聚普体侧信息一体化软件 V1.0

（使用说明书）

目录

1. 引言
   1. 编写目的
   2. 背景
   3. 参考资料
2. 用途
   1. 功能
   2. 性能
3. 运行环境
   1. 硬件设备
   2. 支持软件
4. 使用过程
   1. kmash功能的使用说明
   2. GeneScan功能的使用说明

**1引言**

**1.1编写目的**

本说明书为指导分析潜在的基因突变或者基因激活靶点，提供了扫描全基因组SNP数据的一般方法，用于基因集关联分析，基因集的检验统计量基于广义线性模型的得分统计，并利用基因本体论的有向无环图结构来创建基因集。获得能对临床用药和预后有指导意义的运算程序和模块。

**1.2背景**

这一系统由昆山聚普生物信息技术有限公司开发，使用者是广州中山医学院的有关分析人员。

**1.3参考资料**

《临床检验掌中宝（第2 版）》广东科技出版社

陈铭，包家立，（2013）《生物信息学》（第二版）科学出版社

林标杨 ，（2012）《系统生物学》浙江大学出版社

**2用途**

**2.1功能**

聚普体侧信息一体化软件的功能和特点有，一，构建了一个单任务并行计算框架匹配已知的关键基因，并把一些已知的身体检测信息分析程序，比如BLAST放入此计算框架，减少运算所需时间。二，对匹配到的基因，找出已经发现的关联和功能，列出这些功能的描述论文，和常用的模型，产生模型的预测结果。比如对于基因GZMB，就会将它关联到肺癌，并且引用一篇发表于《自然》杂志上的论文，基于这个基因以及其他一些因子对于治疗作出指引的一个计算模型。三，建立了一个标准化的基因及其他生理信息预测模型实现模板，此模板定义了模型的输入和输出数据。输入数据包括基因数据，常规医学检验数据，等类别。输出数据包括疾病预测风险，建议治疗方案，等类别。当有新的研究成果发表时，工作人员可以把研究成果根据标准，建立新模型，并添加到数据库中。

**3运行环境**

**3.1硬件设备**

PC机要求 ：x86通用计算机系统，4核cpu，8G内存，1T硬盘

打印机 ： Windows支持的打印机

**3.2支持软件**

windows 10 或以上操作系统

开发语言 ：c#，python

**4使用过程**

**4.1 kmash功能的使用说明**

* **kmash的功能：**

在本软件当中，kmash命令可以用来计算 DNA/RNA 或蛋白质序列的MinHash特征值。

* **kmash的运行环境：**

kmash命令需要在Python 2.7.x和Python 3.5+ 下运行。基本要求是 screed 和ijson，以及 C++ 开发环境和CPython开发头文件和库(用于C ++扩展)。

kmash的compare 选项使用numpy，绘图代码使用matplotlib和scipy，但大多数代码在没有这些代码的情况下可以使用。对于 search 选项和 gather 选项，还需khmer版本2.1以上。

* **kmash命令的功能和子功，kmash总共有5个主要的子命令：**

1. compute用来计算特征值
2. compare用来比较两个特征值并且建立距离矩阵
3. plot用于把距离矩阵转成图
4. search用于查询目标特征值
5. gather用于找出非重叠匹配到签名集合中的宏基因组

* **kmash compute**

compute子命令计算并保存一个或多个序列文件中每个序列的签名。它将FASTA或FASTQ文件作为输入，这些文件可以使用gzip或bzip2解压缩或压缩。输出将是一个或多个可与kmash compare子命令一起使用的JSON签名文件

用法：

kmash compute filename [ filename2 ... ]

选项：

--ksizes K1 [，K2，K3]

使用一个或多个k-mer大小;默认值是31。

--force

重新计算现有签名;将非DNA字符转换为N。

--output

将所有签名保存到此文件中。

--track-abundance

计算并保存k-mer丰度。

--name-from-first

根据文件的第一个序列命名签名。

--singleton

不是为每个输入文件计算单个签名，而是为每个序列计算一个。

--merged <name>

为所有输入文件计算单个签名，并将其命名为<name>。

* **kmash compare**

compare 子命令使用估算的 Jaccard 索引比较一个或多个签名文件。默认输出是相似性矩阵的文本显示，其中每个条目[i，j]包含输入签名i和输入签名j之间的估计的 Jaccard 索引。可以使用 --output 将输出矩阵保存到文件中，并可以和kmash plot 子命令一起使用（或使用numpy.load（...）加载。使用 --csv 将输出可以加载到其他语言的CSV文件Python，比如R.

用法：

kmash compare file1.sig [ file2.sig ... ]

选项：

--output

将距离矩阵保存到此文件中（作为numpy二进制矩阵）

--ksize

以此 k-mer大小进行比较。

* **kmashplot**  
  plot子命令从kmash compare – output中计算的距离矩阵中生成两个图 - 树形图和树形图+矩阵。默认输出是两个PNG文件。

用法：

kmash plot <matrix>

选项：

--pdf

输出PDF文件。

--labels

在绘图上显示签名名称（默认情况下为文件名） - 指示 - 关闭绘图上的索引显示。

--vmax

热图的最大值（默认值为1.0）。

--vmin

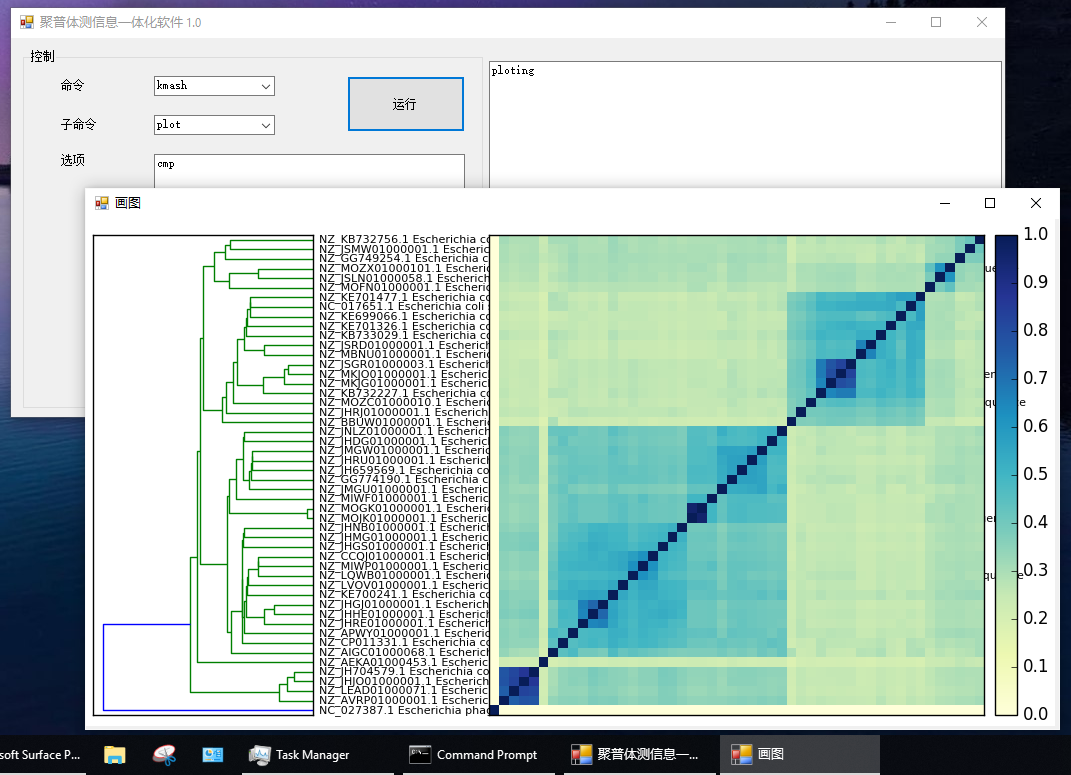
热图的最小值（默认值为0.0）。

--subsample = <N>

绘制随机选择的最大<N>个样本。

--subsample-seed = <seed>

伪随机数生成器的种子。



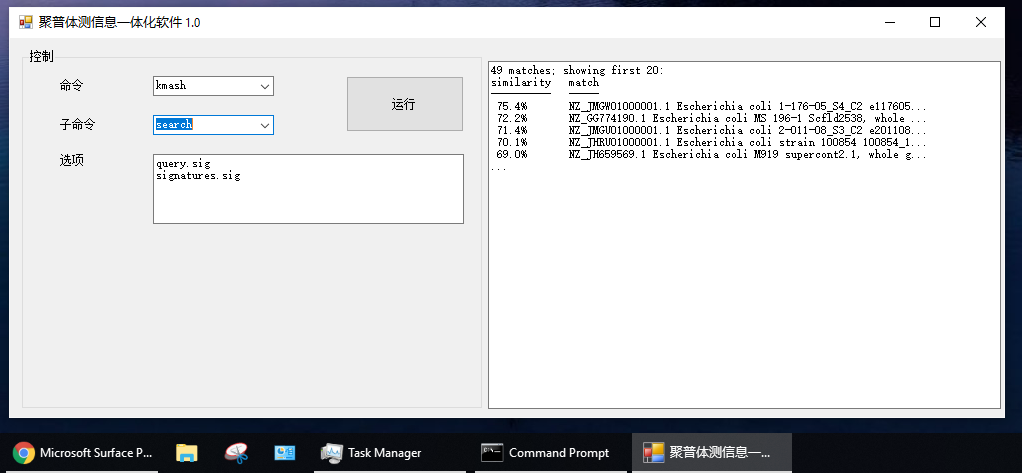
* **kmash search**

search子命令搜索一系列特征值以查找与查询特征值匹配的结果。它可以搜索具有高Jaccard相似性或包含的匹配。本命令默认使用Jaccard相似性。当指定了—containment选项时，则只考虑包含。 -o 选项将把结果输出到CSV文件。

search子命令会将所有提供的特征值加载到内存中，对于大型集合来说，这可能会很慢并且使用非常多的内存。您可以使用kmashindex子命令创建可以在磁盘上快速搜索的序列绽放树（SBT）。这与我们提供GenBank和其他数据库的格式相同。

用法：

kmash search query.sig [签名或SBT列表]



* **kmash gather**

gather子命令查找查询的所有不重叠的匹配。这特别适用于宏基因组和基因组区分析。（有关可在此处使用的不同方法的更多信息，请参阅签名分类。）

如果使用--track-abundance选项计算输入签名，则输出将被丰富加权（除非指定了--ignore-abundances）。用-o/--output选项输出将创建包含匹配项的CSV文件。

gather子命令和 search子命令一样会将所有提供的特征值加载到内存中。您可以使用kmash index创建可以在磁盘上快速搜索的序列绽放树（SBT）。这与我们提供GenBank和其他数据库的格式相同。

用法：

kmash gather query.sig [签名或SBT列表]

注意：

使用kmash gather将宏基因组分类为没有（或不完整）分类信息的基因组集合。使用kmashlca总结和kmashlca聚集来使用具有分类信息的基因组集合对宏基因组进行分类。

-k

选项控制 k-mer的数值。用户可以一次性构建各种k-mer大小的签名，除非用户使用的是非常大的宏基因组，生成的签名文件仍然可以保持很小的文件大小。k 值我们建议为31或者51，这是基于我们阅读的文章，如Metapalette论文（Koslicki和Falush在2016年发表的论文。文中指出可以使用k=49或k=53，并且得出和k=51十分类似的结果）一般来讲，较大的 k-mer值不容易出现误报。以上k值都选择为奇数。这是为了避免数值计算中的回文偶尔出现轻微的问题，其中k-mer的正向和反向互补是相同的。如果k是奇数，则不会发生这种情况。

--n

选项可以用来指定最多计算多少个哈希值

--scaled

选项可以用来指定压缩比例

我们建议使用 --scaled=1000计算所有签名。这将获得1000比1的压缩比，同时可以检测10kb范围内的相似区域。为了与更传统的MinHash方法（如mash）进行比较，如果有一个5 Mbp的基因组并使用 --scaled=1000，你将提取大约5000个哈希值。因此，缩放1000相当于在5 Mbp基因组上使用-n=5000和mash。

使用-n和--scaled之间的区别在于宏基因组分析：使用-n修复哈希数量会限制您检测稀有生物的能力，或者导致非常大的签名（例如，如果您使用大于10000的n）。 - 尺度将通过宏基因组的多样性来扩展您的分辨率。

kmash可以广泛地用于 Illumina 读数据集和组装的基因组，转录组和宏基因组。 PacBio和Nanopore测序的高错误率对于基于k聚体的方法是有问题的，并且我们还没有探索如何调整这种测序的参数。

在更实际的说明中，kmash计算应该自动检测FASTA，FASTQ，无论它们是未压缩，gzip压缩还是bzip2-ed。用户都可以直接调用kmash命令进行分析计算。

-search

search 选项是用来寻找与其他特征值的高度相似性匹配的特征值。这是这是MinHash搜索的最基本使用方式。该命令采用查询签名和一个或多个搜索签名，并查找它可以超过特定阈值的所有匹配。

默认情况下，搜索将找到具有高Jaccard相似性的匹配，这将考虑两个样本的并集中的所有k-mers。实际上，这意味着如果样本之间存在高重叠且样本之间不相交的k-mers相对较少，则只会找到匹配项。这对于发现相似但很少适用于大小不同的样品的基因组或转录组是有效的。

搜索的一个有用的修改是用--containment选项来更改相似度的计算方法，这将找到主题中包含查询的匹配项，但主题可能包含许多其他k-mers。例如，如果您使用质粒作为查询，用户可以使用--containment

查找包含该质粒的基因组。

**4.2 GeneScan功能的使用说明**

* **GeneScan的功能简述：**

在本软件当中，GeneScan这个功能项将读取一些必须的输入文件和一些手动输入的参数，然后输出相应的基因计算结果。

* **GeneScan的基本概念**

GeneScan提供了扫描全基因组SNP数据的一般方法，用于基因集关联分析，基因集的检验统计量基于广义线性模型的得分统计，并利用基因本体论的有向无环图结构来创建基因集。该方法可以使用其他基因组结构，例如kegg基因（KEGG），或用户定义的组。 我们的方法将SNP结合到基因中，将基因结合到基因组中，但确保对特性的正面和负面影响不会消除。为了控制许多基因组的多重测试，我们使用有效的计算策略来解释基因和基因组之间的连锁不平衡和相关性，并为每个基因组提供准确的降低调整的p值。

* **GeneScan的运行状态**

程序从start到运行时，需要系统内存作为系统内存我们将“磁盘空间”定义为在系统上存储任何文件所需的硬盘空间。我们在各种数据集上测试了GeneScan，每个数据集都在550K SNP芯片上进行基因分型，从1000到2500科目不等。完整SNP得分所需的磁盘空间.在这些研究中可以超过10GB，运行程序所需的内存可以超过12GB。因为机器语言命令用于在32位机器上编译的程序不能索引超过4GB的内存，GeneScan应该只能在64位机器上运行在运行时可用的内存。

* **软件中自带提供**

gene\_snp.B37.coding.dat：

参考文件将dbSNP132 SNP映射到GRCh通过它们的rsid构建37个位置和所有编码基因的起始/终止位置。

edges.csv：

定义GO有向无环图边的文件。

gene2go.human：

映射到GO术语的基因列表。

hsa\_pathway.list：

映射到KEGG人类途径的基因列表。

GeneScan通过rsid将SNP映射到基因集所需的所有文件。我们提供以下一组映射参考文件：

edges.csv

定义GO有向无环图边。

gene2go.human

映射到GO术语的基因列表。

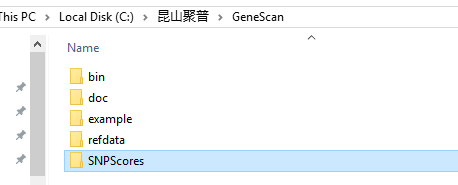
pathway.list

映射到KEGG人类途径的基因列表。

snp.B37.coding.dat

使用来自HapMap参考的SNP的rsid将SNP定位到基因面板更新为NCBI构d的37个位置，映射到Build 37基因开始/停止位置。

* **文件夹结构**



该软件包包含预编译的可执行文件和参考数据文件。 用户最重要的一步是创建一个SNP分数，我们提供脚本和文档，以指导用户完成该过程。该软件具有以下目录结构，以斜体显示子目录的简要说明：

bin\binary

可执行文件 用于将SNP映射到基因组的。

refdata\reference，

输入文件夹。

doc\manual

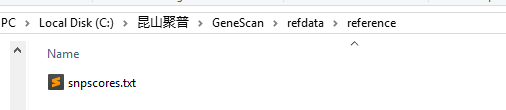
手册 。

example\

示例参数文件和运行程序的示例脚本 用于创建SNP分数。SNPScores\script

运行脚本。

* **输入文件**



用户创建： snpscores.txt：并放在refdata\reference文件夹里。用户创建的SNP评分，测试具有特征的SNP的残差，有或没有调整协变量。提供Shell和R脚本以从plink格式的数据创建这些文件，为每个染色体运行单独的进程，然后将分数连接成一个大文件。

GeneScan程序需要输入SNP得分文件（见下列格式）。这个文件包含了每个受试者和每个SNP到评分SNP与特征的关联的评分统计量。文件里的行用于SNP，列用于主题。第一列是SNP编号，接着是从模型t，Si得到每个SNP每个受试者得分的列; j =（Yi;j Yj）（Gi;j Gi)，Yj是来自逻辑模型的主题j的tted值，其可以包括或不包括

调整协变量。Gi; j是受试者j的SNP i的SNP剂量，通常是次要等位基因计数。G i;：是所有受试者的平均SNP剂量。

SNP得分的一般格式如下：

rs100 0.152 -0.330 0.104 -0.225 ...

rs105 -999.0 0.951 -0.312 0.152 ...

rs222 0.443 0.051 -0.875 -0.342 ...

...

* **辅助生成输入文件的脚本**

GeneScan将丢失的值视为缺失;将缺失值设置为-999.0将在大多数情况下工作。我们提供了我们提供的脚本模式，用于创建SNP-Scores，以及脚本本身如下所述。

用户可以自由使用任何方式来创建SNP分数文件，但我们提供了一些通用脚本。使用R和一些shell脚本来计算染色体的SNP分数。

我们提供的脚本设置为在R中运行逻辑回归以获得二进制大小写控制状态。在SNPScores\script文件夹里有：

unadjustedScores.R一个R脚本来计算单个染色体的SNP分数，而不是调整协变量。该脚本假定可以通过系统命令调用R。

reslectedScores.R一个R脚本，用于计算单个染色体的SNP分数，调整为协变量。该脚本假定可以通过系统命令调用R，并且需要

对模型中要调整的协变量数量进行一些编辑。

startScores.sh一个示例shell脚本，用于在单独的R批处理作业中创建SNP分数。

catScores.sh一个示例shell脚本，用于连接每条染色体的SNP分数进入一个大的SNP得分。个体染色体SNP得分。

* **GeneScan的使用参数**

可以使用程序帮助菜单查看程序选项，如下所示：每个选项都有一个短名称前面有一个或者两个短划线（- /--）。以下是完整的选项列表，列出了默认和简要说明。

-ge goEdges.dat，

GO边缘文件

-gs geneSNP.dat，

映射文件

-gg goGene.dat，

go-gene mapping文件

-kg keggGene.dat，

kegg基因映射文件

-sc score.dat，

分数文件

-r zero，

去掉零，替换缺失的分数

-mv -999，

分数缺失值代码

-m GO，

方法GO。可选择（GENE，GO，KEGG，SNP）

-gl sqrtmean，

基因水平评分（max，mean，sqrtmean）。注意max，mean是实验性 的

-gm nor，

基因水平时刻

-gsl wtdmean，

基因集水平评分（maxZ，meanZ，wtdmean）。注意：maxZ，meanZ是实验性的

-gt all，

GO DAG的类型（全部，biological\_process，cellular\_component，molecular\_function）

-d 50000，

从SNP到基因开始/停止的最大绝对距离，以bp距离表示

-ms 30，

最大数量一组中的基因。注意计算时间随max-size而增加

-ns 1000，

数字的数量

-s -9，

随机数种子的整数（如果<= 0或未指定，则由时钟设置）

-mf output.map，

输出映射文件名

--missval

选项告诉GeneScan如何处理缺少输入分数，

-replace

告诉如何替换缺失值，即在该SNP处为零或平均分。

--nsim

选项告诉GeneScan要执行多少个模拟，

--map-write

选项告诉GeneScan编写可选映射。提供有关SNP和基因如何在分析中映射到基因组的信息。

--dist

选项指定基因开始/停止位置外多远以允许SNP映射到 基因。默认设置为50,000个碱基对，GeneScan允许--dist为任意数字 在0到50,000之间。

--gene-snp

对于基因级测试，我们将方法更改为GENE，更改输出的名称，然后添加，有三个选择。为基因水平分析添加的第一个选项是--gene-snp，它告诉GeneScan在哪里和SNP中的基因映射。其他添加的选项告诉程序如何评分每个基因。对于每个基因的逻辑模型评分，我们将--gene-level选项设置为sqrtmean为SNP分数平均值的平方根与一个基因相结合，为此我们估计假设多元正态分布的基因分布.

--method

对于基因集，我们保留了基因分析设置中的大部分选项更新输出名称。我们需要将--method改为GO并指定哪个GO命名空间使用。然后我们添加选项来指定另外两个引用的名称，给GeneScan一个最大集合大小，并且给出基因集评分基因的说明。

--go-type

选项告诉我们使用三个名称空间中的哪一个：生物过程，分子功能和细胞成分：

biological\_process, 生物过程

cellular\_component, 分子功能

molecular\_function 细胞成分

--go-gene和--go-edges

除了上面基因分析中需要的Gene-SNP之外，我们还要告GeneScan

两个GO数据。第一个是--go-gene将Entrez Gene ID链接到GO术语，我们通过子集化来创建这些术语。 第二个是我们预处理的--go-edges

GO节点和连接它们的边。

--max-size

用户可以将分析限制为不超过max-size指定的基因数量。对于所有的基因集，我们建议--max-size为100。

--gene-set-level wtdmean

给出gene-spec得分，其中权重是基因评分的方差。我们从基因分析中推进基因水平评分，但我们还需要告诉GeneScan如何在基因集内对基因进行评分。我们实施了加权平均值。

-kegg-gene

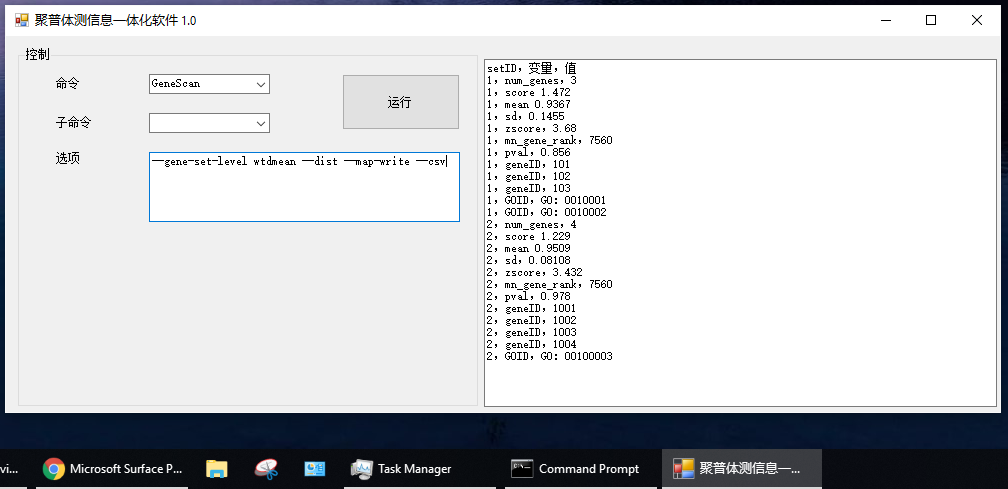
为了分析由kegg基因定义的基因组，我们保持大多数相同的选项如上所述，更新输出名称以反映分析方法。方法选项更改KEGG，我们只需要指定一个对于通过-kegg-gene选项映射到KEGG途径的基因，用于调用并将--max-size更改为5000，因为一些KEGG集很大。

-csv

以csv格式输出。基因集结果的输出提供了更多信息，因此具有更复杂的格式。因为我们提供了每个基因集可以有多个集合ID和多个基因映射到它们.csv 的3列中的所有信息。

第一列是服务的setID作为标识每个集合的行的索引，从1开始，具有最低p值的集合。第二列是一个标志，说明最后一列提供了哪些信息，第三列是标记值。

下面是一个输出的CSV在右边框的例子。每个基因组的一行。



下面这个例子显示的setID与csv中的setID是一样的，它给出了哪些SNP映射到进入基因组的基因。

