Humboldt University

Master Thesis Expose

Development of a Mutation Panel for Neuroendocrine Tumor Research

Author:
Peter Moor
moor@fu-berlin.de

Supervisors:
Dr. Liam Childs
Prof.Dr. Ulf Leser
Prof.Dr.rer.nat. Christine Sers
Prof.Dr. Heike Siebert

December 22, 2014

1 Motivation and Background

With the unprecedented wealth of information available, it's extremely difficult to obtain a whole picture of the genetic basis of diseases [1]. For the past decade great effort has been spent on determining associations between genes and diseases by using the combination of experimental and computational methods [2, 3]. Another issue is the fast increase of literature in the life science domain. No database can keep track of the relevant knowledge that is regularly published [4]. Therefore, text mining algorithms are used to automatically extract the information about the relationships between biomedical entries reported in the literature. However, we are still far from fully understanding disease causation, especially complex diseases like cancer [3].

Neuroendocrine Tumors (NETs) are a rare but clinically important neoplasia, arising from uncontrolled proliferation of neuroendocrine tissue in most organs of the body. The subtypes of NETs share many common pathological features [5, 6]. Basically, there are functioning and non-functioning NETs, depending on whether there is a hormone secretion or not. Terms reflecting the type of the hormone have been applied to these NETs, e.g. insulinoma, glucagonoma and gastrinoma. Several grading systems such as the WHO classification refer to the proliferation and differentiation of the tumor. The distinction of well-differentiated from poorly differentiated NETs is probably one of the most important pathological assessments of these neoplasms.

Being rare, NETs are also vastly under-investigated. In 2014 there were only 335 publications in comparison with 14,126 for breast cancer, 8,209 for lung cancer and 7,013 for prostate cancer. As such, little progress has been made in clinical treatment or understanding the underlying factors involved in NET development. In general, a number of potentially druggable targets have been identified in different human cancers via whole genome or exome sequencing. While a vast amount of information became available, there is no correlation as to what extend such genetic alterations are directly predictive for drug response, except e.g. specific cases in colorectal and lung cancer. Furthermore, it has become clear that while the same major oncogenic pathways are involved in different cancers, mutations, translocations or amplifications are often specific for certain tumor types and need to be identified via a targeted cancer-specific approach.

2 Goal

The goal of this thesis is to develop a sequencing panel for the IonTorrent PGM platform targeting the most frequently mutated genes in NETs. With this approach one may quickly identify mutations specific to this type of tumor. Having an individual mutation profile for each patient enables a wide variety of investigations and can potentially improve the therapeutic success in NETs treatment.

By keeping the panel design pipeline generic (i.e., by parameterizing all tasks), we aim to provide a pipeline that greatly simplifies panel design also for other diseases of interest or to update the panel for NET once it has become better investigated.

3 Related Work

Several studies have successfully identified mutations specific to NET subtypes.

Jiao et al. [7] investigated PNETs. To explore the genetic basis, they determined the exomic sequences of a discovery set of ten patients and then screened the most commonly mutated genes in 58 additional PNETs. Somatic mutations in tumor suppressor genes MEN1, DAXX, ATRX, PTEN, TSC2 and the PIK3CA oncogene were identified and the genetic differences between PNETs and PDAC (pancreatic ductal adenocarcinomas) ware determined.

Banck et al. [8] analyzed 48 Small Intestine Neuroendocrine Tumors (SI-NETs) and normal tissue counterparts by massively parallel exome sequencing. This represents the first exome-wide sequencing study for this tumor type. SI-NETs are the largest group of NETs by organ site and the most common ma-

lignancy of the small bowel. They detected an average of 0.1 protein-altering somatic SNPs per 10^6 nucleotides. This suggests that SI-NETs are genetically stable cancers. Relevant alterations of cancer genes ware found in FGFR2, MEN1, HOOK3, EZH2, MLF1, CARD11, VHL, NONO, SRC, AURKA, EGFR, HSP90, PDGFR, SMAD family as well as AKT1 or AKT2 of PI3K/Akt/mTOR signaling.

Cao et al. [9] analysed insulinomas, the main type of functional PNETs. The major genetic basic of functional PNETs and the differences to non-functional PNETs have not been fully clarified. To explore the unique driver genes in sporadic insulinoma, 10 pairs of tumor and matched blood DNA were selected for whole exome sequencing. They identified 78 somatic mutations, containing recurrent T372R mutations in the transcription factor YY1, which was significantly higher than the background. The screening of additional 103 insulinomas revealed this mutation in 30% of all tumors. In addition, they found somatic mutations in three potential cancer-related genes, including MLL3, H3F3A and LMO2.

Yuan et al. [10] analyzed various PNET associated genes by using high-throughput Sanger sequencing to determine the links between the gene mutations and the clinicopathological features and prognosis of the patients. They identified 133 somatic mutations of the tumor suppressor genes DAXX/ATRX, KRAS, MEN1, mTOR pathway genes PTEN and TSC2, SMAD4/DPC, TP53 and VHL in 37 Chinese patients. In contrast to the results reported by Jiao et al. [7], the KRAS, TP53, PTEN, TSC2 and VHL genes had significantly higher mutation rates. The potential reasons for these differences could be the ethnic background of involved patients (85% Caucasian vs. 100% Chinese).

4 Approach

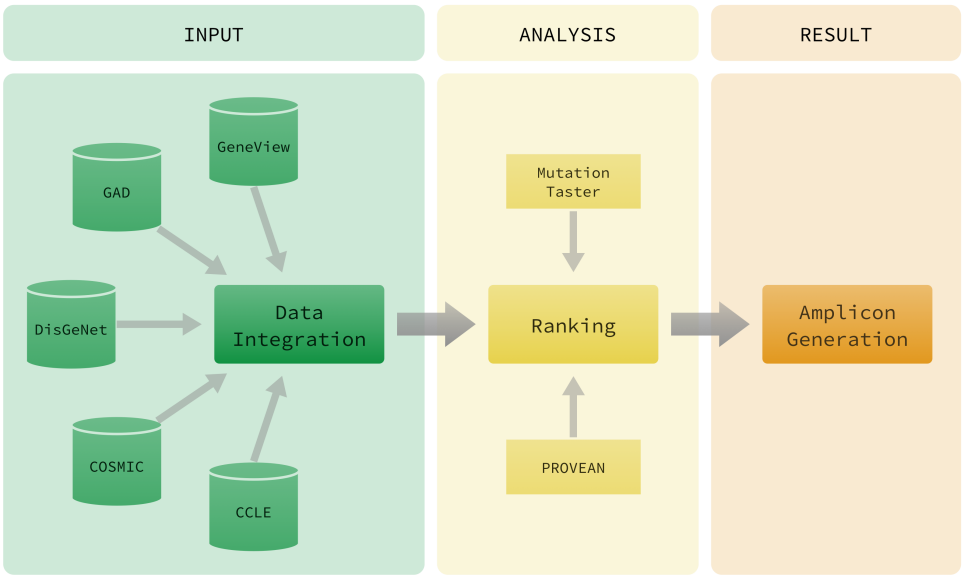

Figure 1: This is a general overview of the pipeline. Due to certain issues like incomplete stocks, wrong entries, spelling errors and much else, data integration from different sources is needed, described in section 4.1. After finding and merging the available information specific to NETs, there may be thousands of mutations. To restrict monetary costs and the amount of DNA needed for a single analysis, it may be necessary to limit the number of mutations of interest. Therefore, it's essential to rank them using gene/mutation prioritisation tools as described in section 4.2. Once there is an authoritative set of ranked mutations, their positions can be submitted to Ion AmpliSeq Designer developed by Life Technologies [11]. This tool provides a custom panel by designing primer pairs in hours for PCR-based target selection, see also section 4.3.

4.1 Data Integration

The necessary data can be obtain from several repositories. There are approaches using text mining tools for literature-based information extraction as well as computational and experimental methods to provide gene associated diseases or a combination of both.

A central problem is that a large body of knowledge is scattered over millions of scientific publications where searching specific information becomes troublesome. Most names of biomedical objects for instance genes and diseases often produce unspecific and too large search results. To address these issues we primary use GeneView [12], a web-based application providing access to a comprehensively annotated version of all articles from PubMed and the PubMed Central open access subset. It uses a variety of state-of-the-art text-mining tools optimized for recognizing mentions from 10 different entity classes i.e. genes (normalized to Entrez Gene IDs), mutations (normalized to dbSNP identifiers), species (normalized to the NCBI taxonomy), chemicals, histone modifications (normalized to the Brno histone modification nomenclature), other mentions of cell-types, diseases, drugs, enzymes and tissues and for automatically identifying protein-protein interactions. However, as it is mostly restricted to PubMed abstracts as source, GeneView doesn't use full text search and could probably miss something. Therefore, integration of other sources is needed to add or doublecheck the available information.

The Genetic Association Database (GAD) [13] provides a standardized molecular nomenclature of archived genetic association study data. Several data fields can be searched such as disease phenotypes, gene-based molecular data, chromosomal and mutation information, sample sizes, significance values, population information and allele description.

DisGeNET [1] has been developed by integrating information from four repositories (OMIM, UNIPROT, PharmGKB, CTD) [14–17] and from literature-derived human gene-disease network (LHGDN [18]). Each of these databases focuses on different phenotype to genotype relations. It allows access to a comprehensive database, including gene-disease associations for Mendelian, complex and environmental diseases.

The Cancer Cell Line Encyclopedia (CCLE) [19] provides public access analysis and visualization of DNA copy number, mRNA expression, mutation data and more.

Catalogue Of Somatic Mutations In Cancer (COSMIC) stores somatic mutation data and associated information extracted form primary literature, TCGA [20] and ICGC [21].

Another source are sequenced data sets of NETs which ware introduced in section 3. Even though they focus on a specific subtype of NETs and the amount of involved genes is small compared to other data sources, they provide exact information about specific mutations.

4.2 Ranking

To pick the most interesting mutations, one of the most important steps is to develop a reliable ranking method. A mutation of interest should occur in NETs and appear in multiple data sources. It should potentially change the codon of an amino acid and destroy the function of the resulting protein. There are several tools that evaluate the impact of a variation e.g. SIFT [22], MutationTaster2 [23], PolyPhen-2 [24] and PROVEAN [25]. MutationTaster2 and PROVEAN predict not only the damaging effects of single amino acid substitutions but also insertions, deletions and multiple amino acid substitutions. It is necessary to calculate a weighted score which combines the different impacts. In the last step of this process, a threshold has to be defined to choose the most interesting mutations.

4.3 Amplicon Generation

To design the panel we will use a web service called Ampliseq provided by IonTorrent[26]. After choosing a definitive list of mutations involved in NETs, we will submit them as "regions of interest" in Browser Extensible Display (BED) format, which contains the chromosome and position of each mutation.

The web service designs a "panel" of primers that target the submitted regions of interest. As it is

possible for primers to interact with each other, interacting primers are separated out into separate pools. However, the more pools there are, the more DNA is required to perform a sequencing run. Thus the service needs to minimise the pools and the interactions among primers within a pool by shifting the positions of the primers around the regions of interest. In doing so, the tool may choose to discard some primers altogether. For the current project, we wish to limit the panel to just two pools due to monetary and material costs.

5 Working Plan

Figure 2: This approach is going to be implemented in Python. The first eight weeks are planned for interface implementations to different sources, collect the data and store it in a standardized structure. In the next six weeks a reliable ranking method has to be developed to focus on the interesting entries. Finally the results have to be validated. This could be done by constructing a data set including a known ratio of NETs and other diseases. The last eight weeks are proposed for the written part of the master thesis.

References

[1] Anna Bauer-Mehren, Michael Rautschka, Ferran Sanz, and Laura I. Furlong. "DisGeNET: a Cytoscape plugin to visualize, integrate, search and analyze gene–disease networks". In: *Bioinformatics* 26.22 (). PMID: 20861032, pp. 2924–2926. ISSN: 1367-4803, 1460-2059. DOI: 10.1093/bioinformatics/btq538. URL: http://bioinformatics.oxfordjournals.org/content/26/22/2924.

[2] David Botstein and Neil Risch. "Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease". In: *Nature Genetics* 33 (), pp. 228–237. DOI: 10.1038/ng1090. URL: http://www.nature.com/ng/journal/v33/n3s/full/ng1090.html.

[3] Maricel G. Kann. "Advances in translational bioinformatics: computational approaches for the hunting of disease genes". In: *Briefings in Bioinformatics* 11.1 (). PMID: 20007728, pp. 96–110. ISSN: 1467-5463, 1477-4054. DOI: 10.1093/bib/bbp048. URL: http://bib.oxfordjournals.org/content/11/1/96.

[4] Murat Cokol, Ivan Iossifov, Chani Weinreb, and Andrey Rzhetsky. "Emergent behavior of growing knowledge about molecular interactions". In: *Nature Biotechnology* 23.10 (), pp. 1243–1247. ISSN: 1087-0156. DOI: 10.1038/nbt1005-1243. URL: http://www.nature.com/nbt/journal/v23/n10/abs/nbt1005-1243.html.

[5] Kjell Öberg and Daniel Castellano. "Current knowledge on diagnosis and staging of neuroendocrine tumors". In: *Cancer and Metastasis Reviews* 30.1 (), pp. 3–7. ISSN: 0167-7659, 1573-7233. DOI: 10.1007/s10555-011-9292-1. URL: http://link.springer.com/article/10.1007/s10555-011-9292-1.

[6] Keith Langley. "The Neuroendocrine Concept Today". In: *Annals of the New York Academy of Sciences* 733.1 (), pp. 1–17. ISSN: 1749-6632. DOI: 10.1111/j.1749-6632.1994.tb17251.x. URL: http://onlinelibrary.wiley.com/doi/10.1111/j.1749-6632.1994.tb17251.x/abstract.

[7] Y. Jiao et al. "DAXX/ATRX, MEN1, and mTOR Pathway Genes Are Frequently Altered in Pancreatic Neuroendocrine Tumors". In: *Science* 331.6021 (), pp. 1199–1203. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1200609. URL: http://www.sciencemag.org/cgi/doi/10.1126/science.1200609.

[8] Michaela S. Banck et al. "The genomic landscape of small intestine neuroendocrine tumors". In: *Journal of Clinical Investigation* 123.6 (), pp. 2502–2508. ISSN: 0021-9738. DOI: 10.1172/JCI67963. URL: http://www.jci.org/articles/view/67963%5C#sd.

[9] Yanan Cao et al. "Whole exome sequencing of insulinoma reveals recurrent T372R mutations in YY1". In: *Nature Communications* 4 (). DOI: 10.1038/ncomms3810. URL: http://www.nature.com/ncomms/2013/131210/ncomms3810/full/ncomms3810.html.

[10] Fei Yuan, Min Shi, Jun Ji, Hailong Shi, Chenfei Zhou, Yingyan Yu, Bingya Liu, Zhenggang Zhu, and Jun Zhang. "KRAS and DAXX/ATRX Gene Mutations Are Correlated with the Clinicopathological Features, Advanced Diseases, and Poor Prognosis in Chinese Patients with Pancreatic Neuroendocrine Tumors". In: *International Journal of Biological Sciences* 10.9 (). PMID: 25210493 PMCID: PMC4159686, pp. 957–965. ISSN: 1449-2288. DOI: 10.7150/ijbs.9773. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4159686/.

[11] Life Technologies. *Ion Ampliseq Designer*. 2014. URL: https://www.ampliseq.com.

[12] Philippe Thomas, Johannes Starlinger, Alexander Vowinkel, Sebastian Arzt, and Ulf Leser. "GeneView: a comprehensive semantic search engine for PubMed". In: *Nucleic Acids Research* 40 (Web Server issue). PMID: 22693219 PMCID: PMC3394277, W585–W591. ISSN: 0305-1048. DOI: 10.1093/nar/gks563. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3394277/.

[13] Kevin G. Becker, Kathleen C. Barnes, Tiffani J. Bright, and S. Alex Wang. "The Genetic Association Database". In: *Nature Genetics* 36.5 (), pp. 431–432. ISSN: 1061-4036. DOI: 10.1038/ng0504-431. URL: http://www.nature.com/ng/journal/v36/n5/full/ng0504-431.html.

[14] Ada Hamosh, Alan F. Scott, Joanna S. Amberger, Carol A. Bocchini, and Victor A. McKusick. "Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders". In: *Nucleic Acids Research* 33 (Database issue). PMID: 15608251 PMCID: PMC539987, pp. D514–517. ISSN: 1362-4962. DOI: 10.1093/nar/gki033.

[15] The UniProt Consortium. "The Universal Protein Resource (UniProt) in 2010". In: *Nucleic Acids Research* 38 (suppl 1). PMID: 19843607, pp. D142–D148. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gkp846. URL: http://nar.oxfordjournals.org/content/38/suppl_1/D142.

[16] T. E. Klein et al. "Integrating genotype and phenotype information: an overview of the PharmGKB project". In: *The Pharmacogenomics Journal* 1.3 (), pp. 167–170. ISSN: 1470-269X. DOI: 10.1038/sj.tpj.6500035. URL: http://www.nature.com/tpj/journal/v1/n3/full/6500035a.html.

[17] Carolyn J. Mattingly, Michael C. Rosenstein, Allan Peter Davis, Glenn T. Colby, John N. Forrest, and James L. Boyer. "The Comparative Toxicogenomics Database: A Cross-Species Resource for Building Chemical-Gene Interaction Networks". In: *Toxicological Sciences* 92.2 (). PMID: 16675512, pp. 587–595. ISSN: 1096-6080, 1096-0929. DOI: 10.1093/toxsci/kfl008. URL: http://toxsci.oxfordjournals.org/content/92/2/587.

[18] Markus Bundschus, Mathaeus Dejori, Martin Stetter, Volker Tresp, and Hans-Peter Kriegel. "Extraction of semantic biomedical relations from text using conditional random fields". In: *BMC Bioinformatics* 9.1 (). PMID: 18433469, p. 207. ISSN: 1471-2105. DOI: 10.1186/1471-2105-9-207. URL: http://www.biomedcentral.com/1471-2105/9/207/abstract.

[19] Jordi Barretina et al. "The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity". In: *Nature* 483.7391 (), pp. 603–607. ISSN: 0028-0836. DOI: 10.1038/nature11003. URL: http://www.nature.com/nature/journal/v483/n7391/full/nature11003.html.

[20] National Human Genome Research Institute (NHGRI) National Cancer Institute (NCI). *The Cancer Genome Atlas*. 2014. URL: http://cancergenome.nih.gov.

[21] Thomas J. Hudson (Chairperson) et al. "International network of cancer genome projects". In: *Nature* 464.7291 (), pp. 993–998. ISSN: 0028-0836. DOI: 10.1038/nature08987. URL: http://www.nature.com/nature/journal/v464/n7291/full/nature08987.html.

[22] Prateek Kumar, Steven Henikoff, and Pauline C. Ng. "Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm". In: *Nature Protocols* 4.7 (). PMID: 19561590, pp. 1073–1081. ISSN: 1750-2799. DOI: 10.1038/nprot.2009.86.

[23] Jana Marie Schwarz, David N. Cooper, Markus Schuelke, and Dominik Seelow. "MutationTaster2: mutation prediction for the deep-sequencing age". In: *Nature Methods* 11.4 (), pp. 361–362. ISSN: 1548-7091. DOI: 10.1038/nmeth.2890. URL: http://www.nature.com/nmeth/journal/v11/n4/full/nmeth.2890.html.

[24] Ivan A. Adzhubei, Steffen Schmidt, Leonid Peshkin, Vasily E. Ramensky, Anna Gerasimova, Peer Bork, Alexey S. Kondrashov, and Shamil R. Sunyaev. "A method and server for predicting damaging missense mutations". In: *Nature Methods* 7.4 (). PMID: 20354512 PMCID: PMC2855889, pp. 248–249. ISSN: 1548-7105. DOI: 10.1038/nmeth0410-248.

[25] Yongwook Choi, Gregory E. Sims, Sean Murphy, Jason R. Miller, and Agnes P. Chan. "Predicting the Functional Effect of Amino Acid Substitutions and Indels". In: *PLoS ONE* 7.10 (), e46688. DOI: 10.1371/journal.pone.0046688. URL: http://dx.doi.org/10.1371/journal.pone.0046688.

[26] Life Technologies. *IonAmpliSeq Designer provides full flexibility to sequence genes of your choice.* 2012. URL: https://tools.lifetechnologies.com/content/sfs/brochures/IonAmpliSeq_CustomPanels_AppNote_CO111038_06042012.pdf.