

Biological Large-scale data analysis —SNPs Disease-association Predictor

Ma Jing

1301213487

School of Chemical Biology & Biotechnology

Peking University

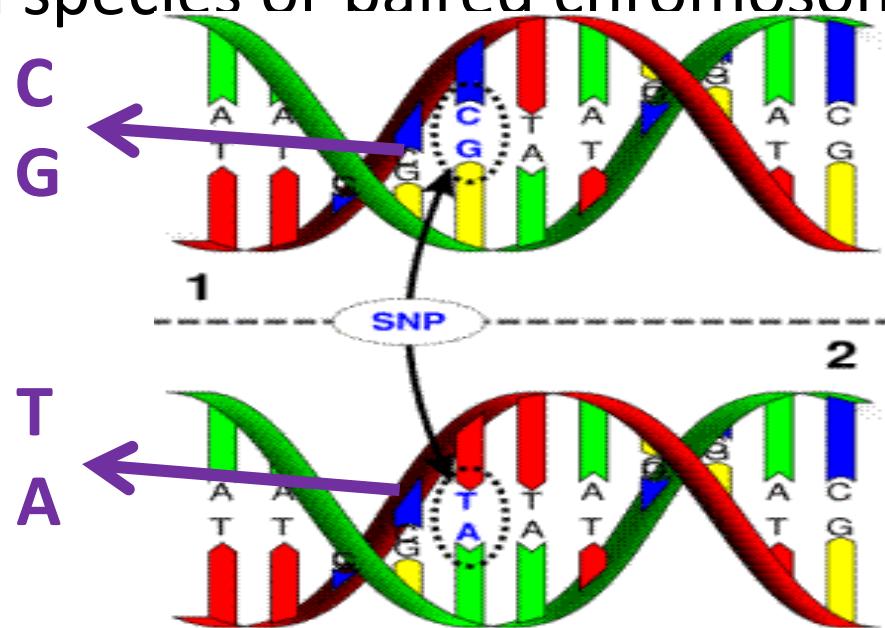
05/01/2014

Outline

- **Background**
- **Database**
- **Data**
- **Methods and Results**
- **Acknowledgement**

Background

- SNP(single nucleotide polymorphism) is a DNA sequence variation occurring when a single nucleotide — A, T, C or G — in the genome (or other shared sequence) differs between members of a biological species or paired chromosomes in a human.



Background

- **The difference between the SNP and the mutation**

SNP ——the variant frequency $\geq 1\%$

Mutation ——the variant frequency $< 1\%$

- **There are many SNPs in the human genome**

Every 1000 basic groups appear a SNP

3×10^6 SNPs in human genome

SCBB PKU



Background

➤ The International HapMap Project HapMap (Haplotype Map)



International HapMap Project

[Home](#) | [About the Project](#) | [Data](#) | [Publications](#) | [Tutorial](#)

[中文](#) | [English](#) | [Français](#) | [日本語](#) | [Yoruba](#)

The International HapMap Project is a partnership of scientists and funding agencies from Canada, China, Japan, Nigeria, the United Kingdom and the United States to develop a public resource that will help researchers find genes associated with human disease and response to pharmaceuticals. See "[About the International HapMap Project](#)" for more information.

Project Information

[About the Project](#)
[HapMap Publications](#)
[HapMap Tutorial](#)
[HapMap Mailing List](#)
[HapMap Project Participants](#)

Project Data

[HapMap Genome Browser release #28 \(Phases 1, 2 & 3 - merged \)](#)

News

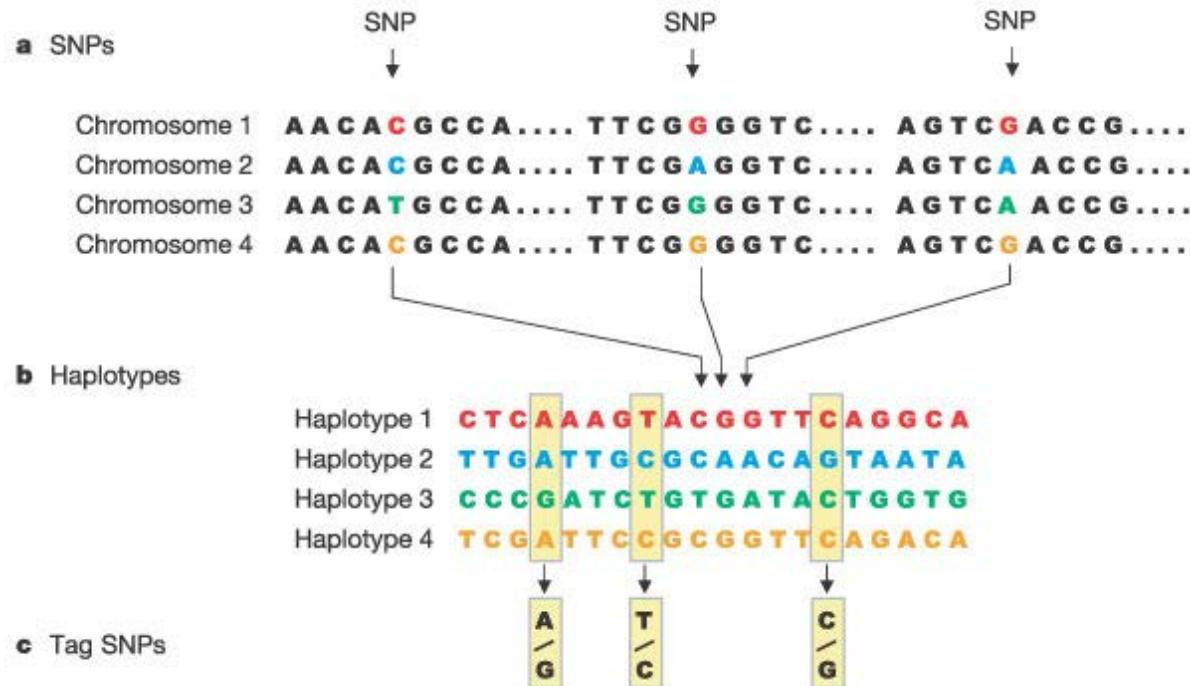
- 2013-06-14: **HapMap data conversion tool**
There are several inquiries for a conversion tool to convert HapMap data into the VCF format. Please take a look of [The Genome Analysis Toolkit](#) (by Broad Institute).
- 2012-12-06: **Downtime for hardware maintenance**
From December 15 – 16, Hapmap site will be taken offline for an internal hardware maintenance. Sorry for the inconvenience.
- 2011-06-13: **HapMap help desk announcement**

<http://hapmap.ncbi.nlm.nih.gov/index.html.en>



Background

➤ SNPs, haplotypes and tag SNPs



Background

➤ 1000 Genomes



The Phase 1 publication, an Integrated map of genetic variation from 1092 human genomes.

Database

➤ Uniprot

The screenshot shows the UniProt homepage. At the top, there's a dark blue header with the UniProt logo and a navigation bar with links: Search, Blast, Align, Retrieve, and ID Mapping. Below the header, there's a search form. It has a dropdown menu labeled "Search in" set to "Protein Knowledgebase (UniProtKB)". To its right is a "Query" input field, followed by a "Search" button, an "Advanced Search" link, and a "Clear" button.

WELCOME

The mission of [UniProt](#) is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information.

What we provide

NEWS

[UniProt release 2013_12 – Dec 11, 2013](#)

The aflatoxin biosynthetic pathway annotated in UniProtKB/Swiss-Prot

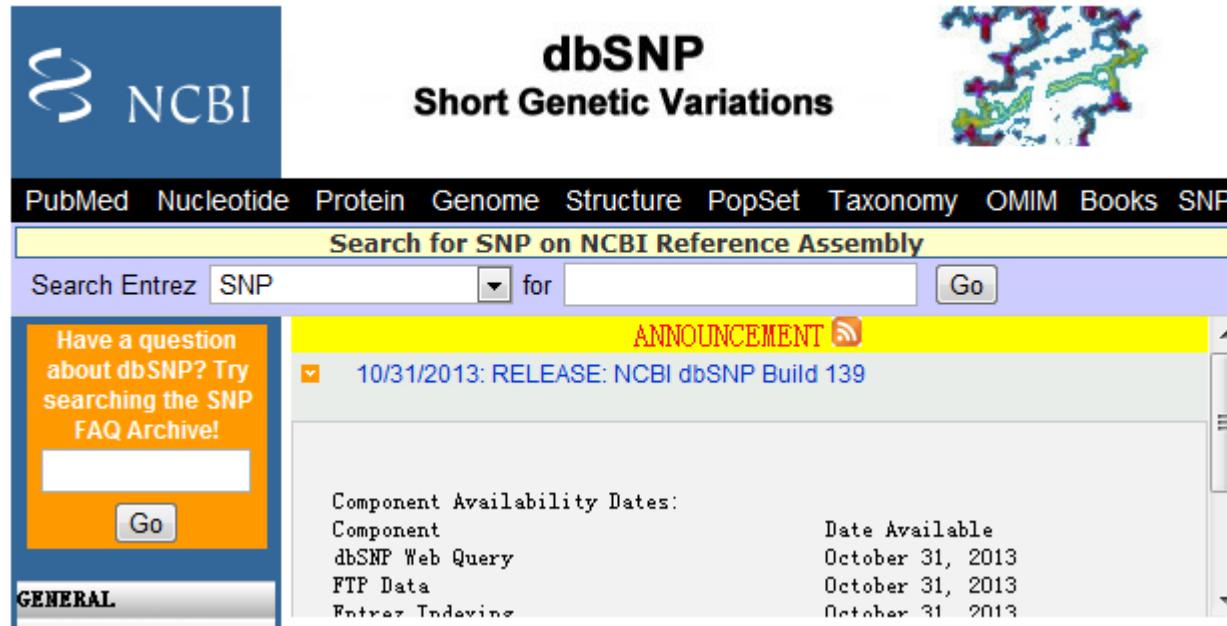
<http://www.uniprot.org/>

SCBB PKU



Database

➤ dbSNP



The screenshot shows the dbSNP homepage. At the top left is the NCBI logo. In the center, the text "dbSNP Short Genetic Variations" is displayed above a small 3D molecular model. A navigation bar at the top includes links for PubMed, Nucleotide, Protein, Genome, Structure, PopSet, Taxonomy, OMIM, Books, and SNP. Below the navigation bar is a search bar with the placeholder "Search for SNP on NCBI Reference Assembly". To the left of the search bar is a sidebar with a yellow background containing a question about the FAQ Archive and a "Go" button. The main content area features an "ANNOUNCEMENT" section with a link to "10/31/2013: RELEASE: NCBI dbSNP Build 139". Below this is a table titled "Component Availability Dates" showing the availability dates for various components.

Component	Date Available
dbSNP Web Query	October 31, 2013
FTP Data	October 31, 2013
Entrez Tandemwise	October 31, 2013

<http://www.ncbi.nlm.nih.gov/projects/SNP/>

SCBB PKU



Database

➤ SwissVar



[HOME](#) | [SEARCH](#) | [STATISTICS](#) | [DOCUMENTATION](#) | [USEFUL LINKS](#) | [CONTACT](#) | [PUBLICATIONS](#)

Search for disease - protein - variant associations



search

Enter a disease (e.g.: cataract), a protein or gene name (e.g.: Plasminogen)

A portal to Swiss-Prot diseases and variants

<http://swissvar.expasy.org/>

SCBB PKU



Database

➤ SwissVar

SwissVar is a portal to search variants in Swiss-Prot entries of the UniProt Knowledgebase (UniProtKB), and gives direct access to the Swiss-Prot Variant pages.

The screenshot shows a yellow sidebar on the left with a green MeSH logo and the text "Browse MeSH Tree Structure". Below it, under the heading "Diseases - (MeSH entry)", is a list of categories with their counts:

- endocrine system diseases (200) +
- congenital, hereditary, and neonatal diseases and abnormalities (1293) +
- disorders of environmental origin (0) +
- wounds and injuries (5) +
- parasitic diseases (0) +
- occupational diseases (3) +
- bacterial infections and mycoses (14) +
- substance-related disorders (0) +

SCBB PKU



Database

 SwissVar

Accession	Entry name	Disease	Variants	3D mapping (variant position)
P31947	1433S_HUMAN		p. Met155Ile	
Q96QU6	1A1L1_HUMAN	breast cancer	p. Gly221Glu p. Ser393Leu	
P13746	1A11_HUMAN		p. Glu43Lys p. Arg89Gly p. Gln94His	1Q94D (19) 1QVOA (65) 1Q94A (70)

National Library of
Medicine - Medical
Subject Headings(MeSH)

ExPASy-
UniProtKB/Swiss-Prot

variant pages

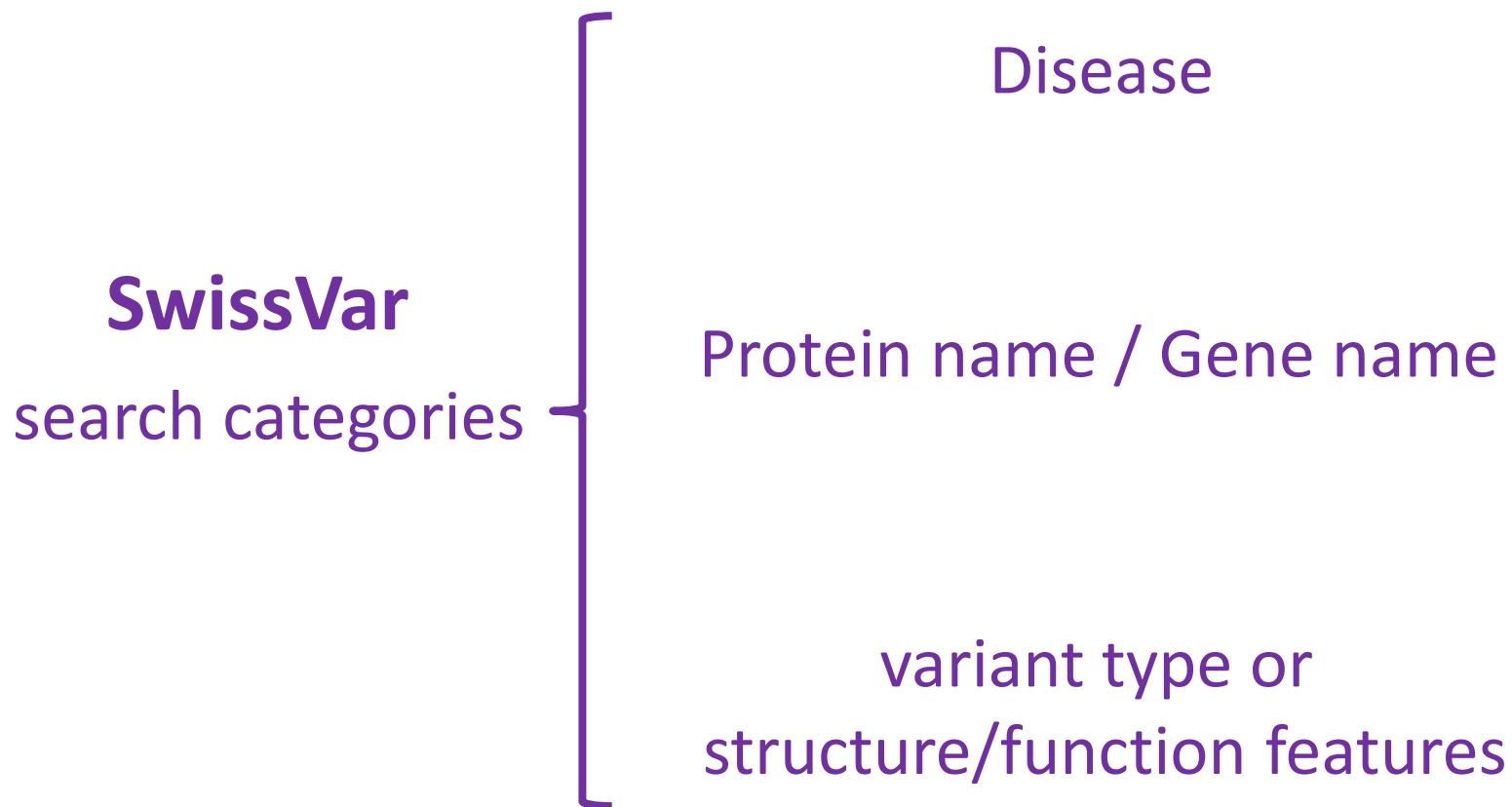
PDB

<http://swissvar.exasy.org/>

SCBB PKU



Database



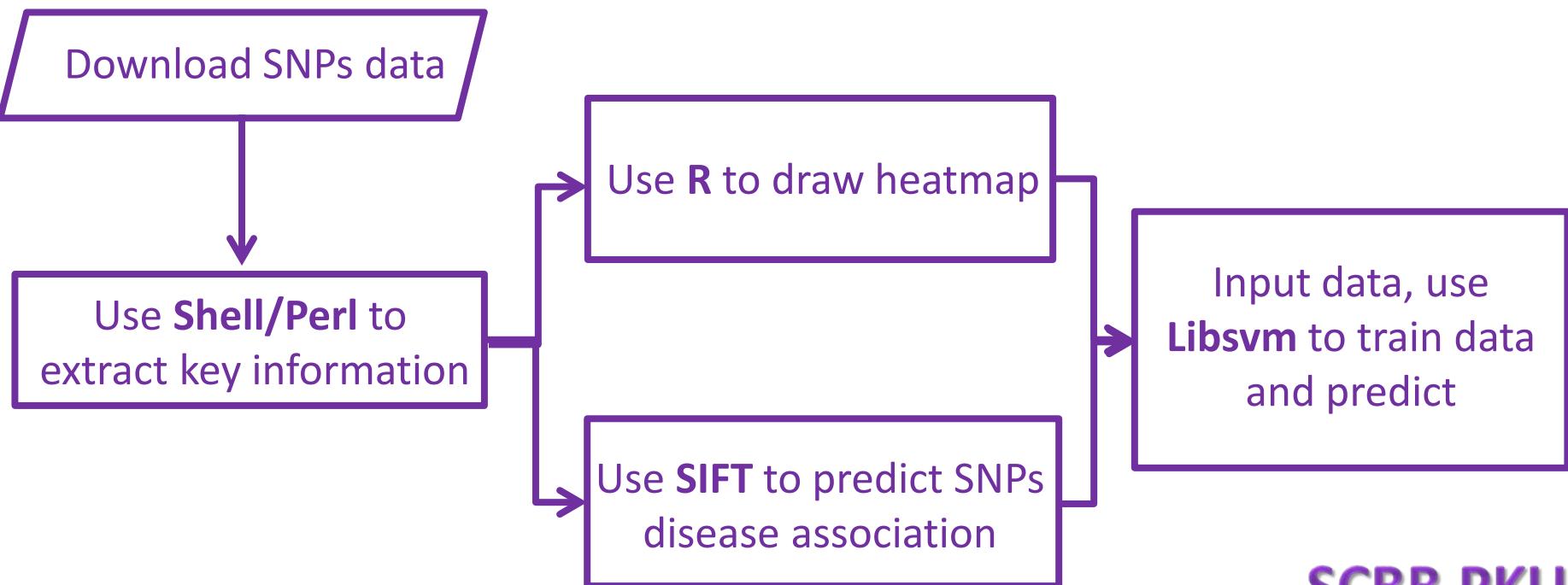
Data

- Download previous knowledgebase
- Get the file name “humsavar.txt”

字段	字段说明	字节
Main gene name	主要基因的名字	1-10 [10]
Swiss-Prot AC	Swiss-Prot登录号（Accession Number）	11-21 [11]
FTId	Swiss-Prot特征标识符	22-33 [12]
AAchange	格式为： p.突变前+位置+突变后	34-48 [15]
Type of variant	变异的类型属于Polymorphism、Disease或Unclassified	49-62 [14]
dbSNP	dbSNP的rs号	63-74 [12]
Disease name	涉及的疾病的名称（有些只是sample）	75-End [...]

Methods and Results

Analysis flow



SCBB PKU



Methods and Results

➤ Statistics and analysis data

- Linux shell script commonly used commands

✓ mkdir	Create a directory
✓ cd	Enter the directory
✓ lftp/get	download the files
✓ ls	View the files in the current directory
✓ tar	Compression and decompression command
✓ cat	View the files
✓ less/more	Selectively view
✓ rm	Delete files or directories

Methods and Results

➤ Statistics and analysis data

- Linux shell script commonly used commands

✓ pwd	Display the current directory path
✓ cp	Copy the file
✓ mv	Move or change the file and directory name
✓ grep	Search a specific string in the files
✓ cut	Interception of a column
✓ perl	Write a perl program at the command line
✓ awk	Write an awk program at the command line
✓ man	View command Manual

Methods and Results

➤ Statistics and analysis data

- Linux Perl script figure out 20*20 amino acids substitution table

*After
substitution*

Before
substitution

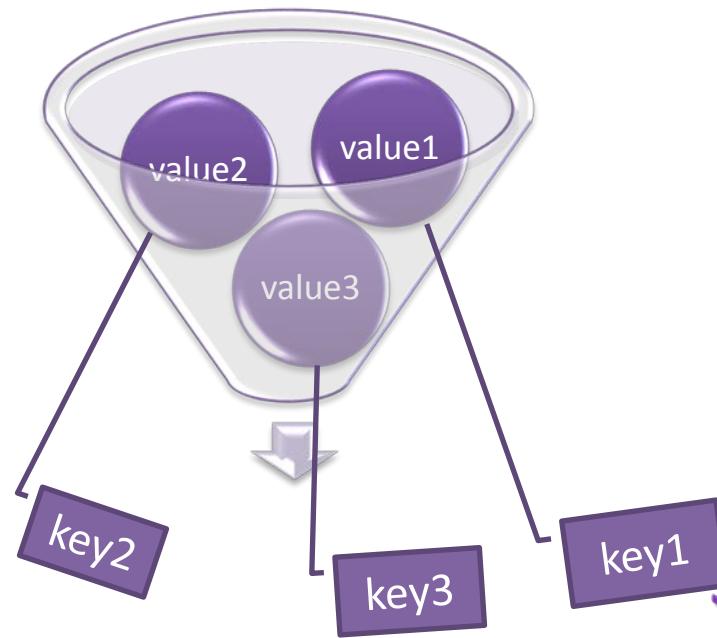
	Ala	Arg	Asp	Asn	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	0	1	116	0	0	117	251	0	1	4	1	1	0	237	308	1261	0	0	996	
Arg	4	0	2	1	700	1160	6	358	1005	38	177	312	30	0	144	213	78	572	0	0
Asp	77	4	0	525	0	2	346	234	143	0	2	0	0	0	0	5	0	0	108	92
Asn	0	0	274	0	0	0	5	0	90	49	0	241	0	1	0	642	98	0	44	0
Cys	0	201	0	0	0	1	0	43	0	0	0	1	0	42	0	136	0	41	174	0
Gln	0	541	2	0	0	0	187	1	334	0	72	124	0	0	104	0	0	3	0	0
Glu	132	2	401	3	0	248	0	310	0	0	4	726	0	0	0	1	2	0	0	85
Gly	229	545	291	1	68	1	245	0	0	0	1	0	0	0	0	570	0	42	0	189
His	0	420	55	68	0	203	1	0	0	0	42	0	0	0	63	1	0	0	241	0
Ile	1	15	0	65	0	0	0	0	0	0	120	11	210	73	0	34	504	0	0	891
Leu	3	114	2	0	0	63	1	0	48	135	0	0	150	431	425	148	3	36	4	440
Lys	2	427	0	254	0	133	358	0	0	16	0	0	37	0	0	0	106	0	0	1
Met	0	31	0	0	0	1	0	0	0	234	132	32	0	0	0	0	316	0	0	444
Phe	0	0	1	2	51	0	0	1	1	47	380	0	0	0	0	149	0	0	62	54
Pro	244	209	0	0	0	99	0	0	102	1	919	0	0	0	0	685	213	0	0	0
Ser	211	221	2	467	204	0	0	347	2	97	318	0	0	247	346	0	298	23	101	1
Thr	744	80	1	121	0	0	5	1	0	492	2	83	584	0	177	307	0	0	0	0
Trp	0	152	0	0	48	4	0	26	0	0	33	0	0	0	1	19	0	0	0	0
Tyr	0	0	38	27	266	1	0	1	204	1	5	0	0	75	0	54	1	0	0	1
Val	562	1	38	0	0	0	60	131	0	1068	392	0	706	83	0	1	0	0	1	0

Methods and Results

➤ Statistics and analysis data

- Linux Perl script figure out 20*20 amino acids substitution tables

hash {
 Key
 |
 | in pairs
 |
 value}



SCBB PKU

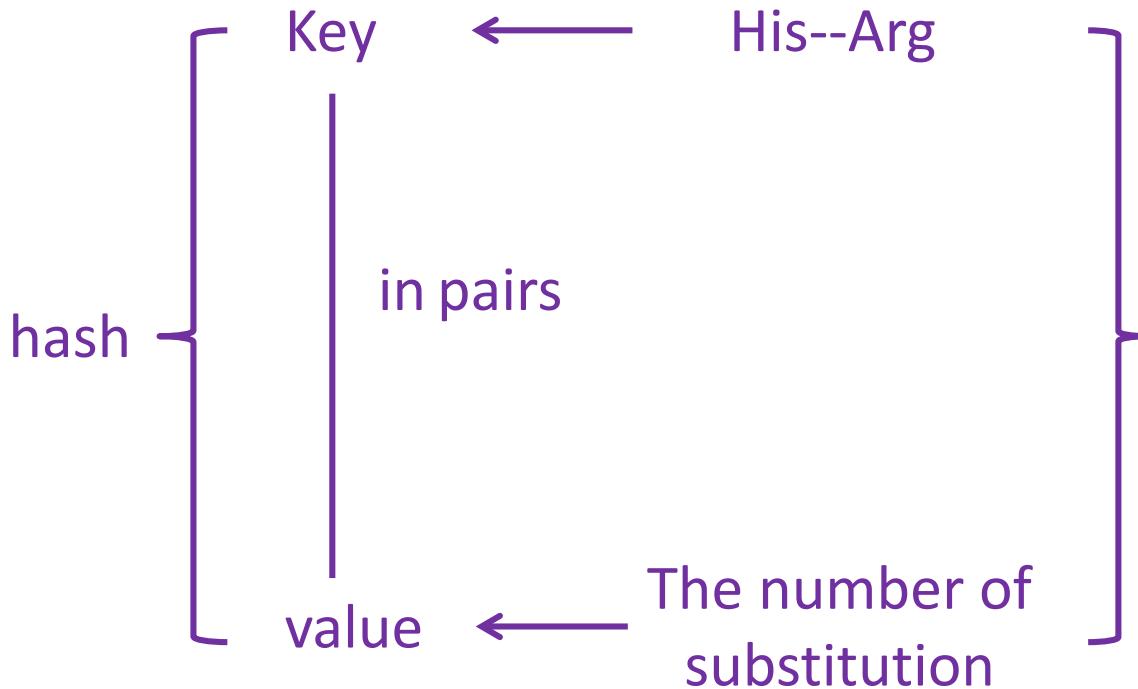


Methods and Results

➤ Statistics and analysis data

A1BG	P04217	VAR_018370	p.His395Arg	Polymorphism	rs2241788	-
A1CF	Q9NQ94	VAR_052201	p.Val555Met	Polymorphism	rs9073	-

substitution tables



	Arg	Met
His	XX	
Val		XX

SCBB PKU

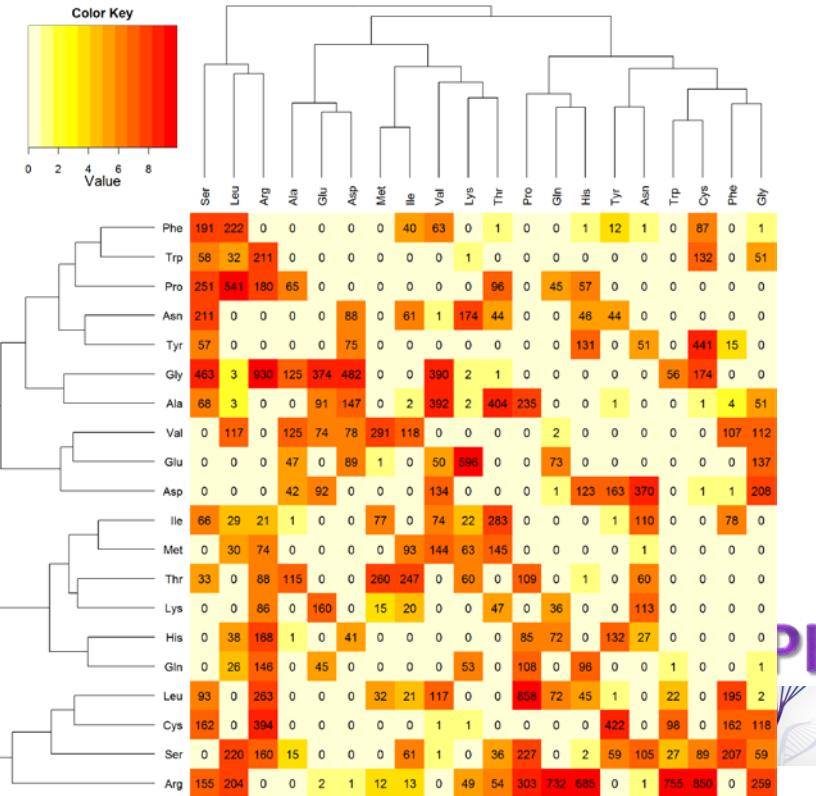


Methods and Results

➤ Draw heatmap

- R draw 20*20 substitution heatmap

	Ala	Arg	Asp	Asn	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	0	0	147	0	1	0	91	51	0	2	3	2	0	4	235	68	404	0	1	392
Arg	0	0	1	1	850	732	2	259	685	13	204	49	12	0	303	155	54	755	0	0
Asp	42	0	0	370	1	1	92	208	123	0	0	0	0	1	0	0	0	0	163	134
Asn	0	0	88	0	0	0	0	0	46	61	0	174	0	0	0	211	44	0	44	1
Cys	0	394	0	0	0	0	0	118	0	0	0	1	0	162	0	162	0	98	422	1
Gln	0	146	0	0	0	0	45	1	96	0	26	53	0	0	108	0	0	1	0	0
Glu	47	0	89	0	0	73	0	137	0	0	0	596	1	0	0	0	0	0	0	50
Gly	125	930	482	0	174	0	374	0	0	0	3	2	0	0	0	463	1	56	0	390
His	1	168	41	27	0	72	0	0	0	0	38	0	0	0	85	0	0	0	132	0
Ile	1	21	0	110	0	0	0	0	0	0	29	22	77	78	0	66	283	0	1	74
Leu	0	263	0	0	0	72	0	2	45	21	0	0	32	195	858	93	0	22	1	117
Lys	0	86	0	113	0	36	160	0	0	20	0	0	15	0	0	0	47	0	0	0
Met	0	74	0	1	0	0	0	0	0	93	30	63	0	0	0	0	145	0	0	144
Phe	0	0	0	1	87	0	0	1	1	40	222	0	0	0	0	191	1	0	12	63
Pro	65	180	0	0	0	45	0	0	57	0	541	0	0	0	0	251	96	0	0	0
Ser	15	160	0	105	89	0	0	59	2	61	220	0	0	207	227	0	36	27	59	1
Thr	115	88	0	60	0	0	0	0	1	247	0	60	260	0	109	33	0	0	0	0
Trp	0	211	0	0	132	0	0	51	0	0	32	1	0	0	0	58	0	0	0	0
Tyr	0	0	75	51	441	0	0	0	131	0	0	0	15	0	57	0	0	0	0	0
Val	125	0	78	0	0	2	74	112	0	118	117	0	291	107	0	0	0	0	0	0



Methods and Results

➤ Draw heatmap

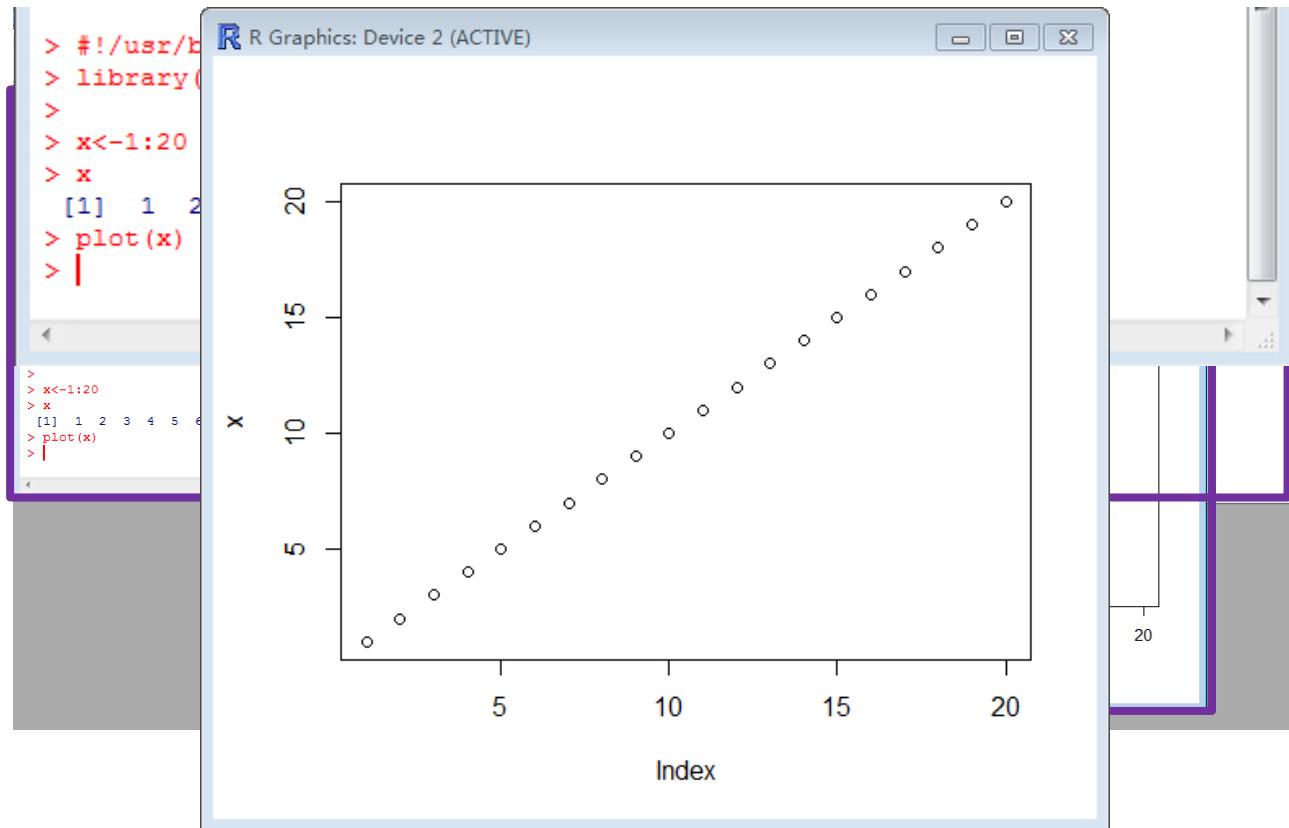
- R draw 20*20 substitution heatmap

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	0	0	0	609	0	0	762	1139	0	1	0	0	0	0	1011	2418	7885	0	0	5937
Arg	0	0	0	0	1624	3661	0	1498	3123	112	684	2680	151	0	503	1158	349	1027	0	0
Asn	0	0	0	1553	0	0	0	0	520	265	0	1237	0	0	0	3021	670	0	250	0
Asp	445	0	2671	0	0	0	2678	1567	380	0	0	0	0	0	0	0	0	0	311	278
Cys	0	703	0	0	0	0	0	230	0	0	0	0	0	0	307	0	649	0	179	840
Gln	1	2046	0	0	0	0	907	0	1707	0	513	989	0	0	610	0	0	0	0	0
Glu	621	0	0	2714	0	930	0	1516	0	0	0	2889	0	0	0	0	0	0	0	395
Gly	1074	2160	0	1747	331	0	1744	0	0	0	0	0	0	0	2970	0	117	0	839	0
His	0	1708	547	261	0	1458	0	0	0	333	0	0	0	301	0	0	0	1393	0	0
Ile	2	100	303	0	0	0	0	0	0	970	110	1239	379	0	333	1930	0	0	4305	0
Leu	0	478	0	0	0	483	0	0	280	1059	0	0	1093	2561	1734	772	0	111	0	1752
Lys	0	2290	1295	0	0	777	1728	0	0	134	0	0	228	0	0	0	634	0	0	0
Met	0	173	0	0	0	0	0	0	1791	898	186	2	0	1	0	1322	0	0	0	1693
Phe	0	0	0	0	235	0	0	0	337	1969	0	0	0	0	0	747	0	0	434	361
Pro	1027	550	0	0	0	644	0	0	424	1	3438	0	0	0	0	4240	1092	0	0	0
Ser	1526	1302	3834	0	856	0	0	1859	0	501	1399	0	0	1276	2690	0	2244	71	337	0
Thr	4888	363	766	0	0	0	0	0	3007	0	664	2083	0	903	2335	0	0	0	0	0
Trp	0	455	0	0	224	0	0	89	0	0	118	0	0	0	66	0	0	0	0	0
Tyr	0	0	223	170	673	0	0	0	911	0	0	0	566	0	242	0	0	0	0	0
Val	3587	0	0	248	0	0	377	679	0	6097	1902	0	2718	429	0	1	0	0	0	0

Methods and Results

➤ Draw heatmap

- R introduction



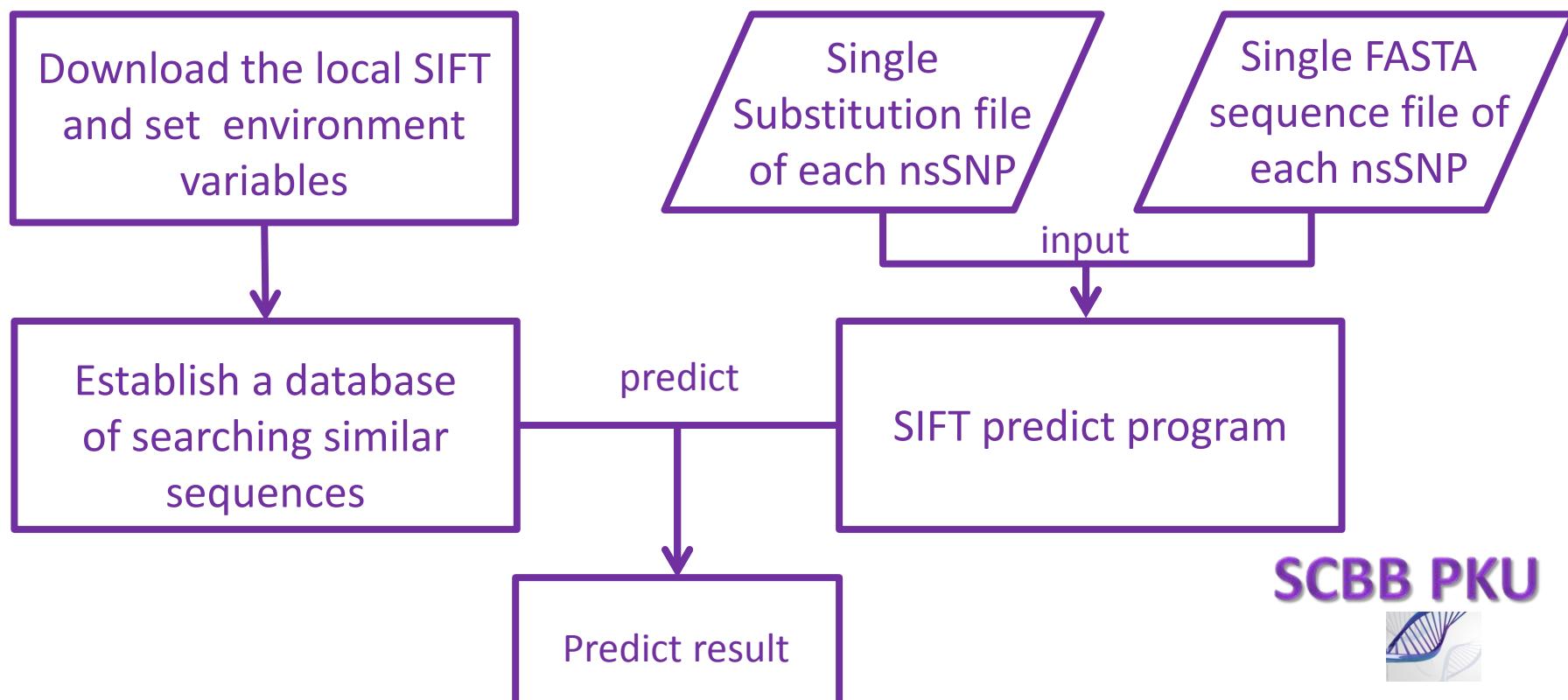
SCBB PKU



Methods and Results

➤ SIFT prediction

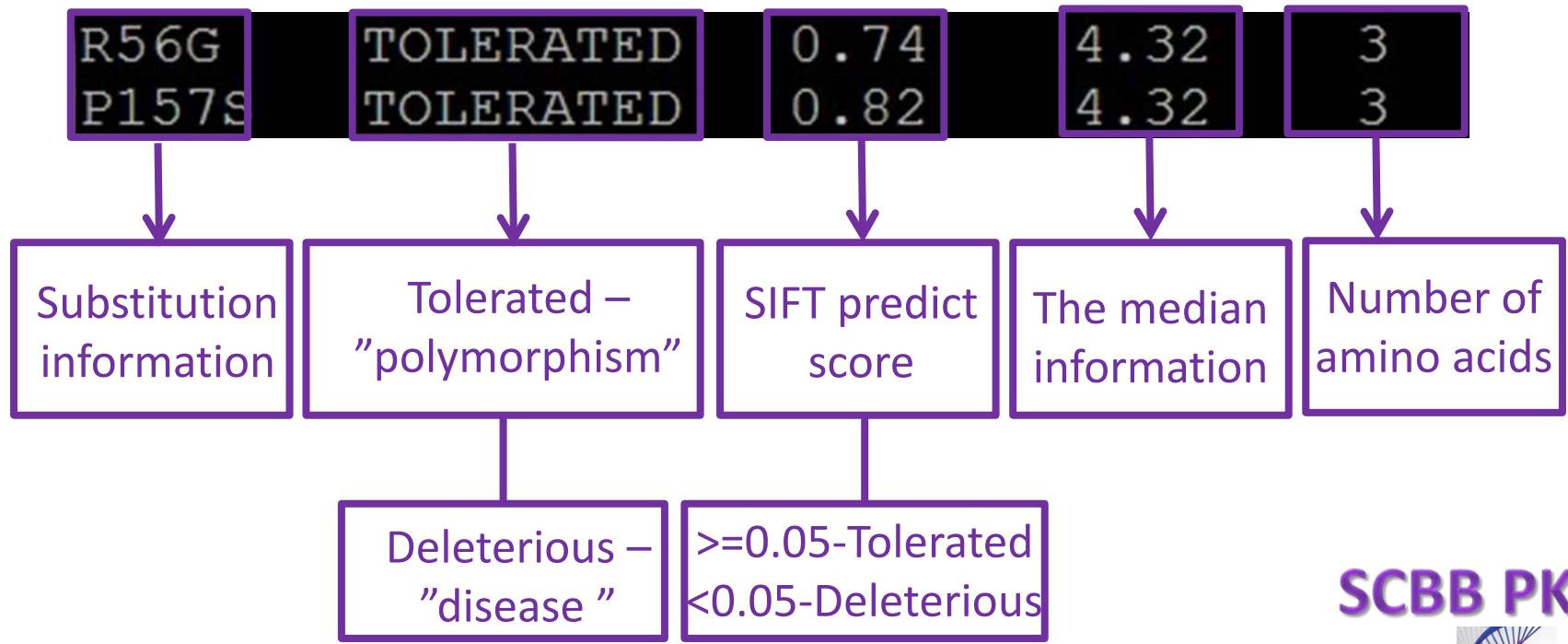
- Local SIFT prediction



Methods and Results

➤ SIFT prediction

- SIFT predict results

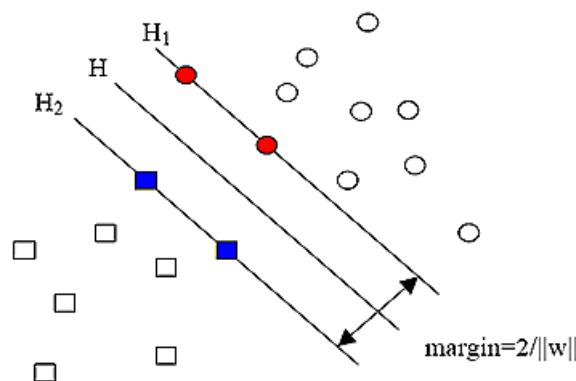
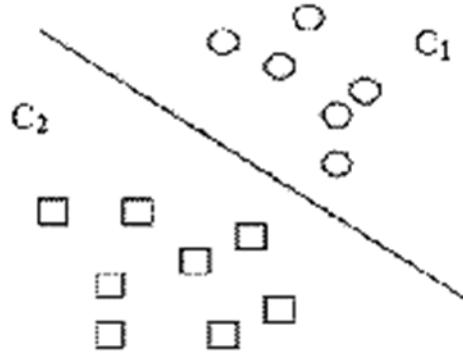


Methods and Results

➤ Libsvm prediction

- SVM introduction

SVM are supervised learning models with associated learning algorithms that analyze data and recognize patterns, used for classification and regression analysis.

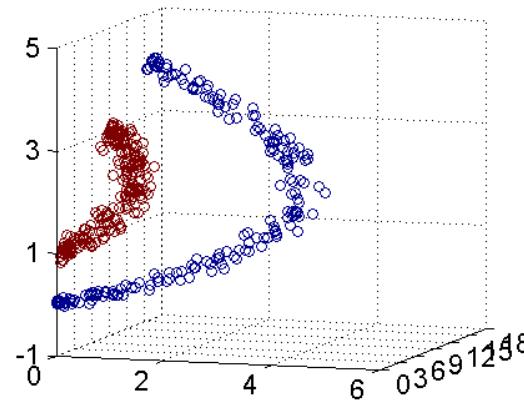
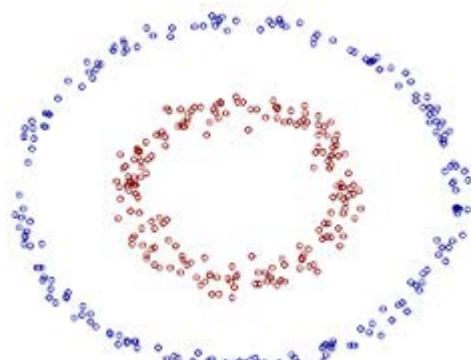


Methods and Results

➤ Libsvm prediction

- SVM introduction

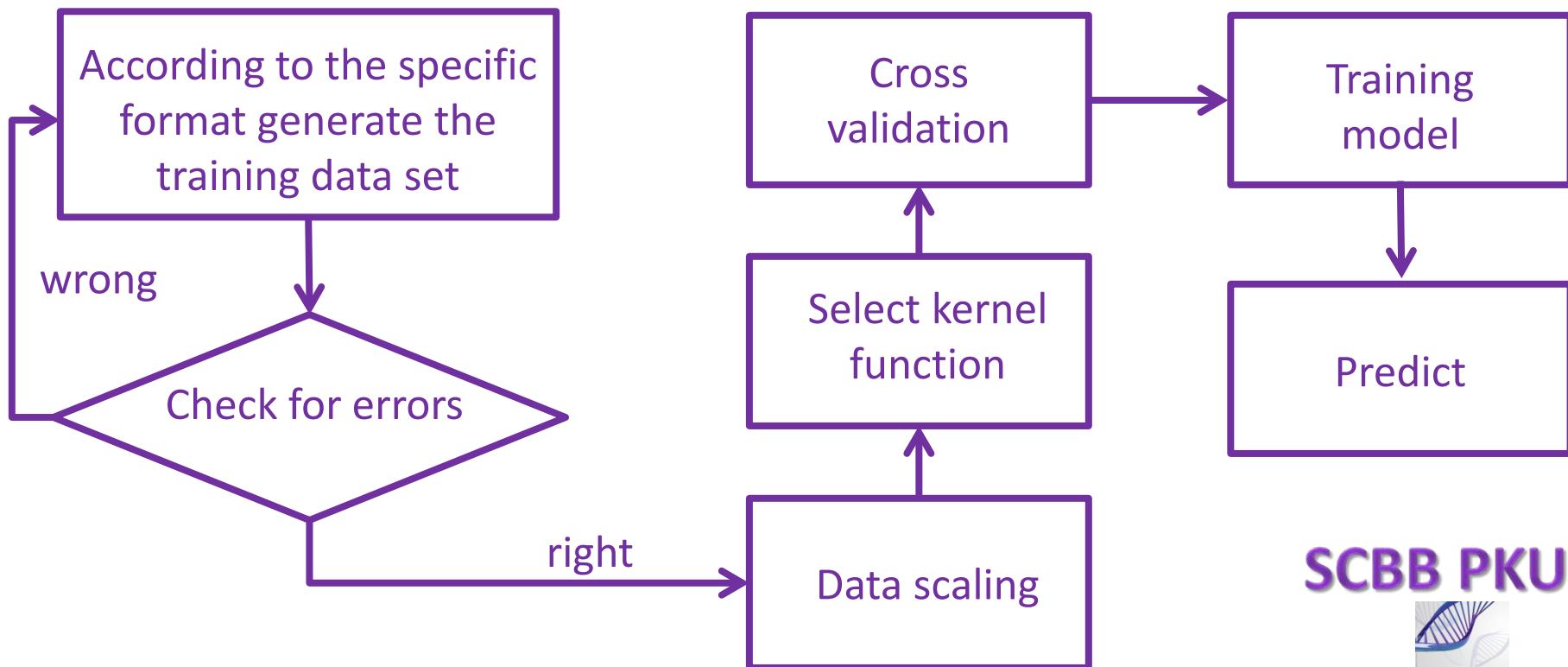
SVMs can efficiently perform a non-linear classification using what is called **the kernel trick**, implicitly mapping their inputs into **high-dimensional feature spaces**.



Methods and Results

➤ Libsvm prediction

- Libsvm training and prediction process



How to process biological data?

➤ Libsvm prediction

- Libsvm training and predicted example

P04217	H52R	VAR_018369	Polymorphism
P04217	H395R	VAR_018370	Polymorphism
Q9NQ94	V555M	VAR_052201	Polymorphism
Q9NQ94	A558S	VAR_059821	Polymorphism
A8K2U0	G207R	VAR_055463	Polymorphism
A8K2U0	C970Y	VAR_055464	Polymorphism
A8K2U0	T1131M	VAR_055465	Polymorphism
A8K2U0	T1412A	VAR_055466	Polymorphism
A8K2U0	D850E	VAR_059083	Polymorphism
A8K2U0	H1229R	VAR_059084	Polymorphism
P01023	R704H	VAR_000012	Polymorphism
P01023	C972Y	VAR_000013	Polymorphism
P01023	I1000V	VAR_000014	Polymorphism
P01023	N639D	VAR_026820	Polymorphism
P01023	L815Q	VAR_026821	Polymorphism
Q9NPC4	M37V	VAR_014296	Polymorphism
Q9NPC4	G187D	VAR_017508	Polymorphism
Q9NPC4	P251L	VAR_017509	Polymorphism
Q9NPC4	Q163R	VAR_022320	Polymorphism
Q9UNA3	A218D	VAR_022096	Polymorphism

GRANTHAM matrix scores

0	29
0	29
1	21
1	99
-2	125
-2	194
-1	81
0	58
2	45
0	29
0	29
-2	194
3	29
1	23
-2	113
1	21
-1	94
-3	98
1	43
-2	126



BLOSUM62 matrix scores

SCBB PKU



How to process biological data?

➤ Libsvm prediction

- Libsvm training and predicted example

-1	1:0	2:29
-1	1:0	2:29
-1	1:1	2:21
-1	1:1	2:99
-1	1:-2	2:125
-1	1:-2	2:194
-1	1:-1	2:81
-1	1:0	2:58
-1	1:2	2:45
-1	1:0	2:29
-1	1:0	2:29
-1	1:-2	2:194
-1	1:3	2:29

Scaling →

-1	1:0.142857	2:-0.771429
-1	1:0.142857	2:-0.771429
-1	1:0.428571	2:-0.847619
-1	1:0.428571	2:-0.104762
-1	1:-0.428571	2:0.142857
-1	1:-0.428571	2:0.8
-1	1:-0.142857	2:-0.27619
-1	1:0.142857	2:-0.495238
-1	1:0.714286	2:-0.619048
-1	1:0.142857	2:-0.771429
-1	1:0.142857	2:-0.771429
-1	1:-0.428571	2:0.8



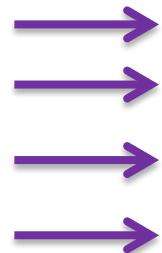
How to process biological data?

➤ Libsvm prediction

- Libsvm training and predicted example

Scaling parameter file

```
X  
-1 1  
1 -4 3  
2 5 215
```

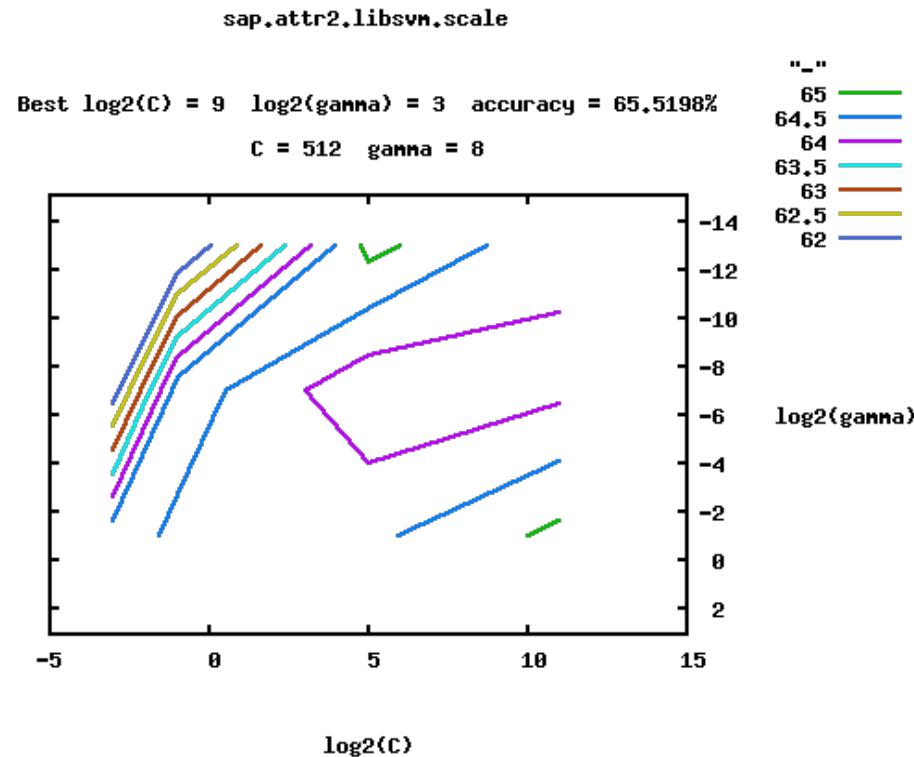


```
Scaling X  
minimum maximum  
feature1: minimum maximum  
feature2: minimum maximum
```

How to process biological data?

➤ Libsvm prediction

- Libsvm training and predicted example-Cross validation result



Acknowledge

- **Thanks for my mentors Dr. Ye and Dr. Gao !**
- **Thanks for my group members!**
- **Thanks for ABC course teacher Dr. Luo and classmates !**

Acknowledge

Thanks for your attention!