

[14] P. J. Stephens, C. D. Greenman, B. Fu, F. Yang, G. R. Bignell, L. J. Mudie, E. D. Pleasance, K. W. Lau, D. Beare, L. A. Stebbings, S. McLaren, M. L. Lin, D. J. McBride, I. Varela, S. Nik-Zainal, C. Leroy, M. Jia, A. Menzies, A. P. Butler, J. W. Teague, M. A. Quail, J. Burton, H. Swerdlow, N. P. Carter, L. A. Morsberger, C. Iacobuzio-Donahue, G. A. Follows, A. R. Green, A. M. Flanagan, M. R. Stratton, P. A. Futreal, and P. J. Campbell. Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell*, 144(1):27–40, Jan 2011.

[15] K. Crasta, N. J. Ganem, R. Dagher, A. B. Lantermann, E. V. Ivanova, Y. Pan, L. Nezi, A. Protopopov, D. Chowdhury, and D. Pellman. DNA breaks and chromosome pulverization from errors in mitosis. *Nature*, 482(7383):53–58, Feb 2012.

[16] J. Peters. The role of genomic imprinting in biology and disease: an expanding view. *Nat. Rev. Genet.*, 15(8):517–530, Aug 2014.

[17] G. Ha, A. Roth, D. Lai, A. Bashashati, J. Ding, R. Goya, R. Giuliany, J. Rosner, A. Oloumi, K. Shumansky, S. F. Chin, G. Turashvili, M. Hirst, C. Caldas, M. A. Marra, S. Aparicio, and S. P. Shah. Integrative analysis of genome-wide loss of heterozygosity and monoallelic expression at nucleotide resolution reveals disrupted pathways in triple-negative breast cancer. *Genome Res.*, 22(10):1995–2007, Oct 2012.

[18] L. Valle, T. Serena-Acedo, S. Liyanarachchi, H. Hampel, I. Comeras, Z. Li, Q. Zeng, H. T. Zhang, M. J. Pennison, M. Sadim, B. Pasche, S. M. Tanner, and A. de la Chapelle. Germline allele-specific expression of TGFBR1 confers an increased risk of colorectal cancer. *Science*, 321(5894):1361–1365, Sep 2008.

[19] B. B. Tuch, R. R. Laborde, X. Xu, J. Gu, C. B. Chung, C. K. Monighetti, S. J. Stanley, K. D. Olsen, J. L. Kasperbauer, E. J. Moore, A. J. Broomer, R. Tan, P. M. Brzoska, M. W. Muller, A. S. Siddiqui, Y. W. Asmann, Y. Sun, S. Kuersten, M. A. Barker, F. M. De La Vega, and D. I. Smith. Tumor transcriptome sequencing reveals allelic expression imbalances associated with copy number alterations. *PLoS ONE*, 5(2):e9317, 2010.

[20] T. Pastinen. Genome-wide allele-specific analysis: insights into regulatory variation. *Nat. Rev. Genet.*, 11(8):533–538, Aug 2010.

7 | Supplementary data

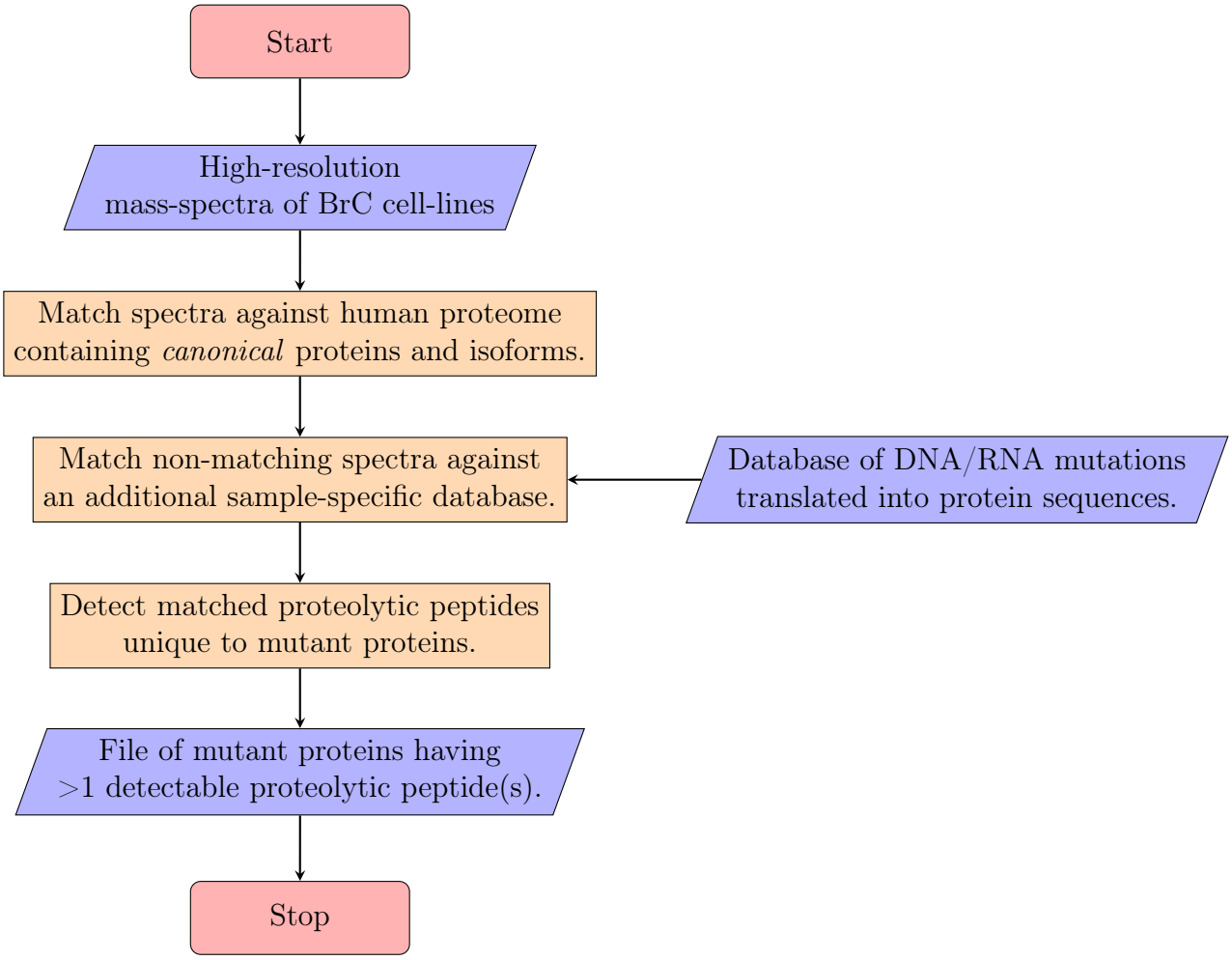
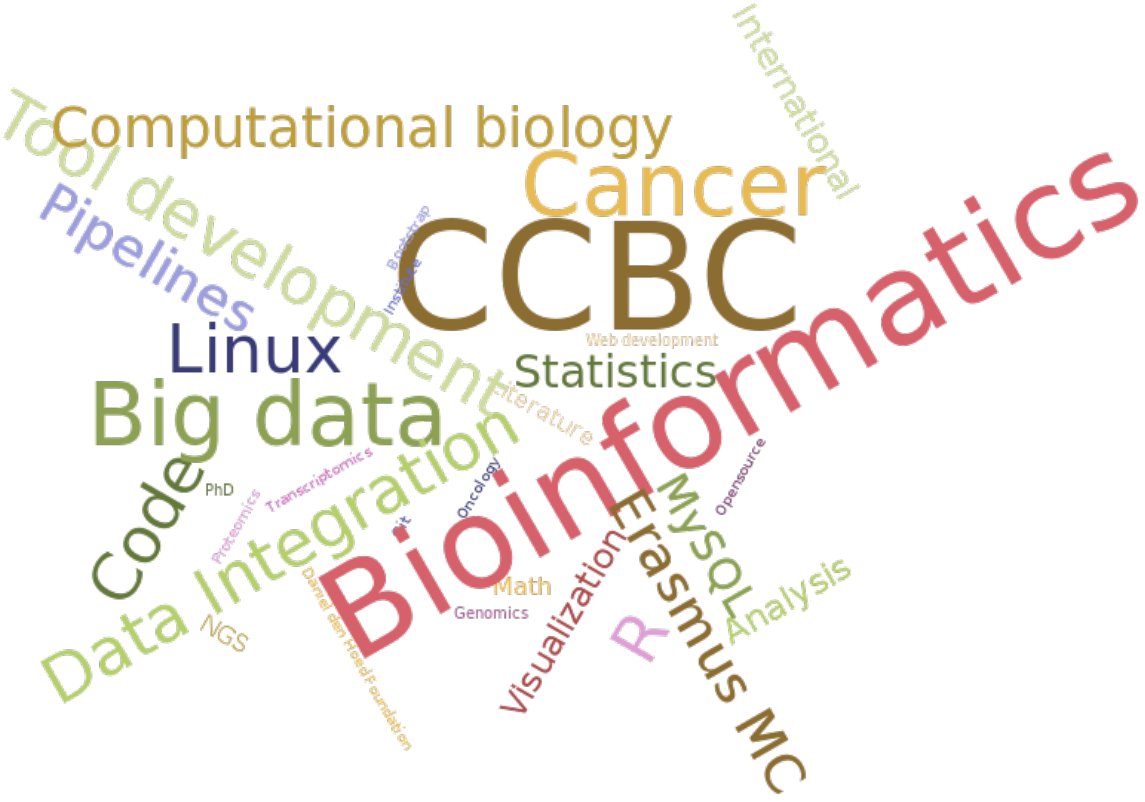


Figure S1: Flowchart depicting the simplified steps in the Erasmus MC proteogenomics pilot study.



Join us!

https://ccbc.erasmusmc.nl