DATABASES AND BIOINFORMATIC TOOLS FOR IDENTIFYING DISEASE GENES

Marili Palover
4.11.2016
Outline

- Introduction
- Databases for identifying disease genes
- Bioinformatic tools to predict genes likely to cause or be associated with disease
- Conclusion aka Take home message
- Homework
DNA → RNA → PROTEIN
CENTRAL DOGMA OF BIOLOGY

Introduction

Katherine Joyce, Woods Hole Oceanographic Institution
SINGLE NUCLEOTIDE POLYMORPHISMS

- SNPs and DNA mutations are defined as DNA variants detectable in >1 % or <1 % of the population, respectively

![Image of DNA sequence with a point mutation highlighted]
DNA $\rightarrow$ RNA $\rightarrow$ PROTEIN

CENTRAL DOGMA OF BIOLOGY

*Introduction*

Katherine Joyce, Woods Hole Oceanographic Institution
GROUP WORK

- Think of a disease you want to study or get more information
- Now, write down all databases and bioinformatic tools you know for doing that
DATABASES

- Provide SNP or disease mutation data
- Many genotype-phenotype databases
DATABASE OF SINGLE NUCLEOTIDE POLYMORPHISMS (dbSNP)

- dbSNP has 14+ million uniquely mapped refSNP (rs) numbers
Online Mendelian Inheritance in Man (OMIM)

- As of November 1, 2016, OMIM reported 3,649 genes with a phenotype-causing mutation, and 5,869 phenotypes with a known molecular basis.
*113705

BREAST CANCER 1 GENE; BRCA1

HGNC Approved Gene Symbol: BRCA1

Cytogenetic location 17q21.31  Genomic coordinates (GRCh38): 17:43,044,294-43,125,482

Gene-Phenotype Relationships

<table>
<thead>
<tr>
<th>Location</th>
<th>Phenotype</th>
<th>Phenotype MIM number</th>
<th>Inheritance</th>
<th>Phenotype mapping key</th>
</tr>
</thead>
<tbody>
<tr>
<td>17q21.31</td>
<td>{Breast-ovarian cancer, familial, 1}</td>
<td>604370</td>
<td>AD, Mu</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>{Pancreatic cancer, susceptibility to, 4}</td>
<td>614320</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

TEXT

Description

BRCA1 plays critical roles in DNA repair, cell cycle checkpoint control, and maintenance of several distinct complexes through association with different adaptor proteins, and each complex manner (Wang et al., 2009).

Cloning and Expression

Miki et al. (1994) identified cDNA sequences corresponding to the BRCA1 gene by positional implicated in familial breast-ovarian cancer syndrome (604370). The deduced 1,863-residue protein near the N terminus. A 7.8-kb mRNA transcript was identified in testes, thymus, breast and...
#114480

**BREAST CANCER**

*Alternative titles: symbols*

**BREAST CANCER, FAMILIAL**

Other entities represented in this entry:

**BREAST CANCER, FAMILIAL MALE, INCLUDED**

### Phenotype-Gene Relationships

<table>
<thead>
<tr>
<th>Location</th>
<th>Phenotype</th>
<th>Phenotype MIM number</th>
<th>Inheritance</th>
<th>Phenotype mapping key</th>
<th>Gene/Locus</th>
<th>Gene/Locus MIM number</th>
</tr>
</thead>
<tbody>
<tr>
<td>1p34.1</td>
<td>{Breast cancer, invasive ductal}</td>
<td>114480</td>
<td>AD</td>
<td>3</td>
<td>RAD54L</td>
<td>603615</td>
</tr>
<tr>
<td>2q33.1</td>
<td>{Breast cancer, protection against}</td>
<td>114480</td>
<td>AD</td>
<td>3</td>
<td>CASP8</td>
<td>601763</td>
</tr>
<tr>
<td>2q35</td>
<td>{Breast cancer, susceptibility to}</td>
<td>114480</td>
<td>AD</td>
<td>3</td>
<td>BARD1</td>
<td>601593</td>
</tr>
<tr>
<td>3q26.32</td>
<td>Breast cancer, somatic</td>
<td>114480</td>
<td>AD</td>
<td>3</td>
<td>PIK3CA</td>
<td>171834</td>
</tr>
<tr>
<td>5q34</td>
<td>{Breast cancer, susceptibility to}</td>
<td>114480</td>
<td>AD</td>
<td>3</td>
<td>HMMR</td>
<td>600936</td>
</tr>
<tr>
<td>6p25.2</td>
<td>{?Breast cancer susceptibility}</td>
<td>114480</td>
<td>AD</td>
<td>1</td>
<td>NQO2</td>
<td>160998</td>
</tr>
<tr>
<td>6q25.1-q25.2</td>
<td>{Breast cancer}</td>
<td>114480</td>
<td>AD</td>
<td>1</td>
<td>ESR1</td>
<td>133430</td>
</tr>
<tr>
<td>8q11.23</td>
<td>Breast cancer, somatic</td>
<td>114480</td>
<td>AD</td>
<td>3</td>
<td>RB1CC1</td>
<td>606837</td>
</tr>
<tr>
<td>11p15.4</td>
<td>Breast cancer, somatic</td>
<td>114480</td>
<td>AD</td>
<td>3</td>
<td>SLC22A1L</td>
<td>602631</td>
</tr>
<tr>
<td>11p15.1</td>
<td>Breast cancer, somatic</td>
<td>114480</td>
<td>AD</td>
<td>3</td>
<td>TSG101</td>
<td>601387</td>
</tr>
<tr>
<td>11q22.3</td>
<td>{Breast cancer, susceptibility to}</td>
<td>114480</td>
<td>AD</td>
<td>3</td>
<td>ATM</td>
<td>607585</td>
</tr>
<tr>
<td>12p13.1</td>
<td>Breast cancer, somatic</td>
<td>114480</td>
<td>AD</td>
<td>3</td>
<td>KRAS</td>
<td>190070</td>
</tr>
</tbody>
</table>

**Clinical Synopsis**
#114480
BREAST CANCER

<table>
<thead>
<tr>
<th>CATEGORY</th>
<th>SUBCATEGORY</th>
<th>FEATURES</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inheritance</td>
<td>-</td>
<td>Autosomal dominant</td>
</tr>
<tr>
<td>Neoplasia</td>
<td>-</td>
<td>Breast carcinoma</td>
</tr>
<tr>
<td>Miscellaneous</td>
<td>-</td>
<td>Genetic heterogeneity</td>
</tr>
<tr>
<td>Molecular Basis</td>
<td>-</td>
<td>Caused by mutation in the breast cancer type 1 gene (BRCA1, 113705.0001)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Caused by mutation in the breast cancer type 2 gene (BRCA2, 600185.0001)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Caused by mutation in the solute carrier family 22, member 1-like gene (SLC22A1L, 602631.0001)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Caused by mutation in the tumor protein p53 gene (TP53, 191170.0023)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Caused by mutation in the BRCA1-associated C-terminal helicase 1 gene (BRIP1, 605882.0001)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Caused by mutation in the homolog of the S. cerevisiae RAD51A gene (RAD51A, 179617.0001)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Susceptibility conferred by mutation in the homolog of the S. pombe checkpoint kinase 2 gene (CHEK2, 604373.0007)</td>
</tr>
</tbody>
</table>

Contributors: Kelly A. Przybyla - revised: 3/16/2000
Creation Date: John F. Jackson: 6/15/1995
Edit History: joanna: 05/17/2011
BIOINFORMATIC TOOLS

- Prediction of genes likely to cause or be associated with the disease
- Gene prioritization
**ENDEAVOUR ALGORITHM**

- Endeavour predicts candidate genes
- It requires a list of training set genes and a list of test set genes.

**Bioinformatic tools**
• Anytime you have a list of candidate genes from which you want to select the most promising genes for further validation.
G2D

Candidate Genes to Inherited Diseases

Welcome to the G2D web server (version 5.0). Here you can use our algorithms to scan a human genomic region for genes related to an inherited disease.

Candidate priorities are automatically established by data mining algorithms that evaluate genes in the chromosomal region where the disease is mapped, and prioritize them for a possible relation to the disease based on the phenotype of the disorder or their similarity to an already known related gene. If the phenotype have been linked to more than one locus, known or inferred interactions between proteins from two loci can be also examined. To know more about g2d.

Use G2D

| PHENOTYPE | KNOWN GENES | INTERACTIONS |

LOCUS BOX

Type a locus and select format and chromosome (maximum 100 Mb)

- Band(s) e.g. q13.2
- Marker(s) e.g. D9S201 D9S298
- Positions e.g. 63950000 73950000

- format Positions ▼
- chromosome: Select ▼

KNOWN GENES BOX
ToppGene: Candidate Gene Prioritization

ToppGene Suite

A one-stop portal for gene list enrichment analysis and candidate gene prioritization based on functional annotations and protein interactions network.

- **ToppFun**: Transcriptome, ontology, phenotype, proteome, and pharmacome annotations based on functional similarity.

  Detect functional enrichment of your gene list based on Transcriptome, Proteome, Regulome (mouse phenotype), Pharmacome (Drug-Gene associations), literature co-citation, and other.

- **ToppGene**: Candidate gene prioritization

  Prioritize or rank candidate genes based on functional similarity to training gene list.

- **ToppNet**: Relative importance of candidate genes in networks

  Prioritize or rank candidate genes based on topological features in protein-protein interaction.

- **ToppGenet**: Prioritization of neighboring genes in protein-protein interaction network

  Identify and prioritize the neighboring genes of the seeds in protein-protein interaction network (ToppNet).

- Prioritize or rank candidate genes based on functional similarity
GROUP WORK

- Write down any new databases or bioinformatic tools to get more information about a disease or a mutation from today's lecture.

- Did you manage to write down something you learned from today's lecture?

- Which ones? Can you use them in your real daily work/searching for answers to your research work?
TAKE HOME MESSAGE

- Integrating genome-wide computational gene prioritization with large-scale genetic screening is a powerful tool for functional gene discovery.

- On the other hand, computational predictions of gene function alone remain far from being accurate enough to be considered high-quality biological data.
Please send homeworks to mariliipalover@gmail.com by 11th November 10:00
HOMEWORK

1. What are the differences between dbSNP and genotype-phenotype databases?

2. Why is it important to work out new molecular tools for gene discovery?

3. What does gene prioritization mean? When and Why do we need gene prioritization?

4. Which of the three bioinformatic tool you like best, why?
HOMEWORK

5. Imagine, you have a dataset to study. You get SNP numbers (on the last slide) from the data, and you want to know:

- In what gene does it occur?
- Does it cause any disease?
- If yes, then what disease(s)?
- Is it associated with single-gene or complex disease?
Homework

6. Sometimes, no mutations are found in suspected disease-related genes, but mutations are found in other genes whose relationship to a particular genetic condition is unknown.

- What database can you overall use to get more information of those genes?
- What bioinformatic tools can you use to predict the association of those genes to specific disease?

6.1 Exercise:
Make a set of training genes (search them from the databases with the given disease that I gave you (on the last slide)) and candidate genes (you can think them up if you want) using any database or bioinformatic tool and write down some of the results you get (doesn’t matter is the result good or bad, just something to see, how tools are working)
SOME USEFUL LINKS

- ENDEAVOUR:  
- G2D:  
  http://g2d2.ogic.ca/
- ToppGene:  
  https://toppgene.cchmc.org/
<table>
<thead>
<tr>
<th>Name</th>
<th>For ex.5</th>
<th>For ex.6.1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sulev Reisberg</td>
<td>rs17244841</td>
<td>Lung Cancer</td>
</tr>
<tr>
<td>Anna Ufliand</td>
<td>rs7412</td>
<td>Multiple Sclerosis</td>
</tr>
<tr>
<td>Liis Kolberg</td>
<td>rs4693075</td>
<td>Albinism</td>
</tr>
<tr>
<td>Ahto Salumets</td>
<td>rs35599367</td>
<td>Polycystic kidney disease</td>
</tr>
<tr>
<td>Kreete Lüll</td>
<td>rs776746</td>
<td>Alzheimer disease</td>
</tr>
<tr>
<td>Timo Tõnis Sikka</td>
<td>rs428785</td>
<td>Dementia</td>
</tr>
<tr>
<td>Alexandra Cicmancova</td>
<td>rs2231142</td>
<td>Parkinson Disease</td>
</tr>
<tr>
<td>Andrii Rozumnyi</td>
<td>rs2032582</td>
<td>Myopathy</td>
</tr>
<tr>
<td>Ants-Oskar Mäesalu</td>
<td>rs2306283</td>
<td>Prostate Cancer</td>
</tr>
<tr>
<td>Gagandeep Singh</td>
<td>rs4149015</td>
<td>Prader-Willi syndrome</td>
</tr>
<tr>
<td>Jakub Dobrik</td>
<td>rs4149056</td>
<td>Stroke</td>
</tr>
<tr>
<td>Sander Tars</td>
<td>rs2630578</td>
<td>Hypercholesterolemia</td>
</tr>
<tr>
<td>Yurii Toma</td>
<td>rs10786775</td>
<td>Bladder Cancer</td>
</tr>
<tr>
<td>Artem Bachynskyi</td>
<td>rs16847897</td>
<td>Cystic Fibrosis</td>
</tr>
</tbody>
</table>