doi:10.3969/j.issn.1672-5565.2014.04.08

基于SNP数据检测染色体拷贝数结果可信度分析

刘小成,刘元宁,夏　红,王明会,李　骜*
(中国科学技术大学电子科学与技术系,安徽 合肥 230027)

摘　要:利用SNP数据检测肿瘤细胞染色体拷贝数变异是癌症相关研究的一个热点,目前已有多种方法可以通过分析SNP array数据检测染色体拷贝数。然而在某些情况下,这些检测方法检测结果与真实拷贝数具有一定错误率。目前并没有方法研究预测结果发生错误的规律。本文分别分析了GPHMM,ASCAT两种检测方法结果信息熵与检测正确率的关系,发现检测正确率与信息熵存在很强的相关性。通过对比不同肿瘤细胞比例下信息熵与正确率关系,本文发现随着肿瘤细胞比例的增大,检测结果信息熵平均值增大,方差减小;同时平均检测正确率也越来越大,方差显著减小。这些结果显示信息熵的大小可以反映出检测结果正确率的高低。最后,本文以高肿瘤细胞比例下拷贝数检测结果为例,研究了在变异类型单一,信息熵小的情况下,染色体倍性检测的正确率。结果表明信息熵可以作为衡量检测结果可信度的指标:即信息熵越高,检测结果越可信。

关键词:生物信息学;SNP array;信息熵;检测结果可信度;拷贝数变异
中图分类号:TP302　　**文献标志码**:A　　**文章编号**:1672-5565(2014)-04-281-06

Confidence level analysis of detecting chromosome copy number detection using SNP array

LIU Xiaocheng, LIU Yuanning, XIA Hong, WANG Minghui, LI Ao*
(*Department of Electronic Science and Technology of China*, *University of Science and Technology of China*, *Anhui Hefei* 230027, *China*)

Abstract: Recently, using SNP arrays to detect chromosomal copy number aberrations of tumor cells gains its popularity. Several methods that devoted for copy number dissection have been proposed. However, there is no study being performed regarding the error rate of results of copy number detection comparing with true copy number profile. In this study, by using GPHMM and ASCAT, which are both devoted for copy number detection, examinations on the relationship between entropy and accuracy are conducted and results show that accuracy and entropy demonstrate a strong correlation. By testing the accuracy and entropy under different tumor cell proportions , results show that with the increase of the proportion of the tumor cells, average entropy of detection results become larger and the variance becomes smaller. Also, study finds that the average rate of correct detection is significantly increasing when the variance is decreasing, indicating that the proportion of tumor cells can affect the accuracy of detection and information entropy at the same time. At last, by taking an error detection case of tumor samples with high proportion of tumor cells, study shows that limited kinds of aberrations and small entropy are likely to cause the occurrence of serious bias in average copy number estimation. In conclusion, all results suggest that entropy can act as confidence level indicator for copy number detection: the higher entropy is likely to produce the better reliability regarding copy number detection.

Keywords: Bioinformatics; SNP arrays; Entropy; Confidence level of detection results; Copy number aberrations

拷贝数变异是指基因组上DNA片段发生缺失或倍增[1]。越来越多的研究表明,肿瘤细胞区别于正常细胞的一个重要特征就是染色体拷贝数变异[2],不仅如此,细胞染色体拷贝数变异还与相当

收稿日期:2014-06-18;**修回日期**:2014-07-07.
基金项目:国家自然科学基金(No.31100955)资助、中央高校基本科研业务费专项资金(No.WK2100230007);高等学校博士学科点专项科研基金(No. 20113402120024)资助。
作者简介:刘小成,男,硕士研究生,研究方向:生物信息学;E-mail:lxcheng@mail.ustc.edu.cn.
*通信作者:李骜,男,博士,副教授,博士生导师,研究方向:生物医学信息处理和生物信息学;E-mail: aoli@ustc.edu.cn.

部分的表型变异(Phenotypic variation)有关(Feuk et al. 2006; Freeman et al. 2006; Eichler et al. 2007; McCarroll and Altshuler 2007),因此细胞染色体拷贝数变异检测对于研究癌症发展具有重要意义[3-5]。随着基因芯片技术的发展,单核苷酸多态性微阵列芯片(Single nucleotide polymorphism array, SNP array)技术可以允许检测数万碱基对长度的 DNA 片段拷贝数变异[6]。目前已有多种利用 SNP array 数据检测肿瘤细胞染色体拷贝数变异的方法,比如 ASCAT[7],GPHMM[8],PennCNV[1]等。通过对信号数据进行统计建模,算法可以正确检测出染色体基因型,并估计样本中正常细胞的比例。但当样本中正常细胞比例过高时,由于肿瘤细胞信号过小,伴随着噪声的影响,这些方法的检测结果会发生错误[9]。当肿瘤细胞比例过低或变异类型很单一时,对于 SNP array 信号的分析则更具挑战性。然而,目前并无研究对这些预测模型检测结果的准确性及错误规律进行详细分析。

本研究中利用拷贝数检测结果信息熵作为标志,对检测结果进行分析。结果显示当信息熵很高时,检测结果很可信,当信息熵过低时,拷贝数检测结果会有很高的概率出错,需要添加先验信息或更改检测手段。

1 材料与方法

以 HapMap NA06991 正常细胞(平均染色体拷贝数为 2)常染色体 SNP array 数据为基础,通过划分变异片段并指定每个变异区段的隐马尔科夫状态(如表 1 所示,在 GPHMM[8]中定义),控制样本中肿瘤细胞比例并添加噪声[9],模拟产生肿瘤细胞 SNP array 数据。取该细胞常染色体 105 872 个 SNP 位点,模拟时将每条染色体每 3 000 个位点划分为一段,这样 22 条染色体被划分成 56 个变异区段。由于 SNP array 数据中的两种信号(即 BAF、LRR 信号)的方差一般在 0.03 与 0.3 左右,故模拟数据添加噪声时控制其方差分别为 0.03,0.3;肿瘤细胞比例在 20%,30%,40%直到 90%这 8 个比例之间随机选取。上述参数设定完成后,只需指定染色体每个变异区段的隐马尔科夫状态即可模拟产生肿瘤细胞 SNP array 数据。通过使用目前流行的 SNP array 信号分析算法,对样本的拷贝数进行检测[10-11]。这里采用了两种指定方法,模拟时每组样本平均拷贝数不高于 4。

(1)随机指定每个片段的变异的隐马尔科夫状态,任意两个片段之间隐含状态保持相互独立,共模拟 450 组;

(2)将 56 个片段划分为两类:主要部分和剩余部分。主要部分所有片段的隐含状态相同,剩余部分片段状态随机指定,保持相互独立。主要部分所包含的片段数目(0~56)随机指定,并随机取一个隐含状态作为共同状态,剩下区段即为剩余部分,这些片段在 56 个片段中的位置也是随机指定的,共拟 400 组;

表 1 隐含状态的定义

Table 1 The definition of hidden states

Hidden states	Copy number	(Tumor genotype, Normal genotype)
1	1	(A,AB),(B,AB),(A,AB),(B,AB)
2	2	(AA,AA),(BB,BB),(AB,AB)
3	2	(AA, AA), (AA, AB), (BB, BB), (BB, AB)
4	3	(AAA, AA), (BBB, BB), (AAB, AB), (ABB, AB)
5	3	(AAA, AA), (AAA, AB), (BBB, BB), (BBB, AB)
6	4	(AAAA, AA), (BBBB, BB), (AAAB, AB), (ABBB, AB)
7	4	(AAAA, AA), (BBBB, BB), (AABB, AB)
8	4	(AAAA, AA), (BBBB, BB), (AAAA, AB), (BBBB, AB)
9	5	(AAAAA, AA), (BBBBB, BB), (AAAAB, AB), (ABBBB, AB)
10	5	(AAAAA,AA),(BBBBB,BB),(AAAAB,AB), (ABBBB,AB)
11	5	(AAAAA, AA), (BBBBB, BB), (AAABB, AB), (AABBB, AB)
12	5	(AAAAA, AA), (BBBBB,BB), (AAAAA,AB), (BBBBB,AB)

使用第一种模拟方法是为了模拟肿瘤细胞的各种变异情形,全面的分析各种可能发生的变异;使用第二种模拟方法是为了模拟变异类型相对单一的情形,这种变异检测难度更大,可以展现拷贝数变异检测方法的检测能力。结合两种模拟手段,可以更全面的分析拷贝数检测方法[12],比较这些方法在不同情形下的检测正确率。

2　结果与讨论

分别利用 ASCAT[7] 和 GPHMM[8] 检测模拟产生的 850 组肿瘤 SNP 数据样本拷贝数,比较拷贝数检测结果与真实拷贝数,计算检测准确率与结果信息熵(检测结果每种拷贝数的比例作为这种拷贝数出现的概率)。用这两种方法检测第 1 种模拟数据样本拷贝数变异,发现这两种检测方法检测平均准确率均超过 95%,在随机模拟下的高准确率说明 GPHMM,ASCAT 检测拷贝数变异具有相当高的可靠性。然而在检测模拟方法 2 中样本时发现,GPHMM 有 57 例(占模拟 2 的 14.25%)检测准确率低于 5%,有 141 例(35.25%)检测准确率低于 50%;ASCAT 有 28 例(占模拟 2 的 7%)无法给出结果,同时有 63 例(占模拟 2 的 15.75%)检测准确率低于 5%,这说明 GPHMM,ASCAT 在肿瘤细胞变异类型相对单一情形下,检测结果不够准确,甚至有可能出现整个细胞染色体变异均检测错误的严重后果。

另一方面,利用 GPHMM 与 ASCAT 检测模拟方法 1 和 2 产生 SNP 数据拷贝数变异,结合两种模拟下检测方法的差异表现,可以对检测准确率与检测结果信息熵之间的关系进行研究。如图 1a 所示,信息熵分布在 0~2.5 之间,GPHMM 检测结果准确率有明显随着信息熵提高而升高的趋势,ASCAT 则两级分化,基本集中在准确率 0 与 1 附近。这两种方法在信息熵很小的时候检测正确率落在 0 或 1 附近,表现不稳定。将信息熵从 0 开始以 0.2 为间隔分为 12 段,计算样本检测结果信息熵落在每段内检测正确率高于 90%所占比重。图 1b 可以看出,随着信息熵增大,该占比显著增高,直到接近 100%。需要注意的是,ASCAT 在 0.6~0.8 这一区段内平均正确检测率有一个反常的增大,这是由于该范围内样本数目过少导致。总体上,比较两种检测方法结果,当信息熵高于 2.0 时,检测正确占比超过 90%,检测结果可靠,当信息熵小于 1 时,检测结果还有较高频率的错误出现,此时检测方法给出的结果置信度较低。从 GPHMM,ASCAT 结果趋势图(图 1b)中可以看出信息熵与检测正确率之间有很强的正相关性:信息熵越高,检测准确率越高。这样检测结果信息熵可以作为衡量拷贝数检测结果可信度的标准,在检测方法给出检测结果时,判断该结果是否可信。信息熵越高,结果越可信,当信息熵很低时,检测结果不可信。

进一步研究了两种造成检测结果信息熵偏低的样本:样本本身肿瘤细胞比例较低,肿瘤细胞信号被噪声掩盖[13-14],变异区域很容易被检测成正常状态[9],结果信息熵很小;样本本身变异类型较少或变异的区段长度很短,这种情形下真实拷贝数与结果信息熵都很小。

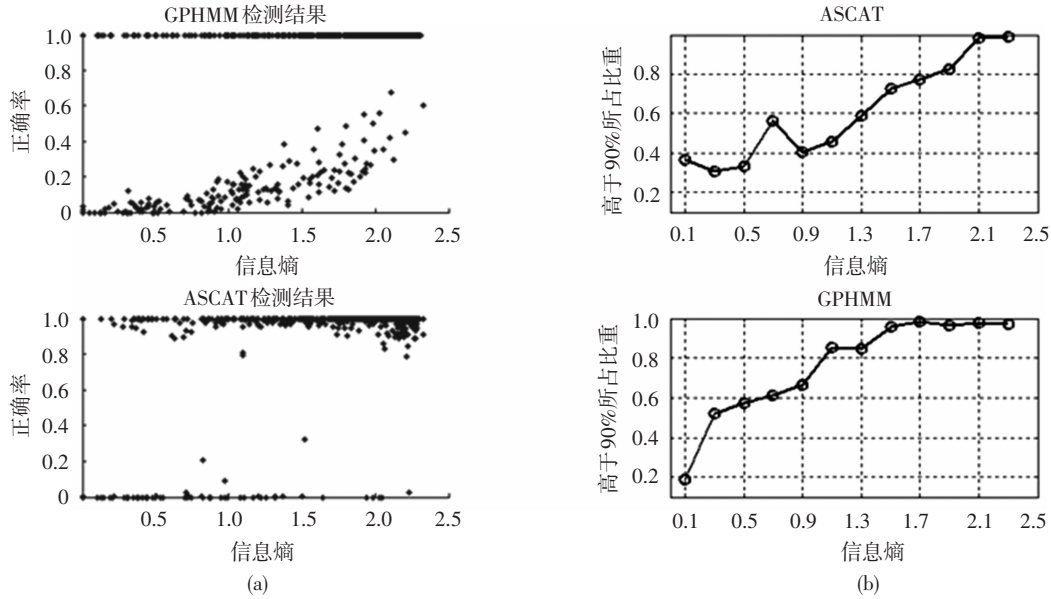

图 1　检测结果

Fig.1　Petection results

注:(a)为 GPHMM,ASCAT 检测正确率与检测结果信息熵关系图;(b)为 GPHMM,ASCAT 结果信息熵以 0.2 为间隔,每段检测正确率高于 90%所占比重折线图,横坐标代表每个信息熵间隔的中值。

Notes: (a) refers to relationship between GPHMM, ASCAT detection accuracy and detection results entropy respectively; (b) refers to dividing entropy intervals of 0.2, the proportion of detection accuracy bigger than 90% in every interval distribution curve. Abscissa stands for mid-value entropy of every interval.

首先研究肿瘤细胞比例对检测结果信息熵的影响。图 2 结果显示,肿瘤比例 0.2,0.5 正确率分布中,比例为 0.5 的样本信息熵分布在 0~2.5 之间,正确率基本都落在 1 附近,并且大部分都集中在右上,即信息熵,正确率均较高的区域,说明在这个比例下检测结果很准确;比例为 0.2 的样本则大部分分布在信息熵在 1 附近,且检测正确率较低区域,这说明很多样本拷贝数都检测错误了。对比肿瘤比例 0.2,0.4两种正确率分布,比例为 0.4 样本信息熵有明显增大趋势,并且检测正确率在 1 附近的样本数目更多。比较 0.2,0.3,0.4,0.5 四种肿瘤细胞比例正确率散点图,可以看出随着肿瘤比例的减小,检测结果信息熵有减小的趋势,同时检测正确率也随之减少,当肿瘤比例低到 20%时,检测结果已明显出现很多错误。这种趋势变化说明信息熵可以反映样本的比例,进而反应检测正确率。

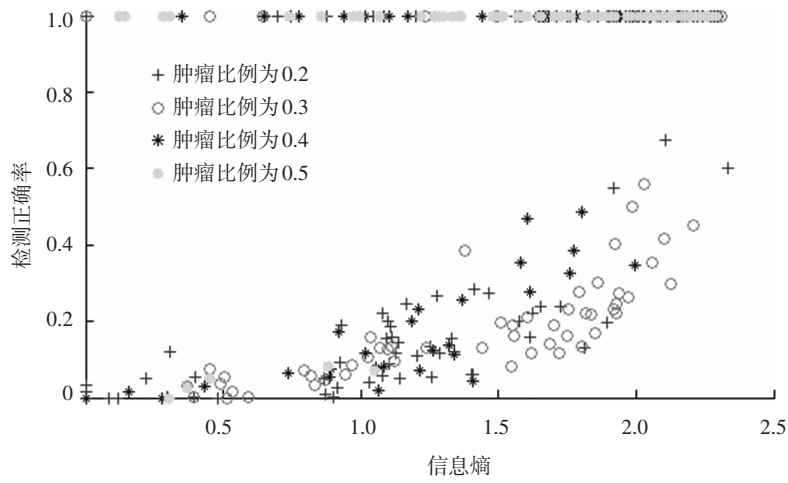

图 2 GPHMM 检测四种肿瘤比例正确率-信息熵关系图

Fig.2 Detection accuracyof GPHMM with respect to four different proportions of cancer cells

注:这四种比例分别为:20%, 30%, 40%, 50%。

Notes: Four different proportions of cancer cell: 20%, 30%, 40%, and 50%.

进一步计算每种肿瘤比例样本平均检测正确率,平均检测结果信息熵,结果如图 3 所示。GPHMM 在肿瘤比例达到 40%时,平均检测准确率很高,方差也明显小于肿瘤比例较低的样本,此时绝大部分变异区段都能正确率的检测出来。在肿瘤比例为 20%时,会有部分肿瘤样本变异检测出错,此时结果信息熵均值低于 1.6,并且平均准确率与信息熵的方差都很大,GPHMM 表现不稳定,置信度不高。比较图 3 左右两图,结果显示随着肿瘤细胞比例增加,信息熵均值逐渐增大,方差逐渐减小,同时检测正确率均值不断提高,且方差越来越小,这表明肿瘤比例越高检测越准确[8],同时信息熵也在逐渐提升,这表明信息熵可以反应样本肿瘤的比例进而反应检测正确率,可以用来度量检测结果是否准确。

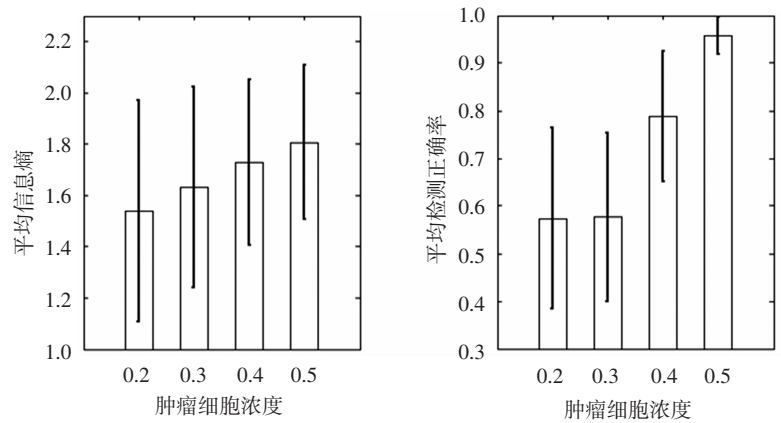

图 3 四种不同肿瘤比例样本信息熵与检测正确率方差棒图

Fig.3 Entropy and GPHMM detection accuracy error bar for four different proportions tumor cell samples

注:柱状图代表平均信息熵或平均检测正确率,竖条代表方差。

Notes: The bars present average entropy or average accuracy rate, the vertical lines represent the variance.

同时,从图 2 可以看出,即使是在高比例(0.5)时,仍有个别检测结果信息熵很小,准确率很低的样本出现。这是因为此时样本变异类型过少或区域很短,真实拷贝数的信息熵很小,此时检测结果信息熵也会很小,会导致许多片段拷贝数检测出错。图 4 给出了其中一个样本的一条染色体,结果显示整条染色体除最后一小段拷贝数估测正确,绝大部分区域估测错误。

这是因为这个样本的变异类型比较单一(绝大部分区域拷贝数为 5,只有一小段区域拷贝数为 4),GPHMM 检测结果信息熵相应的会变得非常小,将倍性 5 错误地估计为 3,最终造成绝大部分检测出错。这说明即使是在肿瘤比例很高的时候,信息熵较小的时候对于肿瘤基因组的倍性估计具有一定难度,最终导致整个样本拷贝数检测错误。

如上所述,本文论述了肿瘤细胞比例与变异类型数目对检测结果信息熵的影响,在肿瘤细胞比例过低,变异类型单一或变异长度过短时,利用 SNP array 数据检测肿瘤细胞染色体拷贝数变异非常困难,此时检测结果信息熵一般呈现偏小的趋势,这也就阐述了信息熵可以作为检测准确率的判断标准的原因。

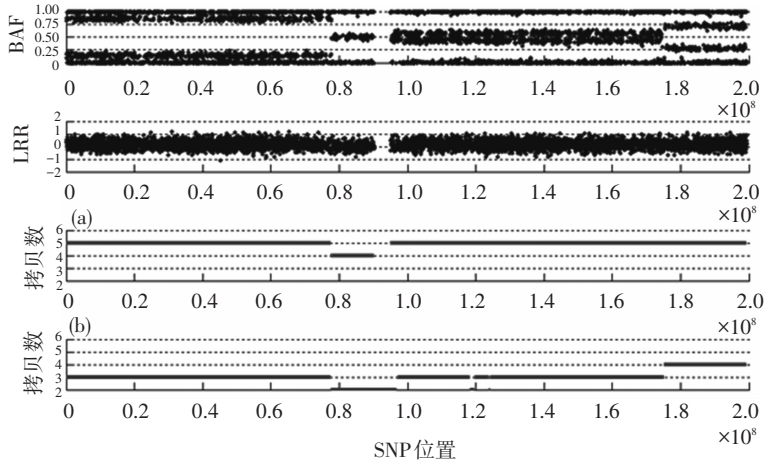

图 4　第 141 个样本第三条染色体 BAF,LRR,真实拷贝数,GPHMM 检测结果拷贝数分布图

Fig.4　BAF, LRR, True copy number, GPHMM detection copy number distribution diagram of chromosome 3 No.141 sample

注:(a)真实拷贝数分布图;(b)GPHMM 检测结果拷贝数分布图,这个样本真实肿瘤比例为 50%。

Notes:(a)refers to real copy number distribution; (b)stands for GPHMM detection results, the proportion of tumor cells in the sample is 50%.

2　结　论

通过研究 GPHMM,ASCAT 两种利用 SNP array 数据检测肿瘤细胞染色体拷贝数变异方法,本文发现检测结果信息熵与检测正确率之间存在很强的相关性,提出了将信息熵作为判断检测结果是否可信的标志:信息熵越高,检测结果越可信,当信息熵很低时,检测结果有很高的概率出错,这也为检测结果准确性的判断提出来一种有效的分析框架。随着 NGS(Next Generation Sequencing)技术发展[15],利用 NGS 数据检测拷贝数方法也越来越多[16-19],同时也面临着一些挑战[20],本文中提出的方法也可以进一步推广到 NGS 数据分析中。

参考文献(References)

[1] WANG K, LI M, HADLEY D, et al. PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data [J]. Genome Research, 2007, 17(11): 1665-1674.

[2] ALBERTSON D G, COLLINS C, MCCORMICK F, et al. Chromosome aberrations in solid tumors [J]. Nature Genetics, 2003, 34(4): 369-376.

[3] SCHAAF C P, WISZNIEWSKA J, BEAUDET A L. Copy number and SNP arrays in clinical diagnostics [J]. Annual Review of Genomics and Human Genetics, 2011, 12: 25-51.

[4] GIRIRAJAN S, CAMPBELL C D, EICHLER E E. Human copy number variation and complex genetic disease [J]. Annual Review of Genetics, 2011, 45: 203-226.

[5] NAVE C, KEREN B, MIGNOT C, et al. Prospective diagnostic analysis of copy number variants using SNP microarrays in individuals with autism spectrum disorders [J]. European Journal of Human Genetics, 2013, 22(1): 71-78.

[6] PEIFFER D A, LE J M, STEEMERS F J, et al. High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping [J]. Genome Research, 2006, 16(9): 1136-1148.

[7] VAN LOO P, NORDGARD S H, LINGJÆRDE O C, et al. Allele-specific copy number analysis of tumors [J]. Proceedings of the National Academy of Sciences, 2010, 107(39): 16910-16915.

[8] LI A, LIU Z, LEZON-GEYDA K, et al. GPHMM: an integrated hidden Markov model for identification of copy number alteration and loss of heterozygosity in complex tumor samples using whole genome SNP arrays [J]. Nucleic Acids Research, 2011, 39(12): 4928-4941.

[9] YAU C, MOURADOV D, JORISSEN R N, et al. A statistical approach for detecting genomic aberrations in heterogeneous tumor samples from single nucleotide polymorphism genotyping data [J]. Genome Biology, 2010, 11(9): R92-R92.

[10] SUN W, WRIGHT F A, TANG Z, et al. Integrated study of copy number states and genotype calls using high-density SNP arrays [J]. Nucleic Acids Research, 2009, 37(16): 5365-5377.

[11] ECKEL-PASSOW J E, ATKINSON E J, MAHARJAN S, et al. Software comparison for evaluating genomic copy number variation for Affymetrix 6.0 SNP array platform [J]. BMC Bioinformatics, 2011, 12(1): 220.

[12] MOSÉN-ANSORENA D, ARANSAY A M, RODRÍGUEZ-EZPELETA N. Comparison of methods to detect copy number alterations in cancer using simulated and real genotyping data [J]. BMC Bioinformatics, 2012, 13(1): 192.

[13] WHEELER E, HUANG N, BOCHUKOVA E G, et al. Genome-wide SNP and CNV analysis identifies common and low-frequency variants associated with severe early-onset obesity [J]. Nature Genetics, 2013, 45(5): 513-517.

[14] TREFF N R, SU J, TAO X, et al. Single-cell whole-genome amplification technique impacts the accuracy of SNP microarray-based genotyping and copy number analyses [J]. Molecular Human Reproduction, 2011, 17(6): 335-343.

[15] MARDIS E R. The impact of next-generation sequencing technology on genetics [J]. Trends in Genetics, 2008, 24(3): 133-141.

[16] YOON S, XUAN Z, MAKAROV V, et al. Sensitive and accurate detection of copy number variants using read depth of coverage [J]. Genome Research, 2009, 19(9): 1586-1592.

[17] BOEVA V, POPOVA T, BLEAKLEY K, et al. Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data [J]. Bioinformatics, 2012, 28(3): 423-425.

[18] CONWAY C, CHALKLEY R, HIGH A, et al. Next-generation sequencing for simultaneous determination of human papillomavirus load, subtype, and associated genomic copy number changes in tumors [J]. The Journal of Molecular Diagnostics, 2012, 14(2): 104-111.

[19] BICKHART D M, HOU Y, SCHROEDER S G, et al. Copy number variation of individual cattle genomes using next-generation sequencing [J]. Genome Research, 2012, 22(4): 778-790.

[20] TEO S M, PAWITAN Y, KU C S, et al. Statistical challenges associated with detecting copy number variations with next-generation sequencing [J]. Bioinformatics, 2012, 28(21): 2711-2718.