Bioinformatics analysis of PDAC subtypes

Do Young Hyeon
Seoul National University https://orcid.org/0000-0002-1825-0097
Dowoon Nam
Korea University
Youngmin Han
Seoul National University College of Medicine
Duk Ki Kim
Seoul National University College of Medicine
Gibeom Kim
Seoul National University
Daeun Kim
Ajou University
Jingi Bae
Korea University
Seunghoon Back
Korea University
Dong-Gi Mun
Korea University
Inamul Hasan Madar
Korea University
Hangyeore Lee
Korea University
Su-Jin Kim
Korea University
Hokeun Kim
Korea University
Sangyeop Hyun
Ajou University
Chang Rok Kim
Seoul National University
Seon Ah Choi
Seoul National University
Yong Ryoul Kim
Seoul National University

Juhee Jeong
Seoul National University College of Medicine
Suwan Jeon
Seoul National University College of Medicine
Yeon Woong Choo
Seoul National University College of Medicine
Kyung Bun Lee
Seoul National University College of Medicine
Wooil Kwon
Seoul National University College of Medicine
Seunghyuk Choi
Hanyang University
Taewan Goo
Seoul National University
Taesung Park
Seoul National University
Young-Ah Suh
Seoul National University College of Medicine
Hongbeom Kim
Seoul National University College of Medicine
Ja-Lok Ku
Seoul National University College of Medicine
Min-Sik Kim
Daegu Gyeongbuk Institute of Science and Technology
Eunok Paek
Hanyang University
Daechan Park (✉ dpark@ajou.ac.kr)
Ajou University
Keehoon Jung (✉ keehoon.jung@snu.ac.kr)
Seoul National University College of Medicine
Sung Hee Baek (✉ sbaek@snu.ac.kr)
Seoul National University
Jin-Young Jang (✉ jangjy4@snu.ac.kr)
Seoul National University College of Medicine
Daehee Hwang (✉ daehee@snu.ac.kr)
Seoul National University
Sang-Won Lee (✉ sw_lee@korea.ac.kr)
Korea University

Method Article

Keywords:

Posted Date: January 19th, 2023

DOI: https://doi.org/10.21203/rs.3.pex-2064/v1

Abstract

The associated publication reports proteogenomic analysis of human pancreatic ductal adenocarcinoma (PDAC), where we provided significantly mutated genes (SMGs)/biomarkers, cellular pathways, and cell types as potential therapeutic targets to improve stratification of patients with PDAC. This protocol describes the detailed methods for bioinformatics analysis of PDAC subtypes, including tumour purity estimation, subtype prediction for tumour samples in previous cohorts, pathway enrichment analysis, kinase activity analysis, and pan-omics analysis.

Introduction

Reagents

Equipment

Procedure

Estimation of tumour cellularity

Tumour cellularity was estimated using KRAS VAF, qpure, ESTIMATE, and histological images. To measure KRAS VAF, we sequenced KRAS amplicons over 1,000,000 × depth for the previously reported mutational hotspots (hg19 chr12: 25398292-25398294, chr12:25398280-25398285, chr12:25380281-25380283, and chr12:25380275-25380277)[1]. Target regions in exons 2 and 3, ranging in size from 150–206 bp, were amplified using the following primers: Exon2 pair 1 (forward, 5'-TTATAAGGCCTGCTGAAAATGA-3'; reverse: 5'-GTATCAAAGAATGGTCCTGCAC-3'); Exon2 pair 2 (forward, 5'-CATTATTTTTATTA TAAGGCCTGCTG-3'; reverse, 5'-CAAGATTTACCTCTATTGTTGGATCA-3'); Exon3 pair 1 (forward, 5'-TGTGTTTCTCCCTTCTCAGGA-3’; reverse, 5'-AAACCCACCTATAATGG TGAATATCT-3'); and Exon3 pair 2 (forward, 5'-TCTCCCTTC-TCAGGATTCCTAC-3'; reverse, 5'-TGGTGAATATCTTCAAATGATTTAGT-3'). Polymerase chain reaction (PCR) products were purified using the QIAquick PCR Purification Kit (Qiagen), and equal quantities from each sample were pooled. Sequencing libraries were prepared from 50 ng of the pooled PCR product and then sequenced with paired ends in 151 bp reads on the Illumina NovaSeq platform. Low-quality sequencing reads were filtered out, and adaptors were trimmed for the remaining reads using Trimmomatic (version 0.36)[2]. The trimmed reads were aligned with the customized KRAS reference using BWA (version 0.7.17) and the MEM algorithm[3]. BAM and ‘mpileup’ files were processed from SAM files using Samtools (version 1.3.1)[4]. Three variant calling methods were used to call mutations in KRAS: GATK3 Mutect2 (version 3.8), GATK4 Mutect2 (version 4.1.9), and BCFtools (version 1.10.2)[5,6]. We then filtered GATK variants at hotspots using allele frequency (AF) ≥ 0.01 and tumor log odds (TLOD) ≥ 300, and BCFtools variants using AF ≥ 0.01. Finally, the VAF for KRAS mutations was defined as the sum of the mutation allele frequencies at the hotspots.

To estimate cellularity by qpure, DNA from tumours and matched blood samples were hybridized to an Infinium Global Screening Array (Illumina). The log R-ratio (LRR) and B-allele frequency (BAF) were determined from the probe intensities using GenomeStudio (version 2). Finally, qpure was applied to the LRR and BAF to estimate the cellularity, as previously described[7]. To estimate the cellularity, we applied ESTIMATE (version 1.0.13)[8] to mRNA data for calculating stromal and immune scores for individual samples based on 137 stromal genes and 140 immune-related genes. Tumour purity was estimated using the default formula of the ESTIMATE score, defined as a combination of stromal and immune scores. Finally, to estimate the cellularity from histological images for each sample, we counted cells on one representative haematoxylin and eosin-stained slide from a resected specimen with a visible tumour area. The cells were counted using QuPath image analysis[9], which was applied to digitally scanned slides at 200× magnification (AT2, Aprio technologies) with a few modifications. Briefly, the image analysis process was as follows: 1) cell segmentation using a cell detection function; 2) manual annotation of tumour cells on a small area of each slide; 3) creation of a cell classifier using cell-feature-based neural networks; 4) calculation of the number of tumour cells, total cells, and tumour area; and 5) error correction and final confirmation by manual inspection.

Integration of RNA signatures with published data

From three previously reported PDAC cohorts, TCGA[10], Australian (PACA-AU)[11], and Canadian (PACA-CA)[12], we downloaded the FPKM data from TCGA and PACA-AU and the normalised expression data from PACA-CA. We then filtered out non-PDAC tumour samples (e.g. intraductal papillary mucinous carcinoma, neuroendocrine carcinoma, and acinar cell carcinoma) from each cohort. From each of the published cohorts (e.g. TCGA cohort), we selected the genes that were identified as ‘expressed mRNAs’ in our cohort and calculated their $\log_2$(FPKM+1) values and $\log_2$ fold-changes with respect to their median values. We then applied quantile normalisation[13] to the $\log_2$ fold-changes across the individual samples so that their distributions would be equal to those in our samples. Subsequently, for each sample, we calculated Pearson’s correlation (ρ) coefficient for the $\log_2$ fold-changes determined for the rna1–3 genes in the published cohort relative to the mean $\log_2$ fold-changes of the rna1–3 genes in our RNA1 samples. The procedure was repeated for RNA2–3, resulting in three ρ values for the three groups (RNA1–3). We estimated an empirical null distribution of ρ for RNA1–3 by performing 10,000 random permutations in our samples. For each sample, we computed p-values for the three observed ρ values of RNA1–3 using the corresponding estimated distributions. Finally, samples with a p-value < 0.05 were classified into the corresponding mRNA cluster. If the samples had a p-value < 0.05, for multiple mRNA clusters, they were classified into the cluster for which they had the minimum p-value.

Integrated pathway analysis

For the integrated pathway analysis, we first identified molecular signatures for Sub1–6, identified from the integrated clustering based on the relationships between Sub1–6 and the clusters identified from the three data types. For example, for Sub6, we identified genes (S6-G) selected for RNA3 from the mRNA data and proteins (S6-P) selected for Prot5 and Phos5 from the proteome and phosphoproteome data, respectively. To identify the pathways represented by the genes and proteins identified for Sub1–6, we performed an enrichment analysis of cellular pathways for the genes and proteins selected for each subtype (e.g. S6-G and S6-P for Sub6) using ConsensusPathDB[14]. The cellular pathways represented by the genes and proteins for each subtype were identified as those with $P < 0.05$, and the number of molecules involved in the pathway ≥ 3.

Kinase activity analysis

We first extracted data for 16,055 kinase-substrate interactions for 473 kinases from PhosphoSitePlus[15], PhosphoELM[16], and SIGNOR[17]. We then selected 45,006 phosphopeptides detected in more than half of the samples for at least one of the six subtypes and generated a 150 × 45,006 matrix containing $\log_2$ fold-changes of the 45,006 phosphopeptides measured in 150 samples. For a kinase, we next estimated the activity in each sample as follows: 1) we selected a set of phosphopeptides measured in the sample with the same phosphorylation sites of its substrates as those in the databases; and 2) we performed GSEA[18] for the selected phosphopeptides in the sample, comparing them with the other phosphopeptides. The 2nd step was performed only when the number of selected phosphopeptides was ≥ 5. Normalised enrichment scores (NESs) from this procedure were included in a 173 × 150 matrix. Next, we performed an empirical *t*-test for each subtype to compare the NESs of the samples in the subtype with the NESs of the other samples. For each subtype, we estimated an empirical null distribution of *t*-statistic values from random permutations of 150 samples, 1,000 times. For a kinase, we computed an adjusted p-value for the subtype using a right-sided test for its observed *t*-statistic value using empirical distribution. Finally, we selected the kinases that were predominantly activated in each subtype as those with p-values ≤ 0.01 and fraction of samples with GSEA p-values < 0.05 as larger than 0.25.

Identification of pan-omics signatures for proteogenomic stratification

We identified the genes that harboured somatic mutations predominantly in each tumour subtype (TS) using the hypergeometric test ($P < 0.05$ and sample enrichment > two-fold): 19 genes for TS1, 3 genes for TS2, 4 genes for TS3, and 4 genes for TS4. Next, we identified mRNAs, proteins, and phosphopeptides that were predominantly up-regulated in TS1–4. For each molecule, we compared the $\log_2$ fold-changes in the samples belonging to each TS with those in the samples belonging to the other TS using the same statistical testing method and cut-offs used to identify molecular subtype signatures (e.g. rna1–3). We used an additional cut-off, *t*-test p-value < 0.05. The cellular pathways represented by these molecular

signatures for each TS were identified as those with $p < 0.05$ and number of molecules involved in the pathway ≥ 3, using ConsensusPathDB[14].

References

1 Biankin, A. V. *et al.* Pancreatic cancer genomes reveal aberrations in axon guidance pathway genes. *Nature* **491**, 399-405, doi:10.1038/nature11547 (2012).

2 Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-2120, doi:10.1093/bioinformatics/btu170 (2014).

3 Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589-595, doi:10.1093/bioinformatics/btp698 (2010).

4 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079, doi:10.1093/bioinformatics/btp352 (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 Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol* **31**, 213-219, doi:10.1038/nbt.2514 (2013).

7 Song, S. *et al.* qpure: A tool to estimate tumor cellularity from genome-wide single-nucleotide polymorphism profiles. *PLoS One* **7**, e45835, doi:10.1371/journal.pone.0045835 (2012).

8 Yoshihara, K. *et al.* Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat Commun* **4**, 2612, doi:10.1038/ncomms3612 (2013).

9 Bankhead, P. *et al.* QuPath: Open source software for digital pathology image analysis. *Sci Rep* **7**, 16878, doi:10.1038/s41598-017-17204-5 (2017).

10 The Cancer Genome Atlas Research Network. Integrated Genomic Characterization of Pancreatic Ductal Adenocarcinoma. *Cancer Cell* **32**, 185-203, doi: 10.1016/j.ccell.2017.07.007 (2017).

11 Scarlett, C. J., Salisbury, E. L., Biankin, A. V. & Kench, J. Precursor lesions in pancreatic cancer: morphological and molecular pathology. *Pathology* **43**, 183-200, doi:10.1097/PAT.0b013e3283445e3a (2011).

12 Zhang, J. *et al.* International Cancer Genome Consortium Data Portal–a one-stop shop for cancer genomics data. *Database (Oxford)* **2011**, bar026, doi:10.1093/database/bar026 (2011).

13 Bolstad, B. M., Irizarry, R. A., Astrand, M. & Speed, T. P. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19**, 185-193, doi:10.1093/bioinformatics/19.2.185 (2003).

14 Kamburov, A., Wierling, C., Lehrach, H. & Herwig, R. ConsensusPathDB–a database for integrating human functional interaction networks. *Nucleic Acids Res* **37**, D623-628, doi:10.1093/nar/gkn698 (2009).

15 Hornbeck, P. V. *et al.* PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. *Nucleic Acids Res* **40**, D261-270, doi:10.1093/nar/gkr1122 (2012).

16 Dinkel, H. *et al.* Phospho.ELM: a database of phosphorylation sites–update 2011. *Nucleic Acids Res* **39**, D261-267, doi:10.1093/nar/gkq1104 (2011).

17 Perfetto, L. *et al.* SIGNOR: a database of causal relationships between biological entities. *Nucleic Acids Res* **44**, D548-554, doi:10.1093/nar/gkv1048 (2016).

18 Zyla, J., Marczyk, M., Weiner, J. & Polanska, J. Ranking metrics in gene set enrichment analysis: do they matter? *BMC Bioinformatics* **18**, 256, doi:10.1186/s12859-017-1674-0 (2017).