基因组学与应用生物学，2021 年，第 40 卷，第 1 期，第 435-441 页

研究报告
Research Report

# 一种基于二代测序拷贝数变异检测的新方法

杨浩[1]　姜丹[2]　方铭[2*]
1 黑龙江八一农垦大学生命科学技术学院, 大庆, 163319; 2 集美大学水产学院, 厦门, 361021
* 通信作者, fangming618@126.com

**摘　要**　区域捕获测序是针对基因组特定区段如对 MHC (Major histocompatibility complex)区域、外显子区域等测序的有效手段，但是由于捕获测序中探针设计不均匀而造成区域内测序深度变异很大，因此，与基于全基因组的测序数据相比，其拷贝数变异的检测难度更大。目前已经出现了捕获测序下拷贝数变异 (copy number variations, CNV)的检测方法，但对 CNV 的检测准确性仍然很低，特别是对于低频率 CNV 来说效果极差。因此，本研究开发了一个新的拷贝数变异检测方法，其特点是：(1)以区域内划分的区间为单位检测区间内的 CNV，而不是直接对每个个体检测 CNV；(2)全面利用群体内所有个体信息，通过区间内 read 深度在群体的分布规律来检测 CNV 的分离规律，假设区间内只有 1 个 CNV，那么区间内的 read 深度将服从三峰的混合正态分布。将该方法应用于 21 327 个银屑病个体区域捕获测序的 CNV 检测中，结果表明，XHMM，ExomeDepth 和本方法跟金标准重叠的窗口总数与金标准总窗口数的百分比(即重叠率)分别是 7%、18%和 62%。与 XHMM 和 ExomeDepth 相比，新方法在区间内 CNV 检测覆盖度可以分别提高 55 个百分点和 44 个百分点。本研究完善拷贝数变异检测方法，为疾病的诊断治疗提供一定的理论依据。

**关键词**　MHC; CNV; 测序; 覆盖度; 银屑病

# A New Method for Detection of Copy Number Variations Under Next Generation Sequencing

Yang Hao[1]　Jiang Dan[2]　Fang Ming[2*]
1 College of Life Science and Technology, Heilongjiang Bayi Agricultural University, Daqing, 163319; 2 Fisheries College, Jimei University, Xiamen, 361021
* Corresponding author, fangming618@126.com
DOI: 10.13417/j.gab.040.000435

**Abstract**　Target region capture sequencing is an effective method for sequencing specific regions on genome, such as MHC region or exon regions. However, due to the uneven design of probes in capture sequencing, the sequencing depth is extremely variable in capture region, compared with whole-genome sequencing data, copy number variation detection is more difficult. So far, several CNV detection methods for target region capture sequencing have been developed, but the detection of CNV, especially for low frequency CNV, is still relative inefficient. Therefore, we developed a new CNV detection method suitable for target region capture sequencing and whole genome sequencing (WGS) in this study. The characteristics of the new method are: (1) It detects the CNV by units of intervals within the region, instead of each individual; (2) The study utilizes all the individual information in the group, and detects the CNV through the distribution of read depth in the region. If there is only one CNV in the interval, the read depth in the interval will follow mixed normal distribution of the three peaks. The study applied the new method to CNV detection of 21 327 psoriasis individual with target sequencing in MHC region, the results showed

基金项目：本研究由国家自然科学基金面上项目(31672399; 31872560)资助
引用格式：Yang H., Jiang D., and Fang M., 2021, A new method for detection of copy number variations under next generation sequencing, Jiyinzuxue yu Yingyong Shengwuxue (Genomics and Applied Biology), 40(1): 435-441. (杨浩, 姜丹, 方铭, 2021, 一种基于二代测序拷贝数变异检测的新方法, 基因组学与应用生物学, 40(1): 435-441.)

(C)1994-2024 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

that the accuracies of XHMM, ExomeDepth and our method is 7%, 18%, and 62%, respectively; and the coverage of CNV detection with the new method is increased by 55 and 44 percentage points than XHMM and ExomeDepth, respectively. The new CNV detection method provided a theoretical basis for the disease diagnosis and treatment.
**Keywords**　MHC; CNV; Sequencing; Coverage; Psoriasis

银屑病俗称牛皮鲜，是由遗传和环境因素引起的炎性皮肤病(Zhou et al., 2018)，它是一种慢性病，影响皮肤和关节，并且有多种表型，其中斑块状银屑病是最常见的形式(Harrington et al., 2017)。经研究发现，银屑病的严重程度与某些基因拷贝数变异(CNV)有明显的相关性(Prans et al., 2013)。大多数 CNV 是正常变异且为良性，而其他 CNV 与疾病有很强的相关性(Haraksingh et al., 2017a)。如大于 1 kb 长度 DNA 片段的扩增与缺失,是已知的导致常见遗传病如银屑病、自闭症、精神分裂症等疾病的重要人类基因组变异(Yao et al., 2017)，因此，检测拷贝数变异十分重要。

目前检测拷贝数变异主要有基于芯片的方法和基于测序的方法。芯片主要有比较基因组杂交芯片(comparative genomic hybridization, CGH)和 SNP (Single nucleotide polymorphism)芯片(Wang and Byers, 2014)。CGH 芯片是双通道芯片，用细菌人工染色体克隆(bacterial artificial chromosome, BAC)和寡聚核苷酸(Oligonucleotide)作为探针，检测出的CNV 长度范围较窄，只能检测长度较长的 CNV。SNP 芯片是单通道芯片，采用寡聚核苷酸探针，可检测出相对较短的 CNV。基于基因组区域捕获结合二代测序技术是当前 MHC 区域 CNV 检测的主要手段(De Groot et al., 2017)，目前对区域捕获测序 CNV 检测常用软件有 XHMM (Fromer et al., 2012)，ExomeDepth (Plagnol et al., 2012)和 CLAMMS (Packer et al., 2016)等。XHMM 软件旨在从目标外显子组序列数据中提取 CNV 的信息，XHMM 的关键步骤包括覆盖深度的计算、数据正态化、CNV 分型等。XHMM 的基本原理是通过主成分分析(PCA)找到深度在多个样本和目标之间变化的主要模式，并且通过剔除影响最大的主成分来控制个体间 read 深度的偏差，进而利用隐马尔可夫模型(HMM)根据所划分区间的 read 深度之间的关系来实现个体水平的 CNV 片段检测。HMM 同时提供了 CNV 检测的质量指标用于评价所检测 CNV 的可靠性(Fromer and Purcell, 2014)。ExomeDepth 也是先将基因组区域划分成小的区间，并在 CNV 检测时选取一些个体作为对照，通常选择 100 个个体作为对照，通过群体内 read 的分布结合隐马尔可夫链方法检测 CNV 片段(Ellingford et al., 2017)。CLAMMS 是目前较新的适用于区域捕获测序的 CNV 检测方法，主要分为三个步骤：根据 GC 含量校正区间测序深度；在参考样本组里利用混合模型将 CNV 划分为不同的状态，如纯合删失，1 个拷贝，2 个拷贝，3 个拷贝，4 个拷贝等，并将所有状态拟合成混合正态分布，再通过最大似然法训练混合分布各组分所占的比例；最后使用隐马尔可夫模型(HMM)在个体水平上检测 CNV 片段(Seiser and Innocenti, 2014)。

上述 3 种软件能够较好地检测大片段，但在稀有拷贝数变异检测方面并不精确，在短片段检测方面有很大缺陷。实际上区域捕获测序的最大研究难点是由于探针设计的不均匀，各个划分的区间之间 read 深度变异很大，虽然进行了各种校正方法，还是无法有效地避免此问题。尽管如此，区间内的测序深度在个体之间呈现高度的相关性。本研究将充分利用这一规律，通过大样本的优势，直接利用群体中所有个体信息来研究每个区间内个体 read 深度的分布规律，在没有 CNV 的情况下，区间内 read 深度服从单峰的正态分布，但是如果存在 CNV 时将服从多峰的混合正态分布，通过检测拟合多峰的混合正态分布可以敏感地检测到 CNV 的存在。最后，通过 24 个个体在基于基因组重测序所检测的 CNV 作为金标准，将新方法所得结果与现有的 XHMM 和 ExomeDepth 两个软件得到的结果进行比较。

## 1 结果与分析

### 1.1 正态化

片段计数受 DNA 片段 GC 含量影响非常大，正态化后和正态化前图形的差别明显(图 1)。结果显示，A 和 C，B 和 D 分别是正态化前和正态化后窗口 14 079，14 251 的 read 直方图(图 1)。通过密度差区分后，A 是单峰，C 是三峰；B 是单峰，D 是三峰。该结果表明正态化是十分必要的，可以明显地从单峰里分离出 CNV 检测所必须的多峰，如 D 的 3 个峰从左到右分别表明纯合的缺失、杂合的缺失和纯合的正常状态。

### 1.2 区间 read 深度的变异系数

正态化后，计算所有窗口的变异系数，绘制每个

(C)1994-2024 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

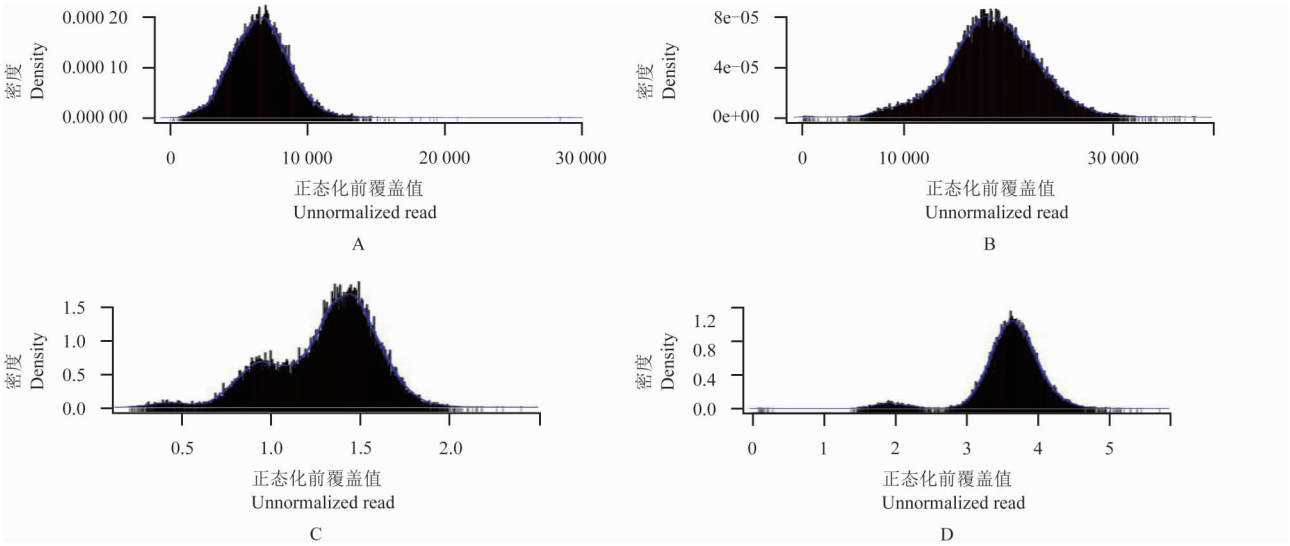


图 1 正态化前和正态化后 read 的窗口对比

注: A: 正态化前的窗口 14 079; B: 正态化前的窗口 14 251; C: 正态化后的窗口 14 079; D: 正态化后的窗口 14 251

Figure 1 Comparison of window read before and after normalization

Note: A: Window 14 079 before normalization; B: Window 14 251 before normalization; C: Window 14 079 after normalization; D: Window 14 251 after normalization

区间的密度分布图(图 2A)。截取横坐标为[0,5]时的图像(图 2B)发现，*CV* 大部分集中于 1.5 之内，而在大于 1.5 后开始趋于平缓，而且变化范围相当大。我们统计分析表明，*CV* 大于等于 1.5 时的典型分布图的 read 严重集中于 0 附近(图 3A)；*CV* 小于 1.5 时的典型分布图的 read 相对分散(图 3B)。因此将针对 *CV*<1.5 和 *CV*≥1.5 两种情况分别研究。

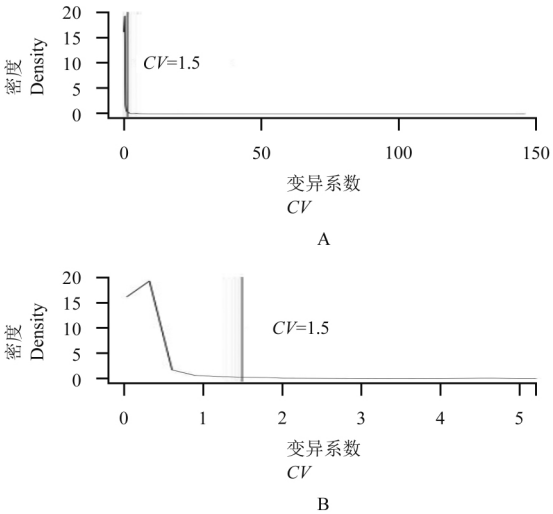


图 2 所有窗口变异系数的密度分布

注: A: 横坐标包含全部 *CV* 值; B: 横坐标包含[0,5]的 *CV* 值

Figure 2 Nuclear density diagram of all window coefficients of variation

Note: A: The abscissa contains all *CV* values; B: The abscissa contains *CV* values of [0,5]

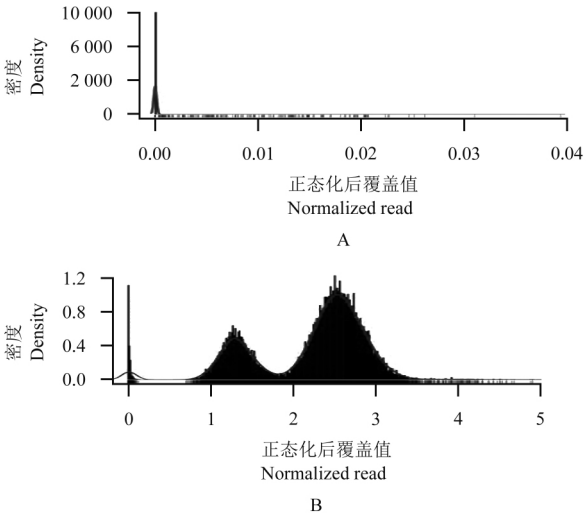


图 3 窗口频率分布

注: A: 窗口 1 107; B: 窗口 13 625

Figure 3 Window frequency distribution

Note: A: Window 1 107; B: Window 13 625

### 1.3 窗口 *CV*<1.5 的 CNV 片段分析

当 *CV*<1.5 时，将差分法得到的多峰窗口按位置聚类，将处于多峰状态的窗口连接成片段，为了保证片段内 CNV 的可靠性及对应的是同一个 CNV，将片段内区间根据变异系数 *CV* 进一步确定，此过程是根据相邻窗口同类型峰的变异系数相似的原理实现的，如果相邻窗口的 CNV 小于跟定阈值 *F*，则认为他们连接在一起。将阈值 *F* 在其可能的范围内设为一

(C)1994-2024 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

系列数值(0.05~0.40)，并统计不同数值下所得出各个区间内图形的峰数的个数，选择处于三峰状态个数最多区间所对应的 *F* 值(图 4A)。*F* 值在0.11，012，0.16所得到的 3 峰数量最多，达到 223 个，*F* 值越大，得到的 CNV 区段越长，越完整，因此选择 0.16。由于密度差分时峰的形状受到函数带宽的影响，带宽大小受 adjust 的影响，因此，对 adjust 值进行训练，得到 adjust 值训练图(图 4B)，当 adjust 等于 1时三峰个数达到最大。在 *F*=0.16，adjust=1 条件下，挑选出图形呈现三峰状态的区间即为部分 CNV 区段。

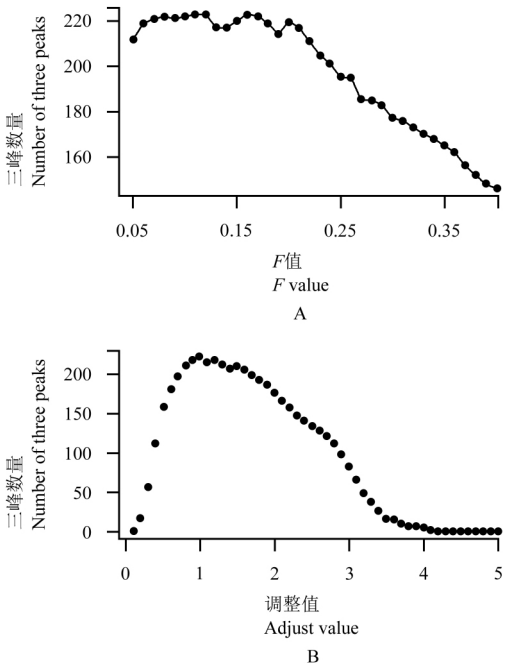


图 4 *F* 值训练和 adjust 值训练

注: A: *F* 值训练; B: adjust 训练(*CV*<*K*)

Figure 4 *F* value training and adjust training

Note: A: *F* value training; B: adjust training (*CV*<*K*)

### 1.4 窗口 *CV*≥1.5 的 CNV 片段分析

*CV*≥1.5 的相邻窗口的变异系数相差较大，不适合用变异系数连接相邻窗口，但是这些窗口中大多测序覆盖度较低。窗口的位置经过层次聚类后挑出 500 bp 及以上的区间，read 区间覆盖度平均值也较低，没有明显多峰状态，因此需要将这些覆盖度低的区间排除。通过覆盖度平均值频率分布图(图 5)中的完整覆盖度平均值频率分布图(图 5A)和截取横坐标为[0, 500]的频率分布图(图 5B)观察发现，曲线从横坐标 120 开始变化平缓，覆盖度平均值较低的区间主要集中在 120 以内，在此阈值设置为 120，将覆盖度总和低于阈值的区间排除。然后对剩余的区间进行密度差分判断峰数，峰数受带宽大小影响，带宽大

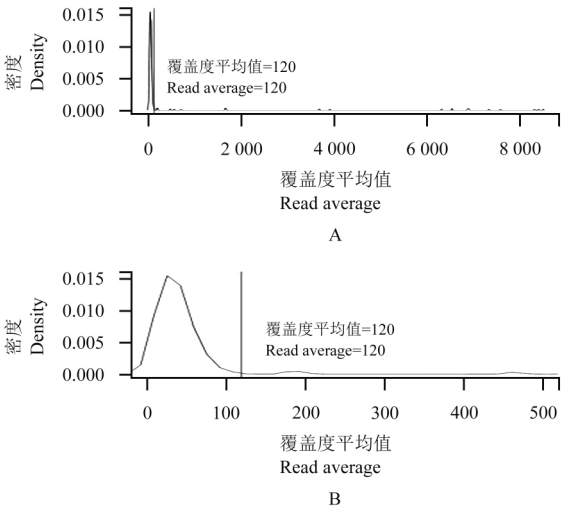


图 5 Read 平均值频率分布

注: A: 横坐标包含全部覆盖度平均值; B: 横坐标包含[0,500]的覆盖度平均值

Figure 5 Read average frequency distribution diagram

Note: A: The abscissa contains all read average values; B: The abscissa contains read average values of [0,500]

小受 adjust 值影响，因此需要对 adjust 值训练(图 6)，当 adjust 取值为 1.6~2.1 时，呈现三峰的区间最多，达到 16 个，此处 adjust 值选择 1.8，符合预期。根据adjust 等于 1.8，然后将覆盖度平均值低于 120 的区间排出后，剩余的区间进行密度差分，挑选图像为三峰的区间即为部分 CNV 区段。

### 1.5 3 种方法所得重叠率对比分析

XHMM、ExomeDepth 和本研究方法计算得出的 24 个个体拷到与金标准重叠的窗口总数及占比(表1)。XHMM 在 HG00436、HG00699、NA18532、NA18542、NA18570、NA18577 和 NA18624 这 7 个样本检测金标准长度为 0，在 HG00653、HG00683 等能检测到的金标准总长也非常短，重叠率非常低。与 XHMM 比较，ExomeDepth 检测的重叠率有所提高。新方法检测到的金标准总长度比前两种方法有明显提升，重

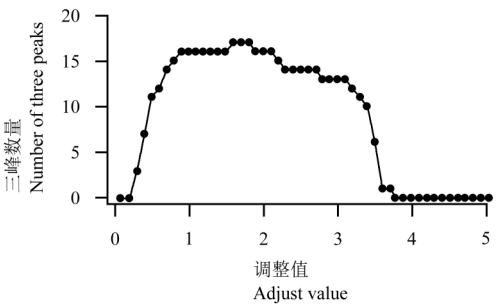


图 6 Adjust 值训练(*CV*≥*K*)

Figure 6 Adjust value training (*CV*≥*K*)

(C)1994-2024 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

表 1 3 种方法拷到的金标准总长度及占比

Table 1 The total length and proportion of the gold standard copied by the three methods

| 样本 Sample | 金标准总长度 Total length of gold standard | 拷到的金标准总长度及占比 Total length and proportion of gold standard copied | | |
|---|---|---|---|---|
| | | XHMM | ExomeDepth | 本方法 Our method |
| HG00421 | 771 | 27 (4%) | 100 (13%) | 488 (63%) |
| HG00428 | 1 300 | 101 (8%) | 103 (8%) | 949 (73%) |
| HG00436 | 1 971 | 0 (0%) | 30 (2%) | 935 (47%) |
| HG00442 | 880 | 248 (28%) | 206 (23%) | 530 (60%) |
| HG00443 | 779 | 30 (4%) | 117 (15%) | 468 (60%) |
| HG00472 | 1 328 | 229 (17%) | 218 (16%) | 962 (72%) |
| HG00473 | 704 | 110 (16%) | 170 (24%) | 435 (62%) |
| HG00478 | 1 380 | 156 (11%) | 115 (8%) | 981 (71%) |
| HG00500 | 714 | 30 (4%) | 146 (20%) | 488 (68%) |
| HG00619 | 785 | 154 (20%) | 86 (11%) | 474 (60%) |
| HG00653 | 784 | 2 (0%) | 454 (58%) | 464 (59%) |
| HG00683 | 771 | 8 (1%) | 206 (27%) | 488 (63%) |
| HG00699 | 796 | 0 (0%) | 321 (40%) | 494 (62%) |
| NA18532 | 783 | 0 (0%) | 100 (13%) | 498 (64%) |
| NA18542 | 634 | 0 (0%) | 70 (11%) | 426 (67%) |
| NA18547 | 811 | 10 (1%) | 70 (9%) | 511 (63%) |
| NA18570 | 641 | 0 (0%) | 105 (16%) | 382 (60%) |
| NA18572 | 1 139 | 139 (12%) | 103 (9%) | 874 (77%) |
| NA18573 | 794 | 92 (12%) | 283 (36%) | 488 (61%) |
| NA18577 | 647 | 0 (0%) | 197 (30%) | 407 (63%) |
| NA18605 | 711 | 44 (6%) | 287 (40%) | 452 (64%) |
| NA18611 | 946 | 52 (5%) | 108 (11%) | 587 (62%) |
| NA18620 | 1 504 | 89 (6%) | 333 (22%) | 486 (32%) |
| NA18624 | 799 | 0 (0%) | 94 (12%) | 501 (63%) |

叠率大大增加。XHMM、ExomeDepth 和本研究方法拷到金标准总长度与金标准总长度的百分比分别是 7%、18%和 62%，可见效果明显，符合预期(图 7)。

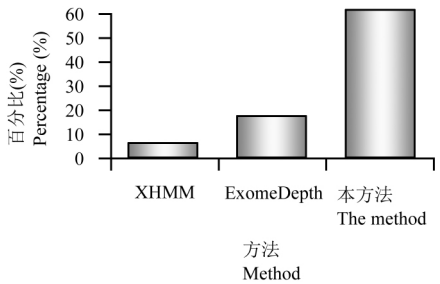


图 7 各方法拷到金标准总长度占金标准总长度的百分比

Figure 7 The percentage of the total length of each method copied to the gold standard

## 2讨论

拷贝数变异是人类遗传变异的主要类别，广泛存在于人类基因组上，与多种遗传疾病和常见表型密切相关。大多数 CNV 构成正常变异并且在功能上是良性的，而其他 CNV 与疾病有很强的相关性。准确检测 CNV 对于生物医学研究和临床诊断都很重要，对人类生命科学的研究与探索具有非常重要的意义。目前现有检测 CNV 的方法中，基于芯片的方法测得结果的好坏严重依赖探针的密度，探针密度大则测得效果就更好(Haraksingh et al., 2017b)。基于二代测序数据的方法，如 XHMM, ExomeDepth，在稀有拷贝数变异检测效果不好，短片段检测并不精确。本研究方法直接运用大数据，通过聚类等方法能够更准确地判断出具有 CNV 特征的窗口，通过参数训练，准确地将相似 CNV 特征的窗口链接起来，形成 CNV 区段。

本研究方法有 4 个潜在的优点：第一，本研究方法采取 21 327 个个体 read 值整体相加，样本量较大，read 值较为稳定，且较稀有的拷贝数经过相加后值变得更大，更容易被检测到，因此可以提高准确度；第二，在判断区间的峰数时采用的差分法，能够更准确地确定区间的峰数，即便窗口不够平滑也更容易被检测出来；第三，本研究中控制带宽大小的 adjust 参数和控制相邻窗口能否被聚集在一起的 $F$ 值都是经过参数训练而得到的最优值；第四，本研究方法采用先找 CNV 区间，再进入个体水平找每个个体的 CNV 区间的方式，即所有个体 read 值相加，找整体的 CNV 区间，然后再在每个 CNV 区间分析每个个体，这样做的优点是，整体水平的 CNV 区间找准确，进入个体水平找 CNV 简单而准确。

虽然本研究方法在整体水平能够准确地检测拷贝数变异的区间并且具有较高的检测性能，但也存在一些局限和不足之处：第一，对于个体水平存在 CNV 而整体水平不具 CNV 特征或窗口图形中的某峰较小，这种情况很难检测到，read 相加前差分法只能将这样的窗口判断为单峰而过滤掉；第二，当窗口的 read 相加后总数还是较少，得到的窗口图形的峰非常稀有，峰之间的距离不大，只出现单峰或者双峰的情况，此种情况不容易被检测，因此排除了双峰区间，但这些区间很可能是 CNV。这些缺陷将纳入接下来的研究中，今后将着重解决，这对拷贝数变异及疾病研究具有重要意义。

(C)1994-2024 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

## 3 材料与方法

### 3.1 实验材料与仪器

基于捕获的银屑病 MHC 区域捕获测序包括 21 327 个个体，采用 BWA 与参考基因组比对和 GATK 软件的 SNP 检测后获得区域内遗传变异的 VCF 文件(Zhou et al., 2016)，为了尽可能捕获更多的 CNV，我们这里的 VCF 文件尽可能地包括测序 read 覆盖的所有的碱基，理论上包括每个碱基上对应的测序深度。计算工作在黑龙江八一农垦大学生物信息实验室和国家超级计算机天津中心进行。选择了当前先进的 CNV 推断软件 XHMM (http://atgu.mgh.harvard.edu/xhmm/tutorial.shtml)和 ExomeDepth (https://github.com/vplagnol//ExomeDepth)进行比较。

### 3.2 新的 CNV 检测算法流程

步骤一：read 值提取和窗口划分

为了便于数据整理，将 vcf 文件的 header 去掉，按行将文件分割成 20 个小文件，从小文件中对每个个体的每个位置提取 read 数，将得到的 read 数的文件按行合并，将窗口大小设为 100 bp，获得每个窗口的 read 个数。

步骤二：正态化(Normalization)

整个 DNA 片段的 GC 含量对个体间 read 深度均一度影响最大(Benjamini and Speed, 2012)，因此要将区域内 read 个数正态化，以此来纠正由 GC 偏差造成的个体间深度过度变异。首先，计算每个窗口的 GC 含量，然后每个个体的窗口都采用公式 Read-norm(w)=Read(w)/mean(Read|GC(w))进行正态化，这里 mean(Cov|GC(w))是个体测序区域内所有窗口中具有相同 GC 含量区间中所对应的 read 平均值，这里 GC 含量从 0 到 1 被划分为 100 个区间。正态化的 read 值被用于后续的 CNV 检测分析。

步骤三：CNV 分布区间的初步筛选

将所有区间内的 read 深度根据差分法获得深度的分布曲线，并获得分布曲线上所有“峰”。为了获得“峰”，首先对分布进行曲线的绘制，这里采用差分法，即利用密度函数 density 对 read 深度计算内核密度估计，纵坐标依次相减(后一纵坐标减前一纵坐标)，这样得到许多纵坐标差值；利用纵坐标差值的正负性，如连续数个正数差值后是数个负数差值，那么在正负差值交替的地方是一个峰的顶点，这样会获得曲线上的多个“峰”；获得的峰中有些峰较平坦，需要过滤掉，本研究方法是选择曲线中纵坐标大于最高峰的纵坐标 *0.05 的峰，这样会过滤掉大部分杂峰。此过程采用 R 软件包实现，使用默认参数。值得一提的是，所获得的曲线未必是单峰的正态分布，而有可能是多峰的分布，而多峰分布正是本研究的 CNV 特征，即 read 深度在个体中形成不同的分布。

步骤四：基于 CNV 的精细检测

计算所有窗口标准化后 read 深度的变异系数，然后做密度图，考察 read 深度的分布特征，计算每个区间 read 的变异系数($CV$)，当 $CV$ 很大时，我们发现 $CV$ 的分布存在 1 个十分接近于 0 的峰，因此，当 $CV \geqslant K$ 时我们单独研究。当 $CV<K$ 时，个体间在特定区间的分布典型地会呈现一种三峰分离状态，分别对应 CNV 的 3 种状态，即纯合的缺失、杂合的单拷贝和纯合双拷贝；或者纯合的双拷贝(纯合的正常态)、杂合的 3 个拷贝(杂合的重复)和纯合的四拷贝(纯合的重复)。将每个区间所拟合的分布采用差分法获得其峰，进而得到峰的个数；挑选出二、三峰的区间(单峰被认为没有 CNV 分离)，并对其进行层次聚类(选择“single”方法，阈值设置为 200 bp)，选择二峰的图形是因为通常 3 个峰中有一个低频率的峰未被检测到，这里的在横坐标 0 附近存在一个低频的峰 (图 1D)；为了进一步验证，按位置聚类所检测片段的可靠性，保留 500 bp 及以上的片段，并计算相邻区间的 CV，并按 CV 将相邻的区间连接在一起：如果相邻区间 CV 的差值在小于设定阈值 $F$ 则将相邻区间合并。为了获得合理的阈值 $F$，设定连续的阈值(0.05~0.40)，统计所有密度分布图，分别尝试不同阈值，具有 3 个峰最多时对应的 $F$ 值即为合理的阈值。此外，本研究也对绘制密度分布图所需的带宽(统计密度分布图像时划分的区段)进行训练和设置，以可靠地检测峰所在的位置并分离多峰，在密度差分时，带宽是通过 adjust*bw 来设置的，这里 bw 是通过公式 1.06*sd(save)*(length(save))^(−1.5)(save 是一组 read 值)来计算的，与上面类似，设置一系列不同的 adjust，选取三峰最多时的 adjust 值并同时提取出所有三峰的区段。对于 $CV \geqslant K$ 的情况，先对剩余的区段按物理位置进行层次聚类，采用“single”方法，并且阈值设置为 200 bp，挑出 500 bp 及以上的区段，并获得频率分布图；其次排除挑出的区段 read 总和较小的区段，即覆盖度较小的区段，方法是先将挑出的区段 read 值相加，画出核密度图，当曲线趋近平坦时，选取一个阈值，小于阈值的区段过滤掉；对频率分布图的带宽参数 adjust 进行训练，选取三峰最多时的 adjust 值，并提取出三峰的区段。

(C)1994-2024 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

### 3.3 方法的比较

XHMM 及 ExomeDepth 软件被选用检测银屑病数据的 CNV，最后，两种方法和新方法的区间与 24 个汉族人金标准比较重叠率(定义为与金标准重叠的窗口总数/对应个体金标准窗口总数)。

## 作者贡献

方铭为项目负责人，负责项目的总体思路、指导相关工作的开展与论文修改；姜丹负责数据处理过程中聚类脚本的编写；杨浩负责试验操作、数据分析及论文撰写。全体作者都阅读并同意最终的文本。

## 致谢

本研究由国家自然科学基金面上项目(31672399; 31872560)资助。

## 参考文献

Benjamini Y., and Speed T.P., 2012, Summarizing and correcting the GC content bias in high-throughput sequencing, Nucleic Acids Res., 40(10): e72.

De Groot N., Stanbury K., De Vos-Rouweler A.J., De Groot N.G., Poirier N., Blancho G., De Luna C., Doxiadis G.G., and Bontrop R.E., 2017, A quick and robust MHC typing method for free-ranging and captive primate species, Immunogenetics, 69(4): 231-240.

Ellingford J.M., Campbell C., Barton S., Bhaskar S., Gupta S., Taylor R.L., Sergouniotis P.I., Horn B., Lamb J.A., Michaelides M., Webster A.R., Newman W.G., Panda B., Ramsden S.C., and Black G.C., 2017, Validation of copy number variation analysis for next-generation sequencing diagnostics, Eur. J. Hum. Genet., 25(6): 719-724.

Fromer M., and Purcell S.M., 2014, Using XHMM software to detect copy number variation in whole-exome sequencing data, Curr. Protoc. Hum. Genet., 81: 7.23.1-7.23.21.

Fromer M., Moran J.L., Chambert K., Banks E., Bergen S.E., Ruderfer D.M., Handsaker R.E., McCarroll S.A., O'Donovan M.C., Owen M.J., Kirov G., Sullivan P.F., Hultman C. M., Sklar P., and Purcell S.M., 2012, Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth, Am. J. Human Genet., 91(4): 597-607.

Haraksingh R.R., Abyzov A., and Urban A.E., 2017a, Comprehensive performance comparison of high-resolution array platforms for genome-wide Copy Number Variation (CNV) analysis in humans, BMC Genomics, 18(1): 321.

Haraksingh R.R., Abyzov A., and Urban A.E., 2017b, Comprehensive performance comparison of high-resolution array platforms for genome-wide Copy Number Variation (CNV) analysis in humans, BMC Genomics, 18(1): 321.

Harrington C.L., Dey A.K., Yunus R., Joshi A.A., and Mehta N. N., 2017, Psoriasis as a human model of disease to study inflammatory atherogenesis, Am. J. Physiol. Heart Circ. Physiol., 312(5): 867-873.

Packer J.S., Maxwell E.K., O'Dushlaine C., Lopez A.E., Dewey F.E., Chernomorsky R., Baras A., Overton J.D., Habegger L., and Reid J.G., 2016, CLAMMS: a scalable algorithm for calling common and rare copy number variants from exome sequencing data, Bioinformatics, 32(1): 133-135.

Plagnol V., Curtis J., Epstein M., Mok K.Y., Stebbings E., Grigoriadou S., Wood N.W., Hambleton S., Burns S.O., Thrasher A.J., Kumararatne D., Doffinger R., and Nejentsev S., 2012, A robust model for read count data in exome sequencing experiments and implications for copy number variant calling, Bioinformatics, 28(21): 2747-2754.

Prans E., Kingo K., Traks T., Silm H., Vasar E., and Koks S., 2013, Copy number variations in IL22 gene are associated with psoriasis vulgaris, Hum. Immunol., 74(6): 792-795.

Seiser E.L., and Innocenti F., 2014, Hidden markov model-based CNV detection algorithms for Illumina genotyping microarrays, Cancer Inform., 13(S17): 77-83.

Wang X., and Byers S., 2014, Copy number variation in chickens: a review and future prospects, Microarrays, 3(1): 24-38.

Yao R.E., Zhang C., Yu T.T., Li N., Hu X.Y., Wang X.M., Wang J., and Shen Y.P., 2017, Evaluation of three read-depth based CNV detection tools using whole-exome sequencing data, Mol. Cytogenet., 10: 30.

Zhou F.S., Cao H.Z., Zuo X.B., Zhang T., Zhang X.G., Liu X.M., Xu R.C.,Chen G., Zhang Y.W., Zheng X.D., Jin X., Gao J. P., Mei J.P., Sheng Y.J., Li Q.B., Liang B., Shen JB., Shen C., Jiang H., Zhu C.H., FanX., Xu F.P., Yue M., Yin X.Y., Ye C., Zhang C.C., Liu X., Yu L.,Wu J.H., Chen M., Zhuang X., Tang LL., Shao H.J., Wu L.M., Li J.,Xu Y., Zhang Y.J., Zhao S.L., Wang Y., Li G., Xu H.S., Zeng L.,Wang J., Bai M.Z., Chen Y.L., Chen W., Kang T., Wu Y.Y., Xu X., Zhu Z.W., Cui Y., Wang Z.X., Yang C.J., Wang P.G., Xiang L.H.,Chen X., Zhang A.P., Gao X.H., Zhang F., Xu J. H., Zheng M.,Zheng J.N., Zhang J.Z., Yu X.P., Li Y.R., Yang S., Yang H.M., Wang J., Liu J.J., Hammarström L., Sun L.D., Wang J., and Zhang X.J., 2016, Deep sequencing of the MHC region in the Chinese population contributes to studies of complex disease, Nat. Genet., 48(7): 740-746.

Zhou F.S., Shen C.B., Hsu Y.H., Gao J., Dou J.F., Ko R., Zheng X.D., Sun L.D., Cui Y., and Zhang X.J., 2018, DNA methylation-based subclassification of psoriasis in the Chinese Han population, Front. Med., 12(6): 717-725.

(C)1994-2024 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net