Supplementary Materials and methods

Data download and Standardization

The expression profile data of pancreatic cancers (IlluminaHiSeq_RNASeq platform) were downloaded from TCGA. In total, 183 samples with corresponding clinical information were downloaded, along with data from multi-omics, including copy number and methylation data (Supplementary Table 1).

Level 3 data from TCGA were standardized from the read count. As the expression level of different genes varied greatly, transcriptional gene expression profiles were preprocessed so that they can be evaluated effectively. First the data were transformed by taking their logarithm, and if the original value was 0, it was replaced with the mean value. Then for each gene, the mean and standard deviation were calculated, and values outside 95% confidence interval were replaced by the boundary value. Last the expression profile matrix were normalized with z-score, so that the expression value of each gene could obey the standard normal distribution [1].

Unsupervised hierarchical clustering

Hierarchical clustering was used for the analysis of the above standardized transcription data. Related pancreatic cancer genes were downloaded from the OMIM database (http://www.omim.org) [2], including 132 Entrez gene IDs, gene name, MIM number, and chromosome location (Supplementary file 1); Entrez gene IDs were converted to gene symbols. All pancreatic cancer samples were analyzed by unsupervised hierarchical clustering[3] with Euclidean distance matrices [4] and the average linkage method [5].

Subgroup clinical characteristics

Clinical information from different subgroup samples were statistical analyzed, including age, sex, smoking and drinking history, and survival time (Supplementary file 2).

Gene distribution in specific subgroups

Through hierarchical clustering, samples with similar molecular basis were clustered into one group, and candidate genes were assigned to each subgroup. To determine whether a specific gene belonged to a particular subgroup, a Student's t-test was used to determine the significance of its expression among subgroups [6]. A gene was assigned to one subgroup if its expression in the subgroup was significantly different from other subgroups ($P < 0.05$). In the end, each subgroup was assigned a specific set of genes (Supplementary file 3, 4).

Identification of specific pathways and genes in subgroups

Specific gene sets were identified by comparing individual genes among different cancer subgroups (Supplementary file 5). Genes within specific gene sets showed different expression patterns in different subgroups (e.g., highly expressed in some subgroups and lowly expressed in others), suggesting that these gene sets were important for distinguishing pancreatic cancer subgroups at the molecular levels. Specific genes for each subtype were unioned, and KEGG pathway enrichment analysis was performed using DAVID [7]; pathways with a P-value of <0.05 were considered statistically significant.

Pathway deviation score

Specific gene set for different subgroups were identified that had different expression patterns in different subgroups, thereby serving a variety of functions at different levels in different subgroups. At different functional levels, these pathways were important for the analysis and individualized treatment of different clinical subgroups of pancreatic cancer patients. These pathways were quantified by gene enrichment according to Formula 1:

$$score(P) = \frac{\sqrt{\sum_{i=1}^{n}(Gi-mean)^2}}{n} \qquad \text{(Formula 1)}$$

wherein pathway *P* is assumed to contain *n* number of enriched genes and *mean* represents the expression mean of a given gene (*Gi*) in all samples. The larger the score(*P*), the more pathway *P* deviated from the normal level and vice versa.

By calculating the Euclidean distance of all genes in pathway P in each subgroup and summing them, the extent by which pathway P deviated from normal in the subgroup was calculated [8]. Finally, by comparing the degree of pathway deviation among different subgroups, functional pathways with specific alterations and genes involved were identified.

Diagnosis and prediction model based on specific pathways
A SVM model was trained to predict the subtype of pancreatic cancer samplers using functional pathways as features and the eigenvector score (Formula 2) as their values. The eigenvector score showed whether there were significant differences in the function of different pancreatic cancer subgroups. The eigenvector score was defined as

$$EVscore = log_2 \frac{\sqrt{\sum_{i=1}^{m}(Gi-\mu)^2}}{\sqrt{\sum_{j=1}^{n}(Gj-\gamma)^2}} \quad \text{(Formula 2)}$$

where *EVscore* is the eigenvector score, *m* is the number of upregulated genes in different pathways, *Gi* is the mean expression value of up-regulated gene *i* in all samples, μ is the mean expression of gene i in control samples, *n* is the number of genes downregulated, *Gj* is the mean expression value of down-regulated gene j in all samples, and γ is the mean expression value of gene *j* in control samples.
Finally, ratios of the distances from normal levels to up-regulated/down-regulated levels were calculated for logarithm; a positive EVscore was considered a positive fluctuation and a negative EVscore was considered a negative fluctuation. A 10 fold cross validation was used to validate the model. All samples were randomly rearranged into 10 parts; nine were used as a training set to train the model and obtain threshold parameters, leaving one for the test set. The trained model was used to predict in the test set, and false-positive/negative rates and prediction accuracy were calculated. The above procedure was repeated 10 times until all samples were predicted in a test set. The receiver operating characteristic curve was then used to evaluate the classification efficiency and robustness of the model.

Multi-omics data analysis
Copy number analysis and methylation data were integrated to analyze genetic variations specific to each pancreatic cancer subgroup from multiple omics levels. Specific genes from all subgroups were integrated, including their variability at regulatory and methylation levels and copy number variation, to determine the reason for abnormal transcription levels in different subgroups.

MicroRNA (miRNA)-long noncoding RNA (lncRNA)-mRNA co-expression analysis
mRNA alone may not explain the differences in pancreatic cancer pathologies at the molecular level. Therefore, mRNA, miRNA, and lncRNA expression profiles were integrated for further analysis. miRNA expression profile data, including 183 pancreatic cancer samples, 1046 miRNA molecules, and lncRNA data containing 178 pancreatic cancer samples and 2441 lncRNA molecules, were gathered from TCGA.
A composite regulatory network was constructed with coexpression analysis. First molecules expressed significantly differently ($P < 0.05$) in at least two subgroups were identified from the expression profile, and coexpression analysis was performed [9]. Co-expression analysis included the following relationships: mRNA-mRNA, miRNA-mRNA, lncRNA-mRNA, and lncRNA-miRNA. These coexpression relationships reflected the interaction among mRNAs, the regulation of mRNA by miRNA/lncRNA, and other regulators. The correlation between any relationship pairs was evaluated by the Pearson correlation coefficient. A composite regulatory network was constructed with the coexpression correlation, with mRNA, miRNA, and lncRNA as nodes, with edges indicating their correlations [4, 10]. The correlation coefficient threshold was 0.5, i.e., there was an edge between two nodes if and only if the correlation coefficient was greater than 0.5.

Supplementary Table 1. Pancreatic cancer data downloaded from The Cancer Genome Atlas (TCGA) database.

Data	Expression profile	CNV copy number	Methylation profile
Platform	IlluminaHiSeq_RNASeqV2	segmented_scna_minus_germline_cnv	PAAD.Methylation_Preprocess
Number of samples	183	183	183
Data level	3	3	3

Supplementary Table 2. Functional annotation analysis.

Term	Count	*P*-Value
Aminoacyl-tRNA biosynthesis	8	2.34E-09
Antigen processing and presentation	8	2.19E-08
Cell adhesion molecules	12	7.97E-08
Cytokine-cytokine receptor interaction	6	8.81E-06
Focal adhesion	6	3.62E-05
Glutamate metabolism	5	3.79E-05
Glycosphingolipid biosynthesis–ganglio-series	5	1.03E-04
GnRH signaling pathway	5	2.54E-04
Hematopoietic cell lineage	4	3.12E-04
Jak-STAT signaling pathway	5	3.17E-04
Leukocyte transendothelial migration	5	3.71E-04
Natural killer cell-mediated cytotoxicity	5	7.13E-04
Nonhomologous end joining	6	2.84E-03
Pathogenic *Escherichia coli* infection-EHEC	5	1.35E-02
Purine metabolism	4	1.97E-02
Actin cytoskeletal regulation	3	3.00E-02

Supplementary Table 3. Pathway deviation score.

Pathway	Subgroup 1	Subgroup 2	Subgroup3
Aminoacyl-tRNA biosynthesis	0.3	1.44	1.02
Antigen processing and presentation	0.87	0.67	0.52
Cell adhesion molecules	0.76	0.57	0.49
Cytokine-cytokine receptor interaction	0.77	0.69	0.56
Focal adhesion	0.73	1.15	1.22
Glutamate metabolism	0.3	1.44	1.02
Glycosphingolipid biosynthesis–ganglio-series	0.94	1.44	1.37

GnRH signaling pathway	1.12	1.14	1.11
Hematopoietic cell lineage	1.07	0.62	0.97
Jak-STAT signaling pathway	1.11	1.12	0.87
Leukocyte transendothelial migration	0.77	0.67	0.59
Natural killer cell-mediated cytotoxicity	0.84	0.7	0.57
Nonhomologous end joining	1.26	1.24	1.31
Pathogenic *Escherichia coli* infection-EHEC	1.13	0.58	0.93
Purine metabolism	1.26	1.01	1.04
Actin cytoskeletal regulation	0.88	0.6	0.83

Supplementary Table 4. Analysis of copy number variation.

Gene	Subgroup 1	Subgroup 2	Subgroup 3
STARD13	0.035714	−0.0125	0.142857
EEF1A1	−0.19048	−0.575	0.285714
ANO7	0.083333	0.05	0.071429
FAM84B	0.416667	0.6	−0.07143
CASC3	0.202381	−0.025	0.142857
PBOV1	−0.27381	−0.575	0
ADH1B	−0.15476	−0.0625	−0.07143
CASC5	−0.09524	−0.15	−0.21429
BLCAP	0.154762	0.3	0.214286
STEAP2	0.25	0.4	0.357143
TFDP3	0.059524	−0.0875	0
NDC80	−0.14286	−0.0125	−0.28571
BCAS1	0.190476	0.3375	0.214286

Supplementary Table 5. Analysis of the methylation spectrum.

Gene	Subgroup 1	Subgroup 2	Subgroup 3

ANO7	0.740217	1.005277	0.754314
CASC3	0.822278	1.199015	0.812373
ADH1B	0.818939	1.184669	0.833363
CASC5	0.047211	0.068949	0.050742
BLCAP	0.482866	0.658869	0.482699
BCAS1	0.460157	0.652133	0.455947

Supplementary Table 6. Correlation analysis of copy number and the transcriptome.

Gene	*P*-value	R
STARD13	0.036296	0.639427
EEF1A1	0.412647	2.47E-01
ANO7	0.052716	0.496066
FAM84B	0.017453	0.821803
CASC3	0.055112	0.476666
PBOV1	0.087464	0.258158
ADH1B	0.008338	0.914325
CASC5	0.011056	0.886559
BLCAP	0.119864	0.120599
STEAP2	0.026326	0.734043
TFDP3	0.014922	0.847309
NDC80	0.019118	0.805132
BCAS1	0.089485	0.247274

Supplementary Table 7. Correlation analysis of methylation and the transcriptome.

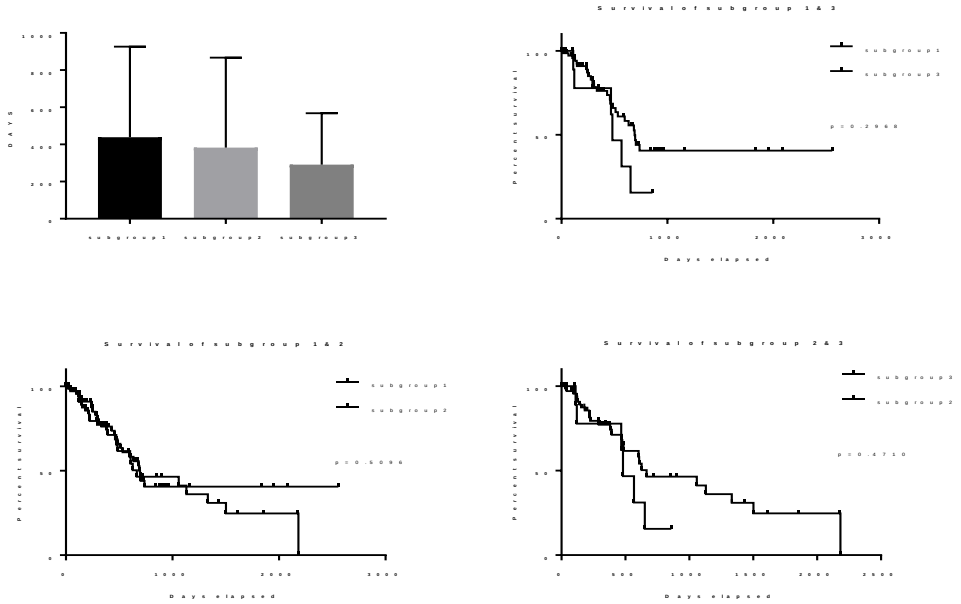
gene	*P*-value	R
ANO7	0.06599593	−0.6333894
CASC3	0.09286334	0.02622384
ADH1B	0.07561739	−0.547573
CASC5	0.01819148	−0.3235922

BLCAP	0.09636655	−0.205003
BCAS1	0.03743212	−0.7420775

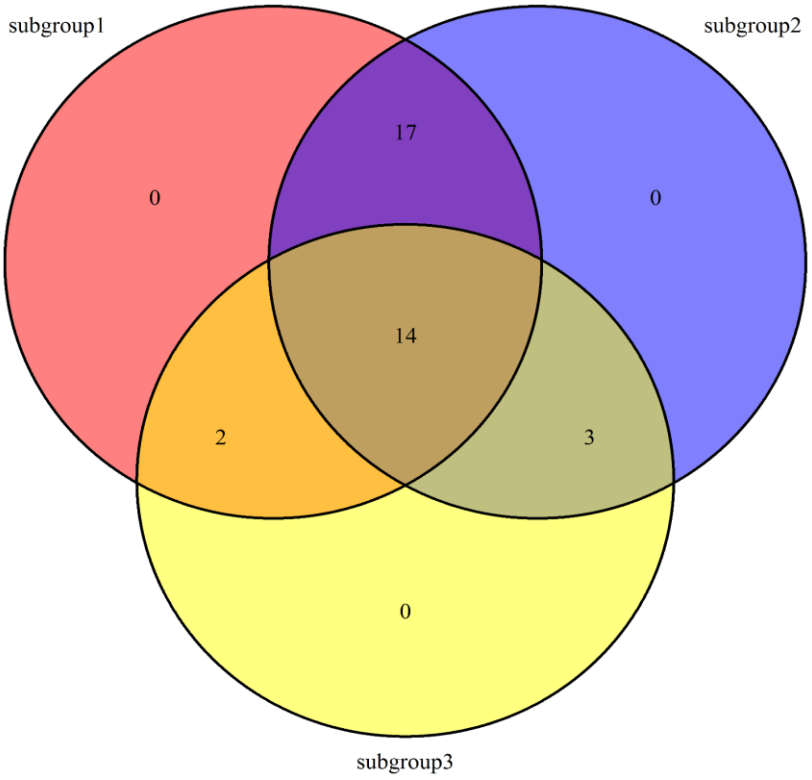
Supplementary Table 8. Topology analysis of the top 10 highest degrees of mRNA.

Name	Degree	ASPL	CC	NC	TC
SYT7	126	3.101078	0.511746	81.46825	0.305035
ADH1A	67	3.684636	0.048847	9.164179	0.056897
DLEC1	54	3.719677	0.410901	51.2963	0.260582
TUNAR	50	3.669811	0.914286	99.96	0.51
TUSC3	42	3.440701	0.498258	63.14286	0.282511
MIR7-3HG	38	3.58221	0.786629	93.76316	0.413054
TFDP3	38	3.97035	0.998578	98.10526	0.645429
MEG3	35	3.634771	0.678992	84.51429	0.34637
VWA5A	35	3.834232	0.904202	76.28571	0.412355
DPH1	29	3.52965	0.768473	74.31034	0.347245
ave	5.79	4.64	0.32	31.49	0.23

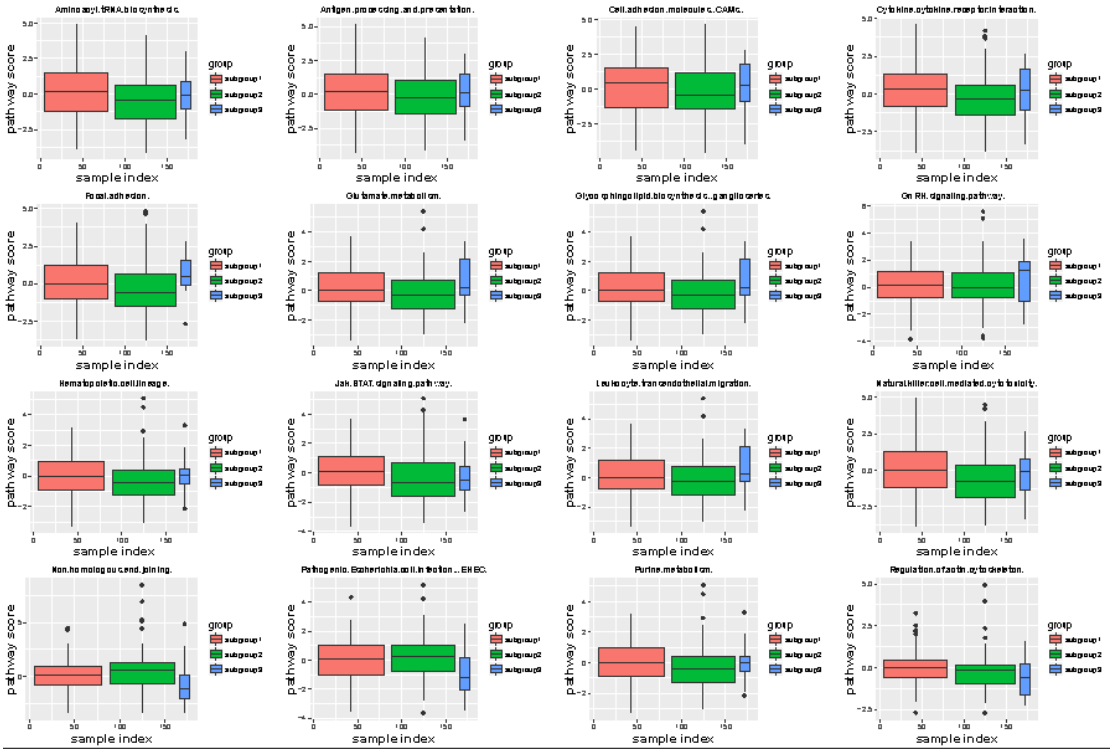
Degree, node degree; ASPL, average shortest path length; CC, clustering coefficient; NC, neighborhood connectivity; TC, topological coefficient; ave, the mean of each topological property.



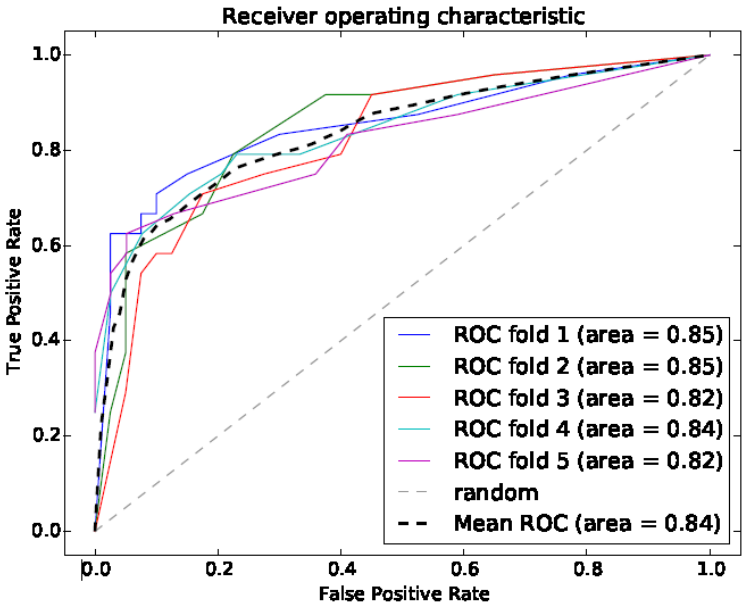
Supplementary Figure 1. Survival analysis. There was no statistically significant difference in survival time among the three subgroups.



Supplementary Figure 2. Venn diagram of the three pancreatic cancer subgroups. The number of specific genes shared by subgroups 1 and 2 was 31, by subgroups 2 and 3 was 17, and by subgroups 1 and 3 was 16; there were 14 overlapping genes in all three subgroups.



Supplementary Figure 3. Boxplot of pathway deviation scores. Overall, these data show that there were functional differences among the three subgroups.



Supplementary Figure 4. Receiver operating characteristic (ROC) curve. Five predicted results were randomly selected for statistics in the ROC curve. The average accuracy was 0.84. These data confirm that the 16 functional pathways identified could effectively differentiate the pancreatic cancer subgroups and that the model had high precision and robustness.

1. Foo LC, Mafauzy M: **Does the use of mean or median Z-score of the thyroid volume indices provide a more precise description of the iodine deficiency disorder status of a population?** *Eur J Endocrinol* 1999, **141**(6):557-560.

2. Martelli PL, Fariselli P, Savojardo C, Babbi G, Aggazio F, Casadio R: **Large scale analysis of protein stability in OMIM disease related human protein variants**. *BMC Genomics* 2016, **17 Suppl 2**:397.
3. Taamneh M, Taamneh S, Alkheder S: **Clustering-based classification of road traffic accidents using hierarchical clustering and artificial neural networks**. *Int J Inj Contr Saf Promot* 2016:1-8.
4. Ghosh A, Barman S: **Application of Euclidean distance measurement and principal component analysis for gene identification**. *Gene* 2016, **583**(2):112-120.
5. Liu CH, Li M, Feng YQ, Hu YJ, Yu BY, Qi J: **Determination of Ruscogenin in Ophiopogonis Radix by High-performance Liquid Chromatography-evaporative Light Scattering Detector Coupled with Hierarchical Clustering Analysis**. *Pharmacogn Mag* 2016, **12**(45):13-20.
6. Lu K: **Distribution of the two-sample t-test statistic following blinded sample size re-estimation**. *Pharm Stat* 2016, **15**(3):208-215.
7. Goncalves L, Filipe M, Marques S, Salgueiro AM, Becker JD, Belo JA: **Identification and functional analysis of novel genes expressed in the Anterior Visceral Endoderm**. *Int J Dev Biol* 2011, **55**(3):281-295.
8. Chen X, Liu L, Wang Y, Liu B, Zeng D, Jin Q, Li M, Zhang D, Liu Q, Xie H: **Identification of breast cancer recurrence risk factors based on functional pathways in tumor and normal tissues**. *Oncotarget* 2017, **8**(13):20679-20694.
9. Mori K, Haraguchi S, Hiori M, Shimada J, Ohmori Y: **Tumor-associated macrophages in oral premalignant lesions coexpress CD163 and STAT1 in a Th1-dominated microenvironment**. *BMC Cancer* 2015, **15**:573.
10. Soul J, Dunn SL, Hardingham TE, Boot-Handford RP, Schwartz JM: **PhenomeScape: a cytoscape app to identify differentially regulated sub-networks using known disease associations**. *Bioinformatics* 2016, **32**(24):3847-3849.