
Table of Contents

preface	1.1
genotyping	1.2
workflow	1.2.1
prospect	1.2.2
drug-reaction	1.3
introduction	1.3.1
workflow	1.3.2
prospect	1.3.3
ancestry	1.4
ancestry-composition	1.4.1
haplotyper-wrapper	1.4.2
workflow	1.4.3
prospect	1.4.4
hereditary-disease	1.5
workflow	1.5.1
prospect	1.5.2
database	1.6
Others	1.7
genecard	1.7.1
gene-imputation	1.7.2

芯片介绍

Affy PMRA

Axiom® PMRA芯片是Affymetrix推出的适用于全球人种精准医学研究的性价比非常高的商业化基因分型芯片，包含903K个标记，涵盖了全球五大祖先超过25个民族，精选全球人在千人基因组三期计划中最小等位基因频率范围的位点，覆盖从公共数据库，包括ClinVar、CPIC、PharmaGKB、PharmaADM选择的最综合的临床相关变异和重要的且经过验证的药物基因组学的内容，是全球人种全基因组关联分析和转化研究的强有力工具。

Infinium Global Screening Array(GSA)

GSA芯片是illumina推出的覆盖全球人种的用于转化研究的非常经济的基因分型芯片，包含660K个标记。芯片内容参考千人基因组三期计划的数据，MAF>1%。

芯片好处

- 单位样本成本低
- 快速、高通量和多重样本的处理
- 快速、直观的分析流程
- 提供专家精选、或者定制和半定制的形式

比较

	PMRA	GSA
价格\$	29	40
单次上机样本数	96	24
位点数	903K	660K
基因分型工具	支持Linux的apt工具	GenomeStudio

分析内容比较

截止2017-12-07，PMRA和GSA芯片可分析项目类型个数：

	PMRA	GSA
祖源分析	3	3
药物反应	193	179
健康风险	108	105
遗传疾病	1052	1037
生活指导	54	46

注意：遗传疾病是按基因进行的统计，如果按rs位点仅有553个；

上述统计仅为可以做的项目，实际转换为GSA时，数据库还需要新增和修改的工作

数据分型

过滤条件

样本过滤

- $DQC \geq 0.92$
- 样本CallRate $\geq 97\%$

探针过滤

- 过滤掉ConversionType为CallRateBelowThreshold
- 多个探针对应同一rs号，按PolyHighResolution, NoMinorHom, MonoHighResolution, Hemizygous, OTV的优先级和检出率选择最终保留的探针
- 探针CallRate $\geq 96\%$
- 单基因病位点杂合率 $> 5\%$

workflow

Illumina GSA分型目前只提供Windows版本，并没有实现自动化

Requirements

- perl
 - Modern::Perl
 - Data::Dumper
 - IO::All
 - Cwd
 - FindBin
 - Getopt::Long::Descriptive
 - File::Basename
 - JSON;
 - MCE::Loop
 - DBI
- rust
 - XSV

使用方法

```
perl ~/workdir/SNParray/data_processing_9800/run.pl

必选参数：
-i STR --indir STR    下机CEL文件路径
-o STR --outdir STR   结果输出路径

可选参数：
-p STR --para STR     分析用的参数文件
-s STR --step STR      选择步骤，默认全选
      1: 样本质控QC
      2: 初步分型
      3: OTV Caller
      4: 探针过滤和选择
      5: 样本分型和格式化输出
```

结果目录

```
├── cel_list1.txt #下机CEL文件列表
├── cel_list2.txt #样本DQC过滤后合格的CEL文件列表
└── cel_list3.txt #样本CallRate过滤后合格的CEL文件列表

├── QC #质控结果
│   ├── apt-genotype-qc.txt
│   ├── AxiomGT1.calls.txt
│   ├── AxiomGT1.confidences.txt
│   └── AxiomGT1.report.txt

└── AxiomGT1.calls.txt #初次分型结果
├── AxiomGT1.confidences.txt
└── AxiomGT1.report.txt
```

```
|--- AxiomGT1.snp-posteriors.txt  
|--- AxiomGT1.summary.txt  
|--- Ps.performance.txt  
  
|--- raw_data #初次分型对应的样本数据  
|--- raw.txt  
  
|--- *.ps #各种ConversionType对应的探针列表  
  
|--- OTV_Caller #OTV分型结果  
  
|--- cmd.sh #运行命令  
|--- data #最终结果  
|   |--- final 最终过滤后的样本结果  
|   |--- used 分析中用到位点的样本结果  
|--- final  
|   |--- sample.callrate.txt #样本检出率以及杂合率的最终统计结果  
|   |--- probe.callrate.txt #探针检出率以及ConversionType的最终统计结果
```

展望

流程

- 绘制位点的聚类分布图（SNPolisher: Ps_Visualization）
- 分型文件结果格式调整（染色体位置）
- 与后续流程联合，实现下机数据->数据质控->数据分型->导入数据库->常规分析->报告和线上系统更新等一套流程的自动化；

质控

- 目前位点检出率为0.96,当样品数从96降至23或更低时，这个标准还有待商榷
- 评估不同的位点检出率对健康风险和单基因病结果判定的影响

数据库

- 建立结果统计数据库，包含位点的频率、分型、检出率等以及样品的检出率、杂合率等相关信息；
- 建立数据的结果关系库（样本信息），样本重复测序和谱系关系；

Illumina 数据处理

- Illumina GSA数据处理以及相关质控条件的确定

药物反应

药物反应模块，涉及19个大类，共计194种常见药物。

孟鲁司特 (Montelukast)

药物分类

平喘药 >> 白三烯受体阻断剂

药物介绍

孟鲁司特适用于15岁及15岁以上成人哮喘的预防和长期治疗，包括预防白天和夜间的哮喘症状，治疗对阿斯匹林敏感的哮喘患者以及预防运动诱发的支气管收缩；也适用于15岁及15岁以上成人以减轻季节性过敏性鼻炎引起的症状。

检测结果

根据目前的研究进展，提示您的检测结果中：

- 药物有效性相关证据有2项，其中2项有效性较差。
- 药物毒性未见相关研究。
- 药物代谢未见相关研究。
- 药物剂量未见相关研究。
- 药物血药浓度未见相关研究。

综上所述，针对该药物建议您慎重使用。

优选等级： ★★☆☆☆

尚属
研究
阶段

检测详情

位点名	基因名称	基因型	用药提示	提示类型	证据等级
CL_DC_183844	ABCC1	C/C	携带CC基因型的哮喘患者，相比CT基因型，基于强力呼气容积测定(FEV1)，药效可能不会提升	有效性	3
CL_DC_159332	-	G/G	携带GG基因型的哮喘患者，相比GT或TT基因型，药物反应可能较差	有效性	3

药物证据等级说明：

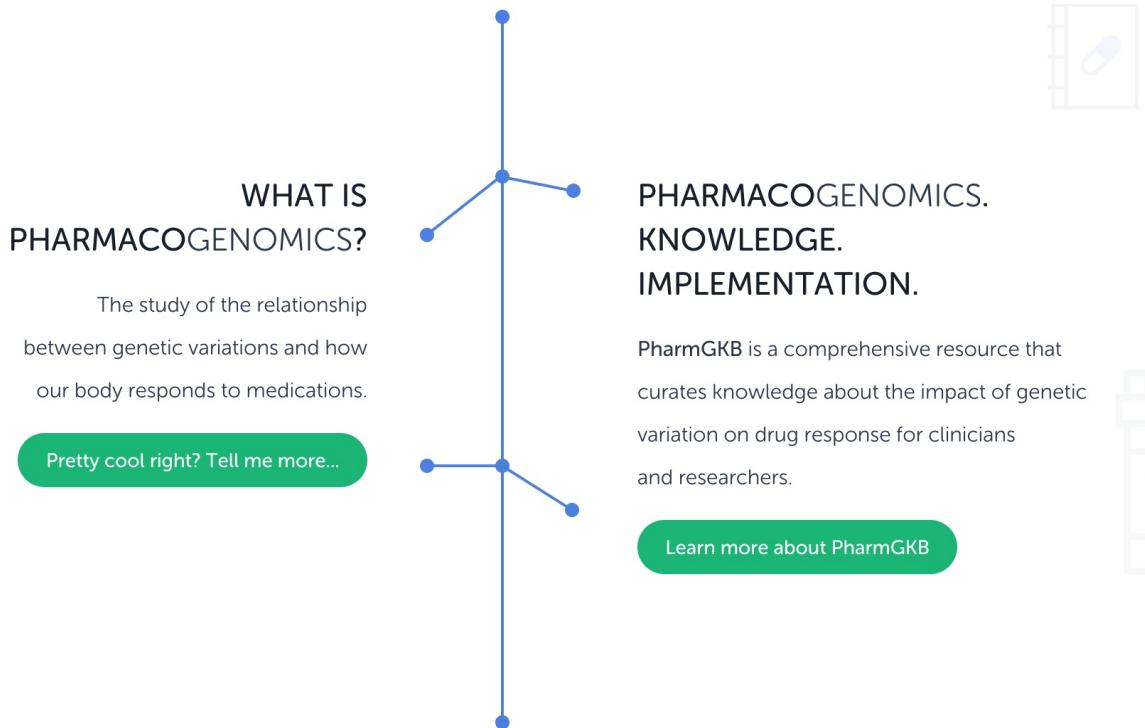
1A：由临床药物基因组学实施联盟(CPIC)或遗传药理学指南确认，或在国际遗传药理学研究网(PGRN)及其他主要卫生系统中已有应用；
 1B：多项证据支持药物与变异之间的相关性，研究具有显著性差异且影响规模较大，结论可重复；
 2A：多项证据支持药物与变异之间的相关性，且基因是功能明确、意义重大的药物代谢基因；
 2B：多项证据支持药物与变异之间的相关性，但其中一些研究无显著性差异和/或影响规模较小，结论可重复；
 3：单一显著性差异研究中支持药物与变异之间的相关性，或有多项研究支持其相关性但尚未达成一致结论。

相关介绍

数据库

PharmGKB

PharmGKB (药物遗传学和药物基因组学知识库) 由美国国立卫生研究院 (NIH) 创建，收集了史上最完整的与药物基因组相关的基因型和表型信息，并将这些信息系统地归类。



截止2017-10-25，共收录的信息统计如下：



截止2017-10-25，pharmGKB手工注释的信息统计如下：

Annotations			
Clinical		Research	
	DOSING GUIDELINES	96	
	DRUG LABELS	476	
	CLINICAL ANNOTATIONS	3,330	VIPs (Very Important Pharmacogenes)
			VARIANT ANNOTATIONS
			18,496

- 网址: <https://www.pharmgkb.org>
- 文献: <https://www.ncbi.nlm.nih.gov/pubmed/22992668>
- 下载: <https://api.pharmgkb.org/>

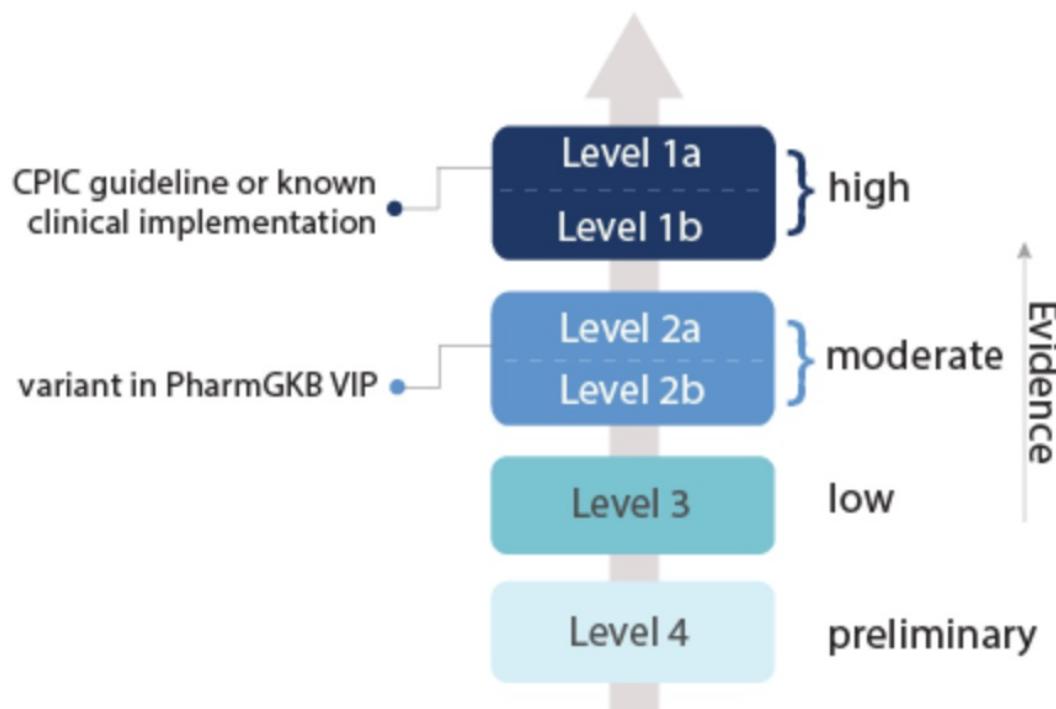
专业词汇

PGx

PharmGKB annotates drug labels containing pharmacogenetic information approved by FDA, EMA, PMDA, HCSC. PharmGKB annotations provide a brief summary of the PGx in the label, an excerpt from the label.

- **Testing required:** PharmGKB considers labels that state the variant is an indication for the drug, as implying a test requirement.
- **Testing recommended:** The label states or implies that some sort of gene, protein or chromosomal testing, including genetic testing, functional protein assays, cytogenetic studies, etc., is recommended before using this drug.
- **Actionable PGx:** The label may mention contraindication of the drug in a particular subset of patients but does not require or recommend gene, protein or chromosomal testing.
- **Informative PGx:** The label mentions a gene or protein is involved in the metabolism or pharmacodynamics of the drug, but there is no information to suggest that variation in these genes/proteins leads to different response.

证据等级



- **Level 1A:** Annotation for a variant-drug combination in a CPIC or medical society-endorsed PGx guideline, or implemented at a PGRN site or in another major health system.
- **Level 1B:** Annotation for a variant-drug combination where the preponderance of evidence shows an association. The association must be replicated in more than one cohort with significant p-values, and preferably will have a strong effect size.
- **Level 2A:** Annotation for a variant-drug combination that qualifies for level 2B where the variant is within a VIP (Very Important Pharmacogene) as defined by PharmGKB. The variants in level 2A are in known pharmacogenes, so functional significance is more likely.
- **Level 2B:** Annotation for a variant-drug combination with moderate evidence of an association. The association must be replicated but there may be some studies that do not show statistical significance, and/or the effect size may be small.

- **Level 3:** Annotation for a variant-drug combination based on a single significant (not yet replicated) study or annotation for a variant-drug combination evaluated in multiple studies but lacking clear evidence of an association.
- **Level 4:** Annotation based on a case report, non-significant study or in vitro, molecular or functional assay evidence only.

芯片pharmGKB注释结果对比

证据等级	Illumina	Affy
1A	29	21
1B	15	12
2A	66	46
2B	80	49
3	1863	788
4	214	124
总计	2267	1040

国内综合型数据库

涵盖药物研发（药物注册以及转让、专利信息、研发阶段）、生产检验（药品标准、参比制剂）、合理用药（药品说明书、医保目录、药物相互作用）、市场信息（政策法规、售价、上市信息）以及化学结构等一系列过程。

- 药智: <https://db.yaozh.com/>
- Insight: <http://db.dxy.cn/v5/home>

分析原理

数据库过滤

1. 过滤掉证据等级为4的位点；
2. **to be continued**

优选等级判定

- 根据一种药物的不同等级的毒性和有效性个数，确定x、y、z

个数	有效性等级	毒性等级
x	好	小
y	中间等级	中间等级
z	差	大

- 根据x、y、z，分别进行有效性(a)和毒性(b)的评分

	有效性(a)	毒性(b)
$x=y=z=0$	0	0
$x \geq y+z \ \&\& x \neq z$	10	5
$x+y < z$	2	1

x+y >=z	8	4
---------	---	---

- 根据有效性和毒性的评分，判定优选等级L

$$L = \text{floor}((a + b)/4) + 2$$

行业动态

23andMe

共12项

参考网址: <https://www.23andme.com/en-gb/health/reports>

wegene

共11项

参考网址: <https://www.wegene.com/demo/female/report/detail/1481>

23魔方

6大类，共57种。

参考网址: <https://www.23mofang.com/sample/drug>

Workflow ReadMe

NAME

Drug::Reaction - Drug Reaction Analysis in 9800

VERSION

version 1.000

SYNOPSIS

```
use Drug::Reaction;
my $instance = Drug::Reaction->new(
    para => "dist_dir/share",
    step => "1,2",
    genotype => "sample1.csv",
    gender => "gender.csv",
    outdir => "drug-reaction_results"
);
$instance->drug_reaction_main;

or

drug-reaction --help
```

ATTRIBUTES

genotype

affy genotype csv file or dir included csv files, csv file formate as "rs_id,chemical_id,pharmgkb_id,Plate ID,Sample ID,Call Code"

gender

gender csv file, formate as "sample_id,gender", gender must be chinese

para

share directory, included essential files for analysization

init db

- pharmgkb-alleles-description-20170810.csv
- pharmgkb-alleles.csv
- pharmgkb-annotations.csv
- pharmgkb-chemicals-summary.csv
- pharmgkb-chemicals.csv
- pharmgkb-evidences.csv

analysis

- chemical_info.csv

step

- 0: database initialize
- 1: run the drug-reaction process, default is 1
- 2: put the results to database
- 1,2: run step 1 and step 2

db_url

drug-reaction database url, default is 'postgres://postgres:123456@192.168.1.205:5439/pharmgkb'

outdir

analysis output directory

METHODS

init_db

drug_reaction database initialize, must provide para directory

drug_reaction

drug reaction analysis, must provide genotype and gender

upload_to_db

execute sql within db_url

drug_reaction_main

drug reaction analysis pipeline, has three steps:

- 0 initialize db
- 1 drug-reaction analysis
- 2 upload results to database

AUTHOR

Su Min sumin@cheerlandgroup.com

COPYRIGHT AND LICENSE

This software is copyright (c) 2017 by CheerLand Group.

This is free software; you can redistribute it and/or modify it under the same terms as the Perl 5 programming language system itself.

展望

优化方向

分析内容

1.位点筛选

- 目前只考虑了有rs编号的位点，后续考虑其他位点也整合进来(正在进行)
- 种族信息是否做为位点筛选的一个标准
- PGx是否也做为药物筛选的标准

2.分析判定

- 利用连锁信息，对更多位点的分型进行补全
- 优选等级判定是否将证据等级考虑进去

3.内容新增

- Dosing Guidelines
- 与疾病和表型（健康风险和遗传疾病）关联起来

4.展示内容

- 新增一个汇总的列表
- 考虑将药物说明书的某些内容（适应症，不良反应，用法用量，作用机理）加进去（正在进行药物机理内容的搜集）
- 医保类型

5.报告版本升级

6.流程结构与速度

- rust改写
- 数据库结构优化，拆分为单个药物

目前分析局限性

- 未能覆盖到药物在数据库中的所有重要位点（目前正在单倍体型信息提取以及分析流程）
- 种族差异没有考虑

相关文献

- <https://www.ncbi.nlm.nih.gov/pubmed/29040422>
- <https://www.ncbi.nlm.nih.gov/pubmed/29026329>
- <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5477417/>
- <https://www.nature.com/articles/s41598-017-08468-y>
- <https://www.ncbi.nlm.nih.gov/pubmed/23376192>

祖源分析

相关公司

国外

- 23andMe : <https://www.23andme.com/>
- Ancestry: <https://www.ancestry.com/>
- Family Tree DNA: <https://www.familytreedna.com>
- MyHeritage: <https://www.myheritage.com> (有中国区服务)

国内

- WEGENE
- 23魔方

位点使用情况

	23andMe	乐土
民族成分	Autosomal DNA + X-DNA(956,000)	Autosomal DNA (3191)
父系祖源	2,000	8
母系祖源	2,500	120

主要内容

- 民族成分
- 母系祖源
- 父系祖源

民族成分

参考资料: <https://www.23andme.com/ancestry-composition-guide/>

数据来源

数据库

	23andMe	乐土
HGDP-CEPH	√	√
1000Genomes	√	√
HapMap	√	✗
iControlDB	√	✗
PGG	✗	✗
others	√	✗

- PGG.Population: 包含来自107个国家356个民族的7122个人
- HapMap : 2016-06-16 宣布有安全漏洞, NCBI停止维护, 1000Genomes提供更多的信息
- iControlDB是Illumina的一个样本数据库

训练集规模

	民族	人数
23andMe	31	10,418
乐土	57	1,730

数据过滤

	23andMe	乐土
样本	1.挑选有4个祖父母在同一个地区（除美国、加拿大、澳大利亚等移民国家）的人; 2.去除亲缘关系较近的人; 3.去除统计结果和填写的民族不一致的人	暂无
民族	1.挑选500年前就已存在的民族; 2.历史上没有移民	人工判断, 关系较近或移民混合去除
偏好	主要为欧洲人	亚洲人

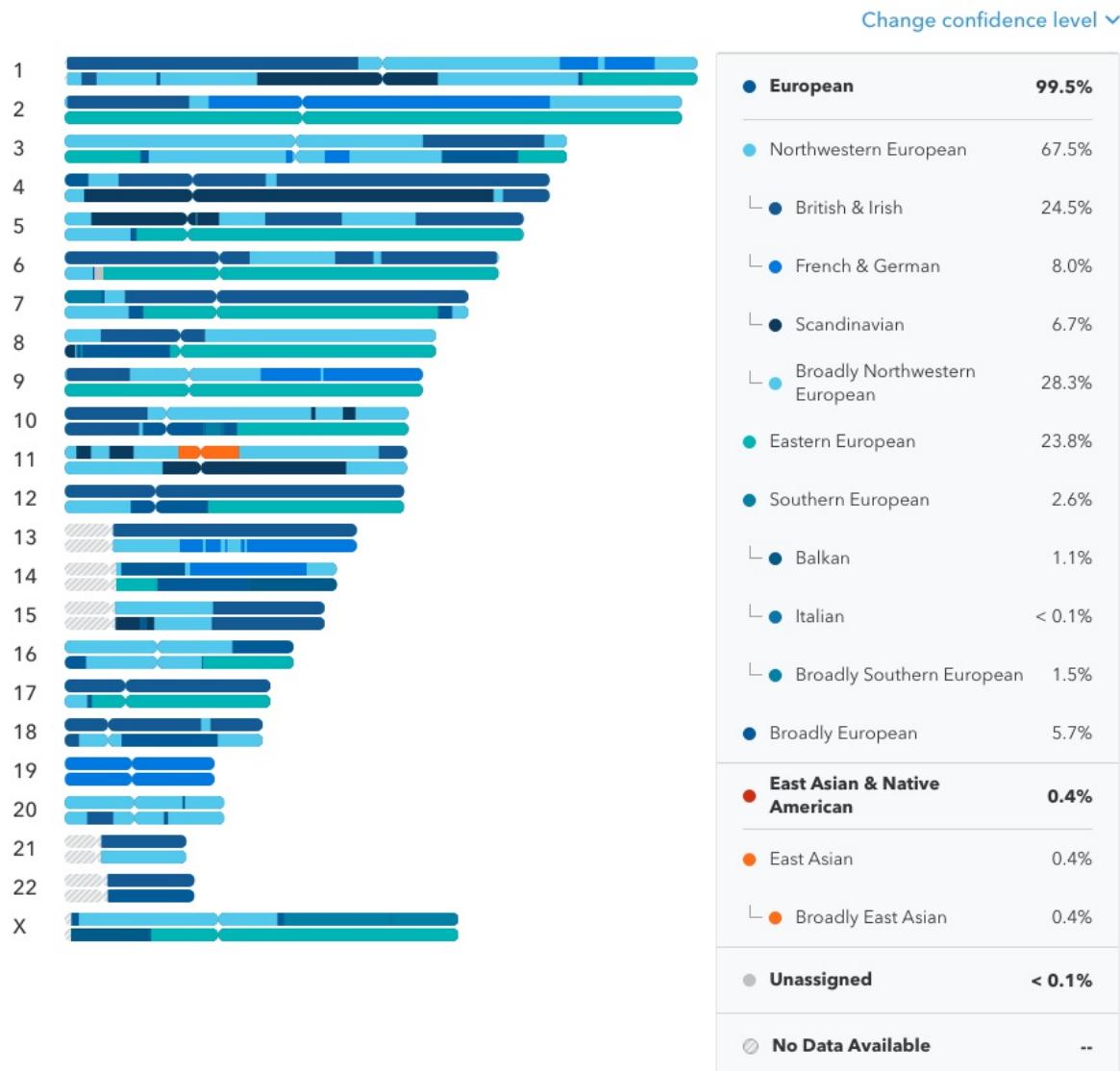
分析步骤

乐土

1. 挑选民族, 构建训练集;
2. 根据位点频率, 挑选差异显著的位点用于下一步的分析;
3. 随机森林确定各个民族的成分;

23andMe

1. Phasing: Finch (BEAGLE的23andMe版) 确定哪些marker组成一个染色体
2. Window Classification: 100个Marker做一个窗口, SVM判断每个窗口的民族来源;
3. Smooth: Hidden Markov Models进行拟合 (unusual assignment; switch error)
4. Re-calibration: 纠正训练集所造成的偏好
5. Aggregation & Reporting: 各个窗口结果加和



Using Close Family Members, you will get a very high-quality chromosome phasing result

Ancestry Timeline (2017 23andMe)

<https://customercare.23andme.com/hc/en-us/articles/115004342967>



假设前提： each ancestry was inherited from a single ancestor. It's not true for highly admixed populations. 结果

分析原理： The Ancestry Timeline feature analyzes the pattern of ancestry in your genome by looking at both the number and size of segments that came from a particular ancestry as well as their distribution across your chromosomes.

颇受争议：

1. <https://dna-explained.com/2017/01/17/calling-hogwash-on-23andmes-ancestry-timeline/>
2. <http://www.rootsandrecombinantdna.com/2017/01/new-23andme-ancestry-timeline-tool.html>

单倍群分析

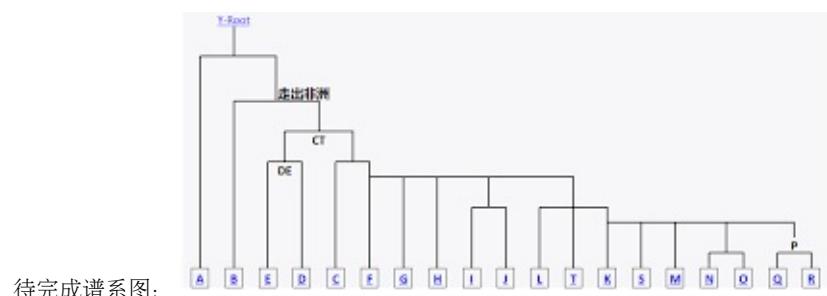
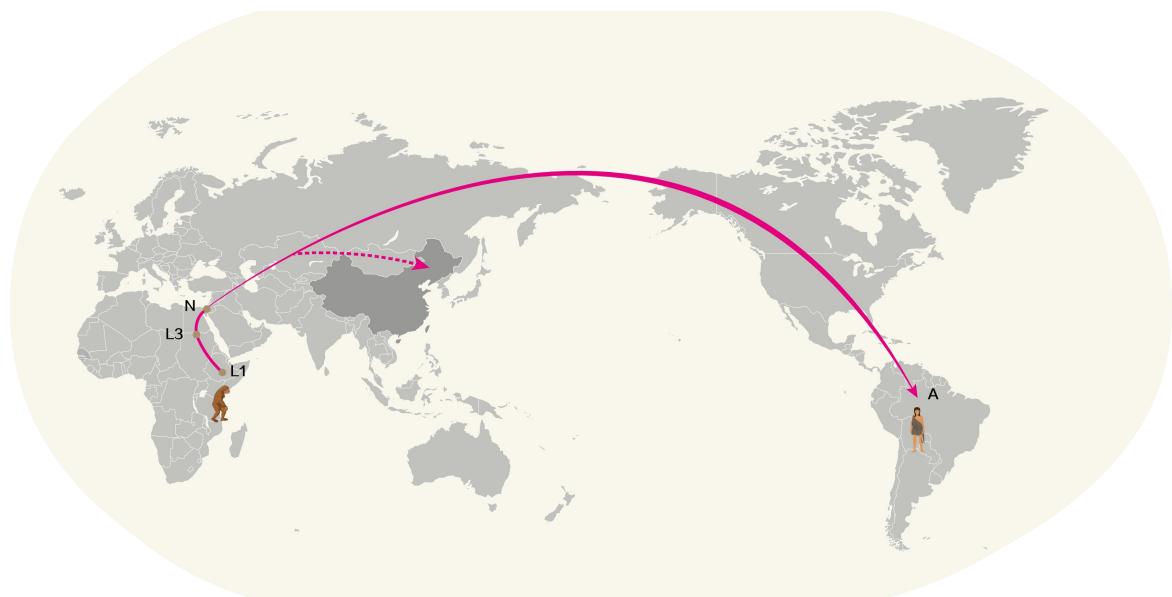
在分子进化的研究中,单倍群或单倍型类群是一组类似的单倍型,它们有一个共同的单核苷酸多态性祖先。可以用于反映祖先的迁移路径。

数据库来源

- Y-DNA Haplogroup: <https://isogg.org/tree/>
- MT-DNA Haplogroup: <http://www.phylotree.org/index.htm>

结果

MT单倍群分析图展示:



分析流程

Requirements

- perl
 - Modern::Perl
 - Data::Dumper
 - IO::All
 - Cwd
 - FindBin
 - Getopt::Long::Descriptive
 - File::Basename
 - JSON;
 - MCE::Loop
 - DBI
- python
 - numpy
 - sklearn

主流程

```
perl ancestry_analysis_csv.pl

Usage:
Ancestry Analysis [-ciov] [long options...] <some-arg>
  -c STR --csv STR    芯片样本的csv文件
  -i STR --step STR   步骤参数:
    0: 结果分析
    1: 分析结果更新至数据库
  -o STR --outdir STR  结果输出目录
```

Y单倍群

```
perl bin/generate_Y_group.pl

Usage:
Y haplogroup [-bjov] [long options...] <some-arg>
  -j STR --json STR  样本分型的json文件
  -o STR --out STR   Y单倍体群结果

  -b STR --base STR  Y tree位点基因型信息
  -t STR --tree STR  Y tree结构的json文件
```

MT单倍群

```
perl bin/generate_MT_group.pl

Usage:
```

```
MT haplogroup [-bjotv] [long options...] <some-arg>
  -j STR --json STR 样本分型的json文件
  -o STR --out STR MT单倍体群结果
  -b STR --base STR MT tree位点基因型信息
  -t STR --tree STR MT tree结构的json文件
```

民族成分

```
python bin/random_forest.py <train.input> <population.txt> <test.input> <test.region.out>

Usage:
  train.input 训练集样本基因型结果
  population.txt 训练集样本的民族信息
  test.input 样本的基因型结果
  test.region.out 样本的民族成分分析结果
```

分析结果上传数据库

```
perl bin/create_ancestry_db.pl

Usage:
Create or Update Ancestry Database [-abcmnrsrvYy] [long options...] <some-arg>
  -r STR --region STR 民族成分结果
  -y STR -Y STR      Y单倍体群结果
  -m STR --MT STR    MT单倍体群结果
  -s INT --step INT   步骤说明:
    0: 数据库初始化
    1: 分析结果上传数据库
```

结果目录

```
├── *.ancestry.csv          #样本在祖源分析位点的基因型信息
├── A01_10003836.ancestry.json #样本在祖源分析位点的基因型信息的json文件
├── A01_10003836.region.out #样本的民族成分结果
├── A01_10003836.MT.out     #样本的MT单倍体群结果
├── A01_10003836.Y.out      #样本的Y单倍体群结果
└── RF                       #民族成分分析的中间文件
    ├── A01_10003836.test.input
    └── A01_10003836.train.input
```

展望

民族成分

- 考虑更多的数据库如Pgg等，来增加训练集的样品数；
- 增加民族和样本的过滤条件；
- BEAGLE来对进行phased；
- 基于划窗法的算法修改，为后续的Chromosome_Painting和Ancestry Timeline分析铺垫；
- Ancestry Timeline流程搭建。

单倍群

- 提供一个单倍群分析中使用的位点列表，方便报告审核；
- 建立单倍群的进化图；

GSA芯片

- 基于GSA芯片数据的流程修改

数据库

- 祖源结果数据库存储结构调整

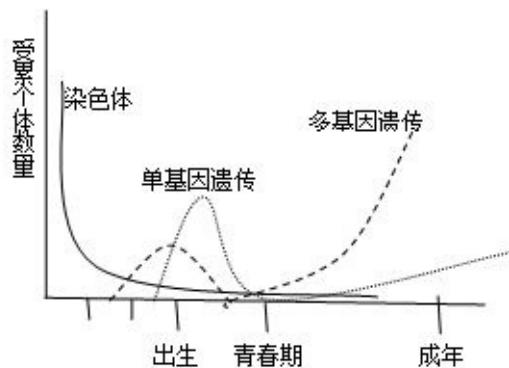
报告

- 新版本报告实现

遗传疾病

概念

单基因病：是指由一对等位基因突变引起的疾病，符合孟德尔遗传方式，也称孟德尔遗传病。目前乐土9800展示的主要为单基因病，发病主要在青春期之前，新生儿的发病率约1/100（WHO）。



分类

- 常染色体显性遗传 (AD) : eg. 软骨发育不全
- 常染色体隐性遗传 (AR) : eg. 白化病
- X连锁显性遗传 (XD) : eg. 抗维生素D佝偻病
- X连锁隐性遗传 (XR) : eg. 法布雷病
- Y连锁遗传: eg. 外耳道多毛症

内容

- 目前包含约1000种疾病及亚型
- 覆盖15个系统（免疫系统、内分泌系统、呼吸系统、心脑血管系统、泌尿系统、消化系统、生殖系统、皮肤系统、眼耳鼻喉、神经系统、肌肉系统、肿瘤、血液系统、遗传代谢、骨骼系统）

检测意义

- 遗传疾病筛查，了解自己携带的隐性致病基因
- 对尚未有表型的迟发型遗传疾病尽早采取措施

相关数据库

- OMIM
- ClinVar

流程

位点过滤

1. 对于本批次杂合率 > 0.05 的位点进行删除
2. 对芯片参考数据中 Minor Allele Frequency > 0.01 进行过滤
3. 疾病的 ClinVar ClinicalSignificance 不是 Pathogenic, Likely pathogenic, PATHOGENIC

结果判断

统计基因型匹配 Alt 的次数 N

常染色体显性

- 可能致病: ≥ 1
- 正常

常染色体隐性

- 可能致病: ≥ 2
- 携带: ≥ 1
- 正常

X连锁显性

- 可能致病: ≥ 1
- 正常

X连锁隐性

- 男
 - 可能致病: ≥ 1
 - 正常
- 女
 - 可能致病: ≥ 2
 - 携带: ≥ 1
 - 正常

线粒体

- 可能致病: ≥ 1
- 正常

尚不明确

- 未知: > 1

展望

数据库

目前进展

1. 下载基因每个位点的遗传信息，需爬虫；
2. 通过ClinVar的rs信息，确定基因位点对应的表型MIM；
3. 对表型进行过滤
 - phenotype prefix, 保留‘Number Sign’；
 - phenotype mapping key, 保留‘3’；
 - phenotype inheritance, 过滤掉‘digenic recessive’, ‘isolated cases’, ‘isolated cases, somatic mutation’, ‘somatic mosaicism’, ‘somatic mutation’, 以及包含‘multifactorial’的条目；此列‘空白’的条目保留；
 - Phenotype full name, 过滤掉包含‘{}’‘[]’‘?’的条目；
4. 位点与目前芯片位点取交集；
5. clinical significance过滤，只保留包括‘pathogenic’、‘likely pathogenic’和‘空白’
6. 删除phenotype包含‘somatic’的条目
7. variant phenotype MIM空白的进行人工补充和修正(在clinVar中未找到)
8. 过滤表型没有遗传模式

局限性

- 数据库整理时对位点跟芯片取了交集，当换GSA或是分型补全得到更多位点时就会有很多信息遗漏
- 建议按照疾病来进行梳理

分析流程

- 重新搭建单基因病数据库
- 考虑位点个数以及位点遗传模式，重写流程

展示

- 总体结果图

报告

- 新版报告实现

结果数据库

分析结果上传以及更新

数据上传

- 导入数据库

```
mongoimport -h 192.168.1.205 --port 27019 -d Annotation -c probesets --type=csv --headerline {}.csv
```

- 数据库样品信息删除

```
mongo mongodb://192.168.1.205:27019/Annotation <<< 'db.probesets.deleteMany({ "Sample ID": "A02_10002781" })'
```

系统更新

```
cd SNParray/SNParray  
node bin/report-upgrade.js -i A01 -S 123:1 -S 1234:1 -E 123
```

参数解析:

- -i 样品的芯片编号
- -S 风险修改（针对健康风险），高风险2，中风险1，低风险0
- -E 表型删除

报告生成

```
yarn start  
yarn global add serve  
serve -s build
```

展望

- 数据库结构调整
- 定期备份的docker实现

其他相关项目

- 基因身份证
- 分型补全

基因身份证

利用成熟的DNA指纹（DNA分型）技术，对若干个固定的基因位点进行鉴定，以进行个体鉴定、区分；

技术原理

DNA多态性，主要包括3大类，即片段长度多态性（限制性片段长度多态性，RFLP）、重复序列多态性（短串联重复序列，STR）、单核苷酸多态性（SNP）。

意义：基于个体鉴定、亲缘关系鉴定，体现在财产继承、试管婴儿、骨髓移植、意外事故、失散等。

相关概念

哈德温伯格平衡

一个群体在理想情况下（不受外界情况影响，如非随机交配、选择、迁移、突变或群体大小有限），经过多个时代，基因频率与基因型频率保持很定并处于稳定的平衡状态。

$$p^2 + 2pq + q^2 = 1$$

\$\$

主要用于描述群体中等位基因频率和基因型频率之间的关系。

FST

观测数据与期望杂合子比例偏离的两个主要因素，即非随机配对和群体结构。通常使用Fst来量化群体结构。

$$\text{亚群间的期望杂合度: } H_S = \frac{1}{n} \sum_{i=1}^n 2p_i q_i$$

$$\text{总体的期望杂合度: } H_T = 2p_i q_i$$

$$FST = \frac{H_T - H_S}{H_T}$$

\$\$

实际计算中fst要考虑很多参数的影响，如等位基因2个以上，多个位点Fst取平均，不用亚群样品数不一样，基因型缺失等

参考文献：<http://onlinelibrary.wiley.com/doi/10.1111/j.1558-5646.1984.tb05657.x/full>

常用软件：Haplovew、SNPstat

连锁计算

遗传距离指两个基因在同一染色体上的相对距离，通常用重组率表示。

其数值以重组率的数值去掉%表示，单位cM。

$$\text{重组率 } r = \frac{\text{重组型配子数}}{\text{总配子数}} \times 100\%$$

\$\$

通常认为 $r < 50\%$ 连锁， $r = 50\%$ 不连锁

已知每种分型子代的数量，可用最大似然法计算重组率

$$LOD = \log_{10} \frac{\text{发生连锁的概率}}{\text{不发生连锁的概率}}$$

\$\$

一般认为 LOD>3 可以判定连锁，LOD<2可以确定不连锁

常用软件：VCF-tools

位点筛选和过滤

1. 杂合率Het-ratio > 0.45
2. Fst < 0.01
3. 符合哈德温伯格定律
4. 不连锁
5. 相互距离 > 1MB
6. 不在重复区

分析步骤

1. 按照上述标准用1000Genomes对90万个芯片位点进行过滤，得到171个位点；
2. 对实际的芯片数据（去除家系和重复检测样本）进行样本和位点的检出率统计，以及位点HWE平衡的检验；
3. 对HWE(\$\$p < 0.05\$\$)和检出率（\$\$CallRate < 98\%\$\$）的位点进行过滤，得到141个位点；
4. 样本数据进行两两比较，不考虑缺失统计共同位点数；
5. 对位点进一步优化精简，确定区分两个样本的最少位点数；

参考文献

- Developing a SNP panel for forensic identification of individuals
- Forensic validation of the SNPforID 52-plex assay

分型补全

方案设计

1. 比较BEAGLE、GeneImp软件（后续可测试Impute2、Minimac3等）对当前芯片数据的适用性与准确度评估；
2. 搭建芯片数据的补全流程；

GeneImp

dependencies

- gfortran
- curl-config
- xml2-config
- libmysqlclient

```
sudo apt-get install gfortran libcurl4-gnutls-dev libxml2-dev libmysqlclient-dev
```

数据要求

- 参考集样本数大于200
- 二等位
- vcf文件必须包含PL列（芯片数据转化时参考confidence信息）
- vcf必须用bgzip压缩，用tabix创建index

运行命令

```
docker run --name imputation --mount type=bind,source=/home/sumin/workdir/GenotypeImputation/data,target=/data
imputation Rscript /data/test.R
```

参数测试

	准确率	时间
kl=10	94.89%	224min
kl=20(default)	93.68%	59min
kl=50	90.22%	35min

	准确率	时间
haps=20	92.59%	28min
haps=100	93.57%	41min
haps=200(default)	93.68%	59min
haps=500	93.71%	122min
haps=1000	93.72%	260min

	准确率	时间
flank=0.1	94.64%	124min
flank=0.5(default)	93.73%	59min
flank=0.8	%	min

BEAGLE

数据要求

- ID列不为空
- vcf已排好序
- 1Mb内有marker

运行命令

```
java -jar bref.08Jun17.d8b.jar ref.vcf.gz
java -jar beagle.08Jun17.d8b.jar ref=ref.bref gt=target.vcf.gz out=prefix
```

参数测试

	准确率	时间
markers=10M	90.11%	18min
markers=50M(default)	90.40%	22min
markers=100M	90.36%	25min

	准确率	时间
iteration=5(default)	90.40%	22min
iteration=20	90.42%	47min
iteration=100	90.38%	180min

	准确率	时间
err=0.00001	90.11%	18min
err=0.0001(default)	90.40%	22min
err=0.001	90.44%	22min
err=0.01	90.21%	22min

结果比较

	GeneImp	BEAGLE
运行时间	59min	28min
准确率	93%	90%

资源消耗	400G	8G
------	------	----