Analysis of tumor heterogeneity using next-generation sequencing data

Synopsis of PhD thesis

Lőrinc Sándor Pongor

Semmelweis University
Doctoral School of Pathological Sciences

Supervisor: Dr. Balázs Győrffy, M.D., Ph.D., D.Sc.

Official reviewers: Dr. István Kocsis, M.D., Ph.D.,
Dr. Bálint Mészáros, Ph.D.

Head of the Final Examination Committee:
Dr. Gábor Veres, M.D., Ph.D., D.Sc.

Members of the Final Examination Committee:
Dr. András Újházy, M.D., Ph.D.,
Dr. Anita Alexa, Ph.D.

Budapest
2018

1. Introduction

Genetic composition of tumors dynamically changes during tumor growth. Novel genetic alterations are acquired which leads to genetic diversity of tumors between patients, as well as spatial heterogeneity within one tumor. The acquired alterations may affect efficiency of therapies, as well as survival of patients. Examples include emergent resistance against receptor tyrosine kinase treatment in patients with mutations in the KRAS oncogene, as well as worse overall survival in case of high expression of the KI67 proliferation marker.

Next-generation sequencing techniques are increasingly used in clinical diagnostics. These techniques usually involve analysis of thousands of genes or whole genomes, enabling identification of previously unknown, and clinically relevant alterations.

Biomarkers are generally used to predict therapeutic efficiency and patient survival. Many recent studies have shown that mutations of one gene usually can not efficiently predict response to targeted therapies, since alterations in other genes can similarly affect the same pathway. For instance, resistance against receptor tyrosine kinase treatment can emerge via alterations in BRAF, NRAS or PI3KCA genes, which, same as KRAS, are all downstream in the growth signaling pathway. Similarly, PARP inhibitors seem to be a promising therapy in breast and ovarian cancer patients with deficient homologous recombination, which may be caused by mutation of BRCA1/2, PALB2, ATM or CHEK2 genes.

2. Objectives

The objectives of my PhD work are as follows:

- Analyzing the effects of cell dispersal on mutation detection using *in vitro* experiments via
  1. Detection of cellular composition shift from *in vitro* cell line invasion assays using next-generation sequencing
  2. Assessing the reproducibility of next-generation sequencing and data analysis by sequencing cell line mixtures with known composition.
  3. Analyzing the effect of sequenced tumor sample size on the detected genetic composition in next-generation sequencing of ovarian cancer patients.

- Identify a relationship between genetic mutations and gene expression changes by:
  4. Examining the effect of mutation-induced gene expression changes on the survival of breast cancer patients
  5. Analyzing and interpreting gene expression changes associated with TP53 mutations

3. Methods

3.1. Database construction

Whole exome sequencing data and RNA-seq data for breast cancer patients were obtained from The Cancer Genome Atlas (TCGA). The data obtained consisted of 762 patients from the breast cancer cohort, and 555 samples from the non-small cell lung cancer cohort. Mutations were identified using the *Mutect* algorithm. The mutations identified were annotated with known genes using the *SNPeff* program. Raw gene expression data deriving from RNA-seq experiments were normalized using RSEM.

Microarray gene expression data for 5,934 patients were obtained from the EGA (European Genome-Phenome Archive) and GEO (Gene Expression Omnibus) databases. The downloaded raw Affymetrix CEL files were normalized using the MAS5 algorithm from the Bioconductor Affy package in R.

3.2. Genotype-2-outcome algorithm

The algorithm utilizes three datasets: TCGA gene mutation, TCGA gene expression, and independent microarray data for survival analysis. There are two major steps in the analysis, by which the algorithm takes an input gene name, and outputs Kaplan-Meier survival plots.

In the first step, the algorithm identifies genes with altered expression levels based on the mutation status of the selected input gene. This is achieved by comparing the expression level of each gene between the two patient groups (mutant vs, wild type with respect to the input gene) using an ROC (Receiver Operating Characteristic) analysis.

The second step involves survival analysis based on the independent microarray data. Genes with altered expression in a given patient are merged into a “surrogate gene” and the median expression value of the surrogate gene is used to split patients into ‘high’ and ‘low’ expression categories for survival analysis.

3.3. Cell line invasion and calibration experiments

We selected four melanoma cell lines for the experiments (A375, MEL-JUSO, SK-MEL-28 and MEWO). Three cell line specific mutations were validated using Sanger sequencing in each case. For the next-generation sequencing experiments, we selected A375 / MEWO and SK-MEL-28 / MEWO pairs, and the A375 / MEL-JUSO cell line pairs for fluorescent video microscopy experiments.

The ring invasion experiments were performed using FlexiPERM ® conB cell exclusion silicone rings. The first cell line was added to the inner segment of the silicone ring, after incubation and attachment, the ring was removed, and the second cell line was added to the entire surface of the plate. DNA was isolated from around the perimeter of the silicone ring, and a control sample was isolated from the perimeter around the edge of the plate after a 72h incubation and invasion period.

To examine the reproducibility of sequencing, we created a calibration set using two cell lines. During the experiments, two cell lines were combined in 2%, 5%, 10%, 25% and 50% compositions, were the MEWO cell line was selected as the major clone.

The DNA isolated during the invasion and calibration experiments was sequenced on an Ion 314 chip using an Ion PGM 200 sequencer with a mean coverage of 600x reads. Sequencing was performed using a custom targeted panel of 25 cell line specific mutations.

3.4. Sample collection and preparation from ovarian cancer patients

Tumor regions were collected from each of five ovarian cancer patients from the National institute of Oncology (Budapest). In case of each patient, three spatially separated tumor regions and one (control) normal blood sample was obtained. DNA isolation was performed using the DNeasy Blood and Tissue Kit from Qiagen.

Tumor DNA isolation was performed in three manners: 1) biopsy sample representing the clinical setting; 2) local sample, where DNA was extracted from three segments adjacent to the biopsy sample; 3) Global sample where DNA isolation was performed from all three regions of the tumor and combined prior to sequencing.

3.5. Bioinformatics analysis of next-generation sequencing data

Quality control of raw sequencing reads was performed using the *FastQC* program. Reads were trimmed using the *trimmomatic* package. Alginment was performed using the *BWA MEM* algorithm against the human reference genome. Aligned reads were processed by sorting and conversion to BAM format using *samtools*.

The sorted and aligned reads were deduplicated using the picard-tools software package. Indel realignment was performed using the GATK *realignertargetcreator* and *indelrealigner* in order to increase sensitivity of indel identification.

Mutations from the cell line sequencing data were identified using the *samtools mpileup* program. In short, we selected regions where mutations were expected from the *mpileup* output using reads with alignment quality above 2. Mutation frequencies were calculated based on the number of reads supporting either the reference or alteration group.

Somatic mutations were identified using the *GATK mutect2* algorithm for the ovarian cancer patient sequencing data. Since multiple samples were available for each patient, we created a program that performs joint somatic genotyping based on the individual mutations identified by mutect2. Germline mutations were identified using the *GATK haplotypecaller* algorithm. Functional mutation annotation was performed using *SNPeff*. Identified mutations were further annotated with the ClinVar database. Copy-number alteration analysis, tumor purity and ploidy estimation were performed using the *sequenza* program.

4. Results

4.1. Analyzing effects of cellular movement on next-generation sequencing

By utilizing next-generation sequencing, we were capable of following cell composition changes caused by cell motility and invasion. Based on the detected mutation frequencies, invasion of A375 into the MEWO section reached 18,6%, while invasion of SK-MEL-28 into the MEWO was only 8,6%. Interestingly, while cellular movement (velocity and displacement) of A375 and SK-MEL-28 was grossly similar in monoculture conditions, the two cell lines displayed significant differences in the invasion experiments (p=0,011) when comparing homozygous mutation frequencies. In contrast, this difference was not captured when utilizing heterozygous mutations for the analysis (p=0,39).

In order to better understand the relationship between mutation frequencies and composition, we performed calibration sequencing using known mixtures of the two cell lines. We found that standard deviation of mutation frequencies increased as the composition of the cell lines evened out. Highest standard deviations were calculated when the composition of the two cell lines was 50%-50%, which further increased when utilizing only heterozygous mutations.

Standard deviation of mutation frequencies between biological replicates reached 17% in the invasion and calibration sequencing as well. When comparing single mutation frequencies between technical replicates, deviations decreased to 5,6%. Our *in silico* results also showed that in cases where sequencing coverage decreased below 50x, standard deviations could reach 12,9%, while increasing coverage to 1000x decreased deviations below 2%.

4.2. Effects of sequenced tumor size on somatic mutations

We identified heterozygous germline mutations in genes associated to homologous recombination repair in three patients. In these cases, somatic loss of heterozygosity was identified in the tumor, and in addition, the “signature 3” somatic mutation signature associated with BRCA-deficiency was identified. Copy-number analysis revealed that these patients harbored many deletions and amplifications that affected smaller segments of chromosomes. Clonal somatic TP53 mutation was present in each patients' tumor, which is commonly partnered with DNA-repair deficiency.

One further patient was affected only by one major copy-number alteration event on chromosome 1. Somatic mutations of the tumor displayed the signature associated with the APOBEC deaminase mutation signature resulting in higher frequency of C>T and C>G mutations. Multiple somatic mutations were identified in genes associated to the PI3K pathway, such as the inactivation of the PTEN tumor suppressor, a common (activating) mutation of the FGFR2 receptor tyrosine kinase gene, as well as two heterozygous inactivating mutations in the PIK3R1 gene.

In the last patient we identified multiple large copy number alterations, affecting several complete chromosomes. This tumor had a hypermutating phenotype as it had three times the commonly expected level of somatic mutations. Two signatures with unknown aetiology were identified in the somatic mutations. In addition, two clonal mutations were identified, the most common p.G12D mutation in the KRAS gene, and the most common p.E454K mutation in the PI3KCA gene.

In case of non-hypermutating tumors, the percentage of common mutations was substantially higher compared to the hypermutating tumor, ranging between 69,3%-93,8%. The mean number of identified mutations was 149. Of the four non-hypermutating tumor patients, we found that more mutations could be identified in the biopsy in two cases, and more in the global sample in two cases.

We identified 688 mutations in the biopsy sample of the hypermutating tumor, of which 15,5% were identifiable in the other tumor regions as well. As sample size increased, the percentage of common mutations increased to 25%, while the overall mutation count decreased to 392 mutations. In case of the global sample, we identified 140 mutations, a substantial decrease compared to the biopsy sample, of which 71,4% were common in the other tumor regions.

4.3. Identifying gene expression changes associated to gene mutations

By utilizing gene mutation and gene expression data, the Genotype2Outcome (G-2-O) algorithm is capable of performing survival analysis by identifying genes displaying altered expression based on mutation, and performing survival analysis on an independent microarray dataset. Survival analysis based on the surrogate expression associated to the gene mutation of GATA3 (HR=1,66; p<E-16) outperforms survival analysis performed solely using gene mutation (p=1,3E-08) or gene expression

(HR=0,71; p=1,3E-08) status in breast cancer. Similarly, the surrogate expression associated to MAP3K1 outperforms (HR=1,8; p<E-16) the survival analysis performed only based on its expression (HR=1,6; p<E-16).

In the case of TP53 mutation we identified 23 genes with increased expression, which were associated to the cell cycle pathway in breast (p=1,3E-16) as well as in lung cancer data (p=1,1E-23). A subsequent survival analysis of these genes showed that in most cases, high expression of identified genes was associated with worse prognosis. By combining the top 10 genes into a surrogate gene, survival analysis showed substantially worse prognosis (HR=2,43, p<E-16) as compared to survival analysis performed on individual genes (HR<2.1, p<E-16). Interestingly, when performing the analysis on patient subgroups, we found a similar trend was observed on only wild type TP53 patients (HR=2,36, p=1,8E-3), but an inverted result was obtained when only TP53 mutant patients were included (HR=0,52, p=3,7E-2). I.e. in the latter case high expression was associated with better survival.

5. Conclusions

1. Somatic mutations usually affect not only a single gene, but may also indirectly affect gene expression levels and signaling pathways. Using the Genotype2Outcome system, we can identify prognostic gene expression changes (as surrogate genes) associated to a genetic alteration. We found that survival analysis using the surrogate gene could outperform survival analysis solely based on a single genes' expression or mutation status, thus the system can serve as an alternative for identification of biomarkers in cancer patients.

2. In breast cancer patients harboring TP53 mutations, the expression of several genes associated to cell cycle is increased. High expression of these genes was previously reported in publications as a prognostic factor in breast cancer patients. In our survival analyses, this association was found only in patients carrying a wild type TP53 gene, but better survival was found when only TP53 mutant patients were included in the analysis. In other terms, the effect of the expression of these genes on survival is apparently strongly influenced by the mutation status of TP53.

3. Using next-generation sequencing, we were able to follow the invasion of cell lines in *in vitro* experiments. Composition detection was more reliable when one of the cell lines was predominant.

4. The standard deviation of detected mutation frequencies was high in case of biological and technical replicates analyzed during the invasion and calibration experiments. High deviations can affect interpretation, and post-analysis (such as inferring of tumor evolution based on identified mutations). These results were also found in the *in silico* models, where deviations decreased as sequencing coverage increased, thus sequencing with high coverages can help misinterpretation in clinical sequencing.

5. Increasing the size of the sequenced tumor size does not strongly affect detected mutations during clinical sequencing. In other terms, current sampling practices used in clinical sequencing are satisfactory even though high sequence coverage seems to be important for optimal results. .

6. Bibliography of the candidate's publications

6.1. Publications related to the thesis:

1. **Pongor LS**, Kormos M, Hatzis C, Pusztai L, Szabo A, Gyorffy B. (2015) A genome-wide approach to link genotype to clinical outcome by utilizing next generation sequencing and gene chip data of 6,697 breast cancer patients. Genome Medicine, 7: 104. **IF=5,846**

2. Jiang T, Shi W, Wali VB, **Pongor LS**, Li C, Lau R, Gyorffy B, Lifton RP, Symmans WF, Pusztai L, Hatzis C. (2016) Predictors of Chemosensitivity in Triple Negative Breast Cancer: An Integrated Genomic Analysis. PLoS Medicine, 13: e1002193. **IF= 11,862**

3. Harami-Papp H*, **Pongor LS***, Munkacsy G, Horvath G, Nagy AM, Ambrus A, Hauser P, Szabo A, Tretter L, Gyorffy B. (2016) TP53 mutation hits energy metabolism and increases glycolysis in breast cancer. Oncotarget, 7: 67183-67195. **IF= 5,168**

4. Nagy A, **Pongor LS**, Szabo A, Santarpia M, Gyorffy B. (2017) KRAS driven expression signature has prognostic power superior to mutation status in non-small cell lung cancer. International Journal of Cancer, 140: 930-937. **IF=6,513**

5. **Pongor LS**, Harami-Papp H, Mehes E, Czirok A, Gyorffy B. (2017) Cell Dispersal Influences Tumor Heterogeneity and Introduces a Bias in NGS Data Interpretation. Scientific Reports, 7: 7358. **IF= 4,259**

6.2. Disszertációtól független publikációk jegyzéke

1. **Pongor LS**, Pinter F, Petak I. (2013) HeurAA: accurate and fast detection of genetic variations with a novel heuristic amplicon aligner program for next generation sequencing. PloS One, 8: e54294. **IF= 3.534**

2. **Pongor LS**, Vera R, Ligeti B. (2014) Fast and sensitive alignment of microbial whole genome sequencing reads to large sequence datasets on a desktop PC:

application to metagenomic datasets and pathogen identification. PloS One, 9: e103441. **IF= 3.234**

3. Cai W, Xiong Chen Z, Rane G, Satendra Singh S, Choo Z, Wang C, Yuan Y, Zea Tan T, Arfuso F, Yap CT, **Pongor LS**, Yang H, Lee MB, Cher Goh B, Sethi G, Benoukraf T, Tergaonkar V, Prem Kumar A. (2017) Wanted DEAD/H or Alive: Helicases Winding Up in Cancers. Journal of the National Cancer Institute, 109. **IF= 12.589**