

Point mutation analysis of four human colorectal cancer exomes

Lisa Thalheim, thalheim@informatik.hu-berlin.de

January 10, 2012

1 Introduction

As the affordability, performance and availability of Next Generation Sequencing (NGS) increases, sequencing data accumulates and with it the need to analyse it for useful information about the biological systems under study. Among the disciplines looking to gain insights from sequencing data is cancer research, which seeks to answer questions on cancer development, progression and treatment options. All three of these are linked to mutations in cancer cells, and so identifying and interpreting mutations in cancer genomes has been a major task in cancer genomics (Meyerson et al. [2010], Stratton [2011], Pfeifer and Hainaut [2011]).

This work aims to identify and analyse point mutations in NGS data from four different human cancer exomes. Point mutations are one particular type of mutation where exactly one base of the genome is altered with respect to the reference genome. To limit the scope of this work, this analysis excludes single nucleotide insertions and deletions (indels).

2 Materials

The data to be analysed in this work was generated by sequencing the exomes of four different established human colorectal cancer cell lines by the Charite Institute of Pathology. The four data sets consist of 50nt single-end quality-tagged reads sequenced using Applied Biosystems SOLiD technology. The number of reads in each data set is given in the table below. The data are in the form of colour space FASTQ files, which contain an identifier, a colour-space sequence string and a quality string for each read. The quality string encodes the Phred score for each position in the sequence. Heat maps that give an overview overall distribution of quality values across all reads and read positions are attached to this document.

Cell line ID	# reads total	# high-quality reads	# low-quality reads
caco	92849949	64945592	16340116
geo	99143727	21925427	62877746
lim	88664822	22509138	52988731
rko	79135892	15194586	53477080

Table 1: Summary of available reads for each cell line. High-quality reads are defined as reads with an average quality value of 20 or greater, low-quality reads are defined as having an average quality value of 11 or lower.

3 Methods

The proposed work consists of four major steps: alignment, identifying point mutations, characterization of point mutations, and clustering and comparison of these characterizations.

In the alignment step, the software tool SHRiMP (Rumble et al. [2009]) will be used to map the raw sequence reads to positions on the hg19 reference genome. Since the process of conversion from colour space to nucleotide space sequence data can lose some information, the alignment will be done on the colour space reads. The sequence reads will be converted to nucleotide space further downstream in the analysis pipeline for the tools that do not support colour space reads.

Identifying point mutations with reasonable confidence requires a good quality alignment, so the parameters for the alignment tool will be chosen such that a lower number of mapped reads at high confidence will be favoured over a higher number of mapped reads at lower confidence. This step also includes recording statistics about both the raw sequencing data and the alignment, namely number of reads, distribution of read lengths and qualities, and genome coverage of the final mapping.

The alignment then serves as input to the software tool SNVMix (Goya et al. [2010]), which was selected because it was designed specifically to identify point mutations in cancer genomes¹. The output of SNVMix will be filtered against the dbSNP database (Sherry et al. [2001]) of known single nucleotide polymorphisms. The polymorphisms recorded in this database are variations between healthy human genomes. They are thus not of immediate interest to the study of cancer cells' mutational landscape and will be excluded from further analysis.

During the characterization step, a number of characteristics will be recorded for each identified point mutation. These characteristics are: type of mutation (synonymous/nonsynonymous, nonsense/missense), the gene this mutation occurs in, predicted

¹Quoting Goya et al. [2010]: "Although tools exist for SNV discovery from NGS data, none are specifically suited to work with data from tumors, where altered ploidy and tumor cellularity impact the statistical expectations of SNV discovery."

impact of the mutation on this gene's product, pathways this gene is involved in, the mutation's CHASM score², and the gene's function according to GO annotation. A sketch of the approach for determining each of these characteristics is given below.

- Type of mutation: The start of the open reading frame (ORF) will be looked up in the most recent available genome annotation, specifically through the Ensembl API (Stabenau et al. [2004]). Knowing the ORF start, a simple codon lookup will be used to determine the type of mutation.
- Surrounding gene: This, too, will be looked up in the most recent available annotation through the Ensembl API.
- Predicted impact: There are three software tools available for use here - SIFT (Kumar et al. [2009]), SNAP (Bromberg and Rost [2007]) and PolyPhen (Sunyaev et al. [2001]). Whether all of these will be used complementarily or whether only a subset of them will be used remains to be determined after some initial experimentation with the actual data.
- Pathways: There are several databases available that assign genes to pathways. Which one of these databases will be used remains to be determined.
- CHASM score: If the mutation is a missense mutation, its CHASM score will be recorded.
- Gene Ontology: There are several databases available for use that assign genes to ontological groups. Which one of these databases will be used remains to be determined.

Additionally, the identified point mutations will be checked against the Catalogue of Somatic Mutations in Cancer (COSMIC) to see whether they have previously been described (Forbes et al. [2008]).

In the final clustering and comparison step, we will be looking for conspicuous clusters of point mutations in pathways and ontological groups. This "landscape" of clusters will be compared pairwise amongst the four cell line data sets, and with prior publications (Wood et al. [2007]).

Aside from obtaining the data analysis results as described above, the objective of this work is also to create an analysis pipeline which accepts a sequenced exome in the form of FASTQ read files and outputs a report containing the analysis results. This will permit repeating the same analysis on new datasets as they become available. The analysis pipeline will most sensibly be written in one scripting language.

²CHASM (Cancer-specific High-throughput Annotation of Somatic Mutations) is a tool that attempts to predict whether a given mutation is a driver or a passenger mutation, see (Carter et al. [2009]).

4 Attachments

Attached to this document are four PNG image files, caco.qual.png, geo.qual.png, lim.qual.png, and rko.qual.png, one for each of the cell lines caco, geo, lim, and rko. Each image shows a heat map that reflects the frequency of a certain quality value in a certain read position, aggregated over all available reads for this cell line.

References

- Yana Bromberg and Burkhard Rost. SNAP: predict effect of non-synonymous polymorphisms on function. *Nucl. Acids Res.*, 35, 2007.
- H Carter et al. Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations. *Cancer Research*, 69, 2009.
- S A Forbes et al. The Catalogue of Somatic Mutations in Cancer. *Current Protocols in Human Genetics*, 2008.
- R Goya, MG Sun, RD Morin, G Leung, G Ha, KC Wiegand, J Senz, A Crisan, MA Marra, M Hirst, D Huntsman, KP Murphy, S Aparicio, and SP Shah. SNVMix: predicting single nucleotide variants from next-generation sequencing of tumors. *Bioinformatics*, 26, 2010.
- P Kumar, S Henikoff, and PC Ng. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc.*, 4, 2009.
- Matthew Meyerson, Stacey Gabriel, and Gad Getz. Advances in understanding cancer genomes through second-generation sequencing. *Nature Review Genetics*, 11, 2010.
- Gerd P Pfeifer and Pierre Hainaut. Next-generation sequencing: emerging lessons on the origins of human cancer. *Current Opinions in Oncology*, 23, 2011.
- Stephen M. Rumble, Phil Lacroute, Adrian V. Dalca, Marc Fiume, Arend Sidow, and Michael Brudno. Shrimp: Accurate mapping of short color-space reads. *PLoS Comput Biol*, 5(5):e1000386, 05 2009. doi: 10.1371/journal.pcbi.1000386.
- ST Sherry, MH Ward, M Kholodov, J Baker, L Phan, EM Smigielski, and K Sirotkin. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research*, 29:308–311, 2001.
- Arne Stabenau, Graham McVicker, Craig Melsopp, Glenn Proctor, Michele Clamp, and Ewan Birney. The ensembl core software libraries. *Genome Research*, 14:929–933, 2004.
- Michael R Stratton. Exploring the genomes of cancer cells: progress and promise. *Science*, 331, 2011.

Shamil Sunyaev, Vasily Ramensky, Ina Koch, Warren Lathe III, Alexey S Kondrashov, and Peer Bork. Prediction of deleterious human alleles. *Hum. Mol. Genet.*, 10, 2001.

Laura D. Wood et al. The genomic landscapes of human breast and colorectal cancers. *Science*, 318, 2007.