NIH/NCI R21-CA209848

ALGORITHMS FOR LITERATURE-GUIDED MULTI-PLATFORM IDENTIFICATION OF CANCER SUBTYPES

Dongjun Chung, Ph.D.

Medical University of South Carolina (MUSC)

ITCR 2017 Annual Meeting, Santa Cruz, CA

BACKGROUND

- Cancer subtype identification:
  - Can offer opportunities for more personalized & targeted cancer treatment.

- Great achievements have been made:
  - The Cancer Genome Atlas (TCGA):
    - Integrative approach using multiple genomic data types.
    - e.g., mRNA expression, somatic mutation, copy number alteration, DNA methylation, …
  - iCluster+ (Mo et al., 2013, *PNAS*).
    - A statistical framework for integrative analysis of multiple data types to identify cancer subtypes.

BACKGROUND

- Cancer subtype identification is often implemented at the gene level:
  - Gene-level findings are sometimes not reproducible between different studies (Glaab, 2015, *Brief Bioinform*).
  - Genes with weak effects might be missed in gene-level analyses (Tyekucheva et al., 2011, *Genome Biol*).

- Pathway-level analysis:
  - Pathway-level findings have been reported to be more robust & reproducible (Glaab, 2015, *Brief Bioinform*).
  - Aggregation of signals within a pathway can potentially improve statistical power to identify key pathways.

BACKGROUND

- Challenges in pathway-level analyses:
  - Incompleteness of pathway knowledge.
  - Heterogeneity in completeness & quality among existing pathway databases.
  - Optimal strategies to combine pathway knowledge from multiple databases remain to be explored, especially when we combine databases for different aspects of biology.

- Biomedical literature:
  - Can potentially supplement incompleteness of pathway annotations because it provides comprehensive information about the relationship among genes.
  - Can potentially serve as a common knowledgebase to integrate multiple existing pathway databases.

BACKGROUND

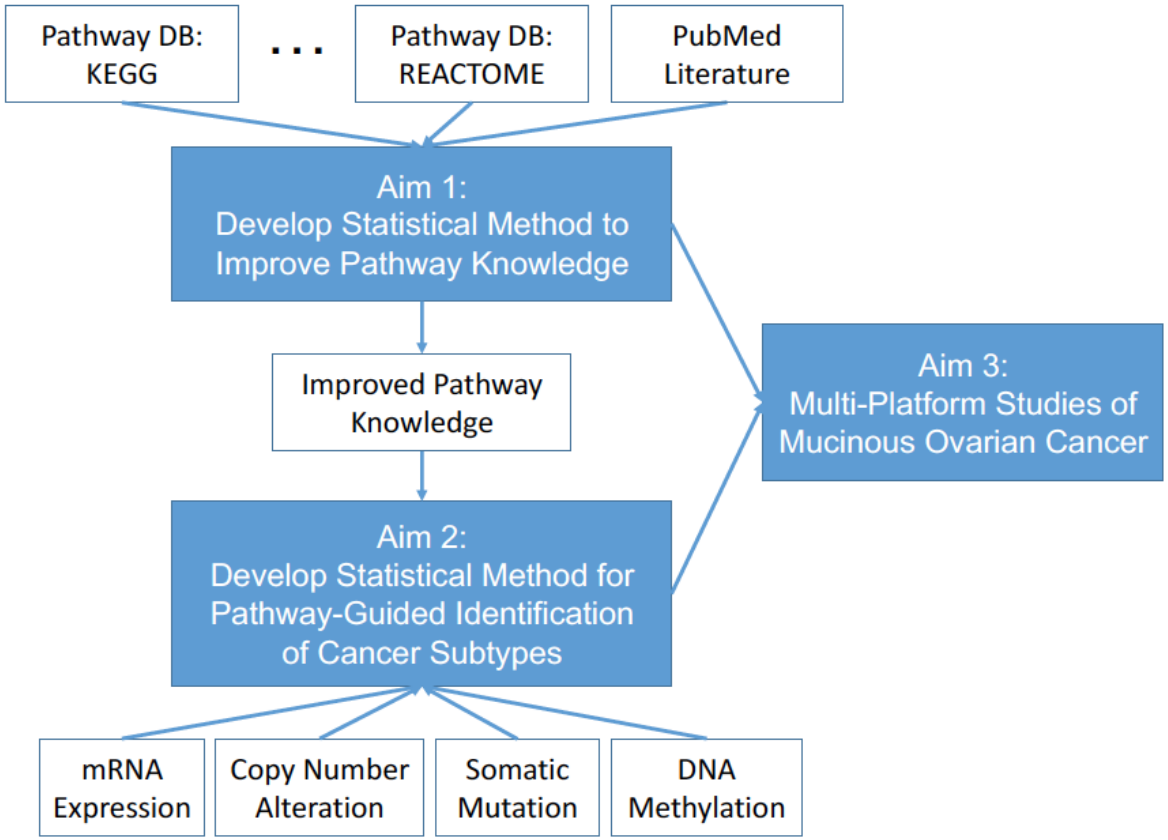
- Current challenges:
  - Pathway knowledge in existing databases & biomedical literature are not fully utilized and investigated for the cancer subtype identification.
  - There is a need for an effective statistical approach to improve pathway knowledge by integrating multiple existing pathway databases, also along with biomedical literature.

- NIH/NCI R21-CA209848 (MPI: D Chung & L Kelemen):
  - Aims to develop novel algorithms to improve robustness & interpretability in identification of cancer subtypes & key molecular features.
  - Integration of multiple genomic data types, biomedical literature, & existing pathway databases.

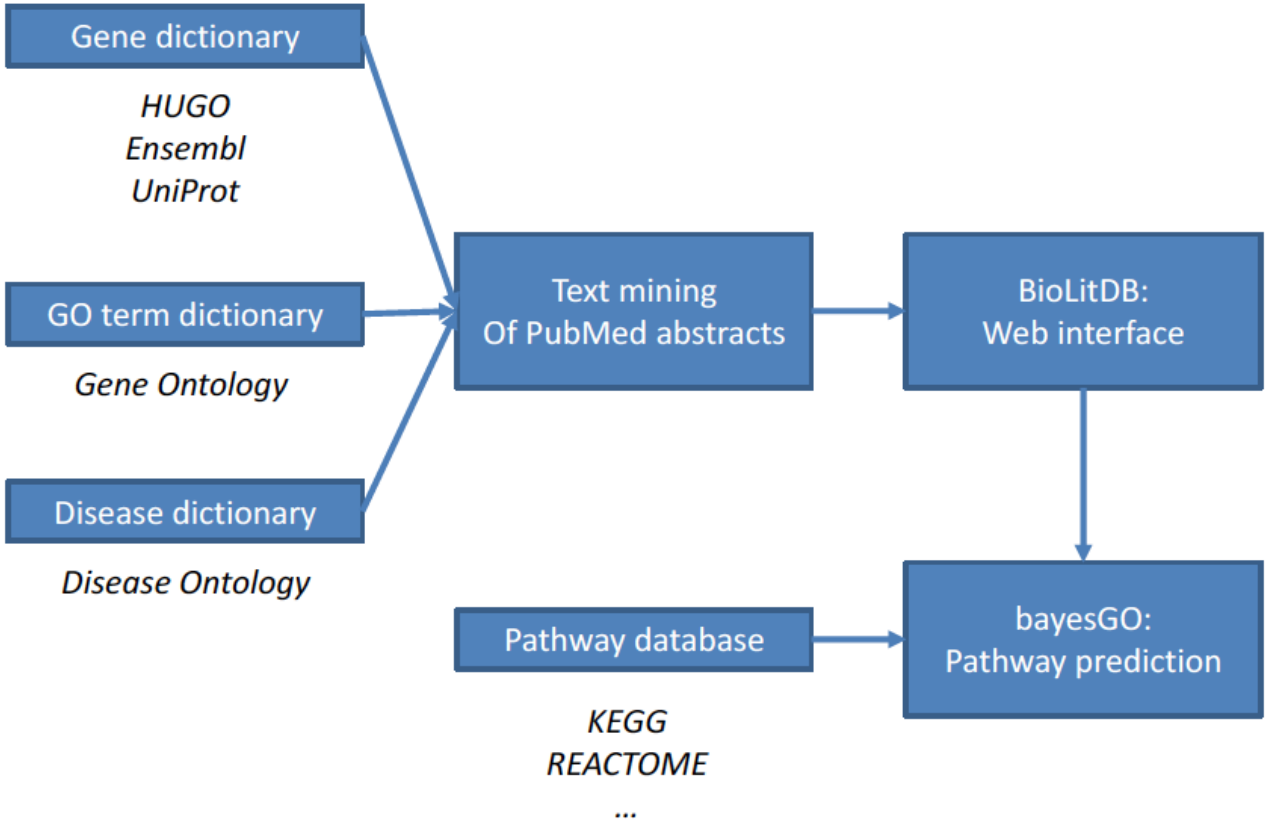
SPECIFIC AIMS

Improve pathway knowledge by integrating biomedical literature with multiple existing pathway databases.

Improve identification of cancer subtypes & key molecular features, by integrating pathway annotation with multiple genomic data types.

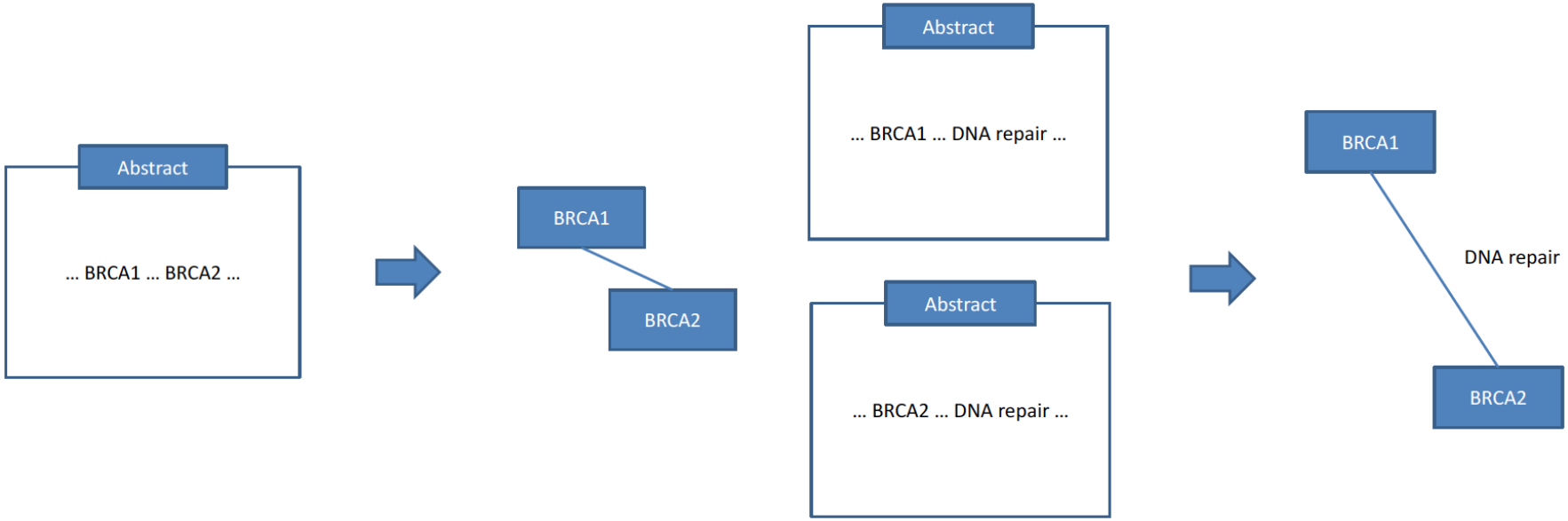


AIM 1. BIOMEDICAL LITERATURE + EXISTING PATHWAY DATABASES



AIM 1. BIOMEDICAL LITERATURE + EXISTING PATHWAY DATABASES

- **Ontology fingerprint**:
  - Collaboration with W. Jim Zheng, UT Health at Houston.
  - Qin et al. (2014), *NAR*.



AIM 1. BIOMEDICAL LITERATURE + EXISTING PATHWAY DATABASES

- **Ontology fingerprint**:
  - Association measures between genes and GO terms using hypergeometric tests.
  - *p*-values become smaller as more abstracts are shared.
  - Take into account how much each gene & GO term have been studied in biomedical literature.

Ontology fingerprint of BRCA1

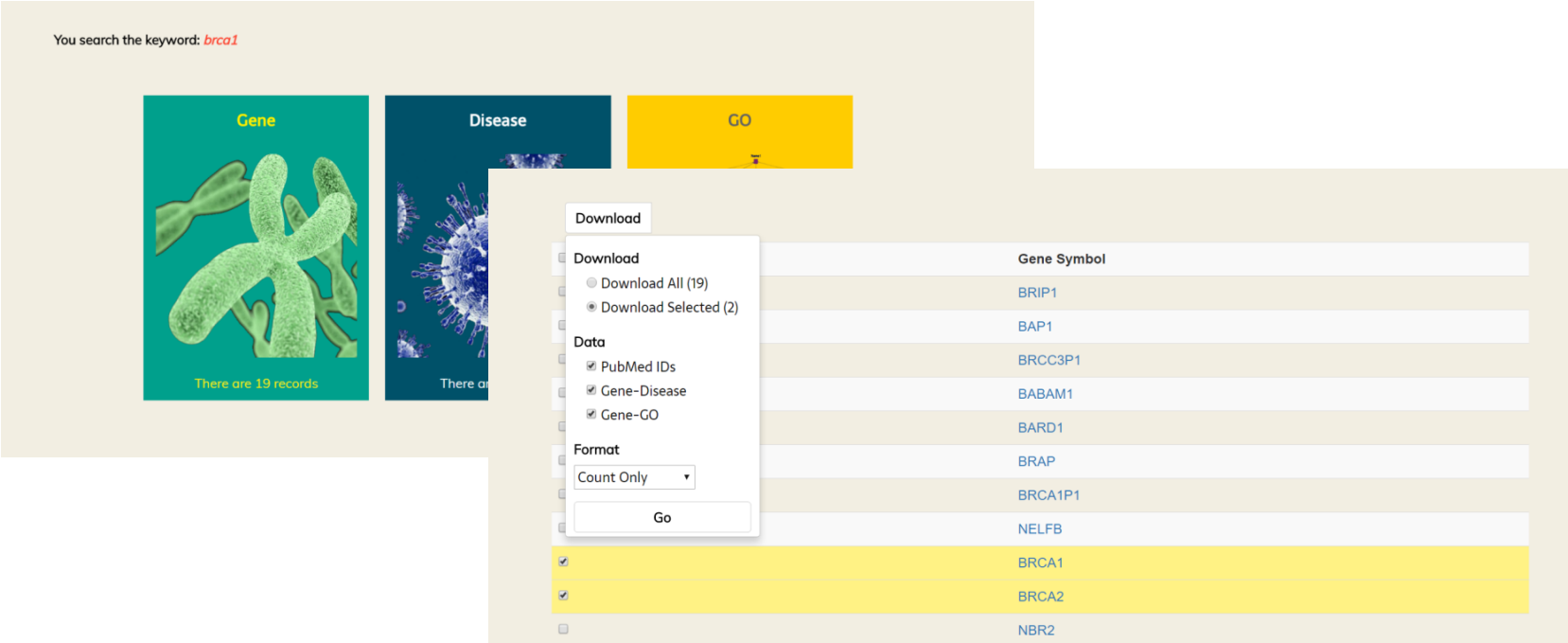
DNA repair	3.48e-158
Double-strand break repair	1.20e-25
Methylation	1.21e-18
Cell cycle checkpoint	3.64e-18
Mismatch repair	8.95e-13

Ontology fingerprint of BRCA2

DNA repair	1.95e-36
Strand invasion	8.32e-13
DNA recombination	5.86e-12
Double-strand break repair	1.61e-08
Recombination repair	2.37e-07

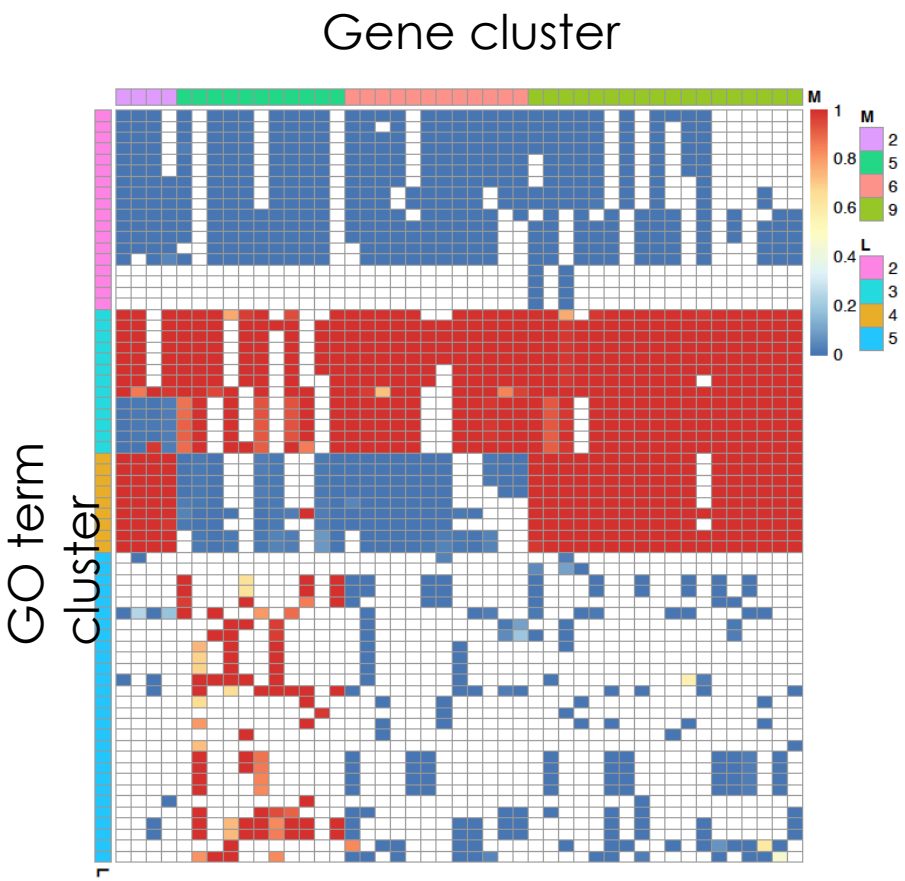
AIM 1. LITERATURE + EXISTING PATHWAY DATABASE

- **BioLitDB**: Web interface for literature mining.
  - Gene names, disease names, & gene ontology terms.
  - Currently internally developed & tested.
  - Plan public dissemination by end of this year.



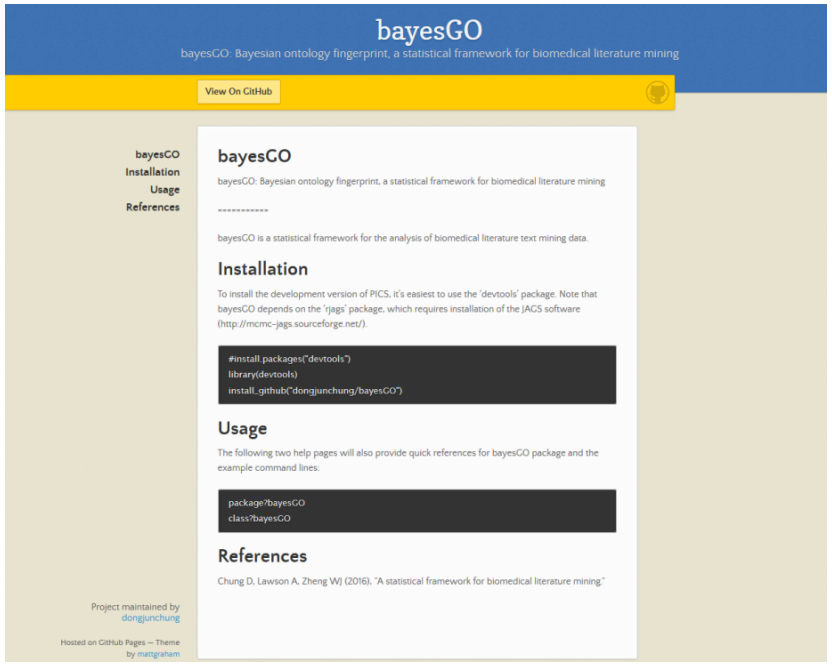
AIM 1. BIOMEDICAL LITERATURE + EXISTING PATHWAY DATABASES

- **bayesGO**:
  - Bayesian bi-clustering approach to identify novel pathways using the ontology fingerprint data.
  - Take care of redundancy & inter-correlation among GO terms.
  - Facilitate interpretation of novel pathways by automatically assigning groups of GO terms to each novel pathway.



AIM 1. BIOMEDICAL LITERATURE + EXISTING PATHWAY DATABASES

- **bayesGO**:
  - Chung et al. (2017), To appear in *Statistics in Medicine*.
  - **R package 'bayesGO'.**
    - bayesGO(): Fit model.
    - predict(): Gene and GO term clustering.
    - plot(): Plot the association heatmap.



https://dongjunchung.github.io/bayesGO/

AIM 1. BIOMEDICAL LITERATURE + EXISTING PATHWAY DATABASES

- Work in Progress:
- **BioLitDB**:
  - Improve user interface & public dissemination.
  - Incorporate pathway identification & visualization tools.
- **bayesGO**:
  - Integrate biomedical literature with multiple existing pathway databases.
    - A semi-supervised clustering approach that utilizes multiple existing pathway databases as prior knowledge for the pathway prediction based on biomedical literature data.
  - Utilize the GO tree structure information.

AIM 2. CANCER SUBTYPE IDENTIFICATION

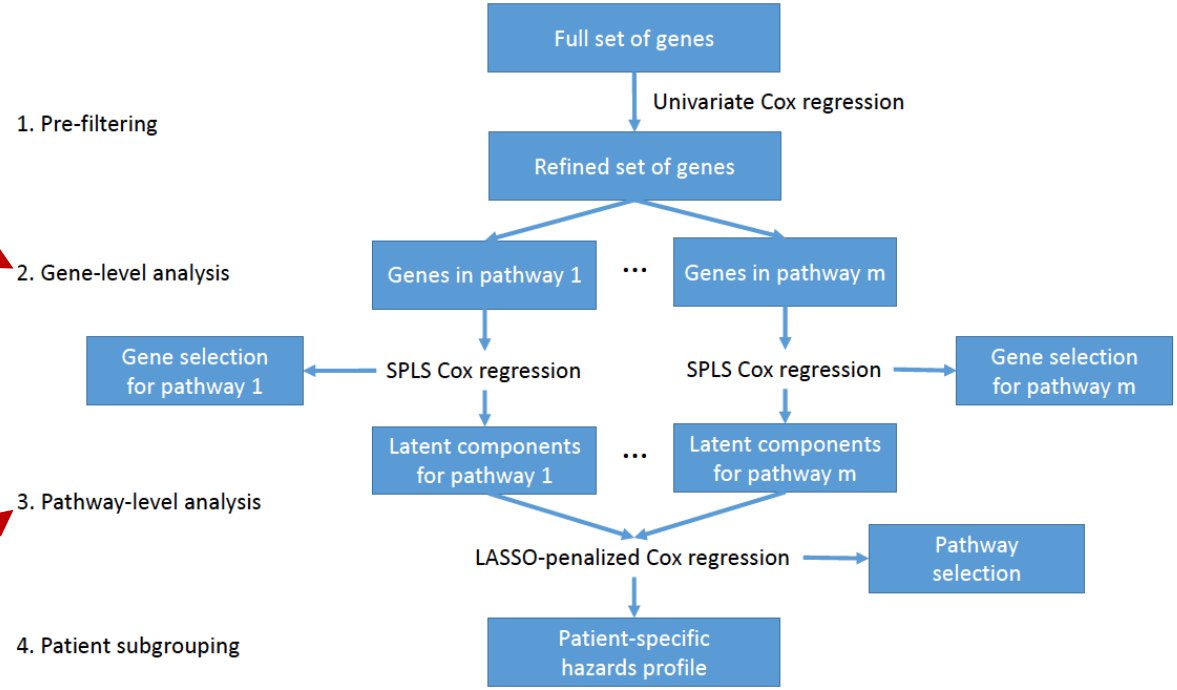
- **pathclust**:
  - Simultaneous identification of patient subgroups & key molecular features.
    - Identify key pathways, along with key genes in each pathway, associated with patient subgrouping.
  - Utilization of pathway information.
    - Improve robustness & stability in patient subgrouping.
  - Utilization of survival outcomes, if available.
    - Guide identification of patient subgroups and key molecular features using their association with survival.

AIM 2. CANCER SUBTYPE IDENTIFICATION

- **pathclust**:
  - Multi-step approach integrating sparse partial least squares (SPLS) and LASSO Cox regression approaches.

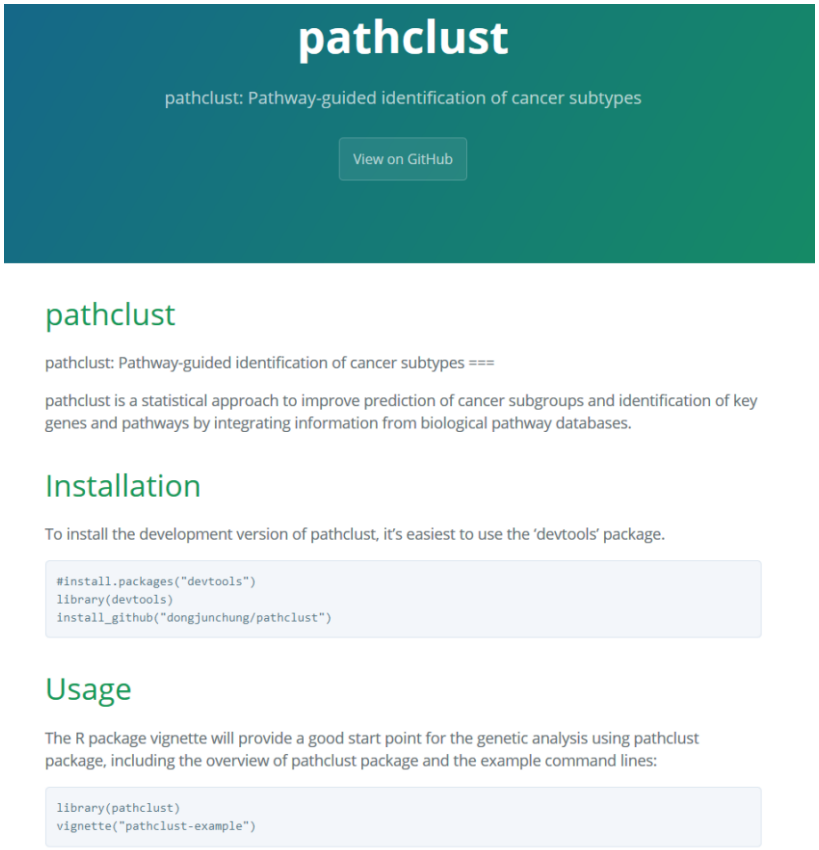
Simultaneous gene selection & dimension reduction for each pathway, using SPLS

Identification of parsimonious set of key pathways, using LASSO



AIM 2. CANCER SUBTYPE IDENTIFICATION

https://dongjunchung.github.io/pathclust/



- **R package 'pathclust'**
  - prefilter(): Prefiltering.
  - selectGene(): Gene selction. -> coef()
  - selectPath(): Pathway selection. -> coef()
  - predict(): Patient subgroup prediction.
  - plot(): Plot Kaplan-Meier, Hazard ratio, and ROC curves for predicted subgroups.

- YouTube tutorial video:
  - https://youtu.be/0qaovmMJPpY

AIM 2. CANCER SUBTYPE IDENTIFICATION

- Work in progress:
- **pathclust**:
  - Joint analysis of multiple genomic data types.
  - Simultaneous utilization of multiple pathway databases.
  - Incorporate various approaches to handle the issue of gene overlap between pathways.
  - Implement a Bayesian approach to unify the framework & to incorporate various prior knowledge.
- Application:
  - Utilize novel pathway knowledge generated from Aim 1, which integrates biomedical literature with multiple existing pathway databases.

AIM 3. MUCINOUS OVARIAN CANCER SUBTYPE IDENTIFICATION

- Mucinous ovarian cancer (MOC):
  - Still relatively less studied; its subtypes remain poorly characterized in spite of its severity.
  - In need of improved treatments targeted to MOC.

- Planned work:
  - The statistical methods developed in Aims 1 & 2 will be applied to our cancer genomic data for MOC patients.
    - BioLitDB, bayesGO, & pathclust.
  - Will also be integrated with corresponding TCGA data for colorectal cancer, endometrial cancer, & gastro-esophageal adenocarcinoma.

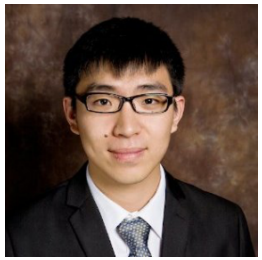
NATIONAL CANCER INSTITUTE
Informatics Technology for Cancer Research

ACKNOWDGEMENT

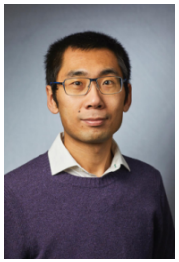
Chung lab, MUSC

Dongjun Chung (Biostatistics)

Zhenning Yu



Zequn Sun



Wei Wei (Yale Cancer Center)

Linda Kelemen (Cancer Epidemiology)

Collaborators

Andrew Lawson (Biostatistics)

Gary Hardiman (Bioinformatics, Center for Genomic Medicine)

- **Questions?**
  - Lab website: https://sites.google.com/site/statdchung/
  - GitHub: https://github.com/dongjunchung/
  - e-mail: chungd@musc.edu.