高雄醫學大學
Kaohsiung Medical University

高階人才研習營

次世代定序及生物資訊分析研習營/課程編號D5

高教深耕計劃－構面二＆核心議題三
發展學校特色＆培育特色研發人才系列活動

Date	Time	Topic	Speaker	Place
July 10 2019	09:00~09:30	Opening	Vice-president Hung Director Hsu	IR334
	09:30~10:30	Sequencing technologies (lecture)	Dr. H. Sunny Sun	
	10:30~12:30	**Exome sequencing data analysis:** Data cleaning, mapping, variant calling and annotation (practical course)	Dr. Alan Tang	
	12:30~14:00	Lunch		
	14:00~16:00	**Exome sequencing data analysis:** structural variant identification and analysis (practical course)	Dr. Shuen-Lin Jeng	IR334
	16:00~16:30	Questions and Answers	Drs. H. Sunny Sun Dr. Alan Tang Dr. Shuen-Lin Jeng	

Date	Time	Topic	Speaker	Place
July 11 2019	09:00~09:30	Opening	Vice-president Hung Director Hsu	Audiovisual center
	09:30~10:00	**High-throughput genomic data analysis:** challenges and opportunities	Dr. I-Shou Chang	
	10:00~10:30	Study design and technical issues on RNA sequencing	Dr. Shih Sheng Jiang	
	10:30~11:20	Preprocessing for high throughput genomic data analysis	Dr. Chung-Hsing Chen	
	11:20~12:00	What after RNA-seq?	Dr. Shih Sheng Jiang	
	12:00~14:00	Lunch		
	14:00~14:30	RNA sequencing data analysis (demonstration)	Ms. Fang-Yu Tsai	IR334
	14:30~16:00	Pathway analysis for transcriptome profiling studies(array and RNA-seq) using GSEA (hands-on practice)	Ms. Fang-Yu Tsai Dr. Chung-Hsing Chen Dr. Shih Sheng Jiang	
	16:00~16:30	Questions and Answers	Drs. I-Shou Chang Dr. Shih Sheng Jiang Dr. Chung-Hsing Chen Ms. Fang-Yu Tsai	

主辦單位：高雄醫學大學研究資源整合發展中心
協辦單位：科技部生技醫藥生物資訊核心設施、衛福部第三期癌症研究跨機構合作平台及其整合應用計畫

Bioinformatics for some cancer research

I-Shou Chang (張憶壽)
National Health Research Institutes
July 11, 2019

Cancer research

- Basic, clinical, epidemiology, translational
- Mechanism study, association study, causal inference
- Reproducibility
- Clinical trials, observational studies

Quantitative methods in cancer research

- Epidemiology, clinical trials; biostatistics; **study design**, inference, prediction
- Bioinformatics, data from high-throughput experiments
- Genomic epidemiology
- Facilitate communication or collaboration

Issues with high-throughput data

- High-throughput data : Gene expression, SNP, CNV, DNA methylation, microRNA, long non-coding RNA, etc. (molecular)
- High-throughput experiments : Array technologies, sequencing technologies
- Data preprocessing
- Data quality opportunity: sample quality, SNP quality, **batch effects**, etc.
- Multiple-testing : p-value (Bonferroni correction), false discovery rate.

Three parts for this lecture

- Common study designs in cancer studies (Statistics for epidemiology, N. P. Jewell)
- Genome-wide association study (GWAS)
- Integrative analysis

Epidemiology/biostatistics

- Study of traits or diseases in **populations**
- Priority setting for health sciences research and care
- Prediction, biology, diagnosis, drug target///
- Incidence, prevalence, mortality, survival; surveillance
- Risk factors, confounder; lifestyle/environmental, genetic
- Effect, odds ratio, hazard ratio
- Design; population study, cohort study, case-control study

Epidemiology/biostatistics

- Explanation or prediction
- Population parameters
- Fitting the data using regression models
- Bayesian or Frequentist approach
- Hypothesis testing, estimation, model check
- Sample size, statistical power
- Number of variables

Basic study designs

- Study the association between a disease and risk factors in a population.
- Both risk factor and disease have **two** states, to simplify the discussion.
- **Target population**, **Study population**
- Three basic study designs: **population-based studies, cohort studies, case-control studies**.

Population-based studies

- Take a simple random sample from the study population
- **Subsequently**, measure the exposure (E) and disease (D) status for each individual in the randomly selected samples.

Population-based studies

- Example 1: HALST (Healthy aging longitudinal study in Taiwan).
- Example 2: Genetic Modifiers of Progression-Free Survival in Never-Smoking Lung Adenocarcinoma Patients Treated with First-Line Tyrosine Kinase Inhibitors.(AJRCCM, 2017, Chang, Jiang, Yang, et al.)

Cohort studies

- Exposure-based sampling
- Identify two subgroups of the population on the basis of presence or absence of the exposure.
- Take a simple random sample from each of these two groups.
- Measure the disease status for each of the individuals in these two random samples.
- Clinical trials

Mok et al., *N Engl J Med* 2009

- Gefitinib vs carboplatin-paclitaxel in Lung Adenocarcinoma

Study subjects

Advanced lung adenocarcinoma

Nonsmokers or former light smokers

Untreated

2 arms

Gefitinib 250 mg per day

Carboplatin + Paclitaxel

Primary end point

Progression-free survival

Case-control studies

- Disease-based sampling
- Identify two subgroups of the population on the basis of the presence or absence of the disease.
- Take a simple random sample from each of these two subgroups.
- **Subsequently**, measure the presence or absence of the exposure for each individual in both random samples.

April 29, 2004 NEJM

- Activating mutations in *EGFR* underlying responsiveness of NSCLC to Gefitinib
- **Mutations in *EGFR* were found in 8/9 vs 0/7, in gefitinib responsive vs nonresponsive lung cancer.**
- Mutations found in 2/25 not exposed in gefitinib.
- All mutations are heterozygous and identical mutation appeared in multiple cancers.

Sept. 2004 PNAS

- *EGFR* gene mutations are common in lung cancers from never smokers and are associated with sensitivity of tumors to gefitinib and erlotinib.
- Mutations found in 7/10 vs 0/8, according to gefitinib sensitivity.
- Mutations found in 5/7 vs 0/10, according to erlotinib sensitivity.
- Mutations found in 7/15 vs 4/81, according to smoking status.

15

Case-control studies: rare disease

$$\frac{P(D \mid E)}{P(D \mid \overline{E})} \times \frac{P(D \mid E) / P(\overline{D} \mid E)}{P(D \mid \overline{E}) / P(\overline{D} \mid \overline{E})} = \frac{P(E \mid D) / P(\overline{E} \mid D)}{P(E \mid \overline{D}) / P(\overline{E} \mid \overline{D})}$$

Odds, odds ratio, D disease status, E exposure status

2018/7/11

16

Many variants of case-control designs

- Suppose we have follow-up data of 500 lung cancer patients. We would like to know if any germline SNPs are associated with the time to cancer progression. Suppose whole genome sequencing of one sample costs $2000. It's too costly to include all the patients for study.
- Compare those with short PFS (progression free survival) with those with long PFS.
- **Nested case-control design**
- **How cases and controls are selected.**

Nested case-control design

- Include all the patients who experienced tumor progression.
- Order all these patients according to his/her time to progression. Call them $T_1 \leq T_2 \leq \ldots.$
- At the progression time of each patient, select one or more patients having not experienced progression at the time but having experienced the same amount of follow-up time. Matching may be considered.

Case-control design, an example

- GELAC (genetic epidemiology study of lung adenocarcinoma in Taiwan), Hsiung et al. PLOS Genetics 2010.
- Incident lung ADC patients were recruited from 6 hospitals in Taiwan from 2002 to 2012. For each patient, one healthy control with the same sex and age was recruited from the health examination center of the same hospital.
- Nested case-control design, with different consideration. Though.

Analysis of data from case-control design

- Assuming a logistic regression model in the population. Data from case-control design can be used to estimate odds ratio in the population without bias, using **unconditional likelihood** approach. (Prentice and Pyke, 1978)
- Case-control study with a matching variable, which assumes many levels, a **conditional likelihood** approach may be used.

Genome-wide association studies (GWAS)

- Looking for genetic variants associated with traits or diseases.
- Genotyping by array technologies or sequencing technologies
- 2005

GWAS (**data preprocess**)

- Reproducibility is high in GWAS.
- Standardized analysis procedures (nearly)
- Illumina SNP array 610K reduces to 450K for analysis. Call rate, population, Linkage disequilibrium, Sex check. Relatedness between samples. Contaminated samples. PCA (principal component analysis) for population substructure adjustment.
- Bonferroni correction (0.05/500,000)
- Replication

GWAS

- Thousands of strong associations between loci and traits have been reported (p-value less than 5E-8).
- Reproducible.
- Polygenic for most traits.
- Pleiotropy is common.
- Common variants contribute to non-trivial heritability.
- Risk prediction.

Post-GWAS

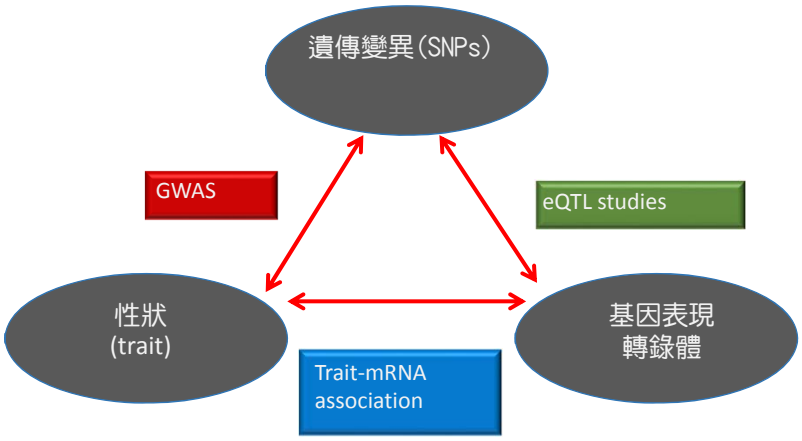
- Missing heritability compared with family studies.
- Over 90% of the variants are in non-coding regions.
- Need to know the path from the SNPs to the traits.
- What the intermediate effects of these variants are that eventually lead to disease. the mechanisms?
- *FTO* and obesity, *MHC* and Schizophrenia
- eQTL (expression quantitative trait loci).
- **Integrative analysis.**
- Am J Hum Genet 2017, Visscher et al.

Never-smoking females with TKI as first-line treatment

- TKI=tyrosine kinase inhibitor, target therapy
- First generation: Iressa, Tarceva.
- Progression free survival
- For *EGFR*-mutated patients, first-line treatment by TKI results in better PFS, compared with first-line treatment by standard chemotherapy. (Four phase 3 clinical trials, 2010-2011)
- Question: Are any germline SNPs associated with PFS?

Integrative analysis

- Look for SNPs that are associated with time to tumor recurrence. (PFS)
- Look for genes whose expression levels in lung cancers are associated with time to tumor recurrence.
- Are the expression levels of these genes associated with these SNPs?
- Interpretation: gene-set (pathway) enrichment analysis (GSEA)



2018/7/11 27

An example (AJRCCM Chang, Jiang, Yang et al. 2017)

- We found loci in 4q12 associated with progression-free-survival among never-smoking women treated with first-line TKI at genome-scale . Its minor allele is correlated with shorter PPS. (1)
- This locus is also associated with the expression of *EGFR* in lung cancer tissue. Its minor allele is correlated with higher level of *EGFR* expression. (2)
- The expression level of *EGFR* in lung cancer tissue is higher than that in the adjacent normal tissue. (3)
- Higher expression of *EGFR* in lung cancer correlates with shorter time to recurrence. (4)

2018/7/11 28

An example

- *BRAF, PDCL2, SRD5A3*
- *CLOCK, NMU*
- Consistency of the 4 types of association studies increases our confidence in each of them.
- Each type association needs validation based on different datasets/cohorts
- Hypothesis generation
- Biomarker, therapeutic Target

Genetic Effects on gene expression across human tissues (Nature 2017)

- The Genotype-Tissue Expression (GTEx) project, v6p analysis freeze.
- 449 donors, 44 human tissues, 7051 post-moterm samples.
- A median depth of 78 million reads for the RNA-seq samples.
- Genotype at 2.2 million sites and inputed to 12.5 million sites.

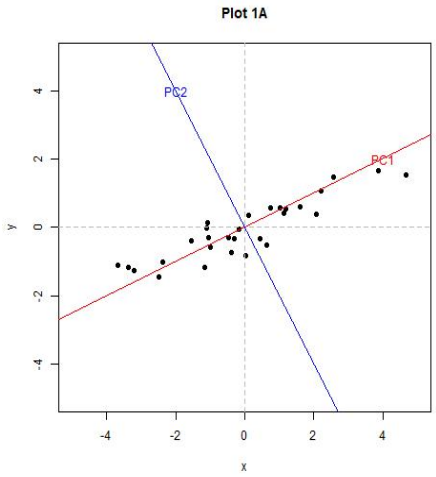
GTXe v6p (**Resources in public domain**)

- Cis-eQTL (expression quantitative trait loci)
- Considering all the tissues, a total of 152,869 cis-eQTL for 19,725 genes. 86.1% of protein-coding genes, 50.3% long intergenic non-coding RNA (lincRNA).
- A median of 2,816 eGenes at 5% FDR within each tissue.
- 41.2% of protein-coding genes having multiple conditionally independent eVariants in at least one tissue.
- Sample size matters.

Principal component analysis

- 76 points in the Euclidean space of dimension 770.
- In expression array, one may consider 100 points in a space of dimension 25K.
- In GWAS, one may consider 2000 points in a space of dimension 500K.
- Imagine data in high-dimension space.

Principal component analysis in 2-dimensional space

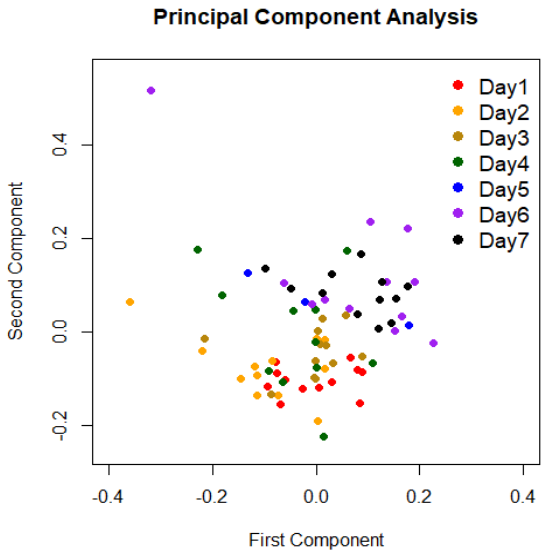


Principal component analysis

- First component is the direction (variable) that shows the largest variability.
- Second component is the direction orthogonal to the first and showing largest variability in the complimentary space.
- Data reduction, interpretation
- Question: are the first few components have large variability?

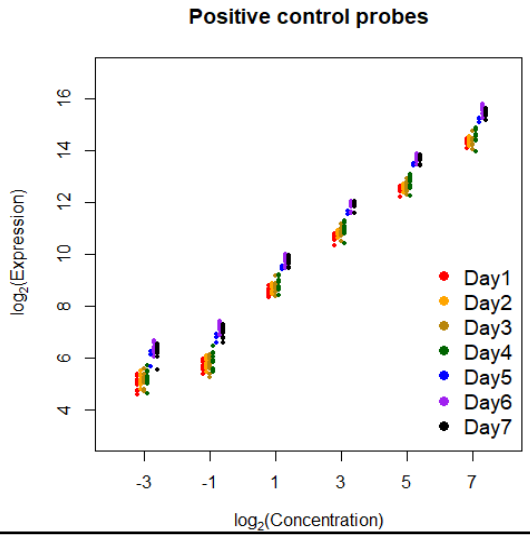
Principal component analysis

- Are they correlated with any of the clinical variables we are interested in?
- If not, treat them as batch effects and adjust for them as confounders in later statistical analysis.
- Previous high-throughput data are of dimension higher than 20K.
- Here it is 770, can we detect batch effect?



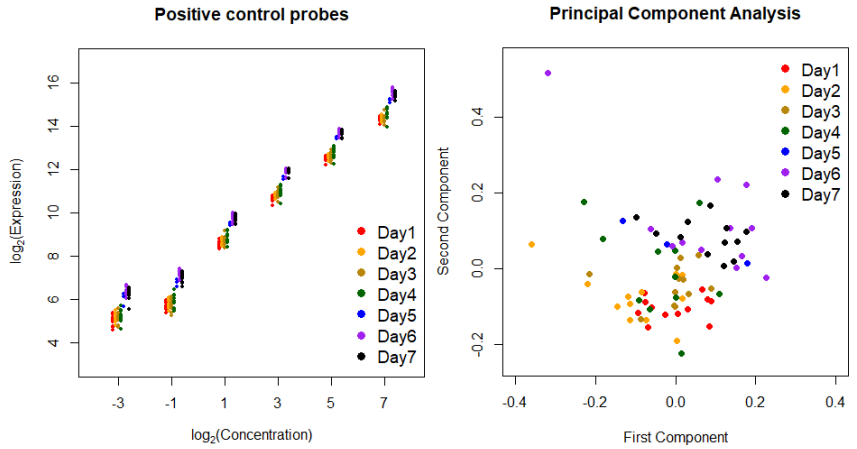
Positive control

- For each positive control, in vitro transcribed RNA targets are pre-mixed with the Reporter CodeSet during manufacturing. Six different transcripts are provided, each at one of the following concentrations in the 30 µL hybridization reaction: 128 fM, 32 fM, 8 fM, 2 fM, 0.5 fM, and 0.125 fM.



Bioinformatics Core Facility TMBD

OSCC (FFPE) Raw data

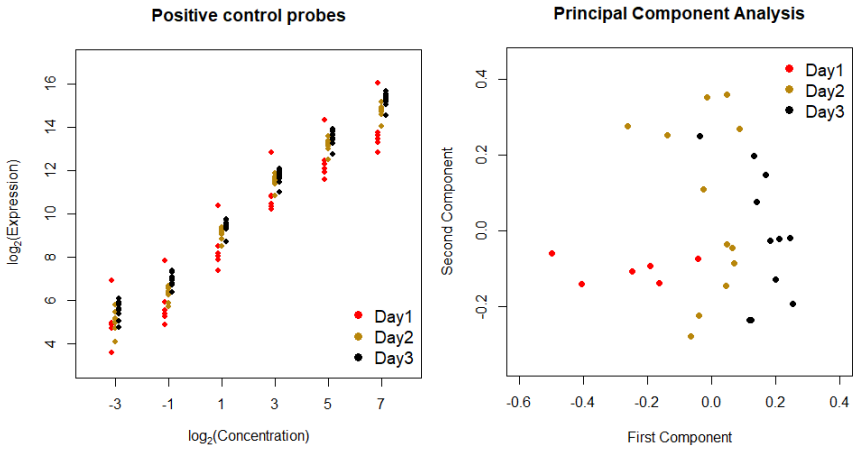


LADC (GSE84797)
PanCancer Immune Profiling Panel

Cartridge	Reaction	Time
1	6	20151016
2	12	20160118
3	11	20160515

Bioinformatics Core Facility TMBD

LADC (FF) Raw data



April 2018

Batch effects by PCA

- PCA can be used to detect batch effects in Nanostring PanCancer Immune Profiling Panel with 770 genes.
- When PCAs are not correlated with any clinical variables we are interested in, they can be treated as confounders. Surrogate variable analysis. (Leek and Storey, PLOS Genetics, 2007)
- PEER (Probabilistic estimation of expression residuals) Nature Protoc. 2012, Stegle et al.,2012

Collaborators

- Bioinformatics Core, supported by MOST, since 2002
- NHRI
- Dr. 江士昇
- Drs. 陳中興, 簡立欣, 陳孜玕
- RA 蔡芳榆

Thank you for your attention

Bioinformatics Core Facility TMBD

Technical and follow-up issues for RNA-sequencing experiment

Workshop on Bioinformatics tools for NGS-data analysis

Shih Sheng Jiang (江士昇), Ph.D.
National Institute of Cancer Research, NHRI

July 2019

Bioinformatics Core Facility TMBD

General issues

- Aims first, **study design** next, and then methods.
- Aims are not feasible when **budget** is a major concern.
- What aims to be addressed in a biological or clinical study?
- Two examples.
- RNA-seq *versus* microarray.
- Why RNA-seq?

July 2019

Bioinformatics Core Facility TMBD

Technical issues

- Type of RNA (mRNA, ncRNA).
- Enrichment of RNA.
- Sources of RNA samples (fresh, freshly frozen, or FFPE).
- Purity/heterogeneity of cell types.
- Splicing: genes or transcripts?
- Genome-wide or targeted?
- Depth of sequencing.

July 2019

Illumina's Suite of RNA Library Prep Solutions

Ribosome Profiling	Whole Transcriptome	mRNA-Seq/ GEx Profiling		Targeted Profiling	miRNA Analysis
 TruSeq Ribo-Profiling	 TruSeq Stranded Total RNA	 TruSeq Stranded mRNA	 TruSeq Stranded RNA Access	 TruSeq Targeted RNA Expression	 TruSeq small RNA
• Sequence RNA fragments that are being actively translated into proteins	• Coding + ncRNA • Transcript-level abundance • Splicing Analysis • Fusion Discovery • FFPE compatible • Any species	• Coding RNA • Transcript-level abundance • Splicing Analysis • Fusion Discovery	• Human Exonic RNA • Transcript-level abundance • Splicing Analysis • Fusion Discovery • FFPE Compatible	• 10s-1,000s of targets • Max 384 spls • Coding + ncRNA • Transcript-level abundance • Fusion Validation • FFPE Compatible	• miRNA abundance • isomiR detection

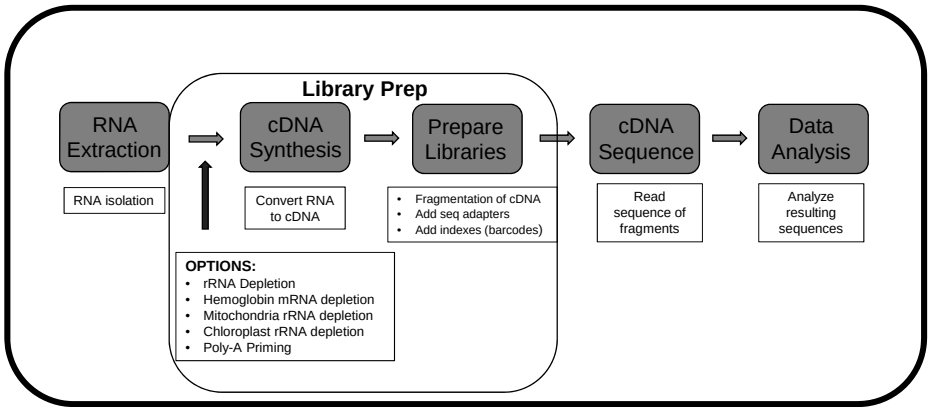
For Research Use Only. Not for use in diagnostic procedures.

Focus on Illumina RNA Kits

	TruSeq Stranded Total RNA	TruSeq Stranded mRNA	TruSeq RNA Access
Assay Time	~8 hours	~9 hours	~2 Days
Hands-on Time	5.5 hours	~5.5 hours	~11 hours
Input Quantity	0.1 – 1 ug high-quality total RNA. Lower-quality samples may require further optimization	0.1 – 1 ug total RNA or 10 - 100 ng previously isolated mRNA (from species with polyA tails)	10 ng total RNA from fresh/frozen samples, or 20 ng total RNA from FFPE samples
Content Specifications	coding RNA (mRNA) + non-coding RNA (no miRNA)	mRNA	mRNA / RNA exome
Mechanism of Action	Bead-based rRNA depletion, cDNA synthesis, and PCR	Oligo-dT beads capture polyA tails	Biotinylated capture probes that target coding RNA. Does not require RNA with poly-A tails.
Multiplexing	LT: Set A – 12; Set B – 12 Single HT: 96 dual	LT: Set A – 12; Set B – 12 Single HT: 96 dual	Set A: 12; Set B – 12 4-plex enrichment
Applications	Expression and splicing analysis; Variants – SNPs, Gene Fusions, Novel Transcripts, Transcript Variants	Expression and splicing analysis; Variants – (SNPs), Gene Fusions, Novel Transcripts, Transcript Variants; Gene expression profiling	Transcript-level abundance; Splicing Analysis; Fusion Discovery; Variants - SNPs, Novel Transcripts, Transcript Variants
Specialized Sample Types	FFPE	High quality only	FFPE, Low Input
Method	Whole-Transcriptome Sequencing	mRNA Sequencing	mRNA Sequencing, Exome Sequencing
Species	Mouse, Human, Rat	Other, Mammalian, Bovine, Mouse, Human, Rat	Human
Automation Capability	Liquid Handling Robots	Liquid Handling Robots	Liquid Handling Robots

For Research Use Only. Not for use in diagnostic procedures.

RNA Sequencing Workflow



Sequencing Depth

- The figures below are general estimates. Please use only as a reference.
- Check Encode guidelines

Differential Expression	**10-20M Reads**	**Based on human-sized transcriptomes**
Allele Specific Expression	50 – 100M Reads	
Splice variation	50 - 100M Reads	Also applies to RNA fusion transcripts in cancer
Complete Annotation	100M - 1B Reads	
Transcript Based Assembly	50 – 200M Reads	

***** Applies to poly A-selected libraries**

- **Ribo-Zero, high quality RNA libraries: requires ~ 2X more reads**
- **Ribo-Zero, FFPE libraries: requires ~ 4X more reads**

For Research Use Only. Not for use in diagnostic procedures.

Min Number of Reads for different Applications

Translating min reads/sample to max numbers of samples per flow cell

Application	Read Length	Reads / Sample	MiniSeq/ MiSeq® 25M	NextSeq® 130M	NextSeq 400M	HiSeq 2000M	NovaSeq S2 3300M
Gene expression profiling (GEX) • High Quality RNA	1x50 bp 1x75 bp	**≥10M**	2	12	40	200	-
mRNA-Seq • High Quality RNA	2x75 bp	**≥25M**	1	5	16	80	-
Exome Enriched RNA-Seq • High Quality, degraded or FFPE	2x75 bp	**≥25M**	1	5	16	80	-
Total RNA-Seq • High Quality RNA	2x75 bp	**≥50M**		2	8	40	66
Total RNA-Seq • Degraded RNA • Less efficient rRNA Depletion	2x75 bp	**≥100M**		1	4	20	33
Total RNA-Seq • FFPE Samples • Less efficient rRNA Depletion • Increased intronic sequences	2x75 bp	**≥200M**			2	10	16

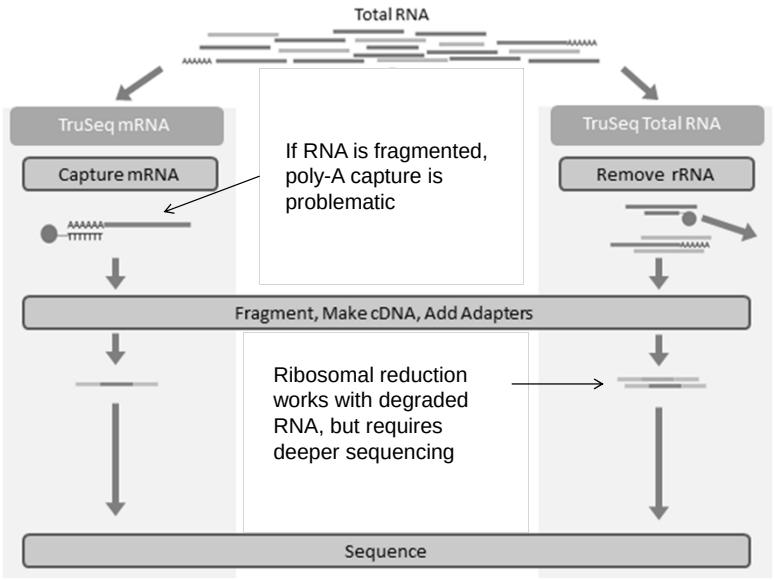
For Research Use Only. Not for use in diagnostic procedures.

Illumina's Current Suite of RNA Library Prep Solutions

Kit	Applications	Molecule Interrogated	FFPE/ Degraded	Min Input
TruSeq Stranded Total RNA	• Expression and splicing analysis • Discover variants, fusions, isoforms	mRNA + ncRNA (excludes miRNA)	Yes	100 ngs
TruSeq Stranded mRNA TruSeq RNA V2	• Expression and splicing analysis • Discover variants, fusions, isoforms • Gene expression profiling	mRNA	No	100 ngs
TruSeq RNA Access	• Expression and splicing analysis • Discover variants, fusions, isoforms • Gene expression profiling	mRNA	Yes	10 ngs
TruSeq small RNA	• Expression analysis • Discover variants, isomiRs	miRNA	No	1ug
TruSeq Targeted RNA Expression	• Profiling/validation of expression and splicing changes • Profiling/validation of variants, fusions, isoforms	Up to 2,000 mRNA + ncRNA (excludes miRNA)	Yes	50 ngs

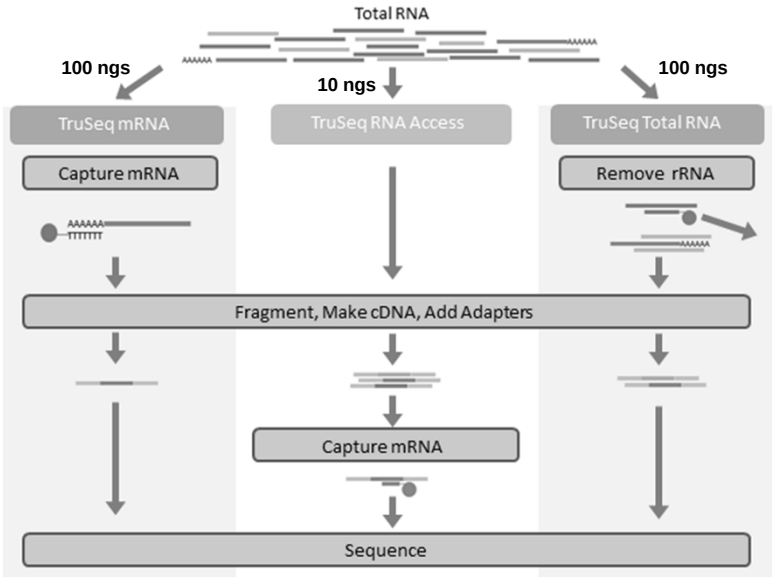


TruSeq® RNA Library Prep Overview

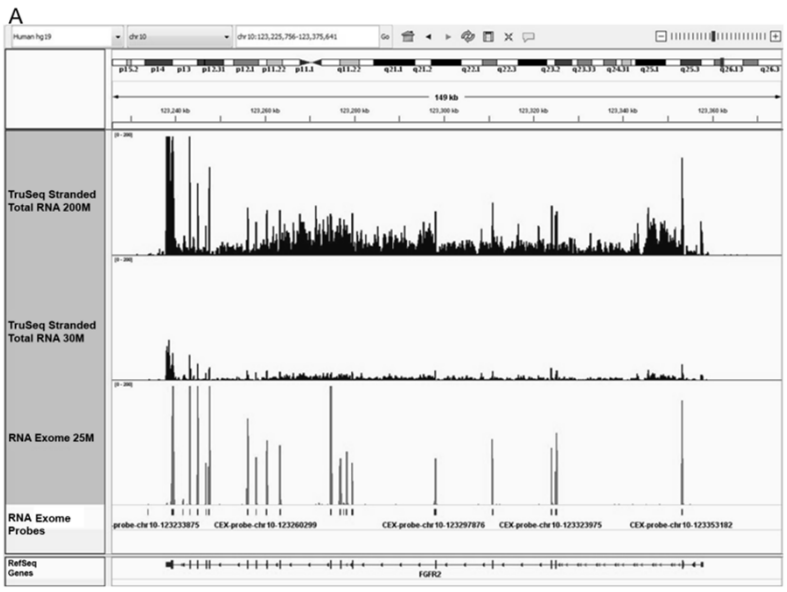


For Research Use Only. Not for use in diagnostic procedures.

TruSeq® RNA Library Prep Overview



For Research Use Only. Not for use in diagnostic procedures.



Bioinformatics Core Facility TMBD

What after RNA-seq

- Sequencing alignment.
- Data preprocessing.
- Basic data output (FPKM, RPKM).
- Validation at individual genes/transcripts level.
- Searching of supporting evidences/consistent findings in the public domain.
- Validation/annotation at pathway level.

July 2019

Bioinformatics Core Facility TMBD

What after RNA-seq

- Sequencing alignment.
- Data preprocessing.
- Basic data output (FPKM, RPKM).
- Validation at individual genes/transcripts level.
- Searching of supporting evidences/consistent findings in the public domain.
- Validation/annotation at pathway level.

July 2019

Bioinformatics Core Facility TMBD

Technical validation using qPCR

Significant correlation between RNA-seq data and RT-qPCR data:
R=0.6~0.8



Fresh frozen samples-cDNA

No correlation between RNA-seq data and RT-qPCR data:
R<0.1

X

FFPE samples-cDNA

July 2019

Bioinformatics Core Facility TMBD

NanoString assay of FFPE samples: a comparison with array data

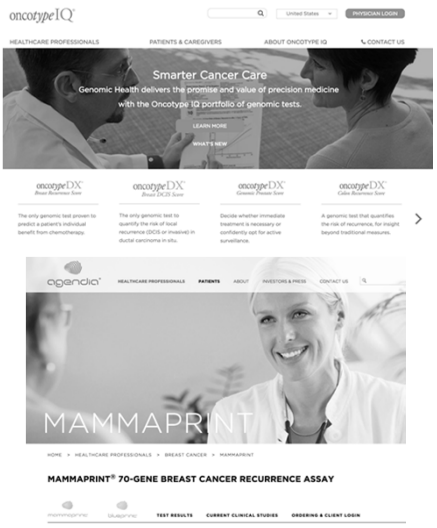
Various data processing protocol

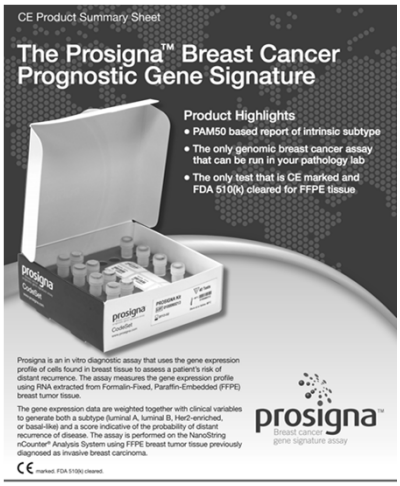
Tumor tissue			A		B		G		H	
Illumina		NanoString	cor	p	cor	p	cor	p	cor	p
CD80	ILMN_1716736	CD80	0.3991	0.01185	0.5826	9.98E-05	0.6106	9.23E-03	0.6087	9.51E-03
CD86	ILMN_1782560	CD86	-0.0175	0.9157	-0.0436	0.7920	-0.2354	6.53E-01	-0.2350	6.54E-01
CD86	ILMN_1714602	CD86	0.0539	0.7445	0.0976	0.5546	0.0460	9.31E-01	0.0471	9.29E-01
CD86	ILMN_1672097	CD86	-0.0478	0.7726	-0.0849	0.6072	-0.1750	7.40E-01	-0.1682	7.50E-01
CTLA4	ILMN_2261627	CTLA4	0.2996	0.06393	0.4270	0.0067	0.4210	1.32E-02	0.4154	1.46E-02
CTLA4	ILMN_2348905	CTLA4	0.3870	0.01495	0.5196	0.0007	0.5130	1.92E-03	0.5108	2.02E-03
CTLA4	ILMN_1763487	CTLA4	0.3248	0.04367	0.6177	2.80E-05	0.6616	2.04E-05	0.6568	2.46E-05
ICOS	ILMN_1669927	ICOS	0.3898	0.01415	0.5714	0.0001	0.5466	1.04E-02	0.5486	1.00E-02

Fresh frozen FFPE

July 2019

Bioinformatics Core Facility TMBD





July 2019

Bioinformatics Core Facility TMBD

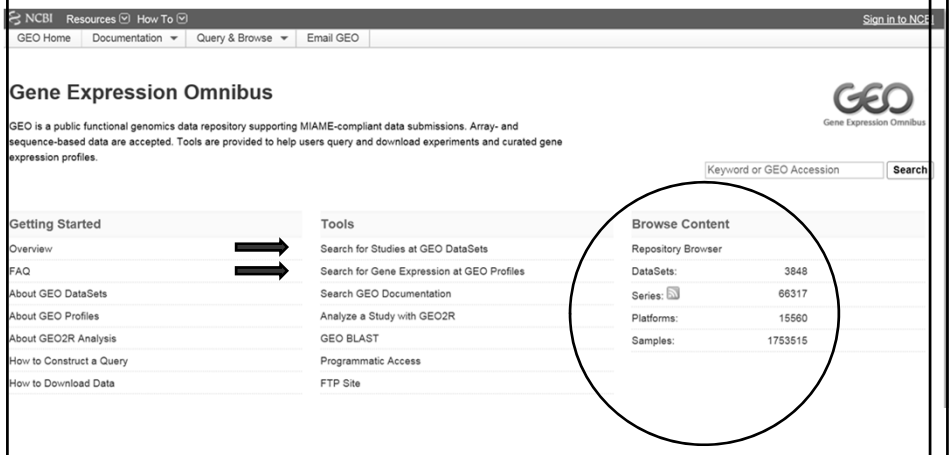
What after RNA-seq

- Sequencing alignment.
- Data preprocessing.
- Basic data output (FPKM, RPKM).
- Validation at individual genes/transcripts level.
- Searching of supporting evidences/consistent findings in the public domain.
- Validation/annotation at pathway level.

July 2019

Bioinformatics Core Facility TMBD

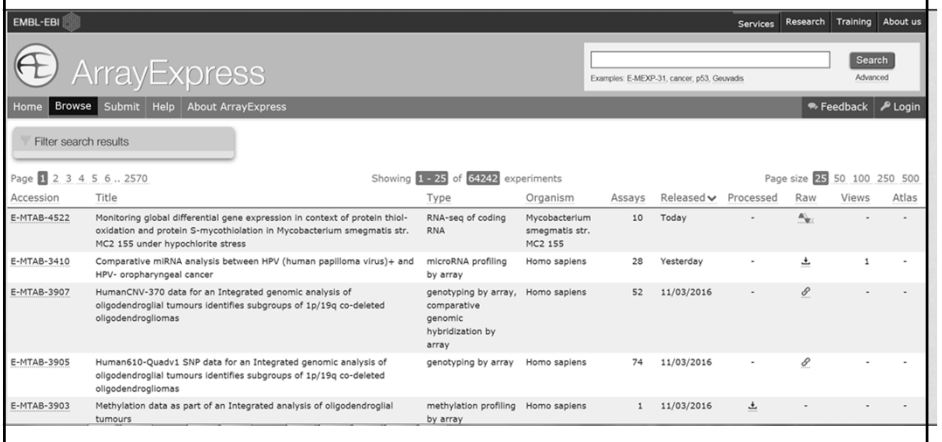
GEO (http://www.ncbi.nlm.nih.gov/geo/)



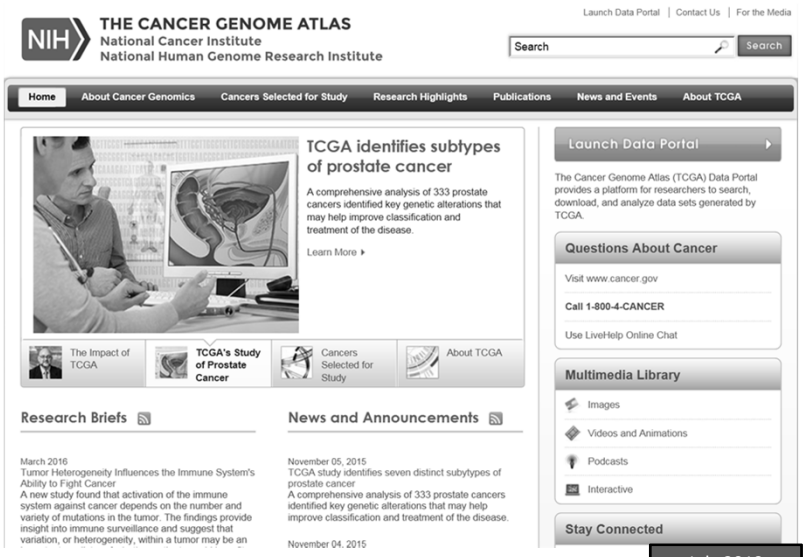
Bioinformatics Core Facility TMBD

ArrayExpress

(https://www.ebi.ac.uk/arrayexpress/browse.html)



TCGA (The Cancer Genome Atlas)



TCGA (The Cancer Genome Atlas)

CANCER TISSUES COLLECTED FOR STUDY

Last Updated: March 04, 2016

Expand All | Collapse All

- Breast
- Central Nervous System
- Endocrine
- Gastrointestinal
- Gynecologic
- Head and Neck
- Hematologic
- Skin
- Soft Tissue
- Thoracic
- Urologic

Cancers Selected for Study

The Cancer Genome Atlas (TCGA) has chosen cancers for study based on specific criteria that include:

- Poor prognosis and overall public health impact
- Availability of human tumor and matched-normal tissue samples that meet TCGA standards for patient consent, quality and quantity

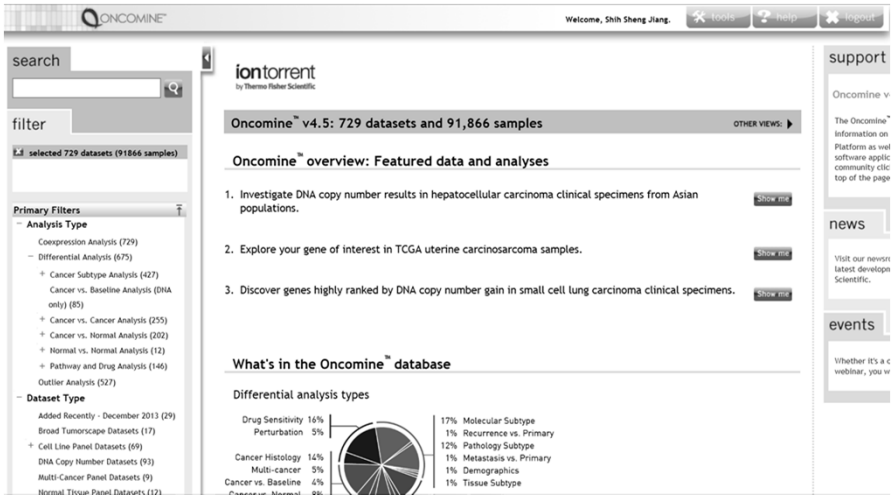
Below is a list of cancers that have been selected for study based on the criteria outlined above. The TCGA program has collected the necessary quality and quantity of samples for these cancers to move them into the TCGA project pipeline. Scheduling for sequencing and characterization for these cancers will be dependent on the capacity of the project pipeline. To view TCGA's clinical data forms, please visit the Biospecimen Core Resource's website.

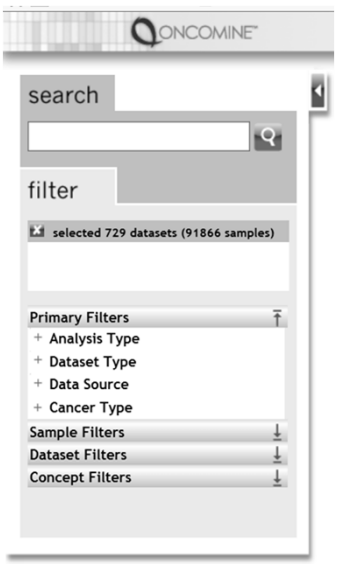
TCGA is no longer accepting additional samples for characterization.

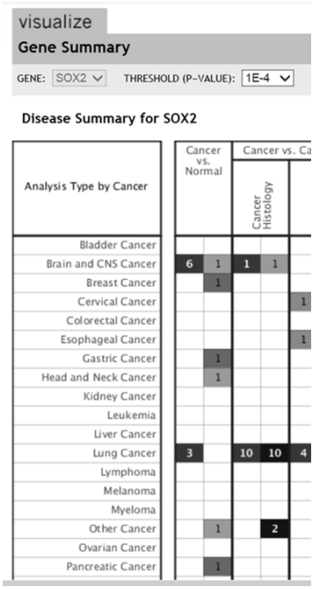
Analysis is underway in the following Analysis Working Groups (AWGs):

- Cervical cancer
- Cholangiocarcinoma
- Esophageal carcinoma
- Liver hepatocellular carcinoma
- Mesothelioma
- Pancreatic ductal adenocarcinoma
- Paraganglioma & Pheochromocytoma
- Sarcoma
- Testicular germ cell cancer
- Thymoma
- Uterine carcinosarcoma
- Uveal melanoma

For more information about the AWGs, please email tcga@mail.nih.gov.


ONCOMINE
Welcome, Shih Sheng Jiang.
tools
help
logout
search
filter
selected 729 datasets (91866 samples)
Primary Filters
Analysis Type
Coexpression Analysis (729)
Differential Analysis (675)
Cancer Subtype Analysis (427)
Cancer vs. Baseline Analysis (DNA only) (85)
Cancer vs. Cancer Analysis (255)
Cancer vs. Normal Analysis (202)
Normal vs. Normal Analysis (12)
Pathway and Drug Analysis (146)
Outlier Analysis (527)
Dataset Type
Added Recently - December 2013 (29)
Broad Tumorscape Datasets (17)
Cell Line Panel Datasets (69)
DNA Copy Number Datasets (93)
Multi-Cancer Panel Datasets (9)
iontorrent
by Thermo Fisher Scientific
Oncomine™ v4.5: 729 datasets and 91,866 samples
OTHER VIEWS:
Oncomine™ overview: Featured data and analyses
1. Investigate DNA copy number results in hepatocellular carcinoma clinical specimens from Asian populations.
Show me
2. Explore your gene of interest in TCGA uterine carcinosarcoma samples.
Show me
3. Discover genes highly ranked by DNA copy number gain in small cell lung carcinoma clinical specimens.
Show me
What's in the Oncomine™ database
Differential analysis types
Drug Sensitivity 16%
Perturbation 5%
Cancer Histology 14%
Multi-cancer 5%
Cancer vs. Baseline 4%
17% Molecular Subtype
1% Recurrence vs. Primary
12% Pathology Subtype
1% Metastasis vs. Primary
1% Demographics
1% Tissue Subtype
support
news
events


ONCOMINE
search
filter
selected 729 datasets (91866 samples)
Primary Filters
Analysis Type
Dataset Type
Data Source
Cancer Type
Sample Filters
Dataset Filters
Concept Filters


visualize
Gene Summary
GENE: SOX2
THRESHOLD (P-VALUE): 1E-4
Disease Summary for SOX2
Analysis Type by Cancer
Cancer vs. Normal
Cancer vs. Ca
Cancer Histology
Bladder Cancer
Brain and CNS Cancer
Breast Cancer
Cervical Cancer
Colorectal Cancer
Esophageal Cancer
Gastric Cancer
Head and Neck Cancer
Kidney Cancer
Leukemia
Liver Cancer
Lung Cancer
Lymphoma
Melanoma
Myeloma
Other Cancer
Ovarian Cancer
Pancreatic Cancer
OTHER VIEWS:
Overview
Dataset View
Concept View
Gene Summary View
Concept Summary View

Bioinformatics Core Facility TMBD

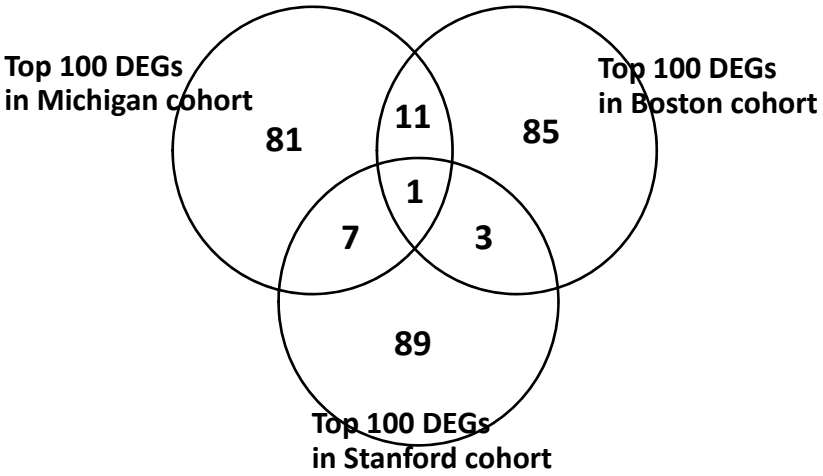
What after RNA-seq

- Sequencing alignment.
- Data preprocessing.
- Basic data output (FPKM, RPKM).
- Validation at individual genes/transcripts level.
- Searching of supporting evidences/consistent findings in the public domain.
- Validation/annotation at pathway level.

July 2019

Bioinformatics Core Facility TMBD

Findings at gene level

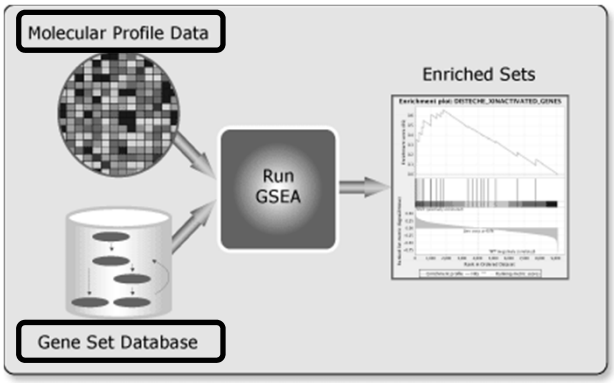


July 2019

Bioinformatics Core Facility TMBD

Gene set enrichment analysis

- Molecular profile data: Microarray, RNA-seq, …
- Gene set database: Molecular signatures database



https://software.broadinstitute.org/gsea/index.jsp

July 2019

Bioinformatics Core Facility TMBD

MSigDB

- The Molecular signatures database (MSigDB) gene sets are divided into 8 major collections

H **hallmark gene sets** are coherently expressed signatures derived by aggregating many MSigDB gene sets to represent well-defined biological states or processes.

C1 **positional gene sets** for each human chromosome and cytogenetic band.

C2 **curated gene sets** from online pathway databases, publications in PubMed, and knowledge of domain experts.

C3 **motif gene sets** based on conserved cis-regulatory motifs from a comparative analysis of the human, mouse, rat, and dog genomes.

C4 **computational gene sets** defined by mining large collections of cancer-oriented microarray data.

C5 **GO gene sets** consist of genes annotated by the same GO terms.

C6 **oncogenic gene sets** defined directly from microarray gene expression data from cancer gene perturbations.

C7 **immunologic gene sets** defined directly from microarray gene expression data from immunologic studies.

http://software.broadinstitute.org/gsea/msigdb/index.jsp

July 2019

Bioinformatics Core Facility TMBD

GSEA

- GSEA is a computational method that applies the Kolmogorov–Smirnov test to compare the distribution of gene scores within **an a priori defined gene set** with one without that gene set.
- A gene score is defined by the difference of gene expressions between categorical phenotypes or correlation between gene expression and a quantitative phenotype.

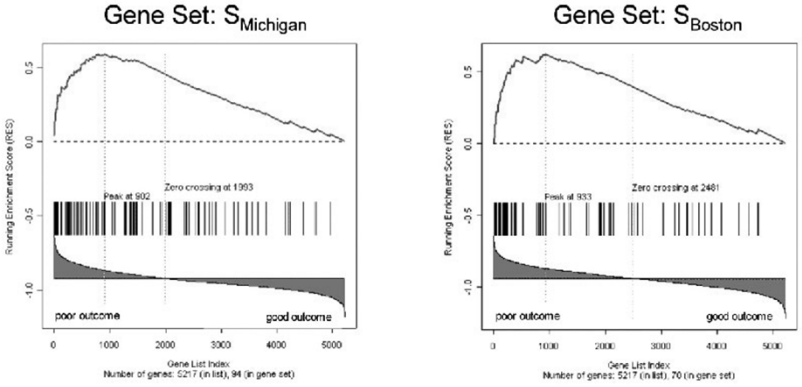
July 2019

Bioinformatics Core Facility TMBD

Results: LADC enrichment score

Boston Dataset

Michigan Dataset



July 2019

Bioinformatics Core Facility TMBD

Results: Lung cancer outcome

- For GSEA using C2, at FDR 0.25, half overlapping gene sets were found.

Boston study	
Enriched in poor outcome	FDR
Hypoxia and p53 in the cardiovascular system	0.050
Aminoacyl tRNA biosynthesis	0.144
Insulin upregulated genes	0.118
tRNA synthetases	0.157
Leucine deprivation down-regulated genes	0.144
Telomerase up-regulated genes	0.128
Glutamine deprivation down-regulated genes	0.146
Cell cycle checkpoint	0.216

Michigan study	
Enriched in poor outcome	FDR
Glycolysis gluconeogenesis	0.006
vegf pathway	0.028
Insulin up-regulated genes	0.147
Insulin signalling	0.170
Telomerase up-regulated genes	0.188
Glutamate metabolism	0.200
Ceramide pathway	0.204
p53 signalling	0.179
tRNA synthetases	0.225
Breast cancer estrogen signalling	0.250
Aminoacyl tRNA biosynthesis	0.229

July 2019

Bioinformatics Core Facility TMBD

Trends and our future works

- Novel Immunotherapy targets in the tissue microenvironment.
- Single cell RNA-seq.
- High quality service and/or collaboration.

July 2019

Acknowledgements

- Dr. I-Shou Chang
- Dr. Wen-Chun Hung
- Dr. Li-Tzong Chen
- Dr. Chung-Hsing Chen
- Ms. Fang-Yu Tsai and other capable RAs
- TBI Bioinformatics core supported by NSC
- 三期癌症計畫

Thank you for your attention

Bioinformatics Core Facility

Introduction to pathway analysis and *in silico* tumor-infiltrating leukocytes assay based on transcriptome data

Chung-Hsing Chen (陳中興), Ph.D.
Institute of Population Health Sciences, NHRI

July 2019

Bioinformatics Core Facility

Transcriptome data source

- Probe-based
  Gene expression microarray.
- Sequencing-based
  RNA-seq (next generation sequencing)

July 2019 2

Mostly addressed questions based on transcriptome data

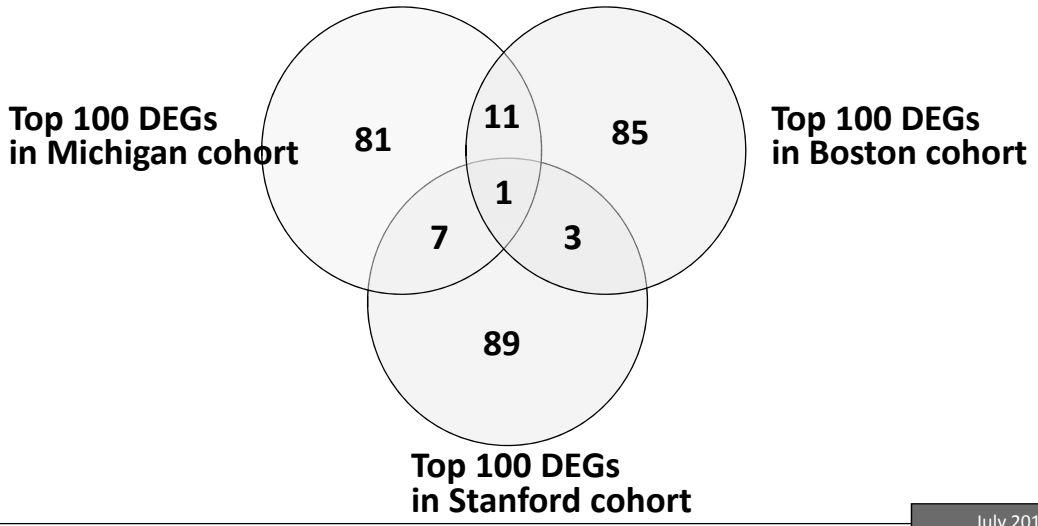
- **Differentially expressed genes** between tumor and adjacent normal tissues.
- Cluster analysis that discovers **tumor subtypes**
- Clinical **association study** (genes whose expression levels are associated with time to recurrence / death)

- Pathway analysis
- Tumor-infiltrating leukocytes

Motivation from LADC studies

- Gene expression profiles in tumor samples from LADC patients and their clinical outcome data, classified as “good” or “poor” outcome.
- Differentially expressed genes (DEGs) in **3 cohorts**: Boston, Michigan, and Stanford.
- Common DEGs?

Problems from LADC study: little or no common genes



Gene set enrichment analysis

PNAS

SEE COMMENTARY

Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles

Aravind Subramanian[a,b], Pablo Tamayo[a,b], Vamsi K. Mootha[a,c], Sayan Mukherjee[d], Benjamin L. Ebert[a,e], Michael A. Gillette[a,f], Amanda Paulovich[g], Scott L. Pomeroy[h], Todd R. Golub[a,e], Eric S. Lander[a,c,i,j,k], and Jill P. Mesirov[a,k]

[a]Broad Institute of Massachusetts Institute of Technology and Harvard, 320 Charles Street, Cambridge, MA 02141; [c]Department of Systems Biology, Alpert 536, Harvard Medical School, 200 Longwood Avenue, Boston, MA 02446; [d]Institute for Genome Sciences and Policy, Center for Interdisciplinary Engineering, Medicine, and Applied Sciences, Duke University, 101 Science Drive, Durham, NC 27708; [e]Department of Medical Oncology, Dana–Farber Cancer Institute, 44 Binney Street, Boston, MA 02115; [f]Division of Pulmonary and Critical Care Medicine, Massachusetts General Hospital, 55 Fruit Street, Boston, MA 02114; [g]Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, C2-023, P.O. Box 19024, Seattle, WA 98109-1024; [h]Department of Neurology, Enders 260, Children's Hospital, Harvard Medical School, 300 Longwood Avenue, Bosto Technology, Cambridge, MA 02142; and [i]Whitehead Institute for Biomedical Re

Contributed by Eric S. Lander, August 2, 2005

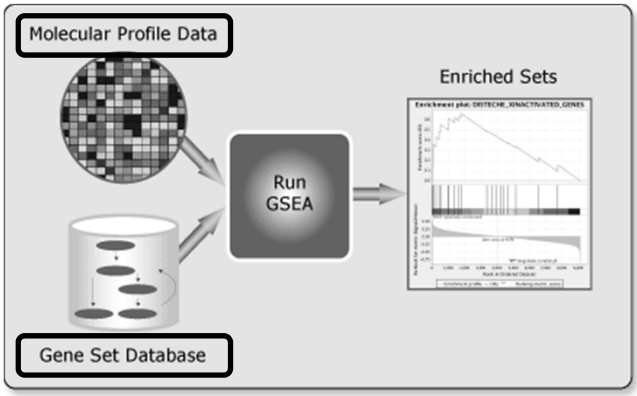
有關 Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles 的學術文章

... set enrichment analysis: **a** knowledge-based approach ... - **Subramanian** - 被引用 16877 次

Gene set enrichment analysis: A knowledge-based approach ... - PNAS
https://www.pnas.org/content/102/43/15545 - 翻譯這個網頁
由 A Subramanian 著作 - 2005 被引用 16877 次 - 相關文章
2005年9月30日 - Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Aravind Subramanian, Pablo ...

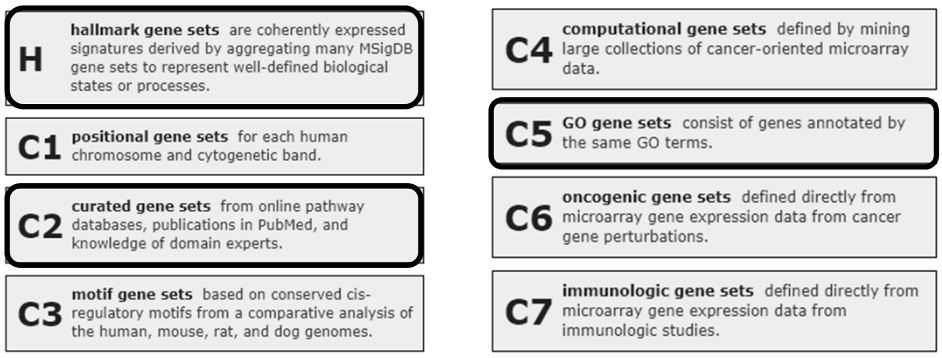
GSEA illustration

- Molecular profile data (user): Microarray, RNA-seq, …
- Gene set database (a priori definition): Molecular signatures database



MSigDB

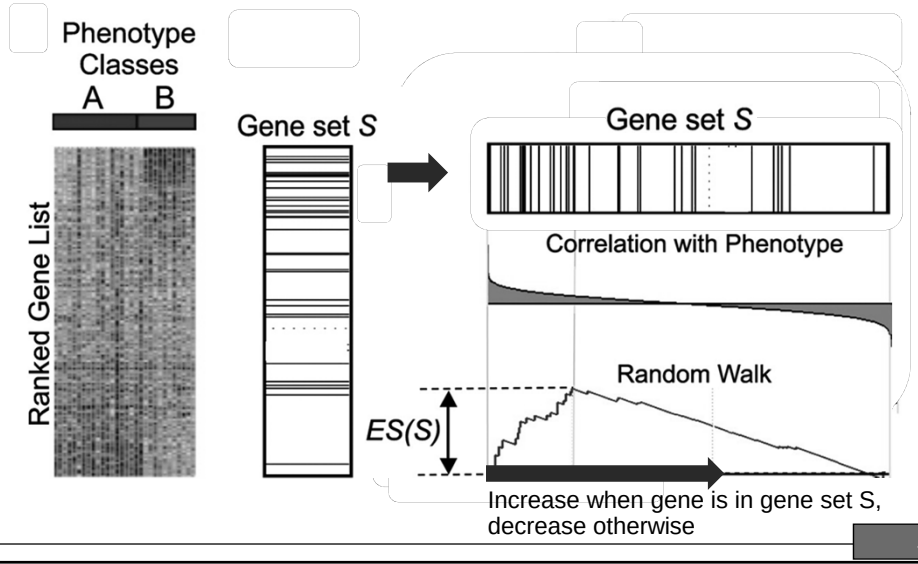
- The Molecular signatures database (MSigDB) gene sets are divided into 8 major collections



GSEA

- GSEA is a computational method that applies the Kolmogorov–Smirnov test to compare the distribution of **gene scores** (metrics) within an a priori defined gene set with one without that gene set.
- A gene score is defined by the difference of gene expressions between categorical phenotypes or correlation between gene expression and a quantitative phenotype.
- **Fold changes** for each gene between samples from tumor and normal tissues is one type of gene score.

How dose it work?



Enrichment Score (ES)

- For all gene i ($\boldsymbol{p}$ is a parameter)

$$P_{\text{hit}}(S, i) = \sum_{\substack{g_j \in S \\ j \le i}} \frac{|r_j|^p}{N_R}, \quad \text{where } N_R = \sum_{g_j \in S} |r_j|^p$$

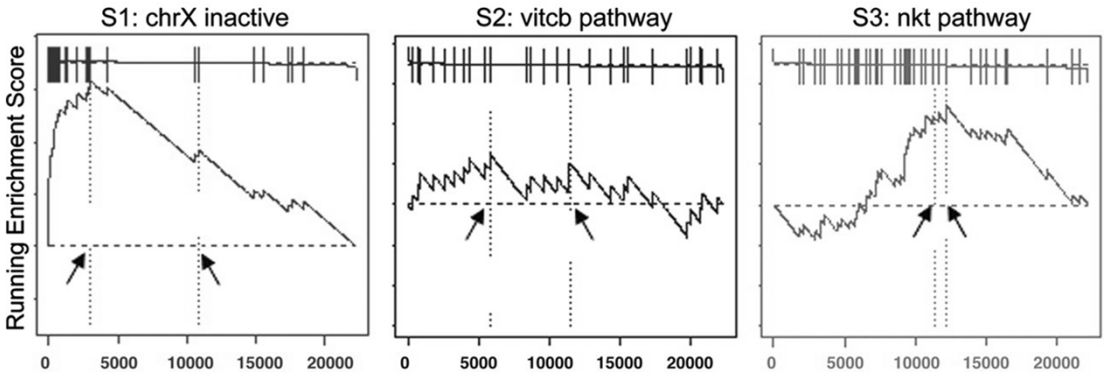
$$P_{\text{miss}}(S, i) = \sum_{\substack{g_j \notin S \\ j \le i}} \frac{1}{(N - N_H)}.$$

- The ES is the largest value for P_{hit}–P_{miss}

Remarks

- p = 0: Kolmogorov-Smirnov statistic
- p = 1: the default in GSEA (recommendation)
- Larger p penalize genes having small correlation with the phenotype.

Enrichment score behavior



Bioinformatics Core Facility

Many commonly enriched gene sets

Boston study	
Enriched in poor outcome	FDR
Hypoxia and p53 in the cardiovascular system	0.050
Aminoacyl tRNA biosynthesis	0.144
Insulin upregulated genes	0.118
tRNA synthetases	0.157
Leucine deprivation down-regulated genes	0.144
Telomerase up-regulated genes	0.128
Glutamine deprivation down-regulated genes	0.146
Cell cycle checkpoint	0.216

Michigan study	
Enriched in poor outcome	FDR
Glycolysis gluconeogenesis	0.006
vegf pathway	0.028
Insulin up-regulated genes	0.147
Insulin signalling	0.170
Telomerase up-regulated genes	0.188
Glutamate metabolism	0.200
Ceramide pathway	0.204
p53 signalling	0.179
tRNA synthetases	0.225
Breast cancer estrogen signalling	0.250
Aminoacyl tRNA biosynthesis	0.229

Bioinformatics Core Facility

Example: pathways associated with p53 mutation

- Gene expression profiles from the 50 cancer cell lines, with 33 being classified as carrying p53 mutation and 17 as normal in gene.
- Identifying targets of the transcription factor p53.

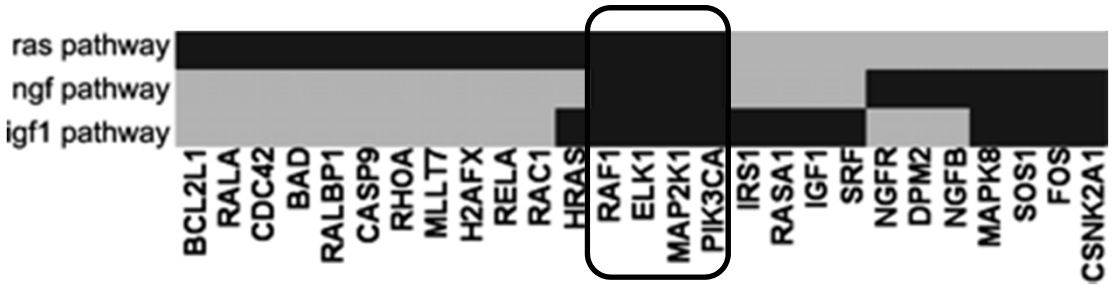
Bioinformatics Core Facility

Results: novel (MAPK) pathways revealed

p53 status in NCI-60 cell lines	
Enriched in p53 mutant	FDR
Ras signaling pathway	0.171
igf1 pathway	0.257
Upregulated by hoxa9	0.275
ngf pathway	0.295
Enriched in p53 wild type	
Hypoxia and p53 in the cardiovascular system	<0.001
Stress induction of HSP regulation	<0.001
p53 signaling pathway	<0.001
p53 up-regulated genes	0.013
Radiation sensitivity genes	0.078

Bioinformatics Core Facility

Leading-edge analysis

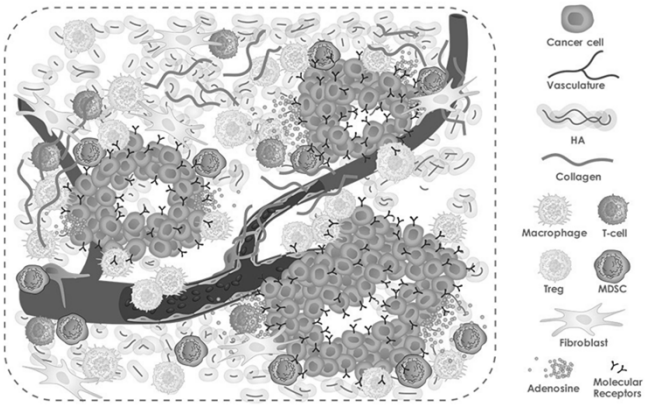


- A common subgroup of genes consists of MAP2K1, PIK3CA, ELK1, and RAF1 and represents a subsection of the MAPK pathway.
- These genes (MAPK pathway) can be further studied.

- Pathway analysis
- Tumor-infiltrating leukocytes

TME

The Tumor Microenvironment



- Tumor purity = % of tumor cells in the sample

http://www.halozyme.com/science/default.aspx

Tumor-infiltrating leukocytes

- The tumor microenvironment is a complex milieu consisting of factors that promote growth and inhibit it, as well as nutrients, chemokines, and other non-cancerous cell types, including fibroblasts, **immune cell**, normal endothelial cell, and epithelial cells.
- The quantification of the different types of tumor-infiltrating leukocytes (TILs) can shed light on the mechanisms underlying the anticancer immune response.
- Immunohistochemistry, immune fluorescence, and flow cytometry
- Computational approaches based on high/medium-throughput genomic data

xCell

Aran *et al. Genome Biology* (2017) 18:220
DOI 10.1186/s13059-017-1349-1

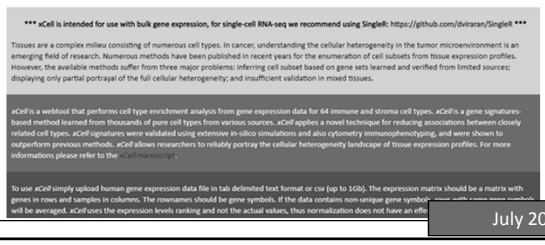
Genome Biology

METHOD Open Access

xCell: digitally portraying the tissue cellular heterogeneity landscape

Dvir Aran*, Zicheng Hu and Atul J. Butte*

- RNA-seq and microarray data



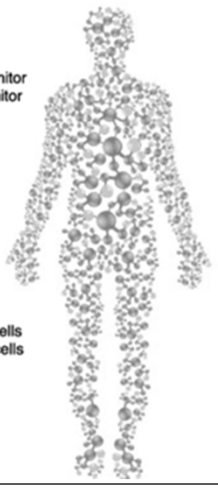
64 cell types

Compendium of human cell type gene signatures

Lymphoids
CD4+ memory T-cells
CD4+ naive T-cells
CD4+ T-cells
Tcm cells
Tem cells
CD4+ Tcm
CD4+ Tem
CD8+ T-cells
CD8+ naive T-cells
CD8+ Tcm
CD8+ Tem
Tregs
Th1 cells
Th2 cells
Tgd cells
NK cells
NKT
B-cells
naive B-cells
Memory B-cells
Class-switched memory B-cells
pro B-cells
Plasma cells

Stem Cells
Hematopoietic stem cells
Common Lymphoid Progenitor
Common myeloid progenitor
Granulocyte-macrophage progenitor
Megakaryocyte-erythroid progenitor
Multipotent progenitors
Megakaryocytes
Erythrocytes
Platelets

Myeloids
Monocytes
Macrophages
Macrophages M1
Macrophages M2
Dendritic cells
Conventional dendritic cells
Plasmacytoid dendritic cells
Immature dendritic cells
Neutrophils
Eosinophils
Mast cells
Basophils



Stromal cells
Mesenchymal stem cells
Adipocytes
Preadipocytes
Stromal cells
Fibroblasts
Pericytes
Endothelial cells
Microvascular endothelial cells
Lymphatic endothelial cells
Smooth muscle cells
Chondrocytes
Osteoblasts
Skeletal muscle cells
Myocytes

Others
Epithelial cells
Sebocytes
Keratinocytes
Mesangial cells
Hepatocytes
Melanocytes
Keratocytes
Astrocytes
Neurons

xCell development pipeline

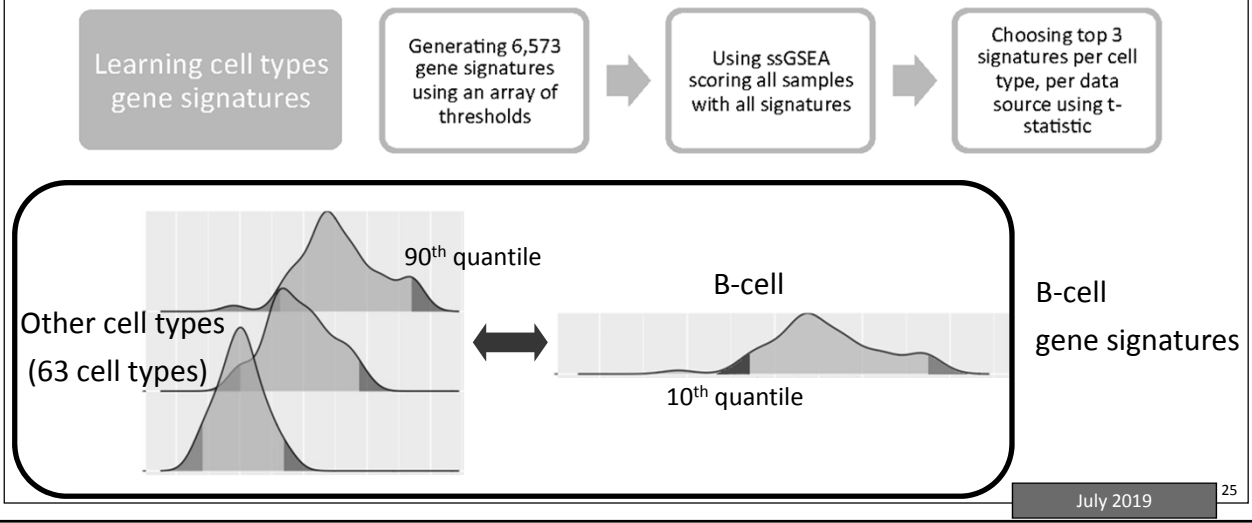
Collecting primary cell types samples → Collecting 1,822 primary cell types from 6 data sources → Manually annotating to 64 cell types → Leave out 97 samples for testing

Data sources for generating gene signatures

FANTOM5	ENCODE project	Blueprint epigenomics	IRIS (Immune response in silico)	Novershtern et al. (Cell 2013)	Human Primary Cells Atlas
757 samples 41 cell types	115 samples 17 cell types	144 samples 28 cell types	107 samples 11 cell types	188 samples 22 cell types	585 samples 39 cell types

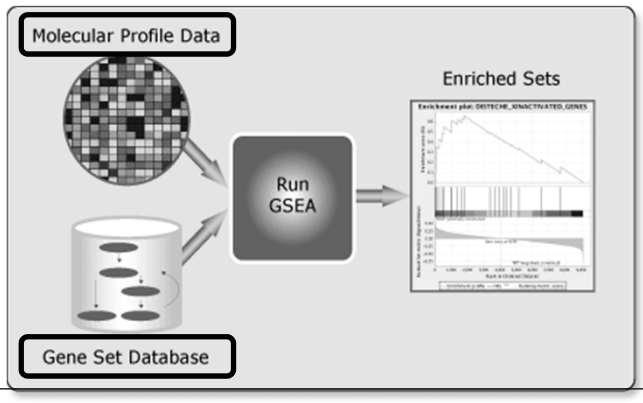
Sequencing-based | Array-based

xCell development pipeline (cont.)



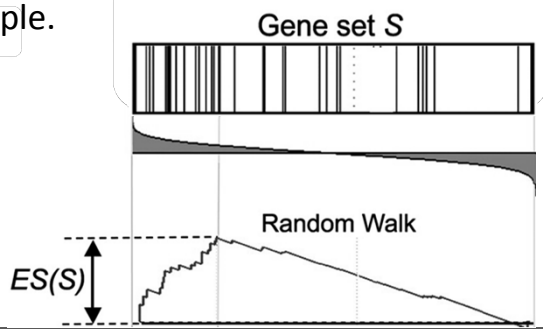
Another GSEA: single-sample GSEA

- Molecular profile data (user): data per sample
- Gene set database (a priori definition): gene set for each cell type



ssGSEA

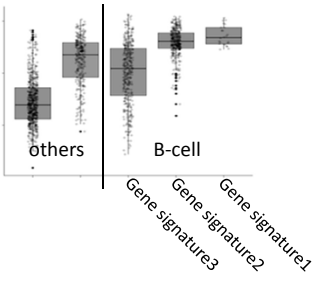
- Single-sample GSEA (ssGSEA) calculates separate enrichment scores **for each pairing of a sample and gene set**.
- Each ssGSEA enrichment score represents the degree to which the genes in a particular gene set are coordinately up- or down-regulated within a sample.



xCell development pipeline (cont.)



- All together we choose 489 signatures corresponding to 64 cell types (across the 6 data sources we have 163 cell types)



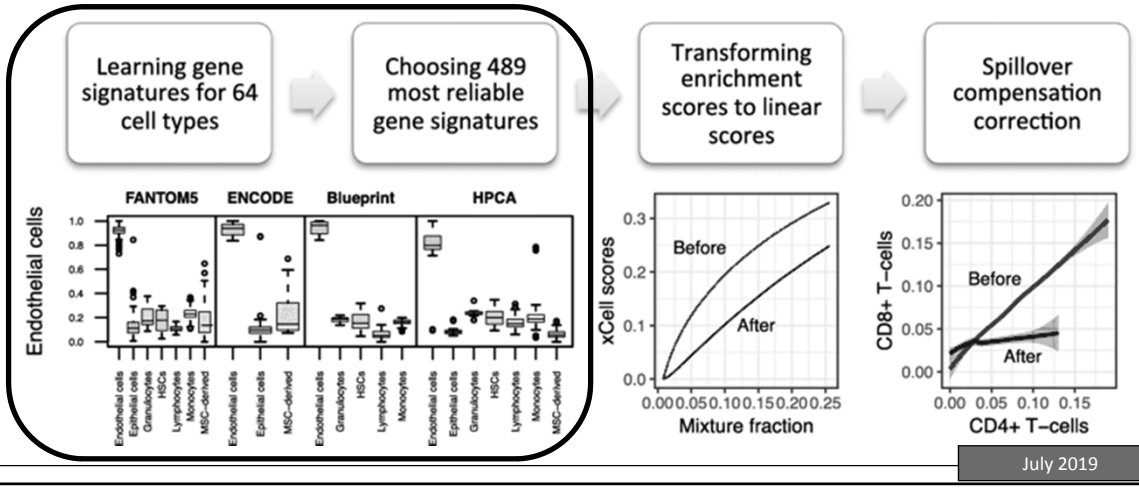
489 gene signatures

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Celltype_Source_ID	# of genes													
2	B-cells_FANTOM_1	16	BLK	CD19	MS4A1	CD22	CD37	CD79A	PNOC	SNX2	MBD4	STAG3	PWP1	SP140	GGA2
3	B-cells_FANTOM_2	12	BLK	CD19	CD22	CD37	CD79A	PNOC	MBD4	STAG3	SP140	GGA2	FCRL2	SMC6	
4	B-cells_FANTOM_3	21	BLK	BTK	CD19	MS4A1	CD22	CD37	CD79A	CSNK1G3	PHKB	PNOC	SNX2	TRAF3	MBD4
5	B-cells_HPCA_1	44	BLK	BTK	CD19	MS4A1	CD22	CD37	CD79A	CSNK1G3	CD180	PHKB	PNOC	POU2F1	PRKCB
6	B-cells_HPCA_2	10	CD19	MS4A1	CD22	CD37	CD79A	SNX2	PRDM4	GGA2	FCRL2	SMC6			
7	B-cells_HPCA_3	8	BLK	CD19	HLA-DOA	SPIB	TNFRSF13B	STAP1	VPREB3	FCRL2					
8	B-cells_NOVERSHTERN_1	65	ACTN2	TNFRSF17	BLK	CXCR5	CD19	MS4A1	CD22	CD37	CD53	CD72	CD79A	CD79B	CNR1
9	B-cells_NOVERSHTERN_2	76	TNFRSF17	BLK	CXCR5	BMP8B	CD19	MS4A1	CD37	CD53	CD72	CD79A	CD79B	CHAD	CCR6
10	B-cells_NOVERSHTERN_3	55	ACTN2	TNFRSF17	BLK	CXCR5	CD19	MS4A1	CD22	CD37	CD53	CD72	CD79A	CD79B	CNR1
11	CD4+ memory T-cells_FANTOM	200	AAMP	ACTL6A	ADSL	AKT2	ANXA7	ARL2	ATF1	BTF3	CCNC	CD2	CD5	CD6	CD28
12	CD4+ memory T-cells_FANTOM	11	CD6	CD28	GPR183	EIF3E	ITK	PKP4	PREPL	DNAJA2	PTGES3	CD2AP	ICOS		
13	CD4+ memory T-cells_FANTOM	139	ADSL	SLC25A6	ANXA7	ATF1	BTF3	CD5	CD6	CD28	CCR4	COX7C	CSNK2A2	CTBP1	DAD1
14	CD4+ memory T-cells_IRIS_1	13	CD3G	CD28	CD40LG	CCR4	CTLA4	GPR15	LIMS1	RBL2	DLEC1	CXCR6	PDCD10	TRAT1	ARHGA
15	CD4+ memory T-cells_IRIS_2	15	CD28	CD40LG	CCR4	CTLA4	GPR15	GZMA	GZMK	HMGB2	PDCD1	RBL2	CD226	GPR171	TRAT1
16	CD4+ memory T-cells_IRIS_3	21	ADSL	CD28	CD40LG	CCR4	CTLA4	GZMA	GZMK	HMGB2	LIMS1	MEN1	PTPN4	ZNF236	DLEC1
17	CD4+ naive T-cells_FANTOM_	15	CD3E	CD6	CD7	CCR7	DSC1	JAK3	PLCL1	PLCO1	ZAP70	CABIN1	PLXDC1	GPSM3	ANKRD
18	CD4+ naive T-cells_FANTOM_	56	APBB1	CD3E	CD6	CD7	CD27	TNFSF8	CCR7	DSC1	GRK6	GZMM	PRMT2	IDUA	INSL3
19	CD4+ naive T-cells_FANTOM_	9	CD27	CLC	CTSW	DNAJB1	RBL2	HAUS3	ANKRD55	ZNF394	CHMP7				
20	CD4+ naive T-cells_HPCA_1	35	CD6	CCR7	DSC1	GLG1	LY9	MAK	PLCL1	RBL2	RBMS1	ATXN7	CUL1	MTRF1	BMS1
21	CD4+ naive T-cells_HPCA_2	20	CCR7	FKTN	IL16	KRT2	NPAT	RBMS1	VPS52	MTRF1	CEPT1	SORCS3	UTP20	PHF20L1	POP5
22	CD4+ naive T-cells_HPCA_3	24	CD6	CCR7	FKTN	IL16	KRT2	NPAT	PLCL1	RBMS1	VPS52	MTRF1	CEPT1	AAK1	SORCS3
23	CD4+ naive T-cells_IRIS_1	17	CD3G	CD4	CD5	CD7	HMOX2	NPAT	RPA3	RPLP2	RPL14	SNPH	TBC1D5	RAB3GAP1	MLH3
24	CD4+ naive T-cells_IRIS_2	12	CD4	CD5	CD7	CD40LG	INPP4A	NPAT	TRAF1	TBC1D5	SIRPG	PARP11	MKL1	GIMAP6	
25	CD4+ naive T-cells_IRIS_3	47	CD2	CD3G	CD247	CD4	CD5	CD7	CD27	CD40LG	CCR7	HMOX2	PRMT2	INPP4A	NPAT
26	CD4+ naive T-cells_NOVERSH	39	CASP8	CD3E	CD4	CD27	CD28	CD40LG	CDK1	CCR7	CTLA4	HMOX2	LAIR2	PLCL1	RBL2
27	CD4+ naive T-cells_NOVERSH	9	APBB1	PRMT2	TATDN2	DIDO1	REV1	RNF216	PLXDC1	NDFIP1	CHMP7				
28	CD4+ naive T-cells_NOVERSH	15	CD3E	CD6	CD7	CCR7	DSC1	JAK3	PLCL1	PLCO1	ZAP70	CABIN1	PLXDC1	GPSM3	ANKRD
29	CD4+ T-cells_BLUEPRINT_1	34	BAD	CD2	CD3G	CD5	CD28	CD40LG	CCR4	CCR7	CTLA4	GOLGA4	HMOX2	MGAT2	NFE2L2
30	CD4+ T-cells_BLUEPRINT_2	67	BAD	CD2	CD3E	CD3G	CD5	CD6	CD28	CD40LG	CCR4	CCR7	CTLA4	DNAH6	GOLGA

xCell_signatures.txt

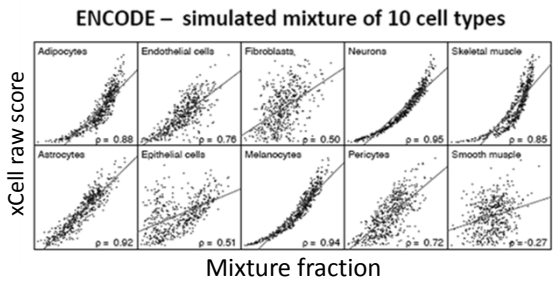
xCell development pipeline (cont.)

xCell pipeline: Adjusted gene expression enrichment analysis



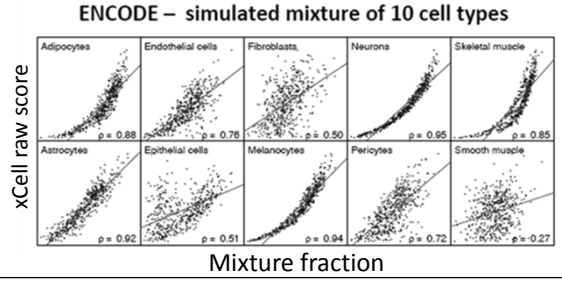
Raw score transformation

- We created a synthetic mixture using the expression profile of the cell type X and an additional “control” cell type. (0.8% cell X and 99.2% control, 1.6% cell X and 98.4% control, etc.)



Problems with the raw scores

- ssGSEA scores have different distributions between different signatures and a score from signatures cannot thus be compared with a score from another signature.
- In sequencing-based data, the association between the underlying levels of the cell type was not **linearly associated** with the score.

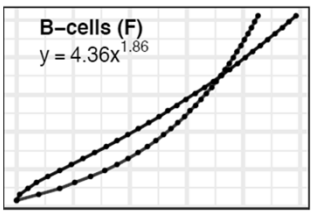


Transforming ES to linear scores

B-cell "control" Raw ES shifted the scores to 0 and divided by 5000 transforming to linear scores

$$\begin{bmatrix} y_{1i}^{(0.8)} \\ y_{2i}^{(0.8)} \\ \vdots \\ y_{Gi}^{(0.8)} \end{bmatrix} = 0.8\% \times \begin{bmatrix} y_{1i} \\ y_{2i} \\ \vdots \\ y_{Gi} \end{bmatrix} + 99.2\% \times \begin{bmatrix} y_{1c} \\ y_{2c} \\ \vdots \\ y_{Gc} \end{bmatrix} \Rightarrow S_{i,0.8}^{*} \Rightarrow S_{i,0.8} = \frac{(S_{i,0.8}^{*} - S_{i,0.8}^{*})}{5000} \Rightarrow V1_i \times S_{i,0.8}^{P_i}$$

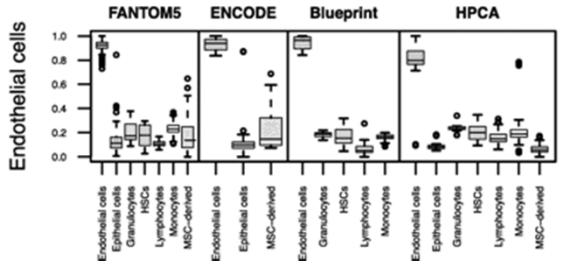
$$\begin{bmatrix} y_{1i}^{(1.6)} \\ y_{2i}^{(1.6)} \\ \vdots \\ y_{Gi}^{(1.6)} \end{bmatrix} = 1.6\% \times \begin{bmatrix} y_{1i} \\ y_{2i} \\ \vdots \\ y_{Gi} \end{bmatrix} + 98.4\% \times \begin{bmatrix} y_{1c} \\ y_{2c} \\ \vdots \\ y_{Gc} \end{bmatrix} \Rightarrow S_{i,1.6}^{*} \Rightarrow S_{i,1.6} = \frac{(S_{i,1.6}^{*} - S_{i,0.8}^{*})}{5000} \Rightarrow V1_i \times S_{i,1.6}^{P_i}$$

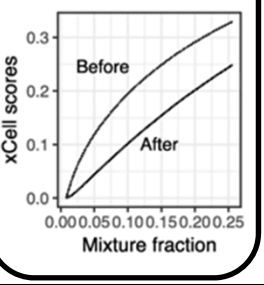


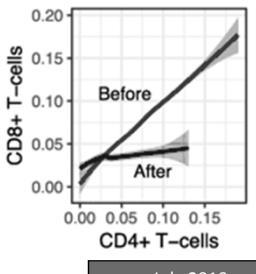
xCell development pipeline (cont.)

xCell pipeline: Adjusted gene expression enrichment analysis

Learning gene signatures for 64 cell types → Choosing 489 most reliable gene signatures → Transforming enrichment scores to linear scores → Spillover compensation correction







Spillover compensation reference matrix

- Another limitation that was observed in the mixtures is the dependencies between closely related cell type (scores that predict enrichment of one cell type also predict enrichment of another cell type)

$$\begin{bmatrix} y^*_{1i} \\ y^*_{2i} \\ \vdots \\ y^*_{Gi} \end{bmatrix} = 25\% \times \begin{bmatrix} y_{1i} \\ y_{2i} \\ \vdots \\ y_{Gi} \end{bmatrix} + 75\% \times \begin{bmatrix} y_{1c} \\ y_{2c} \\ \vdots \\ y_{Gc} \end{bmatrix} \Rightarrow \begin{bmatrix} y^*_{11} & y^*_{12} & \cdots & y^*_{1M} \\ y^*_{21} & y^*_{22} & \cdots & y^*_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ y^*_{G1} & y^*_{G2} & \cdots & y^*_{GM} \end{bmatrix}$$

(cell type i; "control" cell type)

$$\Rightarrow \quad S^* = \begin{array}{c} \\ \text{cell type 1} \\ \text{cell type 2} \\ \vdots \\ \text{cell type } M \end{array} \begin{array}{c} \begin{array}{cccc} \text{sample 1} & \text{sample 2} & \cdots & \text{sample } M \end{array} \\ \begin{bmatrix} S^*_{11} & S^*_{12} & \cdots & S^*_{1M} \\ S^*_{21} & S^*_{22} & \cdots & S^*_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ S^*_{M1} & S^*_{M2} & \cdots & S^*_{MM} \end{bmatrix} \end{array}$$

$$\Downarrow \quad S_{ij} = \frac{1}{V1_i}\left(\frac{S^*_{ij} - \min_j(S^*_i)}{5000}\right)^{P_i}$$

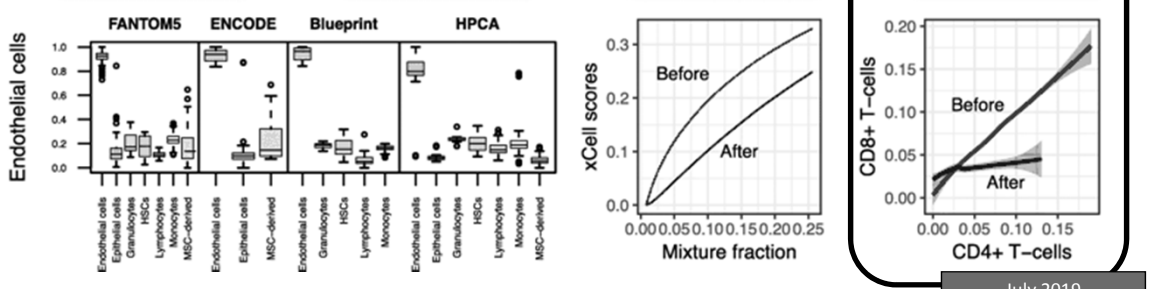
$$\text{spillover matrix } K \equiv \begin{bmatrix} S_{11}/S_{11} & S_{12}/S_{11} & \cdots & S_{1M}/S_{11} \\ S_{21}/S_{22} & S_{22}/S_{22} & \cdots & S_{2M}/S_{22} \\ \vdots & \vdots & \ddots & \vdots \\ S_{M1}/S_{MM} & S_{M2}/S_{MM} & \cdots & S_{MM}/S_{MM} \end{bmatrix}$$

$$S = \begin{array}{c} \\ \text{cell type 1} \\ \text{cell type 2} \\ \vdots \\ \text{cell type } M \end{array} \begin{array}{c} \begin{array}{cccc} \text{sample 1} & \text{sample 2} & \cdots & \text{sample } M \end{array} \\ \begin{bmatrix} S_{11} & S_{12} & \cdots & S_{1M} \\ S_{21} & S_{22} & \cdots & S_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ S_{M1} & S_{M2} & \cdots & S_{MM} \end{bmatrix} \end{array}$$

xCell development pipeline (cont.)

xCell pipeline: Adjusted gene expression enrichment analysis

Learning gene signatures for 64 cell types → Choosing 489 most reliable gene signatures → Transforming enrichment scores to linear scores → Spillover compensation correction

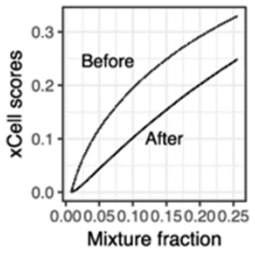


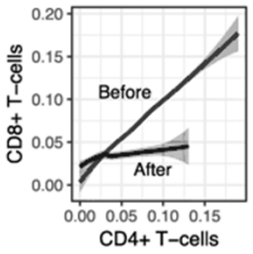
Summary xCell pipeline

xCell pipeline: Adjusted gene expression enrichment analysis

Learning gene signatures for 64 cell types → Choosing 489 most reliable gene signatures → Transforming enrichment scores to linear scores → Spillover compensation correction

$$T_{ij} = {(A_{ij} - \min(A_i))/5000}^{P_i} \Big/ (V1_i \cdot V2_i)$$





xCell input data remark

- A gene expression data set normalized to gene length (such as FPKM or TPM), where rows are genes and columns are samples
- Average duplicate gene names
- xCell uses a set of 10,808 genes for the algorithm (the xCell web tool requires intersection of at least 5000 genes)

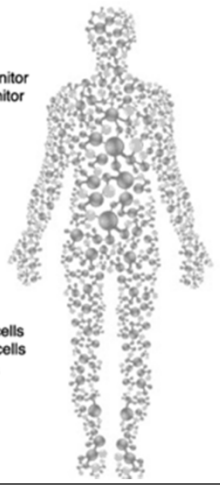
64 cell type gene signatures

Compendium of human cell type gene signatures

Lymphoids
CD4+ memory T-cells
CD4+ naive T-cells
CD4+ T-cells
Tcm cells
Tem cells
CD4+ Tcm
CD4+ Tem
CD8+ T-cells
CD8+ naive T-cells
CD8+ Tcm
CD8+ Tem
Tregs
Th1 cells
Th2 cells
Tgd cells
NK cells
NKT
B-cells
naive B-cells
Memory B-cells
Class-switched memory B-cells
pro B-cells
Plasma cells

Stem Cells
Hematopoietic stem cells
Common Lymphoid Progenitor
Common myeloid progenitor
Granulocyte-macrophage progenitor
Megakaryocyte-erythroid progenitor
Multipotent progenitors
Megakaryocytes
Erythrocytes
Platelets

Myeloids
Monocytes
Macrophages
Macrophages M1
Macrophages M2
Dendritic cells
Conventional dendritic cells
Plasmacytoid dendritic cells
Immature dendritic cells
Neutrophils
Eosinophils
Mast cells
Basophils

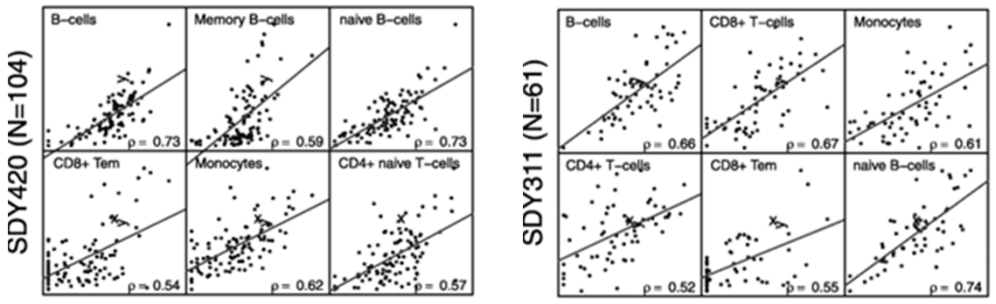


Stromal cells
Mesenchymal stem cells
Adipocytes
Preadipocytes
Stromal cells
Fibroblasts
Pericytes
Endothelial cells
Microvascular endothelial cells
Lymphatic endothelial cells
Smooth muscle cells
Chondrocytes
Osteoblasts
Skeletal muscle cells
Myocytes

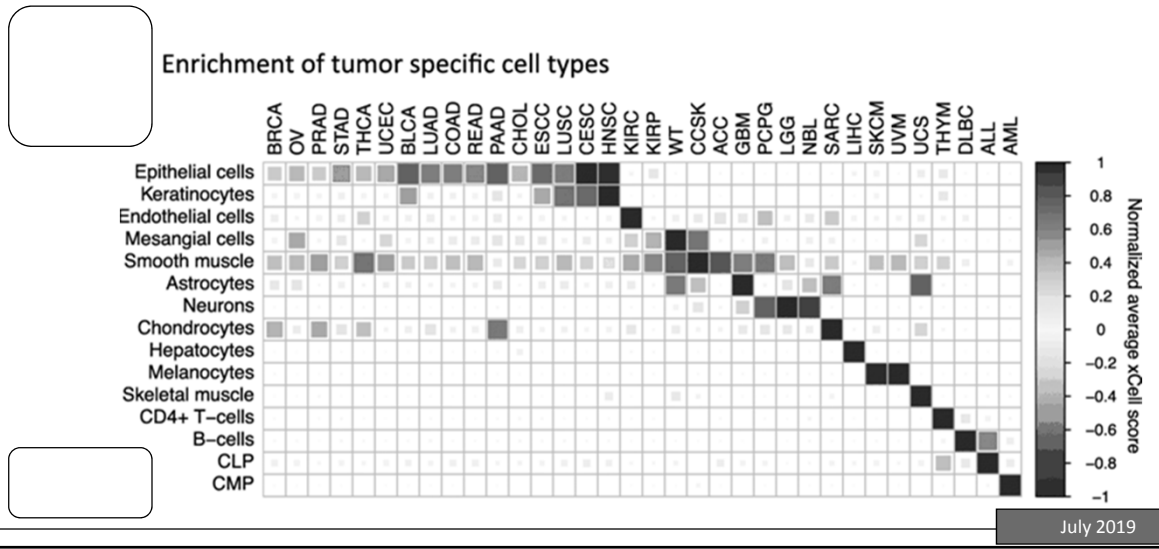
Others
Epithelial cells
Sebocytes
Keratinocytes
Mesangial cells
Hepatocytes
Melanocytes
Keratocytes
Astrocytes
Neurons

Validation

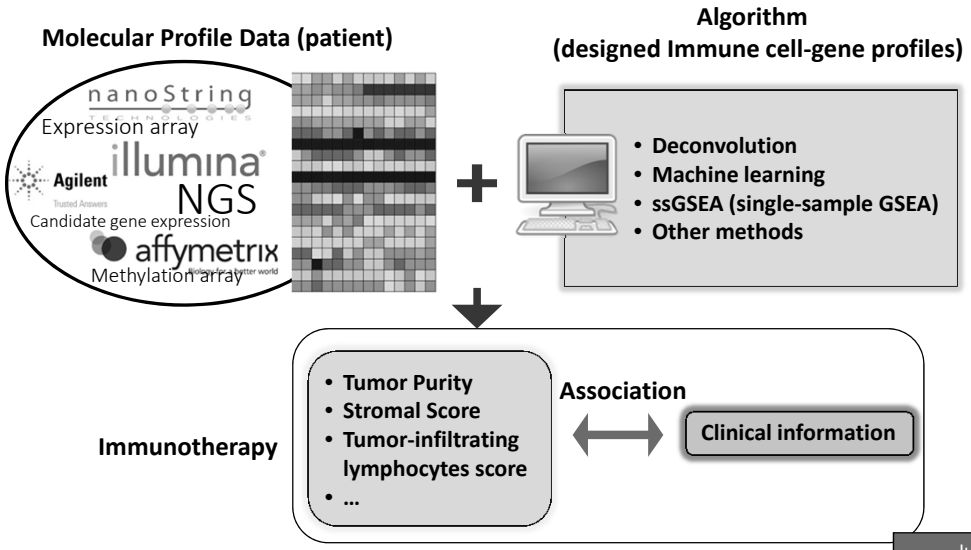
- 165 individuals were studied for both gene expression from the whole blood and flow cytometry across 18 cell subsets from PBMC.



Cell type enrichment analysis in TCGA



Workflow



Acknowledgements

- National Institute of Cancer Research, NHRI
  I-Shou Chang, Shih-Sheng Jiang
- Institute of Population Health Sciences, NHRI
  Chao Hsiung, Tzu-Yu Chen
- Capable RAs
- TBI Bioinformatics core supported by NSC

Thank you for your attention

Advantages of GSEA

- A gene scores definition (p-values, fold changes, hazard ratios)
- Rank-based approach
- The gene-gene dependency

Advantages of xCell

- Gene signatures are rank-based
  (agnostic to normalization methods; concerns related to batch effects)
- No decline in performance with increasing numbers of cell types
  (a partial portrayal of the TME may result in misleading findings)
- Gene signatures are simple and can easily be adjusted.

Comparison

- The deconvolution approaches (CIBERSORT, TIMER)
- Training dataset, gene expression platforms; gene detection efficiencies (FFPE vs. frozen tissue / reagent batches)
- Tumors can express genes typically associated with immune cells

- Single-sample GSEA (xCell)
- Interpretation

GSEA:
Gene Set Enrichment Analysis
軟體操作

江士昇　陳中興　蔡芳榆

1

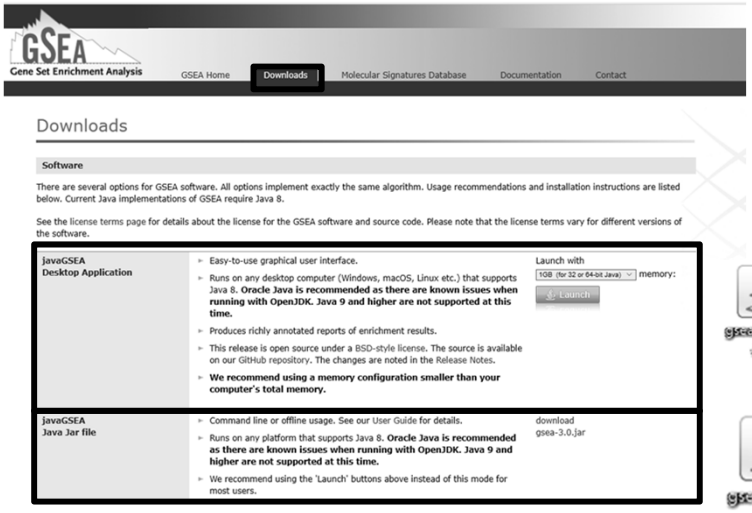
Steps

- 1.Download GSEA
- 2.Prepare data
- 3. Loading data
- 4. Running analysis
- 5. Viewing analysis results

2

Step1.Download GSEA

http://software.broadinstitute.org/gsea/index.jsp



3

Step2.Prepare data

需準備的實驗數據, 樣本資料

Data File	Content	Format	Source
Expression dataset	Contains features (genes or probes), samples, and an expression value for each feature in each sample. Expression data can come from any source (Affymetrix, Stanford cDNA, and so on).	res, gct, pcl, or txt	You create the file.
Phenotype labels	Contains phenotype labels and associates each sample with a phenotype.	cls	You create the file or have GSEA create it for you.
Gene sets	Contains one or more gene sets. For each gene set, gives the gene set name and list of features (genes or probes) in that gene set.	gmx or gmt	You use the files on the Broad ftp site, export gene sets from the Molecular Signature Database (MSigDb) or create your own gene sets file.
Chip annotations	Lists each probe on a DNA chip and its matching HUGO gene symbol. Optional for the gene set enrichment analysis.	Chip	You use the files on the Broad ftp site, download the files from the GSEA web site, or create your own chip file.

可用GSEA內建的 Gene Set, 實驗平台資訊(probe-gene)

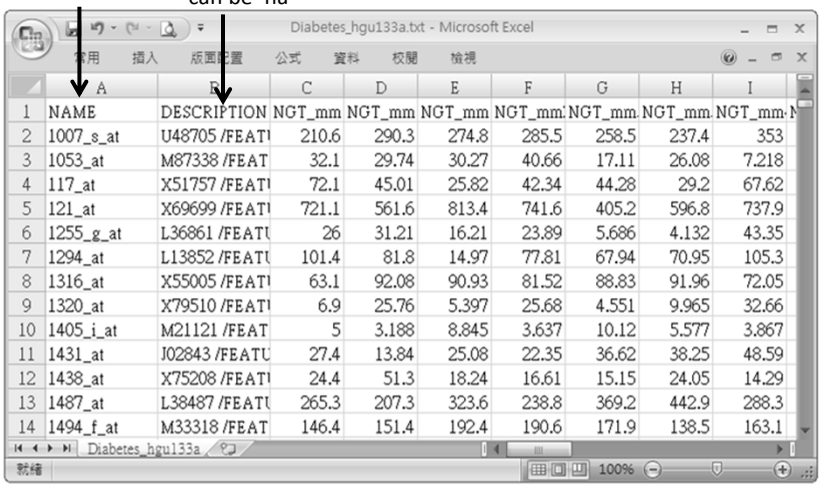
4

Expression data format: .txt

The **first line** contains the labels **NAME** and **DESCRIPTION** followed by the identifiers for each sample in the dataset.

Probe name

can be 'na'



Diabetes_hgu133a.txt - Microsoft Excel

	A	B	C	D	E	F	G	H	I
1	NAME	DESCRIPTION	NGT_mm	NGT_mm	NGT_mm	NGT_mm	NGT_mm	NGT_mm	NGT_mm
2	1007_s_at	U48705 /FEATI	210.6	290.3	274.8	285.5	258.5	237.4	353
3	1053_at	M87338 /FEAT	32.1	29.74	30.27	40.66	17.11	26.08	7.218
4	117_at	X51757 /FEATI	72.1	45.01	25.82	42.34	44.28	29.2	67.62
5	121_at	X69699 /FEATI	721.1	561.6	813.4	741.6	405.2	596.8	737.9
6	1255_g_at	L36861 /FEATU	26	31.21	16.21	23.89	5.686	4.132	43.35
7	1294_at	L13852 /FEATU	101.4	81.8	14.97	77.81	67.94	70.95	105.3
8	1316_at	X55005 /FEATI	63.1	92.08	90.93	81.52	88.83	91.96	72.05
9	1320_at	X79510 /FEATI	6.9	25.76	5.397	25.68	4.551	9.965	32.66
10	1405_i_at	M21121 /FEAT	5	3.188	8.845	3.637	10.12	5.577	3.867
11	1431_at	J02843 /FEATU	27.4	13.84	25.08	22.35	36.62	38.25	48.59
12	1438_at	X75208 /FEATI	24.4	51.3	18.24	16.61	15.15	24.05	14.29
13	1487_at	L38487 /FEATU	265.3	207.3	323.6	238.8	369.2	442.9	288.3
14	1494_f_at	M33318 /FEAT	146.4	151.4	192.4	190.6	171.9	138.5	163.1

Diabetes_hgu133a

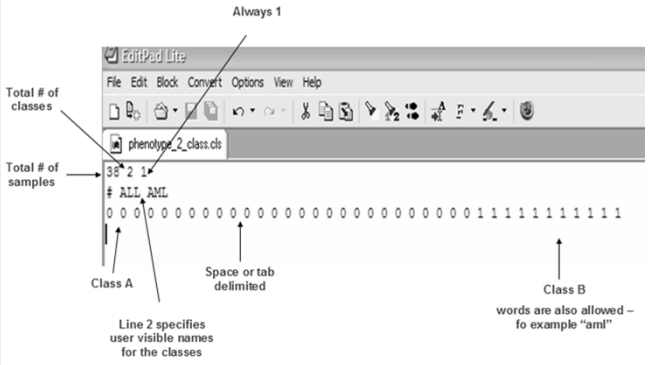
就緒

100%

5

Phenotype data format: .cls

Categorical class file format (e.g NGT vs DMT; tumor vs normal)



Example of a 3 class cls file

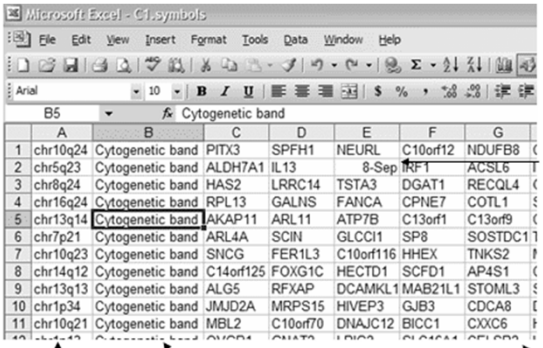


Continuous file format (e.g time-series or gene profile)

```
#numeric
#AFFX-BioB-5_st
206.0 31.0 252.0 -20.0 -169.0 -66.0 230.0 -23.0 67.0 173.0 -55.0 -20.0 469.0 -201.0 -117.0 -162.0 -5.0 -86.0
350.0 74.0 -215.0 193.0 506.0 183.0 350.0 113.0 -17.0 29.0 247.0 -131.0 358.0 561.0 24.0 524.0 167.0 -56.0
176.0 320.0
#AFFX-BioDn-5
75.0 142.0 32.0 109.0 -38.0 -80.0 62.0 39.0 196.0 -42.0 199.0 49.0 171.0 327.0 115.0 -71.0 85.0 80.0 270.0
182.0 208.0 -94.0 292.0 233.0 34.0 0.0 59.0 233.0 48.0 466.0 -7.0 -96.0 297.0 38.0 208.0 -15.0 30.0 357.0
```

Gene set database formats : .gmt

Each row represents one gene set



	A	B	C	D	E	F	G
1	chr10q24	Cytogenetic band	PITX3	SPFH1	NEURL	C10orf12	NDUFB8
2	chr5q23	Cytogenetic band	ALDH7A1	IL13	8-Sep	IRF1	ACSL6
3	chr8q24	Cytogenetic band	HAS2	LRRC14	TSTA3	DGAT1	RECQL4
4	chr16q24	Cytogenetic band	RPL13	GALNS	FANCA	CPNE7	COTL1
5	chr13q14	Cytogenetic band	AKAP11	ARL11	ATP7B	C13orf1	C13orf9
6	chr7p21	Cytogenetic band	ARL4A	SCIN	GLCCI1	SP8	SOSTDC1
7	chr10q23	Cytogenetic band	SNCG	FER1L3	C10orf116	HHEX	TNKS2
8	chr14q12	Cytogenetic band	C14orf125	FOXG1C	HECTD1	SCFD1	AP4S1
9	chr13q13	Cytogenetic band	ALG5	RFXAP	DCAMKL1	MAB21L1	STOML3
10	chr1p34	Cytogenetic band	JMJD2A	MRPS15	HIVEP3	GJB3	CDCA8
11	chr10q21	Cytogenetic band	MBL2	C10orf70	DNAJC12	BICC1	CXXC6

If editing in excel, watch out for its tendency to auto-format gene sets (SEP8 becomes 8-Sep)

First column are gene set names. Duplicates are not allowed

Second column contains a brief description. Its optional – you can fill in a dummy field (e.g. "na")

Unequal lengths (i.e # of genes) *is* allowed

GMT format is convenient to store large databases of gene sets. For a handful of sets (<256) the gmx format offers greater excel-editability

7

Microarray chip annotation formats : .chip

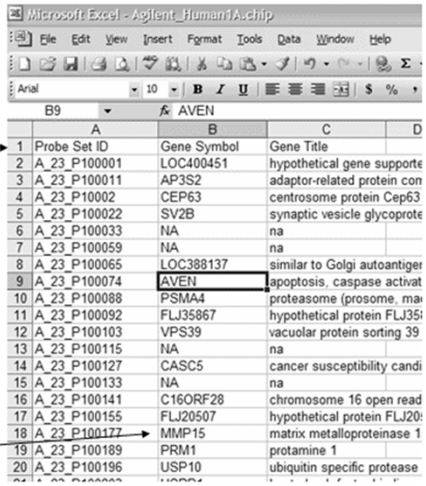
Columns can appear in any order (and there can be more columns which will be ignored).

The column headers must be identical to listed:

Probe Set ID

Gene Symbol

Gene Title



	A	B	C	D
1	Probe Set ID	Gene Symbol	Gene Title	
2	A_23_P100001	LOC400451	hypothetical gene supporte	
3	A_23_P100011	AP3S2	adaptor-related protein con	
4	A_23_P10002	CEP63	centrosome protein Cep63	
5	A_23_P100022	SV2B	synaptic vesicle glycoprote	
6	A_23_P100033	NA	na	
7	A_23_P100059	NA	na	
8	A_23_P100065	LOC388137	similar to Golgi autoantiger	
9	A_23_P100074	AVEN	apoptosis, caspase activat	
10	A_23_P100088	PSMA4	proteasome (prosome, ma	
11	A_23_P100092	FLJ35867	hypothetical protein FLJ35	
12	A_23_P100103	VPS39	vacuolar protein sorting 39	
13	A_23_P100115	NA	na	
14	A_23_P100127	CASC5	cancer susceptibility candi	
15	A_23_P100133	NA	na	
16	A_23_P100141	C16ORF28	chromosome 16 open read	
17	A_23_P100155	FLJ20507	hypothetical protein FLJ20	
18	A_23_P100177	MMP15	matrix metalloproteinase 1	
19	A_23_P100189	PRM1	protamine 1	
20	A_23_P100196	USP10	ubiquitin specific protease	

Gene symbols are considered to be case-INsensitive

Column 1: probe set identifiers. These correspond to the features on the microarray. Must be unique

Column 2: Gene symbol of the probe set. Enter NA if there is none. Need *not be unique*

Column 3: Gene title of the probe set. Enter NA if there is none

8

Step3. Loading data

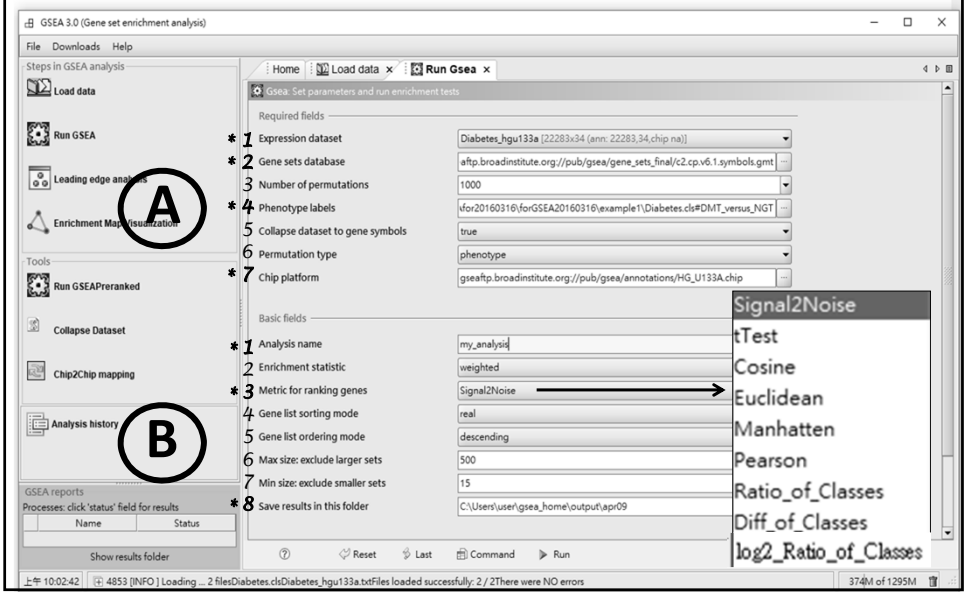


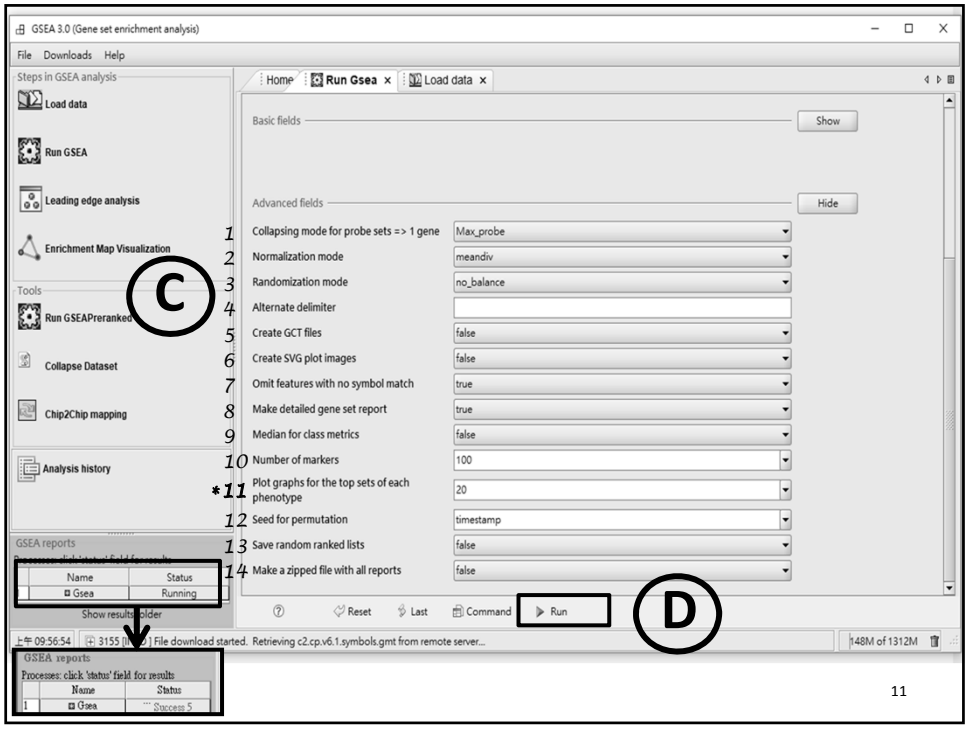
資料格式有錯誤, 需再檢查

資料格式正確

9

Step4. Running analysis





11

- Metrics for ranking genes

For categorical phenotypes

Signal2Noise

$$\frac{\mu_A - \mu_B}{\sigma_A + \sigma_B}$$

tTest

$$\frac{\mu_A - \mu_B}{\sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}}$$

Ratio_of_Classes

$$\frac{\mu_A}{\mu_B}$$

Diff_of_Classes

$$\mu_A - \mu_E$$

log2_Ratio_of_Classes

$$\log2\left(\frac{\mu_A}{\mu_B}\right)$$

For continuous phenotypes

Pearson

Cosine

Manhattan

Euclidean

12

Step5. Viewing analysis results

GSEA Report for Dataset Diabetes_hgu133a

Enrichment in phenotype: NGT (17 samples)

- 109 / 305 gene sets are upregulated in phenotype **NGT**
- 3 gene sets are significant at FDR < 25%
- 1 gene sets are significantly enriched at nominal pvalue < 1%
- 3 gene sets are significantly enriched at nominal pvalue < 5%
- Snapshot of enrichment results
- → Detailed enrichment results in html format
- Detailed enrichment results in excel format (tab delimited text)
- Guide to interpret results

Enrichment in phenotype: DMT (17 samples)

- 196 / 305 gene sets are upregulated in phenotype **DMT**
- 0 gene sets are significantly enriched at FDR < 25%
- 2 gene sets are significantly enriched at nominal pvalue < 1%
- 7 gene sets are significantly enriched at nominal pvalue < 5%
- Snapshot of enrichment results
- → Detailed enrichment results in html format
- Detailed enrichment results in excel format (tab delimited text)
- Guide to interpret results

13

3.SIZE: Number of genes in the gene set after filtering out those genes not in the expression dataset

4.ES: Enrichment score for the gene set; that is, the degree to which this gene set is overrepresented at the top or bottom of the ranked list of genes in the expression dataset.

5.NES: Normalized enrichment score; that is, the enrichment score for the gene set after it has been normalized across analyzed gene sets.

6.NOM p-val: Nominal p value; that is, the statistical significance of the enrichment score. The nominal p value is not adjusted for gene set size or multiple hypothesis testing; therefore, it is of limited use in comparing gene sets.

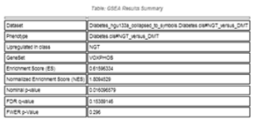
7.FDR q val: False discovery rate; that is, the estimated probability that the normalized enrichment score represents a false positive finding. (GSEA建議小於0.25)

Table: Gene sets enriched in phenotype NGT (17 samples) ***[plain text format]***

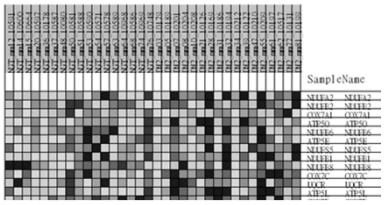
	1 GS follow link to MSigDB	2 GS DETAILS	3 SIZE	4 ES	5 NES	6 NOM p-val	7 FDR q-val	8 FWER p-val	9 RANK AT MAX	10 LEADING EDGE
1	P53_DOWN	→ Details ...	15	0.68	1.86	0.002	0.163	0.186	1988	tags=53%, list=15%, signal=63%
2	VOXPHOS	Details ...	77	0.62	1.81	0.016	0.154	0.296	3094	tags=62%, list=23%, signal=81%

14

Click “GS DETAILS” and you can see ...



GSEA Results Summary



Heat map

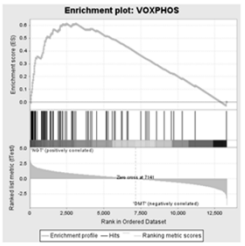


Fig 1: Enrichment plot: VOXPHOS
Profile of the Running ES Score & Positions of GeneSet Members on the Rank Ordered List

Enrichment plot

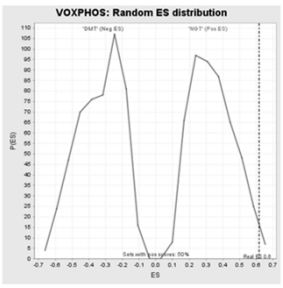
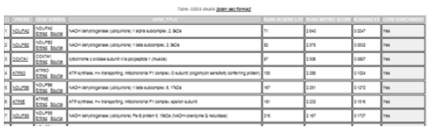


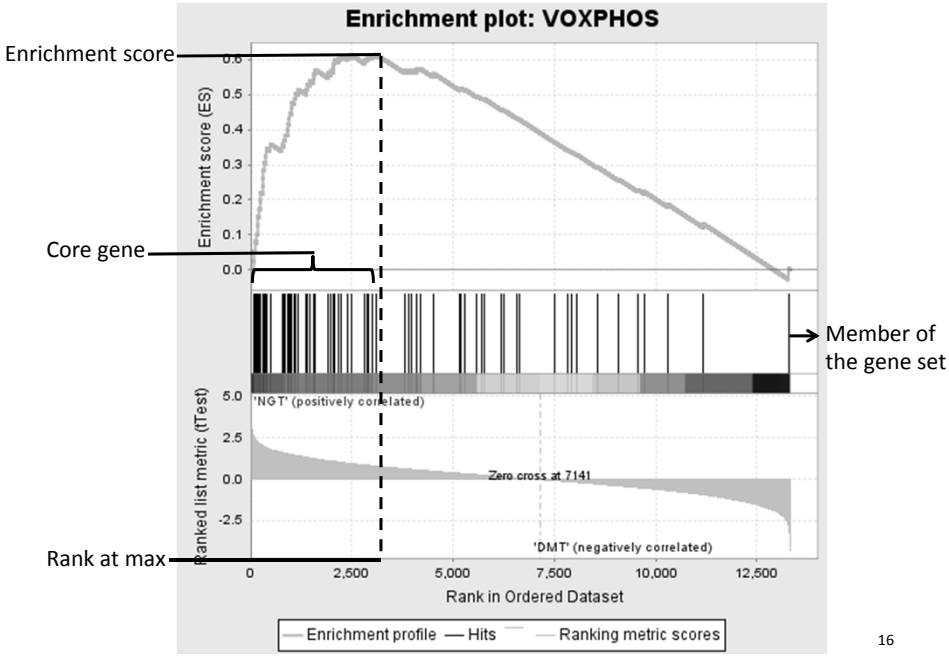
Fig 3: VOXPHOS: Random ES distribution
Gene set null distribution of ES for VOXPHOS

Random ES distribution



GSEA details

15



16

Reference

- *Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP.* (2005) Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. PNAS 102 (43) 15545-15550.
- http://software.broadinstitute.org/gsea/index.jsp

17

Bioinformatics Core Facility

Introduction to xCell

Chung-Hsing Chen (陳中興), Ph.D.
Institute of Population Health Sciences, NHRI

July 2019

Bioinformatics Core Facility

xCell

Aran *et al. Genome Biology* (2017) 18:220
DOI 10.1186/s13059-017-1349-1

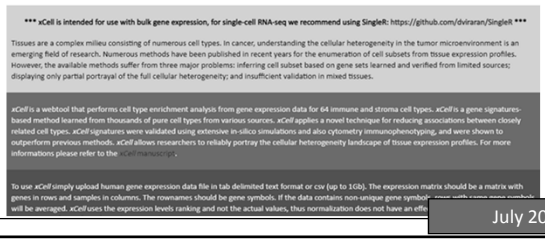
Genome Biology

METHOD Open Access

xCell: digitally portraying the tissue cellular heterogeneity landscape

Dvir Aran*, Zicheng Hu and Atul J. Butte*

- RNA-seq and Microarray


*** xCell is intended for use with bulk gene expression, for single-cell RNA-seq we recommend using SingleR: https://github.com/dviraran/SingleR ***

Tissues are a complex milieu consisting of numerous cell types. In cancer, understanding the cellular heterogeneity in the tumor microenvironment is an emerging field of research. Numerous methods have been published in recent years for the enumeration of cell subsets from tissue expression profiles. However, the available methods suffer from three major problems: inferring cell subset based on gene sets learned and verified from limited sources; displaying only partial portrayal of the full cellular heterogeneity; and insufficient validation in mixed tissues.

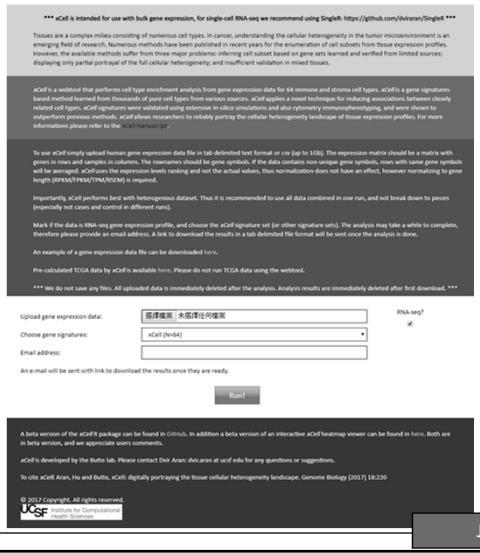
xCell is a webtool that performs cell type enrichment analysis from gene expression data for 64 immune and stroma cell types. xCell is a gene signatures-based method learned from thousands of pure cell types from various sources. xCell applies a novel technique for reducing associations between closely related cell types. xCell signatures were validated using extensive in-silico simulations and also cytometry immunophenotyping, and were shown to outperform previous methods. xCell allows researchers to reliably portray the cellular heterogeneity landscape of tissue expression profiles. For more informations please refer to the "xCell manuscript".

To use xCell simply upload human gene expression data file in tab delimited text format or csv (up to 1Gb). The expression matrix should be a matrix with genes in rows and samples in columns. The rownames should be gene symbols. If the data contains non-unique gene symbols, rows with same gene symbols will be averaged. xCell uses the expression levels ranking and not the actual values, thus normalization does not have an effe

July 2019

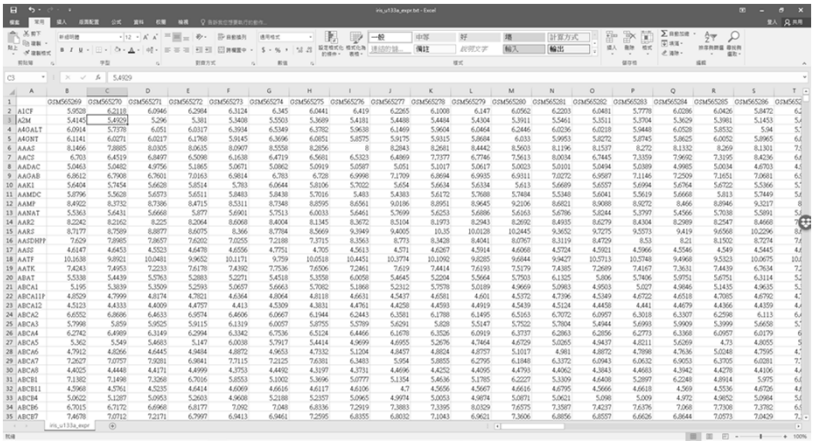
xCell website

- http://xcell.ucsf.edu/
- Need E-mail address



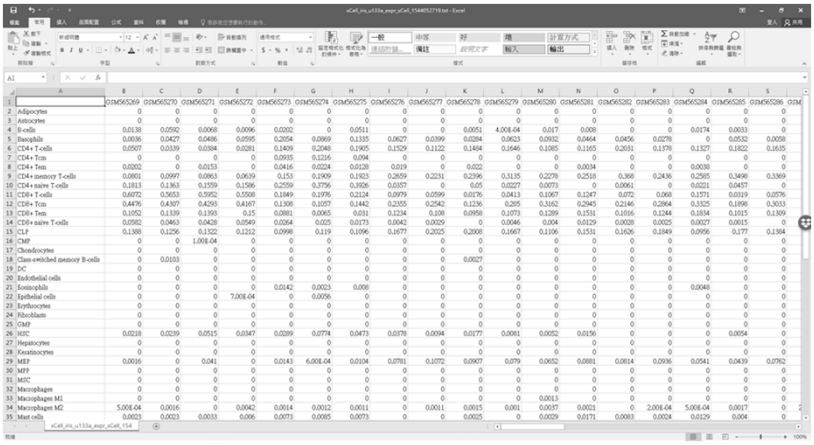
xCell - Input format

- Rows are genes
- Columns are samples



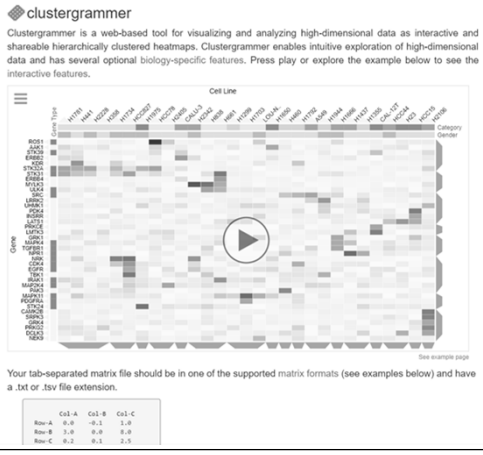
xCell - Output

- Rows are 64 cell types
- Columns are samples



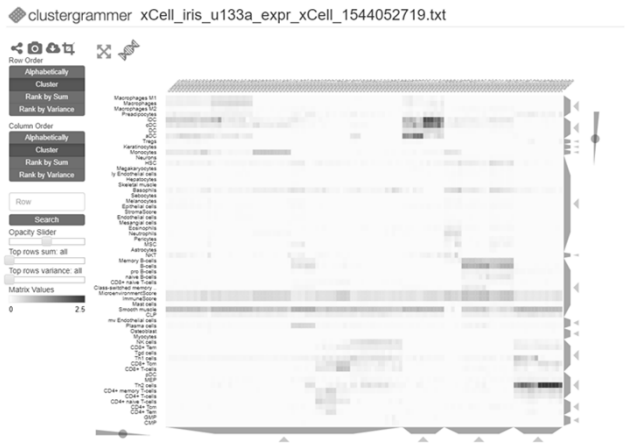
Clustergrammer

- http://amp.pharm.mssm.edu/clustergrammer/Columns are samples



Clustergrammer - output

- http://amp.pharm.mssm.edu/clustergrammer/Columns are samples



xCell Advantages

- Gene signatures are rank-based
  (agnostic to normalization methods; concerns related to batch effects)
- No decline in performance with increasing numbers of cell types
  (a partial portrayal of the TME may result in misleading findings)
- Gene signatures are simple and can easily be adjusted.