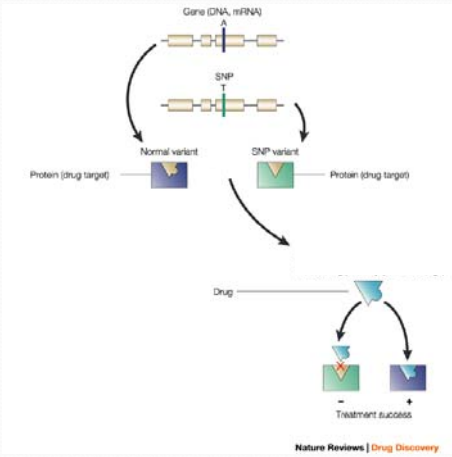
SNP Data Integration and Analysis for Drug-Response Biomarker Discovery

By
Chen Jieqi Pauline

Contents

Introduction

- Single Nucleotide Polymorphisms
  - Point mutations occurring >1% in general population
  - Biallelic, triallelic SNPs
  - Alter enzyme properties and drug response



Introduction

- Pharmacogenomics
  - Genetic factors contributing to variation in drug response
  - "Personalized drugs"
- Problem Statement
  - Methods for SNP biomarker analysis require genotyping
  - Incompleteness of drug-enzyme association databases

Motivation

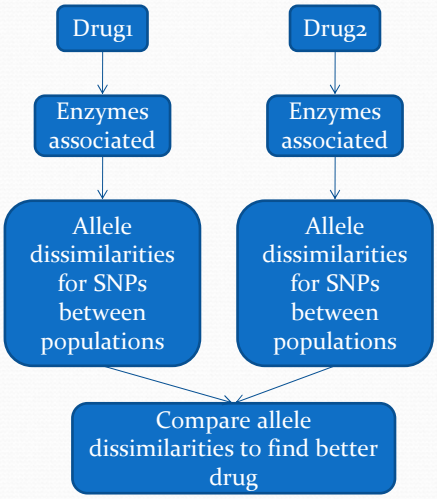
- Discovery of SNPs as drug-response biomarkers
  - Response to drugs and drug marketability
    - Ethnicity differences
    - Individual differences
- Incompleteness of drug-enzyme database
  - Existing databases (e.g. PharmGKB) do not capture information for all drugs
  - Important for determining regions of genome to study

System Overview

- Browser-based tool
- PHP, MySQL
- Main input sources: PharmGKB, International HapMap, DrugBank
- Components
  - Drug-response SNP biomarker discovery
    - SNPs with allelic dissimilarities that affect marketability
    - Single SNP biomarkers
    - Multiple SNP biomarkers
  - Drug-enzyme association discovery

SNP drug-response biomarker discovery - I

- SNPs with allelic dissimilarity
  - Aim: Find the better between 2 drugs to market
  - Motivation: Drug response differs with allelic dissimilarities between ethnicities
  - Methodology: F_{ST} index



SNP drug-response biomarker discovery - I

- About F_{ST} index
  - Measure of population divergence based on polymorphism data
  - 4 categories of values:
    - < 0.05 = little genetic differentiation
    - 0.05 – 0.15 = moderate genetic differentiation
    - 0.15 – 0.25 = great genetic differentiation
    - > 0.25 = very great genetic differentiation

$$H_{exp,\,i} = 1 - (F_{A,i}^{\,2} + F_{B,i}^{\,2})$$

$$H_S = (\Sigma_i\, H_{exp,i} \times n_i) / \Sigma_i\, n_i$$

$$H_T = 1 - (\bar{p}^2 + \bar{q}^2)$$

$$F_{ST} = (H_T - H_S) / H_T$$

SNP drug-response biomarker discovery - I

Search for a drug

Search: doxorubicin

Search!

Go back to main

Search for a drug

Search: exemestane

Search!

Go back to main

SNP drug-response biomarker discovery - I

Results for Search

Your search for 'doxorubicin' has returned 3 result(s).

To view all genes associated with the selected drug and do statistical analysis, choose a drug and click 'Select this drug'.

	PharmGKB ID	Name of Drug	Alternate Names for Drug
○	PA449412	doxorubicin	"Adriamycin PFS","Adriamycin RDF","Rubex",
○	PA449413	doxorubicin hydrochloride	"Adriamycin Pfs","Adriamycin Rdf","Caelyx","Doxil","Doxorubicin Hcl","Rubex",
○	PA449414	doxorubicin hydrochloride liposome	

Select this drug

Results for Search

Your search for 'exemestane' has returned 1 result(s).

To view all genes associated with the selected drug and do statistical analysis, choose a drug and click 'Select this drug'.

	PharmGKB ID	Name of Drug	Alternate Names for Drug
○	PA449563	exemestane	"Aromasin",

Select this drug

SNP drug-response biomarker discovery - I

MSH2	HERE	mutS homolog 2, colon cancer, nonpolyposis type 1 (E. coli)	mutS (E. coli) homolog 2 , mutS (E. coli) homolog 2 (colon cancer, nonpolyposis type 1) ,
TOP2A	HERE	topoisomerase (DNA) II alpha 170kDa	DNA topoisomerase II, 170 kD , topoisomerase (DNA) II alpha (170kD) ,
CBR1	HERE	carbonyl reductase 1	carbonyl reductase (NADPH) , carbonyl reductase (NADPH) 1 ,

Provide statistics for comparison | Frequency Deviations

SULT2A1	HERE	sulfotransferase family, cytosolic, 2A, dehydroepiandrosterone (DHEA)-preferring, member 1	hydroxysteroid sulfotransferase , sulfotransferase family 2A, dehydroepiandrosterone (DHEA) -preferring, member 1 , sulfotransferase family 2A, dehydroepiandrosterone (DHEA)-preferring, member 1 , sulfotransferase family, cytosolic, 2A, dehydroepiandrosterone (DHEA) -preferring, member 1 ,

Provide statistics for comparison | Frequency Deviations

SNP drug-response biomarker discovery - I

Please select the range of FST values for viewing:

Choose a Range... Go

Current view range: Greater than 0.25

RSID	Gene	Allele 1	Allele 2	Fst	Race	A	T	C	G
rs2242480	CYP3A4	C	T	0.2765	African	0	0.679	0.321	0
					Gujarati	0	0.31	0.69	0
					Han(Beijing)	0	0.262	0.738	0
					Japanese	0	0.297	0.703	0
					Luhya	0	0.872	0.128	0
					Maasai	0	0.56	0.44	0
					Mexican	0	0.418	0.582	0
					Utah	0	0.073	0.927	0
					Yoruban	0	0.857	0.143	0
rs3735451	CYP3A4	T	C	0.2562	African	0	0.377	0.623	0
					Chinese(Denver)	0	0.724	0.276	0
					Gujarati	0	0.705	0.295	0
					Han(Beijing)	0	0.661	0.339	0
					Japanese	0	0.698	0.302	0
					Luhya	0	0.156	0.844	0
					Maasai	0	0.462	0.538	0

SNP drug-response biomarker discovery - I

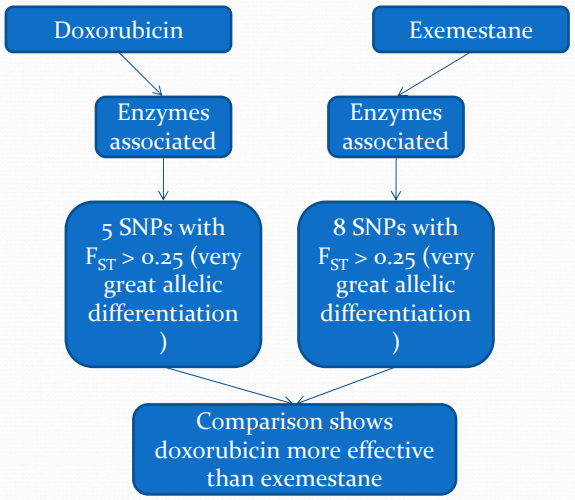
Please select the range of FST values for viewing:

Choose a Range... Go

Current view range: Greater than 0.25

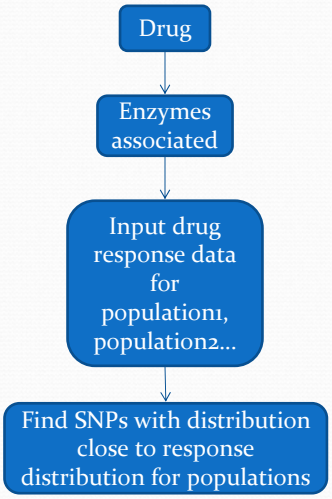
RSID	Gene	Allele 1	Allele 2	Fst	Race	A	T	C	G
rs1056836	CYP1B1	C	G	0.3138	African	0	0	0.74	0.26
					Chinese(Denver)	0	0	0.171	0.829
					Gujarati	0	0	0.176	0.824
					Han(Beijing)	0	0	0.083	0.917
					Japanese	0	0	0.095	0.905
					Luhya	0	0	0.767	0.233
					Maasai	0	0	0.671	0.329
					Mexican	0	0	0.31	0.69
					Toscans	0	0	0.403	0.597
					Utah	0	0	0.447	0.553
					Yoruban	0	0	0.876	0.124
rs1056837	CYP1B1	A	G	0.2848	African	0.736	0	0	0.264
					Chinese(Denver)	0.171	0	0	0.829
					Gujarati	0.176	0	0	0.824
					Han(Beijing)	0.089	0	0	0.911
					Japanese	0.11	0	0	0.89
					Luhya	0.728	0	0	0.272
					Maasai	0.657	0	0	0.343

SNP drug-response biomarker discovery - I



SNP drug-response biomarker discovery - II

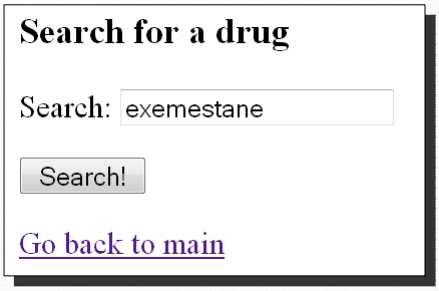
- SNPs affecting drug response
  - Aim: Find SNPs as biomarkers for drug response
  - Motivation: SNPs affect drug response
  - Methodology: Chi-square goodness-of-fit test



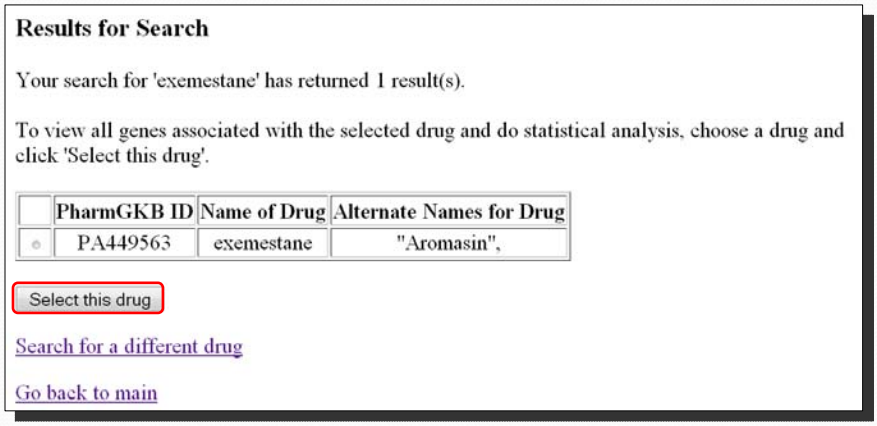
SNP drug-response biomarker discovery - II

- More about chi-square goodness-of-fit
  - Correlation between distribution of allele and drug response
    - 10% allele A, 90% allele a
      => 10% non-responsive, 90% responsive OR
      => 10% responsive, 90% non-responsive
  - Chi-square test at 90% significance level
    - Apply in individual populations
    - Apply over all populations

SNP drug-response biomarker discovery - II



SNP drug-response biomarker discovery - II



SNP drug-response biomarker discovery - II

SULT1A3	HERE	sulfotransferase family, cytosolic, 1A, phenol-preferring, member 3	preferring sulfotransferase , monoamine-sulfating phenosulfotransferase , placental estrogen sulfotransferase , sulfotransferase family 1A, phenol-preferring, member 3 , thermolabile (monoamine, M form) phenol sulfotransferase , thermolabile phenol sulfotransferase ,
SULT2A1	HERE	sulfotransferase family, cytosolic, 2A, dehydroepiandrosterone (DHEA)-preferring, member 1	DHEA-ST , alcohol sulfotransferase , alcohol/hydroxysteroid sulfotransferase , dehydroepiandrosterone sulfotransferase , hydroxysteroid sulfotransferase , sulfotransferase family 2A, dehydroepiandrosterone (DHEA) -preferring, member 1 , sulfotransferase family 2A, dehydroepiandrosterone (DHEA)-preferring, member 1 , sulfotransferase family, cytosolic, 2A, dehydroepiandrosterone (DHEA) -preferring, member 1 ,

Provide statistics for comparison | Frequency Deviations

SNP drug-response biomarker discovery - II

Input statistics directly for determining significant SNPs in affecting drug response.
Only input the values for the races you have information for. For the other races, please leave them blank.

	React to drug	Does not react to drug
African		
Chinese(Denver)		
Gujarati		
Han(Beijing)		
Japanese		
Luhya		
Maasai		
Mexican		
Toscans		
Utah		
Yoruban		

Submit

SNP drug-response biomarker discovery - II

Significant SNPs

Please refer to this link for more information on the calculations.

Enzymes related to drug to be used for calculation: CYP1A1 CYP1A2 CYP1B1 CYP3A CYP19A1 HSD17B1 STS SULT1E1 SULT1A1 SULT1A3 SULT2A1

Response and Non Response input values:

	React to drug	Does not react to drug
African	964	36
Luhya	900	100
Maasai	970	30
Yoruban	960	40

SNP drug-response biomarker discovery - II

Probability threshold:

You are currently using a threshold probability of 0.1. This probability corresponds to having at most the calculated chi-value assuming that the observed values correspond to the expected values which have been calculated using the MAF of each SNP. An SNP will be considered significant if it has a probability smaller than the threshold.

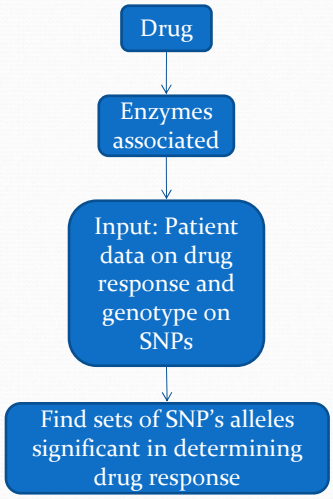
Use this threshold!

The following are the results for individual races:

RSID	Gene	Allele 1	Allele 2	Probability	Chi-square Value	Race	A	T	C	G
rs4545755	CYP19A1	G	A	0.0000	0.0000	Yoruban	0.04	0	0	0.96
rs4775933	CYP19A1	T	C	0.0000	0.0000	Yoruban	0	0.04	0.96	0
rs6493487	CYP19A1	G	A	0.0000	0.0000	Yoruban	0.96	0	0	0.04
rs6493493	CYP19A1	A	G	0.0000	0.0000	Yoruban	0.04	0	0	0.96
rs4986879	CYP1A1	T	C	0.0000	0.0000	Yoruban	0	0.96	0.04	0
rs7891417	STS	G	A	0.0836	0.0110	Luhya	0.101	0	0	0.899
rs1220716	SULT1E1	T	C	0.0836	0.0110	Luhya	0	0.101	0.899	0
rs2547231	SULT2A1	C	A	0.0000	0.0000	Luhya	0.9	0	0.1	0

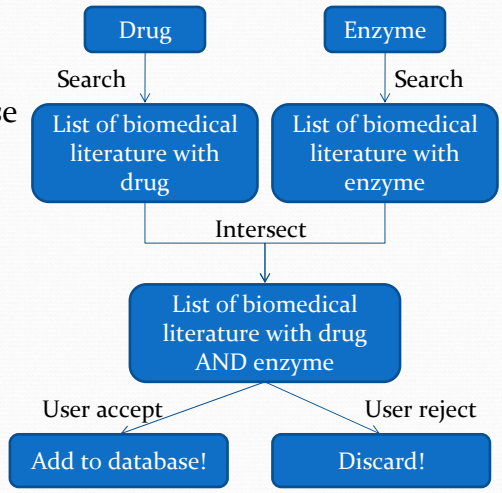
SNP drug-response biomarker discovery - III

- Multiple SNPs affecting drug response
  - Aim: Find sets of SNPs as biomarkers
  - Motivation: Single SNPs usually have small effects on drug response
  - Methodology: Mining emerging patterns (Algorithm from: 'Mining Statistically Important Equivalence Classes and Delta-Discriminative Emerging Patterns')



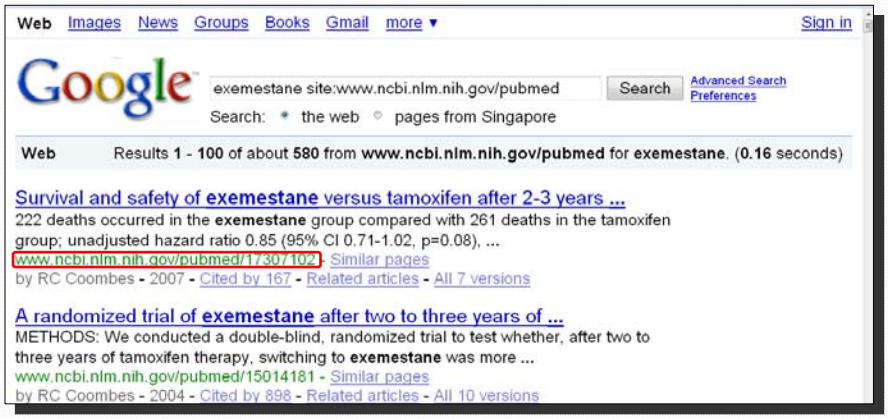
Drug-enzyme association discovery

- Drug-enzyme association discovery
  - Aim: Augment database for drug-enzyme associations
  - Motivation: Incompleteness of current downloaded databases
  - Methodology: Text association in biomedical abstracts



Drug-enzyme association discovery

- Drug-enzyme association discovery



Drug-enzyme association discovery

- Drug-enzyme association discovery

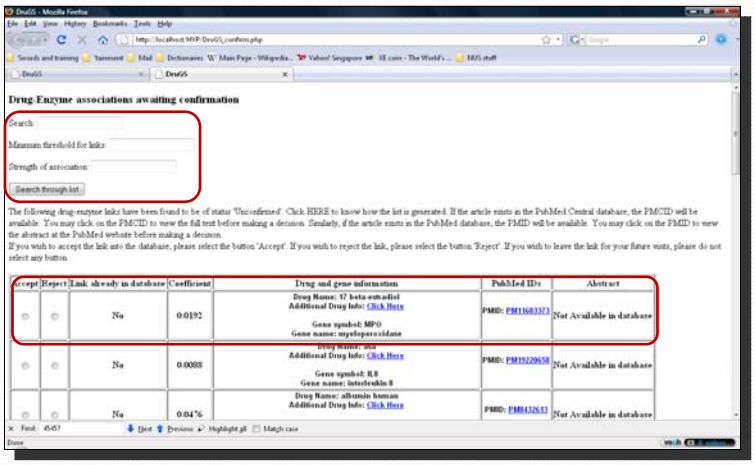
Drug	PMID
Exemestane	17307102
Exemestane	32643
Exemestane	1534647
Exemestane	1374543

Intersect

Gene	PMID
CYP1A1	164532
CYP1A1	1364332
CYP1A1	422141
CYP1A1	1534647

Drug	Gene	PMID
Exemestane	CYP1A1	1534647

Drug-enzyme association discovery



Drug-enzyme association discovery

- Jaccard Index

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}.$$

Drug	Gene	PMID
Exemestane	CYP1A1	1534647

Drug	PMID
Exemestane	17307102
Exemestane	32643
Exemestane	1534647
Exemestane	1374543

intersection

Gene	PMID
CYP1A1	164532
CYP1A1	1364332
CYP1A1	422141
CYP1A1	1534647

Drug-enzyme association discovery

- Results
  - 14 overlaps with known associations found
  - 8 with accurate supporting literature
- Results can be subjective
- Curation function for user feedback

Drug-enzyme association discovery

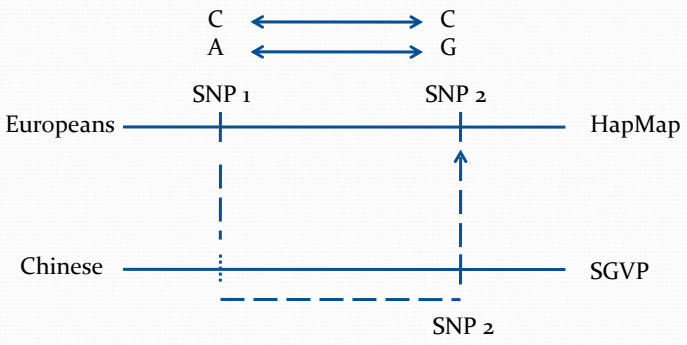
- Drawbacks:
  - False hits from Google
  - Drug name same as enzyme name

Conclusion

- SNPs and their importance in drug response
- Need for supplementing drug-enzyme association databases
- Development of methods to approach them

Future Improvements

- Linkage disequilibrium studies



Acknowledgements

I would like to thank Professor Wong Limsoon for his constant support, comments and patience throughout the development of this project.

THANK YOU!


Q & A