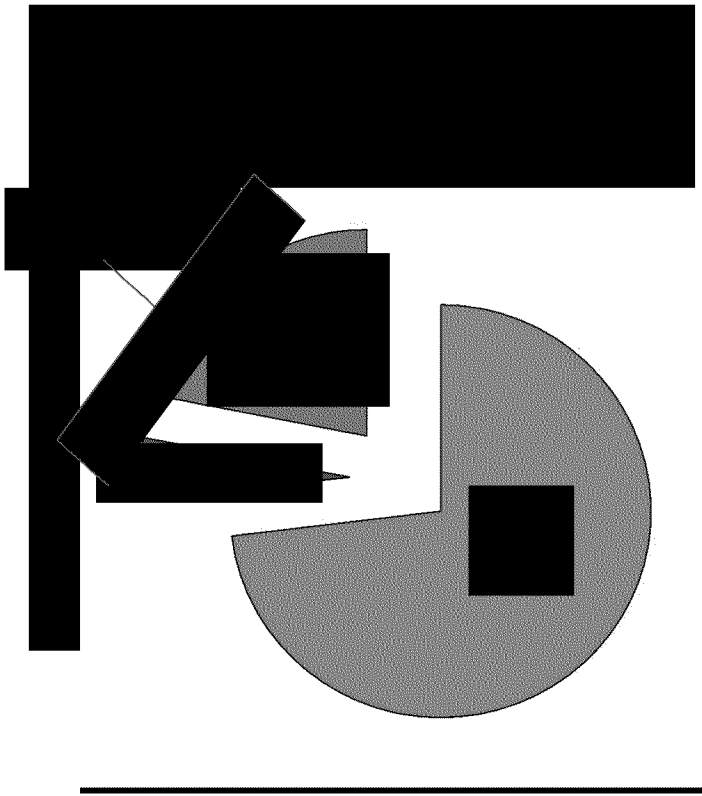
Sean Watford
Research Proposal
April 6 2017



- Hormone Replacement Therapy
- Diethylstilbestrol (DES)
- Alcohol

Recent evidence indicates that estrogenic chemicals may increase breast cancer risk via alterations in *gene expression* and other mechanisms

- How do we identify key genes involved in chemically-influenced breast carcinogenesis?



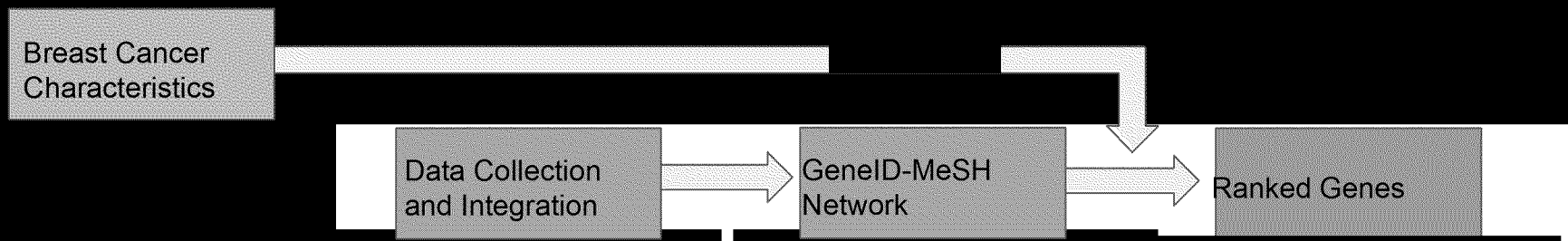
ED_001795_00002843-00002

[redacted]

[redacted]

# information retrieval to identify and rank mammary and cancer-related genes

- Identify keywords related to breast cancer and associated cancer characteristics, or hallmarks
- Use search algorithm to probe publicly available biomedical literature databases
- Identify and rank genes based on overrepresentation within breast cancer-related literature

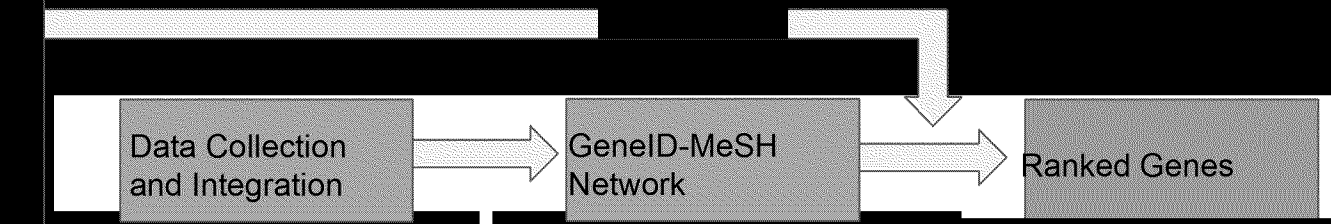





ED_001795_00002843-00004

# information retrieval to identify and rank mammary and cancer-related genes

- Identify keywords related to breast cancer and associated cancer characteristics, or hallmarks
- Use search algorithm to probe publicly available biomedical literature databases
- Identify and rank genes based on overrepresentation within breast cancer-related literature



[redacted]

## Characteristics of carcinogenesis and selected MeSH terms

| Characteristic | Hanahan | Smith | Schwarzman | Goodson | MeSH [MeSH Identifier] |
|---|---|---|---|---|---|
| Altered Peptide (Growth) Hormone Activity | X | | X | X | Growth Hormone [D013006] |
| Angiogenesis | X | | | X | Neovascularization -- Physiologic [D018919] |
| | | | | | Neovascularization -- Pathologic [D009389] |
| Cell Cycle Changes | | X | X | | Cell Cycle [D002453] |
| Epigenetics | X | X | X | | Epigenomics [D057890] |
| Estrogenic and Other Hormone Mediated Effects | | X | X | | Gonadal Steroid Hormones [D012739] |
| Evading Apoptosis | X | X | X | X | Apoptosis [D017209] |
| Genotoxicity | X | X | X | X | DNA Repair [D004260] |
| | | | | | DNA Damage [D004249] |
| Immortalization | X | X | X | X | Cell Survival [D002470] |
| Immune Modulation | X | X | X | X | Immune System [D007107] |
| Inflammation | X | X | X | X | Inflammation [D007249] |
| Mammary | | | X | | Mammary Glands -- Human [D042361] |
| | | | | | Breast Diseases [D001941] |
| Metabolic Activation | X | X | X | | Xenobiotics [D015262] |
| Oxidative Stress | | X | X | | Oxidative Stress [D018384] |
| Cell Proliferation | X | X | X | X | Cell Proliferation [D049109] |

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

Mol Cell Endocrinol. 2009 May 25;304(1-2):63-8. doi: 10.1016/j.mce.2009.02.016. Epub 2009 Mar 9.

**The pancreatic beta-cell as a target of estrogens and xenoestrogens: Implications for blood glucose homeostasis and diabetes.**

Nadal A[1], Alonso-Magdalena P, Soriano S, Quesada I, Ropero AB.

**Author information**

**Abstract**

The estrogen receptor ERalpha is emerging as a key molecule involved in glucose and lipid metabolism. The main functions of pancreatic beta-cells are the biosynthesis and release of insulin, the only hormone that can directly decrease blood glucose levels. Estrogen receptors ERalpha and ERbeta exist in beta-cells. The role of ERbeta is still unknown, yet ERalpha plays an important role in the regulation of insulin biosynthesis, insulin secretion and beta-cell survival. Activation of ERalpha by 17beta-estradiol (E2) and the environmental estrogen bisphenol-A (BPA) promotes an increase of insulin biosynthesis through a non-classical estrogen-activated pathway that involves phosphorylation of ERK1/2. The activation of ERalpha by physiological concentrations of E2 may play an important role in the adaptation of the endocrine pancreas to pregnancy. However, if ERalpha is over stimulated by an excess of E2 or the action of an environmental estrogen such as BPA, it will produce an excessive insulin signaling. This may provoke insulin resistance in the liver and muscle, as well as beta-cell exhaustion and therefore, it may contribute to the development of type II diabetes.

PMID: 19433249 DOI: 10.1016/j.mce.2009.02.016

[PubMed - indexed for MEDLINE]

**Publication Types, MeSH Terms, Substances**

**LinkOut - more resources**

[redacted]

# information retrieval to identify and rank mammary and cancer-related genes

- Identify keywords related to breast cancer and associated cancer characteristics, or hallmarks
- Use search algorithm to probe publicly available biomedical literature databases
- Identify and rank genes based on overrepresentation within breast cancer-related literature

ED_001795_00002843-00009

[redacted]

| | | |
|---|---|---|
| Gene2pubmed | 998,833 | 4,625,706 |
| Gene Reference into Function (GeneRIF) | 628,432 | 84,929 |
| Comparative Toxicogenomics Database (CTD) | 55,568 | 40,862 |
| Universal Protein Resource (UniProt/Swiss-Prot) | 874,281 | 23,1051 |
| Reactome | 15,045 | 10,850 |
| Rat Genome Database (RGD) | 58,144 | 30,172 |
| Mouse Genome Informatics (MGI) | 177,324 | 42,125 |
| **Total Unique Articles** | **1,178,220** | **4,649,012** |

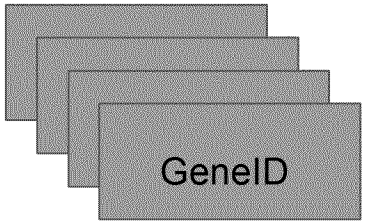
| | | |
|---|---|---|
| Mouse Genome Informatics (MGI) | 177,324 | 42,125 |
| **Total Unique Articles** | **1,178,220** | **4,649,012** |

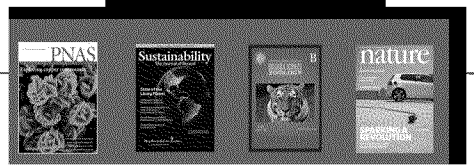
| eneIDs |
|---|
| 4,625,706 |
| 84,929 |
| 40,862 |
| 23,1051 |
| 10,850 |
| 30,172 |

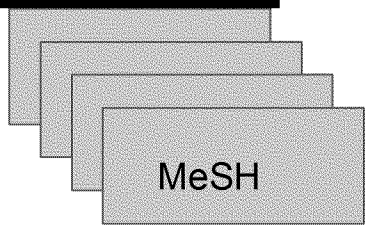
| ism |
|---|
| sapiens |
| sapiens |
| urus |
| rofa |
| atus |
| caballus |
| norvegicus |
| usculus |
| cricetus auratus |
| gallus |
| ppygia guttata |

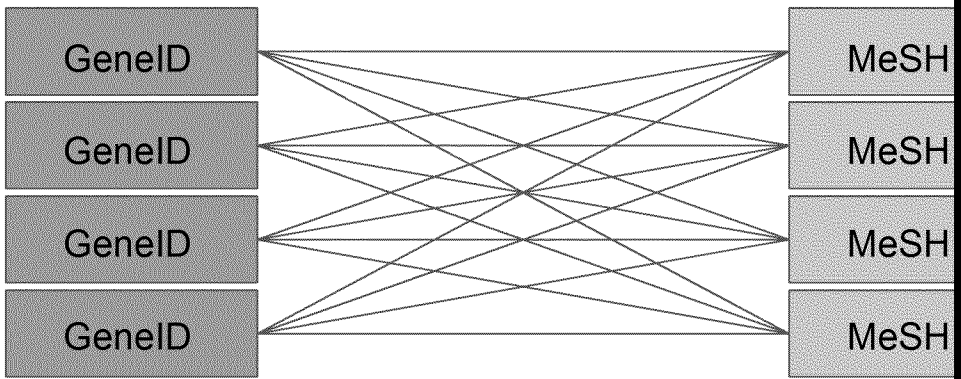
[redacted]

[redacted]



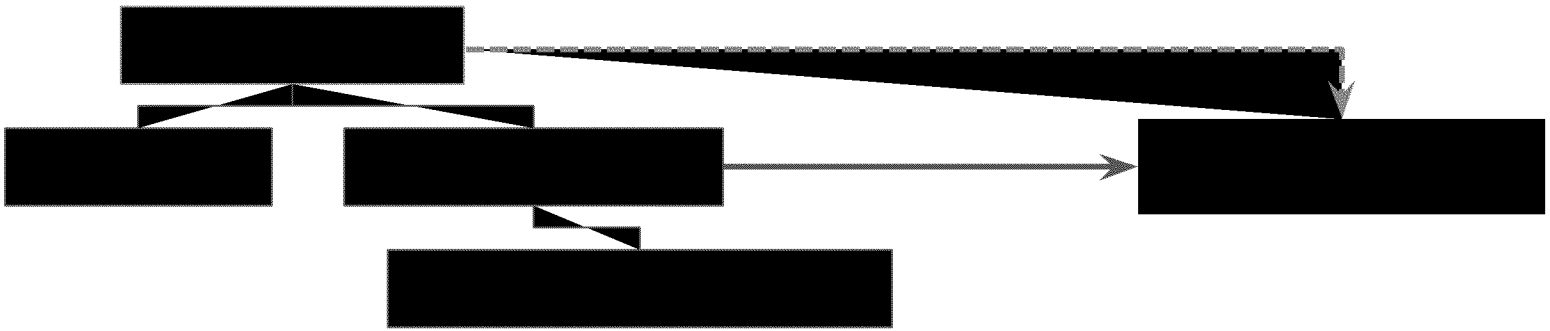




[redacted]

[redacted]



[redacted]

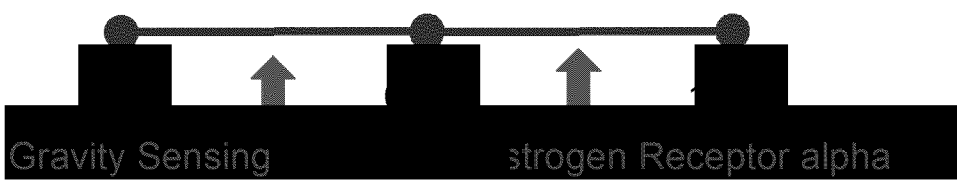
[redacted]

Equation1: $pmi(g;m) = \log \frac{p(g,m)}{p(g)p(m)}$

Equation 2: $npmi(g;m) = \frac{pmi(g;m)}{-\log p(g,m)}$

[redacted]

[redacted]

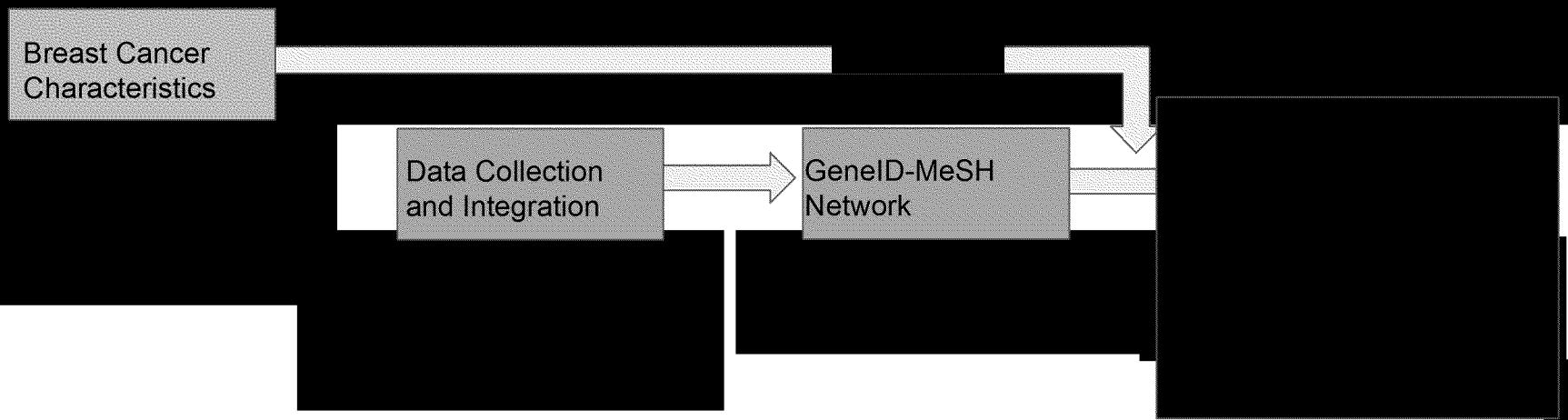


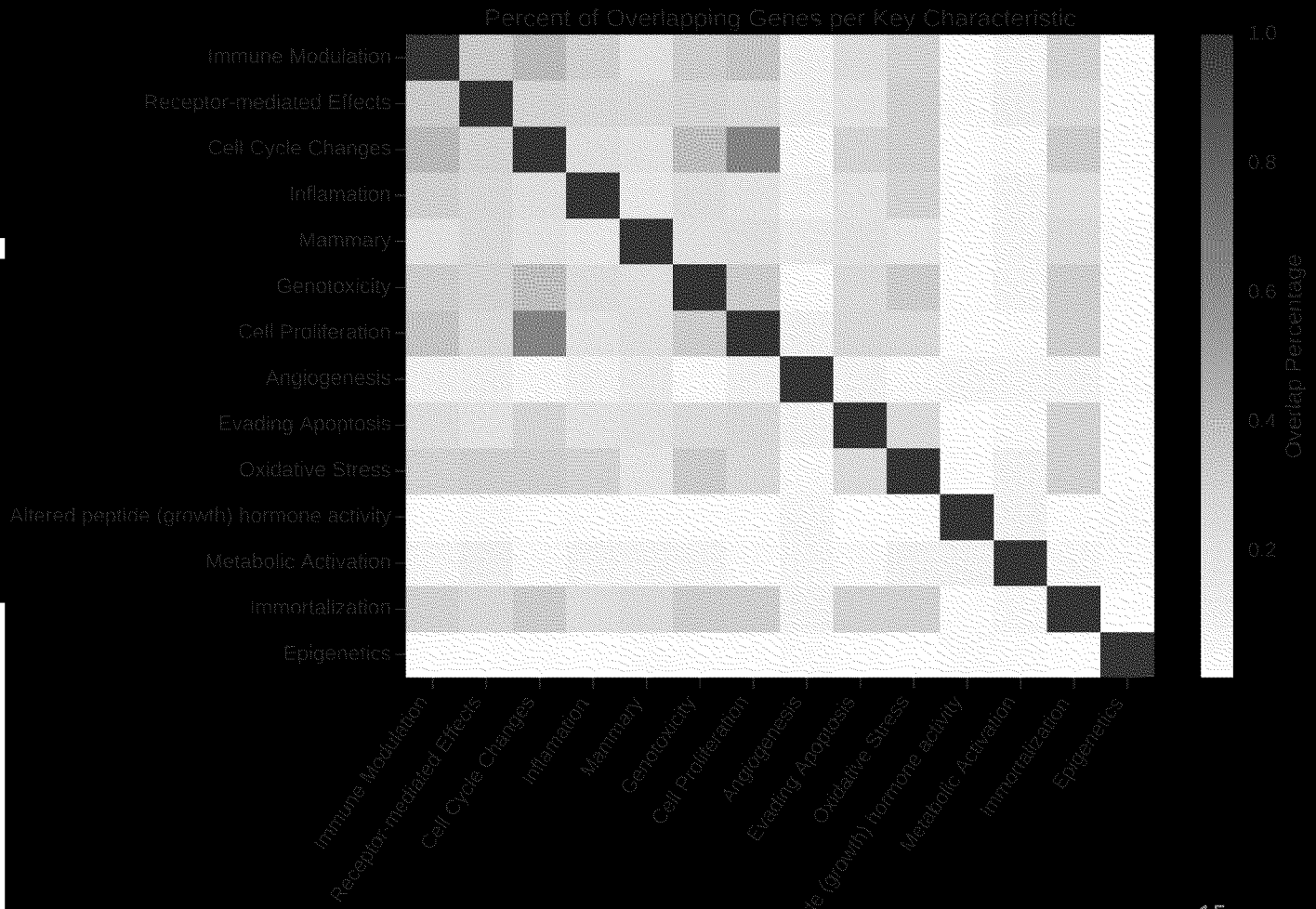

| Gene | MeSH Term | NPMI |
|---|---|---|
| Estrogen Receptor alpha (ESR1) | Estrogen Receptor alpha | 0.4304 |
| Estrogen Receptor alpha (ESR1) | Receptors, Estrogen | 0.4163 |
| Estrogen Receptor alpha (ESR1) | Estrogen Receptor beta | 0.3809 |
| Estrogen Receptor alpha (ESR1) | Receptors, Steroid | 0.3496 |

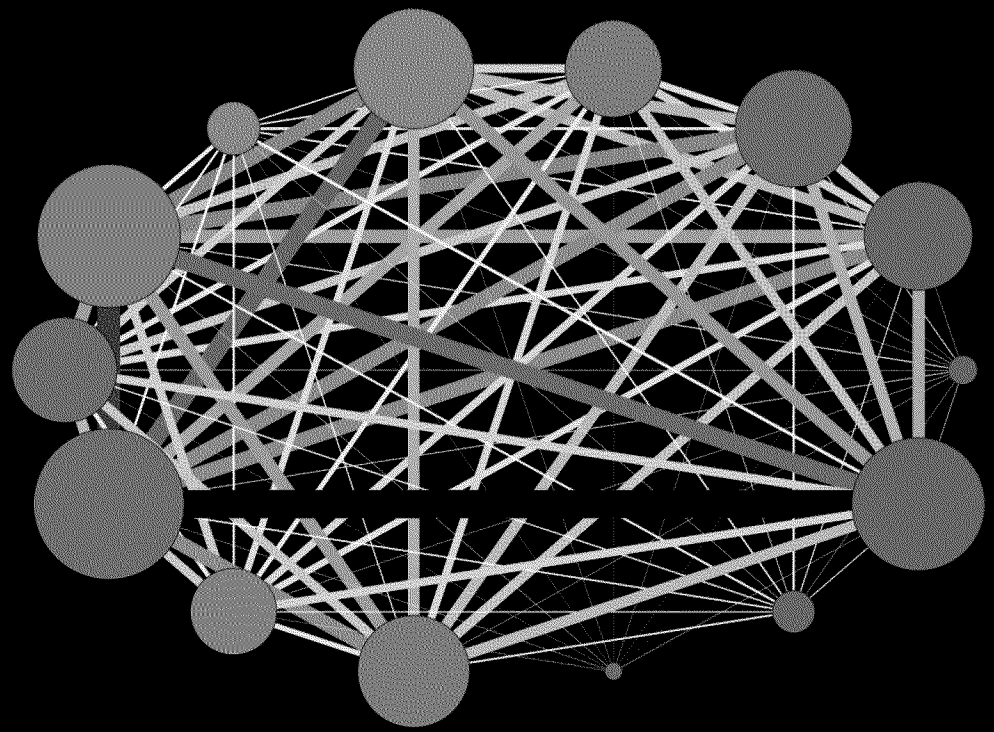
[redacted]

# information retrieval to identify and rank mammary and cancer-related genes

- Identify keywords related to breast cancer and associated cancer characteristics, or hallmarks
- Use search algorithm to probe publicly available biomedical literature databases
- Identify and rank genes based on overrepresentation within breast cancer-related literature




Percent of Overlapping Genes per Key Characteristic
Immune Modulation
Receptor-mediated Effects
Cell Cycle Changes
Inflamation
Mammary
Genotoxicity
Cell Proliferation
Angiogenesis
Evading Apoptosis
Oxidative Stress
Altered peptide (growth) hormone activity
Metabolic Activation
Immortalization
Epigenetics
1.0
0.8
0.6
0.4
0.2
Overlap Percentage

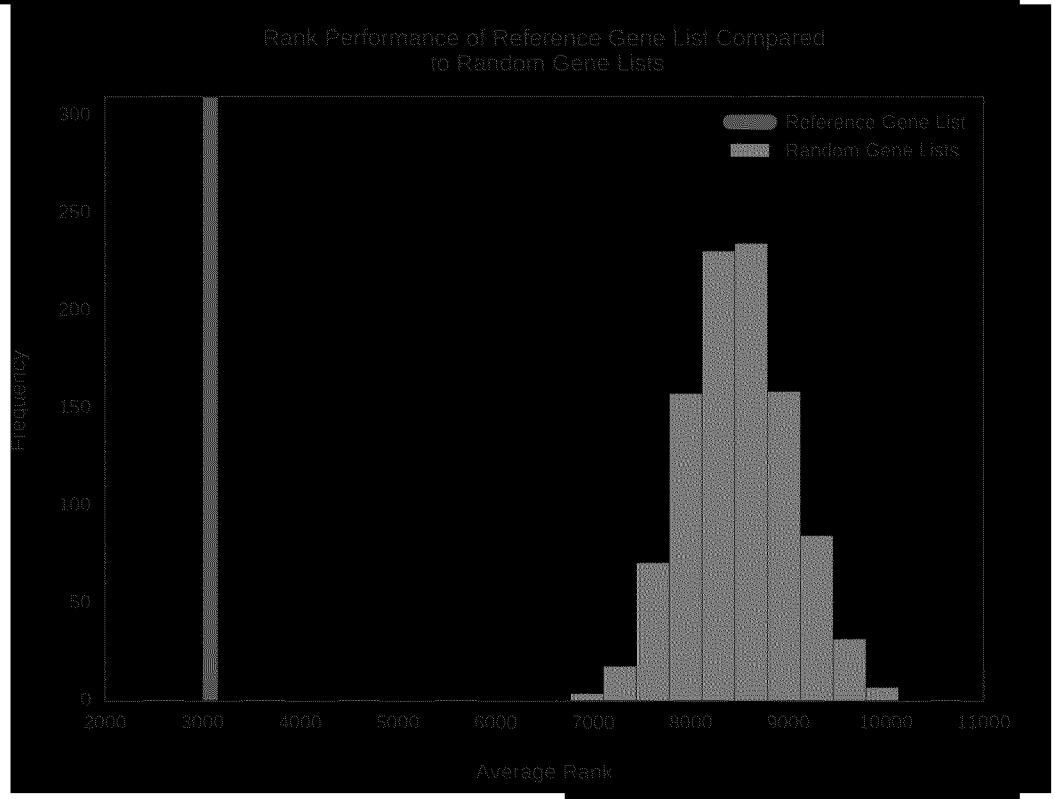


- Cell Cycle Changes
- Immune Modulation
- Cell Proliferation
- Genotoxicity
- Receptor-mediated Effects
- Oxidative Stress
- Immortalization
- Inflamation
- Evading Apoptosis
- Mammary
- Angiogenesis
- Metabolic Activation
- Altered peptide (growth) hormone activity
- Epigenetics

0 800 1600 2400 3200 4000 4800 5600 6400 7200

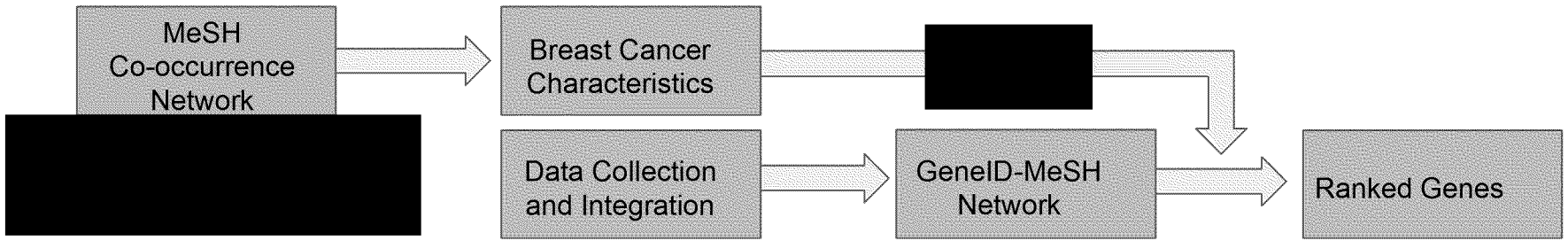
Number of Overlapping Genes

[redacted]


Rank Performance of Reference Gene List Compared to Random Gene Lists
Reference Gene List
Random Gene Lists
Frequency
300
250
200
150
100
50
0
2000
3000
4000
5000
6000
7000
8000
9000
10000
11000
Average Rank

- Relies on manual selection of MeSH terms

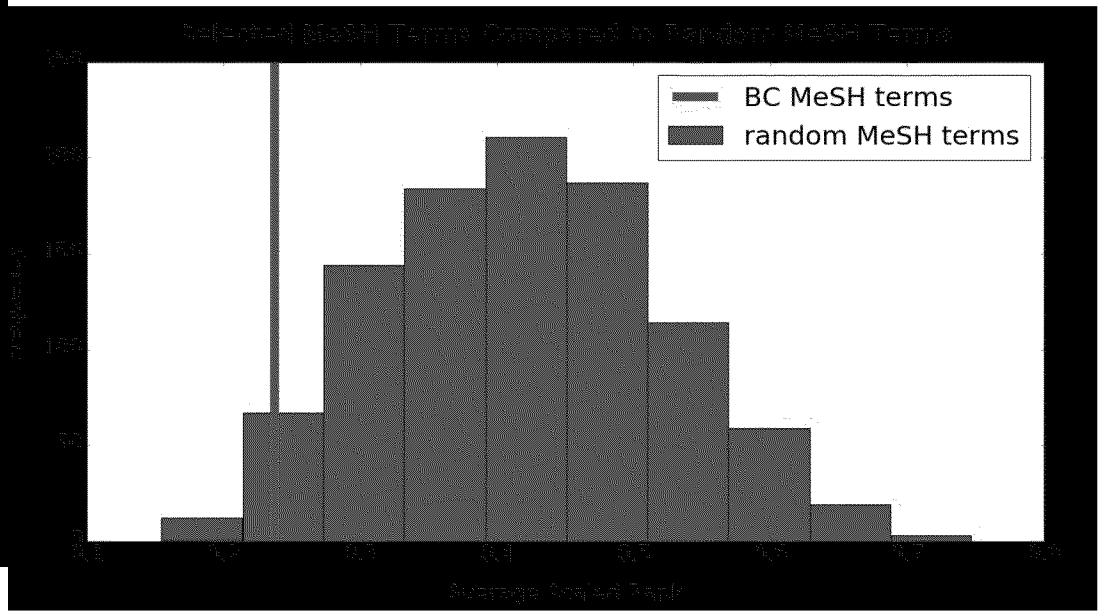
- Use keyword association algorithm that incorporates publication context, i.e. groups of keywords used define a concept like breast neoplasms
  Refine and optimize the search for breast cancer-related genes



[redacted]

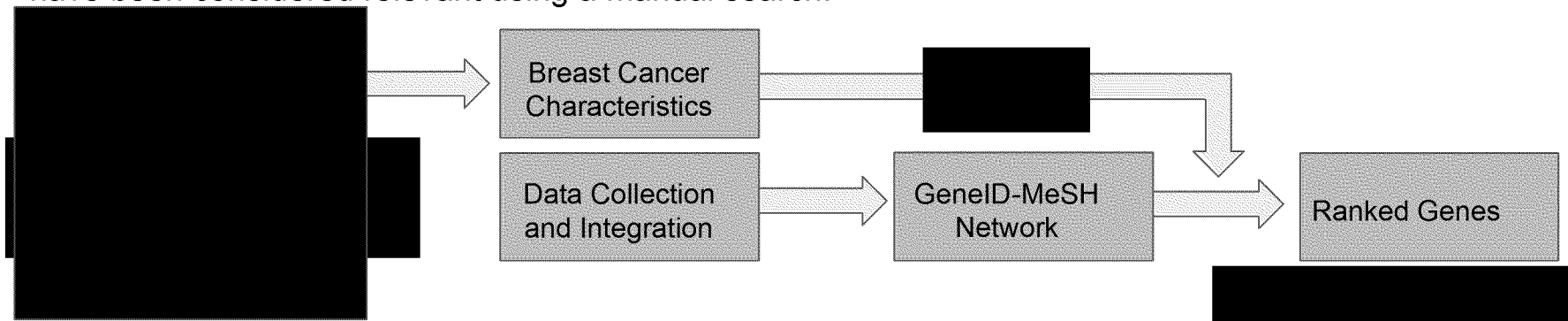
# optimized for breast cancer biomarker

[redacted]




**outperform all randomly selected MeSH term seeds (p<0.034)**

- Use keyword association algorithm that incorporates publication context, i.e. groups of keywords used define a concept like breast neoplasms
  Refine and optimize the search for breast cancer-related genes



[redacted]

[redacted]

- Ranked using NPMI

MeSH co-occurrence examples

| MeSH terms tagged to an article | MeSH cooccurrences |
|---|---|
| Estradiol | Estradiol -- Estrogen Receptor alpha |
| Estrogen Receptor alpha | Estradiol -- Estrogen Receptor beta |
| Estrogen Receptor beta | Estradiol -- Insulin |
| Insulin | Estradiol -- Insulin-Secreting Cells |
| Insulin-Secreting Cells | Estradiol -- Pregnancy |
| Pregnancy | Estradiol -- Signal Transduction |
| Signal Transduction | Estrogen Receptor alpha -- Estrogen Receptor beta |
| | Estrogen Receptor alpha -- Insulin |
| | Estrogen Receptor alpha -- Insulin-Secreting Cells |
| | Estrogen Receptor alpha -- Pregnancy |
| | Estrogen Receptor alpha -- Signal Transduction |
| | Estrogen Receptor beta -- Insulin |
| | Estrogen Receptor beta -- Insulin-Secreting Cells |
| | Estrogen Receptor beta -- Pregnancy |
| | Estrogen Receptor beta -- Signal Transduction |
| | Insulin -- Insulin-Secreting Cells |
| | Insulin -- Pregnancy |
| | Insulin -- Signal Transduction |
| | Insulin-Secreting Cells -- Pregnancy |
| | Insulin-Secreting Cells -- Signal Transduction |
| | Pregnancy -- Signal Transduction |

Top overrepresented co-occurring MeSH terms with "breast"

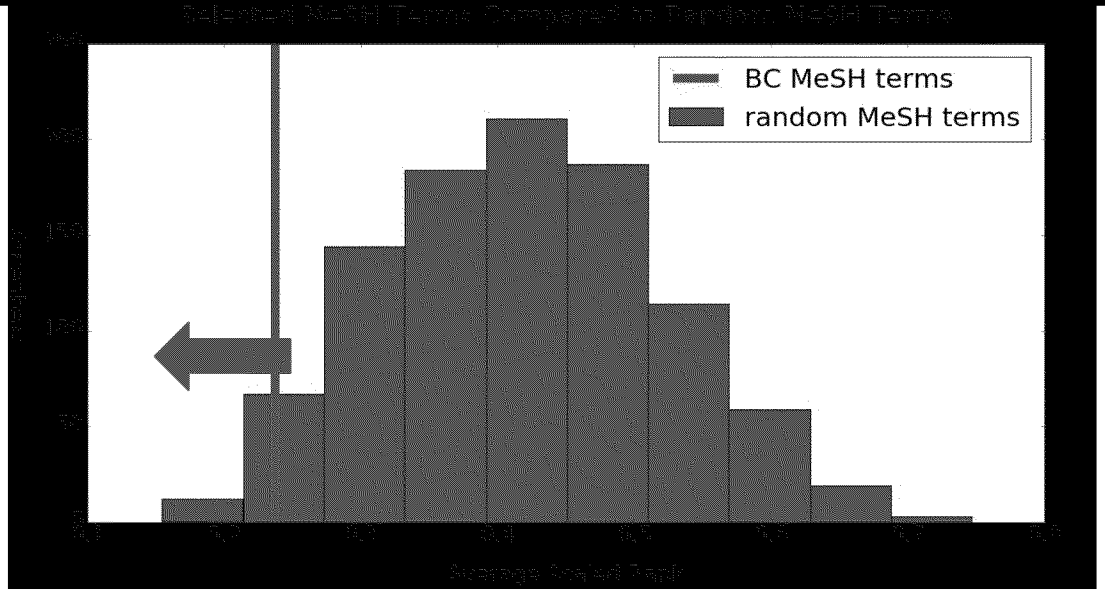
| Search MeSH Term | Cooccurring MeSH Term | NPMI | Number of Shared Articles |
|---|---|---|---|
| breast | breast diseases | 0.586202 | 3096 |
| breast | breast neoplasms | 0.551437 | 17627 |
| breast | nipples | 0.540973 | 1274 |
| breast | mammography | 0.538149 | 3926 |
| breast | fibrocystic breast disease | 0.516292 | 853 |
| breast | ultrasonography -- mammary | 0.514345 | 1019 |
| breast | mammaplasty | 0.499622 | 1537 |
| breast | mastectomy | 0.490098 | 2302 |
| breast | fibroadenoma | 0.479796 | 458 |
| breast | mammary glands -- human | 0.472608 | 668 |

- Use keyword association algorithm that incorporates publication context, i.e. groups of keywords used define a concept like breast neoplasms
  Refine and optimize the search for breast cancer-related genes

MeSH
Co-occurrence
Network

[redacted]

[redacted]




[redacted]

[redacted]

[redacted]

[redacted]

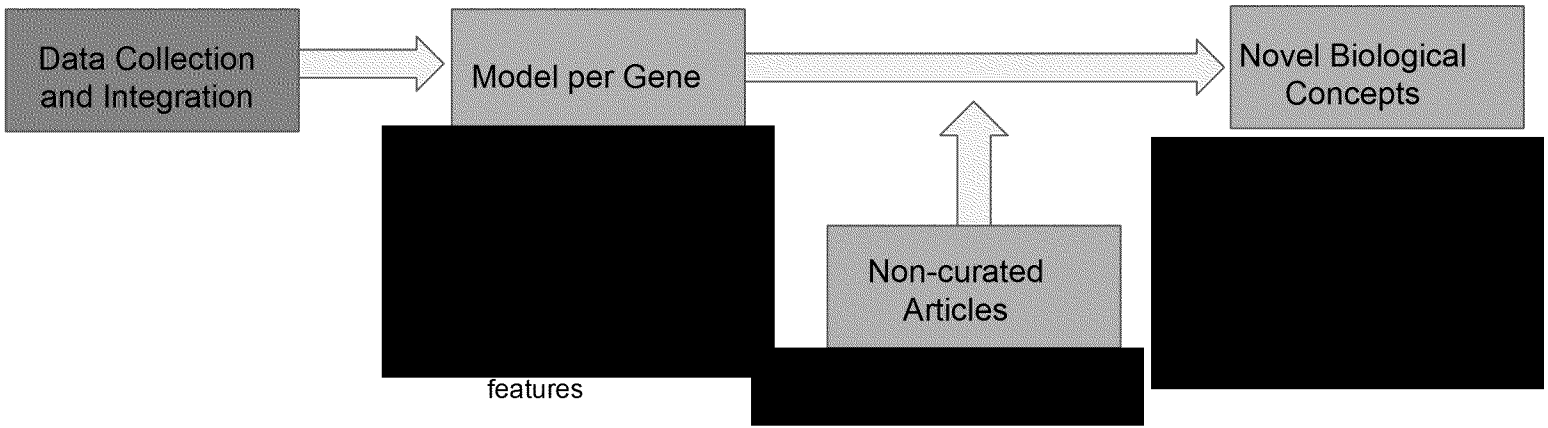
# discover novel biological relationships through machine learning model predictions

[redacted]

- Develop machine learning models that predict previously unknown associations to articles and corresponding biological concepts

[redacted]

develop the model.



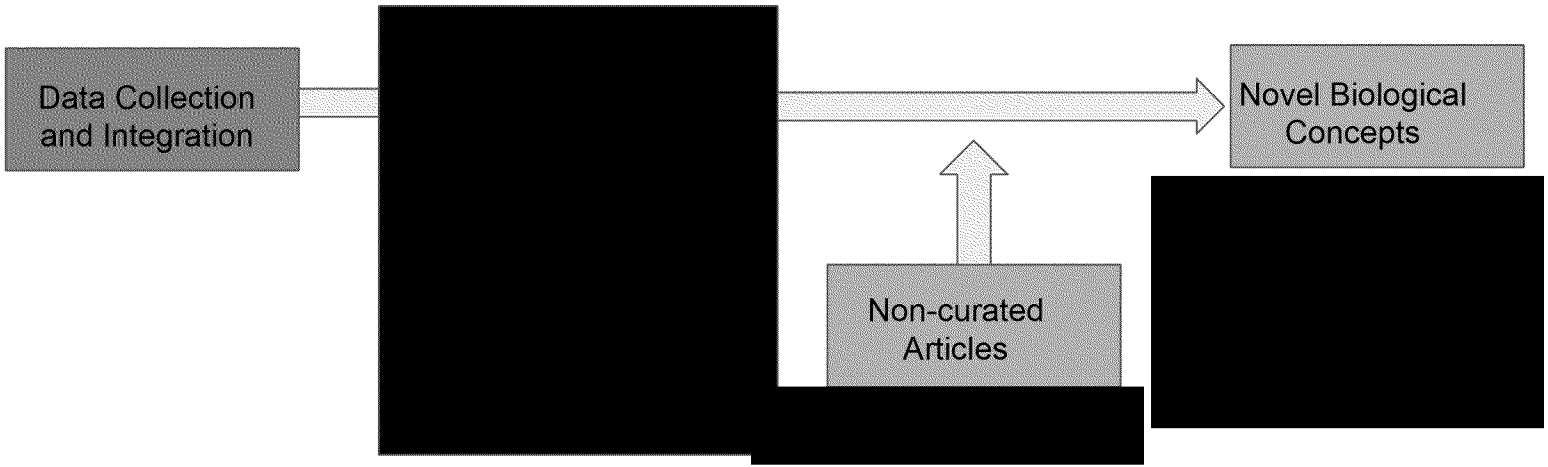
[redacted]

# discover novel biological relationships through machine learning model predictions

- Develop machine learning models that predict previously unknown associations to articles and corresponding biological concepts

[redacted]

develop the model.



[redacted]

[redacted]

[redacted]

| | | D000328 | D000375 | D000595 | D001483 | D015152 | D003001 | D005260 | D005333 | D0 |
|---|---|---|---|---|---|---|---|---|---|---|
| pmid | curated | | | | | | | | | |
| 19259135 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| 9245787 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 |
| 9310354 | 1 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 |
| 7781776 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 16169070 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 22261194 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 12393739 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 18496706 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| 14574404 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 15661163 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 12454288 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

- 80/20 split (cross-validation procedure for variability/stability of the prediction)

[redacted]

# discover novel biological relationships through machine learning model predictions

- Develop machine learning models that predict previously unknown associations to articles and corresponding biological concepts

[redacted]

develop the model.

| Gene Symbol | GeneID | MeSH term | Number of curated articles | Mean Accuracy |
|---|---|---|---|---|
| ESR1 | 2099 | Estrogen receoptor alpha, human | 2631 | 0.95 |
| LAMA1 | 284217 | - | 50 | 0.90 |
| KLC1 | 3831 | - | 75 | 0.79 |
| PPARG | 5468 | PPAR gamma | 1729 | 0.96 |



ED_001795_00002843-00030

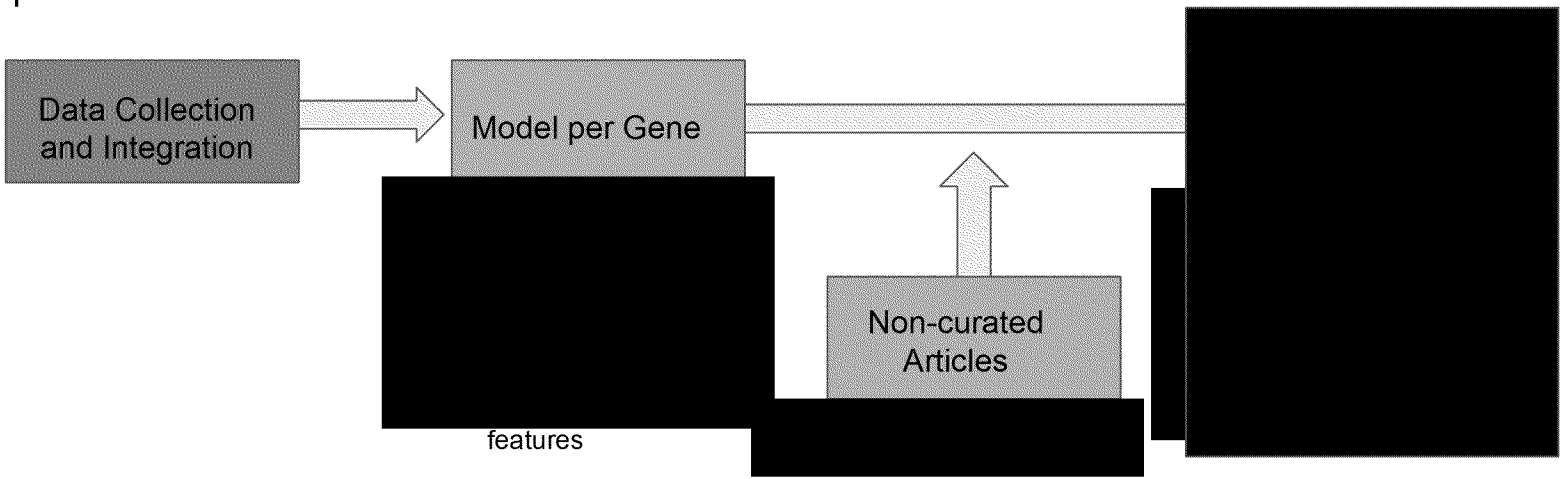
[redacted]

# discover novel biological relationships through machine learning model predictions

- Develop machine learning models that predict previously unknown associations to articles and corresponding biological concepts

[redacted]

develop the model.



[redacted]

[redacted]

| Aim | Publications | Timeline |
|---|---|---|
| Aim 1 | Draft paper in preparation: Novel Approach of integrating and mining biomedical literature for the identification of non-inherited genes that play a role in breast carcinogenesis<br>Authors: **Sean Watford**, Rachel Grashow, Vanessa DeLaRosa, Ruthann Rudel, Matt Martin | Fall 2017 |
| | Draft paper in preparation: A gene expression panel targeting cellular processes involved in breast carcinogenesis for chemical screening<br>Authors: Rachel Grashow, Vanessa DeLaRosa, **Sean Watford**, Janet Ackerman, Ruthann Rudel | |
| Aim 2 | Working title: Utility of MeSH co-occurrence network to define publication context<br>Authors: **Sean Watford**, Matt Martin | Winter 2017 |
| | Working title: Identification of overrepresented breast cancer-related keywords for complex searches<br>Authors: **Sean Watford**, Rachel Grashow, Matt Martin | |
| Aim 3 | Working title: Systematic approaches for expanding associations between genes and articles<br>Authors: **Sean Watford**, Matt Martin | Spring 2018 |
| | Working title: Predicted gene associations to articles elucidate novel biological processes related to breast cancer<br>Authors: **Sean Watford**, Rachel Grashow, Matt Martin | |



ED_001795_00002843-00033

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]