How to analyze human RME

1. Background

区分印记基因和单等位基因

1.1 RME

在二倍体生物体中，一般认为每个基因的两个等位基因都会在相似的时间和水平上表达。然而，一些基因可以优先表达或严格从一个单一的等位基因，这一过程被称为单等位基因表达。单等位基因表达可能是由于等位基因之间的 DNA 序列多态性，例如在增强子或启动子序列中的多态性可能影响基因转录的效率，或基因组大部分的拷贝数变异(copy-number variations，CNVs)。在没有DNA序列多态性的情况下，也会出现单等位基因表达，最为典型的例子是雌性哺乳动物的X染色体失活（XCI）过程，它保证了XX雌性和XY雄性之间的X连锁基因剂量补偿\citep{wutz2011gene}。另一个典型的单等位基因调控的程序化例子则是印记基因。

印记基因是一种亲本起源特异性的单等位基因表达。印迹基因的单等位基因表达受生殖系等位基因特异性甲基化(allele-specific methylation，ASM，传统上称为差异甲基化区域(differentially methylated region，DMR))控制\citep{tycko2010allele, lo2003allelic}。在这些生殖系遗传系(gASMs)中，一个亲本等位基因的印迹控制区的 CpGs 是甲基化的，而另一个等位基因的 CpGs 是非甲基化的。当用亚硫酸氢盐Sanger测序检测时，这些区域的 PCR 克隆有一半是高甲基化的，另一半是低甲基化的，因此呈现出双峰式甲基化模式。重要的是，这些 gASMs/DMR 在胚胎发育早期的整体去甲基化和再甲基化周期中被认为是稳定的\citep{bartolomei25mammalian}。一旦由于精子和/或卵母细胞的暴露，这些遗传性组织被改变，这种改变可以作为“表观遗传记忆”进入体细胞。这种表观遗传改变的水平或许能在个体中维持数十年。此前，在一项针对跨代表观遗传的研究中，研究者们发现，出生前即暴露于荷兰饥荒的个体在出生60年后，印记的 IGF2基因的 DNA 甲基化水平仍然低于未暴露的同性兄弟姐妹\citep{heijmans2008persistent}。目前发现，大约超过100个基因座受父母起源依赖性表达的影响。许多印迹基因在出生前和出生后的新陈代谢中都起着重要作用，它们的双等位基因表达可以导致严重的表型\citep{barlow2011genomic}。不过，在某些组织比如大脑中，处于胎儿出生后神经发育的需要，神经干细胞会发生印记基因的沉默释放（release from silencing）\citep{ferron2011postnatal}。

除了亲本起源特异性单等位基因表达（parent-of-origin-specific monoallelic expression），哺乳动物基因组有大量的基因显示随机单等位基因表达(random monoallelic expression，RME)\citep{lo2003allelic, deng2014single, gendrel2014developmental}。旧观点认为，常染色体上的非印迹基因要么双细胞表达，要么双等位基因抑制。单等位基因表达被认为只发生在位于印记位点或女性 X 染色体的基因（其中一个来自父系或母系的X染色体在女性中被随机灭活）\citep{tycko2010allele}。但是，最近的研究发现，非印迹的常染色体基因的单等位基因的表达似乎不是一个偶发现象，而是在人类和小鼠基因组中的一个保守特征\citep{lo2003allelic, deng2014single, gendrel2014developmental}。虽然RME现象似乎经常发生，但表观遗传学机制好像无法对其发生进行解释\citep{gendrel2014developmental}。

这些随机单等位基因表达一般发生在常染色体基因座上\citep{chess2012mechanisms}，包括大基因家族的重要成员。这些随机单等位基因以高度组织特异性的方式表达，一般参与感觉或免疫系统功能(如人类白细胞抗原HLA)，如淋巴细胞中的免疫球蛋白（immunoglobulin）和T细胞受体基因（T cell receptor genes）、嗅觉神经元中的气味剂受体（odorant receptor，OR）基因、浦肯野神经元（Purkinje neurons）中的原珠蛋白（protocadherins，Pcdhs）等\citep{chess2012mechanisms}。这种单基因和单配位表达（monogenic and monoallelic expression）的现象（也称为等位基因排异（allelic exclusion））被认为对明确细胞身份和保证细胞多样性至关重要。事实上，研究者认为，等位基因排斥保证了每个 b 细胞或 t 细胞只产生一个单一的抗原受体，这种方式有助于免疫多样性\citep{cedar2008choreography}。等位基因排斥似乎代表了一种普遍的表观遗传现象。重要的是，单等位基因的表达在细胞谱系和个体之间存在差异。因此，RME 可能产生多样性的基因表达模式，对细胞的命运和生理有重要影响\citep{ohlsson2007widespread}。

单等位基因表达状态可以稳定地保持在几代细胞中。

单等位表达基因在发育过程中起着重要作用。此前的研究发现，许多常染色体基因的单等位基因表达发生在体内，在正常发育过程中，并趋向于高度组织特异性\citep{gendrel2014developmental}。

RME导致的表观遗传改变与许多人类疾病的发生密切相关。例如，表观遗传学研究表明，癌细胞具有全基因组低甲基化和区域特异性低甲基化的特点，而等位基因失衡(allelic imbalance，AI)与癌症发生风险上升相关。乳腺癌和卵巢癌散发病例中，由于启动子过度甲基化导致的 BRCA1表达缺失与两种癌症的发病相关\citep{esteller2000promoter, thrall2006brca1}。先前的一项研究报道，13个人类基因中有6个，包括 BRCA1和 p53，在两个等位基因中表达有显著差异，并且这种差异是通过孟德尔定律遗传的\citep{yan2002allelic}。

此外，人类常染色体显性遗传性疾病涉及单等位基因表达。

How to analyze ASE in RNA-seq?

\citep{borel2015biased}

> To investigate the extent of allele-specific transcription of autosomal human protein-coding genes, we used single-cell RNA sequencing (RNA-seq) technology to study 203 single cells from two different human primary fibroblast cell lines.
>
> By analyzing informative single-nucleotide variants (SNVs), we determined the relative mRNA abundance of each of the two alleles. For most of the actively transcribed genes, our results revealed that one allele was predominantly detected in a single cell at a particular point in time, whereas the second allele was at low levels or undetectable.

> We observed a stochastic process given that equal numbers of single cells expressed one or the other allele and a minority of single cells expressed both alleles. Interestingly, we detected only a few genes with an equal mRNA level from both alleles in all single cells. Detailed genomic characterization of these “single-cell biallelic” genes revealed that they express high levels of mRNA in a large number of cells.

method

> 1. Single-Cell Capture
> 2. cDNA Synthesis and Pre-amplification of Single Cells
> 3. Total RNA Extraction from Bulk Cell Samples
> 4. mRNA-Seq Library Preparation
> 5. Whole-Genome Sequencing
> 6. Spike-In Experiment
>
>   由于不同文库测序深度不同，比较前当然要进行均一化！用总reads进行均一化可能最简单，其基于以下两个基本假设：
>
>   - 绝大多数的gene表达量不变；
>   - 高表达量的gene表达量不发生改变；
>
>   但在转录组中，通常一小部分极高丰度基因往往会贡献很多reads，如果这些“位高权重”的基因还是差异表达的，则会影响所有其它基因分配到的reads数，而且，两个样本总mRNA量完全相同的前提假设也过于理想了。那如何比较呢，各个方家使出浑身解数，有用中位数的，有用75分位数的，有用几何平均数的，有用TMM(trimmed mean of Mvalues)的等等，总之要**找一个更稳定的参考**值。
>
>   *House-keeping gene(s)*
>
>   矫正的思路很简单，就是在变化的样本中寻找不变的量
>
>   那么在不同RNA-seq样本中，那些是不变的量呢？一个很容易想到的就是**管家基因** (House-keeping gene(s))
>
>   使用Housekeeping gene的办法来进行相对定量，这种办法在一定程度上能够解决我们遇到的问题。但其实这种办法有一个**非常强的先验假设**：housekeeping gene的表达量不怎么发生变化。其实housekeeping gene list有几千个，这几千个基因有一定程度上的变化是有可能的。
>
>   An RNA spike-in is an RNA transcript of known sequence and quantity used to calibrate measurements in RNA hybridization assays, such as DNA microarray experiments, RT-qPCR, and RNA-Seq.
>
>   A spike-in is designed to bind to a DNA molecule with a matching sequence, known as a control probe. This process of specific binding is called hybridization. A known quantity of RNA spike-in is mixed with the experiment sample during preparation. The degree of hybridization between the spike-ins and the control probes is used to normalize the hybridization measurements of the sample RNA.
>
>   在RNA-Seq建库的过程中掺入一些预先知道序列信息以及序列绝对数量的内参。这样在进行RNA-Seq测序的时候就可以通过不同样本之间内参（spike-in）的量来做一条标准曲线，就可以非常准确地对不同样本之间的表达量进行矫正。在这种操作下，可以一定程度上认为是一种绝对定量。（类似于housekeeping gene）

举例说明：

通过在样品制备过程中，混入指定数量的spike-in，我们就可以知道不同样本中的基因绝对比表达值。如等细胞数的样本A和样本B，在每个样本中，我加入了等量的spike-in。最后分析发现，spike-in占样本A的1%，但是占样本B的5%。这表明样本A的RNA表达量也许普遍比样本B的表达量高五倍左右。

ERCC control RNA

ERCC = External RNA Controls Consortium

ERCC就是一个专门为了定制一套spike-in RNA而成立的组织，这个组织早在2003年的时候就已经宣告成立。主要的工作就是设计了一套非常好用的spike-in RNA，方便microarray，以及RNA-Seq进行内参定量。

在RNA-Seq中增加ERCC的绝对量是可以获得FPKM的绝对量的增加，并且两者成非常好的线性关系。这也是我们能够对RNA-Seq样本进行掺入内参的一个基本前提。

R中用来处理带有ERCC spike-in的RNA-Seq数据的包：RUVSeq \citep{risso2014normalization}

7. RPSM Calculation

   RPSM stands for reads at a single-nucleotide position per sequencing read length (in kb) and per million mapped reads. The formula for RPSM is (10^6 × A) / (B × C), where A is the number of mappable reads at a nucleotide position, B is the total number of mappable reads of the sample, and C is the sequencing read length (in kb; C = 0.199).

8. Read Mapping for RNA-Seq Samples

9. ASE analysis

   ASE analysis was performed as in \citep{lappalainen2013transcriptome} In brief, we considered heterozygous sites obtained from whole-genome sequencing with DNA reads supporting both alleles. We used a minimum site quality call of 200. We excluded sites susceptible to allelic mapping bias, namely (1) sites with 50 bp mappability < 1 according to the UCSC mappability track (implying that the 50 bp flanking region of the site is not unique in the genome), and (2) sites where overlapping simulated RNA-seq reads showed a >5% mapping difference between those that carried the reference allele and those that carried the non-reference allele (see the methods in \citep{lappalainen2013transcriptome}).

   In all analyses, we only used uniquely mapped RNA-seq reads (GEM mapping quality > 150) and sites with base quality > 20 and support from at least 16 reads. Using information from SAMtools (v.0.1.19) mpileup\citep{li2009sequence}, we obtained for each site and each sample the number of reads mapping in the reference, the number of alternative alleles, and the sum of both. Each site was then annotated with the overlapped genomic feature in GENCODE annotation v.15 or the novel exons from the de novo assemblies of each sample. For each site, the number of single cells (and non-single cells) where the site was assessed was also counted. The distribution of allelic ratios for all samples is reported in Figure S9.

等位基因特异性表达不平衡（Allelic Expression Imbalance，AEI）可以作为表型用于寻找功能顺式作用多态性，同时可以作为转录区的分子标记来观察同一个体中不同等位基因的特异性表达差异。在杂合子个体中，分别来自父亲与母亲的等位基因，它们在同一细胞中，处于相同的外部环境中，在没有顺式作用多态性（或特定基因表观遗传修饰）的情况下，他们的表达量应该是一样的。与此相反，个体的顺式作用多态性会影响基因的表达及mRNA的加工，导致等位基因有不同的mRNA表达量水平，即AEI(等位基因表达不平衡)。它可以作为一个综合所有顺式作用因子的定量测量。另外，当目的SNP位点位于转录区时，可以直接在总RNA反转录cDNA中观察到SNP 不同等位基因的特异性表达差异，也就是用这个SNP作为Marker对两条不同allele分别进行表达定量。从而观察在同一杂合子个体中两种不同的等位基因型是否对表达水平造成了影响，也就是说排除了个体差异和环境影响之后是否存在由该位点不同基因型所造成的表达差异。该实验需要此SNP位点分型为杂合子的个体的基因组DNA和目的组织细胞抽提得到的总RNA（或者已经反转录好的cDNA）。实验中我们将用基因组DNA的两条allele作为1:1的校正内参，观察目的组织细胞中的RNA中两条allele的比例是否偏离1:1，最终判断是否存在AEI现象，进一步确定该位点与表达水平的关系。

（需要检测的对象：杂合子的个体的基因组DNA和目的组织细胞抽提得到的总RNA）

10. Gene Quantification and De Novo Assembly

    We used the software Cufflinks (v.2.1.1)\citep{trapnell2012differential, trapnell2010transcript} with default parameters and GENCODE v.12 as a reference annotation\citep{harrow2012gencode}. On the basis of Cufflinks transcript (170,086) quantifications, we selected for further analysis single cells that passed the arbitrary threshold of 12,000 transcripts expressed at FPKM (fragments per kilobase of exon per million reads mapped) > 0.3. We retained 163 UCF1014 single-cell samples expressing an average of 15,807 transcripts (the remaining samples expressed an average of 4,998 transcripts). Additionally, for each sample we performed de novo assembly to identify novel transcripts (Figure S4) without using the reference annotation. We then used the program cuffcompare to compare the assembled transcripts with the GENCODE reference annotation (v.15). Finally, for the four bulk RNA samples, we merged the four assemblies into a merged bulk RNA assembly. We compared each single-cell de novo assembly against the merged bulk RNA assembly to identify novel single-cell-specific transcripts. The program intersectBed from bedtools28 was used for this last comparison.

softwares（主要针对ASE分析）

1. SAMtools (v.0.1.19)
2. GENCODE annotation v.15
3. GENCODE v.12
4. Cufflinks(v. 2.1.1)
5. cuffcompare
6. GENCODE reference annotation (v.15).
7. intersectBed （from bedtools）
8. gemtools v. 1.6.2
9. RUVSeq(R) (处理spike-in内参) \citep{borel2015biased}
10. GATK ASEReadCounter \citep{mckenna2010genome}

如何利用SNP信息构建伪亲代基因组，然后把子代的测序信息mapping上去(用STAR）？

输入：BAM files (with proper headers) to be analyzed for ASE；A VCF file with specific sites to process

输出：A table of allele counts at the given sites. By default, it is formatted as a tab-delimited text file that is readable by R and compatible with Mamba, a downstream tool developed for allele-specific expression analysis.

11. Bedtools
12. STAR
13. Mamba

    a tool for further analysis data by GATK

    ```
    git clone https://git.code.sf.net/p/mambas/mambas mambas-mambas
    #install
    ```

14. MBASED(利用RNA-seq数据进行ASE检测，将多个单核苷酸变异位点的信息聚合在一起，以获得ASE的基因水平测量，即使在事先没有相位信息的情况下也可以进行) \citep{mayba2014mbased}
15. GeneiASE(仅使用 RNA-seq 数据，不需要已知或估计的单倍型，并且可以作为可下载的软件包使用。问题是，不能确定真正的单倍型，不能确定基因究竟是来自于父系等位基因还是母系等位基因的表达。) \citep{edsgard2016geneiase}
16. QuASAR(R包，当基因型数据无法获取时，可以用此方法。 i)从下一代测序读取基因分型，ii)对杂合子位点的等位基因不平衡进行推断。测序数据可以是 RNA-seq、 DNase-seq、 ATAC-seq 或任何其他类型的高通量测序数据。) \citep{harvey2015quasar}
17. ASEP（只用RNA-seq数据检测ASE）\citep{fan2020asep}
18. SCALE（需要提前确定来自父母系的等位基因，R中的输入数据是矩阵格式，分为母系基因的read count matrix和父系基因的read count matrix，输出是某个基因的父系与母系等位基因表达量/频率）\citep{borel2015biased}
19. phASER \citep{castel2016rare}

    输入：VCF文件（VCF，或 Variant Call Format，它是一种标准化的文本文件格式，用于表示 SNP、 indel 和结构变化调用。VCF 规范过去由千人基因组计划卫生组织维护，但其管理和进一步开发已被基因组学与健康全球联盟的基因组数据工具包团队接管。） tabix index for the VCF，BAM format file containing RNA-seq reads，index for the BAM file。

eQTL analysis

Expression quantitative trait loci (eQTL) 基因表达数量性状基因座

eQTL 研究成本高昂，需要大样本量和对每个样本进行全基因组分型

To date, most eQTL studies have considered the effects of genetic variation on expression within a single tissue (typically blood).

分析软件/程序：

MT-eQTL

ASE analysis

> analysis of allele-specific expression (ASE) 分析等位基因特异性表达
>
> 通常是通过计算与杂合位点上每个等位基因相匹配的 RNA-seq 读数和检验1:1等位基因比率的零假设来实现的
>
> 原则上，当基因型信息不容易获得时，可以从 RNA-seq 读数直接推断。然而，在考虑基因型调用的不确定性的情况下，目前还没有联合推断基因型和进行 ASE 推断的方法。

利用SNP来定位不同来源的等位基因 \citep{xie2019modeling} BLMRM方法（R包）

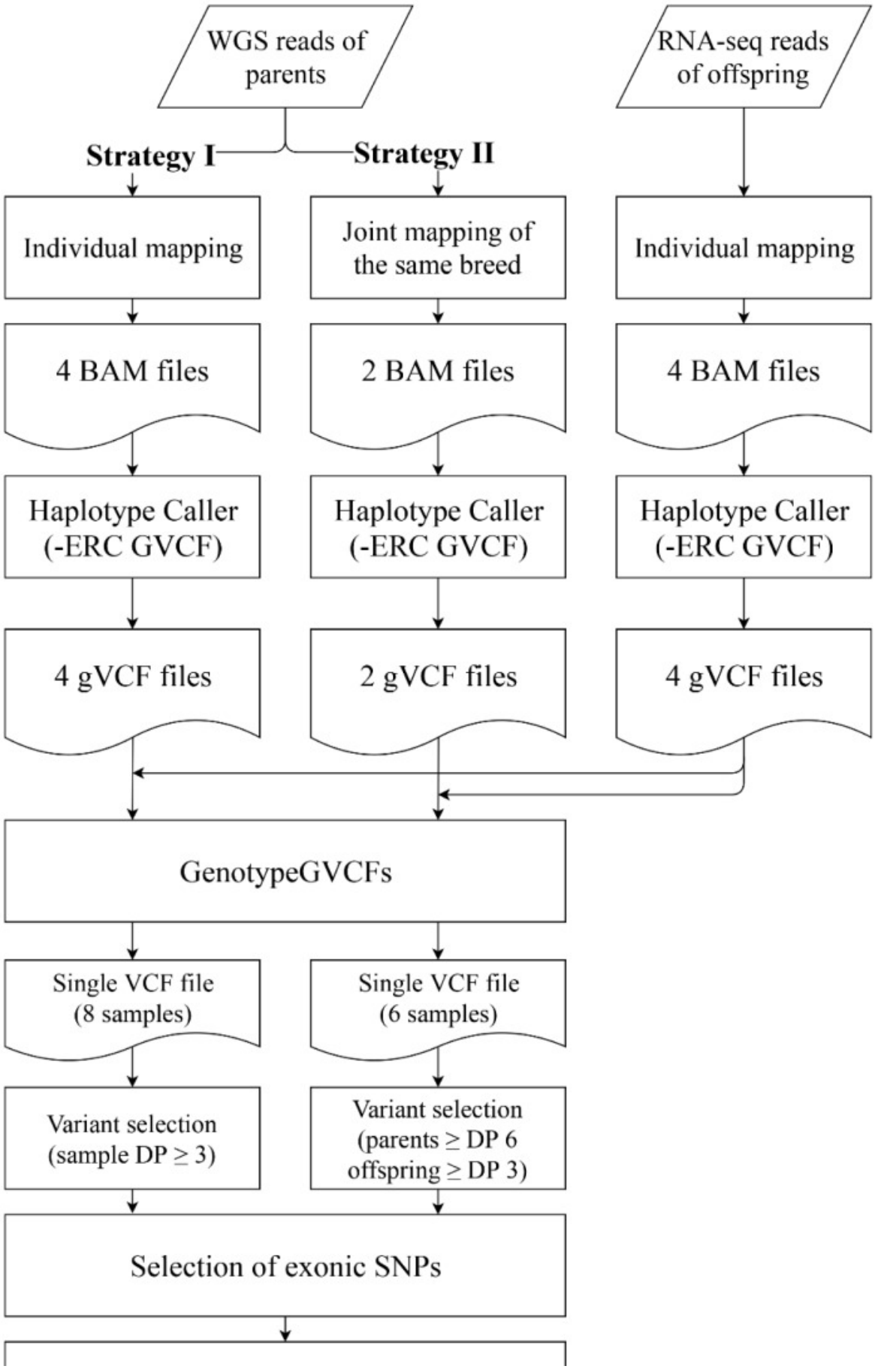
> 公牛的 DNA 经过了下一代测序(DNA-seq)，以确定他的基因组和母牛的参考基因组之间的所有 SNPs。
>
> 然后应用基因组分析工具包(Genome Analysis Toolkit，GATK)\citep{mckenna2010genome}和 SAMtools \citep{li2011statistical}来调用 SNP，只使用两个管道所识别的 SNPs 来生成伪 基因组。
>
> 最后，利用 HISAT2 \citep{kim2015hisat}和 BWA \citep{li2009fast}将来自母牛 × 公牛 F1代的 RNA-seq 片段定位到二倍体基因组，并保留两种方法鉴定的变异体，以减少假阳性的可能影响。

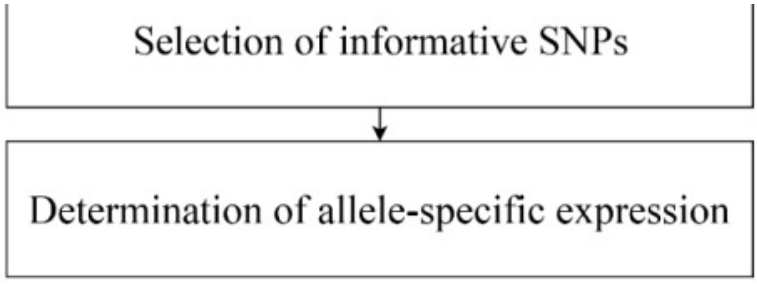
利用亲本特异性 SNPs 分析等位基因特异表达基因。\citep{ahn2019analysis}

> Method: 进行了父母基因组和后代转录组测序使用下一代测序。
>
> 随后，利用单核苷酸多态性(single nucleotide polymorphism，SNPs)的个体基因组定位和用于正反杂交的同一品种亲本的联合基因组定位，采用两种不同的方法对 snp 进行了基因组尺度识别。
>
> 利用亲本特异性 SNPs 分析等位基因特异表达基因。
>
> Result: 由于测序结果的基因组覆盖率较低(约4 ×)，大多数 SNPs 对于亲本谱系鉴定后代表达的等位基因没有信息价值，因此被排除在我们的分析之外。因此，包含336个基因的436个单核苷酸多态性可用于检测父系等位基因在后代中的不平衡表达。通过计算双亲等位基因在后代中的阅读比例，我们鉴定了7个表现等位基因偏向表达的基因($p < 0.05$)，其中包括以前报道的3个基因和本研究中新发现的4个基因。
>
> Read mapping: 下载该物种的参考基因组。使用 BWA MEM 包(版本0.7.17-r1188) \citep{li2009fast}将序列读取与引用对齐。SAM 文件使用 samtools 索引转换为 BAM 格式，并使用 samtools sort \citep{li2009sequence}进行排序。RNA-seq 读取到参考基因组的映射是使用带有默认选项和2-pass 模式的 STAR 包(版本2.5.3 a)进行的。
>
> Variant calling and filtration： 使用 Picard 工具中的 markduplices (2.15.0版， https://broadinstitute.github.io/Picard/)删除了映射读取的副本。对于 RNA-seq 读操作，read 组使用 GATK AddOrReplaceReadGroups 添加到映射读操作中，而使用 GATK Split NCigarReads \citep{poplin2017scaling}删除内含区域中映射的溢出读操作。基本质量的阅读是重新校准使用 GATK BaseRecalibrator (版本3.8)的全基因组测序和 RNA-seq 的结果，质量调整的阅读是使用 GATK PrintReads。父母基因组和后代转录组的变异体分别用 GATK 和 HaplotypeCaller 命名，而所谓的变异体用 GATK GenotypeGVCFs \citep{poplin2017scaling}进行联合基因分型。基因型变体用 NCBI dbSNP 150 \citep{sherry2001dbsnp}和 ENSEMBL 注释(发行版92)注释，使用 GATK VariantAnnotator 和 snpEff \citep{cingolani2012program}。随后，使用 GATK variantfilling 和 selectvariant 去除强链偏倚变异(Fisher strand > 30)、低质量深度变异(< 2)和35bp 窗口内3个或3个以上单核苷酸多单核苷酸多态性(SNP)变异体。我们还过滤了低深度变异(个体映射读深度[DP]

< 3，同一品种联合映射读深度 < 6)。最后，利用 gatkselectvariant 和 snpSift \citep{ruden2012using}筛选出外显 snp。

Workflow:





> 母体读数的比例计算为1 - 父体读数的比例。在该研究中，采用的任意判断等位基因表达偏倚的标准对于任何给定的等位基因都是 < 0.3或 > 0.7。当比值在0.3和0.7之间时，则认为它是双等位基因表达。

利用千人基因组计划鉴别等位基因表达\citep{jadhav2019rna}

> 数据来源：65 trios from the HapMap / 1000 genomes projects with RNA-Seq data from lymphoblastoid cell lines (LCLs)，and 131 trios from the Genome-of-the-Netherlands （查资料发现好像是血液样本的WGS测序？）
>
> 目前思路：
>
> 先选用4个CHS家庭（一共8位父母的基因组测序数据low coverage，4个孩子的RNA-seq数据，其中两个男孩，两个女孩）试着分析

family trio 1(SH007)

Collection UUID:qr1hi-4zz18-hxa1at658d85c1r（用于在run program时使用）

*Content address:*d41d8cd98f00b204e9800998ecf8427e+0

HG00421(father) https://www.ebi.ac.uk/ena/browser/view/SRR1606795 (downloaded) (fastq-dump ing) 59520 (bam downloaded)
HG00422(mother) https://www.ebi.ac.uk/ena/browser/view/SRR1606537 (downloaded)(fastq-dump ing) 59419 (bam downloaded)
HG00423(female child, Lymphoblastoid cell total RNA) https://www.ebi.ac.uk/ena/browser/view/SRX2432923 （downloaded）(bam)

Family trio 2(SH002)

HG00406(father) https://www.ebi.ac.uk/ena/browser/view/SRR1602073 (fastq-dump ing) 58908 (bam downloaded)
HG00407(mother) https://www.ebi.ac.uk/ena/browser/view/SRR1602075 (downloading) 42363 (bam downloaded)
HG00408(female child, Lymphoblastoid cell total RNA) https://www.ebi.ac.uk/ena/browser/view/SRX2432921 (bam)

Family trio3(SH014)

HG00442(father) https://www.ebi.ac.uk/ena/browser/view/SRR1607190 (fastq-dumping)18181 (downloaded)
HG00443(mother) https://www.ebi.ac.uk/ena/browser/view/SRR1606504 (fastq-dumping)65250 (downloading bam) 11075
HG00444(male child, Lymphoblastoid cell total RNA) https://www.ebi.ac.uk/ena/browser/view/SRX2432926 (bam)

family trio 4(SH021)

HG00463(father) https://www.ebi.ac.uk/ena/browser/view/SRR1596842 (mapping) 49012 (downloaded)
HG00464(mother) https://www.ebi.ac.uk/ena/browser/view/SRR1596844 (prefetching) 13266 (downloaded)
HG00465(male child, Lymphoblastoid cell total RNA) https://www.ebi.ac.uk/ena/browser/view/SRX2432931 (bam)

QC by fastqc

```
fastqc
/Users/keqinliu/studying_document/bioinfo/human_RME/raw_data/SH007/SRX2432923/SRR5117458/SRR5117458.fastq.gz --extract -o
/Users/keqinliu/studying_document/bioinfo/human_RME/raw_data/SH007/fastqc_report

fastqc
/Users/keqinliu/studying_document/bioinfo/human_RME/raw_data/SH007/SRX2432923/SRR5117459/SRR5117459.fastq.gz --extract -o
/Users/keqinliu/studying_document/bioinfo/human_RME/raw_data/SH007/fastqc_report #failed
# in SH007/fastqc_report
```

GATK preprocess

```
samtools faidx new_hg19.fa #generate .fai
samtools dict new_hg19.fa -o new_hg19.dict # generate .dict
```

mapping by STAR/hisat2

```
 #generate index
 /Users/keqinliu/STAR-2.7.6a/bin/MacOSX_x86_64/STAR --runThreadN 6 --runMode genomeGenerate \
 --genomeDir /Users/keqinliu/studying_document/bioinfo/human_RME/STAR/SH007 \
```

```
 --genomeFastaFiles /Users/keqinliu/Downloads/hg38.fa \
 --sjdbGTFfile /Users/keqinliu/Downloads/human.gtf \
 --sjdbOverhang 100

 # do mapping
 STAR --runThreadN 20 --genomeDir ~/reference/index/STAR/mm10/ \
 --readFilesIn SRR3589959_1.fastq SRR3589959_2.fastq \
 --outSAMtype BAM SortedByCoordinate \
 --outFileNamePrefix ./SRR3589959


# the problem is the processing seems never end!
# and STAR took too much storage
# change to hisat2
# ref genome

hisat2 -x /Users/keqinliu/Downloads/hg38/genome \
-1
/Users/keqinliu/studying_document/bioinfo/human_RME/raw_data/SH007/SRR1606795.f
astq.gz \
-S
/Users/keqinliu/studying_document/bioinfo/human_RME/raw_data/SH007/SRR1606795.s
am
# report error
# Error: Must specify at least one read input with -U/-1/-2
# Overall time: 00:00:00
# (ERR): hisat2-align exited with value 1

hisat2 -x /Users/keqinliu/Downloads/hg38/genome \
-U
/Users/keqinliu/studying_document/bioinfo/human_RME/raw_data/SH007/SRR1606795.f
astq.gz \
-S
/Users/keqinliu/studying_document/bioinfo/human_RME/raw_data/SH007/SRR1606795.s
am # it works
```

Mac turn off/on

```
sudo pmset -b sleep 0; sudo pmset -b disablesleep 1 # sleep function off

sudo pmset -b sleep 5; sudo pmset -b disablesleep 0 # sleep function on
```

report

```
# SRR2432923.sam(child)
2599467 reads; of these:
```

```
  2599467 (100.00%) were unpaired; of these:
    149548 (5.75%) aligned 0 times
    1933123 (74.37%) aligned exactly 1 time
    516796 (19.88%) aligned >1 times
94.25% overall alignment rate

# SRR1606795.sam(mother)
3638957 reads; of these:
  3638957 (100.00%) were unpaired; of these:
    727664 (20.00%) aligned 0 times
    2715042 (74.61%) aligned exactly 1 time
    196251 (5.39%) aligned >1 times
80.00% overall alignment rate

# SRR1606537.sam(father)

# SRR1602073.sam (father)
2507162 reads; of these:
  2507162 (100.00%) were unpaired; of these:
    302285 (12.06%) aligned 0 times
    2065134 (82.37%) aligned exactly 1 time
    139743 (5.57%) aligned >1 times
87.94% overall alignment rate

# SRR1602075.sam (mother)
gzip:
/Users/keqinliu/studying_document/bioinfo/human_RME/raw_data/SH002/SRR1602075/S
RR1602075.fastq.gz: unexpected end of file
gzip:
/Users/keqinliu/studying_document/bioinfo/human_RME/raw_data/SH002/SRR1602075/S
RR1602075.fastq.gz: uncompress failed
6341028 reads; of these:
  6341028 (100.00%) were unpaired; of these:
    290072 (4.57%) aligned 0 times
    5817396 (91.74%) aligned exactly 1 time
    233560 (3.68%) aligned >1 times
95.43% overall alignment rate

# SRR5117454.sam(child)
```

sam to bam

```
# SH007
samtools view -S SRR2432923.sam -b > SRR2432923.bam #child
samtools view -S SRR1606795.sam -b > SRR1606795.bam #mother
samtools view -S SRR1606537.sam -b > SRR1606537.bam #father

# SH002
samtools view -S SRR1602073.sam -b > SRR1602073.bam #father
samtools view -S SRR1602075.sam -b > SRR1602075.bam #mother
samtools view -S SRR5117454.sam -b > SRR5117454.bam
```

GATK hyplotypeCaller

```
/Users/keqinliu/Downloads/gatk-4.1.9.0/gatk --java-options "-Xmx4g"
HaplotypeCaller  \
   -R
/Users/keqinliu/Downloads/resources_broad_hg38_v0_Homo_sapiens_assembly38.fasta
\
   -I
/Users/keqinliu/studying_document/bioinfo/human_RME/raw_data/SH007/SRX2432923/S
RR2432923.bam \ # child
   -O
/Users/keqinliu/studying_document/bioinfo/human_RME/raw_data/SH007/SRR2432923.g
.vcf.gz \
   -ERC GVCF
   -G Standard \
   -G AS_Standard

# report error!
#

# then try to check whether bam file is damaged
samtools view -c -f 1 -F 12
/Users/keqinliu/studying_document/bioinfo/human_RME/raw_data/SH007/SRX2432923/S
RR2432923.bam
# ruturn 0
samtools view -c -f 1 -F 12
/Users/keqinliu/studying_document/bioinfo/human_RME/raw_data/SH002/SRR1602073.b
am
samtools view -c -f 1 -F 12
/Users/keqinliu/studying_document/bioinfo/human_RME/raw_data/SH002/SRR5117454.b
am
# all return 0
# so bam file is damaged

# check whether sam file is damaged
samtools view -c -f 1 -F 12
/Users/keqinliu/studying_document/bioinfo/human_RME/raw_data/SH002/SRR1602073.s
am
```

```
# return 0
# so sam file is dameged
```

GATK GenotypeGVCFs

```

```

```
ascp -k 1 -QT -l 300m -P33001 -i ~/.aspera/connect/etc/asperaweb_id_dsa.openssh
era-fasp@fasp.sra.ebi.ac.uk:$SRR1606795.

ascp -i asperaweb_id_dsa.openssh -Tr -Q -l 6M -P33001 -L- -k1 era-
fasp@fasp.sra.ebi.ac.uk:/vol1/fastq/SRR160/007/SRR1606537/SRR1606537.fastq.gz

ascp -k 1 -QT -l 300m -P33001 -i ~/.aspera/connect/etc/asperaweb_id_dsa.openssh
anonftp@ftp-private.ncbi.nlm.nih.gov://sra/sra-
instant/reads/ByExp/sra/SRX/SRX727/SRX727828/SRR1606795/SRR1606795.sra.

ascp -v -QT -l 400m -P33001 -k1 -i
~/.aspera/connect/etc/asperaweb_id_dsa.openssh
anonftp@ftp.ncbi.nlm.nih.gov:/sra/sra-
instant/reads/ByRun/sra/SRR/SRR160/SRR1606795/SRR1606795.sra.

ftp.sra.ebi.ac.uk/vol1/fastq/SRR160/007/SRR1606537/SRR1606537.fastq.gz
```

```
/data/mouse/human_rme/sratoolkit.2.10.8-ubuntu64/bin/fastq-dump -I --split-
files

/data/mouse/human_rme/sratoolkit.2.10.8-ubuntu64/bin/prefetch SRR1602075
```

```
STAR --genomeDir /data/mouse/human_rme/genome_index \
    --runThreadN 20 \
    --readFilesIn sample_r1.fq.gz sample_r2.fq.gz \
    # --readFilesCommand zcat \
    --outFileNamePrefix sample \
    --outSAMtype BAM SortedByCoordinate \
    --outBAMsortingThreadN 10
```

```
/data/mouse/human_rme/genome_index/gatk-4.1.9.0/gatk --java-options "-Xmx4g"
HaplotypeCaller  \
   -R /data/mouse/human_rme/genome_index/hg19.fa \
   -I /data/mouse/human_rme/SH007/SRR5117458.1.bam \ # child
   -O /data/mouse/human_rme/SH007/SRR5117458.1.g.vcf.gz \
   -ERC GVCF
   -G Standard \
   -G AS_Standard

https://github.com/samtools/samtools/releases/download/1.11/samtools-
1.11.tar.bz2
```

```
@RG ID:HG00423  PL:illumina PU:H0164ALXX140820.2    LB:Solexa-272222    PI:0
DT:2014-08-20T00:00:00-0400 SM:NA12878  CN:BI
```

use GATK to do with 1000genome bam file

```
/data/mouse/human_rme/genome_index/gatk-4.1.9.0/gatk --java-options "-Xmx4g"
HaplotypeCaller  \
-R /data/mouse/human_rme/genome_index/hg19.fa \
-I HG00421.mapped.ILLUMINA.bwa.CHS.low_coverage.20130415.bam \
-O /data/mouse/human_rme/SH007/HG00421.vcf.gz \
-ERC GVCF
```

error report:

```
A USER ERROR has occurred: Input files reference and reads have incompatible
contigs: No overlapping contigs found.
  reference contigs = [chr1, chr2, chr3, chr4, chr5, chr6, chr7, chrX, chr8,
chr9, chr10, chr11]
  reads contigs = [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17,
18, 19, 20, 21, 22, X, Y, MT, GL000207.1, GL000226.1, GL000229.1, GL000231.1,
GL000210.1, GL000239.1, GL000235.1, GL000201.1, GL000247.1, GL000245.1,
GL000197.1, GL000203.1, GL000246.1, GL000249.1, GL000196.1, GL000248.1,
GL000244.1, GL000238.1, GL000202.1, GL000234.1, GL000232.1, GL000206.1,
GL000240.1, GL000236.1, GL000241.1, GL000243.1, GL000242.1, GL000230.1,
GL000237.1, GL000233.1, GL000204.1, GL000198.1, GL000208.1, GL000191.1,
GL000227.1, GL000228.1, GL000214.1, GL000221.1, GL000209.1, GL000218.1,
GL000220.1, GL000213.1, GL000211.1, GL000199.1, GL000217.1, GL000216.1,
GL000215.1, GL000205.1, GL000219.1, GL000224.1, GL000223.1, GL000195.1,
GL000212.1, GL000222.1, GL000200.1, GL000193.1, GL000194.1, GL000225.1,
GL000192.1, NC_007605, hs37d5]
```

Use hg19 ref genome

```
/data/mouse/human_rme/genome_index/gatk-4.1.9.0/gatk --java-options "-Xmx4g"
HaplotypeCaller  \
-R /data/mouse/human_rme/genome_index/hg19.fa \
-I HG00421.mapped.ILLUMINA.bwa.CHS.low_coverage.20130415.bam \
-O /data/mouse/human_rme/SH007/HG00421.vcf.gz \
-ERC GVCF
```

error report:

```
A USER ERROR has occurred: Input files reference and reads have incompatible
contigs: No overlapping contigs found.
  reference contigs = [chr1, chr2, chr3, chr4, chr5, chr6, chr7, chrX, chr8,
chr9, chr10, chr11]
  reads contigs = [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17,
18, 19, 20, 21, 22, X, Y, MT, GL000207.1, GL000226.1, GL000229.1, GL000231.1,
GL000210.1, GL000239.1, GL000235.1, GL000201.1, GL000247.1, GL000245.1,
GL000197.1, GL000203.1, GL000246.1, GL000249.1, GL000196.1, GL000248.1,
GL000244.1, GL000238.1, GL000202.1, GL000234.1, GL000232.1, GL000206.1,
GL000240.1, GL000236.1, GL000241.1, GL000243.1, GL000242.1, GL000230.1,
GL000237.1, GL000233.1, GL000204.1, GL000198.1, GL000208.1, GL000191.1,
GL000227.1, GL000228.1, GL000214.1, GL000221.1, GL000209.1, GL000218.1,
GL000220.1, GL000213.1, GL000211.1, GL000199.1, GL000217.1, GL000216.1,
GL000215.1, GL000205.1, GL000219.1, GL000224.1, GL000223.1, GL000195.1,
GL000212.1, GL000222.1, GL000200.1, GL000193.1, GL000194.1, GL000225.1,
GL000192.1, NC_007605, hs37d5]
```

use hg38 get error report

```
A USER ERROR has occurred: Input files reference and reads have incompatible
contigs: No overlapping contigs found.
  reference contigs = [chr1, chr10, chr11, chr11_KI270721v1_random, chr12,
chr13, chr14, chr14_GL000009v2_random, chr14_GL000225v1_random,
chr14_KI270722v1_random, chr14_GL000194v1_random, chr14_KI270723v1_random,
chr14_KI270724v1_random, chr14_KI270725v1_random, chr14_KI270726v1_random,
chr15, chr15_KI270727v1_random, chr16, chr16_KI270728v1_random, chr17,
chr17_GL000205v2_random, chr17_KI270729v1_random, chr17_KI270730v1_random,
chr18, chr19, chr1_KI270706v1_random, chr1_KI270707v1_random,
chr1_KI270708v1_random, chr1_KI270709v1_random, chr1_KI270710v1_random,
chr1_KI270711v1_random, chr1_KI270712v1_random, chr1_KI270713v1_random,
chr1_KI270714v1_random, chr2, chr20, chr21, chr22, chr22_KI270731v1_random,
chr22_KI270732v1_random, chr22_KI270733v1_random, chr22_KI270734v1_random,
chr22_KI270735v1_random, chr22_KI270736v1_random, chr22_KI270737v1_random,
chr22_KI270738v1_random, chr22_KI270739v1_random, chr2_KI270715v1_random,
chr2_KI270716v1_random, chr3, chr3_GL000221v1_random, chr4,
chr4_GL000008v2_random, chr5, chr5_GL000208v1_random, chr6, chr7, chr8, chr9,
chr9_KI270717v1_random, chr9_KI270718v1_random, chr9_KI270719v1_random,
```

```
chr9_KI270720v1_random, chr1_KI270762v1_alt, chr1_KI270766v1_alt,
chr1_KI270760v1_alt, chr1_KI270765v1_alt, chr1_GL383518v1_alt,
chr1_GL383519v1_alt, chr1_GL383520v2_alt, chr1_KI270764v1_alt,
chr1_KI270763v1_alt, chr1_KI270759v1_alt, chr1_KI270761v1_alt,
chr2_KI270770v1_alt, chr2_KI270773v1_alt, chr2_KI270774v1_alt,
chr2_KI270769v1_alt, chr2_GL383521v1_alt, chr2_KI270772v1_alt,
chr2_KI270775v1_alt, chr2_KI270771v1_alt, chr2_KI270768v1_alt,
chr2_GL582966v2_alt, chr2_GL383522v1_alt, chr2_KI270776v1_alt,
chr2_KI270767v1_alt, chr3_JH636055v2_alt, chr3_KI270783v1_alt,
chr3_KI270780v1_alt, chr3_GL383526v1_alt, chr3_KI270777v1_alt,
chr3_KI270778v1_alt, chr3_KI270781v1_alt, chr3_KI270779v1_alt,
chr3_KI270782v1_alt, chr3_KI270784v1_alt, chr4_KI270790v1_alt,
chr4_GL383528v1_alt, chr4_KI270787v1_alt, chr4_GL000257v2_alt,
chr4_KI270788v1_alt, chr4_GL383527v1_alt, chr4_KI270785v1_alt,
chr4_KI270789v1_alt, chr4_KI270786v1_alt, chr5_KI270793v1_alt,
chr5_KI270792v1_alt, chr5_KI270791v1_alt, chr5_GL383532v1_alt,
chr5_GL949742v1_alt, chr5_KI270794v1_alt, chr5_GL339449v2_alt,
chr5_GL383530v1_alt, chr5_KI270796v1_alt, chr5_GL383531v1_alt,
chr5_KI270795v1_alt, chr6_GL000250v2_alt, chr6_KI270800v1_alt,
chr6_KI270799v1_alt, chr6_GL383533v1_alt, chr6_KI270801v1_alt,
chr6_KI270802v1_alt, chr6_KB021644v2_alt, chr6_KI270797v1_alt,
chr6_KI270798v1_alt, chr7_KI270804v1_alt, chr7_KI270809v1_alt,
chr7_KI270806v1_alt, chr7_GL383534v2_alt, chr7_KI270803v1_alt,
chr7_KI270808v1_alt, chr7_KI270807v1_alt, chr7_KI270805v1_alt,
chr8_KI270818v1_alt, chr8_KI270812v1_alt, chr8_KI270811v1_alt,
chr8_KI270821v1_alt, chr8_KI270813v1_alt, chr8_KI270822v1_alt,
chr8_KI270814v1_alt, chr8_KI270810v1_alt, chr8_KI270819v1_alt,
chr8_KI270820v1_alt, chr8_KI270817v1_alt, chr8_KI270816v1_alt,
chr8_KI270815v1_alt, chr9_GL383539v1_alt, chr9_GL383540v1_alt,
chr9_GL383541v1_alt, chr9_GL383542v1_alt, chr9_KI270823v1_alt,
chr10_GL383545v1_alt, chr10_KI270824v1_alt, chr10_GL383546v1_alt,
chr10_KI270825v1_alt, chr11_KI270832v1_alt, chr11_KI270830v1_alt,
chr11_KI270831v1_alt, chr11_KI270829v1_alt, chr11_GL383547v1_alt,
chr11_JH159136v1_alt, chr11_JH159137v1_alt, chr11_KI270827v1_alt,
chr11_KI270826v1_alt, chr12_GL877875v1_alt, chr12_GL877876v1_alt,
chr12_KI270837v1_alt, chr12_GL383549v1_alt, chr12_KI270835v1_alt,
chr12_GL383550v2_alt, chr12_GL383552v1_alt, chr12_GL383553v2_alt,
chr12_KI270834v1_alt, chr12_GL383551v1_alt, chr12_KI270833v1_alt,
chr12_KI270836v1_alt, chr13_KI270840v1_alt, chr13_KI270839v1_alt,
chr13_KI270843v1_alt, chr13_KI270841v1_alt, chr13_KI270838v1_alt,
chr13_KI270842v1_alt, chr14_KI270844v1_alt, chr14_KI270847v1_alt,
chr14_KI270845v1_alt, chr14_KI270846v1_alt, chr15_KI270852v1_alt,
chr15_KI270851v1_alt, chr15_KI270848v1_alt, chr15_GL383554v1_alt,
chr15_KI270849v1_alt, chr15_GL383555v2_alt, chr15_KI270850v1_alt,
chr16_KI270854v1_alt, chr16_KI270856v1_alt, chr16_KI270855v1_alt,
chr16_KI270853v1_alt, chr16_GL383556v1_alt, chr16_GL383557v1_alt,
chr17_GL383563v3_alt, chr17_KI270862v1_alt, chr17_KI270861v1_alt,
chr17_KI270857v1_alt, chr17_JH159146v1_alt, chr17_JH159147v1_alt,
chr17_GL383564v2_alt, chr17_GL000258v2_alt, chr17_GL383565v1_alt,
```

```
chr17_KI270858v1_alt, chr17_KI270859v1_alt, chr17_GL383566v1_alt,
chr17_KI270860v1_alt, chr18_KI270864v1_alt, chr18_GL383567v1_alt,
chr18_GL383570v1_alt, chr18_GL383571v1_alt, chr18_GL383568v1_alt,
chr18_GL383569v1_alt, chr18_GL383572v1_alt, chr18_KI270863v1_alt,
chr19_KI270868v1_alt, chr19_KI270865v1_alt, chr19_GL383573v1_alt,
chr19_GL383575v2_alt, chr19_GL383576v1_alt, chr19_GL383574v1_alt,
chr19_KI270866v1_alt, chr19_KI270867v1_alt, chr19_GL949746v1_alt,
chr20_GL383577v2_alt, chr20_KI270869v1_alt, chr20_KI270871v1_alt,
chr20_KI270870v1_alt, chr21_GL383578v2_alt, chr21_KI270874v1_alt,
chr21_KI270873v1_alt, chr21_GL383579v2_alt, chr21_GL383580v2_alt,
chr21_GL383581v2_alt, chr21_KI270872v1_alt, chr22_KI270875v1_alt,
chr22_KI270878v1_alt, chr22_KI270879v1_alt, chr22_KI270876v1_alt,
chr22_KI270877v1_alt, chr22_GL383583v2_alt, chr22_GL383582v2_alt,
chrX_KI270880v1_alt, chrX_KI270881v1_alt, chr19_KI270882v1_alt,
chr19_KI270883v1_alt, chr19_KI270884v1_alt, chr19_KI270885v1_alt,
chr19_KI270886v1_alt, chr19_KI270887v1_alt, chr19_KI270888v1_alt,
chr19_KI270889v1_alt, chr19_KI270890v1_alt, chr19_KI270891v1_alt,
chr1_KI270892v1_alt, chr2_KI270894v1_alt, chr2_KI270893v1_alt,
chr3_KI270895v1_alt, chr4_KI270896v1_alt, chr5_KI270897v1_alt,
chr5_KI270898v1_alt, chr6_GL000251v2_alt, chr7_KI270899v1_alt,
chr8_KI270901v1_alt, chr8_KI270900v1_alt, chr11_KI270902v1_alt,
chr11_KI270903v1_alt, chr12_KI270904v1_alt, chr15_KI270906v1_alt,
chr15_KI270905v1_alt, chr17_KI270907v1_alt, chr17_KI270910v1_alt,
chr17_KI270909v1_alt, chr17_JH159148v1_alt, chr17_KI270908v1_alt,
chr18_KI270912v1_alt, chr18_KI270911v1_alt, chr19_GL949747v2_alt,
chr22_KB663609v1_alt, chrX_KI270913v1_alt, chr19_KI270914v1_alt,
chr19_KI270915v1_alt, chr19_KI270916v1_alt, chr19_KI270917v1_alt,
chr19_KI270918v1_alt, chr19_KI270919v1_alt, chr19_KI270920v1_alt,
chr19_KI270921v1_alt, chr19_KI270922v1_alt, chr19_KI270923v1_alt,
chr3_KI270924v1_alt, chr4_KI270925v1_alt, chr6_GL000252v2_alt,
chr8_KI270926v1_alt, chr11_KI270927v1_alt, chr19_GL949748v2_alt,
chr22_KI270928v1_alt, chr19_KI270929v1_alt, chr19_KI270930v1_alt,
chr19_KI270931v1_alt, chr19_KI270932v1_alt, chr19_KI270933v1_alt,
chr19_GL000209v2_alt, chr3_KI270934v1_alt, chr6_GL000253v2_alt,
chr19_GL949749v2_alt, chr3_KI270935v1_alt, chr6_GL000254v2_alt,
chr19_GL949750v2_alt, chr3_KI270936v1_alt, chr6_GL000255v2_alt,
chr19_GL949751v2_alt, chr3_KI270937v1_alt, chr6_GL000256v2_alt,
chr19_GL949752v1_alt, chr6_KI270758v1_alt, chr19_GL949753v2_alt,
chr19_KI270938v1_alt, chrM, chrUn_KI270302v1, chrUn_KI270304v1,
chrUn_KI270303v1, chrUn_KI270305v1, chrUn_KI270322v1, chrUn_KI270320v1,
chrUn_KI270310v1, chrUn_KI270316v1, chrUn_KI270315v1, chrUn_KI270312v1,
chrUn_KI270311v1, chrUn_KI270317v1, chrUn_KI270412v1, chrUn_KI270411v1,
chrUn_KI270414v1, chrUn_KI270419v1, chrUn_KI270418v1, chrUn_KI270420v1,
chrUn_KI270424v1, chrUn_KI270417v1, chrUn_KI270422v1, chrUn_KI270423v1,
chrUn_KI270425v1, chrUn_KI270429v1, chrUn_KI270442v1, chrUn_KI270466v1,
chrUn_KI270465v1, chrUn_KI270467v1, chrUn_KI270435v1, chrUn_KI270438v1,
chrUn_KI270468v1, chrUn_KI270510v1, chrUn_KI270509v1, chrUn_KI270518v1,
chrUn_KI270508v1, chrUn_KI270516v1, chrUn_KI270512v1, chrUn_KI270519v1,
chrUn_KI270522v1, chrUn_KI270511v1, chrUn_KI270515v1, chrUn_KI270507v1,
```

```
chrUn_KI270517v1, chrUn_KI270529v1, chrUn_KI270528v1, chrUn_KI270530v1,
chrUn_KI270539v1, chrUn_KI270538v1, chrUn_KI270544v1, chrUn_KI270548v1,
chrUn_KI270583v1, chrUn_KI270587v1, chrUn_KI270580v1, chrUn_KI270581v1,
chrUn_KI270579v1, chrUn_KI270589v1, chrUn_KI270590v1, chrUn_KI270584v1,
chrUn_KI270582v1, chrUn_KI270588v1, chrUn_KI270593v1, chrUn_KI270591v1,
chrUn_KI270330v1, chrUn_KI270329v1, chrUn_KI270334v1, chrUn_KI270333v1,
chrUn_KI270335v1, chrUn_KI270338v1, chrUn_KI270340v1, chrUn_KI270336v1,
chrUn_KI270337v1, chrUn_KI270363v1, chrUn_KI270364v1, chrUn_KI270362v1,
chrUn_KI270366v1, chrUn_KI270378v1, chrUn_KI270379v1, chrUn_KI270389v1,
chrUn_KI270390v1, chrUn_KI270387v1, chrUn_KI270395v1, chrUn_KI270396v1,
chrUn_KI270388v1, chrUn_KI270394v1, chrUn_KI270386v1, chrUn_KI270391v1,
chrUn_KI270383v1, chrUn_KI270393v1, chrUn_KI270384v1, chrUn_KI270392v1,
chrUn_KI270381v1, chrUn_KI270385v1, chrUn_KI270382v1, chrUn_KI270376v1,
chrUn_KI270374v1, chrUn_KI270372v1, chrUn_KI270373v1, chrUn_KI270375v1,
chrUn_KI270371v1, chrUn_KI270448v1, chrUn_KI270521v1, chrUn_GL000195v1,
chrUn_GL000219v1, chrUn_GL000220v1, chrUn_GL000224v1, chrUn_KI270741v1,
chrUn_GL000226v1, chrUn_GL000213v1, chrUn_KI270743v1, chrUn_KI270744v1,
chrUn_KI270745v1, chrUn_KI270746v1, chrUn_KI270747v1, chrUn_KI270748v1,
chrUn_KI270749v1, chrUn_KI270750v1, chrUn_KI270751v1, chrUn_KI270752v1,
chrUn_KI270753v1, chrUn_KI270754v1, chrUn_KI270755v1, chrUn_KI270756v1,
chrUn_KI270757v1, chrUn_GL000214v1, chrUn_KI270742v1, chrUn_GL000216v2,
chrUn_GL000218v1, chrX, chrY, chrY_KI270740v1_random, chr1_KQ031383v1_fix,
chr1_KQ983255v1_alt, chr1_KN538361v1_fix, chr1_KQ458383v1_alt,
chr1_KN196473v1_fix, chr1_KZ208904v1_alt, chr1_KN196472v1_fix,
chr1_KZ208905v1_alt, chr1_KQ458382v1_alt, chr1_KV880763v1_alt,
chr1_KN196474v1_fix, chr1_KN538360v1_fix, chr1_KZ208906v1_fix,
chr1_KQ458384v1_alt, chr2_KQ031384v1_fix, chr2_KZ208907v1_alt,
chr2_KQ983256v1_alt, chr2_KZ208908v1_alt, chr2_KN538363v1_fix,
chr2_KN538362v1_fix, chr3_KV766192v1_fix, chr3_KN196475v1_fix,
chr3_KQ031385v1_fix, chr3_KN538364v1_fix, chr3_KZ208909v1_alt,
chr3_KQ031386v1_fix, chr3_KN196476v1_fix, chr4_KQ090013v1_alt,
chr4_KQ090014v1_alt, chr4_KQ090015v1_alt, chr4_KV766193v1_alt,
chr4_KQ983257v1_fix, chr4_KQ983258v1_alt, chr5_KZ208910v1_alt,
chr5_KN196477v1_alt, chr5_KV575243v1_alt, chr5_KV575244v1_fix,
chr6_KZ208911v1_fix, chr6_KQ090017v1_alt, chr6_KQ031387v1_fix,
chr6_KN196478v1_fix, chr6_KQ090016v1_fix, chr6_KV766194v1_fix,
chr7_KV880764v1_fix, chr7_KV880765v1_fix, chr7_KZ208912v1_fix,
chr7_KZ208913v1_alt, chr7_KQ031388v1_fix, chr8_KZ208915v1_fix,
chr8_KV880767v1_fix, chr8_KV880766v1_fix, chr8_KZ208914v1_fix,
chr9_KQ090018v1_alt, chr9_KQ090019v1_alt, chr9_KN196479v1_fix,
chr10_KN538367v1_fix, chr10_KQ090020v1_alt, chr10_KN196480v1_fix,
chr10_KQ090021v1_fix, chr10_KN538366v1_fix, chr10_KN538365v1_fix,
chr11_KQ759759v1_fix, chr11_KN538368v1_alt, chr11_KV766195v1_fix,
chr11_KQ090022v1_fix, chr11_KN196481v1_fix, chr12_KQ090023v1_alt,
chr12_KZ208916v1_fix, chr12_KN538369v1_fix, chr12_KN196482v1_fix,
chr12_KZ208918v1_alt, chr12_KQ759760v1_fix, chr12_KZ208917v1_fix,
chr12_KN538370v1_fix, chr13_KN538372v1_fix, chr13_KQ090024v1_alt,
chr13_KN196483v1_fix, chr13_KN538373v1_fix, chr13_KQ090025v1_alt,
chr13_KN538371v1_fix, chr14_KZ208920v1_fix, chr14_KZ208919v1_alt,
```

```
chr15_KN538374v1_fix, chr15_KQ031389v1_alt, chr16_KQ090026v1_alt,
chr16_KV880768v1_fix, chr16_KQ090027v1_alt, chr16_KZ208921v1_alt,
chr16_KQ031390v1_alt, chr17_KV766196v1_fix, chr17_KV575245v1_fix,
chr17_KV766198v1_alt, chr17_KV766197v1_alt, chr18_KQ458385v1_alt,
chr18_KQ090028v1_fix, chr18_KZ208922v1_fix, chr19_KQ458386v1_fix,
chr19_KN196484v1_fix, chr19_KV575246v1_alt, chr19_KV575247v1_alt,
chr19_KV575248v1_alt, chr19_KV575249v1_alt, chr19_KV575250v1_alt,
chr19_KV575251v1_alt, chr19_KV575252v1_alt, chr19_KV575253v1_alt,
chr19_KV575254v1_alt, chr19_KV575255v1_alt, chr19_KV575256v1_alt,
chr19_KV575257v1_alt, chr19_KV575259v1_alt, chr19_KV575260v1_alt,
chr19_KV575258v1_alt, chr22_KN196485v1_alt, chr22_KQ458387v1_alt,
chr22_KQ458388v1_alt, chr22_KN196486v1_alt, chr22_KQ759761v1_alt,
chr22_KQ759762v1_fix, chrX_KV766199v1_alt, chrY_KZ208923v1_fix,
chrY_KZ208924v1_fix, chrY_KN196487v1_fix, chr1_KZ559100v1_fix,
chr3_KZ559104v1_fix, chr3_KZ559105v1_alt, chr3_KZ559103v1_alt,
chr3_KZ559102v1_alt, chr3_KZ559101v1_alt, chr7_KZ559106v1_alt,
chr8_KZ559107v1_alt, chr11_KZ559109v1_fix, chr11_KZ559108v1_fix,
chr11_KZ559111v1_alt, chr11_KZ559110v1_alt, chr12_KZ559112v1_alt,
chr16_KZ559113v1_fix, chr17_KZ559114v1_alt, chr18_KZ559116v1_alt,
chr18_KZ559115v1_fix]
  reads contigs = [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17,
18, 19, 20, 21, 22, X, Y, MT, GL000207.1, GL000226.1, GL000229.1, GL000231.1,
GL000210.1, GL000239.1, GL000235.1, GL000201.1, GL000247.1, GL000245.1,
GL000197.1, GL000203.1, GL000246.1, GL000249.1, GL000196.1, GL000248.1,
GL000244.1, GL000238.1, GL000202.1, GL000234.1, GL000232.1, GL000206.1,
GL000240.1, GL000236.1, GL000241.1, GL000243.1, GL000242.1, GL000230.1,
GL000237.1, GL000233.1, GL000204.1, GL000198.1, GL000208.1, GL000191.1,
GL000227.1, GL000228.1, GL000214.1, GL000221.1, GL000209.1, GL000218.1,
GL000220.1, GL000213.1, GL000211.1, GL000199.1, GL000217.1, GL000216.1,
GL000215.1, GL000205.1, GL000219.1, GL000224.1, GL000223.1, GL000195.1,
GL000212.1, GL000222.1, GL000200.1, GL000193.1, GL000194.1, GL000225.1,
GL000192.1, NC_007605, hs37d5]

************************************************************************
Set the system property GATK_STACKTRACE_ON_USER_EXCEPTION (--java-options '-
DGATK_STACKTRACE_ON_USER_EXCEPTION=true') to print the stack trace.
```

Use human_g1k_v37.fasta as reference genome

```
/data/mouse/human_rme/genome_index/gatk-4.1.9.0/gatk --java-options "-Xmx4g"
HaplotypeCaller  \
-R /data/mouse/human_rme/genome_index/human_g1k_v37.fasta \
-I HG00421.mapped.ILLUMINA.bwa.CHS.low_coverage.20130415.bam \
-O /data/mouse/human_rme/SH007/HG00421.vcf.gz \
-ERC GVCF
```

error report

```
A USER ERROR has occurred: Traversal by intervals was requested but some input
files are not indexed.
Please index all input files:

samtools index
/data/mouse/human_rme/SH007/HG00421.mapped.ILLUMINA.bwa.CHS.low_coverage.201304
15.bam
```

then try to index by samtools

```
samtools index
/data/mouse/human_rme/SH007/HG00421.mapped.ILLUMINA.bwa.CHS.low_coverage.201304
15.bam
```

Error report

```
[W::bam_hdr_read] EOF marker is absent. The input is probably truncated
[E::bgzf_read_block] Failed to read BGZF block data at offset 4612181361
expected 18560 bytes; hread returned 12925
[E::bgzf_read] Read block operation failed with error 4 after 0 of 4 bytes
samtools index: failed to create index for
"/data/mouse/human_rme/SH007/HG00421.mapped.ILLUMINA.bwa.CHS.low_coverage.20130
415.bam": No such file or directory
```

Guess: bam file is probably truncated, check bam file

```
tail HG00421.mapped.ILLUMINA.bwa.CHS.low_coverage.20130415.bam | hexdump -C
```

return

```
00000000  c1 c5 58 42 f1 3a a6 89  41 9b 13 31 bb da cd 6a  |..XB.:..A..1...j|
00000010  eb 60 5d cc 8a 1e 33 ec  cf cd 36 53 54 e0 ac c1  |.`]...3...6ST...|
00000020  cd 6f 09 1c f8 32 db c1  99 6d 91 c6 02 b2 b2 55  |.o...2...m.....U|
00000030  64 ab 0e 63 25 d2 70 d3  61 19 37 84 d8 b7 d9 8d  |d..c%.p.a.7.....|
00000040  5a b8 02 f1 31 fe 33 3f  07 68 e7 79 94 2b ff f6  |Z...1.3?.h.y.+..|
00000050  8a a2 03 2a 47 e8 45 85  9d 79 39 8e 4b 8e fe 91  |...*G.E..y9.K...|
00000060  e9 74 72 6e cf 35 bd 3e  1f 9c 3b a3 fe 00 98 43  |.trn.5.>..;....C|
00000070  97 69 bb 07 12 40 64 98  2e 00 70 59 b6 8a db 86  |.i...@d...pY....|
00000080  0a 65 bb 34 c4 8a a1 e9  34 6b 9c 36 e3 45 22 b6  |.e.4....4k.6.E".|
00000090  c6 83 d9 40 a9 32 5d 08  1c 07 40 8d 3a 05 17 4d  |...@.2]...@.:..M|
000000a0  d1 14 a8 18 16 96 e1 f7  ae ef aa e9 49 f4 f6 77  |............I..w|
000000b0  56 a4 65 11 e8 c5 c9 ee  79 39 8e eb 9a 33 77 22  |V.e.....y9...3w"|
000000c0  e4 5c cc a5 44 43 cc 0b  88 db 1d 00 99 18 bd 2d  |.\..DC.........-|
000000d0  13 1b db 02 85 0b 4f b1  a1 c4 f6 a3 27 cc 76 eb  |......O.....'.v.|
000000e0  59 49 4a cd 70 87 93 15  38 10 01 44 7f 5c e0 45  |YIJ.p...8..D.\.E|
000000f0  53 5a f9 08 b3 14 c0 bb  e4 da 96 85 27 6b 58 be  |SZ.........'kX.|
00000100  3e 47 71 27 de f6 b3 ef  4c 27 b0 26 de b6 dd 84  |>Gq'....L'.&....|
```

```
00000110  38 b9 51 1b af 61 b9 0a  ce 0c 13 06 46 c5 39 8e  |8.Q..a......F.9.|
00000120  35 b1 24 2a 2b 94 ea 06  17 9a 98 28 d2 81 77 48  |5.$*+......(..wH|
00000130  3d 57 88 40 a4 79 96 f3  7c 0f 62 05 4a e2 44 c7  |=W.@.y..|.b.J.D.|
00000140  34 a4 0e d9 df 0b 0a 49  18 07 7c cf f5 fd d2 3c  |4......I..|....<|
00000150  ab ba a2 80 f2 fb 8a d5  24 d2 fd e5 15 19 1c 94  |........$.......|
00000160  33 59 dc 55 38 95 2e c1  15 2c 28 fc d4 35 d7 e1  |3Y.U8....,(..5..|
00000170  01 23 b0 ce 4b d9 92 08  71 69 6c 82 1b 41 22 88  |.#..K...qil..A".|
00000180  74 51 15 8a 07 52 f9 ea  28 94 05 3f 0c b0 f0 ce  |tQ...R..(..?....|
00000190  45 6c 94 44 59 65 42 12  85 a2 df 9e 2e 45 52 15  |El.DYeB......ER.|
000001a0  88 17 72 e6 1d 6c dd 99  9f 7d 45 61 a8 56 f3 fe  |..r..l...}Ea.V..|
000001b0  6a 7d 87 e7 1a e9 fe ca  0a 77 8b 5a 26 72 51 76  |j}.......w.Z&rQv|
000001c0  72 5e b7 e4 ca 77 bf ec  19 08 d3 cc a8 1f 45 0e  |r^...w........E.|
000001d0  ca 6a 01 85 f2 0b 45 08  b4 24 2b d0 58 c4 a2 e8  |.j....E..$+.X...|
000001e0  c6 64 cb f3 85 36 41 b8  0d cf 00 d2 89 00 8e e6  |.d...6A.........|
000001f0  ef 72 16 e1 f6 e8 84 fb  be 1f 45 c5 62 d1 09 0b  |.r........E.b...|
00000200  3b bb 9e 50 14 42 8c 4d  96 bf 41 e9 dd bb 91 88  |;..P.B.M..A.....|
00000210  c2 dd 5d 42 bd 48 f1 b5  9d 09 9a 8c 75 a3 1c 1f  |..]B.H......u...|
00000220  4e 50 dc 9a 33 6d 94 ed  23 8b 01 d8 44 a2 ef 47  |NP..3m..#...D..G|
00000230  6f 6d 6c fc 02 90 2e 07  27 5a 46 c3 fe 68 30 0e  |oml.....'ZF..h0.|
00000240  75 73 7d 38 bf 4c 8f c3  9f 51 87 f5 c6 e7 5d 0f  |us}8.L...Q....].|
00000250  ce bb f3 c1 e4 bc eb 3f  21 07 a3 fb 2a 74 c4 0b  |.......?!...*t..|
00000260  85 00 4e 47 11 ed 6e 17  b6 44 bc 15 3b 5b 3b 9b  |..NG..n..D..;[;.|
00000270  f0 d8 75 24 77 15 f3 fd  e0 ee 7e 29 a4 39 30 eb  |..u$w.....~).90.|
00000280  0f 20 d8 cd 07 ed 7b 27  0f 4e ee 55 2f 66 28 71  |. ....{'.N.U/f(q|
00000290  88 e2 f6 dc 10 c5 2b cf  46 e6 7f 06 fa af 3d bc  |......+.F.....=.|
000002a0  28 16 f3 a3 ef 2c 8a 16  34 20 18 6b b4 4e 9a 67  |(....,..4 .k.N.g|
000002b0  1b 27 77 32 bc 1f c9 21  7a f3 1c a2 2b bb 83 61  |.'w2...!z...+..a|
000002c0  ae 0b de a4 d7 c9 6d 87  77 dd cd 1b 47 bb 31 67  |......m.w...G.1g|
000002d0  b9 98 c5 07 71 a9 78 58  22 bb 3b 81 5f 88 f8 81  |....q.xX".;._...|
000002e0  47 92 42 71 87 45 31 c0  aa b7 5c ae 76 b9 b8 01  |G.Bq.E1...\.v...|
000002f0  a1 05 b0 bc 30 74 e2 03  6c 26 01 5f 9e 73 e0 3c  |....0t..l&._.s.<|
00000300  8b 25 7d a1 74 70 e7 c8  15 b1 24 a2 28 19 1c 9f  |.%}.tp....$.(...|
00000310  4c 92 85 c0 e2 75 bb e5  fb 41 ad 75 bf b6 1e 52  |L....u...A.u...R|
00000320  48 48 5c a6 0b 03 be 39  d7 07 35 39 f0 50 cc af  |HH\....9..59.P..|
00000330  b0 10 01 4f 84 fb f0 bc  bb 9a 33 d9 cd 47 1c f0  |...O......3..G..|
00000340  3c d8 73 5c 56 88 14 d9  f3 75 24 72 81 70 9c 2d  |<.s\V....u$r.p.-|
00000350  a2 a4 91 c2 8f 4a 48 bd  8a 3a f5 f7 62 e6 8b 94  |.....JH..:..b...|
00000360  ee 0a d7 a3 5b 41 bc b5  e7 09 f0 40 94 14 73 3e  |....[A.....@..s>|
00000370  09 c1 69 3f cf dd 0a 9d  19 07 fb e6 0a 0e f6 c5  |..i?............|
00000380  0b 4e 61 22 66 64 c5 95  af e3 77 87 3d c7 e9 31  |.Na"fd....w.=..1|
00000390  26 34 66 b7 44 c2 50 46  80 a9 58 94 b2 85 8e d8  |&4f.D.PF..X.....|
000003a0  cd 64 58 1c 10 29 ac d4  11 4a d0 2a ec 3a 60 b6  |.dX..)...J.*.:`.|
000003b0  af 29 29 27 2c 10 28 b0  01 d1 b1 4c 7c cc 4a a2  |.))',.(....L|.J.|
000003c0  f2 a7 b0 5d ea 89 48 6d  07 02 c4 d1 97 35 f7 af  |...]..Hm.....5..|
000003d0  47 9a f1 b7 57 64 bf 50  d0 64 49 5c 79 5e bc e4  |G...Wd.P.dI\y^..|
000003e0  ca f8 62 4f aa 93 77 87  1d 81 8d cf 38 c6 63 a7  |..bO..w.....8.c.|
000003f0  a3 6c a1 b2 6c 7b bd b0  0f 21 d3 58 4f 6d 89 d2  |.l..l{...!.XOm..|
00000400  58 d9 01 9c 59 01 a6 cb  b3 65 b1 65 0c 2e 2a c7  |X...Y....e.e..*.|
00000410  c7 38 eb 32 d9 27 3b 99  a6 ca 76 d8 e0 5a df b5  |.8.2.';...v..Z..|
```

```
00000420  9d a6 fa d2 8d 99 b4 c0  32 d0 28 41 20 16 9b 4d  |........2.(A ..M|
00000430  27 72 03 b6 d9 74 98 b9  d3 de cc 9d 76 7a d8 81  |'r...t......vz..|
00000440  da e9 b0 51 cf e9 e0 70  77 7f 34 34 03 d3 e1 72  |...Q...pw.44...r|
00000450  c8 fa 5d 87 71 5e a2 31  f1 71 6d 15 c4 65 01 86  |..].q^.1.qm..e..|
00000460  0d d8 ad c7 53 23 8d dd  aa 8b 18 26 00 b2 6d d5  |....S#.....&..m.|
00000470  b3 43 6c b8 e2 02 30 56  d8 a3 5a b6 ab 84 50 4f  |.Cl...0V..Z...PO|
00000480  19 c7 81 34 70 66 0c 8c  51 5a 39 b1 55 67 bd 28  |...4pf..QZ9.Ug.(|
00000490  44 5b bf 77 d6 ba 52 57  cd 47 3e f2 91 a7 8f f4  |D[.w..RW.G>.....|
000004a0  c4 a4 e5 db 17 47 5f 7f  fe 9d e5 55 41 cd 56 bd  |.....G_....UA.V.|
000004b0  56 ad 9d 8d 9b f3 1e 6d  bb 80 33 7c 3b 1a 75 47  |V......m..3|;.uG|
000004c0  7c 98 e7 80 f1 90 75 4c  77 14 62 9f 02 a6 14 88  ||.....uLw.b.....|
000004d0  c6 cc 39 e6 1e ca 76 9d  12 6e 2a 4e 55 12 43 50  |..9...v..n*NU.CP|
000004e0  0c 4e 05 f7 8a 61 24 8c  95 76 34 75 cd 13 c1 29  |.N...a$..v4u...)|
000004f0  4e 14 49 df e3 10 b5 01  43 66 56 d0 52 79 4c 65  |N.I.....CfV.RyLe|
00000500  8b 85 ca 29 8b 50 ac e7  12 c1 b9 e7 39 10 b0 0a  |...).P......9...|
00000510  e8 99 8f c6 01 ef 0f 2e  49 95 cc 0d 73 5f 1f 68  |........I...s_.h|
00000520  6d f7 2f a5 29 b3 b6 89  ed 0b 65 74 af a8 22 95  |m./.).....et.."."|
00000530  b5 3c d8 64 50 c5 b6 96  e2 c7 6c 7a db 26 29 92  |.<.dP.....lz.&).|
00000540  2c 92 c6 b6 a6 6c 08 ce  a0 5c 3e b5 5d 10 da 2e  |,....l...\>.]...|
00000550  e3 b5 ed a6 fc 12 dd ca  f5 68 37 fd ae b7 2f d6  |.........h7.../.|
00000560  82 7e fd 9d 65 a9 92 7a  f5 b4 7a d6 a8 8f 4f c3  |.~..e..z..z...O.|
00000570  2b fb 8c fc 10 57 f5 76  fa 21 41 59 76 ad 43 a2  |+....W.v.!AYv.C.|
00000580  70 be d8 76 e5 00 58 dc  6a 71 05 dc 58 f5 2e 5c  |p..v..X.jq..X..\|
00000590  8c 25 50 29 5f a6 4a 85  12 eb 15 26 64 d2 30 14  |.%P)_.J....&d.0.|
000005a0  0c b7 63 56 60 c8 a8 62  80 89 0c 9f 69 db 82 aa  |..cV`..b....i...|
000005b0  a5 56 a9 cc 0b b6 b6 02  a1 13 a0 1b 2b 5c c7 5f  |.V..........+\._|
000005c0  7c 67 59 f1 b6 de 3c ad  d7 cf c6 a3 6e 57 b7 64  ||gY...<.....nW.d|
000005d0  86 6b b2 4c ae 17 71 6e  db 77 8c 5d 2f 62 a5 ab  |.k.L..qn.w.]/b..|
000005e0  33 ef 2a 75 2a 6c d3 3a  6e 8f 4d 15 b3 da e0 d8  |3.*u*l.:n.M.....|
000005f0  ad 43 68 56 4e 4b c7 83  70 01 45 7c a5 cc 96 6b  |.ChVNK..p.E|...k|
00000600  80 a9 0b 1c 30 c6 46 00  15 70 dc c2 b1 9e 52 5e  |....0.F..p....R^|
00000610  f7 7a 53 a0 db 2b 2c fa  e7 2f 16 36 4e da 67 a7  |.zS..+,../.6N.g.|
00000620  b5 7a e6 3a ae 65 cc dd  73 a7 d3 0d 85 e2 a8 0e  |.z.:.e..s.......|
00000630  8c 4d 26 06 05 d1 62 b0  6f 29 b0 6f 11 55 70 13  |.M&...b.o).o.Up.|
00000640  eb 77 c5 58 e8 a1 5c f6  89 56 e3 5c 27 ae 32 4e  |.w.X..\..V.\'.2N|
00000650  a3 08 65 d1 0c b3 5b 35  70 f7 37 b3 63 dc f0 2b  |..e...[5p.7.c..+|
00000660  18 c7 82 73 a2 04 5f 4f  a0 5f 9c 25 8d bf 7b 85  |...s.._O._.%..{.|
00000670  45 ff 0c 00 fd 4b 4b 65  e4 b3 56 eb 64 5c 46 be  |E....KKe..V.d\F.|
00000680  2e cc 79 16 31 2c 14 63  43 19 90 08 6c 41 07 18  |..y.1,.cC...lA..|
00000690  2b 70 f2 d9 45 d2 76 7b  11 a0 4a 8d 04 9a 17 33  |+p..E.v{..J....3|
000006a0  6d 67 61 70 0e 00 d0 2c  53 4c 7a 26 40 10 31 09  |mgap...,SLz&@.1.|
000006b0  8a 29 7b 6e b2 bc 73 aa  65 b9 9c 09 b9 db f1 8c  |.){n..s.e.......|
000006c0  75 07 ba b3 e2 30 44 11  0e b3 74 18 ce 09 6e 5c  |u....0D...t...n\|
000006d0  13 65 70 27 d9 38 dc 78  96 c8 a6 df 11 1c 5c 09  |.ep'.8.x......\.|
000006e0  69 8b 18 65 7b 40 1e 57  80 fd 99 63 7b 56 aa 0a  |i..e{@.W...c{V..|
000006f0  5a ad 55 0c c4 d1 22 6c  f3 c5 45 bd 49 36 2d 0e  |Z.U..."l..E.I6-.|
00000700  8c 5b e9 d4 2e 81 c4 a5  5e 98 ad 8f d6 74 c8 a5  |.[......^....t..|
00000710  3b 1d 09 f8 ec db 17 55  00 51 a4 ce 5b 8c 58 26  |;.......U.Q..[.X&|
00000720  82 74 ce ed ac 46 b7 32  48 31 17 bc 76 7f e6 b5  |.t...F.2H1..v...|
```

```
00000730  9f 30 55 b4 f4 88 02 9c  46 2c e5 ef 6e b9 e4 ee  |.0U.....F,..n...|
00000740  91 bf                                            |..|
00000742
```

seems 28 byte empty BGZF block as an EOF marker is absent

Try

```
samtools view -b -q 20
HG00421.mapped.ILLUMINA.bwa.CHS.low_coverage.20130415.bam >
HG00421.mapped.ILLUMINA.bwa.CHS.low_coverage.20130415.new.bam
```

return

```
[W::bam_hdr_read] EOF marker is absent. The input is probably truncated
[E::bgzf_read_block] Failed to read BGZF block data at offset 4612181361
expected 18560 bytes; hread returned 12925
[E::bgzf_read] Read block operation failed with error 4 after 0 of 4 bytes
[main_samview] truncated file.
```

then try to use HG00421.mapped.ILLUMINA.bwa.CHS.low_coverage.20130415.new.bam by samtools index:

```
samtools index HG00421.mapped.ILLUMINA.bwa.CHS.low_coverage.20130415.new.bam
```

(No error return!)

try GATK

```
/data/mouse/human_rme/genome_index/gatk-4.1.9.0/gatk --java-options "-Xmx4g"
HaplotypeCaller  \
-R /data/mouse/human_rme/genome_index/human_g1k_v37.fasta \
-I HG00421.mapped.ILLUMINA.bwa.CHS.low_coverage.20130415.new.bam \
-O /data/mouse/human_rme/SH007/HG00421_new.vcf.gz \
-ERC GVCF
```

mapping by STAR (没有gtf文件时的做法：)

```
STAR --runThreadN 6 --runMode genomeGenerate \
--genomeDir /data/mouse/human_rme/genome_index \
--genomeFastaFiles /data/mouse/human_rme/genome_index/human_g1k_v37.fasta
#generate index

STAR --genomeDir /data/mouse/human_rme/genome_index \
    --runThreadN 20 \
    --readFilesIn sample_r1.fq.gz sample_r2.fq.gz \
    # --readFilesCommand zcat \
    --outFileNamePrefix sample \
    --outSAMtype BAM SortedByCoordinate \
    --outBAMsortingThreadN 10 #do mapping
```

12.2 进度：

SH007

genotypeGVCF done HG00422 (5110)

genotypeGVCF HG00421 ready

```
#GATK genotypeGVCF
/data/mouse/human_rme/genome_index/gatk-4.1.9.0/gatk --java-options "-Xmx4g"
GenotypeGVCFs \
-R /data/mouse/human_rme/genome_index/human_g1k_v37.fasta \
-V HG00422_new.vcf.gz \
-O HG00422_genotypeGVCF.vcf.gz
```

HG00423 can't do gatk missing RG

```
samtools view -H HG00444Aligned.sortedByCoord.out.bam | grep '^@RG'
```

return : none!

```
java -jar /data/mouse/human_rme/genome_index/picard.jar
```

try AddOrReplaceReadGroups

```
java -jar /data/mouse/human_rme/genome_index/picard.jar AddOrReplaceReadGroups
\
I=HG00423Aligned.sortedByCoord.out.bam \
O=HG00423Aligned.sortedByCoord.out.new.bam \
RGID=4 \
RGLB=lib1 \
RGPL=illumina \
RGPU=unit1 \
RGSM=20
```

then try samtools index

```
samtools index HG00423Aligned.sortedByCoord.out.new.bam
```

(it seems work!!!)

gatk doing HG00423 65457 (shuting down)

```
/data/mouse/human_rme/genome_index/gatk-4.1.9.0/gatk --java-options "-Xmx4g"
HaplotypeCaller  \
-R /data/mouse/human_rme/genome_index/human_g1k_v37.fasta \
-I HG00423Aligned.sortedByCoord.out.new.bam \
-O /data/mouse/human_rme/SH007/HG00423_new.vcf.gz \
-ERC GVCF

# return
Using GATK jar /data/mouse/human_rme/genome_index/gatk-4.1.9.0/gatk-package-
4.1.9.0-local.jar
Running:
    java -Dsamjdk.use_async_io_read_samtools=false -
Dsamjdk.use_async_io_write_samtools=true -
Dsamjdk.use_async_io_write_tribble=false -Dsamjdk.compression_level=2 -Xmx4g -
jar /data/mouse/human_rme/genome_index/gatk-4.1.9.0/gatk-package-4.1.9.0-
local.jar HaplotypeCaller -R
/data/mouse/human_rme/genome_index/human_g1k_v37.fasta -I
HG00423Aligned.sortedByCoord.out.new.bam -O
/data/mouse/human_rme/SH007/HG00423_new.vcf.gz -ERC GVCF
15:21:05.494 INFO  NativeLibraryLoader - Loading libgkl_compression.so from
jar:file:/data/mouse/human_rme/genome_index/gatk-4.1.9.0/gatk-package-4.1.9.0-
local.jar!/com/intel/gkl/native/libgkl_compression.so
Dec 02, 2020 3:21:05 PM
shaded.cloud_nio.com.google.auth.oauth2.ComputeEngineCredentials
runningOnComputeEngine
INFO: Failed to detect whether we are running on Google Compute Engine.
15:21:05.775 INFO  HaplotypeCaller - ----------------------------------------
------------------
15:21:05.776 INFO  HaplotypeCaller - The Genome Analysis Toolkit (GATK)
v4.1.9.0
```

```
15:21:05.776 INFO  HaplotypeCaller - For support and documentation go to
https://software.broadinstitute.org/gatk/
15:21:05.832 INFO  HaplotypeCaller - Executing as zhuyufei@bdp-svr05 on Linux
v5.3.1-1.el7.elrepo.x86_64 amd64
15:21:05.832 INFO  HaplotypeCaller - Java runtime: OpenJDK 64-Bit Server VM
v1.8.0_152-release-1056-b12
15:21:05.832 INFO  HaplotypeCaller - Start Date/Time: 2020年12月2日 下午03时21分
05秒
15:21:05.832 INFO  HaplotypeCaller - ----------------------------------------
-----------------
15:21:05.832 INFO  HaplotypeCaller - ----------------------------------------
-----------------
15:21:05.833 INFO  HaplotypeCaller - HTSJDK Version: 2.23.0
15:21:05.834 INFO  HaplotypeCaller - Picard Version: 2.23.3
15:21:05.834 INFO  HaplotypeCaller - HTSJDK Defaults.COMPRESSION_LEVEL : 2
15:21:05.834 INFO  HaplotypeCaller - HTSJDK
Defaults.USE_ASYNC_IO_READ_FOR_SAMTOOLS : false
15:21:05.834 INFO  HaplotypeCaller - HTSJDK
Defaults.USE_ASYNC_IO_WRITE_FOR_SAMTOOLS : true
15:21:05.834 INFO  HaplotypeCaller - HTSJDK
Defaults.USE_ASYNC_IO_WRITE_FOR_TRIBBLE : false
15:21:05.834 INFO  HaplotypeCaller - Deflater: IntelDeflater
15:21:05.834 INFO  HaplotypeCaller - Inflater: IntelInflater
15:21:05.834 INFO  HaplotypeCaller - GCS max retries/reopens: 20
15:21:05.835 INFO  HaplotypeCaller - Requester pays: disabled
15:21:05.835 INFO  HaplotypeCaller - Initializing engine
15:21:06.227 INFO  HaplotypeCaller - Done initializing engine
15:21:06.228 INFO  HaplotypeCallerEngine - Tool is in reference confidence mode
and the annotation, the following changes will be made to any specified
annotations: 'StrandBiasBySample' will be enabled. 'ChromosomeCounts',
'FisherStrand', 'StrandOddsRatio' and 'QualByDepth' annotations have been
disabled
15:21:06.235 INFO  HaplotypeCallerEngine - Standard Emitting and Calling
confidence set to 0.0 for reference-model confidence output
15:21:06.235 INFO  HaplotypeCallerEngine - All sites annotated with PLs forced
to true for reference-model confidence output
15:21:06.248 INFO  NativeLibraryLoader - Loading libgkl_utils.so from
jar:file:/data/mouse/human_rme/genome_index/gatk-4.1.9.0/gatk-package-4.1.9.0-
local.jar!/com/intel/gkl/native/libgkl_utils.so
15:21:06.249 INFO  NativeLibraryLoader - Loading libgkl_pairhmm_omp.so from
jar:file:/data/mouse/human_rme/genome_index/gatk-4.1.9.0/gatk-package-4.1.9.0-
local.jar!/com/intel/gkl/native/libgkl_pairhmm_omp.so
15:21:06.301 INFO  IntelPairHmm - Using CPU-supported AVX-512 instructions
15:21:06.301 INFO  IntelPairHmm - Flush-to-zero (FTZ) is enabled when running
PairHMM
15:21:06.301 INFO  IntelPairHmm - Available threads: 64
15:21:06.301 INFO  IntelPairHmm - Requested threads: 4
15:21:06.301 INFO  PairHMM - Using the OpenMP multi-threaded AVX-accelerated
native PairHMM implementation
```

```
15:21:06.336 INFO  ProgressMeter - Starting traversal
15:21:06.337 INFO  ProgressMeter -        Current Locus  Elapsed Minutes
Regions Processed   Regions/Minute
15:21:16.341 INFO  ProgressMeter -           1:11753701              0.2
          39180         235056.5
15:21:26.339 INFO  ProgressMeter -           1:28001701              0.3
          93340         279992.0
15:21:36.340 INFO  ProgressMeter -           1:42413701              0.5
         141380         282731.7
15:21:46.341 INFO  ProgressMeter -           1:57197701              0.7
         190660         285961.4
15:21:59.455 INFO  ProgressMeter -           1:70118701              0.9
         233730         264017.2
15:22:09.454 INFO  ProgressMeter -           1:84668701              1.1
         282230         268292.2
15:22:19.454 INFO  ProgressMeter -           1:99485701              1.2
         331620         272128.2
15:22:29.454 INFO  ProgressMeter -          1:111821701              1.4
         372740         269071.3
15:22:39.455 INFO  ProgressMeter -          1:125609701              1.6
         418700         269786.7
15:22:49.617 INFO  ProgressMeter -          1:139055701              1.7
         463520         269279.6
15:22:59.643 INFO  ProgressMeter -          1:153017701              1.9
         510060         270106.4
15:23:09.768 INFO  ProgressMeter -          1:163760701              2.1
         545870         265350.4
15:23:19.767 INFO  ProgressMeter -          1:176396701              2.2
         587990         264403.8
15:23:29.983 INFO  ProgressMeter -          1:188480701              2.4
         628270         262424.3
15:23:39.984 INFO  ProgressMeter -          1:201497701              2.6
         671660         262288.6
15:23:50.582 INFO  ProgressMeter -          1:213197701              2.7
         710660         259609.7
15:24:00.582 INFO  ProgressMeter -          1:225782701              2.9
         752610         259155.8
15:24:10.583 INFO  ProgressMeter -          1:236975701              3.1
         789920         257240.1
15:24:20.582 INFO  ProgressMeter -          1:247268701              3.2
         824230         254595.0
15:24:46.185 INFO  HaplotypeCaller - Shutting down engine
[2020年12月2日 下午03时24分46秒]
org.broadinstitute.hellbender.tools.walkers.haplotypecaller.HaplotypeCaller
done. Elapsed time: 3.68 minutes.
Runtime.totalMemory()=4291821568
Exception in thread "main" java.lang.OutOfMemoryError: Java heap space
        at java.util.stream.Nodes$DoubleArrayNode.<init>(Nodes.java:1429)
```

```
        at java.util.stream.Nodes$DoubleFixedNodeBuilder.<init>
(Nodes.java:1589)
        at java.util.stream.Nodes.doubleBuilder(Nodes.java:279)
        at
java.util.stream.DoublePipeline.makeNodeBuilder(DoublePipeline.java:164)
        at
java.util.stream.AbstractPipeline.evaluate(AbstractPipeline.java:543)
        at
java.util.stream.AbstractPipeline.evaluateToArrayNode(AbstractPipeline.java:260
)
        at java.util.stream.DoublePipeline.toArray(DoublePipeline.java:506)
        at
org.broadinstitute.hellbender.utils.MathUtils.median(MathUtils.java:841)
        at
org.broadinstitute.hellbender.utils.variant.writers.GVCFBlock.getMedianDP(GVCFB
lock.java:75)
        at
org.broadinstitute.hellbender.utils.variant.writers.HomRefBlock.createHomRefGen
otype(HomRefBlock.java:73)
        at
org.broadinstitute.hellbender.utils.variant.writers.GVCFBlock.toVariantContext(
GVCFBlock.java:49)
        at
org.broadinstitute.hellbender.utils.variant.writers.GVCFBlockCombiner.emitCurre
ntBlock(GVCFBlockCombiner.java:177)
        at
org.broadinstitute.hellbender.utils.variant.writers.GVCFBlockCombiner.signalEnd
OfInput(GVCFBlockCombiner.java:227)
        at
org.broadinstitute.hellbender.utils.variant.writers.GVCFWriter.close(GVCFWriter
.java:70)
        at
org.broadinstitute.hellbender.tools.walkers.haplotypecaller.HaplotypeCaller.clo
seTool(HaplotypeCaller.java:216)
        at
org.broadinstitute.hellbender.engine.GATKTool.doWork(GATKTool.java:1053)
        at
org.broadinstitute.hellbender.cmdline.CommandLineProgram.runTool(CommandLinePro
gram.java:140)
        at
org.broadinstitute.hellbender.cmdline.CommandLineProgram.instanceMainPostParseA
rgs(CommandLineProgram.java:192)
        at
org.broadinstitute.hellbender.cmdline.CommandLineProgram.instanceMain(CommandLi
neProgram.java:211)
        at
org.broadinstitute.hellbender.Main.runCommandLineProgram(Main.java:160)
        at org.broadinstitute.hellbender.Main.mainEntry(Main.java:203)
        at org.broadinstitute.hellbender.Main.main(Main.java:289)
```

SH014

HG00442 vcf ready

HG00443 vcf ready

gatk HaplotypeCaller doing HG00444 but shutting down (1299)

```
/data/mouse/human_rme/genome_index/gatk-4.1.9.0/gatk --java-options "-Xmx4g"
HaplotypeCaller  \
-R /data/mouse/human_rme/genome_index/human_g1k_v37.fasta \
-I HG00444Aligned.sortedByCoord.out.new.bam \
-O /data/mouse/human_rme/SH014/HG00444_new.vcf.gz \
-ERC GVCF

# return
Using GATK jar /data/mouse/human_rme/genome_index/gatk-4.1.9.0/gatk-package-
4.1.9.0-local.jar
Running:
    java -Dsamjdk.use_async_io_read_samtools=false -
Dsamjdk.use_async_io_write_samtools=true -
Dsamjdk.use_async_io_write_tribble=false -Dsamjdk.compression_level=2 -Xmx4g -
jar /data/mouse/human_rme/genome_index/gatk-4.1.9.0/gatk-package-4.1.9.0-
local.jar HaplotypeCaller -R
/data/mouse/human_rme/genome_index/human_g1k_v37.fasta -I
HG00444Aligned.sortedByCoord.out.new.bam -O
/data/mouse/human_rme/SH014/HG00444_new.vcf.gz -ERC GVCF
15:51:47.944 INFO  NativeLibraryLoader - Loading libgkl_compression.so from
jar:file:/data/mouse/human_rme/genome_index/gatk-4.1.9.0/gatk-package-4.1.9.0-
local.jar!/com/intel/gkl/native/libgkl_compression.so
Dec 02, 2020 3:51:48 PM
shaded.cloud_nio.com.google.auth.oauth2.ComputeEngineCredentials
runningOnComputeEngine
INFO: Failed to detect whether we are running on Google Compute Engine.
15:51:48.116 INFO  HaplotypeCaller - ----------------------------------------
------------------
15:51:48.116 INFO  HaplotypeCaller - The Genome Analysis Toolkit (GATK)
v4.1.9.0
15:51:48.117 INFO  HaplotypeCaller - For support and documentation go to
https://software.broadinstitute.org/gatk/
15:51:48.225 INFO  HaplotypeCaller - Executing as zhuyufei@bdp-svr05 on Linux
v5.3.1-1.el7.elrepo.x86_64 amd64
15:51:48.225 INFO  HaplotypeCaller - Java runtime: OpenJDK 64-Bit Server VM
v1.8.0_152-release-1056-b12
```

```
15:51:48.225 INFO  HaplotypeCaller - Start Date/Time: 2020年12月2日 下午03时51分47秒
15:51:48.225 INFO  HaplotypeCaller - ------------------------------------------------------------
15:51:48.225 INFO  HaplotypeCaller - ------------------------------------------------------------
15:51:48.226 INFO  HaplotypeCaller - HTSJDK Version: 2.23.0
15:51:48.226 INFO  HaplotypeCaller - Picard Version: 2.23.3
15:51:48.227 INFO  HaplotypeCaller - HTSJDK Defaults.COMPRESSION_LEVEL : 2
15:51:48.227 INFO  HaplotypeCaller - HTSJDK Defaults.USE_ASYNC_IO_READ_FOR_SAMTOOLS : false
15:51:48.227 INFO  HaplotypeCaller - HTSJDK Defaults.USE_ASYNC_IO_WRITE_FOR_SAMTOOLS : true
15:51:48.227 INFO  HaplotypeCaller - HTSJDK Defaults.USE_ASYNC_IO_WRITE_FOR_TRIBBLE : false
15:51:48.227 INFO  HaplotypeCaller - Deflater: IntelDeflater
15:51:48.227 INFO  HaplotypeCaller - Inflater: IntelInflater
15:51:48.227 INFO  HaplotypeCaller - GCS max retries/reopens: 20
15:51:48.228 INFO  HaplotypeCaller - Requester pays: disabled
15:51:48.228 INFO  HaplotypeCaller - Initializing engine
15:51:48.606 INFO  HaplotypeCaller - Done initializing engine
15:51:48.608 INFO  HaplotypeCallerEngine - Tool is in reference confidence mode and the annotation, the following changes will be made to any specified annotations: 'StrandBiasBySample' will be enabled. 'ChromosomeCounts', 'FisherStrand', 'StrandOddsRatio' and 'QualByDepth' annotations have been disabled
15:51:48.615 INFO  HaplotypeCallerEngine - Standard Emitting and Calling confidence set to 0.0 for reference-model confidence output
15:51:48.615 INFO  HaplotypeCallerEngine - All sites annotated with PLs forced to true for reference-model confidence output
15:51:48.631 INFO  NativeLibraryLoader - Loading libgkl_utils.so from jar:file:/data/mouse/human_rme/genome_index/gatk-4.1.9.0/gatk-package-4.1.9.0-local.jar!/com/intel/gkl/native/libgkl_utils.so
15:51:48.632 INFO  NativeLibraryLoader - Loading libgkl_pairhmm_omp.so from jar:file:/data/mouse/human_rme/genome_index/gatk-4.1.9.0/gatk-package-4.1.9.0-local.jar!/com/intel/gkl/native/libgkl_pairhmm_omp.so
15:51:48.687 INFO  IntelPairHmm - Using CPU-supported AVX-512 instructions
15:51:48.687 INFO  IntelPairHmm - Flush-to-zero (FTZ) is enabled when running PairHMM
15:51:48.688 INFO  IntelPairHmm - Available threads: 64
15:51:48.688 INFO  IntelPairHmm - Requested threads: 4
15:51:48.688 INFO  PairHMM - Using the OpenMP multi-threaded AVX-accelerated native PairHMM implementation
15:51:48.725 INFO  ProgressMeter - Starting traversal
15:51:48.725 INFO  ProgressMeter -        Current Locus  Elapsed Minutes    Regions Processed   Regions/Minute
15:51:58.727 INFO  ProgressMeter -           1:11600701              0.2          38670         232020.0
```

```
15:52:08.726 INFO  ProgressMeter -           1:27983701              0.3
          93280         279826.0
15:52:18.727 INFO  ProgressMeter -           1:42827701              0.5
         142760         285501.0
15:52:28.747 INFO  ProgressMeter -           1:57899701              0.7
         193000         289340.9
15:52:39.822 INFO  ProgressMeter -           1:70169701              0.9
         233900         274654.1
15:52:49.822 INFO  ProgressMeter -           1:85727701              1.0
         285760         280629.2
15:52:59.823 INFO  ProgressMeter -          1:101075701              1.2
         336920         284328.7
15:53:09.824 INFO  ProgressMeter -          1:113090701              1.4
         376970         278896.2
15:53:19.906 INFO  ProgressMeter -          1:127433701              1.5
         424780         279518.8
15:53:29.910 INFO  ProgressMeter -          1:142109701              1.7
         473700         280899.8
15:53:39.985 INFO  ProgressMeter -          1:155774701              1.9
         519250         280019.8
15:53:49.996 INFO  ProgressMeter -          1:166265701              2.0
         554220         274205.7
15:54:00.097 INFO  ProgressMeter -          1:178859701              2.2
         596200         272297.5
15:54:10.227 INFO  ProgressMeter -          1:191456701              2.4
         638190         270608.7
15:54:20.303 INFO  ProgressMeter -          1:204050701              2.5
         680170         269237.4
15:54:30.432 INFO  ProgressMeter -          1:216647701              2.7
         722160         267951.3
15:54:40.525 INFO  ProgressMeter -          1:229241701              2.9
         764140         266870.8
15:54:50.526 INFO  ProgressMeter -          1:240734701              3.0
         802450         264835.0
15:55:26.303 INFO  HaplotypeCaller - Shutting down engine
[2020年12月2日 下午03时55分26秒]
org.broadinstitute.hellbender.tools.walkers.haplotypecaller.HaplotypeCaller
done. Elapsed time: 3.64 minutes.
Runtime.totalMemory()=4291821568
Exception in thread "main" java.lang.OutOfMemoryError: Java heap space
        at java.util.stream.Nodes$DoubleArrayNode.<init>(Nodes.java:1429)
        at java.util.stream.Nodes$DoubleFixedNodeBuilder.<init>
(Nodes.java:1589)
        at java.util.stream.Nodes.doubleBuilder(Nodes.java:279)
        at
java.util.stream.DoublePipeline.makeNodeBuilder(DoublePipeline.java:164)
        at
java.util.stream.AbstractPipeline.evaluate(AbstractPipeline.java:543)
```

```
        at
java.util.stream.AbstractPipeline.evaluateToArrayNode(AbstractPipeline.java:260
)
        at java.util.stream.DoublePipeline.toArray(DoublePipeline.java:506)
        at
org.broadinstitute.hellbender.utils.MathUtils.median(MathUtils.java:841)
        at
org.broadinstitute.hellbender.utils.variant.writers.GVCFBlock.getMedianDP(GVCFB
lock.java:75)
        at
org.broadinstitute.hellbender.utils.variant.writers.HomRefBlock.createHomRefGen
otype(HomRefBlock.java:73)
        at
org.broadinstitute.hellbender.utils.variant.writers.GVCFBlock.toVariantContext(
GVCFBlock.java:49)
        at
org.broadinstitute.hellbender.utils.variant.writers.GVCFBlockCombiner.emitCurre
ntBlock(GVCFBlockCombiner.java:177)
        at
org.broadinstitute.hellbender.utils.variant.writers.GVCFBlockCombiner.signalEnd
OfInput(GVCFBlockCombiner.java:227)
        at
org.broadinstitute.hellbender.utils.variant.writers.GVCFWriter.close(GVCFWriter
.java:70)
        at
org.broadinstitute.hellbender.tools.walkers.haplotypecaller.HaplotypeCaller.clo
seTool(HaplotypeCaller.java:216)
        at
org.broadinstitute.hellbender.engine.GATKTool.doWork(GATKTool.java:1053)
        at
org.broadinstitute.hellbender.cmdline.CommandLineProgram.runTool(CommandLinePro
gram.java:140)
        at
org.broadinstitute.hellbender.cmdline.CommandLineProgram.instanceMainPostParseA
rgs(CommandLineProgram.java:192)
        at
org.broadinstitute.hellbender.cmdline.CommandLineProgram.instanceMain(CommandLi
neProgram.java:211)
        at
org.broadinstitute.hellbender.Main.runCommandLineProgram(Main.java:160)
        at org.broadinstitute.hellbender.Main.mainEntry(Main.java:203)
        at org.broadinstitute.hellbender.Main.main(Main.java:289)
```

Solution: java程序溢出

```
-Xmx75000M
```

SH021

HG00463 vcf ready

gatk doing HG00464 but shutting down (633)

```
Exception in thread "main" java.lang.OutOfMemoryError: Java heap space
```

gatk HaplotypeCaller doing HG00465 but shutting down(729)

```
/data/mouse/human_rme/genome_index/gatk-4.1.9.0/gatk --java-options "-Xmx4g"
HaplotypeCaller  \
-R /data/mouse/human_rme/genome_index/human_g1k_v37.fasta \
-I HG00465Aligned.sortedByCoord.out.new.bam \
-O /data/mouse/human_rme/SH021/HG00465_new.vcf.gz \
-ERC GVCF

# return
Using GATK jar /data/mouse/human_rme/genome_index/gatk-4.1.9.0/gatk-package-
4.1.9.0-local.jar
Running:
    java -Dsamjdk.use_async_io_read_samtools=false -
Dsamjdk.use_async_io_write_samtools=true -
Dsamjdk.use_async_io_write_tribble=false -Dsamjdk.compression_level=2 -Xmx4g -
jar /data/mouse/human_rme/genome_index/gatk-4.1.9.0/gatk-package-4.1.9.0-
local.jar HaplotypeCaller -R
/data/mouse/human_rme/genome_index/human_g1k_v37.fasta -I
HG00465Aligned.sortedByCoord.out.new.bam -O
/data/mouse/human_rme/SH021/HG00465_new.vcf.gz -ERC GVCF
15:35:01.927 INFO  NativeLibraryLoader - Loading libgkl_compression.so from
jar:file:/data/mouse/human_rme/genome_index/gatk-4.1.9.0/gatk-package-4.1.9.0-
local.jar!/com/intel/gkl/native/libgkl_compression.so
Dec 02, 2020 3:35:02 PM
shaded.cloud_nio.com.google.auth.oauth2.ComputeEngineCredentials
runningOnComputeEngine
INFO: Failed to detect whether we are running on Google Compute Engine.
15:35:02.092 INFO  HaplotypeCaller - ----------------------------------------
------------------
15:35:02.092 INFO  HaplotypeCaller - The Genome Analysis Toolkit (GATK)
v4.1.9.0
15:35:02.093 INFO  HaplotypeCaller - For support and documentation go to
https://software.broadinstitute.org/gatk/
15:35:02.192 INFO  HaplotypeCaller - Executing as zhuyufei@bdp-svr05 on Linux
v5.3.1-1.el7.elrepo.x86_64 amd64
15:35:02.192 INFO  HaplotypeCaller - Java runtime: OpenJDK 64-Bit Server VM
v1.8.0_152-release-1056-b12
```

```
15:35:02.192 INFO  HaplotypeCaller - Start Date/Time: 2020年12月2日 下午03时35分01秒
15:35:02.192 INFO  HaplotypeCaller - ------------------------------------------------------------
15:35:02.193 INFO  HaplotypeCaller - ------------------------------------------------------------
15:35:02.194 INFO  HaplotypeCaller - HTSJDK Version: 2.23.0
15:35:02.194 INFO  HaplotypeCaller - Picard Version: 2.23.3
15:35:02.194 INFO  HaplotypeCaller - HTSJDK Defaults.COMPRESSION_LEVEL : 2
15:35:02.194 INFO  HaplotypeCaller - HTSJDK Defaults.USE_ASYNC_IO_READ_FOR_SAMTOOLS : false
15:35:02.194 INFO  HaplotypeCaller - HTSJDK Defaults.USE_ASYNC_IO_WRITE_FOR_SAMTOOLS : true
15:35:02.194 INFO  HaplotypeCaller - HTSJDK Defaults.USE_ASYNC_IO_WRITE_FOR_TRIBBLE : false
15:35:02.195 INFO  HaplotypeCaller - Deflater: IntelDeflater
15:35:02.195 INFO  HaplotypeCaller - Inflater: IntelInflater
15:35:02.195 INFO  HaplotypeCaller - GCS max retries/reopens: 20
15:35:02.195 INFO  HaplotypeCaller - Requester pays: disabled
15:35:02.195 INFO  HaplotypeCaller - Initializing engine
15:35:02.568 INFO  HaplotypeCaller - Done initializing engine
15:35:02.569 INFO  HaplotypeCallerEngine - Tool is in reference confidence mode and the annotation, the following changes will be made to any specified annotations: 'StrandBiasBySample' will be enabled. 'ChromosomeCounts', 'FisherStrand', 'StrandOddsRatio' and 'QualByDepth' annotations have been disabled
15:35:02.576 INFO  HaplotypeCallerEngine - Standard Emitting and Calling confidence set to 0.0 for reference-model confidence output
15:35:02.576 INFO  HaplotypeCallerEngine - All sites annotated with PLs forced to true for reference-model confidence output
15:35:02.589 INFO  NativeLibraryLoader - Loading libgkl_utils.so from jar:file:/data/mouse/human_rme/genome_index/gatk-4.1.9.0/gatk-package-4.1.9.0-local.jar!/com/intel/gkl/native/libgkl_utils.so
15:35:02.590 INFO  NativeLibraryLoader - Loading libgkl_pairhmm_omp.so from jar:file:/data/mouse/human_rme/genome_index/gatk-4.1.9.0/gatk-package-4.1.9.0-local.jar!/com/intel/gkl/native/libgkl_pairhmm_omp.so
15:35:02.642 INFO  IntelPairHmm - Using CPU-supported AVX-512 instructions
15:35:02.642 INFO  IntelPairHmm - Flush-to-zero (FTZ) is enabled when running PairHMM
15:35:02.642 INFO  IntelPairHmm - Available threads: 64
15:35:02.642 INFO  IntelPairHmm - Requested threads: 4
15:35:02.642 INFO  PairHMM - Using the OpenMP multi-threaded AVX-accelerated native PairHMM implementation
15:35:02.671 INFO  ProgressMeter - Starting traversal
15:35:02.672 INFO  ProgressMeter -        Current Locus  Elapsed Minutes    Regions Processed   Regions/Minute
15:35:12.675 INFO  ProgressMeter -           1:13535701              0.2            45120         270720.0
```

```
15:35:22.673 INFO  ProgressMeter -           1:28421701              0.3
          94740         284220.0
15:35:32.672 INFO  ProgressMeter -           1:42599701              0.5
         142000         284000.0
15:35:42.673 INFO  ProgressMeter -           1:56069701              0.7
         186900         280343.0
15:35:59.292 INFO  ProgressMeter -           1:70610701              0.9
         235370         249425.1
15:36:09.292 INFO  ProgressMeter -           1:83174701              1.1
         277250         249699.8
15:36:19.293 INFO  ProgressMeter -           1:97466701              1.3
         324890         254413.3
15:36:29.294 INFO  ProgressMeter -          1:108578701              1.4
         361930         250696.1
15:36:39.295 INFO  ProgressMeter -          1:121907701              1.6
         406360         252337.4
15:36:49.407 INFO  ProgressMeter -          1:135599701              1.8
         452000         254087.2
15:36:59.411 INFO  ProgressMeter -          1:148913701              1.9
         496380         255131.7
15:37:10.510 INFO  ProgressMeter -          1:157865701              2.1
         526220         246978.2
15:37:20.743 INFO  ProgressMeter -          1:168830701              2.3
         562770         244556.8
15:37:30.823 INFO  ProgressMeter -          1:180791701              2.5
         602640         244064.5
15:37:40.934 INFO  ProgressMeter -          1:192755701              2.6
         642520         243591.0
15:37:50.934 INFO  ProgressMeter -          1:205016701              2.8
         683390         243687.8
15:38:00.935 INFO  ProgressMeter -          1:217073701              3.0
         723580         243543.5
15:38:10.935 INFO  ProgressMeter -          1:228584701              3.1
         761950         242835.8
15:38:20.935 INFO  ProgressMeter -          1:239054701              3.3
         796850         241149.4
15:38:49.666 INFO  HaplotypeCaller - Shutting down engine
[2020年12月2日 下午03时38分49秒]
org.broadinstitute.hellbender.tools.walkers.haplotypecaller.HaplotypeCaller
done. Elapsed time: 3.80 minutes.
Runtime.totalMemory()=4292345856
Exception in thread "main" java.lang.OutOfMemoryError: Java heap space
        at java.util.stream.Nodes$DoubleArrayNode.<init>(Nodes.java:1429)
        at java.util.stream.Nodes$DoubleFixedNodeBuilder.<init>
(Nodes.java:1589)
        at java.util.stream.Nodes.doubleBuilder(Nodes.java:279)
        at
java.util.stream.DoublePipeline.makeNodeBuilder(DoublePipeline.java:164)
```

```
        at
java.util.stream.AbstractPipeline.evaluate(AbstractPipeline.java:543)
        at
java.util.stream.AbstractPipeline.evaluateToArrayNode(AbstractPipeline.java:260
)
        at java.util.stream.DoublePipeline.toArray(DoublePipeline.java:506)
        at
org.broadinstitute.hellbender.utils.MathUtils.median(MathUtils.java:841)
        at
org.broadinstitute.hellbender.utils.variant.writers.GVCFBlock.getMedianDP(GVCFB
lock.java:75)
        at
org.broadinstitute.hellbender.utils.variant.writers.HomRefBlock.createHomRefGen
otype(HomRefBlock.java:73)
        at
org.broadinstitute.hellbender.utils.variant.writers.GVCFBlock.toVariantContext(
GVCFBlock.java:49)
        at
org.broadinstitute.hellbender.utils.variant.writers.GVCFBlockCombiner.emitCurre
ntBlock(GVCFBlockCombiner.java:177)
        at
org.broadinstitute.hellbender.utils.variant.writers.GVCFBlockCombiner.signalEnd
OfInput(GVCFBlockCombiner.java:227)
        at
org.broadinstitute.hellbender.utils.variant.writers.GVCFWriter.close(GVCFWriter
.java:70)
        at
org.broadinstitute.hellbender.tools.walkers.haplotypecaller.HaplotypeCaller.clo
seTool(HaplotypeCaller.java:216)
        at
org.broadinstitute.hellbender.engine.GATKTool.doWork(GATKTool.java:1053)
        at
org.broadinstitute.hellbender.cmdline.CommandLineProgram.runTool(CommandLinePro
gram.java:140)
        at
org.broadinstitute.hellbender.cmdline.CommandLineProgram.instanceMainPostParseA
rgs(CommandLineProgram.java:192)
        at
org.broadinstitute.hellbender.cmdline.CommandLineProgram.instanceMain(CommandLi
neProgram.java:211)
        at
org.broadinstitute.hellbender.Main.runCommandLineProgram(Main.java:160)
        at org.broadinstitute.hellbender.Main.mainEntry(Main.java:203)
        at org.broadinstitute.hellbender.Main.main(Main.java:289)
```

SH002

gatk HaplotypeCaller doing HG00407 (1060)

gatk HaplotypeCaller doing HG00406(1182) (but shutting down at Current Locus 1:246638732; try agin then shutting Current Locus down 1:241940732)

gatk HaplotypeCaller doing HG00408 (1455) (but shutting down at Current Locus 1:240860701)

imprinting gene

基因铭印\parencite{reik2001genomic}是在真兽类哺乳动物中观察到的一种表观遗传现象。对于大多数常染色体基因，两个亲本拷贝都是转录或沉默。然而，在一小组基因中，其中一个拷贝以亲本特有的方式被关闭，从而导致单等位基因表达。这些基因被称为“印迹”，因为沉默的基因拷贝在卵子或精子中具有表观遗传标记或印迹。

印迹基因在胎儿和胎盘组织中发挥重要作用，在产前和产后的发育和生长中发挥重要作用\parencite{morison2005census}。有趣的是，母方表达的基因会限制胚胎的生长，而父方表达的基因会促进胚胎的生长。Igf2和 Igf2r在小鼠体内的拮抗作用是这种突出场景的一个典型案例。父方表达Igf2基因的缺失导致了宫内生长迟缓。另一方面，缺失母方表达的Igf2r基因，导致过度生长\parencite{lau1994loss}。

母方和父方表达的基因的拮抗作用引发了一系列进化理论的争论，旨在解释在“自然选择”过程中遗传印记的起源。目前最为科学接受的理论是亲属理论\parencite{haig1989selective, moore1991genomic}。简单地说，这个理论认为在多配偶的哺乳动物物种中，沉默来自母系的生长抑制基因可以导致胚胎的生长。这与营养需求增加有关，从而导致以后代为代价开发母亲资源，而后代可能是另一个男性的后代。

一个基因调控机制的进化优先沉默一个基因的亲本等位基因意味着父方和母方表达基因在进化过程中经历不同的选择压力。这一假设得到了两个群体揭示了不同的序列保守模式的支持。父系表达基因的蛋白质编码 DNA 序列在不同哺乳动物中保存得很好，而母系表达基因则差异很大。是否父方和母方表达的基因在分子功能和基因调控上也不同，这是一个尚未详细研究的问题。许多研究表明，印迹基因不仅在胚胎发育过程中起重要作用，而且还具有出生后的功能。因此，以产前发育为中心的亲属理论或许可以解释基因铭印进化的某些方面，但不是全部\parencite{hamed2012cellular}。

在出生后的发育过程中，基因铭印会影响内分泌网络、能量代谢和行为。对小鼠的敲除研究表明，表达Peg1和 Peg3基因的两个父方基因具有明显的行为表型\parencite{lefebvre1998abnormal}。从父亲那里遗传到这些基因的非等位基因的雌性表现出缺乏母性照顾行为，包括吞食胎盘和筑巢以及幼崽聚集。

另一个有趣的事实是，胎盘哺乳动物如老鼠和人类，以及有袋动物如负鼠和袋鼠，都有基因铭印。卵生哺乳动物，如鸭嘴兽和针鼹鼠，似乎缺乏印记基因。胎盘哺乳动物和有袋类动物区别于卵生哺乳动物的一个生殖策略是允许胚胎直接影响用于自身生长的母性资源的数量。相反，在卵子中发育的胚胎不能直接影响母体资源。大多数无脊椎动物和脊椎动物使用产卵繁殖策略。值得注意的是，它们也可以进行孤雌生殖——种繁殖形式，即雌配子不经过雄配子受精而发育成一个新的二倍体个体(注意，孤雌生殖胚胎来自同一母体基因组的复制，而图2中描述的雌核发育胚胎来自两个不同的母体基因组)。生物体进行单性生殖的能力很可能表明基因铭印的完全缺失，因为这表明父系基因组是可有可无的。然而，在哺乳动物中，印迹基因表达控制胎儿生长的直接结果是孤雌生殖是不可能的。双亲都必须产生可存活的后代，使哺乳动物完全依赖有性生殖进行繁殖。因此，哺乳动物几乎不存在孤雌生殖现象\parencite{renfree2009evolution}。

为什么基因铭印只在一些哺乳动物中进化，而在脊椎动物中却没有？基因组印记的三个特征——许多印记基因的生长调节功能，印记基因对胎盘哺乳动物和有袋哺乳动物的限制，以及父系基因组对胎儿发育的必要性，为两个同样有吸引力的假说提供了证据。

第一个假设提出，基因铭印的进化是为了回应“父母冲突”的情况\parencite{moore1991genomic}。这源于母体和父体基因组的对立利益: 胚胎的发育依赖于父母一方，但是受到胚胎的影响，胚胎的基因组来自父母双方。父方表达的印记基因被认为可以促进胚胎发育，从而最大限度地提高拥有特定父方基因组的个体后代的适应性。母方表达的印迹基因被认为可以抑制胎儿的生长。这将使得母系资源更平等地分配给所有后代，并增加母系基因组向多个后代的传递，这些后代可能有不同的父系基因组。

第二个假设被称为“滋养层防御”\parencite{varmuza1994genomic}。这就提出，如果自发的卵母细胞激活导致胚胎的完全发育，那么母体基因组就有可能因为具备内部繁殖的解剖学条件而面临风险。由于男性缺乏必要的内部繁殖解剖设备，他们不共享相同的风险，应自发激活精子发生。因此，印记被认为可以抑制母染色体上促进胎盘发育的基因，或者激活限制这一过程的基因。因此，胎盘侵入母体子宫血管所必需的基因只能在受精后由父体基因组表达。

这两种假说都指出印迹基因在调节胎盘发育和功能方面的作用，然而，无论是亲代冲突还是滋养层防御模型都不能为所有的数据提供完整的解释\parencite{wilkins2003good}。有趣的是，植物胚乳中也发现了印迹基因，这种组织被比作胎盘，因为它将营养资源从亲本植物转移到胚胎中\parencite{grossniklaus2014transcriptional}。这一发现加强了关于基因铭印是作为调节父母和子女之间营养物质转移的手段进化而来的论点。有可能的是，并不是一个集群中的所有基因都是刻意为之的印记机制的目标，有些基因可能只是这个过程的“无辜旁观者”，而且它们的功能也不能提供有关基因铭印的信息。受印迹机制影响的无辜旁观者基因的存在可以令人满意地解释印迹基因的奇特丰富性，但在发育过程中没有明显的生物学功能\parencite{bartolomei1997genomic}。

method design

artical: https://bmcbiol.biomedcentral.com/articles/10.1186/s12915-019-0674-0

这篇文章的主要思路是通过SNV来确定子代基因来自于父母双方的哪一方。

印记基因假说：

> Genomic imprinting is a special case of mono-allelic expression where genes are expressed in a parent-of-origin (PofO)-specific manner. Although several hypotheses exist to explain why genomic imprinting occurs, the parental conflict hypothesis that imprinted genes evolved from a parental battle between males and females to influence the allocation of maternal resources to offspring. This type of mono-allelic expression can be observed in mammals at different developmental stages and is dependent on stage, cell, and tissue type.

Material

165 trio from HapMap/1000 Genomes Projects with RNA-Seq data from lymphoblastoid cell lines (LCLs) and 131 trios from the Genome-of-the-Netherlands

Focus

complete imprinting (exclusive expression of the paternal or maternal allele) and incomplete imprinting (bias in expression towards the maternal or paternal allele)

Method

allele-specific RNA-Seq analysis of parent-offspring trios:

used phased genotypes to compute the relative expression from the maternal and paternal alleles in RNA-Seq reads at expressed heterozygous single nucleotide variants (SNVs);

summed the paternal and maternal counts for all heterozygous SNVs contained in a gene(irrespective of their exonic or intronic nature);

statistical tests to check for consistent parental expression bias of autosomal genes within the populations: Wilcoxon signed-rank (WSR) test and ShrinkBayes (SB);

1. Strand-specific RNA-Seq in lymphoblastoid cell lines
    1. quality control by fastqc (version 0.11.2)
    2. Over-represented sequences were removed by trimmomatic (version 0.32) [reads ≥ 30 bp in length were kept]
    3. Cleaned reads were mapped to the human reference genome (hg19) with Gencode v16 annotations by STAR aligner (version 2.3.0) [yielding a mean of 79% uniquely mapped reads]
    4. intermediate BAM file processing such as add read groups and sorting and merging BAM files of the same samples by Picard (version 1.112)
    5. correct for mapping errors and biases which can result in false-positive allele-specific read assignments by WASP software (version 0.1)https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4626402/ [resulting in the removal of a mean of 36% of reads that overlapped SNVs in each sample] {WASP input: bam; then identifies mapped reads that overlap known polymorphisms}
    6. SNVs in each offspring were assigned on parental origin
    7. determine allele-specific expression by heterozygous sites
    8. quantified reference and alternate RNA-Seq reads mapped at heterozygous loci by AlleleCounter (v0.2, https://github.com/secastel/allelecounter)
    9. reference and alternate allele counts were used with PofO information to assign counts to the maternal and paternal alleles at each heterozygous site [Reads that did not uniquely map, or had base quality ≤ 10, were discarded.]

10. Use filter to reduce the mapping errors:

    ① removing heterozygous SNVs that had a mappability score < 1 (based on the "CRG GEM Alignability of 50mers with no more than 2 mismatches" track, downloaded from UCSC genome browser)

    ② removing heterozygous SNVs that overlapped CNVs with MAF ≥ 5% identified in samples from the 1000 Genomes and HapMap Projects (ftp://ftp.1000genomes.ebi.ac.uk/vol1/withdrawn/phase3/integrated_sv_map/ and common CNVs

    ③ removing heterozygous SNVs that are segmental duplications

    ④ removing heterozygous SNVs that are simple repeats (both downloaded from "Variation and Repeats" track group of the UCSC genome browser).

    [These filters resulted in the removal of 21% of heterozygous sites, leaving ~ 3.1 million sites for downstream analysis.]

11. Genotyping DNA:

    BEAGLE (https://bmcbiol.biomedcentral.com/articles/10.1186/s12915-019-0674-0#ref-CR57) and IMPUTE2. Using GATK:UnifiedGenotyper as input for BEAGLE, treating all samples as unrelated. SHAPEIT2 and MVNcall19 were then used along with trio information to phase the complete set of SNVs. Each haplotype transmitted to the offspring, and therefore, allelic parental origin was then obtained from the phased haplotypes

2. overall Genotype data processing
    1. get genotype data from smples

       The **GATK** joint genotyping workflow can be applied in genotyping, here list the steps:

       ①Versions 3.0 and above of GATK offer the possibility of calling DNA variants on cohorts of samples using the **HaplotypeCaller** algorithm in Genomic Variant Call Format (GVCF) mode.

       input: RNA data

       output: one GVCF file per sample

       ②variants are called from the GVCF files through a joint genotyping analysis.

    2. quality control: resolving strand inconsistencies, removing multi-allelic SNVs and indels, removing SNVs not present in the 1000 Genomes data
    3. Use PLINK (versions 1.07 and 1.9), vcftools (version 0.1.15) and Beagle Utilities to convert coordinates from hg18 to hg19