LIST OF FIGURES **2**
LIST OF TABLES **2**
KEYWORDS **3**
ABSTRACT **3**
INTRODUCTION **3**
Project Goal 3
Seven Basic Questions 4
A Guiding Principle 4
The Power of Visualization 4
OMIM 4
OMIM Entry Types 6
COSMIC 6
METHODS **6**
Question 1 Method 6
Question 2 Method 7
Setback: 43 histological types of BC and 4 histological types of PC 7
Advance: two histologically types account for 51% of BC and 1 type accounts for 95% of PC 7
Data Sources for Counting Genes 8
The Informal 'Top Offenders' 8
Clarify the Query 10
Question 2 Code: 10
Question 3 Method 11
Setback: Location Data Requires Cleaning 11
Advance: Clean the Data 11
Question 4 Method 12
Question 5 Method 14
Question 6 Method 15
Question 7 Method 16
A Gene Connection 'Microscope' 16
RESULTS **17**
Question 1 Results 17
Question 2 Results 18
Question 3 Results 18
Question 4 Results 20
Question 5 Results 21
Question 6 Results 21
Question 7 Results 22
Other Results – Prostate Cancer Not as Well Researched as BC 24
DISCUSSION / CONCLUSIONS **25**
ACKNOWLEDGEMENTS **25**
REFERENCES **26**

LIST OF FIGURES

Figure 1 - OMIM Gene Entry for PTEN 5
Figure 2 - Text Annotations for OMIM Gene Entries 5
Figure 3 – Setback #1: The 43 Histological Types of BC/PC 7
Figure 4 – Advance #1: Two Histologies Account for 51% of Breast Cancer 7
Figure 5 – Advance #1 Cont'd: Two Histologies Account for 98% of Prostate Cancer 8
Figure 6 – The Informal Top Offender Genes in BC/PC 9
Figure 7 – Transforming the Informal Top Offender Genes in BC/PC 9
Figure 8 - R Notebooks for OMIM Queries and Processing 12
Figure 9 - OMIM Gene 'Discovery' Sorted by Date 14
Figure 10 - Class Chat Excerpt that gave rise to this survey. 14
Figure 11 - Computation of Smooth Derivatives for Rate of Discovery 15
Figure 12 - kg: The "Knowledge Gazer" Graph Visualization Program 16
Figure 13 - Who is affected by breast and prostate cancer (BC/PC)? 17
Figure 14 - The Answer to Question 2: How many BC/PC genes are known? 18
Figure 15 - Answer to Question 3: Where are the BC/PC genes in the genome? 18
Figure 16 - Answer to Question 3: Zooming In By Scaling BC/PC Gene Counts 19
Figure 17 - Computed Data for BC/PC Distribution by Chromosome 19
Figure 18 - Summary of Findings for Question 3: Where are BC/PC genes? 20
Figure 19 - Discovery Curves for BC/PC Genes 20
Figure 20 - Rate of Discovery Curves for BC/PC Genes 21
Figure 21 - The Lion's Share of Discoveries Appear Done 21
Figure 22 - Total Genes and Genes in Common for Question 7 22
Figure 23 - Edge Counts for Genes in BC/PC Union and Intersection 22
Figure 24 - A Snapshot of Prostate Cancer Gene Connections 23
Figure 25 - A Ring of Fifteen Cancer Genes 23
Figure 26 - Gene Connection Tallies for BC/PC and Intersection 24
Figure 27 - Finding: Prostate Cancer Not as Researched as Breast Cancer 25

LIST OF TABLES

Table 1 – OMIM Numbered Entry Types 6

A Visual Survey of Breast and Prostate Cancer Genes

L. Van Warren
Department of Information Science and Bioinformatics
University of Arkansas at Little Rock
December 2020

KEYWORDS

Bioinformatics, computational biology, genetics, breast cancer, prostate cancer, oncogenes, tumor suppressor genes, protein structures, selection algebras, science visualization

ABSTRACT

This survey paper compares and contrasts the genes involved in breast and prostate cancer (abbreviated as *BC*/*PC*) by asking seven basic questions. The method of answering these questions consisted of scientific literature review along with the standard {who, what, when, how, why, and how many} questions of investigative reporting. Visualization of gene networks was accomplished using special purpose software to display connections between *BC*/*PC* genes. The OMIM and COSMIC data repositories were used as the primary sources of genetic information for answering all of the questions. The findings herein include the incidental one that prostate cancer has only half of the web engagement of breast cancer, as measured by user queries. The current *BC*/*PC* genes count is 672/434 respectively with 167 genes in common and 939 unique genes total. *BC*/*PC* cancer genes are evenly distributed throughout the genome with a σ of 1%. More statistical findings are explored below along with a visual exploration of their startling interconnection complexity.

INTRODUCTION

The advent of the Human Genome Project, the collaboration between government and private enterprise led by Francis Collins on the government side and Craig Venter on the private side created a windfall of understanding the genes that make us who we are. This was followed by a two-decade spurt of disease gene identification cataloging by OMIM on the germ line side, and by COSMIC on the acquired mutation side.

Project Goal

The goal of the current survey was to develop a 'lay of the land' summary by asking seven basic questions about the nature of breast and prostate cancer, culminating in the last question on the connections their genes have to each other. These are questions whose answers have been significantly clarified – but not completely answered – in the past two decades.

Initially this survey was intended to include breast cancer (BC) only, but it turned out to be reasonable to include prostate cancer (PC) as well. By including this second cancer an advantage was obtained, and that was the ability to compare and contrast one cancer against another. Further enhancing this advantage was that these are largely gender-specific diseases, but males do carry a 1 in 883 chance of developing breast cancer. This same approach could be repeated for any pair of related conditions, or for that matter any single condition without loss of generality.

Seven Basic Questions

This survey paper compares and contrasts the genes involved in breast and prostate cancer by investigating the following seven questions:

1. *Who is affected by breast and prostate cancer (BC/PC)?*
2. *How many BC/PC genes are known?*
3. *Where are the BC/PC genes in the genome?*
4. *When were BC/PC genes discovered?*
5. *What is the rate of discovery of BC/PC genes?*
6. *How many BC/PC genes remain to be discovered?*
7. *What are the connections between genes?*

A Guiding Principle

Separate What from How

Describes the important discipline of separating what we found from how we figured it out. This is equivalent to clearly separating ***methods*** from ***results***.

The Power of Visualization

One can associate breakthroughs in science with the advent of tools that enabled visualization of the underlying phenomena. In clinical lab diagnostics, spectrometers made precise measurements of colorimetric variations that were first observed visually on a bench and extended to non-visible wavelengths. Optical, confocal and electron microscopes enable direct visualization of cells, stained and marked with quantum dots or fluorophores to reveal their inner workings. More recently virtual visualizations enable inspection of the direct structure of individual proteins and nucleic acid sequences. This survey introduces a new kind of visualization, "A Gene Connection Scope". It addresses the complex interconnection of genes within the cell.

In the large visualization:

- Enhances the communication of work, increasing its reach
- Enables access by domain experts, educators and the general public
- Accelerates scientific, medical and social progress

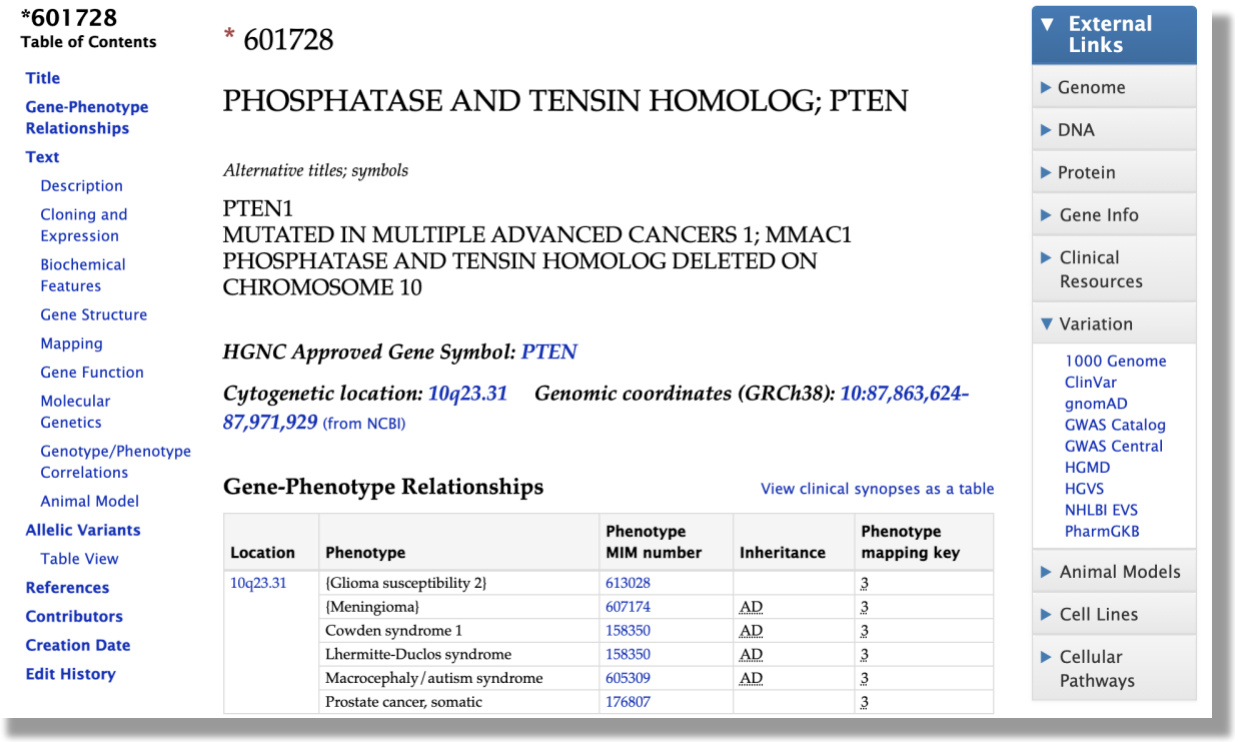
The visualizations herein include:

- Getting to know the genomic landscape of cancer
- The genes most involved in carcinogenesis and their distribution in the genome.

OMIM

According to the OMIM FAQ, "Online Mendelian Inheritance in Man (OMIM) is a continuously updated catalog of human genes and genetic disorders and traits, with a particular focus on the gene-phenotype relationship." This database holds information on all known Mendelian diseases and more than 15,000 of the 20,399 coding genes in the human genome. The database is an outgrowth of a systematic catalog of twelve printed editions that were created by Victor McKusick between 1966 and 1968. OMIM is a collaboration between the National Library of Medicine and the Medical Library at Johns Hopkins which began in 1985 and continues to this day.

A typical OMIM gene entry looks like this:



*601728
Table of Contents

Title
Gene-Phenotype Relationships
Text
Description
Cloning and Expression
Biochemical Features
Gene Structure
Mapping
Gene Function
Molecular Genetics
Genotype/Phenotype Correlations
Animal Model
Allelic Variants
Table View
References
Contributors
Creation Date
Edit History

* 601728

PHOSPHATASE AND TENSIN HOMOLOG; PTEN

Alternative titles; symbols

PTEN1
MUTATED IN MULTIPLE ADVANCED CANCERS 1; MMAC1
PHOSPHATASE AND TENSIN HOMOLOG DELETED ON CHROMOSOME 10

HGNC Approved Gene Symbol: PTEN

Cytogenetic location: 10q23.31 ***Genomic coordinates (GRCh38): 10:87,863,624-87,971,929*** (from NCBI)

Gene-Phenotype Relationships

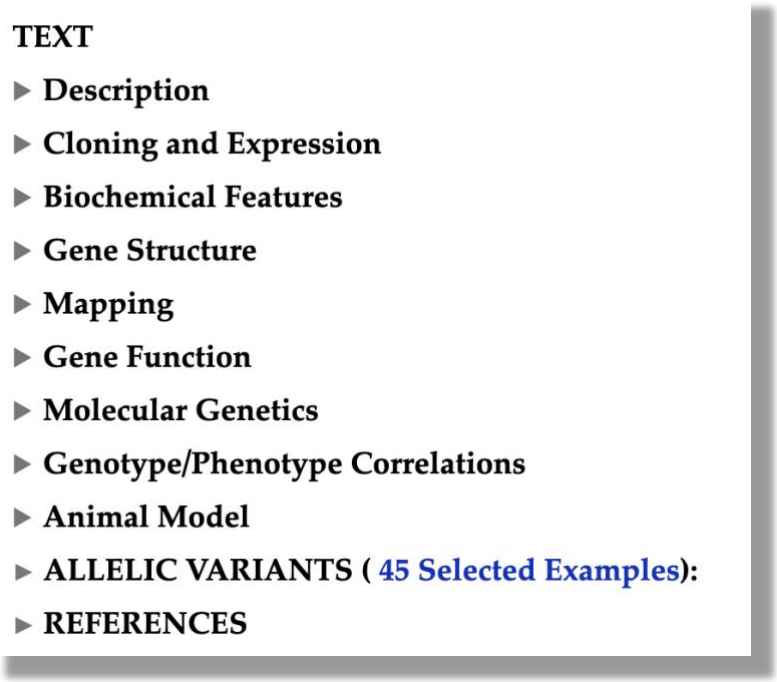
View clinical synopses as a table

Location	Phenotype	Phenotype MIM number	Inheritance	Phenotype mapping key
10q23.31	{Glioma susceptibility 2}	613028		3
	{Meningioma}	607174	AD	3
	Cowden syndrome 1	158350	AD	3
	Lhermitte-Duclos syndrome	158350	AD	3
	Macrocephaly/autism syndrome	605309	AD	3
	Prostate cancer, somatic	176807		3

External Links
Genome
DNA
Protein
Gene Info
Clinical Resources
Variation
1000 Genome
ClinVar
gnomAD
GWAS Catalog
GWAS Central
HGMD
HGVS
NHLBI EVS
PharmGKB
Animal Models
Cell Lines
Cellular Pathways

Figure 1 - OMIM Gene Entry for PTEN

It is followed by a textual description with the following annotations.



TEXT
- Description
- Cloning and Expression
- Biochemical Features
- Gene Structure
- Mapping
- Gene Function
- Molecular Genetics
- Genotype/Phenotype Correlations
- Animal Model
- ALLELIC VARIANTS (45 Selected Examples):
- REFERENCES

Figure 2 - Text Annotations for OMIM Gene Entries

OMIM Entry Types

There are four kinds of numbered entries in OMIM:

Asterisk	*****	**a gene**
Sharp	**#**	**a descriptive entry**
Plus	**+**	**a gene plus phenotype**
Percent	**%**	**a phenotype**

Table 1 – OMIM Numbered Entry Types

For genes we examine we must count the number of Asterisks (*) and Pluses (+) or we will MISS genes like CHEK2 which are only listed under Plus. Descriptive and Phenotypic entries are not used by the automation used here. Victor McKusick's creation of OMIM is discussed in Siddhartha Mukherjee's excellent opus, "The Gene: An Intimate History".

COSMIC

According to About COSMIC – the Catalogue of Somatic Mutations in Cancer – "is the world's largest source of expert manually curated somatic mutation information relating to human cancers." COSMIC consists of curated data from over 27,000 peer reviewed papers which focus on known and suspected cancer genes and genome-wide screening data from 37,000 peer reviewed datasets. COSMIC provides a finer brush when it comes to distinguishing histologically defined cancer types than does OMIM and that solved a major problem for this survey.

METHODS

The overall method used here is to ask the questions effective investigators ask. These are the standard questions you see in descriptive journalism, and they also apply to science. It consists of combining scientific literature review along with the {who, what, when, how, why, and how many} questions of investigative reporting. The specific methods are covered below:

Question 1 Method

Who is affected by breast and prostate cancer (BC/PC)?

This is answered in the Results section below. It was answered by comparing, contrasting and culling the facts presented in these four sources and synthesizing the summary figure presented in the Results section.

http://www.breastcancer.org/symptoms/understand_bc
http://www.cancer.net/cancer-types/prostate-cancer/risk-factors-and-prevention
http://www.cancer.net/cancer-types/prostate-cancer/statistics
http://www.cancer.org/cancer/prostate-cancer/about/key-statistics.html

Question 2 Method

How many BC/PC genes are known?

Setback: 43 histological types of BC and 4 histological types of PC

A problem immediately appears when we consult COSMIC where we discover there are 43 histological types of breast cancer and 4 histological types of prostate cancer as listed in the figure below:

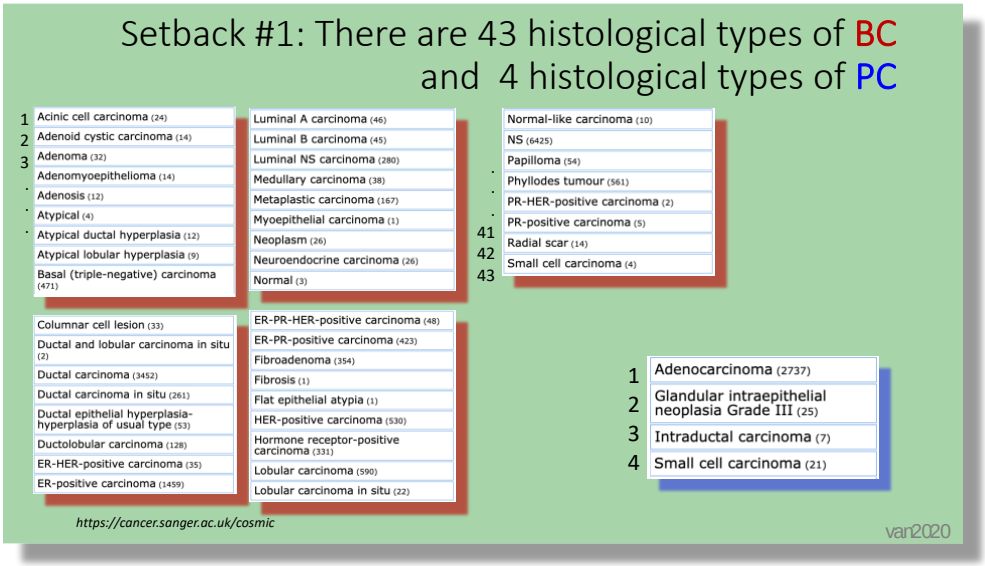


Figure 3 – Setback #1: The 43 Histological Types of BC/PC

Advance: two histologically types account for 51% of BC and 1 type accounts for 95% of PC

This problem is significantly remediated when we cut, paste and tabulate to find that the first two histological types account for 51% of all breast cancer.

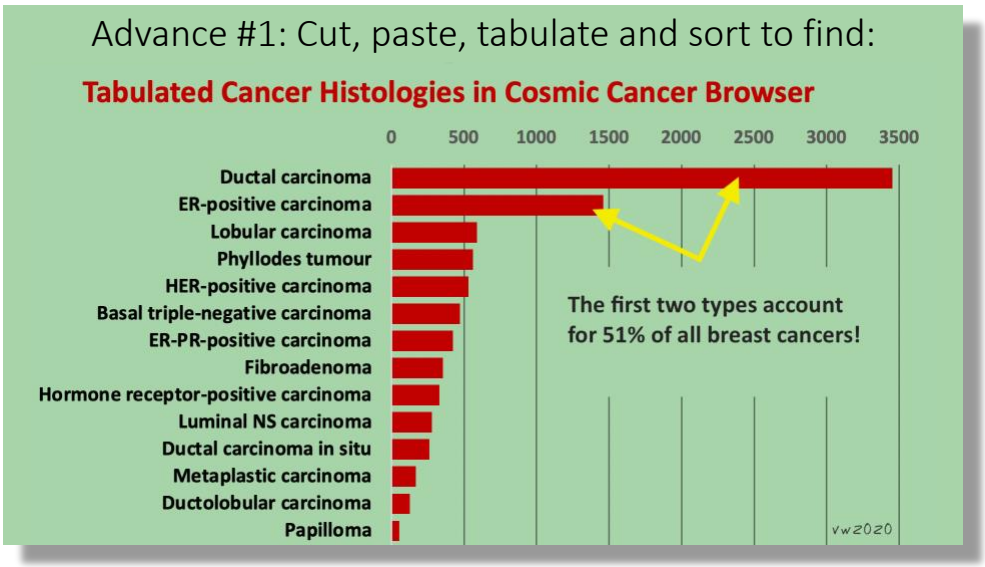


Figure 4 – Advance #1: Two Histologies Account for 51% of Breast Cancer

The situation of histological diversity is even less problematic in prostate cancer when one histological type, adenocarcinoma, accounts for 98% of all cases.

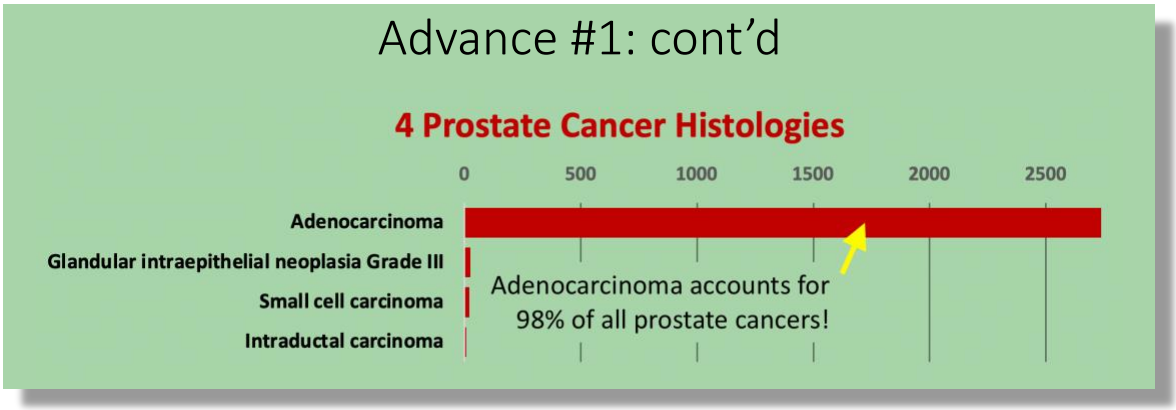


Figure 5 – Advance #1 Cont'd: Two Histologies Account for 98% of Prostate Cancer

This difference in histological diversity was important to keep in mind when considering the connectivity of the breast cancer and prostate cancer gene sets.

Data Sources for Counting Genes

There are three sources to explore when counting the *BC/PC* genes:

- The Informal 'Top Offender Genes' *BC/PC*
- The Germ Line Mutations Catalog OMIM for *BC/PC*
- The Somatic Mutations Cataloged in COSMIC for *BC/PC*

This survey focused on the first two, COSMIC was used to solve the proliferation of histological types problem.

The Informal 'Top Offenders'

First the Informal 'Top Offender Genes' for BC and PC was examined. The intention of the term 'Informal' is to communicate the list of genes which clinical testing panels use.

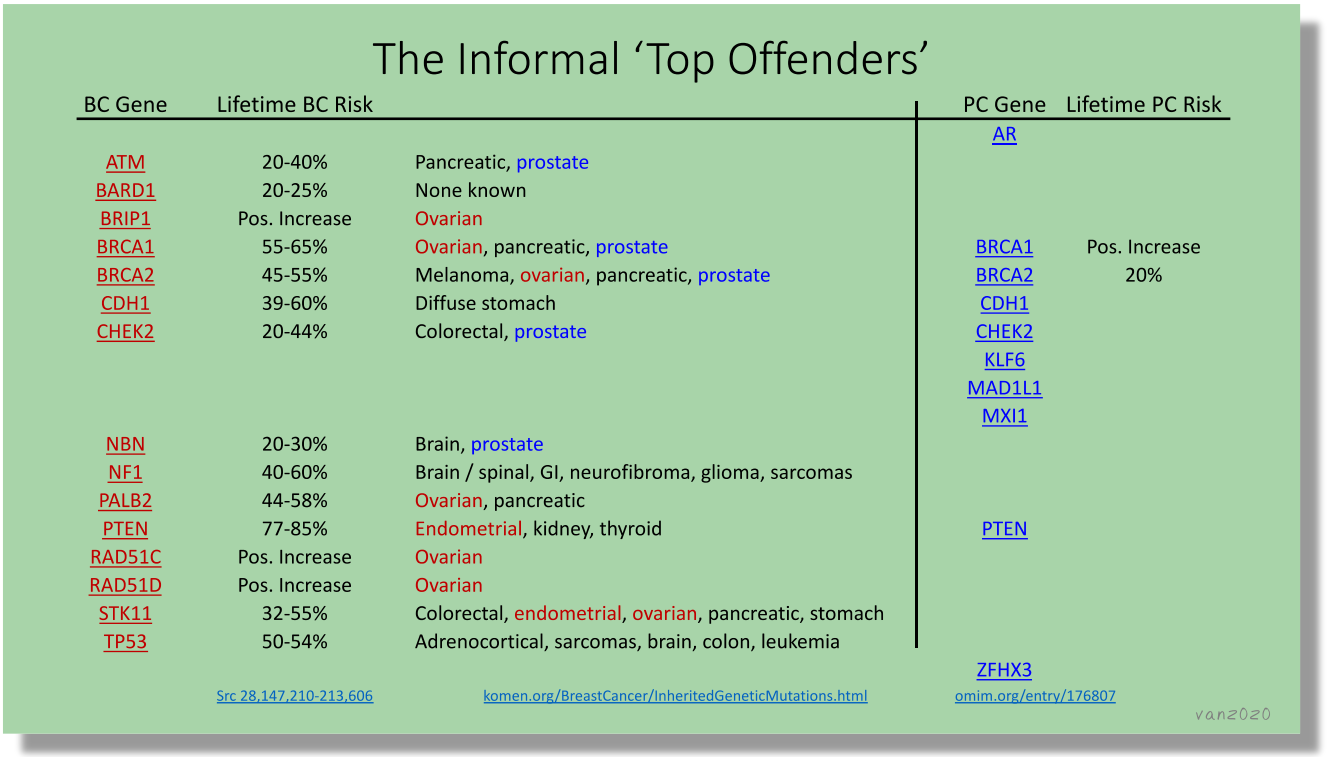


Figure 6 – The Informal Top Offender Genes in BC/PC

Looking at the comparative table suggested a refinement to the list of Figure 4 below:
The structure of this table asks the question, "Could any of the genes not listed in one group actually belong to the other?" This question can be asked visually, and as we shall see in our connections analysis, at least one of them, Androgen Receptor (AR) is! This is a finding that must be included in the results section below.

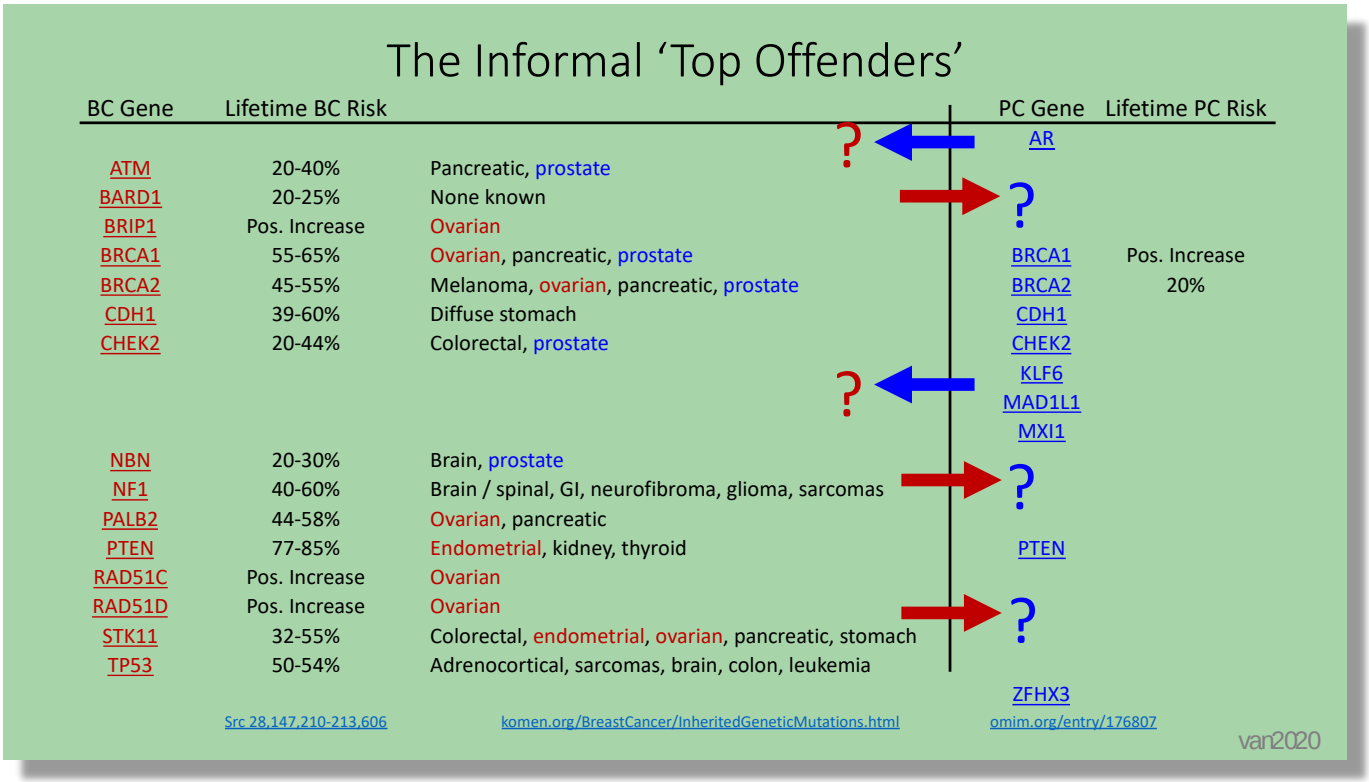


Figure 7 – Transforming the Informal Top Offender Genes in BC/PC

Question 2, "*How many BC/PC genes are known?",* had other twists and turns. The second setback encountered was that subtle differences in query structure produced different answers. Consider the two variants of this question:

Q2: How many *BC/PC genes* are in OMIM?

vs.

Q2: How many *BC/PC entries* are in OMIM?

This setback was remedied or advanced in two ways:

- State the question that is actually *answered*.
- State assumption that makes the answer true for the original question.

Clarify the Query

In this case the assumption is that there is an entry for every gene, the accuracy of which we will explore below. By using the correct tags, star '*' and plus '+' to correctly filter then OMIM entries, the correctness of our gene counts was improved over naïve first runs.

Question 2 Code:

There is no clearer statement of method than source code. The bash script *countOMIM-Genes.bash* performed the count. Internally it looked like this:

```
#!/bin/bash
echo -n 'BC Gene Count:'
postProcessOMIM.bash OMIM-BC-Genes.txt | wc -l
echo -n 'PC Gene Count:'
postProcessOMIM.bash OMIM-PC-Genes.txt | wc -l
```

Running *countOMIM-Genes.bash.bash* produces the gene counts for this survey which were:

```
BC Gene Count:      672
PC Gene Count:      434
```

The files *OMIM-BC-Genes.txt* and *OMIM-PC-Genes.txt* were downloaded from OMIM using the query *breast+cancer* and *prostate+cancer* in the search bar. This can be done directly using the URL's:

BC: *https://www.ncbi.nlm.nih.gov/omim/?term=%22breast+cancer%22*
PC: *https://www.ncbi.nlm.nih.gov/omim/?term=%22prostate+cancer%22*

The input text files (.txt) were downloaded from OMIM and had this format:

```
1.  113721 - BREAST CANCER-RELATED REGULATOR OF TP53
Cytogenetic locations: 17p13.3
OMIM: 113721

2. *613746 - BREAST CANCER ANTIESTROGEN RESISTANCE 4; BCAR4
Cytogenetic locations: 16p13.13
OMIM: 613746
```

Question 3 Method

Where are the BC/PC genes in the genome?

The intention of this question is to find the population distribution of the *BC/PC* genes in the genome, as opposed to their specific cytogenetic locations which are already listed in OMIM. Specifically, the goal was to understand whether particular chromosomes contain clusters of cancer genes.

Setback: Location Data Requires Cleaning

- Some genes have locations but not names.
- Some genes have names but not locations.
- Some entries have extra blank lines.
- There are white space errors in the data.

Advance: Clean the Data

- Some genes have locations but not names. PRESERVED
- Some genes have names but not locations. DELETED
- Some entries have extra blank lines. DELETED
- There are white space errors in the data. REPAIRED ← This was hard!

To make sure that cleaned data was not corrupted on subsequent downloads, a goal was set that no entries would be cleaned manually, but addressing all the edge cases was difficult.

We see this in the *postProcessOMIM.bash* script below, which was also used in gene counting above.

```
#! /bin/bash
cat $1                                          |\ # Set shell
sed 's/INCLUDEDCyto/INCLUDED\                      # Fix white space problem
Cyto/'                                          |\ # macOS newline issue
grep -v OMIM                                    |\ # Suppress spurious lines
sed 'N;/^\n$/D;P;D;'                            |\ # Delete extra newlines
sed 'N;s/\n\s*Cytogenetic/; Cytogenetic/;P;D' |\ # Fuse records to one line
sed '/^$/d'                                     |\ # Remove blank lines
sed 'N;s/\n\s*\([A-Z]\.*\)/; \1/;P;D'           |\ # Preserve GENE name.
grep '\*\|\+'                                   |\ # Find GENES ONLY
sed 's/ - [^;+]*;/;/'                           |\ # Remove verbose descrip.
sed 's/\(; .*;\).*;/\1;/'                       |\ # Trim to gene name
grep Cytogenetic                                       # PRESERVE GENE LOCATION
```

The *postProcessOMIM.bash script* above was called by the *locateOMIM-genes.bash* script below, which actually tallied the locations:

```
#!/bin/bash                     # Set shell
postProcessOMIM.bash $1      |\ # Clean data
sed 's/^.* [\*\+]//'         |\ # Remove preamble
sed 's/Cyto.*locations: //' |\ # Remove extra text
sed 's/.*; //'               |\ # Remove long name
sort -n                      |\ # Sort by location
sed 's/[pq].*//'             |\ # Scope to chromosome
uniq -c                         # Count locations
```

This process produced the answers to question 3 for the project:

Where are the BC/PC genes in the genome?

The graphs of and the tabulating spreadsheet are in the Results Section below.

Question 4 Method

When were BC/PC genes discovered?

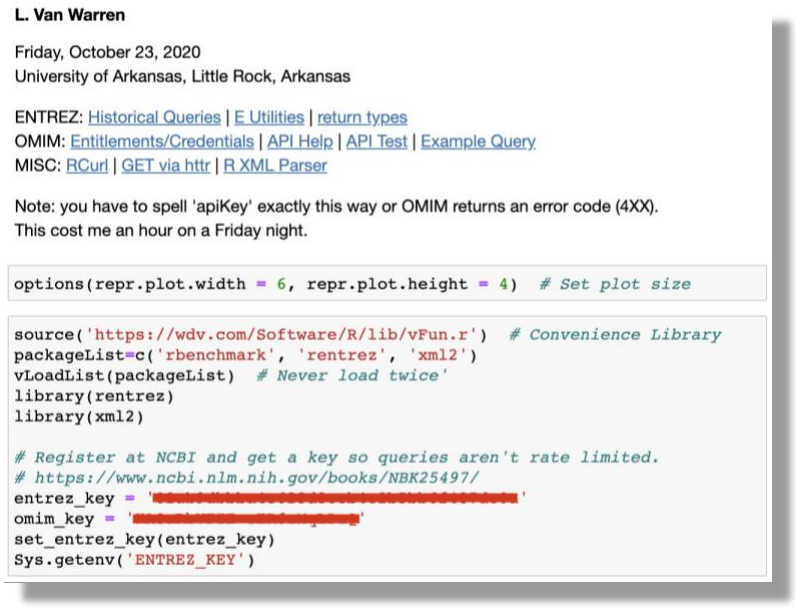
One could say, “What does it matter WHEN the genes were discovered? It matters only that we have them now, so get back to work”. This shortsighted approach neglects the fact that logging the times at which each gene was discovered enables us to estimate how far along we are in that discovery process.

Logging our discovery times allows us to say, “We knew this much at this time.”

Hypothesis: If we can identify a high rate of discovery followed by a lull, we have accumulated a significant amount of available knowledge.

Again, there was an element of imprecision here. The actual discovery time is not as important, as the date at which the specific information for the gene became widely available. OMIM was the perfect vehicle for measuring the *intended* question since it has been logging these genes promptly since before the Human Genome Project enabled such ubiquitous discovery.

If one wants to repeatedly hit the OMIM servers for various gene entries, then an account, and freely available key are required. Otherwise, bad things happen, no data is returned, or it is returned at a throttled rate. OMIM provides an excellent API, described here, which explains how to fetch particular items in a gene entry. The section 'Advanced Counting' allows us to determine when particular gene entries were made in OMIM. This was all done in R, run from a Jupyter notebook which was reproducible and convenient.



L. Van Warren

Friday, October 23, 2020
University of Arkansas, Little Rock, Arkansas

ENTREZ: Historical Queries | E Utilities | return types
OMIM: Entitlements/Credentials | API Help | API Test | Example Query
MISC: RCurl | GET via httr | R XML Parser

Note: you have to spell 'apiKey' exactly this way or OMIM returns an error code (4XX).
This cost me an hour on a Friday night.

```
options(repr.plot.width = 6, repr.plot.height = 4)  # Set plot size
```

```
source('https://wdv.com/Software/R/lib/vFun.r')  # Convenience Library
packageList=c('rbenchmark', 'rentrez', 'xml2')
vLoadList(packageList)  # Never load twice'
library(rentrez)
library(xml2)

# Register at NCBI and get a key so queries aren't rate limited.
# https://www.ncbi.nlm.nih.gov/books/NBK25497/
entrez_key = '[redacted]'
omim_key = '[redacted]'
set_entrez_key(entrez_key)
Sys.getenv('ENTREZ_KEY')
```

Figure 8 - R Notebooks for OMIM Queries and Processing

The data was read from OMIM using the R code listed below:

```
makeURL = function(number)
{
    prolog = 'https://api.omim.org/api/entry?mimNumber='
    epilog = '&apiKey=KAOaRbVFTE-eER6zMqLDuQ&include=creationDate&format=xm
    return(paste(prolog, number, epilog, sep=''))
}
```

```
x_list = as_list(read_xml(makeURL(113705)))
x_list$omim$entryList$entry$creationDate
```

1. 'Victor A. McKusick : 12/20/1990'

```
bcData = read.csv("../GeneCounting/OMIM-BC-Genes-Post.txt", sep=";", header
pcData = read.csv("../GeneCounting/OMIM-PC-Genes-Post.txt", sep=";", header
```

```
length(bcData$Number)
length(pcData$Number)
```

672

434

After the gene counts were confirmed, the fetch ran in a simple loop and wrote into the local user directory.

```
for(i in 1:nrow(bcData))
{
    cat(bcData$Name[i], '; ')
    x_list = as_list(read_xml(makeURL(bcData$Number[i])))
    entry = toString(x_list$omim$entryList$entry$creationDate)
    entry = sub(':', ';', entry)
    cat(entry, '\n')
}
```

```
BCAR4 ; Patricia A. Hartz ; 2/17/2011
 FAM84B ; Patricia A. Hartz ; 7/19/2005
 BCAS1 ; Victor A. McKusick ; 8/14/1998
 BCAR3 ; Paul J. Converse ; 3/20/2000
 AGR3 ; Patricia A. Hartz ; 7/19/2005
 BPIFA4P ; Victor A. McKusick ; 3/14/2003
 LINC01488 ; Patricia A. Hartz ; 09/26/2017
 ANKRD30A ; Patricia A. Hartz ; 3/16/2007
 AKIP1 ; Patricia A. Hartz ; 2/4/2005
 VWA5A ; Rebekah S. Rasooly ; 8/4/1998
 NQO1 ; Victor A. McKusick ; 6/4/1986
```

The resulting data was collected in a spreadsheet for sorting, postprocessing and visualization:

	A	B	C	D	E	F	G
1	1	ABCB1	6/2/1986	Victor A. McKusick	ERBB2	Victor McKusick	6/2/86
2	2	AKT1	6/2/1986	Victor A. McKusick	TP53	Victor McKusick	6/2/86
3	3	ERBB2	6/2/1986	Victor A. McKusick	KRAS	Victor McKusick	6/2/86
4	4	ETS1	6/2/1986	Victor A. McKusick	AKT1	Victor McKusick	6/2/86
5	5	ETS2	6/2/1986	Victor A. McKusick	INHA	Victor McKusick	6/2/86
6	6	HRAS	6/2/1986	Victor A. McKusick	AMD1	Victor McKusick	6/2/86
7	7	IGF1	6/2/1986	Victor A. McKusick	IGF1	Victor McKusick	6/2/86
8	8	IGF1R	6/2/1986	Victor A. McKusick	CD8A	Victor McKusick	6/2/86
9	9	IGF2	6/2/1986	Victor A. McKusick	MUC1	Victor McKusick	6/2/86
10	10	KRAS	6/2/1986	Victor A. McKusick	TF	Victor McKusick	6/2/86
11	11	MTR	6/2/1986	Victor A. McKusick	KIT	Victor McKusick	6/2/86
12	12	MUC1	6/2/1986	Victor A. McKusick	AR	Victor McKusick	6/4/86
13	13	PPY	6/2/1986	Victor A. McKusick	EGFR	Victor McKusick	6/4/86
14	14	REL	6/2/1986	Victor A. McKusick	CYP3A4	Victor McKusick	6/4/86
15	15	SDHB	6/2/1986	Victor A. McKusick	ESR1	Victor McKusick	6/4/86
16	16	TGFB1	6/2/1986	Victor A. McKusick	GHRH	Victor McKusick	6/4/86
17	17	TP53	6/2/1986	Victor A. McKusick	GLS	Victor McKusick	6/4/86
18	18	VIM	6/2/1986	Victor A. McKusick	EGF	Victor McKusick	6/4/86
19	19	AR	6/4/1986	Victor A. McKusick	CTSB	Victor McKusick	6/4/86

Figure 9 - OMIM Gene 'Discovery' Sorted by Date

The plot of gene discovery versus time is presented in the results section below.

Question 5 Method

What is the rate of discovery of BC/PC genes?

This question was the engine that drove the construction of this entire survey! It arose during a side chat in the graduate bioinformatics class taught by Dr. Mary Yang at University of Arkansas, Little /rock. The course itself was launched in cyberspace and telepresence by the Covid-19 pandemic that began in 2020. An excerpt t was recorded here in Bioinformatics 5445 Lecture4, Recording #7.

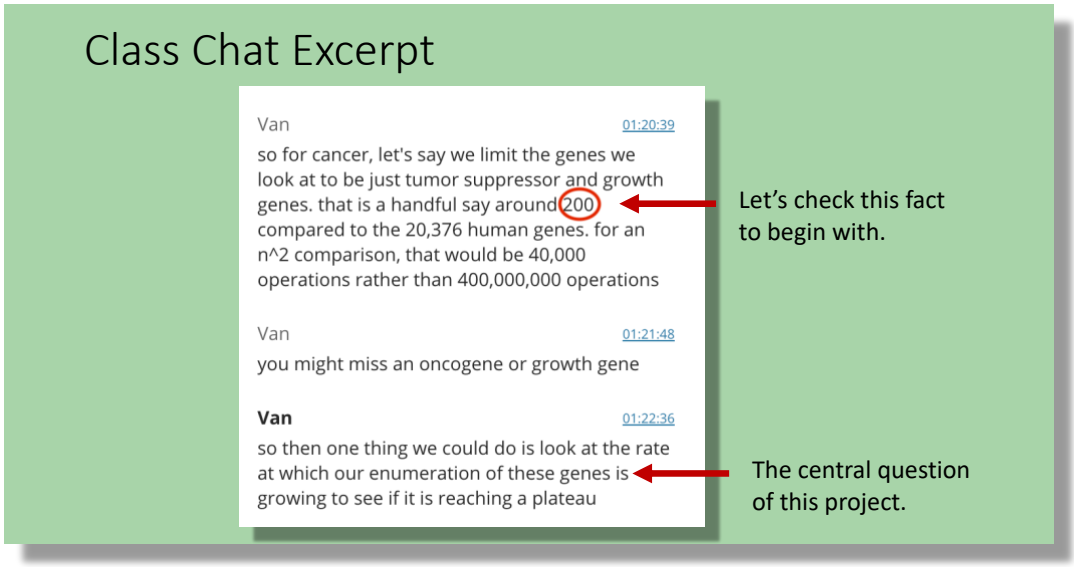


Figure 10 - Class Chat Excerpt that gave rise to this survey.

To answer this question it was necessary to use a 60-entry rolling average to smooth the jitter introduced by computing the derivatives of discoveries over time to obtain the rate.

Q5: What is rate of discovery of BC/PC genes?

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
2	18	VIM	31565.00	Victor A. McKusick	dy	dt	dy/dt	BC_dy/dt		1	EGF	Victor McKusick	6/4/86		dy	dt	dy/dt	PC_dy/dt
3	33	TFF1	6/4/1986	Victor A. McKusick	15	2	7.50			2	SEMG1	Victor McKusick	3/20/89		1	1020	0.0010	
61	106	VCAM1	8/8/1991	Victor A. McKusick	2	15	0.13			62	BIN1	Victor McKusick	5/7/96		1	7	0.1429	
62	107	FPR1	8/21/1991	Victor A. McKusick	1	13	0.08	0.28		63	LIMK1	Victor McKusick	6/27/96		1	51	0.0196	0.0866
63	108	RPA1	9/4/1991	Victor A. McKusick	1	14	0.07	0.15		64	MAP2K4	Victor McKusick	7/5/96		1	8	0.1250	0.0868
64	109	ARID4A	9/12/1991	Victor A. McKusick	1	8	0.13	0.15		65	CDH13	Moyra Smith	8/8/96		1	34	0.0294	0.0866
65	110	CCND1	9/19/1991	Victor A. McKusick	1	7	0.14	0.15		66	KAT5	Mark Paalman	8/31/96		1	23	0.0435	0.0849
66	111	CCNE1	10/4/1991	Victor A. McKusick	1	15	0.07	0.15		67	TUSC3	Alan Scott	9/16/96		1	16	0.0625	0.0851
67	112	TERT	10/30/1991	Victor A. McKusick	1	26	0.04	0.15		68	PPARG	Jennifer Macke	11/4/96		1	49	0.0204	0.0859
68	113	GSTP1	11/15/1991	Victor A. McKusick	1	16	0.06	0.15		69	GRB14	Lori Kelman	11/20/96		1	16	0.0625	0.0851
69	114	LAG3	12/11/1991	Victor A. McKusick	1	26	0.04	0.15		70	EVPL	Moyra Smith	12/19/96		1	29	0.0345	0.0852
70	115	TFDP1	12/19/1991	Victor A. McKusick	1	8	0.13	0.15		71	ESR2	Lori Kelman	2/4/97		1	47	0.0213	0.0854
71	117	RUNX1	1/27/1992	Victor A. McKusick	2	39	0.05	0.15		72	PTEN	Victor McKusick	3/27/97		1	51	0.0196	0.0860
72	119	XBP1	3/3/1992	Victor A. McKusick	2	36	0.06	0.15		73	CASP8	Jennifer Macke	4/18/97		1	22	0.0455	0.0866
73	120	NFKB1	3/6/1992	Victor A. McKusick	1	3	0.33	0.16		75	BPTF	Alan Scott	5/20/97		2	32	0.0625	0.0862
74	121	F2R	5/6/1992	Victor A. McKusick	1	61	0.02	0.15		76	JAG1	Victor McKusick	7/7/97		1	48	0.0208	0.0859
75	122	MMP1	5/28/1992	Victor A. McKusick	1	22	0.05	0.15		77	PAWR	Jennifer Macke	7/24/97		1	17	0.0588	0.0862
76	123	TFF2	8/14/1992	Victor A. McKusick	1	78	0.01	0.15		78	DMBT1	Victor McKusick	9/3/97		1	41	0.0244	0.0883
77	124	IGFBP5	8/27/1992	Victor A. McKusick	1	13	0.08	0.15		79	NCOA4	Victor McKusick	9/10/97		1	7	0.1429	0.0885
78	125	JAK2	9/4/1992	Victor A. McKusick	1	8	0.13	0.16		80	NKX3-1	Jennifer Macke	10/9/97		1	29	0.0345	0.0829
79	126	PIK3CA	10/15/1992	Victor A. McKusick	1	41	0.02	0.16		81	KLF6	Victor McKusick	10/15/97		1	6	0.1667	0.1162
80	127	MDM2	10/16/1992	Victor A. McKusick	1	1	1.00	0.17		83	TMPRSS2	Victor McKusick	10/16/97		2	1	2.0000	0.1193
81	128	EPHA2	10/23/1992	Victor A. McKusick	1	7	0.14	0.17		84	NRP2	Victor McKusick	10/21/97		1	5	0.2000	0.1198
82	129	ACLY	11/4/1992	Victor A. McKusick	1	12	0.08	0.18		85	CHD1	Victor McKusick	11/13/97		1	23	0.0435	0.1193
83	130	GRN	11/24/1992	Victor A. McKusick	1	20	0.05	0.18		86	PER1	Ada Hamosh	1/21/98		1	69	0.0145	0.1206
84	131	BRAF	12/1/1992	Victor A. McKusick	1	7	0.14	0.18		87	WWP1	Rebekah Rasooly	1/31/98		1	10	0.1000	0.1223
85	133	PTPN3	2/1/1993	Victor A. McKusick	2	62	0.03	0.18		88	TERC	Rebekah Rasooly	2/9/98		1	9	0.1111	0.1264
86	134	PAX2	2/25/1993	Victor A. McKusick	1	24	0.04	0.18		89	TGFB1I1	Rebekah Rasooly	2/13/98		1	4	0.2500	0.1262
87	135	LSP1	3/19/1993	Victor A. McKusick	1	22	0.05	0.18		90	DCX	Victor McKusick	3/24/98		1	39	0.0256	0.1268
88	137	PYCR1	5/14/1993	Victor A. McKusick	2	56	0.04	0.17		91	USP7	Patti Sherman	4/14/98		1	21	0.0476	0.1253
89	138	CTTN	5/27/1993	Victor A. McKusick	1	13	0.08	0.16		92	CTBP2	Rebekah Rasooly	5/13/98		1	29	0.0345	0.1274
90	140	TNFAIP3	6/23/1993	Victor A. McKusick	2	27	0.07	0.16		93	FHL2	Jennifer Macke	5/18/98		1	5	0.2000	0.1312
91	142	RHOG	6/24/1993	Victor A. McKusick	2	1	2.00	0.19		94	SPOP	Jennifer Macke	5/22/98		1	4	0.2500	0.1371
92	143	CTNNA1	9/2/1993	Victor A. McKusick	1	70	0.01	0.19		96	TUBB3	Rebekah Rasooly	5/27/98		2	5	0.4000	0.1390
93	145	CXCR4	9/16/1993	Victor A. McKusick	2	14	0.14	0.19		97	FOXO3A	Rebekah Rasooly	6/3/98		1	7	0.1429	0.1468
94	146	RAD51	9/28/1993	Victor A. McKusick	1	12	0.08	0.19		98	GLIPR1	Ethylin Wang Jabs	6/5/98		1	2	0.5000	0.1479
95	147	BMI1	11/1/1993	Victor A. McKusick	1	34	0.03	0.19		99	UTY	Victor McKusick	6/16/98		1	11	0.0909	0.1625
96	148	SOX4	11/5/1993	Victor A. McKusick	1	4	0.25	0.19		100	SCARA3	Victor McKusick	6/17/98		1	1	1.0000	0.1610

Figure 11 - Computation of Smooth Derivatives for Rate of Discovery

Question 6 Method

How many BC/PC genes remain to be discovered?

This question is a conundrum, since if we knew how many genes remained to be discovered we would have already discovered them! This would make the answer zero. So instead, we construct a related question that is amenable to actual analytic estimation. The approximating question is

How can we estimate how many BC/PC genes remain to be discovered?

This slight adjustment in wording makes answering the question more tractable. The results are shown in the results section below. It appears that the "lion's share" of the discoveries have been made.

Question 7 Method

What are the connections between genes?

This question was addressed using a datamining approach in combination with special purpose software and enabling assumptions. This software developed by the author, displayed the connections and relationships between genes of *BC/PC*.

The enabling assumption was ***'guilt by association***'. If the OMIM description of a gene (the 'record gene') included another gene (a 'cited gene') that was also in that specific cancer's gene list, then it was assumed that the two genes were in some way connected. This refined the working assumption to ***'guilt by citation'***. When a cited gene was found the question was asked, "Is this citation gene also in the list of genes associated by OMIM with this particular cancer query?" A number of genes were found that were not in the record gene's list, that could possibly be added in the query result gene list by discovery. For this survey, these extra genes were considered 'false positives' and culled, but a more time-consuming and thorough analysis could add these candidates to the list of cancer genes for the OMIM queried cancer. One shortcoming of the 'guilt by citation' is that it is not possible to discern whether the gene was associated because of its key functional relationship with the gene in question, or whether it was associated because it had been studied more than other genes and thus was more likely to be cited. A remedy for this would be the time-consuming and thorough enumeration of the specific functional relationship that the cited gene has with respect to the record gene. That analysis will have to wait for the sequel. This approach gets us started on a process that can be subsequently refined.

A Gene Connection 'Microscope'

The author has developed a general-purpose network display program over the past few years. It exploits the peculiar power of the human visual system to sort out motion in complex contexts. It has been in a continual state of prototyping, enduring various incarnations of the Java language in which it is written. It seemed useful for depicting the gene networks being explored here and assessing connections between genes. Since motion is involved, a short video with sound was made demonstrating its use and application to this survey.

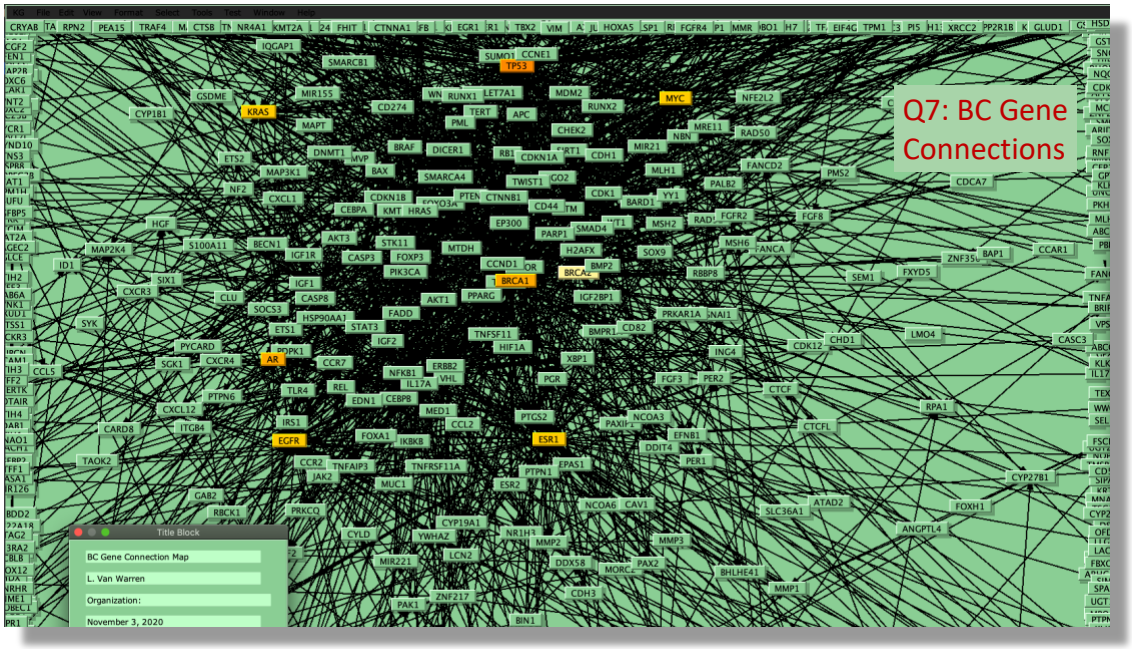


Figure 12 - kg: The "Knowledge Gazer" Graph Visualization Program

RESULTS

Question 1 Results

Who is affected by breast and prostate cancer (BC/PC)?

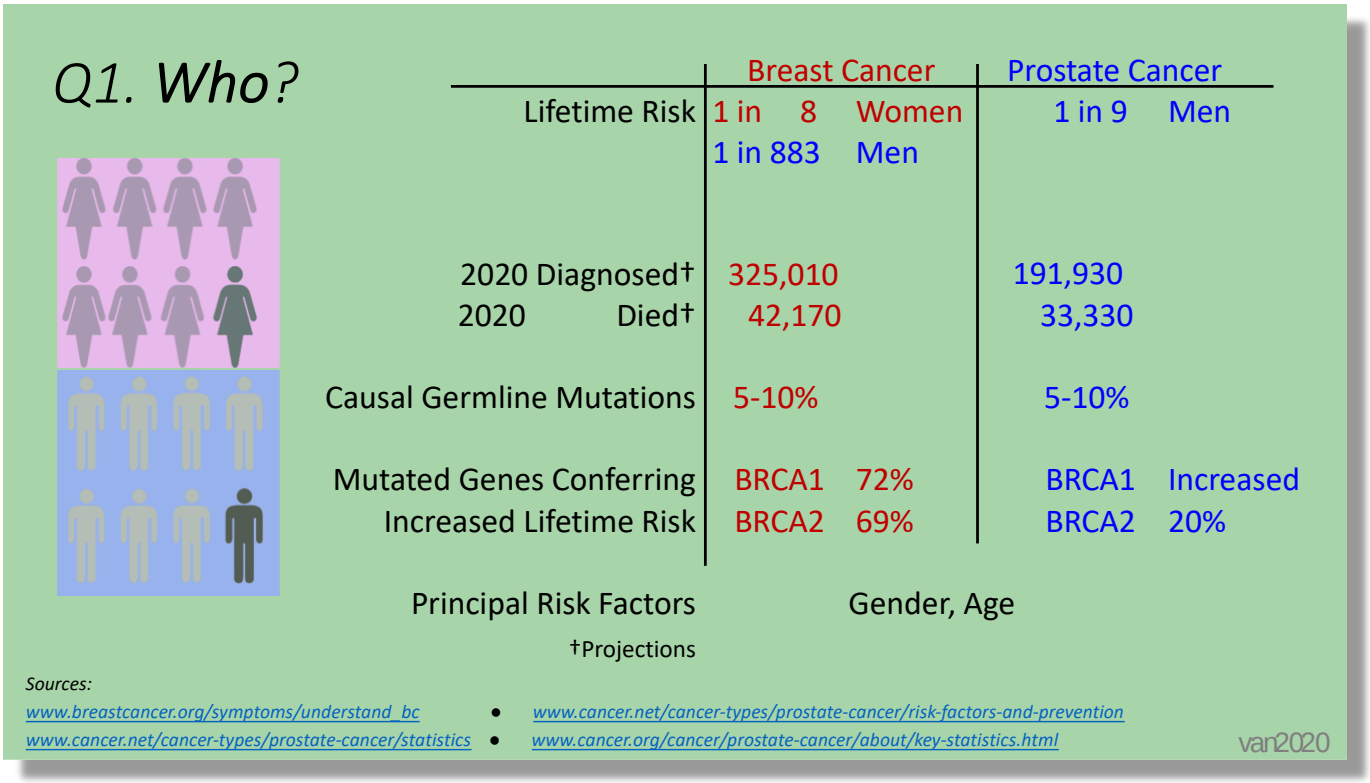


Figure 13 - Who is affected by breast and prostate cancer (BC/PC)?

From Figure 1 we see that women have a lifetime risk of 1 in 8 for breast cancer and men have a lifetime risk of 1 and 9 for prostate cancer. Men also have a breast cancer risk of 1 and 883. The projections for 2020 state that there will have been 325,010 women diagnosed with breast and 191,930 prostate cancer diagnosis in men, figures with comparable orders of magnitude. The referenced projections have further predicted that 42,170 females and 33,330 males will lose their lives to this disease. Germline mutations are believed to account for five to 10% of the risk of these two cancers. Interestingly, both BRCA1 and BRCA2 are associated with increased risk of both kinds of cancer. We shall also see that the phosphorylase PTEN is the most cited gene for both cancers. For BRCA1, the increase in lifetime risk is quantitatively assessed at 72%. with men it's qualitatively assessed as 'increased'. For BRCA2, the increase in lifetime risk is 69 and 20%, respectively. We also know the principal risk factors for these two cancers are gender and age, but young people get them as well.

Question 2 Results

How many BC/PC genes are known?

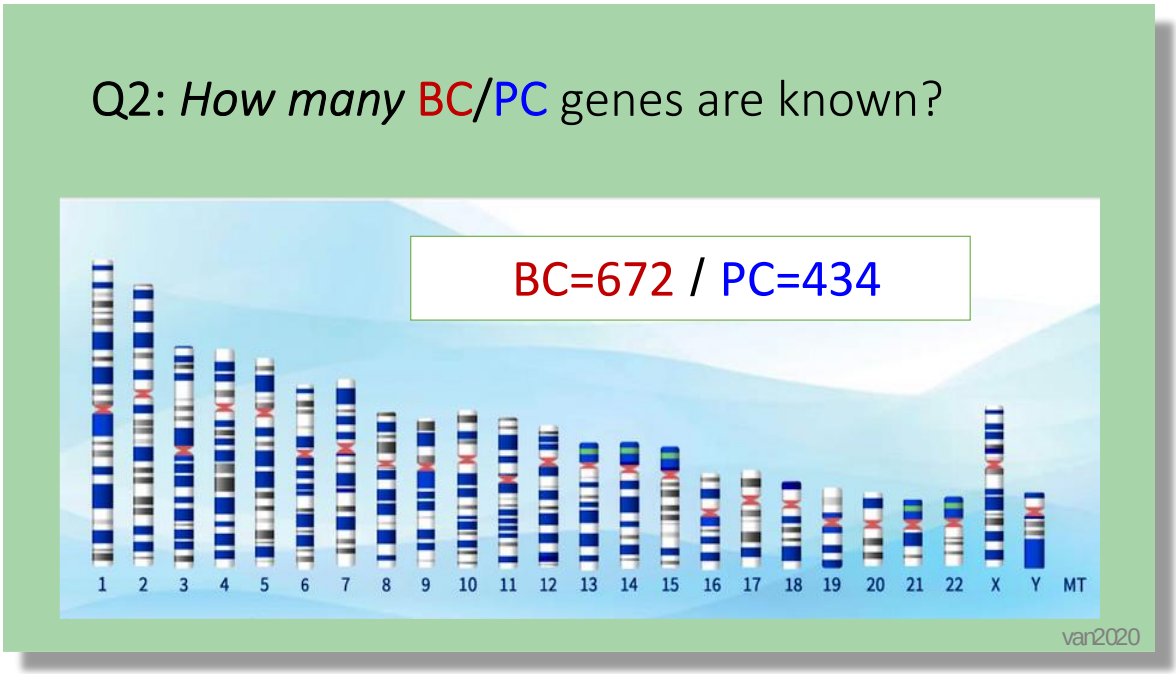


Figure 14 - The Answer to Question 2: How many BC/PC genes are known?

Question 3 Results

Where are the BC/PC genes in the genome?

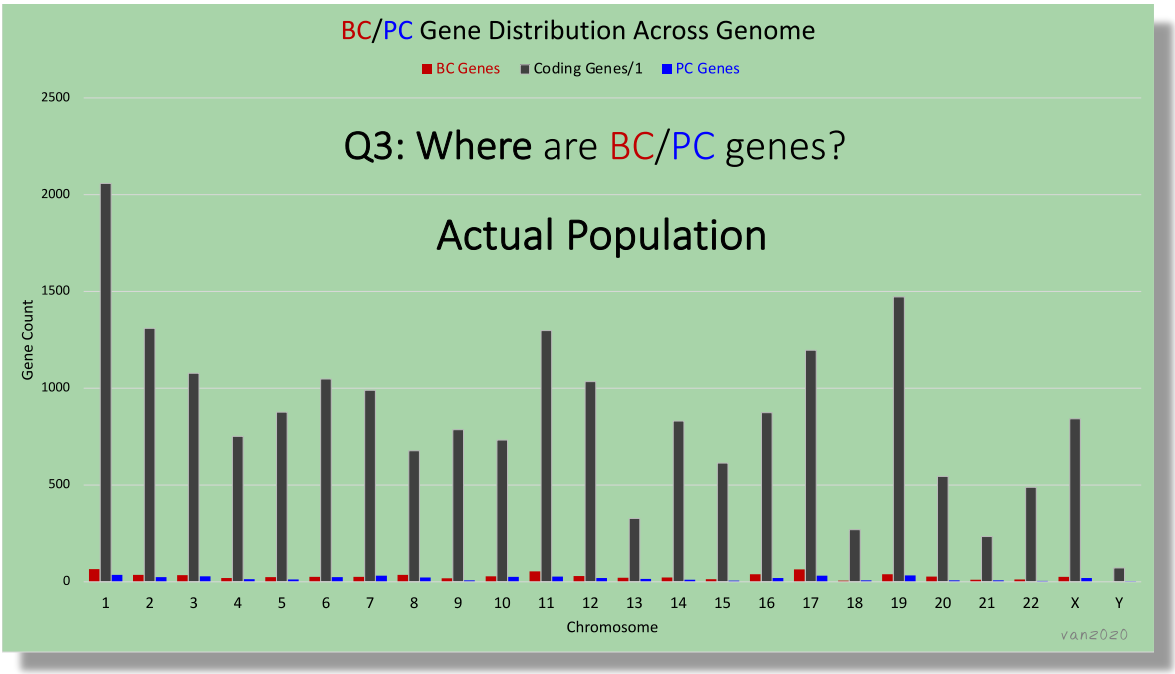


Figure 15 - Answer to Question 3: Where are the BC/PC genes in the genome?

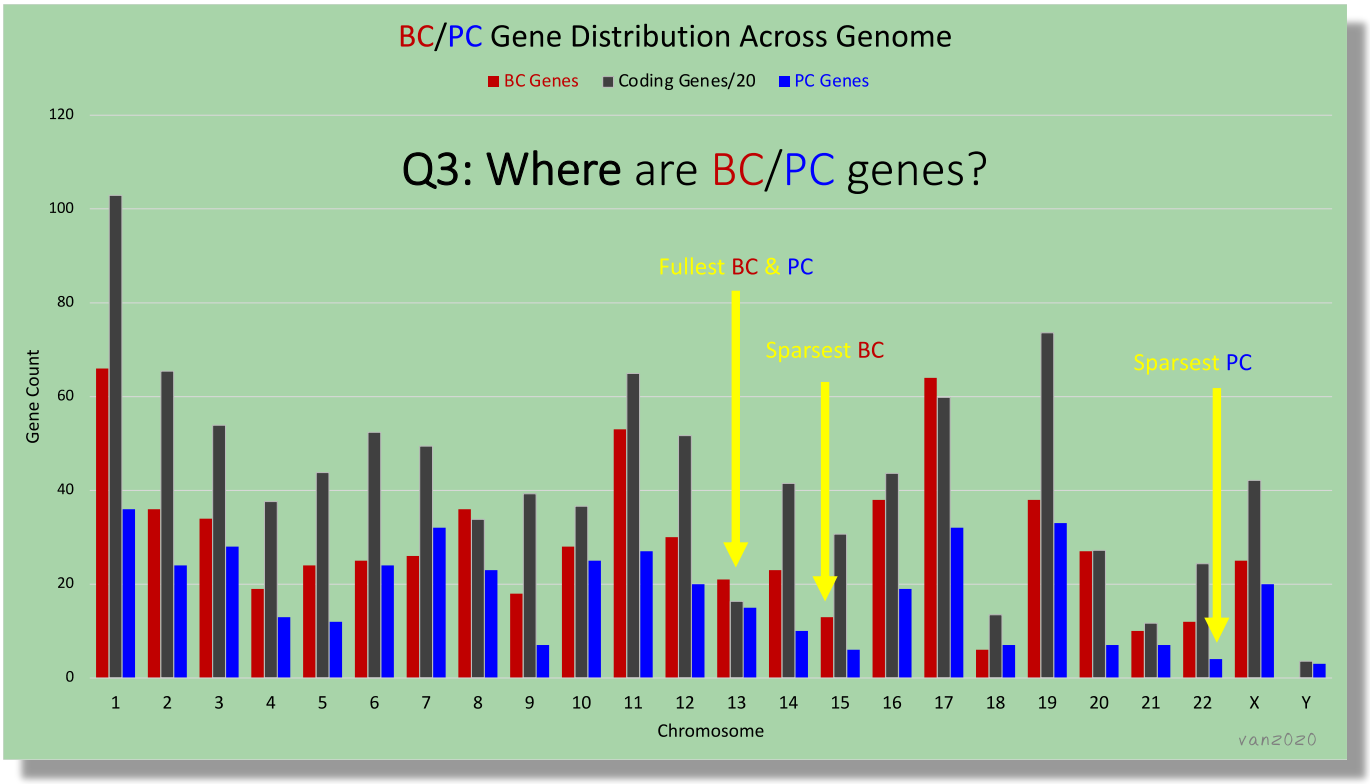


Figure 16 - Answer to Question 3: Zooming In By Scaling BC/PC Gene Counts

Q3: Where are BC/PC genes?

	Coding Genes	Scaled Coding Genes	Chr.	BC Genes	PC Genes	base pairs	BC Genes Per Coding Gene	PC Genes Coding Gene
	2058	103	1	**66**	**36**	248,956,422	3.2%	1.7%
	1309	65	2	36	24	242,193,529	2.8%	1.8%
	1078	54	3	34	28	198,295,559	3.2%	2.6%
	752	38	4	19	13	190,214,555	2.5%	1.7%
	876	44	5	24	12	181,538,259	2.7%	1.4%
	1048	52	6	25	24	170,805,979	2.4%	2.3%
	989	49	7	26	32	159,345,973	2.6%	3.2%
	677	34	8	36	23	145,138,636	5.3%	3.4%
	786	39	9	18	7	138,394,717	2.3%	0.9%
	733	37	10	28	25	133,797,422	3.8%	3.4%
	1298	65	11	53	27	135,086,622	4.1%	2.1%
	1034	52	12	30	20	133,275,309	2.9%	1.9%
	327	16	13	21	15	114,364,328	**6.4%**	**4.6%**
	830	42	14	23	10	107,043,718	2.8%	1.2%
	613	31	15	13	6	101,991,189	*2.1%*	1.0%
	873	44	16	38	19	90,338,345	4.4%	2.2%
	1197	60	17	64	32	83,257,441	5.3%	2.7%
	270	14	18	*6*	7	80,373,285	2.2%	2.6%
	1472	74	19	38	33	58,617,616	2.6%	2.2%
	544	27	20	27	7	64,444,167	5.0%	1.3%
	234	12	21	10	7	46,709,983	4.3%	3.0%
	488	24	22	12	*4*	50,818,468	2.5%	*0.8%*
	842	42	X	25	20	156,040,895	3.0%	2.4%
	71	4	Y		*3*	57,227,415		4.2%
sum	20,399	1,020		672	434	3,088,269,832		
min	71			*6*	*3*	46,709,983	2.1%	0.8%
max	2058			**66**	**36**	248,956,422	**6.4%**	**4.6%**
avg	850			29	18	128,677,910	*3.3%*	*2.1%*
SD	431			15	10	56,594,293	1%	1%

Scale Factor 20

Src: Human Genome Wiki

vanz020

Figure 17 - Computed Data for BC/PC Distribution by Chromosome

Q3: Where are BC/PC genes? Findings:

- Average distribution of BC/PC genes is 3.3% and 2.1% with σ of 1%.
- Chromosome 1: most BC/PC genes with 66 and 36 genes.
- Chromosome 13: most BC/PC genes per coding gene at 6.4% & 4.6%.
- Chromosome 15: least BC genes per coding gene at 2.1%
- Chromosome 22: least PC genes per coding gene at 0.8%
- There are no male BC genes on the Y chromosome.
- There are three PC genes on the Y chromosome.

vanz020

Figure 18 - Summary of Findings for Question 3: Where are BC/PC genes?

Question 4 Results

When were BC/PC genes discovered?

Q4:
When *were BC/PC genes discovered?*

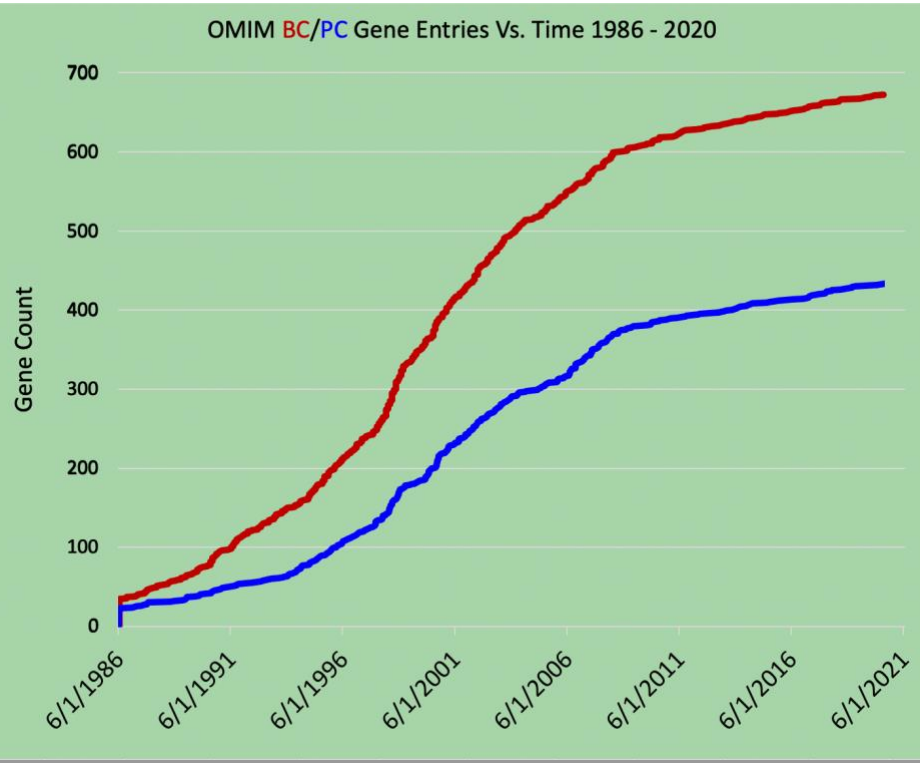


Figure 19 - Discovery Curves for BC/PC Genes

Question 5 Results

What is the rate of discovery of BC/PC genes?

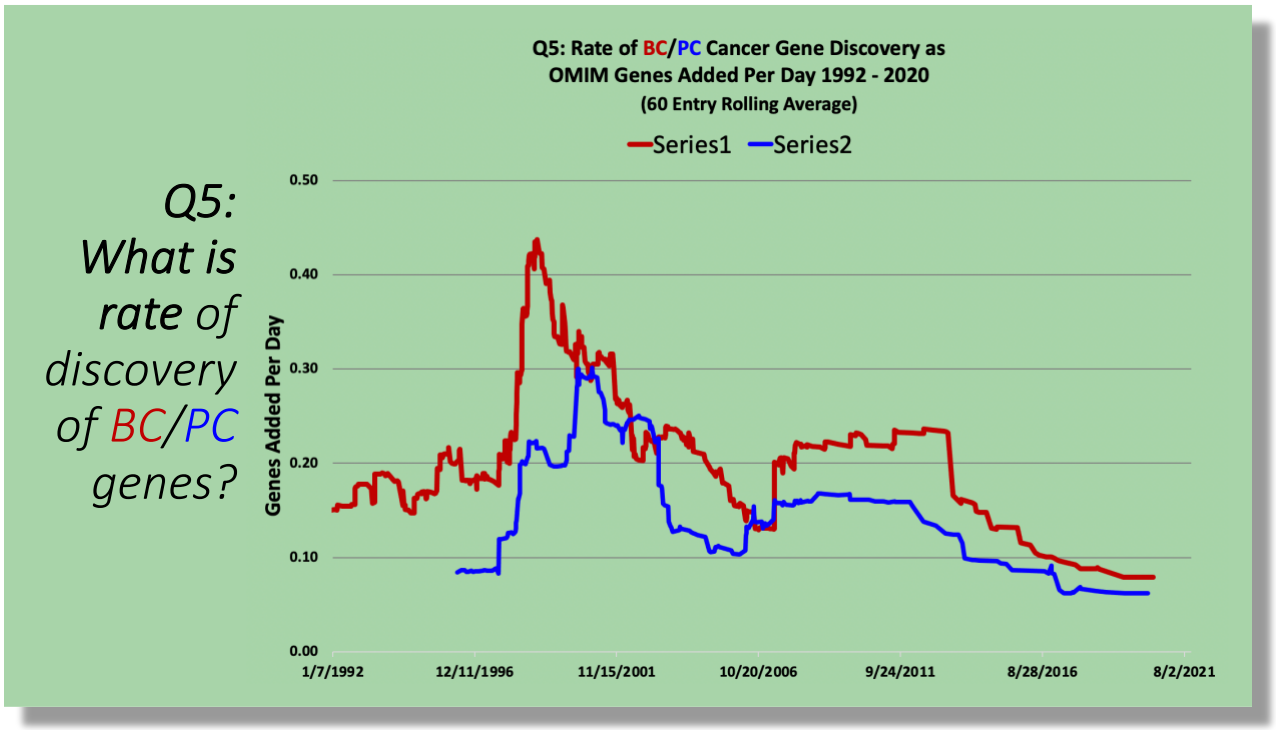


Figure 20 - Rate of Discovery Curves for BC/PC Genes

Question 6 Results

How many BC/PC genes remain to be discovered?

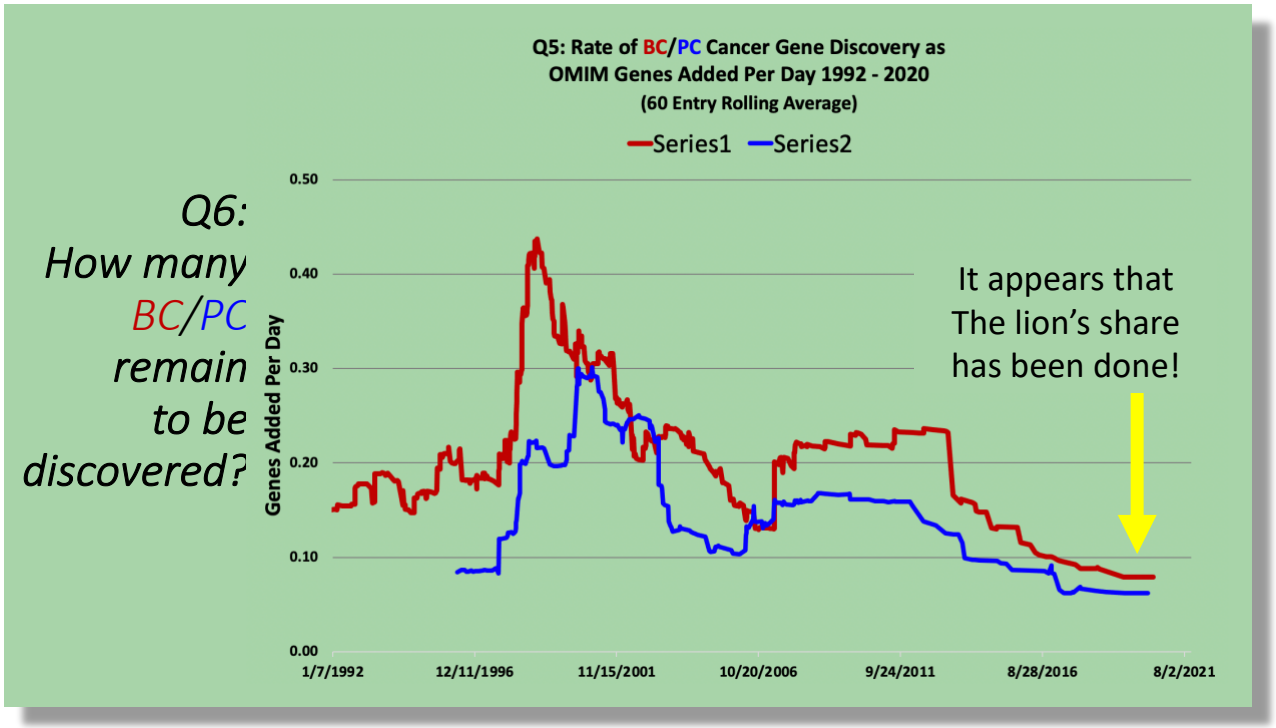


Figure 21 - The Lion's Share of Discoveries Appear Done

Question 7 Results

What are the connections between genes?

Q7: BC/PC Genes Connections? Findings:

- *ls* omimBC | wc –*l* → 672 BC Genes
- *ls* omimPC | wc –*l* → 434 PC Genes
- *ls* omimBCPC | wc –*l* → 939 BC/PC Genes

- Therefore BC and PC have 1106 – 939 = 167 genes in common
- Focusing on these shared genes could lead to single therapeutics that treat **both** cancers and are thus beneficial to both genders.

vanz020

Figure 22 - Total Genes and Genes in Common for Question 7

Q7: BC/PC Genes Connections? Discoveries:

- grep edge graphBC.txt | wc –*l* → 3404 BC Gene-Gene Relationships
- grep edge graphPC.txt | wc –*l* → 1427 PC Gene-Gene Relationships
- grep edge graphBC∪PC.txt | wc –*l* → 4931 BC ∪ PC Gene-Gene Relationships
- grep edge graphBC∩PC.txt | wc –*l* → 563 BC ∩ PC Gene-Gene Relationships

- We must explore these relationships!
- This requires a new kind of instrument to advance our understanding that exploits evolutionary strengths of human pattern recognition:
- A Gene Connection "Microscope", a *GenConScope*, if you will.

vanz020

Figure 23 - Edge Counts for Genes in BC/PC Union and Intersection

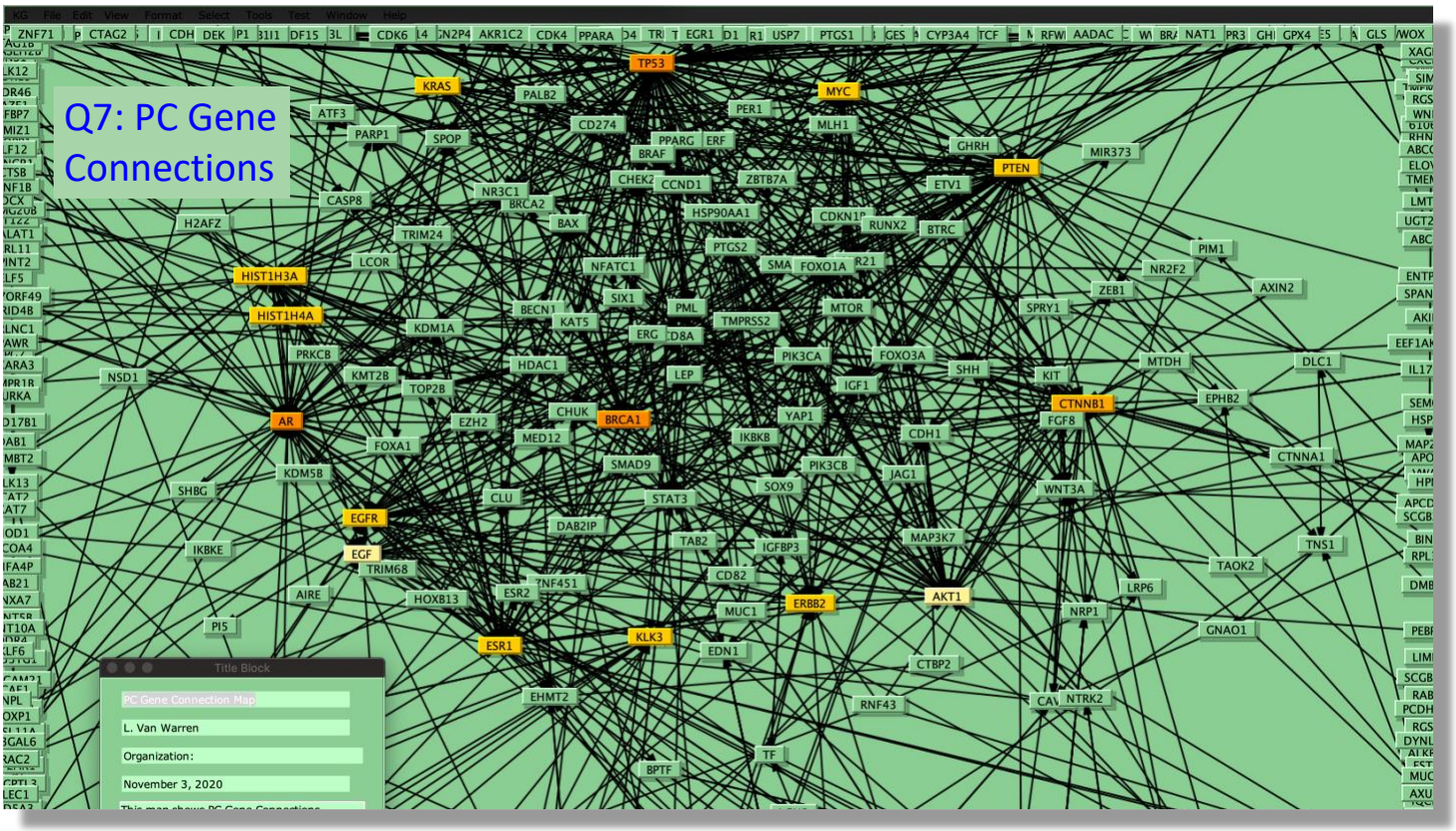


Figure 24 - A Snapshot of Prostate Cancer Gene Connections

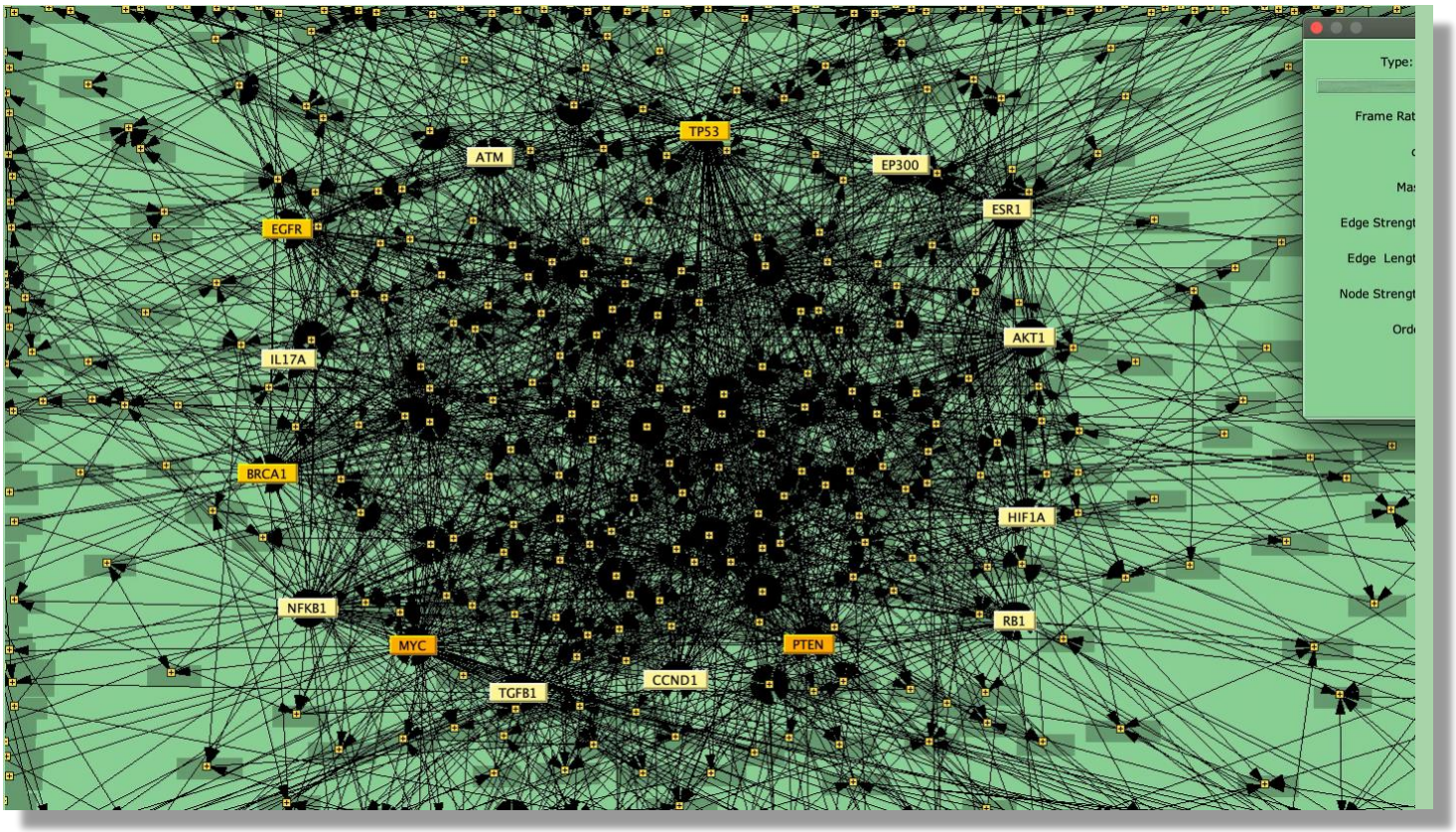


Figure 25 - A Ring of Fifteen Cancer Genes

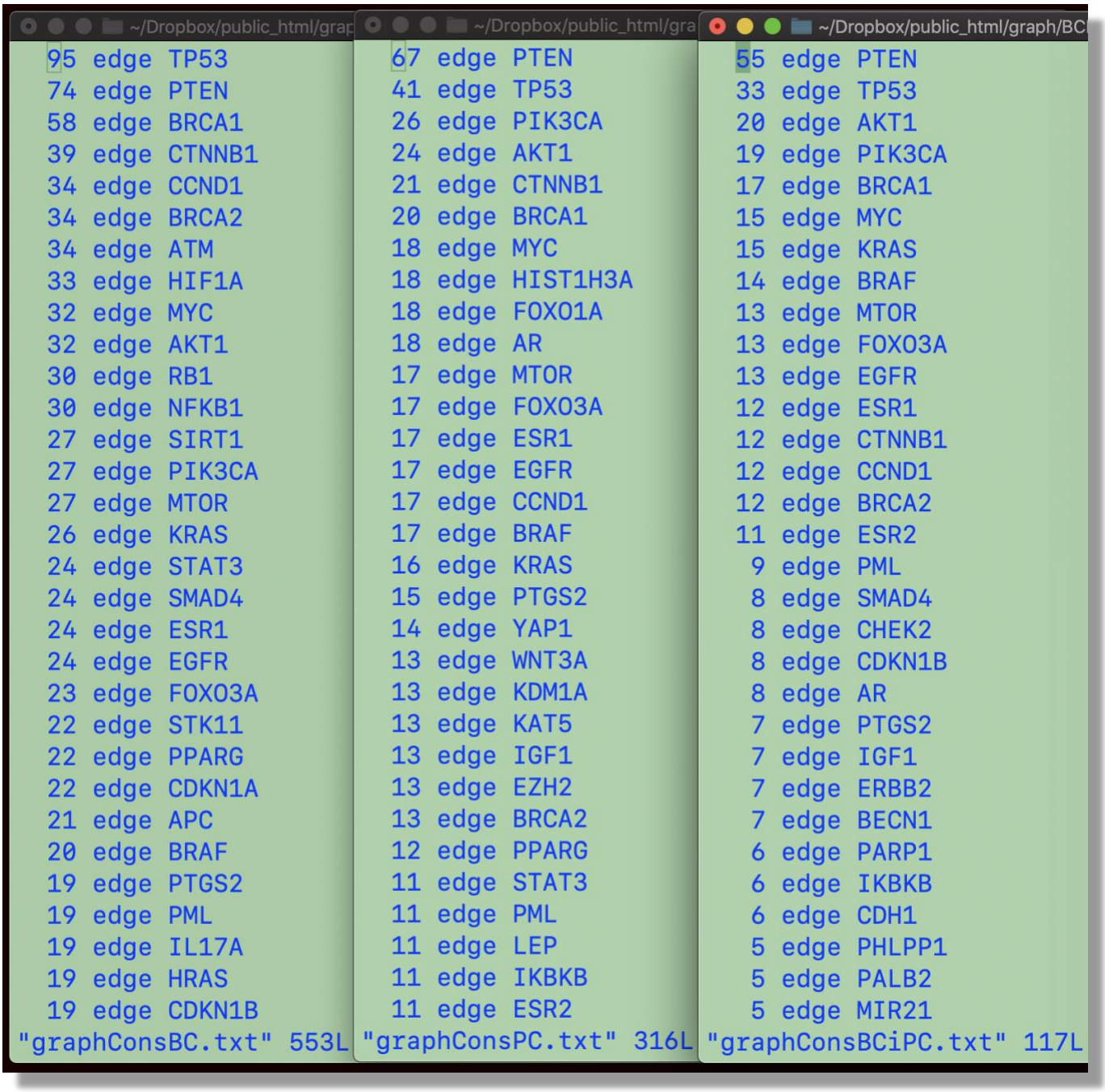


Figure 26 - Gene Connection Tallies for BC/PC and Intersection

Other Results – Prostate Cancer Not as Well Researched as BC

One finding of this report is that prostate cancer is not as researched as breast cancer. There is no Susan Komen Foundation or Race for the Cure for prostate cancer. Consider two leading search engines.

With Bing we get 38 million hits for breast cancer and 11 million hits for prostate cancer. If we look at Google Trends, we can see the yearly increase in interest - probably as a result of the Susan Komen activities, while prostate remains at a lower baseline of interest about half that of breast cancer.

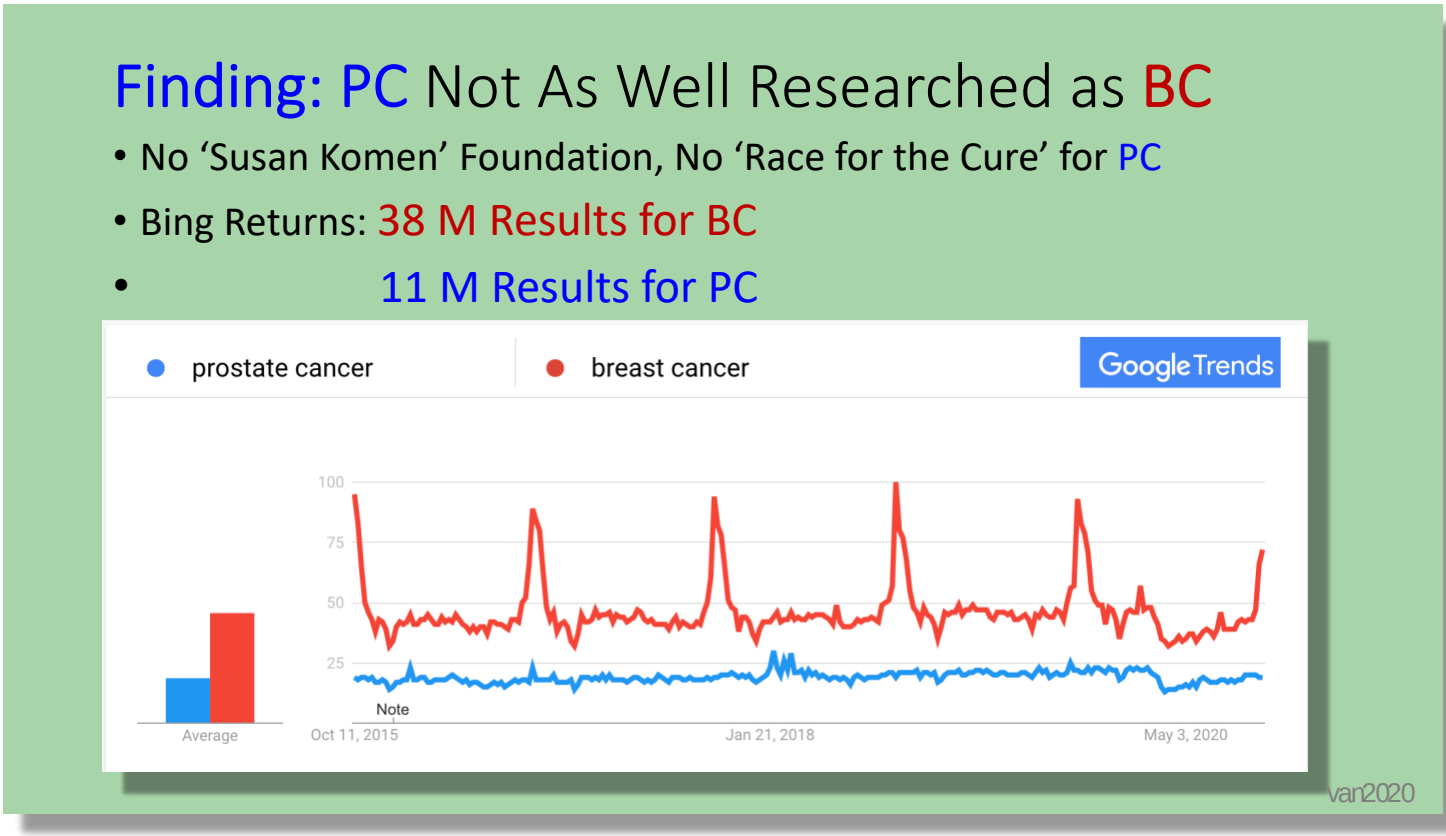


Figure 27 - Finding: Prostate Cancer Not as Researched as Breast Cancer

DISCUSSION / CONCLUSIONS

Much of the foundational work for *BC/PC* discovery has been done. But the understanding of the relationships between *BC/PC* genes is just beginning. Extracting putative relationships using tools of automation and visualization is a focus here, but further work remains in categorization. The visualization of the startling complexity of these relationships is a useful tool for conducting this research. Connection visualization also mitigates against the tendency to over focus on single factor causes and solutions. This approach and software tooling opens a door for greater understanding facilitates making new discoveries which will lead to customized and more effective treatments for these perennial afflictions. The open question of this survey is whether 'guilt by citation' is an adequate screen to distinguish genes that are heavily studied from genes with putative connections. A sequel with non-cancer conditions may answer this. In performing this survey, the author experienced the wonderful sensation of discovering the answers to questions whose answers he did not know beforehand and that was the most fun of all.

ACKNOWLEDGEMENTS

The author would like to thank Dr. Mary Yang for teaching an excellent course in bioinformatics.

The author must also thank each of these people:

- Dr. Elizabeth Pierce — Chair of Department of Information Science
- Dr. Dan Berleant — Professor in Information Science
- Dr. Phil Williams — Professor Bioinformatics and Cloud Computing
- Dr. Victor McKusick — Founder of OMIM
- Joanna Amberger — Scientist at COSMIC

and particularly my spouse Lynn Warren, without whose support nothing would have ever gotten done.

The author would also like to thank my friend Dr. Dan Berleant, who quoting Hillel said,

If not now, when?

The author deeply appreciates the work of the many teachers, scientists, and engineers at the following institutions whose efforts made this work possible.

- UALR Department of Information Science
- OMIM Online Mendelian Inheritance in Man
- COSMIC Catalogue of Somatic Mutations in Cancer
- PDB Protein Data Bank
- R Consortium
- Google
- Bing
- Apache Netbeans
- Otter.ai Voice to Text Translation Services

REFERENCES

Collins FS, Fink L. The Human Genome Project. Alcohol Health Res World. 1995;19(3):190-195.

Online Mendelian Inheritance in Man, OMIM®. McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD), World Wide Web URL: https://omim.org/

"Key Statistics for Prostate Cancer: Prostate Cancer Facts." American Cancer Society, www.cancer.org/cancer/prostate-cancer/about/key-statistics.html.

"Prostate Cancer - Risk Factors and Prevention." Cancer.Net, 11 Aug. 2020, www.cancer.net/cancer-types/prostate-cancer/risk-factors-and-prevention.

"Prostate Cancer - Statistics." Cancer.Net, 28 Feb. 2020, www.cancer.net/cancer-types/prostate-cancer/statistics.
"Understanding Breast Cancer." Breastcancer.org, 25 June 2020, www.breastcancer.org/symptoms/understand_bc.

Mukherjee, Siddhartha. The Emperor of All Maladies: a Biography of Cancer. Gale, Cengage Learning, 2012.
Mukherjee, Siddhartha. The Gene: an Intimate History. Bodley Head, 2016.

Forbes S.A., Beare D., Boutselakis H., Bamford S., Bindal N., Tate J., Cole C.G., Ward S., Dawson E., Ponting L.et al. COSMIC: somatic cancer genetics at high-resolution. Nucleic Acids Res. 2017; 45:D777–D783.

H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne.
(2000) The Protein Data Bank Nucleic Acids Research, 28: 235-242.

PubMed; National Center for Biotechnology Information (NCBI)[Internet]. Bethesda (MD): National Library of Medicine (US); https://www.ncbi.nlm.nih.gov/

MeSH Browser;[Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; 2004 – [cited 2020 Dec 04]. Available from: https:// https://meshb.nlm.nih.gov/