BRCA1, BRCA2與PALB2在台灣乳癌病患的流行病學分析與次世代定序生資處理實務

Next-Generation Sequencing Data Processing: Practices and Epidemiological Analysis of BRCA1, BRCA2, and PALB2 in Taiwanese Breast Cancer Patients

實習單位：台北榮總乳房醫學中心及實驗外科

實習學生：龍紀妘、許容欣　指導老師：盧子彬老師　單位指導：黃其晟醫師

I. Introduction

Cancer is the top 1 of the top 10 causes of death, while breast cancer has the highest incidence of all cancers in women. In order to increase the survival rate of breast cancer, Taiwan actively promotes breast cancer screening for early detection and early treatment. The mutations in BRCA1, BRCA2, PALB2 are risk factors for breast cancer development, and these 3 genes greatly increase the risk of breast cancer.

Research shows that patients with BRCA1 or BRCA2 gene mutations have an approximately 80% lifetime chance of developing breast cancer, and for those with PALB2 gene mutation have an approximately 35% lifetime chance of suffering from breast cancer.

II. Motivation & Purpose

Annotation refers to the process of adding detailed information and context to genetic data. The information is critical to clinical diagnosis while it provides insights into whether a genetic variant is pathogenic or benign, supporting assessment of a patient's genetic predisposition and leading to a more effective treatment.

In this project, we aim to add descriptive markers to genetic data, thereby enriching the dataset's value and facilitating advanced analyses. Furthermore, by investigating and analyzing the insights contained within this genomic data through descriptive statistics, we aspire to uncover some hidden patterns that may have previously gone unnoticed.

III. Materials & Methods

1. Data Resource: Whole Exome Sequencing (WES) data from Taipei Veterans General Hospital Comprehensive Breast Health Center.
2. Methods:
   - **a. Data Filtering:** Filter through variables to find the matching data by following steps
     - Keep "exonic" genome sequence
     - Keep gene"BRCA1", " BRCA2" and "PALB2"
     - Delete variants which clinical significance referred to as "Benign"
   - **b. Data Pairing:** Pair somatic and germline data by patients' ID.
   - **c. Data Annotation:** Find out variants' clinical significance from public domain database.
     - Using Excel data parsing to process Column [AAChange.refGene]
     - Making a new variable "protein changes"
     - Using "protein changes" to do data annotation
     - Using R Studio to compare data with those from ClinVar (data from NCBI)
     - Searching one by one from OncoKB
   - **d. Data Analysis:** Separately analyze the somatic- and germline- patients.

Table1 The total number of two group before/after filtering

Group	Original Data	Filtered Data
Germline Patients	83	18
Germline Variants	11089	27
Somatic Patients	14	3
Somatic Variants	18	6

IV. Results

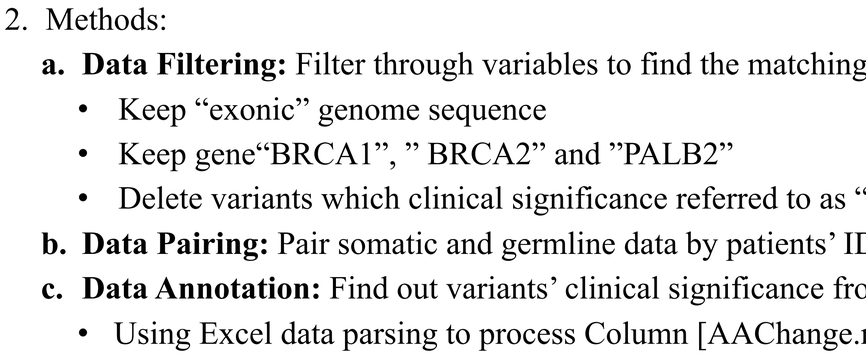

Figure1
BRCA1 Lollipop Chart

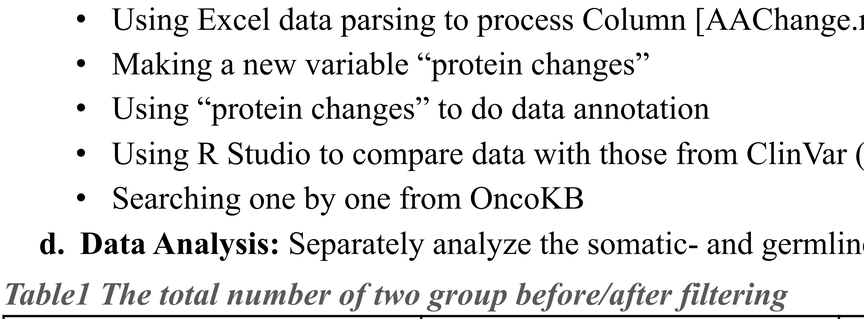

Figure2
BRCA2 Lollipop Chart

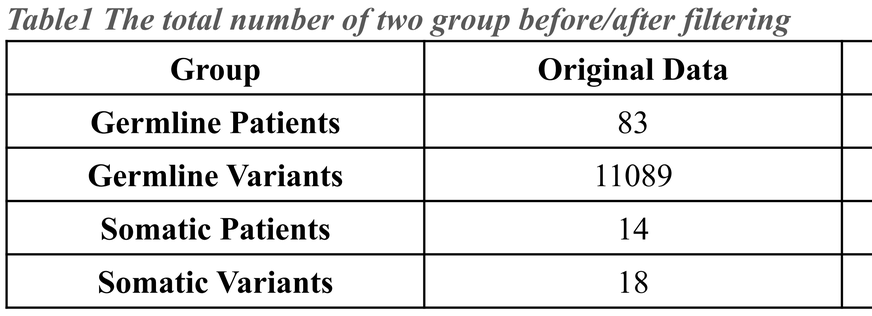

Figure3
PALB2 Lollipop Chart

Table2 Annotation results of somatic variants

Sample	Gene	Protein Change	Variant	OncoKB	ClinVar
#1	PALB2	p.N280Kfs*3	frameshift insertion	Likely Oncogenic	NA
#2	BRCA1	p.K607Sfs*47	frameshift deletion	Likely Oncogenic	NA
#2	BRCA1	p.K654Sfs*47	frameshift deletion	Likely Oncogenic	NA
#2	BRCA1	p.K654Sfs*47	frameshift deletion	Likely Oncogenic	NA
#3	BRCA2	p.Q1562E	nonsynonymous SNV	NA	Uncertain significance
#3	BRCA2	p.Q2384E	nonsynonymous SNV	NA	Uncertain significance

Table3 Annotation results of germline variants

Sample	Gene	Protein Change	Variant	OncoKB	ClinVar
#5	BRCA2	p.G2508S	nonsynonymous SNV	NA	Conflicting interpretations of pathogenicity*
#7	BRCA2	p.S1074C	nonsynonymous SNV	NA	Conflicting interpretations of pathogenicity
#7	BRCA2	p.S1946P	nonsynonymous SNV	NA	Conflicting interpretations of pathogenicity
#8	PALB2	p.T787I	nonsynonymous SNV	NA	Conflicting interpretations of pathogenicity
#9	BRCA2	p.S205Ffs*9	frameshift deletion	Likely Oncogenic	NA
#11	PALB2	p.E352Q	nonsynonymous SNV	NA	Conflicting interpretations of pathogenicity
#12	PALB2	p.P405A	nonsynonymous SNV	NA	Conflicting interpretations of pathogenicity
#13	PALB2	p.L531Ffs*30	frameshift deletion	Likely Oncogenic	NA
#14	BRCA2	p.K1533N	nonsynonymous SNV	NA	Conflicting interpretations of pathogenicity
#14	PALB2	p.R825T	nonsynonymous SNV	NA	Conflicting interpretations of pathogenicity
#15	BRCA2	p.K1533N	nonsynonymous SNV	NA	Conflicting interpretations of pathogenicity
#16	BRCA1	p.I736V	nonsynonymous SNV	NA	Conflicting interpretations of pathogenicity
#16	BRCA1	p.I783V	nonsynonymous SNV	NA	Conflicting interpretations of pathogenicity
#17	BRCA1	p.I783V	nonsynonymous SNV	NA	Conflicting interpretations of pathogenicity
#17	PALB2	p.D498Y	nonsynonymous SNV	NA	Conflicting interpretations of pathogenicity
#18	BRCA2	p.K2496N	nonsynonymous SNV	NA	Uncertain significance
#19	BRCA2	p.S1946P	nonsynonymous SNV	NA	Conflicting interpretations of pathogenicity
#20	PALB2	p.S454P	nonsynonymous SNV	NA	Uncertain significance
#21	PALB2	p.R825T	nonsynonymous SNV	NA	Conflicting interpretations of pathogenicity

V. Conclusion

1. It has come to our attention that "frameshift deletions" and "frameshift insertions", two particular types of genetic alterations, are often categorized as "likely oncogenic" by oncoKB. Moreover, translation is terminated at a downstream stop codon in these cases.
2. Even if a large amount of data after being filtered out many variables and being annotated with public domain database, there are very few cases remained undermined regarding clinical significance. This situation leads to the amount of information that can be provided is limited, and showing the importance of big data analysis for genetic data.
3. We've discovered a phenomenon where certain individuals exhibit simultaneous presence of multiple genetic mutations, referred to as compound mutation.

VI. Discussion

1. The interactions of compound mutations holds the potential to unlock insights into specific biological processes or disease pathways.
2. The versions of human genome which are used by the websites aren't unified, some of them use hg38, while others use hg19. This might increase the difficulty of information searching.

VII. Reference

1. NCBI – ClinVar https://www.ncbi.nlm.nih.gov/clinvar/
2. OncoKB™ - MSK's Precision Oncology Knowledge Base https://www.oncokb.org/