-

使用手册

PopGeneCNV

Rapid detection of population gene CNV

based on bam

基于 bam 极快检测群体基因 CNV 的软件

population gene CNV

快速　bam.bai　多线程

Version 1.01 使用说明文档

2023-01-17

hewm2008@gmail.com / hewm2008@qq.com

微信 打赏　　QQ 入群: ***125293663***　　***微信公众号***

目录

1.简介

PopGeneCNV 是于 bam (sam/cram)文件，快速检测群体（一个样品或多个样品）存在多态性 CNV 的 候选基因出来。即输入样品的 bam 文件及基因信息文件(GFF/GTF)，即可以得到 1 群体的基因表达矩阵 2 群体的基因 CNV 矩阵， 3 候选出群体存在 CNV 多态性的基因

具体主要 功能如下:

1 自动检测是否存在索引 bam.bai 文件，根据基因坐标定位 bam 文件位置，可以极大提高读 bam 的速度，因为基因区域是基因组的 1%，只读取这些区域的 bam 会极大提速度。 同时 程序 自动识别 bam header 样品名
2 多线程，在存在 bam 索引文件时，默认 3 个进程并发加速读 bam
3 自动识别 GFF 还是 GTF， 考虑到有些物种基因的 intron 极大，程序基因坐标是剔除 intron 区域。即主要识别基因的 CDS 区域
4 定义了 CNV 多态性指标，快速过滤出群体中存在 CNV 差异的候选 gene 出来。
5 结果有各样品的基因深度和覆盖度文件，和基因 CNV 矩阵文件，和存在群体多态的 CNV 候选 gene 文件
6 用法简单，输出 bam 文件列表，和基因信息文件(GFF/GTF) ,即可以一步到位出结果。
7 静态编译，linux 直接解压可以

亮点:

1 极快， 索引 多线。
2 用法简单
3 结果精确（初 test 结果感觉结果还不错..）

程序是给一些有基础的生信朋友用的，若是**小白**看不懂就算了。

各版本新功能

A. 第一次 对外公开，v1 Beta 版本，大家试用 有 bug 和更的需求有空再丰富
B.

2.应用场景示例

1 多样品高深度海量数据量中，快速检测群体基因 CNV

2 统计样品基因的深度和覆盖度，即可以用于转录本算基因 count，基因表达矩阵。

脚本用法如下：

./bin/PopGeneCNV -i Bam.list -g ./Ref/MC.gff -o outPrefix -t 5

PopGeneCNV 各位画时
1 遇到问题可以反馈与我
2 同时我希望得到好 可以帮发个贴 教与别人

2.1 检测群体基因 CNV

最简易用法。输入两个基因组的坐标 **bam 文件列表 和基因坐标信息 GFF 文件即可**

3. 下载与安装

3.1 下载网址

这主要是防止大家还在用 beta 版本，还在用可能存在 bug 的程序，强迫***大家定期更新***。

持续更新在这：https://github.com/hewm2008/PopGeneCNV 下载时记得***加个星星***哦

里面有 example1

3.2 预先安装

PopGeneCNV 适用于 Linux/Unix/**macOS** 系统。
本程序

若要到源代码的话，须要在安装之前，请先安装以下 3 使用条件：
1) gcc (>4.8 --std=c++11): 高版本 gcc
2) -lpthread ： 多线程程序库
3) Htslib ： 读 sam/bam/sram 库

3.3 安装

使用者可采用以下直接 chmod 755 *运行:
1)

```
git clone https://github.com/hewm2008/PopGeneCNV
cd PopGeneCNV; chmod -R 755 bin/*
./bin/ PopGeneCNV -h # 直接运行
```

4. 用法和参数说明

4.1 PopGeneCNV 参数

程序 PopGeneCNV 很简单，两个输入和一个输出. 具体如下。

```
[heweiming@cngb-ologin-25 bin]$ ./PopGeneCNV

      Usage: PopGeneCNV  -i  InSortBam.list -g gene.gff -o  outPrefix

           -i   <str>    input SAM/BAM files List, must be sort sort.bam
           -g   <str>    inPut gff/gtf file to stat only gene CDS Depth/Coverage
           -o   <str>    prefix of output file

           -r   <str>    inPut the ref.fa file for cram
           -b   <str>    list of the regions of which the coverage and mean of depth would be given

           -v   <float>  the min variance(S2) of CNV to filter cnv[0.1]
           -p   <float>  the min cnv polymorphic to filter cnv [0.1]

           -q   <int>    the quality to filter reads, default [0]
           -l   <int>    the region extension length,default [300]
           -t   <int>    the thread Num(CPU) for deal bam[3]
           -s            the bam with No bai method

           -h            show more details for help [hewm2008 v1.00]
```

-i 输入 bam 文件列表（建议是建好索引 bai）的，可以提速
-g 输入基因坐标文件，常规 GFF 和 GTF 文件
-o 输出文件名前缀

4.1.1 其它主要配置参数

序号	示例	解析说明
参数 1	-v <float> the min variance(S2) of CNV to filter cnv[0.1]	群体 CNV 方差（S2）的值，默认方差小于 0.1 相当所有样品的 CNV 的方差小于 0.1，这个群体在这个基因是没有多态的，
参数 2	<float> the min cnv polymorphic to filter cnv [0.1]	定义的 cnv 多态性指标，当这个指标小于 0.1 时，默认这个群体在这个基因是没有多态的
参数 3	-q <int> the quality to filter reads, default [0]	过滤 bam 里面 比对质量低的 read
参数 4	-l <int> the region extension length,default [300]	基因区域外延长度，主要保留一些在在边缘的 read
参数 5	-t <int> the thread Num(CPU) for deal bam[3]	多线程参数
参数 6		
参数 7	-b <str> list of the regions of which the coverage and mean of depth would be given	可以自己定义感兴趣的区域。格式为 bed 格式，chr start end
参数 8		
参数 9	-r <str> inPut the ref.fa file for cram	Bam sam 和 cram，若是 cram 须要结果 ref 文件来

4.2 输入文件

两种输入文件

A 一种比对文件（bam/sam/cram）

B 别一种是基因信息文件(GFF/GTF)

4.2.1 比对文件（bam）

Bam 文件见网上众多信息，也可以见 这儿 pdf

常通是 bwa 生成后 samtools 一些系处理。

4.2.2 基因坐标信息文件（GFF/GTF）

GFF 和 GTF 是两种最常用的基因组注释格式，在信息分析中建库时除了需要 fasta 文件一般还会需要这两种文件，提取需要的信息进行注释。

一、GFF

GFF(General Feature Format)是一种用来描述基因组特征的文件，现在我们所使用的大部分都是第三版（gff3）。 具体 GFF 也可以网上查找。

以下截图为好:　　gff 文件格式：统计 Parent

```
Ca_59Chr01	PGA	gene	177061	181260	.	-	.	ID=Ca01g00050
Ca_59Chr01	PGA	mRNA	177061	181260	.	-	.	ID=Ca01g00050.1;Parent=Ca01g00050
Ca_59Chr01	PGA	exon	177061	177985	.	-	.	ID=Ca01g00050.1.exon5;Parent=Ca01g00050.1
Ca_59Chr01	PGA	CDS	177061	177985	.	-	1	ID=Ca01g00050.1.cds5;Parent=Ca01g00050.1
Ca_59Chr01	PGA	exon	178078	178140	.	-	.	ID=Ca01g00050.1.exon4;Parent=Ca01g00050.1
Ca_59Chr01	PGA	CDS	178078	178140	.	-	1	ID=Ca01g00050.1.cds4;Parent=Ca01g00050.1
Ca_59Chr01	PGA	exon	178246	178802	.	-	.	ID=Ca01g00050.1.exon3;Parent=Ca01g00050.1
Ca_59Chr01	PGA	CDS	178246	178802	.	-	0	ID=Ca01g00050.1.cds3;Parent=Ca01g00050.1
Ca_59Chr01	PGA	exon	178918	179015	.	-	.	ID=Ca01g00050.1.exon2;Parent=Ca01g00050.1
Ca_59Chr01	PGA	CDS	178918	179015	.	-	2	ID=Ca01g00050.1.cds2;Parent=Ca01g00050.1
Ca_59Chr01	PGA	exon	181140	181260	.	-	.	ID=Ca01g00050.1.exon1;Parent=Ca01g00050.1
Ca_59Chr01	PGA	CDS	181140	181260	.	-	0	ID=Ca01g00050.1.cds1;Parent=Ca01g00050.1
```

二、GTF

gtf 全称为 gene transfer format，主要是用来对基因进行注释，当前所广泛使用的 gtf 格式为第二版（gtf2）。。
以下截图为好:

以下截图为好:　　gtf 文件格式：统计 transcript_id

```
Ca_59Chr01	PGA	exon	18837	20023	.	+	.	transcript_id "Ca01g00010.1"; gene_id "Ca01g00010";
Ca_59Chr01	PGA	CDS	18930	19574	.	+	0	transcript_id "Ca01g00010.1"; gene_id "Ca01g00010";
Ca_59Chr01	PGA	exon	30366	32373	.	+	.	transcript_id "Ca01g00020.1"; gene_id "Ca01g00020";
Ca_59Chr01	PGA	CDS	30443	31438	.	+	0	transcript_id "Ca01g00020.1"; gene_id "Ca01g00020";
Ca_59Chr01	PGA	exon	45493	50986	.	+	.	transcript_id "Ca01g00030.1"; gene_id "Ca01g00030";
Ca_59Chr01	PGA	CDS	45532	49347	.	+	0	transcript_id "Ca01g00030.1"; gene_id "Ca01g00030";
Ca_59Chr01	PGA	exon	73652	74025	.	-	.	transcript_id "Ca01g00040.1"; gene_id "Ca01g00040";
Ca_59Chr01	PGA	exon	74210	74397	.	-	.	transcript_id "Ca01g00040.1"; gene_id "Ca01g00040";
Ca_59Chr01	PGA	exon	74511	74640	.	-	.	transcript_id "Ca01g00040.1"; gene_id "Ca01g00040";
Ca_59Chr01	PGA	exon	74913	75446	.	-	.	transcript_id "Ca01g00040.1"; gene_id "Ca01g00040";
```

4.3 输出文件

共如下几个文件:

输出文件	说明
out.gene.stat.gz	各个样品各个基因的深度和覆盖度等统计信息
out.GeneDepth.mat.gz	各个样品基因的深度，即基因表达矩阵(readcount)

Out.GeneCNV.Raw.mat.gz	标准化后，群体基因 CNV 的矩阵
Out.GeneCNV.filter.mat.gz	标准化后，存在群体多态 CNV 的候选基因
outPrefix.GeneCNV.Raw.vcf.gz	标准化后，群体基因 CNV 的 VCF 格式文件
outPrefix.GeneCNV.Filter.vcf.gz	标准化后，存在群体多态 CNV 候选基因（VCF 格式）

其中各文件是有 header 的，打开即明

具体以 3 个样品的前面部分截图示例

out.gene.stat.gz　　各个样品各个基因的深度和覆盖度等统计信息

#GeneID SampleName Length CoveredSite TotalDepth Coverage% MeanDepth

#GeneID	SampleName	Length	CoveredSite	TotalDepth	Coverage%	MeanDepth
MCh00g14009	LB	1500	1500	1583610	100.00	1055.74
MCh00g14007	LB	261	161	177751	61.69	681.04
MCh00g14005	LB	246	146	396	59.35	1.61
MCh00g14004	LB	534	135	486	25.28	0.91
MCh00g14002	LB	222	222	141666	100.00	638.14
MCh00g14001	LB	345	345	382105	100.00	1107.55
MCh00g14000	LB	561	491	64992	87.52	115.85
MCh00g13996	LB	666	327	61518	49.10	92.37
MCh00g13993	LB	612	612	135952	100.00	222.14
MCh00g13991	LB	282	282	104015	100.00	368.85
MCh00g13989	LB	225	225	84988	100.00	377.72
MCh00g13987	LB	558	558	200515	100.00	359.35
MCh00g13982	LB	531	531	223205	100.00	420.35
MCh00g13978	LB	612	612	165254	100.00	270.02
MCh00g13977	LB	429	429	111449	100.00	259.79

out.GeneDepth.mat.gz　　各个样品基因的深度，即基因表达矩阵(readcount)

GeneID 样品 1 样品 2 …

#GeneID	LB	LG	LH
MCh00g04914	0.00	0.00	0.00
MCh00g04915	0.00	0.00	0.00
MCh00g04916	0.00	0.00	0.00
MCh00g04917	1066.08	2306.49	1100.18
MCh00g04918	1050.94	2157.08	1061.07
MCh00g04919	439.54	883.84	456.20
MCh00g04920	684.90	1231.55	728.43
MCh00g04921	533.82	1047.94	535.70
MCh00g04922	652.01	1340.84	701.95
MCh00g04923	1083.32	2159.89	1130.36
MCh00g04924	1014.18	2233.47	1064.52
MCh00g04925	2007.63	4480.78	2057.59
MCh00g04926	2089.11	4600.20	2131.89

Out.GeneCNV.Raw.mat.gz　　标准化后，群体基因 CNV 的矩阵

GeneID　群体 CNV 多态性指标 LogPi　　S2 所有样品 CNV 的方差　　样品 1CNV　　样品 2CNV …

#GeneID LogPi	S2	LB	LG	LH	
MCh00g04914	0.00	0.00	0.00	0.00	0.00
MCh00g04915	0.00	0.00	0.00	0.00	0.00
MCh00g04916	0.00	0.00	0.00	0.00	0.00
MCh00g04917	0.00	2.87	31.87	35.18	31.36
MCh00g04918	0.00	1.18	31.42	32.90	30.25
MCh00g04919	0.00	0.04	13.14	13.48	13.00
MCh00g04920	0.00	0.76	20.48	18.79	20.76
MCh00g04921	0.00	0.11	15.96	15.98	15.27
MCh00g04922	0.00	0.15	19.49	20.45	20.01
MCh00g04923	0.00	0.10	32.39	32.95	32.22
MCh00g04924	0.00	3.10	30.32	34.07	30.35
MCh00g04925	0.00	18.35	60.02	68.35	58.65
MCh00g04926	0.00	16.73	62.45	70.17	60.77
MCh00g04927	0.00	4.02	64.83	69.11	64.88

Out.GeneCNV.filter.mat.gz　标准化后，存在群体多态 CNV 的候选基因

格式同上：截图如下

#GeneID LogPi	S2	LB	LG	LH	
MCh02g06373	2.38	0.13	0.00	0.76	0.00
MCh02g06524	0.17	0.11	0.37	1.02	0.28
MCh02g06525	0.13	0.13	0.42	1.18	0.41
MCh02g06552	0.89	0.27	0.07	1.17	0.07
MCh04g17936	0.30	0.26	0.38	1.40	0.24
MCh04g17937	0.22	0.25	0.37	1.44	0.40
MCh04g17938	0.16	0.21	0.46	1.42	0.46
MCh04g17939	0.21	0.15	0.29	1.14	0.31
MCh04g17989	0.85	0.12	0.04	0.80	0.08
MCh04g17991	2.34	0.12	0.00	0.74	0.00
MCh04g17992	2.37	0.13	0.00	0.76	0.00
MCh06g20943	0.14	0.16	0.41	1.33	0.55
MCh07g04806	2.55	0.18	0.00	0.89	0.00
MCh08g15710	1.49	0.11	0.71	0.01	0.70
MCh09g11649	0.12	0.12	0.42	1.15	0.43
MCh09g12628	0.85	0.20	0.91	0.06	1.09
MCh09g12982	0.11	0.13	0.49	1.25	0.50

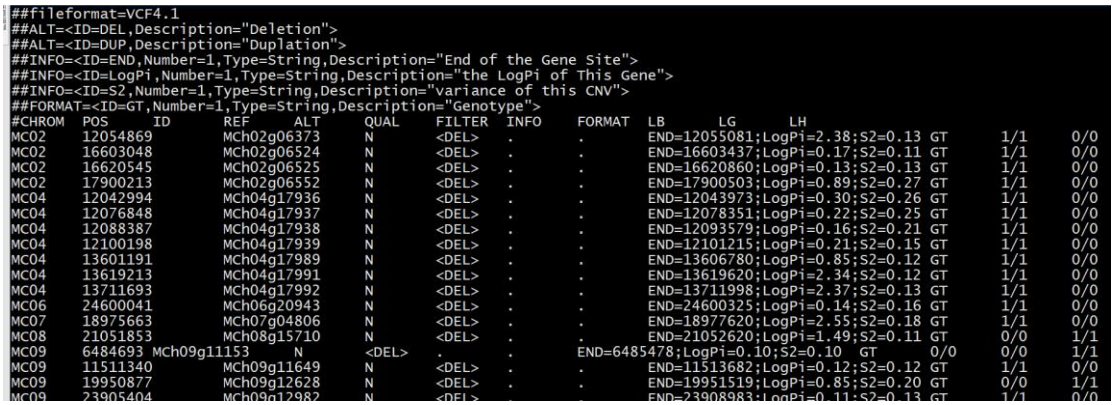
outPrefix.GeneCNV.Raw.vcf.gz　标准化后，群体基因 CNV 的 VCF 格式文件

和

outPrefix.GeneCNV.Filter.vcf.gz　标准化后，存在群体多态 CNV 候选基因（VCF 格式）

都是：

上面文件 转格式，转为 VCF，见如下截图：



```
##fileformat=VCF4.1
##ALT=<ID=DEL,Description="Deletion">
##ALT=<ID=DUP,Description="Duplation">
##INFO=<ID=END,Number=1,Type=String,Description="End of the Gene Site">
##INFO=<ID=LogPi,Number=1,Type=String,Description="the LogPi of This Gene">
##INFO=<ID=S2,Number=1,Type=String,Description="variance of this CNV">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  LB      LG      LH
MC02    12054869        MCh02g06373     N       <DEL>   .       .       END=12055081;LogPi=2.38;S2=0.13 GT      1/1     0/0
MC02    16603048        MCh02g06524     N       <DEL>   .       .       END=16603437;LogPi=0.17;S2=0.11 GT      1/1     0/0
MC02    16620545        MCh02g06525     N       <DEL>   .       .       END=16620860;LogPi=0.13;S2=0.13 GT      1/1     0/0
MC02    17900213        MCh02g06552     N       <DEL>   .       .       END=17900503;LogPi=0.89;S2=0.27 GT      1/1     0/0
MC04    12042994        MCh04g17936     N       <DEL>   .       .       END=12043973;LogPi=0.30;S2=0.26 GT      1/1     0/0
MC04    12076848        MCh04g17937     N       <DEL>   .       .       END=12078351;LogPi=0.22;S2=0.25 GT      1/1     0/0
MC04    12088387        MCh04g17938     N       <DEL>   .       .       END=12093579;LogPi=0.16;S2=0.21 GT      1/1     0/0
MC04    12100198        MCh04g17939     N       <DEL>   .       .       END=12101215;LogPi=0.21;S2=0.15 GT      1/1     0/0
MC04    13601191        MCh04g17989     N       <DEL>   .       .       END=13606780;LogPi=0.85;S2=0.12 GT      1/1     0/0
MC04    13619213        MCh04g17991     N       <DEL>   .       .       END=13619620;LogPi=2.34;S2=0.12 GT      1/1     0/0
MC04    13711693        MCh04g17992     N       <DEL>   .       .       END=13711998;LogPi=2.37;S2=0.13 GT      1/1     0/0
MC06    24600041        MCh06g20943     N       <DEL>   .       .       END=24600325;LogPi=0.14;S2=0.16 GT      1/1     0/0
MC07    18975663        MCh07g04806     N       <DEL>   .       .       END=18977620;LogPi=2.55;S2=0.18 GT      1/1     0/0
MC08    21051853        MCh08g15710     N       <DEL>   .       .       END=21052620;LogPi=1.49;S2=0.11 GT      0/0     1/1
MC09    6484693 MCh09g11153     N       <DEL>   .       .       END=6485478;LogPi=0.10;S2=0.10  GT      0/0     0/0     1/1
MC09    11511340        MCh09g11649     N       <DEL>   .       .       END=11513682;LogPi=0.12;S2=0.12 GT      1/1     0/0
MC09    19950877        MCh09g12628     N       <DEL>   .       .       END=19951519;LogPi=0.85;S2=0.20 GT      0/0     1/1
MC09    23905404        MCh09g12982     N       <DEL>   .       .       END=23908983;LogPi=0.11;S2=0.13 GT      1/1     0/0
```

4.4 群体 CNV 多态性指标

在算基因 x 的这个基因的 CNV 多态时，我们定义了一个群本多态性指标 LogPi

在标准化后，

样品 i 的 CNV 记为 CNV_i，若这个不存在 CNV，那么CNV_i=1，即在 1 附近，

我们记 $Log_i = log10(CNV_i * 100 + 1)$.

那么群体 CNV 多态性标标 LogPi 为

$$\text{LogPi} = \left(\sum_i^n \sum_{j=i+1}^n (Log_i - Log_j)^2\right)/(C_n^2)$$

即为要对两个样品 i 和 j 的 CNV 差别为 $(Log_i - Log_j)^2$．一个样品的 LogPi 为所有样品的 CNV 差异的平均值。

程序默认　当基因的LogPi大于 0.1 时，同时方差 S2 也大于 0.1 时，这个基因被选为存在群体多态 CNV 的基因。

5.实例

上面示例图都有实例，我这

具体数据格式和配置等见这：

./bin/PopGeneCNV　-i Bam.list -g　./Ref/MC.gff　-o　outPrefix -t 5

这儿共提示了 1 个实例，配置文件和输入文件格式等，

PopGeneCNV　各位有问题 欢迎交流。

更多实例 随时更新，见 website 网页，具体见这：

https://github.com/hewm2008/PopGeneCNV 里面的

6.优势

1 快速少内存 。
2 用法简易，友好。一看即会
3 应用场景广泛，如 基因 readcount 矩阵，群体 CNV 等。
4 免安装，使用方便

7.常见问题

7.1 为什么要开发 PopGeneCNV？

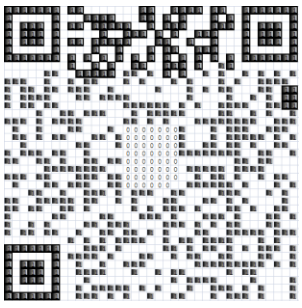
当前没有相关现成相似的功能，群体重测序 样品多 bam 大，极耗计算资源，须要开发一个可以读索引的，只对基因区域进行读取 bam 提速， 同时考虑到用户的等待时间，须要开线程并发。 所以考虑这个须求还是有一定人要用到，故开发本程序。

7.2 联系与打赏

随意 是缘是福，一切随风

- ✉ hewm2008@gmail.com / hewm2008@qq.com
- join the ***QQ Group : 125293663***

微信 打赏



QQ 入群：***125293663***

群名称:Reseqtools (itools)
群　号:125293663

微信公众号