# RNA-sequencing: technical issues, data analysis, and utilization of public domain data

Workshop on Bioinformatics tools for NGS-data analysis

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

# 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?
- RNA-seq *versus* microarray.
- Why RNA-seq?

# Comparison of miRNA profiling studies in a PPG

| | A | B | C | D |
|---|---|---|---|---|
| **Samples** | Stool samples from cases of CRC (n=136) and health (n=266) | Polyps (n=10) | CRC tumor tissues (n=10) | CRC tumor tissue (n=44) and adjacent normal tissues (n=44) |
| **Description** | Biomarkers in stool samples for early detection and diagnosis | unclear | miRNAs differentially expressed between tumors of good response to CCRT (n=5) versus poor response (n=5) | miRNAs differentially expressed between tumor and normal tissues |
| **Platform** | Unclear (no show) | Unclear (no show) | miRNA array | Small RNA sequencing (NGS) |
| **Aim** | Early detection/diagnosis | unclear | Prognosis markers to predict CCRT response | Diagnosis |

# 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.

# 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<br>• Transcript-level abundance<br>• Splicing Analysis<br>• Fusion Discovery<br>• FFPE compatible<br>• Any species | • Coding RNA<br>• Transcript-level abundance<br>• Splicing Analysis<br>• Fusion Discovery | • Human Exonic RNA<br>• Transcript-level abundance<br>• Splicing Analysis<br>• Fusion Discovery<br>• FFPE Compatible | • 10s-1,000s of targets<br>• Max 384 spls<br>• Coding + ncRNA<br>• Transcript-level abundance<br>• Fusion Validation<br>• FFPE Compatible | • miRNA abundance<br>• isomiR detection |

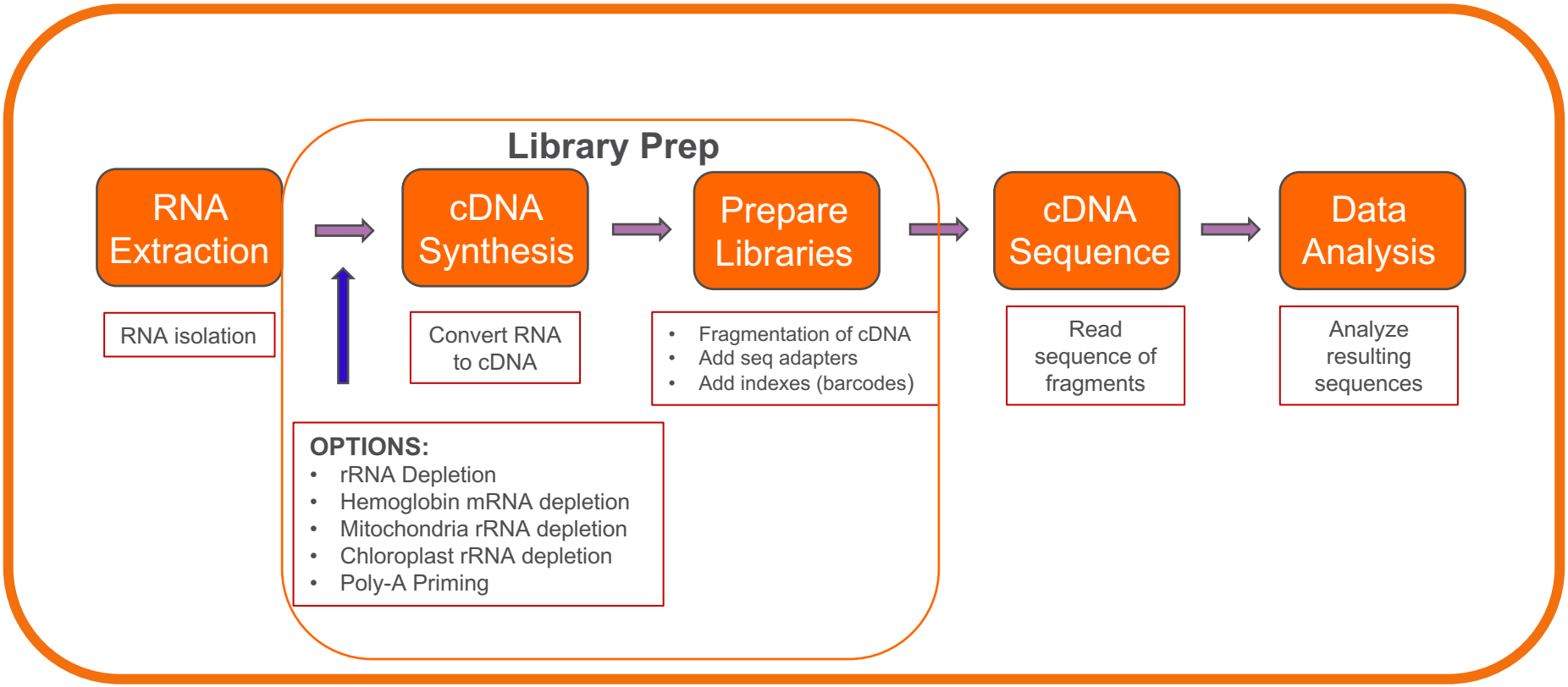
**For Research Use Only. Not for use in diagnostic procedures.**

# 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<br>HT: 96 dual | LT: Set A – 12; Set B – 12 Single<br>HT: 96 dual | Set A: 12; Set B – 12<br>4-plex enrichment |
| Applications | Expression and splicing analysis;<br>Variants – SNPs, Gene Fusions, Novel Transcripts, Transcript Variants | Expression and splicing analysis;<br>Variants – (SNPs), Gene Fusions, Novel Transcripts, Transcript Variants;<br>Gene expression profiling | Transcript-level abundance;<br>Splicing Analysis;<br>Fusion Discovery;<br>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)**<br>• High Quality RNA | 1x50 bp<br>1x75 bp | **≥10M** | 2 | 12 | 40 | 200 | - |
| **mRNA-Seq**<br>• High Quality RNA | 2x75 bp | **≥25M** | 1 | 5 | 16 | 80 | - |
| **Exome Enriched RNA-Seq**<br>• High Quality, degraded or FFPE | 2x75 bp | **≥25M** | 1 | 5 | 16 | 80 | - |
| **Total RNA-Seq**<br>• High Quality RNA | 2x75 bp | **≥50M** | | 2 | 8 | 40 | 66 |
| **Total RNA-Seq**<br>• Degraded RNA<br>• Less efficient rRNA Depletion | 2x75 bp | **≥100M** | | 1 | 4 | 20 | 33 |
| **Total RNA-Seq**<br>• FFPE Samples<br>• Less efficient rRNA Depletion<br>• Increased intronic sequences | 2x75 bp | **≥200M** | | | 2 | 10 | 16 |

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

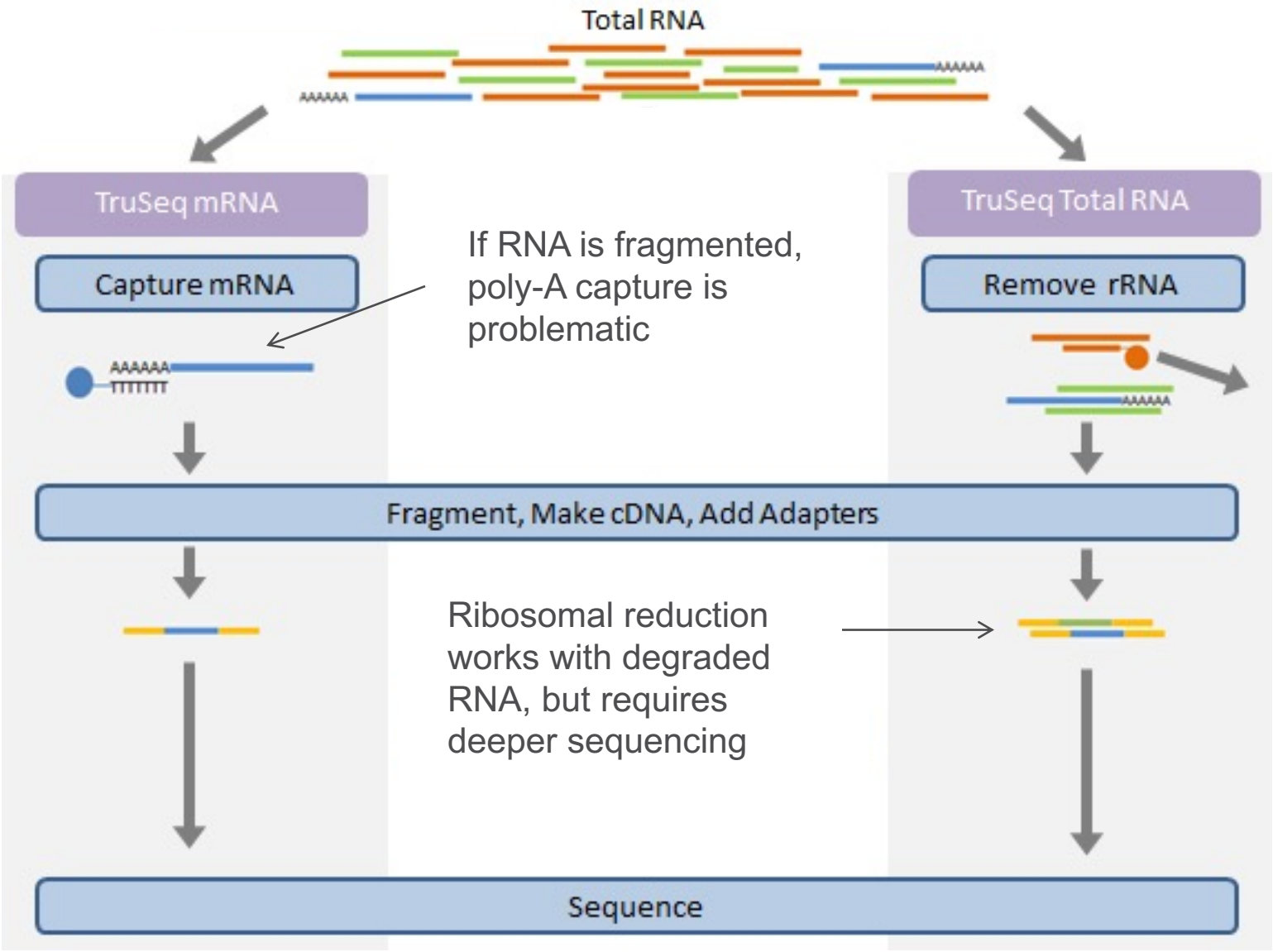
# Illumina's Current Suite of RNA Library Prep Solutions

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



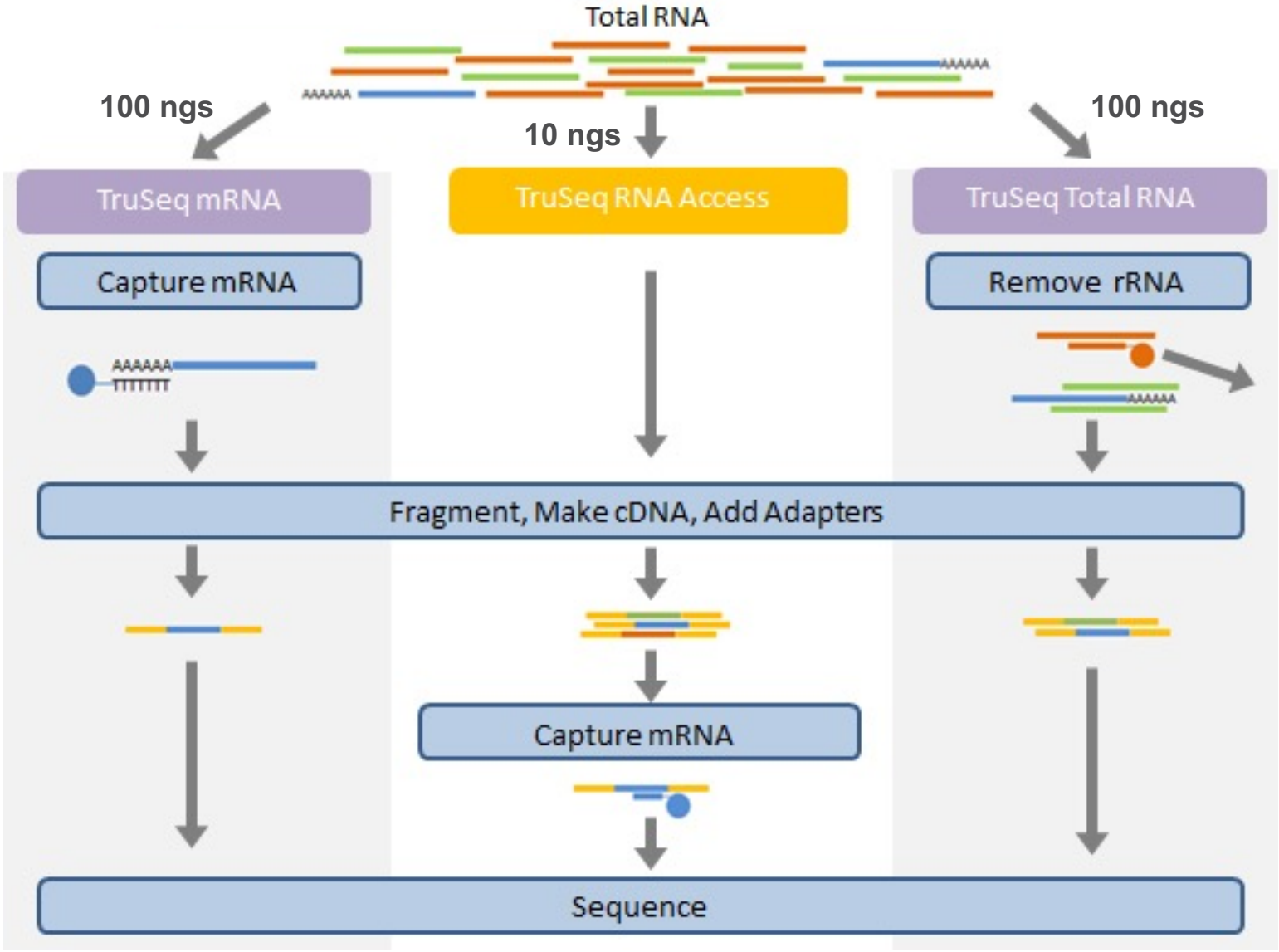



# TruSeq® RNA Library Prep Overview



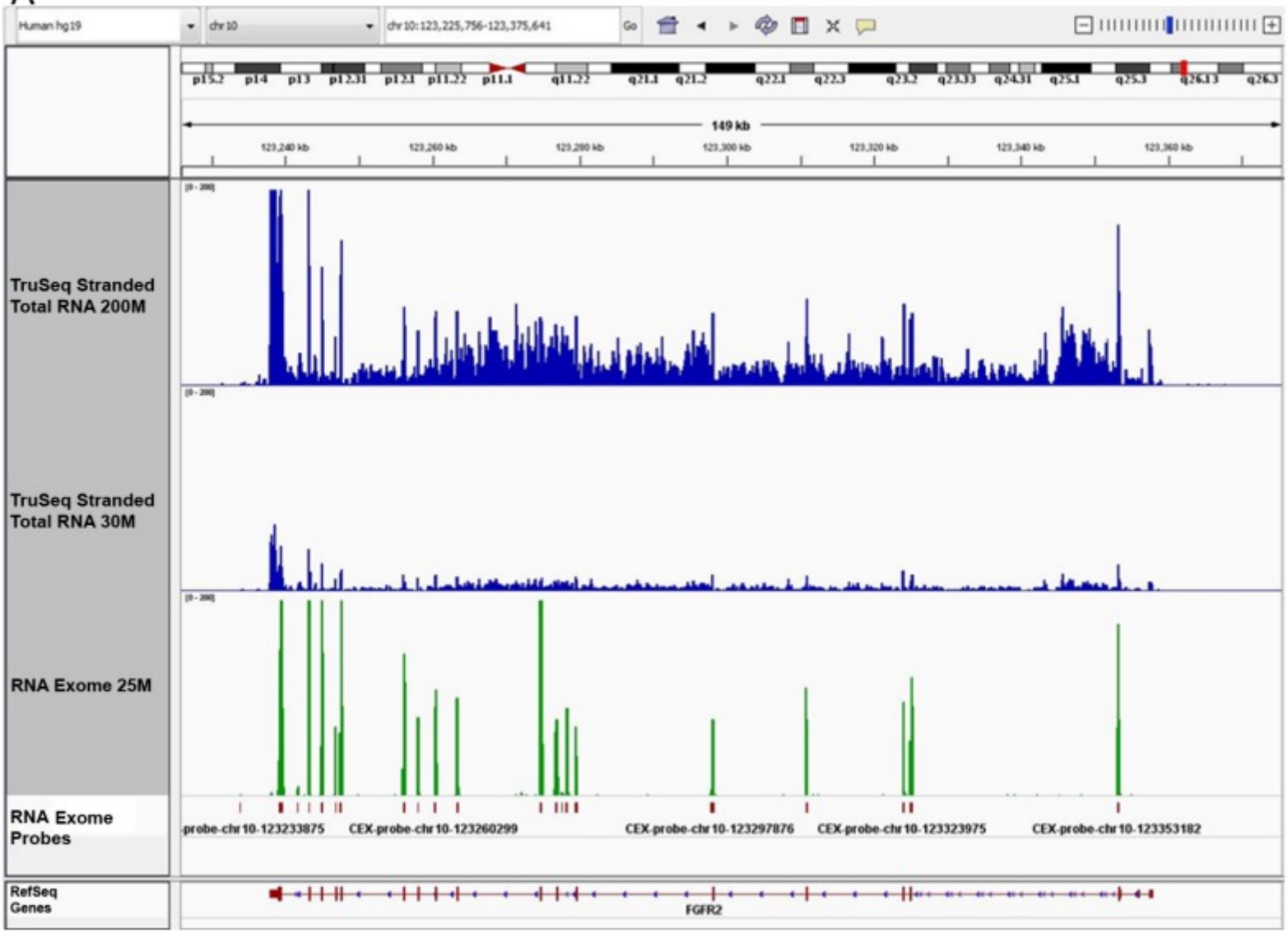

For Research Use Only. Not for use in diagnostic procedures.

# TruSeq® RNA Library Prep Overview




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

A


Human hg19
chr10
chr10:123,225,756-123,375,641
Go
p15.2 p14 p13 p12.31 p12.1 p11.22 p11.1 q11.22 q21.1 q21.2 q22.1 q22.3 q23.2 q23.33 q24.31 q25.1 q25.3 q26.13 q26.3
149 kb
123,240 kb
123,260 kb
123,280 kb
123,300 kb
123,320 kb
123,340 kb
123,360 kb
TruSeq Stranded Total RNA 200M
TruSeq Stranded Total RNA 30M
RNA Exome 25M
RNA Exome Probes
probe-chr10-123233875
CEX-probe-chr10-123260299
CEX-probe-chr10-123297876
CEX-probe-chr10-123323975
CEX-probe-chr10-123353182
RefSeq Genes
FGFR2

# Qiagen's UPX system

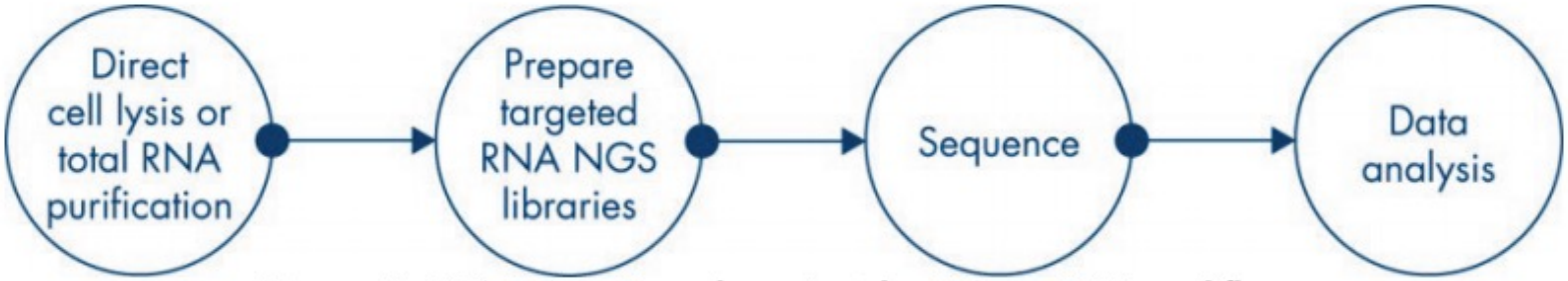



**Figure 1. QIAGEN's Sample to Insight QIAseq UPX workflow.**

# Qiagen's UPX system

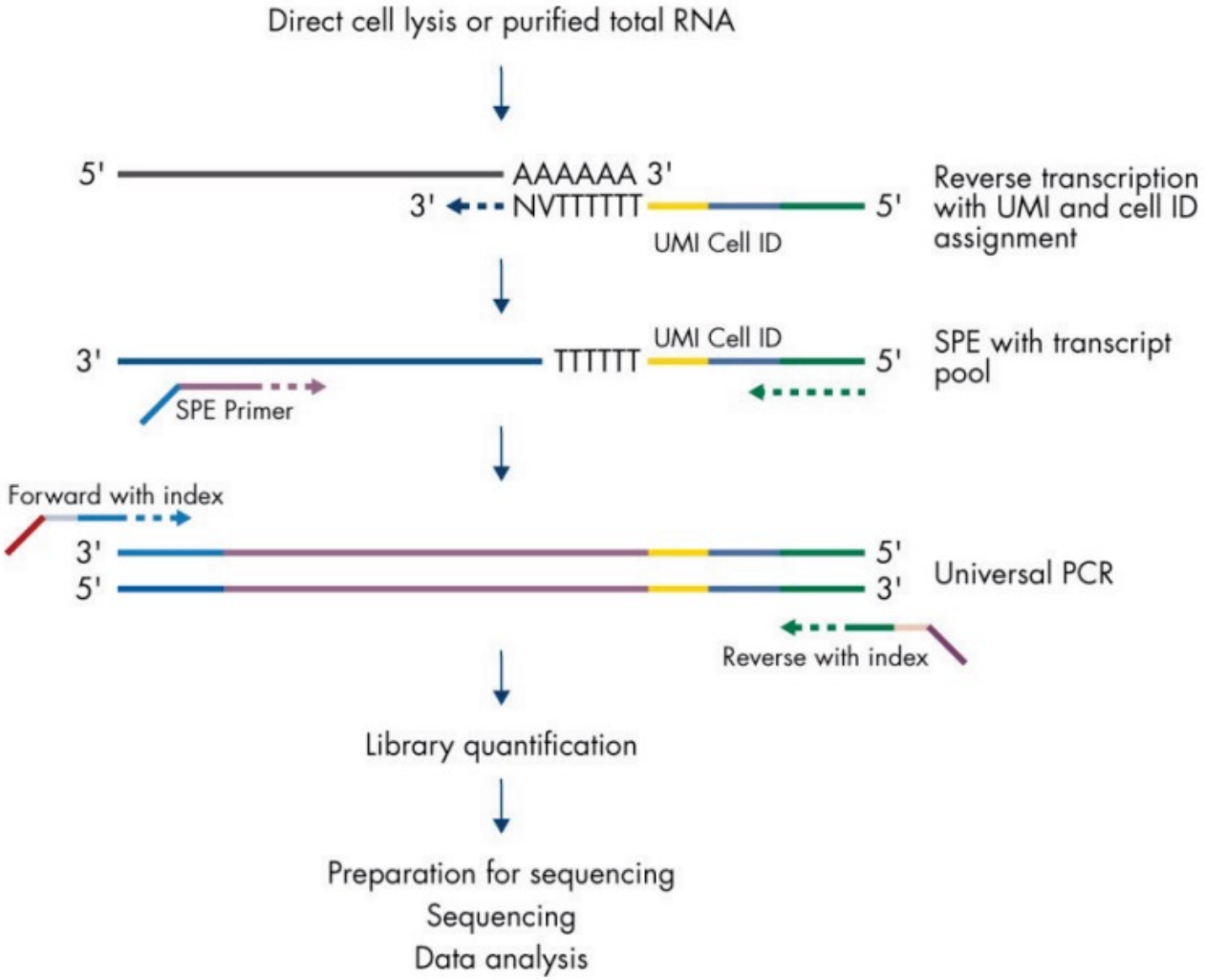
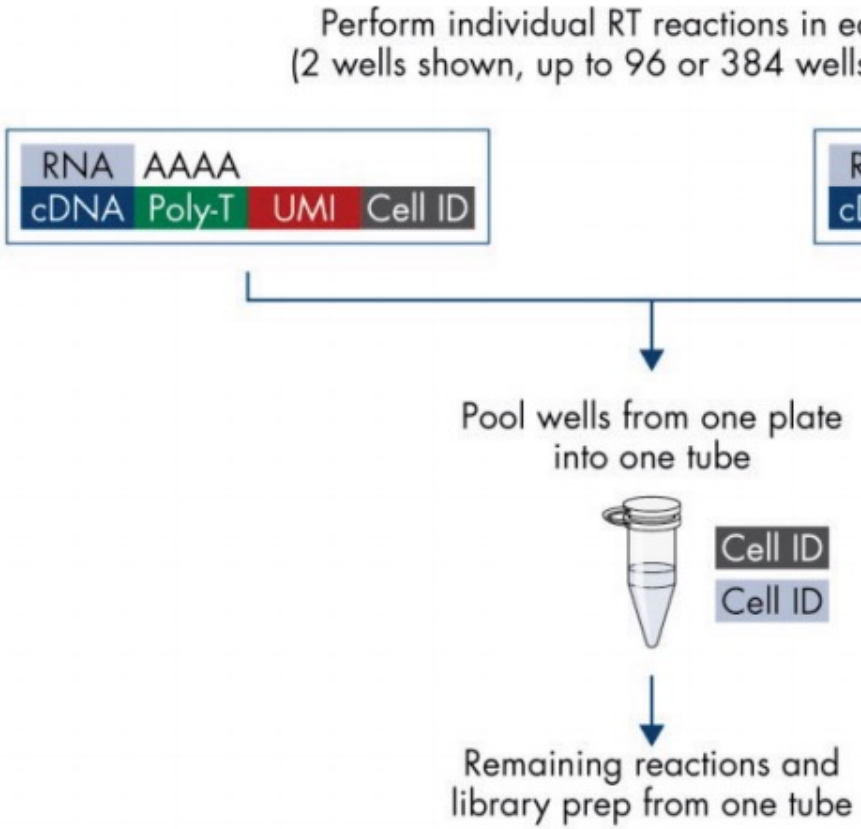



**Figure 2. QIAseq UPX 3' Targeted RNA Panel workflow.**

# Qiagen's UPX system




| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | C1 | C2 | C3 | C4 | C5 | C6 | C7 | C8 | C9 | C10 | C11 | C12 |
| B | C13 | C14 | C15 | C16 | C17 | C18 | C19 | C20 | C21 | C22 | C23 | C24 |
| C | C25 | C26 | C27 | C28 | C29 | C30 | C31 | C32 | C33 | C34 | C35 | C36 |
| D | C37 | C38 | C39 | C40 | C41 | C42 | C43 | C44 | C45 | C46 | C47 | C48 |
| E | C49 | C50 | C51 | C52 | C53 | C54 | C55 | C56 | C57 | C58 | C59 | C60 |
| F | C61 | C62 | C63 | C64 | C65 | C66 | C67 | C68 | C69 | C70 | C71 | C72 |
| G | C73 | C74 | C75 | C76 | C77 | C78 | C79 | C80 | C81 | C82 | C83 | C84 |
| H | C85 | C86 | C87 | C88 | C89 | C90 | C91 | C92 | C93 | C94 | C95 | C96 |

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | C1 | C2 | C3 | C4 | C5 | C6 | C7 | C8 | C9 | C10 | C11 | C12 | C13 | C14 | C15 | C16 | C17 | C18 | C19 | C20 | C21 | C22 | C23 | C24 |
| B | C25 | C26 | C27 | C28 | C29 | C30 | C31 | C32 | C33 | C34 | C35 | C36 | C37 | C38 | C39 | C40 | C41 | C42 | C43 | C44 | C45 | C46 | C47 | C48 |
| C | C49 | C50 | C51 | C52 | C53 | C54 | C55 | C56 | C57 | C58 | C59 | C60 | C61 | C62 | C63 | C64 | C65 | C66 | C67 | C68 | C69 | C70 | C71 | C72 |
| D | C73 | C74 | C75 | C76 | C77 | C78 | C79 | C80 | C81 | C82 | C83 | C84 | C85 | C86 | C87 | C88 | C89 | C90 | C91 | C92 | C93 | C94 | C95 | C96 |
| E | C97 | C98 | C99 | C100 | C101 | C102 | C103 | C104 | C105 | C106 | C107 | C108 | C109 | C110 | C111 | C112 | C113 | C114 | C115 | C116 | C117 | C118 | C119 | C120 |
| F | C121 | C122 | C123 | C124 | C125 | C126 | C127 | C128 | C129 | C130 | C131 | C132 | C133 | C134 | C135 | C136 | C137 | C138 | C139 | C140 | C141 | C142 | C143 | C144 |
| G | C145 | C146 | C147 | C148 | C149 | C150 | C151 | C152 | C153 | C154 | C155 | C156 | C157 | C158 | C159 | C160 | C161 | C162 | C163 | C164 | C165 | C166 | C167 | C168 |
| H | C169 | C170 | C171 | C172 | C173 | C174 | C175 | C176 | C177 | C178 | C179 | C180 | C181 | C182 | C183 | C184 | C185 | C186 | C187 | C188 | C189 | C190 | C191 | C192 |
| I | C193 | C194 | C195 | C196 | C197 | C198 | C199 | C200 | C201 | C202 | C203 | C204 | C205 | C206 | C207 | C208 | C209 | C210 | C211 | C212 | C213 | C214 | C215 | C216 |
| J | C217 | C218 | C219 | C220 | C221 | C222 | C223 | C224 | C225 | C226 | C227 | C228 | C229 | C230 | C231 | C232 | C233 | C234 | C235 | C236 | C237 | C238 | C239 | C240 |
| K | C241 | C242 | C243 | C244 | C245 | C246 | C247 | C248 | C249 | C250 | C251 | C252 | C253 | C254 | C255 | C256 | C257 | C258 | C259 | C260 | C261 | C262 | C263 | C264 |
| L | C265 | C266 | C267 | C268 | C269 | C270 | C271 | C272 | C273 | C274 | C275 | C276 | C277 | C278 | C279 | C280 | C281 | C282 | C283 | C284 | C285 | C286 | C287 | C288 |
| M | C289 | C290 | C291 | C292 | C293 | C294 | C295 | C296 | C297 | C298 | C299 | C300 | C301 | C302 | C303 | C304 | C305 | C306 | C307 | C308 | C309 | C310 | C311 | C312 |
| N | C313 | C314 | C315 | C316 | C317 | C318 | C319 | C320 | C321 | C322 | C323 | C324 | C325 | C326 | C327 | C328 | C329 | C330 | C331 | C332 | C333 | C334 | C335 | C336 |
| O | C337 | C338 | C339 | C340 | C341 | C342 | C343 | C344 | C345 | C346 | C347 | C348 | C349 | C350 | C351 | C352 | C353 | C354 | C355 | C356 | C357 | C358 | C359 | C360 |
| P | C361 | C362 | C363 | C364 | C365 | C366 | C367 | C368 | C369 | C370 | C371 | C372 | C373 | C374 | C375 | C376 | C377 | C378 | C379 | C380 | C381 | C382 | C383 | C384 |

# What after RNA-seq

- Sequence 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.

# 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.

# 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

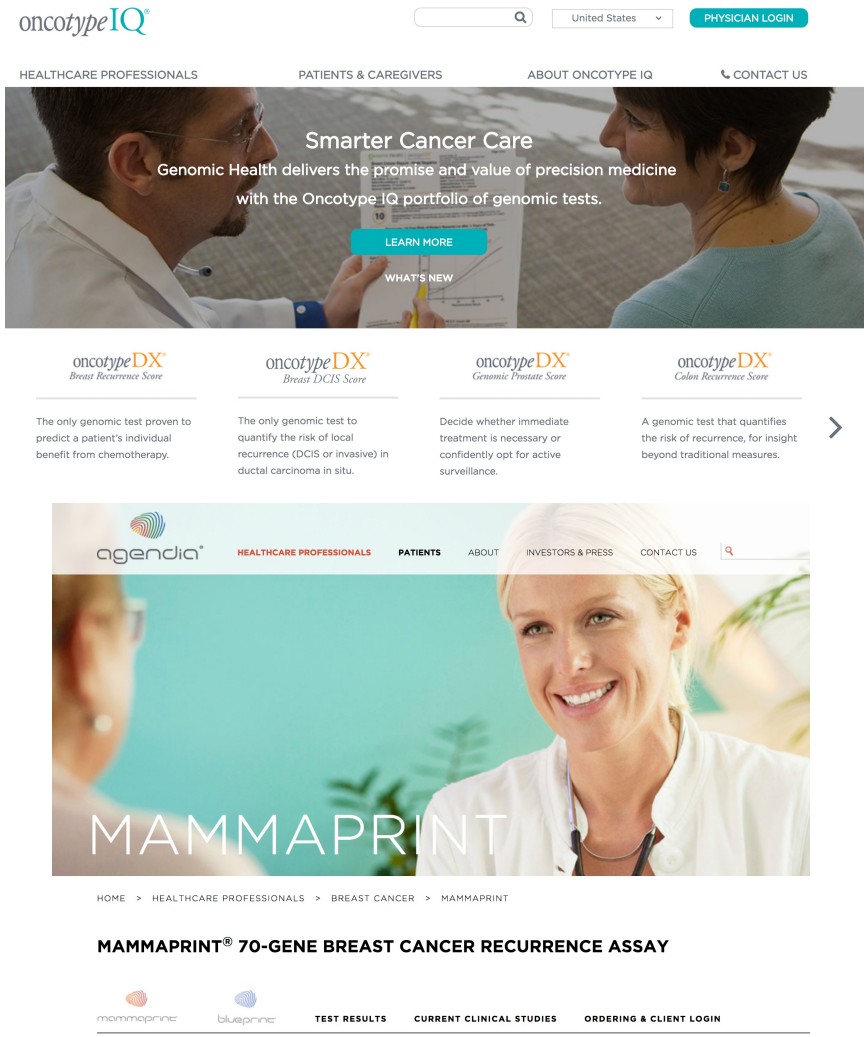
**FFPE samples-cDNA**

# 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

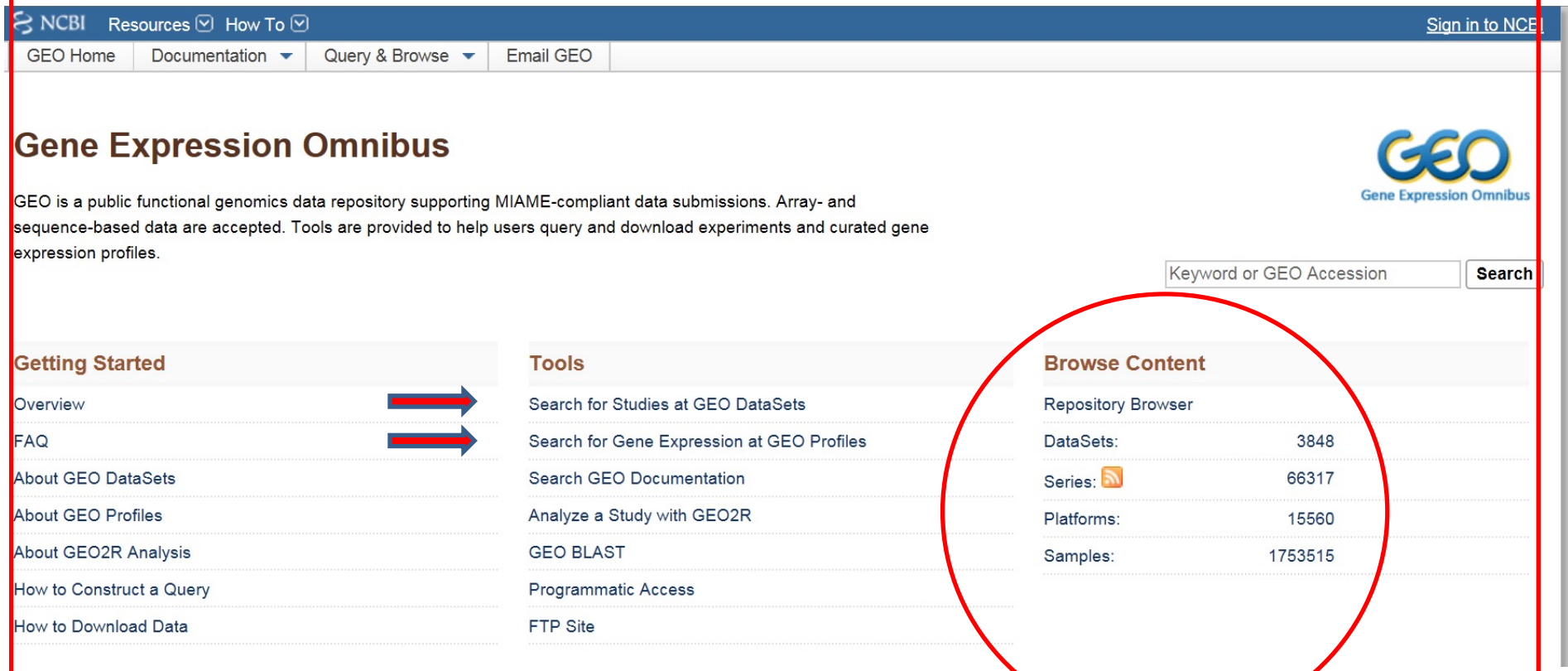

oncotypeIQ
United States
PHYSICIAN LOGIN
HEALTHCARE PROFESSIONALS
PATIENTS & CAREGIVERS
ABOUT ONCOTYPE IQ
CONTACT US
Smarter Cancer Care
Genomic Health delivers the promise and value of precision medicine with the Oncotype IQ portfolio of genomic tests.
LEARN MORE
WHAT'S NEW
oncotypeDX Breast Recurrence Score
The only genomic test proven to predict a patient's individual benefit from chemotherapy.
oncotypeDX Breast DCIS Score
The only genomic test to quantify the risk of local recurrence (DCIS or invasive) in ductal carcinoma in situ.
oncotypeDX Genomic Prostate Score
Decide whether immediate treatment is necessary or confidently opt for active surveillance.
oncotypeDX Colon Recurrence Score
A genomic test that quantifies the risk of recurrence, for insight beyond traditional measures.
agendia
HEALTHCARE PROFESSIONALS
PATIENTS
ABOUT
INVESTORS & PRESS
CONTACT US
MAMMAPRINT
HOME > HEALTHCARE PROFESSIONALS > BREAST CANCER > MAMMAPRINT
MAMMAPRINT® 70-GENE BREAST CANCER RECURRENCE ASSAY
mammaprint
blueprint
TEST RESULTS
CURRENT CLINICAL STUDIES
ORDERING & CLIENT LOGIN


CE Product Summary Sheet
The Prosigna™ Breast Cancer Prognostic Gene Signature
Product Highlights
• PAM50 based report of intrinsic subtype
• The only genomic breast cancer assay that can be run in your pathology lab
• The only test that is CE marked and FDA 510(k) cleared for FFPE tissue
Prosigna is an in vitro diagnostic assay that uses the gene expression profile of cells found in breast tissue to assess a patient's risk of distant recurrence. The assay measures the gene expression profile using RNA extracted from Formalin-Fixed, Paraffin-Embedded (FFPE) breast tumor tissue.
The gene expression data are weighted together with clinical variables to generate both a subtype (luminal A, luminal B, Her2-enriched, or basal-like) and a score indicative of the probability of distant recurrence of disease. The assay is performed on the NanoString nCounter® Analysis System using FFPE breast tumor tissue previously diagnosed as invasive breast carcinoma.
CE marked. FDA 510(k) cleared.
prosigna™
Breast cancer gene signature assay

# What after RNA-seq

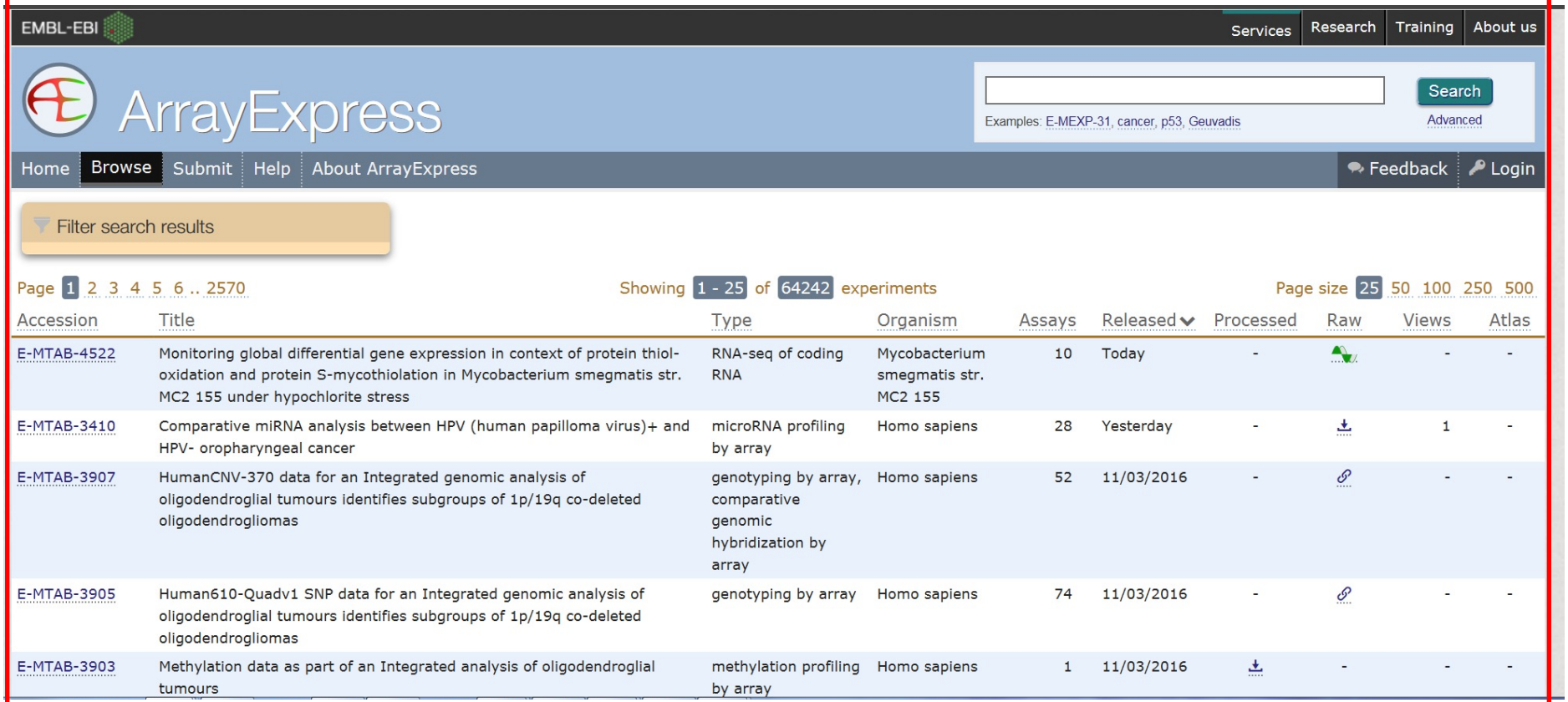
- Sequencing alignment.
- Data preprocessing.
- Basic data output (FPKM, RPKM).
- Validation at individual genes/transcripts level.
- Searching of supporting evidences/consistent findings in the public domain.
- Validation/annotation at pathway level.

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



# ArrayExpress

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



EMBL-EBI

Services | Research | Training | About us

ArrayExpress

Search

Examples: E-MEXP-31, cancer, p53, Geuvadis

Advanced

Home | Browse | Submit | Help | About ArrayExpress | Feedback | Login

Filter search results

Page 1 2 3 4 5 6 .. 2570

Showing 1 - 25 of 64242 experiments

Page size 25 50 100 250 500

| Accession | Title | Type | Organism | Assays | Released | Processed | Raw | Views | Atlas |
|---|---|---|---|---|---|---|---|---|---|
| E-MTAB-4522 | Monitoring global differential gene expression in context of protein thiol-oxidation and protein S-mycothiolation in Mycobacterium smegmatis str. MC2 155 under hypochlorite stress | RNA-seq of coding RNA | Mycobacterium smegmatis str. MC2 155 | 10 | Today | - | | - | - |
| E-MTAB-3410 | Comparative miRNA analysis between HPV (human papilloma virus)+ and HPV- oropharyngeal cancer | microRNA profiling by array | Homo sapiens | 28 | Yesterday | - | | 1 | - |
| E-MTAB-3907 | HumanCNV-370 data for an Integrated genomic analysis of oligodendroglial tumours identifies subgroups of 1p/19q co-deleted oligodendrogliomas | genotyping by array, comparative genomic hybridization by array | Homo sapiens | 52 | 11/03/2016 | - | | - | - |
| E-MTAB-3905 | Human610-Quadv1 SNP data for an Integrated genomic analysis of oligodendroglial tumours identifies subgroups of 1p/19q co-deleted oligodendrogliomas | genotyping by array | Homo sapiens | 74 | 11/03/2016 | - | | - | - |
| E-MTAB-3903 | Methylation data as part of an Integrated analysis of oligodendroglial tumours | methylation profiling by array | Homo sapiens | 1 | 11/03/2016 | | - | - | - |

# 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

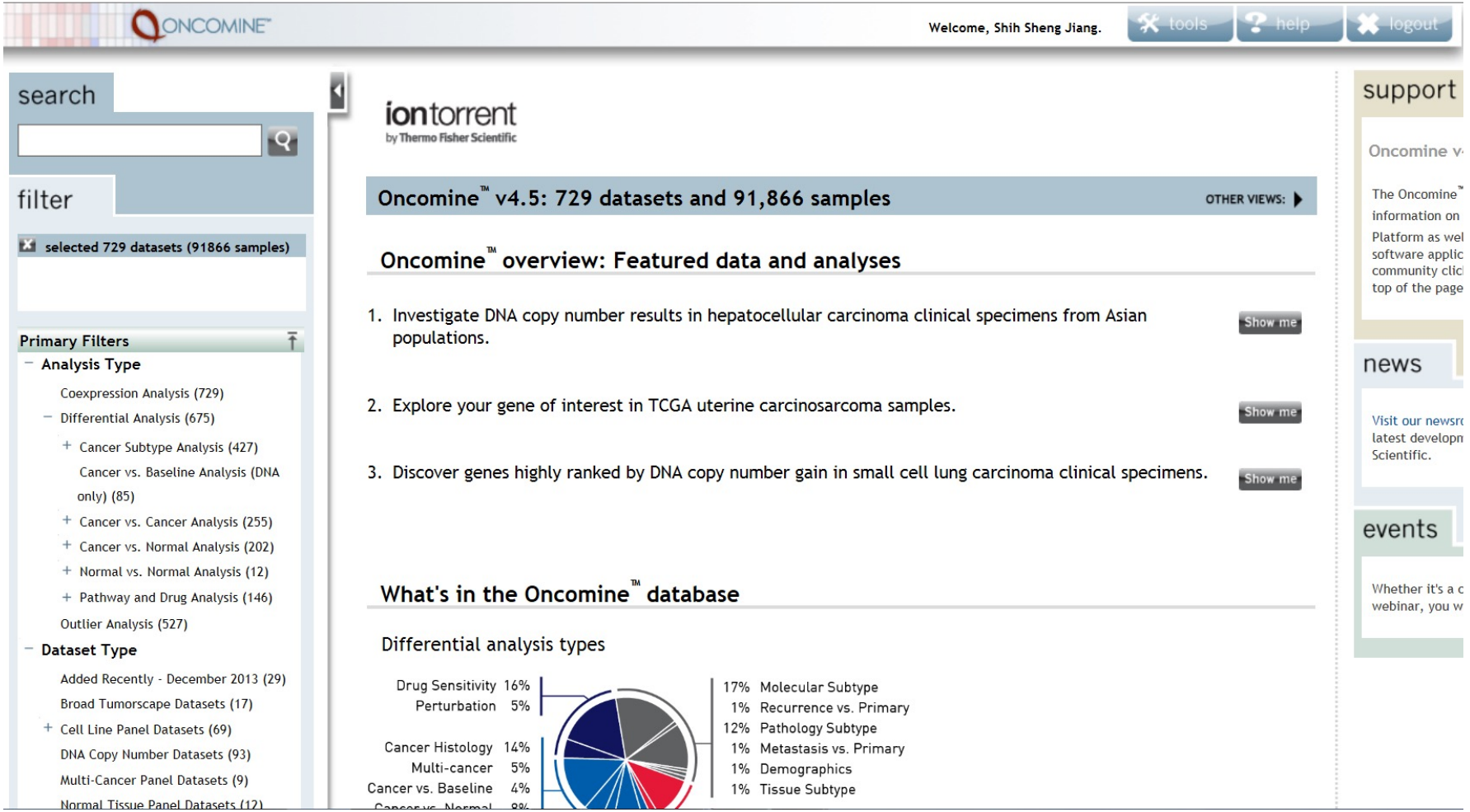
Below is a list of cancers that have been selected for study based on the criteria outlined above. The TCGA program has collected the necessary quality and quantity of samples for these cancers to move them into the TCGA project pipeline. Scheduling for sequencing and characterization for these cancers 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.

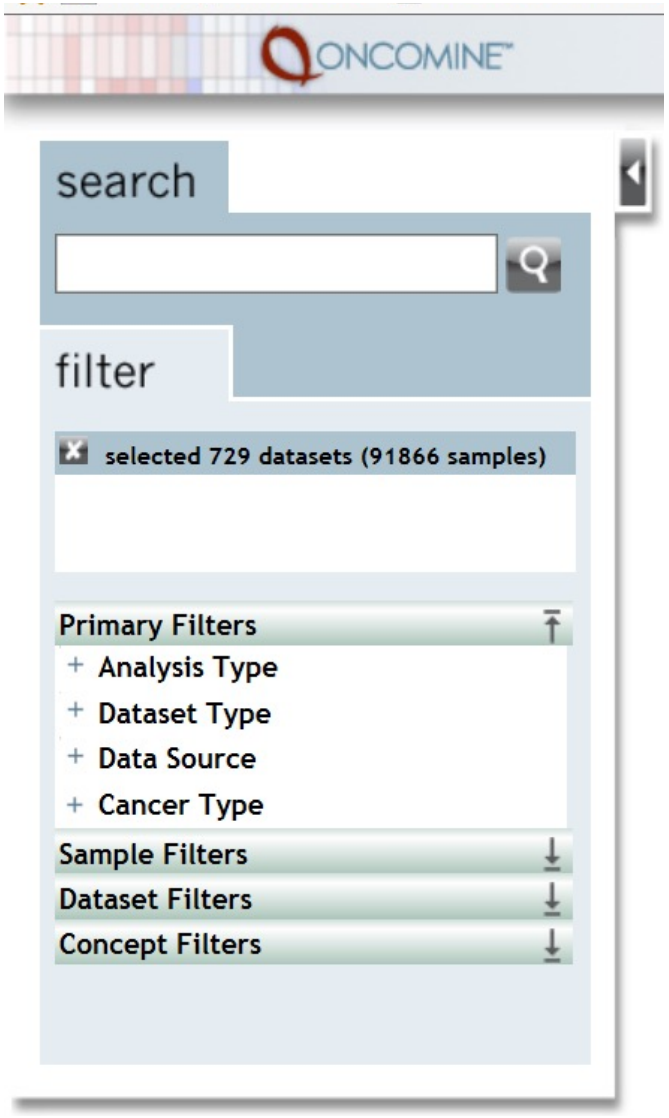
TCGA is no longer accepting additional samples for characterization.

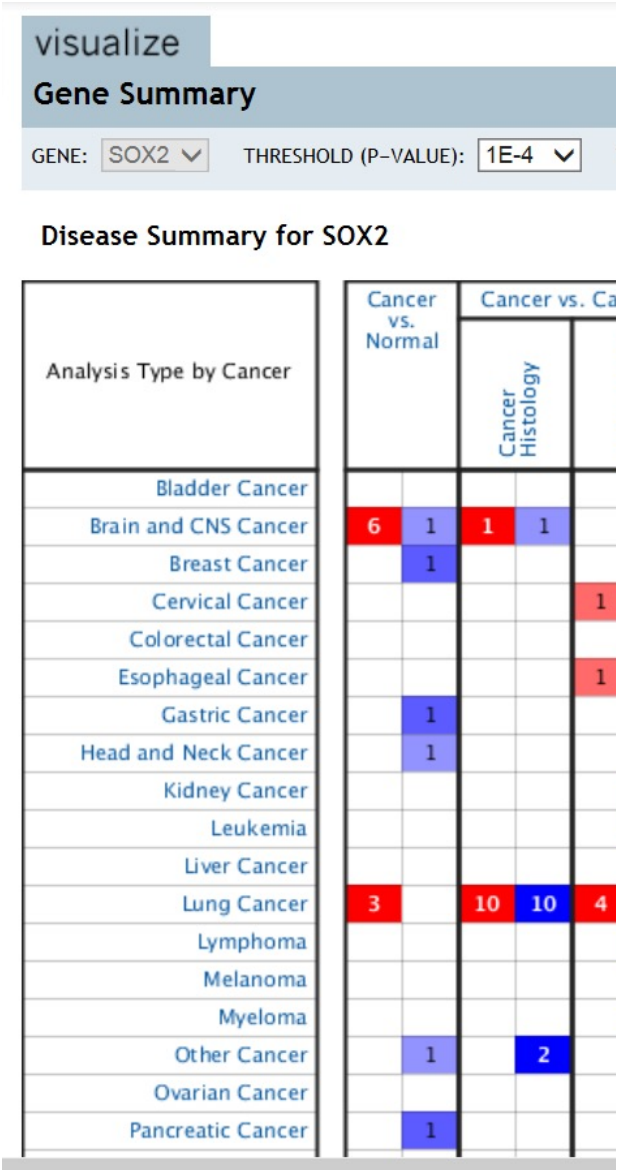
Analysis is underway in the following Analysis Working Groups (AWGs):

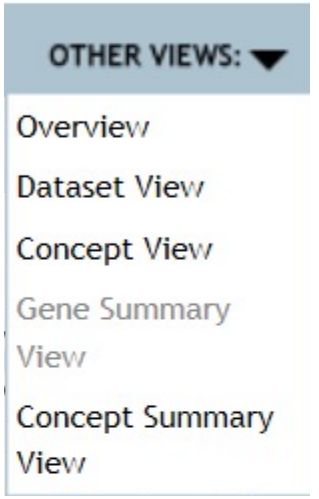
- Cervical cancer
- Cholangiocarcinoma
- Esophageal carcinoma
- Liver hepatocellular carcinoma
- Mesothelioma
- Pancreatic ductal adenocarcinoma
- Paraganglioma & Pheochromocytoma
- Sarcoma
- Testicular germ cell cancer
- Thymoma
- Uterine carcinosarcoma
- Uveal melanoma

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)
Normal Tissue Panel Datasets (12)
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
Oncomine v
The Oncomine™
information on
Platform as wel
software applic
community clic
top of the page
news
Visit our newsr
latest developn
Scientific.
events
Whether it's a c
webinar, you w


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 6 1 1 1
Breast Cancer 1
Cervical Cancer 1
Colorectal Cancer
Esophageal Cancer 1
Gastric Cancer 1
Head and Neck Cancer 1
Kidney Cancer
Leukemia
Liver Cancer
Lung Cancer 3 10 10 4
Lymphoma
Melanoma
Myeloma
Other Cancer 1 2
Ovarian Cancer
Pancreatic Cancer 1


OTHER VIEWS:
Overview
Dataset View
Concept View
Gene Summary View
Concept Summary View

UNIVERSITY OF CALIFORNIA
SANTA CRUZ

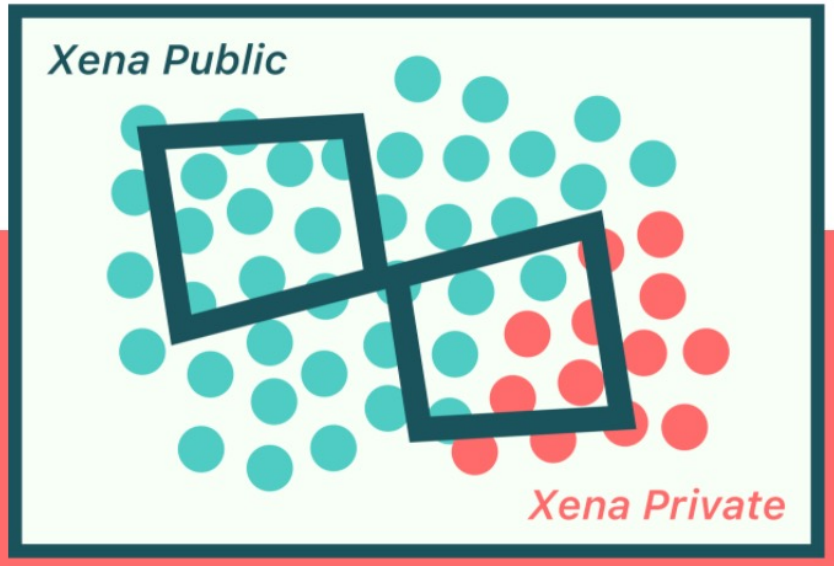
Overview Analysis Tutorials What's New Cite Us Subscribe

# UCSC Xena

*See the bigger picture*

An online exploration tool for public and private, multi-omic and clinical/phenotype data

Launch Xena




## Tutorials and walkthroughs

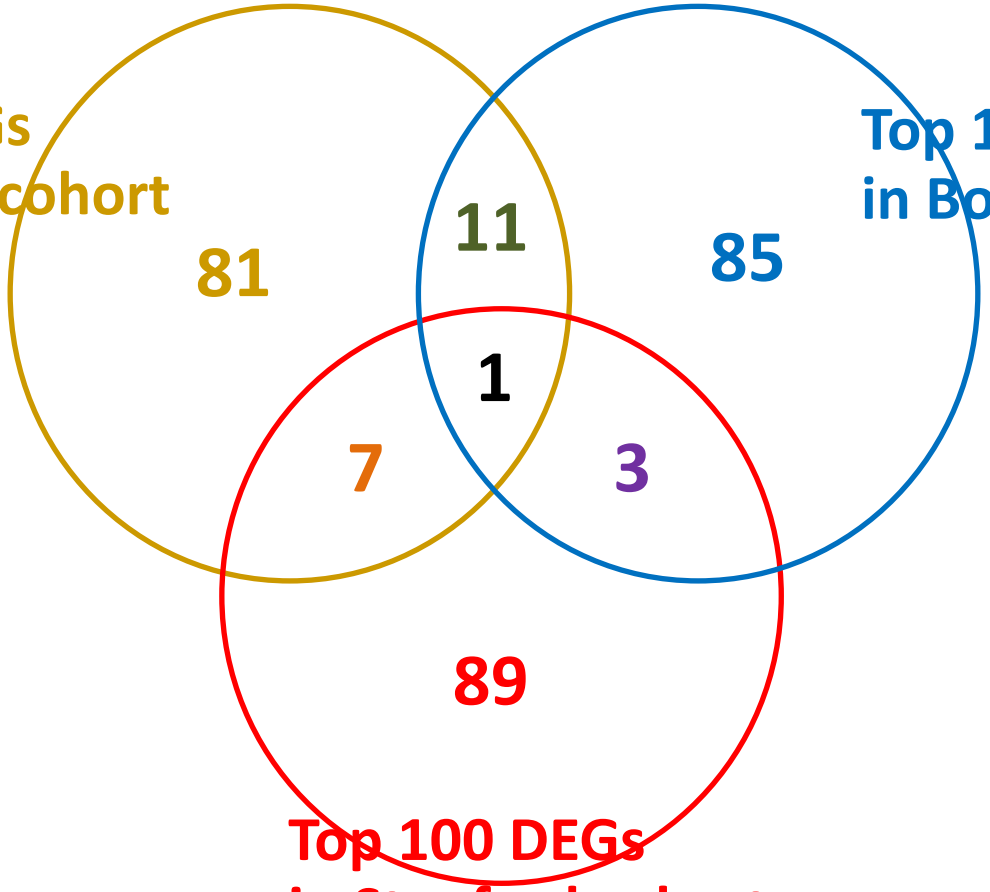
Don't know where to start? Jump in with one of our tutorials or "How do I ..." walkthroughs

Tutorials

# What after RNA-seq

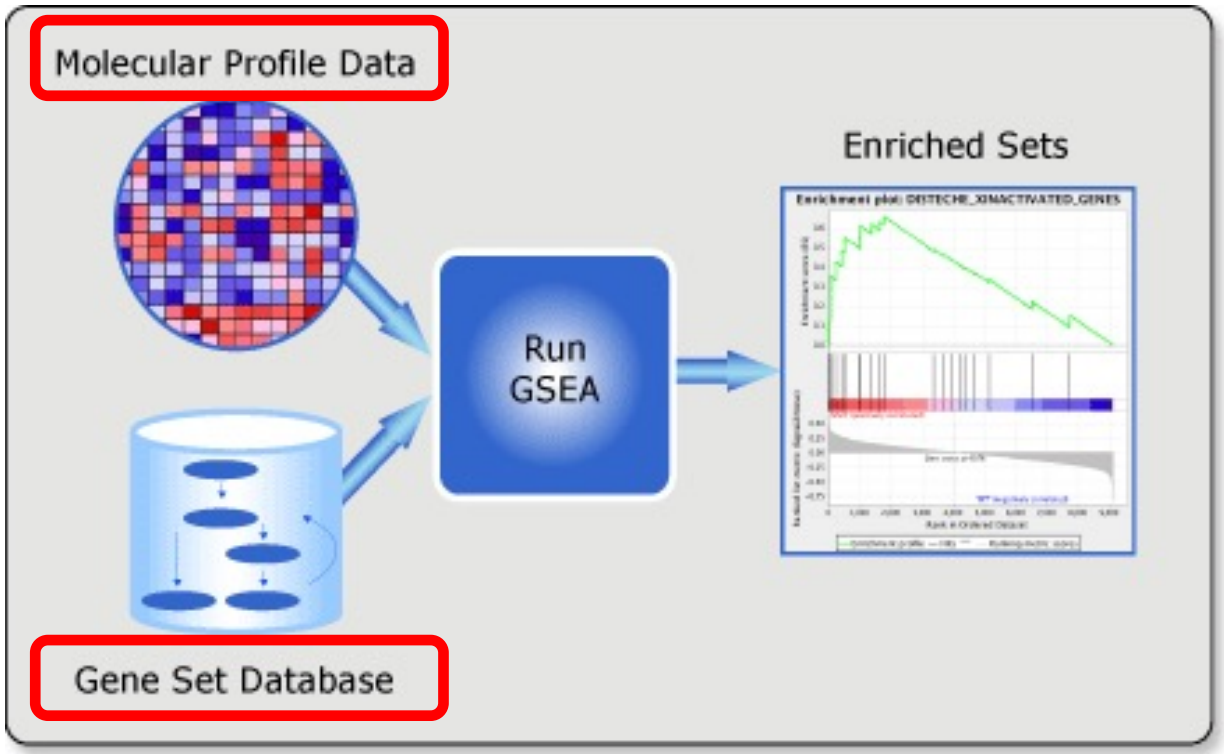
- Sequencing alignment.
- Data preprocessing.
- Basic data output (FPKM, RPKM).
- Validation at individual genes/transcripts level.
- Searching of supporting evidences/consistent findings in the public domain.
- Validation/annotation at pathway level.

# Findings at gene level



# Gene set enrichment analysis

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




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

# 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**

Cell Systems
Report

CellPress

# The Molecular Signatures Database Hallmark Gene Set Collection

**Arthur Liberzon,[1] Chet Birger,[1] Helga Thorvaldsdóttir,[1] Mahmoud Ghandi,[1] Jill P. Mesirov,[1,2,3,4,*] and Pablo Tamayo[1,2,3,4,*]**
[1]Broad Institute of MIT and Harvard, 415 Main Street, Cambridge, MA 02142, USA
[2]Department of Medicine, University of California, San Diego, La Jolla, CA 92093, USA
[3]Moores Cancer Center, University of California, San Diego, La Jolla, CA 92093, USA
[4]Co-senior author
*Correspondence: jmesirov@ucsd.edu (J.P.M.), ptamayo@ucsd.edu (P.T.)
http://dx.doi.org/10.1016/j.cels.2015.12.004

**Table 1. Summary of the Hallmark Gene Sets: Name, Process Category, Description, Number of Founder Sets, and Number of Genes It Contains**

| | Hallmark Name | Process Category | Description | Number of Founder Sets | Number of Genes |
|---|---|---|---|---|---|
| 1 | APICAL_JUNCTION | cellular component | apical junction complex consisting of adherens and tight junctions | 37 | 200 |
| 2 | APICAL_SURFACE | cellular component | membrane proteins in the apical domain | 12 | 44 |
| 3 | PEROXISOME | cellular component | peroxisomes | 28 | 107 |
| 4 | ADIPOGENESIS | development | adipocyte development | 36 | 200 |
| 5 | ANGIOGENESIS | development | blood vessel formation | 14 | 36 |
| 6 | EPITHELIAL_MESENCHIMAL_TRANSITION | development | epithelial mesenchymal transition | 107 | 200 |
| 7 | MYOGENESIS | development | muscle differentiation | 64 | 200 |
| 8 | SPERMATOGENESIS | development | sperm development and male fertility | 24 | 135 |
| 9 | PANCREAS_BETA_CELL | development | genes specific to pancreatic beta cells | 24 | 40 |
| 10 | DNA_REPAIR | DNA damage | DNA repair | 44 | 150 |
| 11 | UV_RESPONSE_DOWN | DNA damage | UV response: downregulated genes | 17 | 144 |
| 12 | UV_RESPONSE_UP | DNA damage | UV response: upregulated genes | 16 | 158 |
| 13 | ALLOGRAFT_REJECTION | immune | allograft rejection | 190 | 200 |
| 14 | COAGULATION | immune | blood coagulation cascade | 71 | 138 |
| 15 | COMPLEMENT | immune | complement cascade | 71 | 200 |
| 16 | INTERFERON_ALPHA_RESPONSE | immune | interferon alpha response | 82 | 97 |
| 17 | INTERFERON_GAMMA_RESPONSE | immune | interferon gamma response | 82 | 200 |
| 18 | IL6_JAK_STAT3_SIGNALING | immune | IL6 STAT3 signaling during acute phase response | 24 | 87 |
| 19 | INFLAMMATORY_RESPONSE | immune | inflammation | 120 | 200 |
| 20 | BILE_ACID_METABOLISM | metabolic | biosynthesis of bile acids | 28 | 112 |
| 21 | CHOLESTEROL_HOMEOSTASIS | metabolic | cholesterol homeostasis | 28 | 74 |

| | | | | | |
|---|---|---|---|---|---|
| 20 | BILE_ACID_METABOLISM | metabolic | biosynthesis of bile acids | 28 | 112 |
| 21 | CHOLESTEROL_HOMEOSTASIS | metabolic | cholesterol homeostasis | 28 | 74 |
| 22 | FATTY_ACID_METABOLISM | metabolic | fatty acid metabolism | 53 | 158 |
| 23 | GLYCOLYSIS | metabolic | glycolysis and gluconeogenesis | 87 | 200 |
| 24 | HEME_METABOLISM | metabolic | heme metabolism and erythroid lineage | 36 | 200 |
| 25 | OXIDATIVE_PHOSPHORYLATION | metabolic | oxidative phosphorylation and citric acid cycle | 93 | 200 |
| 26 | XENOBIOTIC_METABOLISM | metabolic | metabolism of xenobiotics | 124 | 200 |
| 27 | APOPTOSIS | pathway | programmed cell death; caspase pathway | 80 | 161 |
| 28 | HYPOXIA | pathway | response to hypoxia; HIF1A targets | 87 | 200 |
| 29 | PROTEIN_SECRETION | pathway | protein secretion | 74 | 96 |
| 30 | UNFOLDED_PROTEIN_RESPONSE | pathway | unfolded protein response; ER stress | 22 | 113 |
| 31 | REACTIVE_OXYGEN_SPECIES_PATHWAY | pathway | reactive oxygen species pathway | 13 | 49 |
| 32 | E2F targets | proliferation | cell cycle progression: E2F targets | 420 | 200 |
| 33 | G2M_checkpoint | proliferation | cell cycle progression: G2/M checkpoint | 420 | 200 |
| 34 | MYC_TARGETS_V1 | proliferation | MYC targets, variant 1 | 404 | 200 |
| 35 | MYC_TARGETS_V2 | proliferation | MYC targets, variant 2 | 6 | 58 |
| 36 | P53_PATHWAY | proliferation | p53 pathway | 85 | 200 |
| 37 | MITOTIC_SPINDLE | proliferation | cell cycle progression: mitotic spindle assembly | 108 | 200 |
| 38 | ANDROGEN_RESPONSE | signaling | androgen response | 8 | 117 |
| 39 | ESTROGEN_RESPONSE_EARLY | signaling | early estrogen response | 61 | 200 |
| 40 | ESTROGEN_RESPONSE_LATE | signaling | late estrogen response | 61 | 200 |
| 41 | IL2_JAK_STAT5_SIGNALING | signaling | IL2 STAT5 signaling | 13 | 200 |
| 42 | KRAS_SIGNALING_UP | signaling | KRAS signaling, upregulated genes | 14 | 200 |

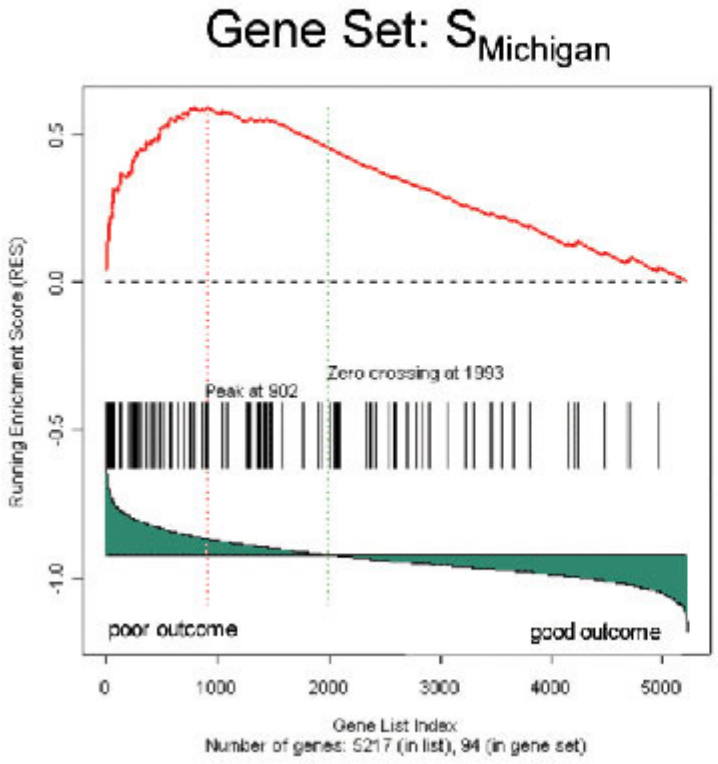
*(Continued on next page)*

# GSEA

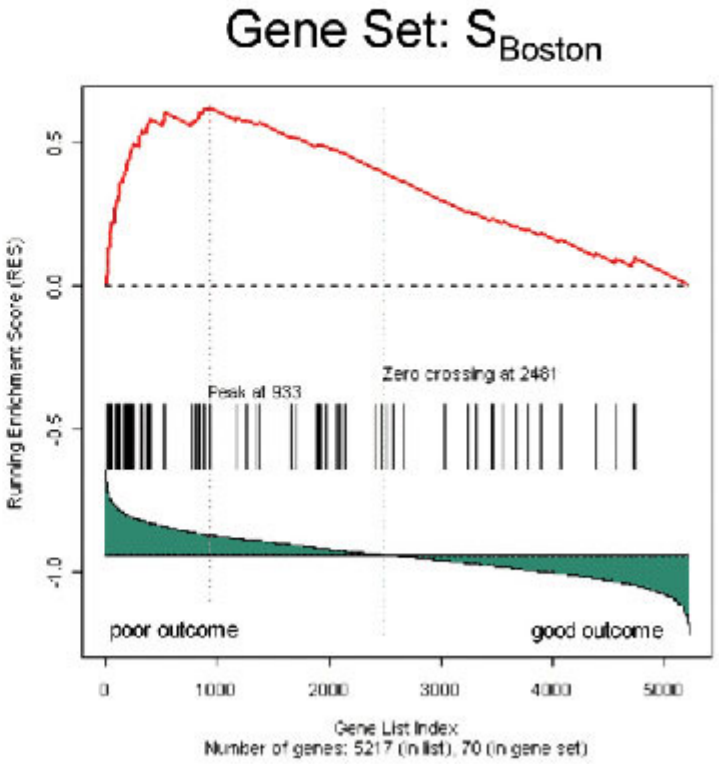
- GSEA is a computational method that applies the Kolmogorov–Smirnov test to compare the distribution of gene scores within **an a priori defined gene set** with one without that gene set.
- A gene score is defined by the difference of gene expressions between categorical phenotypes or correlation between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

# Results: LADC enrichment score

## Boston Dataset




## Michigan Dataset



# 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 |

# Trends and our future works

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

# Acknowledgements


- Dr. I-Shou Chang
- Dr. Wen-Chun Hung
- Dr. Li-Tzong Chen
- Ms. Fang-Yu Tsai
- TBI Bioinformatics core supported by MOST

# Thank you for your attention