Supplementary

1. Comparisons of different settings of *v*.

2. MI based forward searching process for *hyper-gene* selection.

3. The evaluation framework based on classification

4. Prostate cancer related GO annotations.

5. Prostate cancer related KEGG pathways.

1. Comparisons of different settings of *v*

Given a gene (say, g_k) and a gene functional category (say c_i), we estimate $p(g_k|c_i)$ according to $p(g_k \mid c_i) \propto (1 + s(g_k, c_i))^v$.

$s(g_k, c_i)$ is the similarity between g_k and c_i, and is determined by $s(g_k, c_i) = \arg\max_{g_j \in c_i}(s_{kj})$ where s_{kj} is the association between g_k and g_j and is measured using Pearson's correlation. In our investigation, different settings of v were tested on the synthetic and real data. In table S1 and S2, the comparative results are presented.

Table S1. The results on a synthetic example. A percentage value indicates the probability of the corresponding category being selected across 1000 trials.

	(a) The results on the category setting shown in Fig. 2
$v = 2$	$\{S_2, S_5\}$: 80.1%; $\{S_3, S_5\}$: 2.3%; $\{S_4, S_5\}$: 0.4%.
$v = 6$	$\{S_2, S_5\}$: 89.1%; $\{S_3, S_5\}$: 1.2%; $\{S_4, S_5\}$: 0.
$v = 10$	$\{S_2, S_5\}$: 37.5%; $\{S_3, S_5\}$: 0; $\{S_4, S_5\}$: 0.
$v = 18$	$\{S_2, S_5\}$: 20.7%; $\{S_3, S_5\}$: 0; $\{S_4, S_5\}$: 0.
	(b) The results on the category setting shown in Fig. 2
$v = 2$	$\{S_2, S_5\}$: 72%; $\{S_3, S_5\}$: 1.2%; $\{S_4, S_5\}$: 0.3%.
$v = 6$	$\{S_2, S_5\}$: 80.1%; $\{S_3, S_5\}$: 2.1%; $\{S_4, S_5\}$: 0.4%.
$v = 10$	$\{S_2, S_5\}$: 45.3%; $\{S_3, S_5\}$: 0.5%; $\{S_4, S_5\}$: 2.1%.
$v = 18$	$\{S_2, S_5\}$: 8.1%; $\{S_3, S_5\}$: 0; $\{S_4, S_5\}$: 0.

2. MI based forward searching process for *hyper-gene* selection.

This process can be stated as follows.

Step 1. Set the selected gene set, say *S*, with empty.

Step 2. For each hyper-gene, say *g*, compute MI(*g*,*y*) where y is the response.

Step 3. Find the hyper-gene having the maximal MI(*g*,*y*), and place it into *S*.

Step 4. Repeat the following until certain hyper-genes have been selected.

(a) For each unselected hyper-gene, say g, calculate MI(*S*+*g*, *y*).

(b) Identify the one with the maximal MI(*S*+*g*, *y*), and place that hyper-gene into S.

Step 5. Output S.

In the above process,

$$MI(S,y) = \log\left(\sum_y \int p_S(y,x)^2\,dx\right) + \log\left(\sum_y P(y)^2\right)\left(\int p_S(x)^2\,dx\right) - 2\log\left(\sum_y \int p_S(y,x)P(y)p_S(x)dx\right)$$

where

$$P(y) = \frac{|\text{class } y|}{N},$$

$$p_S(x) = \frac{1}{N}\sum_{x_i \in X} p_S(x \mid x_i) = \frac{1}{N}\sum_{x_i \in X} \kappa(x - x_i, h),$$

$$p_S(x) = \frac{1}{N}\sum_{x_i \in \text{class } y} p_S(x \mid x_i) = \frac{1}{N}\sum_{x_i \in \text{class } y} \kappa(x - x_i, h).$$

Furthermore, we have

$$\kappa(x - x_i, h) = G(x - x_i, h) = \frac{1}{(2\pi h^2)^{M/2}} \exp\left(-\frac{1}{2h^2}(x - x_0)(x - x_0)^T\right),$$

$$h = \left\{\frac{4}{(M+2)}\right\}^{1/(M+4)} N^{-1/(M+4)}.$$

X is a given dataset. M and N are the dimensionality and size of X.

3. The evaluation framework based on classification

We evaluated the quality of a selected category set based on its classification capability. In this course, the framework of 5 cross validation was adopted. In Fig S1, this framework is detailed.

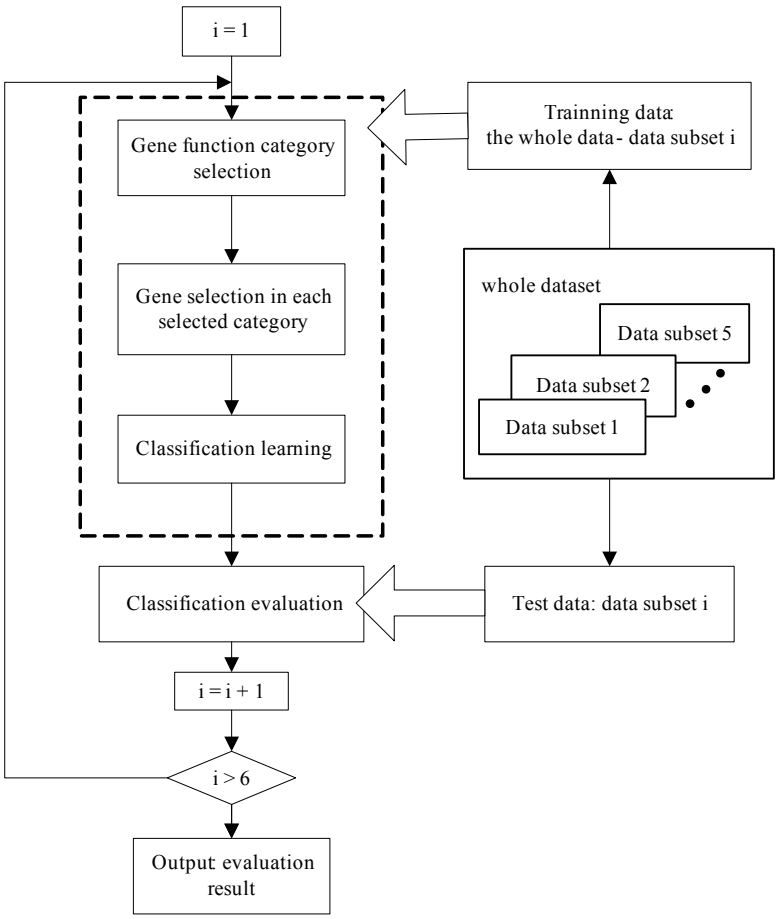

Fig S1 the block diagram of 5CV framework for classification evaluation

4. The prostate-cancer-related GO biological terms

Based on the prostate-cancer-related gene list (http://www.superarray.com/gene_array_product/HTML/OHS-403.html), we identified 58 prostate-cancer-related biological process GO terms.

GO:0016049:cell growth
GO:0048732:gland development
GO:0045893:positive regulation of transcription, DNA-dependent
GO:0030521:androgen receptor signaling pathway
GO:0030518:steroid hormone receptor signaling pathway
GO:0016567:protein ubiquitination
GO:0008637:apoptotic mitochondrial changes
GO:0045884:regulation of survival gene product activity
GO:0007281:germ cell development
GO:0008634:negative regulation of survival gene product activity
GO:0008624:induction of apoptosis by extracellular signals
GO:0006916:anti-apoptosis
GO:0008629:induction of apoptosis by intracellular signals
GO:0000079:regulation of cyclin dependent protein kinase activity
GO:0007050:cell cycle arrest
GO:0001501:skeletal development
GO:0006935:chemotaxis
GO:0006007:glucose catabolism
GO:0007169:transmembrane receptor protein tyrosine kinase signaling pathway
GO:0008286:insulin receptor signaling pathway
GO:0051262:protein tetramerization
GO:0008284:positive regulation of cell proliferation
GO:0016064:humoral defense mechanism (sensu Vertebrata)
GO:0045793:positive regulation of cell size
GO:0001558:regulation of cell growth
GO:0019735:antimicrobial humoral response (sensu Vertebrata)
GO:0045927:positive regulation of growth
GO:0008015:circulation
GO:0008630:DNA damage response, signal transduction resulting in induction of apoptosis
GO:0030330:DNA damage response, signal transduction by p53 class mediator
GO:0045892:negative regulation of transcription, DNA-dependent
GO:0000122:negative regulation of transcription from RNA polymerase II promoter
GO:0000080:G1 phase of mitotic cell cycle
GO:0030308:negative regulation of cell growth
GO:0045792:negative regulation of cell size
GO:0045926:negative regulation of growth
GO:0000082:G1/S transition of mitotic cell cycle
GO:0007265:Ras protein signal transduction
GO:0016044:membrane organization and biogenesis
GO:0016337:cell-cell adhesion
GO:0050678:regulation of epithelial cell proliferation
GO:0030334:regulation of cell migration
GO:0030520:estrogen receptor signaling pathway
GO:0007507:heart development
GO:0007160:cell-matrix adhesion
GO:0007417:central nervous system development
GO:0051301:cell division
GO:0001525:angiogenesis
GO:0008544:epidermis development
GO:0006096:glycolysis
GO:0008016:regulation of heart contraction
GO:0007605:sensory perception of sound
GO:0050954:sensory perception of mechanical stimulus
GO:0045765:regulation of angiogenesis
GO:0006817:phosphate transport
GO:0015698:inorganic anion transport
GO:0007156:homophilic cell adhesion
GO:0007254:JNK cascade

5. The prostate-cancer-related KEGG pathways

Similar to the above section, we identified 13 prostate-cancer-related KEGG pathways.

path:hsa01510 Neurodegenerative Disorders;
path:hsa05030 Amyotrophic lateral sclerosis (ALS);
path:hsa04210 Apoptosis;
path:hsa04010 MAPK signaling pathway;
path:hsa04110 Cell cycle;
path:hsa04510 Focal adhesion;
path:hsa04320 Dorso-ventral axis formation;
path:hsa04810 Regulation of actin cytoskeleton;
path:hsa04310 Wnt signaling pathway;
path:hsa04530 Tight junction;
path:hsa04620 Toll-like receptor signaling pathway;
path:hsa04920 Adipocytokine signaling pathway;
path:hsa04350 TGF-beta signaling pathway.