

一种胎儿染色体异常的检测方法与系统

技术领域

本发明涉及生物技术领域，更具体而言，涉及一种胎儿染色体异常的检测方法与系统。

背景技术

染色体非整倍体疾病是指胎儿体内个别染色体数目出现增加或减少，从而影响正常基因表达的一类严重的遗传疾病，主要包括21-三体综合征、18-三体综合征、13-三体综合征、5p-综合征等疾病等。染色体非整倍体疾病具有较高的致死、致残风险，且没有有效的治疗方法，目前主要是通过产前筛查和产前诊断降低染色体非整倍体疾病患儿出生率。

传统的染色体非整体检测主要包括基于超声诊断学检查或血清学筛查的无创产前筛查和有创取样的产前诊断。基于超声诊断学检查的产前筛查方法通过检查10-14孕周胎儿的颈项透明带（NT）的厚度来判断胎儿染色体是否异常，一般认为NT大于3毫米则胎儿染色体异常风险较高。基于血清学的产前筛查是在13-16孕周时通过检测母体血清中甲型胎儿蛋白（AFP）和绒毛促性腺激素（HCG）的浓度，结合孕妇预产期、年龄和采血时孕周计算胎儿染色体异常的风险系数。基于有创取样的产前诊断方法一般在16-24孕周通过羊水穿刺、脐血穿刺或绒毛膜直接取样获取胎儿样本，来检测胎儿是否出现染色体异常。基于超声诊断学检查和血清学的联合筛查方式不是直接检测胎儿染色体而是推测胎儿患病风险，其检测准确性50%~95%，假阳性率高达3%-7%[1,2]。基于有创取样的方法可直接对胎儿非整倍体进行准确诊断，是胎儿染色体异常检测和诊断的“金标准”，但该方法会导致一定的流产率（0.5%~2%），且患有乙肝等传染病的孕妇因有传染胎儿的风险不适宜做有创取样（如羊水穿刺）。另外，做羊水穿刺需要在B超引导下进行，需要的时间较长，对操作人员的技术要求很高。

随着母体外周血中胎儿游离DNA（cell-free DNA, cfDNA）的发现和新一代高通量测序（Next Generation Sequencing, NGS）技术的成熟、测序成本的显著下降及信息分析技术的发展，基于NGS技术的无创产前检测（Noninvasive Prenatal Testing, NIPT）正成为应用最为广泛的胎儿染色体非整倍体疾病的产前筛查手段。NIPT技术通过采取孕妇外周血，利用NGS技术对母体外周血浆中的游离DNA（包含胎儿游离DNA）进行测序，结合生物信息学分析得到胎儿遗传信息，从而检测胎儿是否患有如21-三体综合征（唐氏综合征）、18-三体综合征（爱德华氏综合征）、13-三体综合征（帕陶氏综合征）等染色体异常疾病。

NIPT 技术有着高敏感性和特异性（T21、T18 和 T13 的灵敏性均在 99%以上）以及较低假阳性率（<0.1%），现已广泛应用于临床[3-5]。NIPT 技术可以降低血清学筛查的假阳性率，避免由于侵入性产前诊断操作（如羊膜穿刺、绒毛膜取样等）带来的胎儿宫内感染及流产风险，是一项安全性高的早中孕期无创性非侵入性的产前筛查技术。

常规基于NGS技术的NIPT通过计算测序的读段数量（Read Count）并使用基准Z检验（Baseline Z-test）来检测胎儿染色体异常[6]。其原理为：首先采取12-22孕周的母体外周血样品，利用NGS技术对外周血样品中游离DNA进行测序并将获取的测序读段与人类参考基因组序列进行比对（同时对GC含量进行读段数量校正）；然后统计每条染色体唯一比对的读段数量并计算其占该样本所有染色体唯一比对的读段数量的比例；接着利用染色体上唯一比对读段的比例减去对照组样品（即正常样品）中对应染色体唯一比对的读段数量的比例的均值，再除以对照组样本中对应染色体唯一比对的读段数量的比例的标准差，得到待检测样本的该染色体的Z值；最后将该Z值与给定阈值比较，超过阈值则判断为存在三体综合征高风险，反之则判断为低风险。这里对照组中正常样品中每条染色体唯一比对的读段

数量的均值即为基准值（Baseline Value）。由此可见，对照组中正常样品数越多，唯一比对读段比例的均值和标准差就越准确，于是得到的Z值也更准确。这里Z值的给定阈值一般是3，这是从统计学上定义的，即99.9%偏离正常预期。

根据基准值不同可以选择不同的统计假设检验。例如，文献[7]中采用的是相关性分析和T检验，其基准值是使用样品中每条染色体在固定大小窗口里的读段数量的中位数代表该条染色体的读段数量，而且使用该样本中所有染色体的读段数量的中位数代表该样品的读段数量；然后使用每条染色体的读段数量除以样品的读段数量得到归一化（Normalization）的对应染色体的读段数量；最后利用对照组中所有样品的每条染色体的归一化的读段数量计算置信区间，当待测样品的值不在此置信区间时则为异常样品。再例如，文献[8]中提出了在已知核型的样品中，对感兴趣的染色体（如21号染色体）找到一个GC含量与之相似的参考染色体，并且以该参考染色体的读段数量为基准值进行Z检验后使得在已知核型的样品检测感兴趣染色体异常的准确率达到最大。这个作为基准值的参考染色体就是所谓的内部染色体（Internal Chromosome）。又例如，文献[9]提出了无创胎儿染色体非整倍体检测（Noninvasive Fetal Trisomy, NIFTY）方法。该方法除了对染色体的读段数量与对照组正常样品进行比较，还考虑了胎儿游离DNA比例。该方法使用了二元假设检验、对数似然比以及FCAPS二值分割算法对检测结果进行判定。NIFTY是一种基于全基因组的方法。该方法通过大人群验证，准确度高，但流程相对复杂。以上所述的基于读段数量的统计假设检验（Z检验或T检验）方法是目前NIPT分析的核心。

前面所述基于读段数量的统计假设检验（如Z检验）的分析方法是目前主流NIPT分析方法，但是这些分析方法具有明显局限性：（1）目前的NIPT分析方法会导致个体样本测序读段分布的偏差，从而导致Z值计算出现不同情况的波动，从而影响最终的结果判定和相关性能指标；（2）目前的NIPT分析方法高度依赖母体外周血浆中胎儿游离DNA的比例，由于孕妇个体差异较大，过低胎儿游离DNA比例（<4%）会增加假阴性的错误检测风险；（3）目前的NIPT分析方法对21-三体综合征的检测表现良好，但是由于孕妇个体差异和GC含量在不同染色体上的偏差，其对18-三体综合征和13-三体综合征检测的准确率较差；（4）目前的NIPT分析方法主要检测以唐氏综合征为代表的常见三体综合征，对检测综合发病率同样较高的染色体微缺失重复综合征，如迪格奥尔格综合征（DiGeorge Syndrome）、小胖威利综合征（Prader-Willi Syndrome）等的临床作用有限[14]。

此外，有人还提出了基于机器学习模型利用NIPT测序结果检测染色体异常的新技术。例如，文献[10]提出了使用支持向量机（Support Vector Machine, SVM）辅助NIPT判定的方法。该方法通过计算不同的基准值得到6种不同的Z值结果，另外加入样本的临床指征训练SVM模型进行染色体异常判定。再例如，文献[11]设计了一种贝叶斯方法进行染色体异常判定。该方法利用胎儿游离DNA比例先验信息，使用隐马尔科夫模型（Hidden Markov Model, HMM）剔除群体水平和母体CNV的干扰，并进行GC含量校正，结合Z检验似然值和由性染色体含量推测出的胎儿游离DNA比例先验值计算贝叶斯因子，同时将孕妇年龄等多项风险因素纳入先验概率对贝叶斯因子进行校正，综合Z值及贝叶斯因子对染色体是否异常进行评估。又例如，专利公开文本[12]提出了利用NIPT测序结果训练简单卷积神经网络模型检测染色体拷贝数变异和染色体非整倍体异常。还例如，专利公开文本[13]提出首先从外周血样品中分离出胎儿游离DNA和母体游离DNA，从分离的游离DNA扩增多个单核苷酸变异（SNV）基因座，对扩增产物进行测序以确定多个SNV基因座的遗传测序数据或遗传阵列数据，然后依据这些基因座的遗传测序数据或遗传阵列数训练人工神经网络模型以检测个体染色体的倍体水平状态（Ploidy State）、组织癌变（Cancer State）或器官移植的排斥情况（Transplantation Rejection State）。

前述的基于机器学习模型利用NIPT测序结果检测染色体异常同样存在以下几个局限：

这些方法大多数基于测序数据的读段数量来计算作为模型训练所需特征；大多数依赖对Z值的计算；要么计算太过复杂(如文献[11])，要么模型设计的太过简单(如专利公开文本[12])，要么需要基于SNV基因座的遗传测序数据或遗传阵列数据(如专利公开文本[13])，其临床应用前景、模型可扩展性和检测准确性有限；检测准确率都有待提高。

发明内容

鉴于现有技术对染色体异常，特别是非整倍体的检测存在问题，为了更加有效地对染色体异常进行检测，本发明的目的至少在于基于深度混合模型进一步提升对于染色体异常的检测准确率。

因此，在第一方面，本发明提供了一种检测胎儿染色体异常的方法，所述方法包括：

(1) 获得待测孕妇的游离核酸片段的测序数据和临床表型特征数据，其中所述测序数据包括若干读段，所述待测孕妇的临床表型特征数据形成孕妇表型特征向量；

(2) 对参考基因组染色体的至少一部分进行窗口划分，获得若干滑动窗口，统计落入所述滑动窗口内的读段，生成所述染色体序列的序列特征矩阵；

(3) 将所述序列特征矩阵输入训练的机器学习模型以提取所述染色体序列的序列特征向量；

(4) 将所述序列特征向量和孕妇表型特征向量组合形成组合特征向量，将所述组合特征向量输入分类检测模型，获得所述待测孕妇的胎儿染色体异常情况。

在一个实施方案中，在(1)中，所述游离核酸片段来自于孕妇外周血浆、孕妇肝脏和/或胎盘。

在一个实施方案中，在(1)中，所述游离核酸片段为游离DNA。

在一个实施方案中，在(1)中，所述测序数据来自超低深度测序；优选地，所述超低深度测序的测序深度为 $1\times$ 、 $0.1\times$ 或 $0.01\times$ 。

在一个实施方案中，在(1)中，将所述读段与参考基因组进行比对，获取唯一比对的读段(优选进行GC含量校正)；优选地，后续步骤使用所述唯一比对的读段(优选经GC含量校正后的读段)进行。

在一个实施方案中，GC含量校正按如下步骤进行：

a. 首先从人类参考基因组的某个染色体上随机采样 m 个长度为 ℓ 的片段；

b. 计算GC含量为 i 的片段数量 N_i ：

$$N_i = \sum_{k=1}^m \mathbb{I}_i(f(k))$$

其中 $\mathbb{I}_i(f(k)) = \begin{cases} 1, & f(k) = i \\ 0, & \text{其他} \end{cases}$ ， $f(k)$ 是片段 k 的GC含量， i 表示GC含量($i =$

0%, 1%, ..., 100%)；

c. 计算GC含量为 i 的测序读段数量 F_i ：

$$F_i = \sum_{k=1}^m c_k \cdot \mathbb{I}_i(f(k))$$

其中 $c_k \cdot \mathbb{I}_i(f(k))$ 表示片段 k 的GC含量， F_i 表示GC含量为 i 且测序读段起始点位与该片段起始点位相同的测序读段的数量；

d. 计算GC含量观测期望比 λ_i ：

$$\lambda_i = \begin{cases} r \cdot \frac{F_i}{N_i}, & F_i > 0 \text{ 且 } N_i > 0 \\ 1, & \text{其他} \end{cases}$$

其中 r 是全局比例因子（Global Scaling），其定义为：

$$r = \frac{\sum_i N_i}{\sum_i F_i}$$

e. 校正测序读段数量：

$$R_i = \frac{1}{\lambda_i}$$

其中 R_i 表示经校正后GC含量为 i 的测序读段的期望数量。

在一个实施方案中，在（1）中，孕妇表型特征数据选自如下的一个或多个的组合：年龄、孕周、身高、体重、BMI、产检生化检验检测结果、超声检查诊断结果和血浆中胎儿游离DNA浓度。

在一个实施方案中，在（1）中，所述孕妇表型特征数据经异常值处理、缺失值处理和/或空值处理。

在一个实施方案中，在（1）中，孕妇样本的表型数据出现如下记录将被判定为异常值，并将这些异常值设置为空值：

- a. $x_{age} < 10$ 或 $x_{age} > 80$ ；
- b. $x_{GW} < 5$ 或 $x_{GW} > 50$ ；
- c. $x_{height} < 40$ 或 $x_{height} > 300$ ；
- d. $x_{weight} < 10$ 或 $x_{weight} > 200$ 。

在一个实施方案中，缺失值和空值使用missForest算法进行填充。

在一个实施方案中，在（2）中，所述染色体为21号染色体、18号染色体、13号染色体和/或性染色体。

在一个实施方案中，（2）中包括，（2.1）使用长度为 b 的窗口以步长 t 对参考基因组上的长度为 L 的染色体序列进行重叠滑动，获得若干滑动窗口，其中 b 为正整数，且 $b=[10000,10000000]$ ， t 为任意正整数， L 为正整数，且 $L \geq b$ ；（2.2）统计落入每一个所述滑动窗口内的读段，生成所述染色体序列的序列特征矩阵。

在一个实施方案中，在（2）中，所述序列特征矩阵包括滑动窗口内读段数量、碱基质量和比对质量。

在一个实施方案中，所述碱基质量包括碱基质量的均值、标准差、偏度和/或峰值。

在一个实施方案中，所述比对质量包括比对质量的均值、标准差、偏度和/或峰值。

在一个实施方案中，在（2）中，所述序列特征矩阵为：

$$\mathbf{X} = (x_{ij})_{h \times w}$$

其中 h 表示窗口的滑动次数， w 表示单个滑动窗口内的序列特征数量， x_{ij} 表示第 i 个滑动窗口中第 j 个序列特征值。

在一个实施方案中，在（3）中，将所述序列特征矩阵进行标准化处理。

在一个实施方案中，在（3）中，使用公式（1）进行所述序列特征矩阵的标准化：

$$z_{i,j}^{(k)} = \frac{x_{i,j}^{(k)} - \mu_{i,j}}{\sigma_{i,j}} \quad (1)$$

其中， $\mathbf{z}_{i,j}^{(k)}$ 为样本 k 经标准化处理后的序列特征矩阵， $\mathbf{x}_{i,j}^{(k)}$ 表示样本 k 的第 i 个滑动窗

口中的第j个序列特征值, $\mu_{i,j}$ 和 $\sigma_{i,j}$ 分别表示所有样本的第i个滑动窗口中的第j个序列特征值的均值和标准差。

在一个实施方案中,在(3)中,所述训练的机器学习模型为神经网络模型或AutoEncoder模型;优选地,所述神经网络模型为深度神经网络模型;更优选地,所述神经网络模型为基于1维卷积的深度神经网络模型。

在一个实施方案中,所述深度神经网络模型的结构包括:

输入层,用于接收所述序列特征矩阵;

前置模块,与所述输入层相连,用于对来自所述输入层的序列特征矩阵进行第一次卷积和激活操作,得到特征图谱;

核心模块,与所述前置模块相连,用于对来自所述前置模块的特征图谱进行进一步抽象和特征提取,通过有效增加神经网络模型深度从而加强神经网络表达能力;

后置模块,与所述核心模块相连,用于对来自所述核心模块的特征图谱进行特征抽象表示;

第一全局平均池化层,与所述后置模块相连,用于将所述特征抽象表示的特征图谱进行向量化,并输出所述染色体序列的序列特征向量。

在一个实施方案中,所述前置模块包括:

(I) 1维卷积层(1D Convolution);

(II) 批标准化层(Batch Normalization),与(I)中所述1维卷积层相连;

(III) ReLU激活层,与(II)中所述批标准化层相连。

在一个实施方案中,所述核心模块由一个或多个结构相同的残差子模块组成,其中每个残差模块的输出是下一个残差模块的输入。

在一个实施方案中,所述残差子模块包括:

(A) 核心模块的前置子模块,每个子模块包括1维卷积层、与1维卷积层相连的Dropout层、与Dropout层相连的批标准化层,以及与批标准化层相连的ReLU激活层;

(B) 第一1维平均池化层(1D Average Pooling),与(A)中所述的核心模块的前置子模块相连;

(C) Squeeze-Excite模块(SE模块),和/或 Spatial Squeeze-Excite 模块(sSE模块),与(B)中所述的第一1维平均池化层相连;

(D) 第一加法层(Add层),与(C)中所述的SE模块和/或sSE模块相连;

(E) 第二1维平均池化层,与所述前置模块中的ReLU激活层相连;

(F) 第二加法层(Add层),与(D)中所述的第一加法层和(E)中所述的第二1维平均池化层相连。

在一个实施方案中,所述SE模块包括:

(a) 第二全局平均池化层,与所述残差子模块中的(B)第一1维平均池化层相连;

(b) Reshape层,与(a)中所述的第二全局平均池化层相连,所述Reshape层输出特征图大小为 $1 \times f$,其中 f 为1维卷积核数量;

(c) 第一全连接层,与(b)中所述Reshape层相连,所述第一全连接层输出神经元个数为 $\frac{f}{r_{SE}}$,其中 f 为1维卷积核数量, r_{SE} 为SE模块下降速度;

(d) 第二全连接层,与(c)中所述第一全连接层相连,所述第二全连接层输出神经元个数为 f ,其中 f 为1维卷积核数量;

(e) Multiply层,与(d)中所述第二全连接层、以及与残差子模块中(B)第一1维平均池化层相连。

在一个实施方案中,所述sSE模块包括:

a. 一个 1×1 的1维卷积层, 与(B)第一1维平均池化层相连, 该 1×1 的1维卷积层使用sigmoid函数作为激活函数;

b. 一个Multiply层, 与之连接的是(B)第一1维平均池化层和a中所述的 1×1 的1维卷积层。

在一个实施方案中, 在(4)中, 将所述序列特征向量与所述孕妇表型特征向量拼接得组合特征向量。

在一个实施方案中, 在(4)中, 所述组合特征向量 x 经标准化处理:

$$x'_i = \frac{x_i - \mu_i}{\sigma_i}$$

其中 x'_i 是标准化后的组合特征向量 x 中的第 i 个序列特征值, x_i 是组合特征向量 x 中的第 i 个序列特征值, μ_i 是组合特征向量中第 i 个序列特征值的均值, σ_i 是组合特征向量中第 i 个序列特征值的标准差。

在一个实施方案中, 在(4)中, 所述分类检测模型为集成学习模型。

在一个实施方案中, 所述集成学习模型为基于堆叠(Stacking)或多数投票(Majority Voting)的集成学习模型; 优选地, 所述集成学习模型为以下的一项或多项: 支持向量机模型、朴素贝叶斯分类器、随机森林分类器、XGBoos和逻辑回归。

在一个实施方案中, 所述染色体异常包括以下至少一项或多项: 21-三体综合征、18-三体综合征、13-三体综合征、5p-综合征、染色体微缺失和染色体微重复。

在第二方面, 本发明提供了一种构建用于检测胎儿染色体异常的分类检测模型的方法, 所述方法包括:

(1) 获得孕妇的游离核酸片段的测序数据和临床表型特征数据, 其中所述测序数据包括若干读段, 并且所述孕妇的胎儿染色体情况已知, 所述孕妇的临床表型特征数据形成孕妇表型特征向量;

(2) 对参考基因组染色体的至少一部分进行窗口划分, 获得若干滑动窗口, 统计落入所述滑动窗口内的读段, 生成所述染色体序列的序列特征矩阵;

(3) 利用所述序列特征矩阵和胎儿染色体情况构建训练数据集, 训练机器学习模型以提取所述染色体序列的序列特征向量;

(4) 将所述序列特征向量和孕妇表型特征向量组合形成组合特征向量, 利用所述组合特征向量和胎儿染色体情况训练分类模型, 获得训练后的所述分类检测模型。

在一个实施方案中, 所述孕妇的胎儿染色体情况为以下的一项或多项: 正常二倍体、染色体非整倍性、部分单体综合征、染色体微缺失和染色体微重复。

在一个实施方案中, 所述染色体非整倍性包括以下至少一项或多项: 21-三体综合征、18-三体综合征和13-三体综合征。

在一个实施方案中, 所述部分单体综合征包括5p-综合征。

在一个实施方案中, 所述孕妇数量大于10个, 并且正常二倍体的胎儿数量与染色体非整倍性的胎儿数量之比为 $\frac{1}{2}$ 至2。

在一个实施方案中, 在(3)中, 所述训练数据集表示为:

$$D_{seq} = \{Z_{i,j}^{(k)}, y^{(k)}\}_{k=1}^N,$$

$$y^{(k)} = \begin{cases} 1, & \text{样本}k\text{的胎儿染色体异常} \\ 0, & \text{样本}k\text{的胎儿染色体正常} \end{cases}$$

其中, N 表示训练样本的数量, 且 N 为 ≥ 1 的整数; $Z_{i,j}^{(k)}$ 为训练样本 k 经标准化处理后

的序列特征矩阵, 且 $k \in [1, N]$, i 为 ≥ 1 的整数, j 为 ≥ 1 的整数。

对于除训练的机器学习模型之外的与本发明第一方面相同的技术特征, 本发明第一方面在实施方案中的限定也适用于此。在本方面中, 所训练的机器学习模块包括输出层。例如, 神经网络模型的结构在所述第一全局平均池化层后包括输出层, 所述输出层与所述第一全局平均池化层相连, 是一个全连接层, 输出神经元个数为1, 用于输出染色体异常情况。

在第三方面, 本发明提供了一种检测胎儿染色体异常的系统, 包括:

数据获取模块, 用于获得待测孕妇样本的游离核酸片段的测序数据和临床表型特征数据, 其中所述测序数据包括若干读段, 所述待测孕妇样本的临床表型特征数据形成孕妇表型特征向量;

序列特征矩阵生成模块, 用于对参考基因组染色体的至少一部分进行窗口划分, 获得若干滑动窗口, 统计落入所述滑动窗口内的读段, 生成所述染色体序列的序列特征矩阵;

序列特征向量提取模块, 用于将所述序列特征矩阵输入训练的机器学习模型以提取所述染色体序列的序列特征向量;

分类检测模块, 用于将所述序列特征向量和孕妇表型特征向量组合形成组合特征向量, 将所述组合特征向量输入分类检测模型, 获得所述待测孕妇的胎儿染色体异常情况。

在一个实施方案中, 所述系统还包括比对模块, 用于将所述测序数据的读段与参考基因组比对, 获取唯一比对的读段。

在一个实施方案中, 在数据获取模块中, 所述游离核酸片段来自于孕妇外周血浆、孕妇肝脏和/或胎盘。

在一个实施方案中, 在数据获取模块中, 所述游离核酸片段为游离DNA。

在一个实施方案中, 在数据获取模块中, 所述测序数据来自超低深度测序; 优选地, 所述超低深度测序的测序深度为 $1\times$ 、 $0.1\times$ 或 $0.01\times$ 。

在一个实施方案中, 在数据获取模块中, 将所述读段与参考基因组进行比对, 获取唯一比对的读段(优选进行GC含量校正); 优选地, 后续步骤使用所述唯一比对的读段(优选经GC含量校正后的读段)进行。

在一个实施方案中, GC含量校正按如下步骤进行:

- 首先从人类参考基因组的某个染色体上随机采样 m 个长度为 ℓ 的片段;
- 计算GC含量为 i 的片段数量 N_i :

$$N_i = \sum_{k=1}^m \mathbb{I}_i(f(k))$$

其中 $\mathbb{I}_i(f(k)) = \begin{cases} 1, & f(k) = i \\ 0, & \text{其他} \end{cases}$, $f(k)$ 是片段 k 的GC含量, i 表示GC含量($i = 0\%, 1\%, \dots, 100\%$);

- 计算GC含量为 i 的测序读段数量 F_i :

$$F_i = \sum_{k=1}^m c_k \cdot \mathbb{I}_i(f(k))$$

其中 $c_k \cdot \mathbb{I}_i(f(k))$ 表示片段 k 的GC含量, F_i 表示GC含量为 i 且测序读段起始点位与该片段起始点位相同的测序读段的数量;

- 计算GC含量观测期望比 λ_i :

$$\lambda_i = \begin{cases} r \cdot \frac{F_i}{N_i}, & F_i > 0 \text{ 且 } N_i > 0 \\ 1, & \text{其他} \end{cases}$$

其中 r 是全局比例因子（Global Scaling），其定义为：

$$r = \frac{\sum_i N_i}{\sum_i F_i}$$

e. 校正测序读段数量：

$$R_i = \frac{1}{\lambda_i}$$

其中 R_i 表示经校正后GC含量为 i 的测序读段的期望数量。

在一个实施方案中，在数据获取模块中，孕妇表型特征数据选自如下的一个或多个的组合：年龄、孕周、身高、体重、BMI、产检生化检验检测结果、超声检查诊断结果和血浆中胎儿游离DNA浓度。

在一个实施方案中，在数据获取模块中，所述孕妇表型特征数据经异常值处理、缺失值处理和/或空值处理。

在一个实施方案中，在数据获取模块中，孕妇样本的表型数据出现如下记录将被判定为异常值，并将这些异常值设置为空值：

- a. $x_{age} < 10$ 或 $x_{age} > 80$;
- b. $x_{GW} < 5$ 或 $x_{GW} > 50$;
- c. $x_{height} < 40$ 或 $x_{height} > 300$;
- d. $x_{weight} < 10$ 或 $x_{weight} > 200$ 。

在一个实施方案中，缺失值和空值使用missForest算法进行填充。

在一个实施方案中，在序列特征矩阵生成模块中，所述染色体为21号染色体、18号染色体、13号染色体和/或性染色体。

在一个实施方案中，在序列特征矩阵生成模块中进行：（2.1）使用长度为 b 的窗口以步长 t 对参考基因组上的长度为 L 的染色体序列进行重叠滑动，获得若干滑动窗口，其中 b 为正整数，且 $b \in [10000, 10000000]$ ， t 为任意正整数， L 为正整数，且 $L \geq b$ ；（2.2）统计落入每一个所述滑动窗口内的读段，生成所述染色体序列的序列特征矩阵。

在一个实施方案中，在序列特征矩阵生成模块中，所述序列特征矩阵包括滑动窗口内读段数量、碱基质量和比对质量。

在一个实施方案中，所述碱基质量包括碱基质量的均值、标准差、偏度和/或峰值。

在一个实施方案中，所述比对质量包括比对质量的均值、标准差、偏度和/或峰值。

在一个实施方案中，在序列特征矩阵生成模块中，所述序列特征矩阵为：

$$\mathbf{X} = (x_{ij})_{h \times w}$$

其中 h 表示窗口的滑动次数， w 表示单个滑动窗口内的序列特征数量， x_{ij} 表示第 i 个滑动窗口中第 j 个序列特征值。

在一个实施方案中，在序列特征向量提取模块中，将所述序列特征矩阵进行标准化处理。

在一个实施方案中，在序列特征向量提取模块中，使用公式（1）进行所述序列特征矩阵的标准化：

$$z_{i,j}^{(k)} = \frac{X_{i,j}^{(k)} - \mu_{i,j}}{\sigma_{i,j}}$$

(I)

其中, $\mathbf{Z}_{i,j}^{(k)}$ 为样本 k 经标准化处理后的序列特征矩阵, $\mathbf{X}_{i,j}^{(k)}$ 表示样本 k 的第 i 个滑动窗口中的第 j 个序列特征值, $\mu_{i,j}$ 和 $\sigma_{i,j}$ 分别表示所有样本的第 i 个滑动窗口中的第 j 个序列特征值的均值和标准差。

在一个实施方案中, 在序列特征向量提取模块中, 所述训练的机器学习模型为神经网络模型或AutoEncoder模型; 优选地, 所述神经网络模型为深度神经网络模型; 更优选地, 所述神经网络模型为基于1维卷积的深度神经网络模型。

对于所述深度神经网络模型, 本发明第一方面在实施方案中的限定也适用于此。

在一个实施方案中, 在分类检测模块中, 将所述序列特征向量与所述孕妇表型拼接获得组合特征向量。

在一个实施方案中, 在分类检测模块中, 所述组合特征向量 $\mathbf{x}$ 经标准化处理:

$$x'_i = \frac{x_i - \mu_i}{\sigma_i}$$

其中 x'_i 是标准化后的组合特征向量 $\mathbf{x}$ 中的第 i 个序列特征值, x_i 是组合特征向量 $\mathbf{x}$ 中的第 i 个序列特征值, μ_i 是组合特征向量中第 i 个序列特征值的均值, σ_i 是组合特征向量中第 i 个序列特征值的标准差。

在一个实施方案中, 在分类检测模块中, 所述分类检测模型为集成学习模型。

在一个实施方案中, 所述集成学习模型为基于堆叠 (Stacking) 或多数投票 (Majority Voting) 的集成学习模型; 优选地, 所述集成学习模型为以下的一项或多项: 支持向量机模型、朴素贝叶斯分类器、随机森林分类器、XGBoos和逻辑回归。

在第四方面, 本发明提供了一种构建用于检测胎儿染色体异常的分类检测模型的系统, 所述系统包括:

数据获取模块, 用于获得孕妇游离核酸片段的测序数据和孕妇的临床表型特征数据, 其中所述测序数据包括若干读段, 并且所述孕妇的胎儿染色体情况已知, 所述孕妇的临床表型特征数据形成孕妇表型特征向量;

序列特征矩阵生成模块, 用于对参考基因组染色体的至少一部分进行窗口划分, 获得若干滑动窗口, 统计落入所述滑动窗口内的读段, 生成所述染色体序列的序列特征矩阵;

序列特征向量提取模块, 用于利用所述序列特征矩阵和胎儿染色体情况构建训练数据集, 训练机器学习模型以提取所述染色体序列的序列特征向量;

分类检测模型获得模块, 用于将所述序列特征向量和孕妇表型特征向量组合形成的组合特征向量, 利用所述组合特征向量和胎儿染色体情况对分类检测模型进行训练, 获得训练后的所述分类检测模型。

在一个实施方案中, 所述系统还包括比对模块, 用于将所述测序数据的读段与参考基因组比对, 获取唯一比对的读段。

对于除训练的机器学习模型之外的与本发明第三方面相同的技术特征, 本发明第三方面在实施方案中的限定也适用于此。在本方面中, 所训练的机器学习模块包括输出层。例如, 深度神经网络模型的结构在所述第一全局平均池化层后包括输出层, 所述输出层与所述第一全局平均池化层相连, 是一个全连接层, 输出神经元个数为1, 用于输出染色体异常情况。本发明的方法和模型基于测序数据的创新型算法, 不基于Z检验, 避免了依赖阈值在结果分值处于“灰区”而难以判断的临床问题。而且, 随着样本数量 (例如, 样本测序数据和对对应孕妇表型数据) 的增多, 本发明所提出的混合模型可以进行自动升级优化, 提高检测准确度。

附图说明

图1示出了根据本发明一个实施方案的基于深度神经网络混合模型检测胎儿染色体异常的方法流程图。

图2示出了根据本发明一个实施方案计算测序数据的特征矩阵。

图3示出了根据本发明一个实施方案的深度神经网络结构。

图4示出了根据本发明一个实施方案的Squeeze-Excite模块（SE模块）。

图5示出了根据本发明一个实施方案的Spatial Squeeze-Excite模块（sSE模块）。

图6示出了根据本发明一个实施方案的表型数据集缺失值填充。

图7示出了根据本发明一个实施方案的基于Stacking的集成学习模型结构。

图8示出了根据本发明一个实施方案的基于Stacking的集成学习模型机型5折交叉验证训练结果的ROC曲线。

图9示出了根据本发明一个实施方案的基于测试集的模型评估的ROC曲线。

图10示出了根据本发明一个实施方案的基于测试集的模型评估的Precision-Recall曲线。

图11示出了根据本发明一个实施方案的决策判定阈值为默认值（即0.5）时的混淆矩阵图。

图12示出了根据本发明一个实施方案的查准率和查全率作为阈值的函数。

图13示出了根据本发明一个实施方案设置最小查全率=0.95（即限制第二类错误）时的混淆矩阵图。

具体实施方式

本发明中，检测胎儿染色体异常的方法可以通过检测胎儿染色体异常的系统来实现；构建用于检测胎儿染色体异常的检测模型的方法可以通过用于检测胎儿染色体异常的检测模型的系统来实现。

在本发明中，数据获取模块用于获得孕妇游离核酸片段的测序数据和孕妇的临床表型特征数据，其中所述测序数据包括若干读段，所述孕妇的胎儿染色体情况已知（训练样本）或未知（待测样本），所述孕妇的临床表型特征数据形成孕妇表型特征向量。数据获取模块可以包括数据接收模块，用于接收上述数据。数据获取模块还可以包括测序仪，通过输入孕妇的游离核酸进行测序，获得测序数据。测序可以使高通量测序，可以使超低深度测序，超低深度测序的测序深度为 $1\times$ 、 $0.1\times$ 或 $0.01\times$ 。孕妇的游离核酸可以来源于孕妇的外周血浆、肝脏和/或胎盘。所述孕妇临床表型特征和所述孕妇的胎儿染色体情况（训练样本）可以通过数据库获取，所述孕妇的胎儿染色体情况可以是染色体非整倍性、微缺失和/或微重复。

在本发明中，比对模块用于将所述读段与参考基因组比对，获取唯一比对的读段。将序列比对到参考基因组上的应用软件可以从开源开发者获取，例如从某些网站在线获取，或者可以自行开发。

在本发明中，序列特征矩阵生成模块用于对参考基因组染色体的至少一部分进行窗口划分，获得若干滑动窗口，统计落入所述滑动窗口内的读段，生成该段染色体序列的序列特征矩阵。这可以通过以固定长度的窗口在染色体序列上滑动，固定长度的窗口可以是10k、100k、1M、10M等。步长可以是任意长度，为了方便计算，一般设置为滑动窗口的一半。染色体序列长度只需要大于滑动窗口长度，可以是10k、100k、1M、10M、100M……直至整条染色体长度大小。染色体可以是靶标染色体，例如检测21三体综合征则对应第21号染色体，检测18三体综合征则对应第18号染色体，检测13三体综合征则对应第13号染色体，检测性染色体异常则对应XY染色体，检测染色体微缺失/微重复则对应全部常染色体。对于每个窗口，统计其参数，包括读段数量、碱基质量（衡量测序准确度）和比对质量（读段比对到参考基因组的准确性度量，比对质量越高说明该读段比对到参考基因组上的位置越

唯一)等,这可以使用计算机软件完成。

在本发明中,序列特征提取模块用于提取染色体序列的序列特征。对于训练数据集而言,序列特征向量生成模块利用所述序列特征矩阵和孕妇的胎儿染色体异常情况构建训练数据集,训练机器学习模型以提取该段染色体序列的序列特征向量。对于测试数据而言,序列特征向量生成模块利用所述序列特征矩阵构建测试数据集,输入训练的机器学习模型例如深度神经网络模型以提取该段染色体序列的序列特征向量。

在本发明中,对于训练数据集而言,分类检测模块例如集成学习模型训练模块用于将序列特征向量和孕妇表型特征向量组合形成的组合特征向量以及胎儿染色体情况对分类检测模型进行训练,获得训练后的分类检测模型。

对于测试数据集而言,分类检测模块用于将序列特征向量和孕妇表型特征向量组合形成组合特征向量作为输入,用上述训练后的分类检测模型检测染色体异常情况。

本发明提出一个完全创新的方法用于检测染色体异常,例如非整倍体、微缺失或微重复。与传统方法不同,本发明不直接基于读段数量和Z值检测非整倍体,无需繁琐的数据预处理和特征提取选择工作,而是通过设计一个机器学习模型对测序数据中生成的序列特征矩阵进行自动提取序列特征向量,并将序列特征向量与孕妇的临床表型特征组合在一起,并利用分类检测模型进行检测,最终得到胎儿染色体是否存在遗传异常的预测结果。

在本发明中,利用机器学习模型从测序数据中自动提取序列特征向量,避免传统人工提取NIPT全基因组序列特征的弊端。本发明的方法不仅充分挖掘测序数据信息,而且充分利用孕妇临床表型信息(可以加入到模型的表型数据信息包括孕妇年龄、孕周、身高、体重、BMI(身体质量指数)、产检生化检验检测结果、超声检查诊断结果如NT值等),并将提取出的序列特征向量与孕妇表型特征向量组合,从而充分挖掘了NIPT测序数据和孕妇临床表型结果中蕴含的丰富特征数据信息,保证了检测结果的高可靠性和有效性。本发明的方法不仅可以用于检测常见的三体综合征,可以用于检测其他染色体缺陷,例如染色体拷贝数变异、染色体微缺失、微重复等。

在本发明中,提取序列特征向量还可以使用基于自动编码器网络(Autoencoder)或变分自动编码器网络(Variational Autoencoder)等深度神经网络模型进行。

在本发明中,训练基于Stacking或多数投票(Majority Voting)的集成学习模型检测染色体异常,充分利用不同分类器对非整倍体识别结果,大大地提高了识别非整倍体的准确度。

在本发明中,参考基因组是指通过例如人类基因组计划产生的染色体为正常二倍体的人类基因组图谱,例如hg38、hg19等。参考基因组可以是一个染色体或多个染色体,或者可以是一个染色体中的一部分。

下面通过具体实施例对本发明进一步说明,但本发明不受实施例限制。

实施例一、构建检测模型的实施例

在一个示例性方案中,构建检测模型的示例性模型方案的流程和步骤具体描述如下。

1. 获取NIPT测序数据和比对结果

利用高通量测序平台BGISEQ500对训练样本进行测序(采用SE35,测序深度 $0.1\times$),即孕妇游离核酸片段,已知所述孕妇的胎儿染色体情况,将测序数据与参考基因组进行比对并过滤重复比对序列,得到唯一比对测序读段结果。

2. 对上述1所获得的测序读段进行预处理,通过GC含量和测序深度的关系重新校正基因组各个覆盖区域的序列覆盖深度,具体过程如下(详细过程请参考文献[15])。

a. 首先从人类参考基因组的某个染色体(如21号染色体)上随机采样 m 个长度为 ℓ 的片段(Fragment)。

b. 计算GC含量为 i 的片段数量 N_i :

$$N_i = \sum_{k=1}^m \mathbb{I}_i(f(k))$$

其中 $\mathbb{I}_i(f(k)) = \begin{cases} 1, & f(k) = i \\ 0, & \text{其他} \end{cases}$, $f(k)$ 是片段 k 的 GC 含量, i 表示 GC 含量 ($i = 0\%, 1\%, \dots, 100\%$)。

c. 计算 GC 含量为 i 的测序读段数量:

$$F_i = \sum_{k=1}^m c_k \cdot \mathbb{I}_i(f(k))$$

其中 $c_k \cdot \mathbb{I}_i(f(k))$ 表示片段 k 的 GC 含量, F_i 表示 GC 含量为 i 且测序读段起始点位与该片段起始点位相同的测序读段的数量。

d. 计算 GC 含量观测期望比:

$$\lambda_i = \begin{cases} r \cdot \frac{F_i}{N_i}, & F_i > 0 \text{ 且 } N_i > 0 \\ 1, & \text{其他} \end{cases}$$

其中 r 是全局比例因子 (Global Scaling), 其定义为:

$$r = \frac{\sum_i N_i}{\sum_i F_i}$$

e. 校正测序读段数量:

$$R_i = \frac{1}{\lambda_i}$$

其中 R_i 表示经校正后 GC 含量为 i 的测序读段的期望数量。

3. 生成序列特征矩阵

上述2的结果计算特征矩阵, 计算过程如下 (如图2所示):

使用长度为 b 的滑动窗口对长度为 L 的目标染色体从起始点位到结束点位进行滑动, 滑动步长为 t 。对每次滑动窗口所覆盖的长度为 b 的区域计算如下特征:

- 该区域内经 GC 校正后的读段数量;
- 该区域内的碱基质量均值 (mean);
- 该区域内的碱基质量标准差 (std);
- 该区域内的碱基质量偏度 (skewness);
- 该区域内的碱基质量峰度 (kurtosis);
- 该区域内的比对质量均值 (mean);
- 该区域内的比对质量标准差 (std);
- 该区域内的比对质量偏度 (skewness);
- 该区域内的比对质量峰度 (kurtosis);

由此得到序列特征矩阵:

$$\mathbf{X} = (x_{ij})_{h \times w}$$

其中, h 表示窗口的滑动次数, 例如, $h = \left\lfloor \frac{L-b}{t} \right\rfloor + 1$;

w 表示单个滑动窗口内的序列特征数量, 例如 $w = 9$ (即每个长度为 b 的滑动窗口计算 9 个不同特征);

x_{ij} 表示第 i 个滑动窗口中的第 j 个序列特征值。

碱基质量 (Base Quality) 是指定量描述测序结果的准确程度; 碱基质量均值、碱基质量标准差、碱基质量偏度和碱基质量峰度分别是指测序读段中所有碱基质量的均值、标准差、偏度以及峰度。比对质量 (Map Quality) 是指给定测序读段比对到参考基因组序列上的可靠程度; 比对质量均值、比对质量标准差、比对质量偏度和比对质量峰度分别是指给定测序读段的比对质量的均值、标准差、偏度以及峰度。

4. 构建深度神经网络模型

4.1 构建数据集

使用3的结果构造训练集 $D_{seq} = \{Z^{(k)}, y^{(k)}\}_{k=1}^N$, 其中 N 表示样本的数量, 且 N 为 ≥ 1 的整数; $Z^{(k)}$ 为样本 k 经标准化处理后的序列特征矩阵 (以下称为标准化序列特征矩阵), 且 $k \in [1, N]$, 定义为:

$Z_{i,j}^{(k)} = \frac{x_{i,j}^{(k)} - \mu_{i,j}}{\sigma_{i,j}}$, 其中 $x_{i,j}^{(k)}$ 表示训练集中样本 k 的第 i 个滑动窗口中的第 j 个序列特征向量, $\mu_{i,j}$ 是训练集中的第 i 个滑动窗口中的第 j 个特征向量的均值, $\sigma_{i,j}$ 是训练集中的第 i 个滑动窗口中的第 j 个特征向量的标准差, i 为 ≥ 1 的整数, j 为 ≥ 1 的整数;

$$y^{(k)} = \begin{cases} 1, & \text{样本 } k \text{ 的胎儿染色体异常} \\ 0, & \text{样本 } k \text{ 的胎儿染色体正常} \end{cases}.$$

4.2 构建深度神经网络模型

构建深度神经网络模型, 其结构如图3所示。所述深度神经网络模型中涉及的卷积层均进行1维卷积操作, 无另外特殊说明所述1维卷积核 (即1维过滤器) 的参数均相同, 即1维卷积核数量为 f , 1维卷积核大小为 k , 1维卷积操作步长为 s , 1维卷积核使用L2正则化且正则化因子为 r_{L2} , 1维卷积核的初始化函数为 g , 设置1维卷积操作的输出特征图 (Feature Map) 大小与输入特征图大小不变, 池化核大小为 p , 池化步长为 p_s 。

所述深度神经网络模型中涉及的Dropout层所使用的丢弃率 (Dropout Ratio) 相同, 设置为 d 。

所述深度神经网络模型结构包括:

4.2.1 输入层

输入层, 用于接收经标准化处理后的序列特征矩阵 $Z^{(k)}$, 其大小为 $h \times w$ 。

4.2.2 前置模块

前置模块与输入层相连, 用于对输入序列特征矩阵进行第一次卷积和激活操作, 得到抽象表示特征图谱 (Feature Map), 该模块包括: 一个1维卷积层 (1D Convolution), 与之相连的一个批标准化层 (Batch Normalization), 与之相连的一个ReLU激活层。

4.2.3 核心模块

核心模块与前置模块相连, 用于对所述特征图谱进行进一步抽象和特征提取, 通过有效增加神经网络模型深度从而加强神经网络表达能力, 该核心模块由结构相同的残差模块重复操作3次组成, 其中每个残差模块的输出是下一个残差模块的输入。所述每个残差模块包括:

(A) 重复两次的核心模块的前置子模块, 每个子模块的结构相同, 包括一个1维卷积层、与之相连的一个Dropout层、与之相连的一个批标准化层和与之相连的一个ReLU激活层;

(B) 一个第一1维平均池化层 (1D Average Pooling), 与 (A) 中所述的第二个前置子模块相连;

(C) 一个Squeeze-Excite模块 (即SE模块) 或者一个Spatial Squeeze-Excite 模块 (即sSE模块), 与 (B) 中所述的第一1维平均池化层相连;

首先设置SE模块的下降速率 (Reduction Ratio) 为 r_{SE} , 如图4所示, 所述SE模块结构包

括（详细描述见参考文献[16]）：

- （a）一个第二全局平均池化层，与（B）中所述的第一1维平均池化层相连；
- （b）与（a）所述的第二全局平均池化层相连的一个Reshape层，输出特征图大小为 $1 \times f$ ，其中 f 为1维卷积核数量；
- （c）与（b）中所述Reshape层相连的一个第一全连接层，其输出神经元个数为 $\frac{f}{r_{SE}}$ ，其中 f 为1维卷积核数量， r_{SE} 为SE模块下降速度；
- （d）与（c）所述第一全连接层相连的一个第二全连接层，其输出神经元个数为 f ，其中 f 为1维卷积核数量；
- （e）一个Multiply层，与之连接的是所述（B）的第一1维平均池化层和所述（d）中的第二全连接层；

如图5所示，所述sSE模块结构包括（详细描述见参考文献[17]）：

- a. 一个 1×1 的1维卷积层，与（B）中所述的第一1维平均池化层相连，该 1×1 的1维卷积层使用sigmoid函数作为激活函数；
- b. 一个Multiply层，与之连接的是（B）中所述的第一1维平均池化层和a中所述的 1×1 的1维卷积层；
- （D）一个第一加法层（Add层），与之相连是（C）中所述的SE模块和所述的sSE模块；
- （E）第二个1维平均池化层，与4.2.2所述前置模块中的ReLU激活层相连；
- （F）一个第二加法层（Add层），与之相连是（D）所述的第一加法层和（E）中所述的第二1维平均池化层；

上述（A）-（D）为所述残差模块左分支，（E）为所述残差模块右分支。

4.2.4 后置模块

后置模块与前置模块的结构相同，唯一区别在于后置模块中的1维卷积核数量设置为 n_{out} ，用于对来自核心模块的特征图谱进行输出前的特征抽象表示。

4.2.5 第一全局平均池化层

一个第一全局平均池化层，与后置模块相连，用于将所述特征抽象表示的特征图谱进行向量化。

4.2.6 输出层

输出层，与所述第一全局平均池化层相连，是一个全连接层，输出神经元个数为1，激活函数为sigmoid函数，用于输出染色体异常情况。

5. 计算序列特征向量

利用训练集对4中所述深度神经网络模型进行训练，利用训练完成后的深度神经网络模型计算样本的序列特征向量，其过程如下：

- （1）按照上述4.1所述计算每个样本的标准化序列特征向量；
- （2）按照（1）中得到的标准化序列特征矩阵输入深度神经网络模型进行计算；
- （3）保存 4.2.5中所述深度神经网络模型的第一全局平均池化层的输出作为输入样本对应生成的序列特征向量 seq ，定义为：

$$seq = [x_1, x_2, \dots, x_{n_{out}}]^T$$

其中 n_{out} 是4.2.4中所述的后置模块中定义的1维卷积核的数量。

6. 获取孕妇样本对应的表型结果

获取孕妇样本对应的表型结果，并构建初始表型特征向量 phe_{init} ，其中包括5个特征，其定义为：

$$phe_{init} = [x_{age}, x_{GW}, x_{height}, x_{weight}, x_{FF}]^T$$

其中 x_{age} 表示孕妇采样时年龄 (岁), x_{GW} 表示孕妇采样时孕周, x_{height} 表示孕妇身高 (厘米), x_{weight} 表示孕妇体重 (千克), x_{FF} 表示孕妇血浆中胎儿游离DNA浓度。

7. 表型数据预处理

对孕妇表型数据集进行预处理, 包括异常值处理和缺失值或空值处理。

(1) 异常值处理

孕妇样本的表型数据出现如下记录将被判定为异常值, 并将这些异常值设置为空值。

- a. $x_{age} < 10$ 或 $x_{age} > 80$;
- b. $x_{GW} < 5$ 或 $x_{GW} > 50$;
- c. $x_{height} < 40$ 或 $x_{height} > 300$;
- d. $x_{weight} < 10$ 或 $x_{weight} > 200$ 。

(2) 缺失值或空值处理

构造表型数据矩阵 P , 其定义如下:

$$P = [phe_{init}^1, phe_{init}^2, \dots, phe_{init}^i, \dots, phe_{init}^N]^T$$

其中, phe_{init}^i 表示训练集中第 i 个样本的表型特征向量 (如6中所述定义), N 表示训练集中样本数量。这里的训练集中的样本与4.1中所述训练集中样本一致, 于是表型数据矩阵 P 是一个大小为 $N \times M$ 的矩阵, 其中 M 是表型特征的数量, 这里 $M = 5$ 。

缺失值填充使用missForest算法, missForest算法是一种基于随机森林的非参缺失值填充算法 (详细描述见参考文献[18]), 其算法如下:

算法1: missForest算法

输入: 矩阵 P ;

迭代终止条件 γ ;

输出: 缺失值填充后的矩阵 P^{imp} ;

1. $k \leftarrow$ 对矩阵 P 中的列的缺失值数量从小到大重新建立列索引;
2. while not γ do:
3. for s in k do:
4. 抽取矩阵 P 的第 s 列 P_s ;
5. 根据 P_s 将矩阵 P 划分为 $\mathbf{y}_{obs}^{(s)}, \mathbf{y}_{mis}^{(s)}, \mathbf{X}_{obs}^{(s)}, \mathbf{X}_{mis}^{(s)}$ 四个部分, 其中 $\mathbf{y}_{obs}^{(s)}$ 表

示 P_s 列中的观测值部分, $\mathbf{y}_{mis}^{(s)}$ 表示 P_s 列中的缺失值部分, $\mathbf{X}_{obs}^{(s)}$ 表示矩阵 P

中除去 P_s 列且与 $\mathbf{y}_{obs}^{(s)}$ 所在行对应的部分, $\mathbf{X}_{mis}^{(s)}$ 表示矩阵 P 中除去 P_s 列且

与 $\mathbf{y}_{mis}^{(s)}$ 所在行对应的部分;

6. 使用观测值训练随机森林模型: $\mathbf{y}_{obs}^{(s)} \sim RF(\mathbf{X}_{obs}^{(s)})$;

7. 使用已训练的随机森林模型预测缺失值: $\mathbf{y}_{mis}^{(s)} = RF(\mathbf{X}_{mis}^{(s)})$;

8. $P^{imp} \leftarrow$ 根据预测值 $\mathbf{y}_{mis}^{(s)}$ 更新矩阵 P ;

9. end for

10.更新迭代终止条件 γ ;

11. end while

12. return P^{imp} ;

(3) 计算BMI

利用缺失值填充后的表型结果计算BMI, 其定义为:

$$BMI = \frac{x_{weigh}}{(x_{height}/100)^2}$$

(4) 将(3)的结果添加到经过缺失值填充后的表型特征向量, 得到最终的表型特征向量:

$$phe = [x_{age}, x_{GW}, x_{height}, x_{weight}, x_{FF}, x_{BMI}]^T$$

8. 生成组合特征向量

将5所述序列特征向量与7所述最终特征向量进行组合, 得到组合特征向量:

$$\mathbf{x} = [seq^T, phe^T]^T = [x_1, x_2, \dots, x_{n_{out}}, x_{age}, x_{GW}, x_{height}, x_{weight}, x_{FF}, x_{BMI}]^T$$

9. 组合特征向量标准化

对8所述组合特征向量进行标准化处理:

$$x'_i = \frac{x_i - \mu_i}{\sigma_i}$$

其中, x'_i 是标准化后的组合特征向量 $\mathbf{x}$ 中的第 i 个序列特征值, x_i 是组合特征向量 $\mathbf{x}$ 中的第 i 个序列特征值, μ_i 是组合特征向量中第 i 个序列特征值的均值, σ_i 是组合特征向量中第 i 个序列特征值的标准差。

10. 构建基于Stacking的集成学习模型

用9中所述结果构造训练集 $D_{seq} = \{\mathbf{z}_{i,j}^{(k)}, y^{(k)}\}_{k=1}^N$,

$$y^{(k)} = \begin{cases} 1, & \text{样本} k \text{ 的胎儿染色体异常} \\ 0, & \text{样本} k \text{ 的胎儿染色体正常} \end{cases}$$

其中, N 表示训练样本的数量, 且 N 为 ≥ 1 的整数; $\mathbf{z}_{i,j}^{(k)}$ 为训练样本 k 经标准化处理后的序列特征矩阵, 且 $k \in [1, N]$, i 为 ≥ 1 的整数, j 为 ≥ 1 的整数; $y=0$ 表示胎儿染色体正常, $y=1$ 表示胎儿染色体异常。

使用基于Stacking的集成学习算法预测非整倍体, 算法描述如下(详细描述见参考文献[19]):

算法2: 基于Stacking的集成学习算法

输入: 训练集 D ;

M 个一级基分类器(base-classifier) $h_1, \dots, h_M$;

1个二级元分类器(meta-classifier) h' ;

输出: 集成学习器 H

1. //第一步: 训练 M 个第一级基分类器

2. for $m \leftarrow 1$ to M do:

3. 基于训练集 D 训练基分类器 h_m

4. end for

5. //第二步: 从训练集 D 构建新数据集 D'

```

6.  $D' = \emptyset$ 
7. for  $i \leftarrow 1$  to  $N$  do:
8. 计算向量  $z_i = [h_1(x'_i), h_2(x'_i), \dots, h_M(x'_i)]^T$ 
9. 将向量  $z_i$  添加到新数据集  $D' = D' \cup \{(z_i, y_i)\}$ 
10. end for
11. //第三步：训练第二级元分类器 $h'$ 
12. 基于新数据集 $D'$ 训练第二级元分类器 $h'$ 
13. return  $H(x) = h'(x) = h'(h_1(x), h_2(x), \dots, h_M(x))$ 

```

实施例二、检测染色体异常的实施例

在一个示例性方案中，本发明提出一种检测胎儿染色体异常的方法，所述方法利用无创产前检测（NIPT）的核酸测序结果和孕妇表型数据共同预测胎儿染色体是否存在遗传异常。在一个具体实施方案中，检测胎儿染色体异常的方法的流程和步骤如图1所示，具体过程描述如下。

1. 获取NIPT测序数据和比对结果

利用高通量测序平台BGISEQ500对待测样本进行测序（采用SE35，测序深度 $0.1 \times$ ），将测序数据与参考基因组进行比对并过滤重复比对序列，得到唯一比对测序读段结果。

2. 对上述1所获得的测序读段进行预处理，通过GC含量和测序深度的关系重新校正基因组各个覆盖区域的序列覆盖深度，具体过程参考实施例一。

3. 生成序列特征矩阵

上述2的结果计算特征矩阵，计算过程如下（如图2所示）：

使用长度为 b 的滑动窗口对长度为 L 的目标染色体从起始点位到结束点位进行滑动，滑动步长为 t 。对每次滑动窗口所覆盖的长度为 b 的区域计算如下特征：

- 该区域内经GC校正后的读段数量；
- 该区域内的碱基质量均值（mean）；
- 该区域内的碱基质量标准差（std）；
- 该区域内的碱基质量偏度（skewness）；
- 该区域内的碱基质量峰度（kurtosis）；
- 该区域内的比对质量均值（mean）；
- 该区域内的比对质量标准差（std）；
- 该区域内的比对质量偏度（skewness）；
- 该区域内的比对质量峰度（kurtosis）；

由此得到序列特征矩阵：

$$X = (x_{ij})_{h \times w}$$

其中， h 表示窗口的滑动次数，例如， $h = \left\lfloor \frac{L-b}{t} \right\rfloor + 1$ ；

w 表示单个滑动窗口内的序列特征数量，例如 $w = 9$ （即每个长度为 b 的滑动窗口计算9个不同特征）；

x_{ij} 表示第 i 个滑动窗口中的第 j 个序列特征值。

碱基质量（Base Quality）是指定量描述测序结果的准确程度；碱基质量均值、碱基质量标准差、碱基质量偏度和碱基质量峰度分别是指测序读段中所有碱基质量的均值、标准差、偏度以及峰度。比对质量（Map Quality）是指给定测序读段比对到参考基因组序列上的可靠程度；比对质量均值、比对质量标准差、比对质量偏度和比对质量峰度分别是指给定测序读段的比对质量的均值、标准差、偏度以及峰度。

4. 利用实施例一中训练完成后的深度神经网络模型计算样本的序列特征向量，其过程如下：

- (1) 按照实施例一中4.1所述计算样本的标准化序列特征矩阵；
- (2) 按照(1)中得到的标准化序列特征矩阵输入深度神经网络模型进行计算；
- (3) 保存实施例一中 4.2.5中所述深度神经网络模型的第一全局平均池化层的输出作为输入样本对应生成的序列特征向量 seq ，定义为：

$$seq = [x_1, x_2, \dots, x_{n_{out}}]^T$$

其中 n_{out} 是4.2.4中所述的后置模块中定义的1维卷积核的数量。

5. 获取待测孕妇样本对应的表型结果

获取待测孕妇样本对应的表型结果，并构建初始表型特征向量 phe_{init} ，其中包括5个特征，其定义为：

$$phe_{init} = [x_{age}, x_{GW}, x_{height}, x_{weight}, x_{FF}]^T$$

其中 x_{age} 表示孕妇采样时年龄（岁）， x_{GW} 表示孕妇采样时孕周， x_{height} 表示孕妇身高（厘米）， x_{weight} 表示孕妇体重（千克）， x_{FF} 表示孕妇血浆中胎儿游离DNA浓度。

6. 表型数据进行异常值处理

待测孕妇样本的表型数据出现如下记录将被判定为异常值，并将这些异常值设置为空值。

- a. $x_{age} < 10$ 或 $x_{age} > 80$ ；
- b. $x_{GW} < 5$ 或 $x_{GW} > 50$ ；
- c. $x_{height} < 40$ 或 $x_{height} > 300$ ；
- d. $x_{weight} < 10$ 或 $x_{weight} > 200$ 。

7. 将4所述序列特征向量与6所述最终特征向量进行组合，得到组合特征向量：

$$x = [seq^T, phe^T]^T = [x_1, x_2, \dots, x_{n_{out}}, x_{age}, x_{GW}, x_{height}, x_{weight}, x_{FF}, x_{BMI}]^T$$

8. 组合特征向量标准化

对7所述组合特征向量进行标准化处理：

$$x'_i = \frac{x_i - \mu_i}{\sigma_i}$$

其中， x'_i 是标准化后的组合特征向量 x 中的第 i 个序列特征值， x_i 是组合特征向量 x 中的第 i 个序列特征值， μ_i 是组合特征向量中第 i 个序列特征值的均值， σ_i 是组合特征向量中第 i 个序列特征值的标准差。

9. 将所述组合特征向量输入实施例一中构建基于Stacking的集成学习模型，获得所述待测孕妇的胎儿染色体情况。

实施例三、验证实施例

1. 样本数量

本实施例使用1205个“21三体（T21）”作为阳性样本和1600个正常染色体（二倍体）样本作为阴性样本。

表1 描述了训练样本和测试样本的样本数量

	样本总数(N)	训练集样本数(90%×N)	测试集样本数(10%×N)
正样本(T21)	1205	1084	121
负样本(正常)	1600	1440	160

2. 按照前述实施例一中2描述的步骤对所有正负样本的测序数据进行预处理，其中随机

采样片段的数量 $m = 50000000$, 片段 $\ell = 180$ 。

3. 按照前述实施例一中3描述的步骤对所有正负样本生成测序序列特征矩阵。参数设置如下:

21号染色体的长度: $L = 46709983$;

滑动窗口大小: $b = 1000000$;

滑动步长: $t = 500000$ 。

于是得到每个样本对应的测序序列特征矩阵, 其大小为 9×93 , 即 $w = 9, h = 93$ 。由于21号染色体起始部分在参考基因组中没有可比对序列, 在本实施例中过滤测序序列特征矩阵前8列, 即实际使用的测序序列特征矩阵大小为 9×85 。

4. 按照3的结果, 使用训练集中对应的测序数据特征矩阵训练深度神经网络模型。

(1) 按照前述实施例一中4.1所述对训练集中测序数据特征矩阵进行标准化, 保存标准化模型。

(2) 按照(1)中所述得到作为训练深度神经网络模型的输入张量, 张量大小为 $2524 \times 85 \times 9$ 。

(3) 按照前述实施例一中4.2所述训练深度神经网络模型, 设置深度神经网络模型参数如下:

1维卷积核数量: $f = 32$,

1维卷积核大小: $k = 8$,

1维卷积操作步长: $s = 1$,

1维卷积核L2正则化因子: $r_{l2} = 0.0004$,

1维卷积核的初始化函数 g 使用文献[20]中描述的“He normalization”初始化函数,

1维卷积操作的输出特征图大小与输入特征图大小不变,

池化核大小: $p = 2$,

池化步长: $p_s = 2$,

Dropout层所使用的丢弃率: $r_d = 0.5$,

SE模块下降速率: $r_{SE} = 16$,

后置模块中1维卷积核数量 $n_{out} = 8$ 。

本实施例基于Tensorflow(version=1.12.2)和keras(version=2.2.4)的GPU版本实现。表2列出了依据所述参数的深度神经网络模型中各层操作、输出特征图大小以及网络连接。

序号	模块				网络层	输出大小	连接序号
1	输入层				Input	(None, 85, 9)	
2	前置模块				Conv1D	(None, 85, 32)	1
3					BatchNormalization	(None, 85, 32)	2
4					ReLU	(None, 85, 32)	3
5	核心模块	第一个残差模块	第一个子模块		Conv1D	(None, 85, 32)	4
6					Dropout	(None, 85, 32)	5
7					BatchNormalization	(None, 85, 32)	6
8					ReLU	(None, 85, 32)	7
9			第二个子模块		Conv1D	(None, 85, 32)	8
10					Dropout	(None, 85, 32)	9
11					BatchNormalization	(None, 85, 32)	10
12					ReLU	(None, 85, 32)	11
13			平均		AveragePooling	(None, 42, 32)	12

			池化层			
14			SE 模块	GlobalAveragePooling	(None, 32)	13
15				Reshape	(None, 1, 32)	14
16				FC (Activation=ReLU)	(None, 1, 2)	15
17				FC (Activation=sigmoid)	(None, 1, 32)	16
18				Multiply	(None, 42, 32)	12和17
19			sSE 模块	Conv1D (Activation=sigmoid)	(None, 42, 1)	12
20				Multiply	(None, 42, 32)	12和19
21			相加	Add	(None, 42, 32)	18和20
22		残差模块右分支	平均池化层	AveragePooling	(None, 42, 32)	3
23		分支相加		Add	(None, 42, 32)	21和22
24	第二个残差模块	残差模块左分支	第一个子模块	Conv1D	(None, 42, 32)	23
25				Dropout	(None, 42, 32)	24
26				BatchNormalization	(None, 42, 32)	25
27				ReLU	(None, 42, 32)	26
28			第二个子模块	Conv1D	(None, 42, 32)	27
29				Dropout	(None, 42, 32)	28
30				BatchNormalization	(None, 42, 32)	29
31				ReLU	(None, 42, 32)	30
32			平均池化层	AveragePooling	(None, 21, 32)	31
33			SE 模块	GlobalAveragePooling	(None, 32)	32
34				Reshape	(None, 1, 32)	33
35				FC (Activation=ReLU)	(None, 1, 2)	34
36				FC (Activation=sigmoid)	(None, 1, 32)	35
37				Multiply	(None, 21, 32)	32和36
38			sSE 模块	Conv1D (Activation=sigmoid)	(None, 21, 1)	32
39				Multiply	(None, 21, 32)	32和38
40			相加	Add	(None, 21, 32)	37和39
41		残差模块右分支	平均池化层	AveragePooling	(None, 21, 32)	23

			支				
42			分支相加	Add	(None, 21, 32)	40和41	
43			第一个子模块	Conv1D	(None, 21, 32)	42	
44				Dropout	(None, 21, 32)	43	
45				BatchNormalization	(None, 21, 32)	44	
46				ReLU	(None, 21, 32)	45	
47			第二个子模块	Conv1D	(None, 21, 32)	46	
48				Dropout	(None, 21, 32)	47	
49				BatchNormalization	(None, 21, 32)	48	
50				ReLU	(None, 21, 32)	49	
51			平均池化层	AveragePooling	(None, 10, 32)	50	
52			SE模块	GlobalAveragePooling	(None, 32)	51	
53				Reshape	(None, 1, 32)	52	
54				FC (Activation=ReLU)	(None, 1, 2)	53	
55				FC (Activation=sigmoid)	(None, 1, 32)	54	
56				Multiply	(None, 10, 32)	51和55	
57			sSE模块	Conv1D	(None, 10, 1)	51	
58				Multiply	(None, 10, 32)	51和57	
59			相加	Add	(None, 10, 32)	56和58	
60			残差模块右分支	平均池化层	(None, 10, 32)	42	
61			分支相加	Add	(None, 10, 32)	59和60	
62			后置模块	Conv1D	(None, 10, 8)	61	
63				BatchNormalization	(None, 10, 8)	62	
64				ReLU	(None, 10, 8)	63	
65			全局平均池化层	GlobalAveragePooling	(None, 8)	64	
66			输出层	FC(Activation=sigmoid)	(None, 1)	65	

(4) 训练集中80%样本用于训练深度神经网络, 20%样本用于验证以便计算准确率。

(5) 设置训练深度神经网络迭代次数epochs=100, 样本批量大小mini_batch=64。梯度下降的优化算法使用Adam算法(参数 $\beta_1=0.9$, $\beta_2=0.999$), 初始学习率(Learning Rate)设置为0.01。训练过程中如果准确率在连续2次迭代后没有提高, 学习率将减小2倍(即乘以0.5); 如果准确率在连续10次迭代后没有提高, 训练停止。

(6) 深度神经网络模型训练过程中引入类权重因子(利用机器学习库scikit-learn(version=0.22.2)中的compute_class_weight()函数计算类权重, 并将该类权重赋予对应类别的样本)。

(7) 保存训练的深度神经网络模型。

5. 按照前述实施例一中5所述计算序列特征向量：

(1) 按照前述实施例一中3所述对整个数据集（包括训练集和测试集）中所有样本计算序列特征矩阵；

(2) 按照4.1所述使用已得到的序列标准化模型对上述（1）中得到的序列特征矩阵进行标准化处理；

(3) 利用4所得神经网络模型，模型输入为上述（2）的结果，模型输出层修改为全局平均池化层（即表2中的第65层）；

(4) 按照（3）的过程得到整个数据集（包括训练集和测试集）中所有样本的序列特征向量。

6. 按照前述实施例一中7所述获整个数据集（包括训练集和测试集）中所有样本的表型特征，并对表型特征异常值进行处理。

7. 按照前述实施例一中7所述对训练集中的表型特征进行缺失值填充处理，保存缺失值填充模型。

8. 按照前述实施例一中7所述对缺失值填充处理后的训练集中的表型特征计算BMI，如图6所示。

9. 按照前述实施例一中8所述将训练集中序列特征向量和对应样本的表型特征向量进行组合操作得到组合特征向量。

10. 按照前述实施例一中9所述将训练集中每个样本的组合特征向量进行标准化处理，得到标准化特征向量，并保存组合特征向量标准化模型。

11. 按照上述7-10的过程，利用保存的缺失值填充模型对测试集中每个样本的表型特征进行缺失值填充，然后测试集中的序列特征向量和对应样本的表型特征向量进行组合得到测试组合特征向量，再利用已保存的组合特征向量标准化模型对测试集中的组合特征向量进行标准化处理。

12. 利用上述10中得到的训练集标准化特征向量训练基于Stacking的集成学习模型，如图7所示。本实施例基于scikit-learn(version=0.22.2)机器学习库实现，其中每个基分类器模型和最终元分类器模型均引入类权重因子，参数无特殊说明均使用默认值。

(1) 按照前述实施例一中10所述，本实施例使用的基分类器包括：

- SVC, 参数为C=0.5, kernel="rbf"
- ν-SVC, 参数为ν=0.25, kernel= kernel="rbf"
- GaussianNB(高斯朴素贝叶斯模型)
- RandomForestClassifier (随机森林分类器) , 参数为 n_estimators=100, criterion="gini", max_depth=5, min_samples_leaf=1, min_samples_split=2
- XGBClassifier (XGBoosting 分类器) , 参数为 n_estimators=100, min_child_weight=1, gamma=0.1, colsample_bytree=0.8, subsample=0.7, reg_alpha=0.01, max_depth=5, learning_rate=0.05,
- LogisticRegression (逻辑回归) , 参数为C=0.5

(2) 按照前述实施例一中10所述的最终元分类器为ExtraTreesClassifier (极限随机树分类器) , 该分类器涉及的参数分别设置为 n_estimators=110, max_depth=6, min_samples_split=3, min_samples_leaf=1。

(3) 对基于Stacking的集成学习模型进行5折交叉验证训练，其结果如图8所示。可以看出5折交叉验证训练模型得到的平均AUC=0.96。

13. 使用测试集对12所述训练得到的基于Stacking的集成学习模型进行验证。

(1) 测试结果ROC曲线如图9所示，其中AUC=0.96；

(2) 测试结果Precision-Recall曲线如图10所示,其中AP=0.95。

(3) 当决策判定阈值为默认值(即0.5)时的混淆矩阵如图11所示,此时查全率和查准率分别为0.83和0.89。

(4) 查准率(Precision)和查全率(Recall)作为阈值的函数如图12所示。

(5) 设置最小查全率=0.95,即限制第二类错误,得到的结果如图12所示,此时查全率和查准率分别为0.96和0.70。

本发明提出了利用机器学习模型(例如深度神经网络)提取NIPT测序数据的序列特征向量,然后通过向量组合的方式将序列特征向量(特征项包括但不限于读段数量、碱基质量和比对质量)和孕妇表型特征向量(孕妇表型特征包括但不限于孕妇年龄、孕周、身高、体重、BMI、产检生化检验检测结果、超声检查诊断结果如NT值等)整合组合在一起,然后使用分类模型(例如基于Stacking的集成学习模型)得到最终预测非整倍体。在本发明中,提取序列特征向量并不限于此此处所使用的方法,还可以使用包括但不限于自动编码器网络(Autoencoder)或变分自动编码器网络(Variational Autoencoder)。本发明所提出的模型结构是一种混合(Hybrid)模型,即模型包括2个阶段,第一阶段使用机器学习模型(例如深度神经网络)计算序列特征向量,第二阶段使用分类模型(例如基于Stacking的集成学习模型)利用整合组合后的序列特征向量和表型特征向量预测非整倍体,也可以使用其他集成学习模型,例如基于多数投票(Majority Voting)。

本发明实施例中使用的经验证的高级深度神经网络模型与其他卷积神经网络相比,其网络设计和架构特点包括:本发明实施例中使用的深度神经网络模型是基于1维卷积模型设计深度神经网络模型;本发明实施例中使用的深度神经网络模型是基于残差网络(Residual Net)网络模型;本发明实施例中使用的深度神经网络模型引入了Squeeze-Excite网络的SE模块。基于这些设计,本发明实施例中使用的神经网络模型层数更多(参见实施例三)并且有效减少模型训练过程中梯度消失的风险和过拟合风险,提高模型稳定性,从而有效提高模型预测结果的准确性。

本发明可以实现为一种计算机可读存储介质,其上存储有计算机程序,所述计算机程序在由处理器执行时导致本发明的方法的步骤被执行。在一个实施例中,所述计算机程序被分布在网络耦合的多个计算机设备或处理器上,以使得所述计算机程序由一个或多个计算机设备或处理器以分布式方式存储、访问和执行。单个方法步骤/操作,或者两个或更多个方法步骤/操作,可以由单个计算机设备或处理器或由两个或更多个计算机设备或处理器执行。一个或多个方法步骤/操作可以由一个或多个计算机设备或处理器执行,并且一个或多个其他方法步骤/操作可以由一个或多个其他计算机设备或处理器执行。一个或多个计算机设备或处理器可以执行单个方法步骤/操作,或执行两个或更多个方法步骤/操作。

本领域的技术人员应当理解,本发明用于检测胎儿染色体异常的方法中各个步骤划分以及次序仅仅是示意性的而非限定性的,本领域的技术人员可以在不背离在附加的权利要求和其等价技术方案阐述的本发明精神和范围的情况下,进行删减、增加、替换、修改和变化。本发明的实施例的各技术特征可以进行任意的组合,为使描述简洁,未对上述实施例中的各个技术特征所有可能的组合都进行描述,然而,只要这些技术特征的组合不存在矛盾,都应当认为是本说明书记载的范围。

尽管已经参考示例性实施例描述了本发明,但是应理解,本发明并不限于上述实施例的构造和方法。相反,本发明意在覆盖各种修改例和等同配置。另外,尽管在各种示例性结合体和构造中示出了本发明所公开的各种元件和方法步骤,但是包括更多、更少的元件或方法的其它组合也落在本发明的范围之内。

参考文献:

[1] Evans, Mark I., Stephanie Andriole, and Shara M. Evans. "Genetics: update on prenatal

- screening and diagnosis." *Obstetrics and Gynecology Clinics* 42.2 (2015): 193-208.
- [2] Norwitz, Errol R., and Brynn Levy. "Noninvasive prenatal testing: the future is now." *Reviews in obstetrics and gynecology* 6.2 (2013): 48.
- [3] Norton, Mary E., et al. "Cell-free DNA analysis for noninvasive examination of trisomy." *New England Journal of Medicine* 372.17 (2015): 1589-1597.
- [4] Langlois, Sylvie, et al. "Current status in non-invasive prenatal detection of Down syndrome, trisomy 18, and trisomy 13 using cell-free DNA in maternal plasma." *Journal of Obstetrics and Gynaecology Canada* 35.2 (2013): 177-181.
- [5] Allyse, Megan, et al. "Non-invasive prenatal testing: a review of international implementation and challenges." *International journal of women's health* 7 (2015): 113.
- [6] Chiu, Rossa WK, et al. "Noninvasive prenatal diagnosis of fetal chromosomal aneuploidy by massively parallel genomic sequencing of DNA in maternal plasma." *Proceedings of the National Academy of Sciences* 105.51 (2008): 20458-20463.
- [7] Fan, H. Christina, et al. "Noninvasive diagnosis of fetal aneuploidy by shotgun sequencing DNA from maternal blood." *Proceedings of the National Academy of Sciences* 105.42 (2008): 16266-16271.
- [8] Lau, Tze Kin, et al. "Noninvasive prenatal diagnosis of common fetal chromosomal aneuploidies by maternal plasma DNA sequencing." *The Journal of Maternal-Fetal & Neonatal Medicine* 25.8 (2012): 1370-1374.
- [9] Jiang, Fuman, et al. "Noninvasive Fetal Trisomy (NIFTY) test: an advanced noninvasive prenatal diagnosis methodology for fetal autosomal and sex chromosomal aneuploidies." *BMC medical genomics* 5.1 (2012): 57.
- [10] Yang, Jianfeng, Xiaofan Ding, and Weidong Zhu. "Improving the calling of non-invasive prenatal testing on 13-/18-/21-trisomy by support vector machine discrimination." *BioRxiv* (2017): 216689.
- [11] Xu, Hanli, et al. "Informative priors on fetal fraction increase power of the noninvasive prenatal screen." *Genetics in Medicine* 20.8 (2018): 817-824.
- [12] Ehrich, Mathias, et al. "Deep learning-based methods, devices, and systems for prenatal testing", Publication number: WO2019191319A1, Filing Date: 27 March 2019.
- [13] Egilsson, Agust, et al. "Methods and systems for calling ploidy status using a neural network". Publication number: WO2020018522A1, Filing date: 16 July 2019.
- [14] Petersen, Andrea K., et al. "Positive predictive value estimates for cell-free noninvasive prenatal screening from data of a large referral genetic diagnostic laboratory." *American journal of obstetrics and gynecology* 217.6 (2017): 691-e1.
- [15] Benjamini, Yuval, and Terence P. Speed. "Summarizing and correcting the GC content bias in high-throughput sequencing." *Nucleic acids research* 40.10 (2012): e72-e72.
- [16] Hu, Jie, Li Shen, and Gang Sun. "Squeeze-and-excitation networks." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.
- [17] Roy, Abhijit Guha, Nassir Navab, and Christian Wachinger. "Concurrent spatial and channel 'squeeze & excitation' in fully convolutional networks." *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, Cham, 2018.
- [18] Stekhoven, Daniel J., and Peter Bühlmann. "MissForest—non-parametric missing value imputation for mixed-type data." *Bioinformatics* 28.1 (2012): 112-118.
- [19] Tang, J., S. Alelyani, and H. Liu. "Data Classification: Algorithms and Applications." *Data*

Mining and Knowledge Discovery Series, CRC Press (2015): pp. 498-500.

[20] He, Kaiming, et al. "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification." Proceedings of the IEEE international conference on computer vision. 2015.

权 利 要 求 书

1、一种检测胎儿染色体异常的方法，所述方法包括：

(1) 获得待测孕妇的游离核酸片段的测序数据和临床表型特征数据，其中所述测序数据包括若干读段，所述待测孕妇的临床表型特征数据形成孕妇表型特征向量；

(2) 对参考基因组染色体的至少一部分进行窗口划分，获得若干滑动窗口，统计落入所述滑动窗口内的读段，生成所述染色体序列的序列特征矩阵；

(3) 将所述序列特征矩阵输入训练的机器学习模型以提取所述染色体序列的序列特征向量；

(4) 将所述序列特征向量和孕妇表型特征向量组合形成组合特征向量，将所述组合特征向量输入分类检测模型，获得所述待测孕妇的胎儿染色体异常情况。

2、根据权利要求1所述的方法，在(1)中，所述游离核酸片段来自于孕妇外周血浆、孕妇肝脏和/或胎盘。

3、根据权利要求1或2所述的方法，在(1)中，所述游离核酸片段为游离DNA。

4、根据权利要求1-3任一项所述的方法，在(1)中，所述测序数据来自超低深度测序；优选地，所述超低深度测序的测序深度为 $1\times$ 、 $0.1\times$ 或 $0.01\times$ 。

5、根据权利要求1-4任一项所述的方法，在(1)中，将所述读段与参考基因组进行比对，获取唯一比对的读段（优选进行GC含量校正）；优选地，后续步骤使用所述唯一比对的读段（优选经GC含量校正后的读段）进行。

6、根据权利要求5所述的方法，GC含量校正按如下步骤进行：

a. 首先从人类参考基因组的某个染色体上随机采样 m 个长度为 ℓ 的片段；

b. 计算GC含量为 i 的片段数量 N_i ：

$$N_i = \sum_{k=1}^m \mathbb{I}_i(f(k))$$

其中 $\mathbb{I}_i(f(k)) = \begin{cases} 1, & f(k) = i \\ 0, & \text{其他} \end{cases}$ ， $f(k)$ 是片段 k 的GC含量， i 表示GC含量($i = 0\%, 1\%, \dots, 100\%$)；

c. 计算GC含量为 i 的测序读段数量 F_i ：

$$F_i = \sum_{k=1}^m c_k \cdot \mathbb{I}_i(f(k))$$

其中 $c_k \cdot \mathbb{I}_i(f(k))$ 表示片段 k 的GC含量， F_i 表示GC含量为 i 且测序读段起始点位与该片段起始点位相同的测序读段的数量；

d. 计算GC含量观测期望比 λ_i ：

$$\lambda_i = \begin{cases} r \cdot \frac{F_i}{N_i}, & F_i > 0 \text{ 且 } N_i > 0 \\ 1, & \text{其他} \end{cases}$$

其中 r 是全局比例因子，其定义为：

$$r = \frac{\sum_i N_i}{\sum_i F_i}$$

e. 校正测序读段数量：

$$R_i = \frac{1}{\lambda_i}$$

其中 R_i 表示经校正后GC含量为*i*的测序读段的期望数量。

7、根据权利要求1-6任一项所述的方法，在（1）中，孕妇表型特征数据选自如下的一个或多个的组合：年龄、孕周、身高、体重、BMI、产检生化检验检测结果、超声检查诊断结果和血浆中胎儿游离DNA浓度。

8、根据权利要求1-7任一项所述的方法，在（1）中，所述孕妇表型特征数据经异常值处理、缺失值处理和/或空值处理。

9、根据权利要求8所述的方法，在（1）中，孕妇样本的表型数据出现如下记录将被判定为异常值，并将这些异常值设置为空值：

- a. $x_{age} < 10$ 或 $x_{age} > 80$;
- b. $x_{GW} < 5$ 或 $x_{GW} > 50$;
- c. $x_{height} < 40$ 或 $x_{height} > 300$;
- d. $x_{weight} < 10$ 或 $x_{weight} > 200$ 。

10、根据权利要求8或9所述的方法，缺失值和空值使用missForest算法进行填充。

11、根据权利要求1-10任一项所述的方法，在（2）中，所述染色体为21号染色体、18号染色体、13号染色体和/或性染色体。

12、根据权利要求1-11任一项所述的方法，（2）中包括，（2.1）使用长度为**b**的窗口以步长 t 对参考基因组上的长度为L的染色体序列进行重叠滑动，获得若干滑动窗口，其中**b**为正整数，且**b**=[10000,10000000]， t 为任意正整数，L为正整数，且L≥**b**；（2.2）统计落入每一个所述滑动窗口内的读段，生成所述染色体序列的序列特征矩阵。

13、根据权利要求1-12任一项所述的方法，在（2）中，所述序列特征矩阵包括滑动窗口内读段数量、碱基质量和比对质量。

14、根据权利要求13所述的方法，所述碱基质量包括碱基质量的均值、标准差、偏度和/或峰值。

15、根据权利要求13所述的方法，所述比对质量包括比对质量的均值、标准差、偏度和/或峰值。

16、根据权利要求1-15任一项所述的方法，在（2）中，所述序列特征矩阵为：

$$X = (x_{ij})_{h \times w}$$

其中**h**表示窗口的滑动次数，**w**表示单个滑动窗口内的序列特征数量， x_{ij} 表示第*i*个滑动窗口中第*j*个序列特征值。

17、根据权利要求1-16任一项所述的方法，在（3）中，将所述序列特征矩阵进行标准化处理。

18、根据权利要求17所述的方法，在（3）中，使用公式（1）进行所述序列特征矩阵的标准化：

$$z_{i,j}^{(k)} = \frac{x_{i,j}^{(k)} - \mu_{i,j}}{\sigma_{i,j}} \quad (1)$$

其中， $z_{i,j}^{(k)}$ 为样本*k*经标准化处理后的序列特征矩阵， $x_{i,j}^{(k)}$ 表示样本*k*的第*i*个滑动窗口中的第*j*个序列特征值， $\mu_{i,j}$ 和 $\sigma_{i,j}$ 分别表示所有样本的第*i*个滑动窗口中的第*j*个序列特征值的均值和标准差。

19、根据权利要求1-18任一项所述的方法，在（3）中，所述训练的机器学习模型为神经网络模型或AutoEncoder模型；所述神经网络模型优选为深度神经网络模型，更优选为基于1维卷积的深度神经网络模型。

20、根据权利要求1-19任一项所述的方法，所述深度神经网络模型的结构包括：

输入层，用于接收所述序列特征矩阵；

前置模块，与所述输入层相连，用于对来自所述输入层的序列特征矩阵进行第一次卷积和激活操作，得到特征图谱；

核心模块，与所述前置模块相连，用于对来自所述前置模块的特征图谱进行进一步抽象和特征提取，通过有效增加神经网络模型深度从而加强神经网络表达能力；

后置模块，与所述核心模块相连，用于对来自所述核心模块的特征图谱进行特征抽象表示；

第一全局平均池化层，与所述后置模块相连，用于将所述特征抽象表示的特征图谱进行向量化，并输出所述染色体序列的序列特征向量。

21、根据权利要求20所述的方法，所述前置模块包括：

（I）1维卷积层；

（II）批标准化层，与（I）中所述1维卷积层相连；

（III）ReLU激活层，与（II）中所述批标准化层相连。

22、根据权利要求20或21所述的方法，所述核心模块由一个或多个结构相同的残差子模块组成，其中每个残差模块的输出是下一个残差模块的输入。

23、根据权利要求20-22任一项所述的方法，所述残差子模块包括：

（A）核心模块的前置子模块，每个子模块包括1维卷积层、与1维卷积层相连的Dropout层、与Dropout层相连的批标准化层，以及与批标准化层相连的ReLU激活层；

（B）第一1维平均池化层，与（A）中所述的核心模块的前置子模块相连；

（C）Squeeze-Excite模块，和/或Spatial Squeeze-Excite模块，与（B）中所述的第一1维平均池化层相连；

（D）第一加法层，与（C）中所述的Squeeze-Excite模块和/或Spatial Squeeze-Excite 模块相连；

（E）第二1维平均池化层，与所述前置模块中的ReLU激活层相连；

（F）第二加法层，与（D）中所述的第一加法层和（E）中所述的第二1维平均池化层相连。

24、根据权利要求23所述的方法，所述Squeeze-Excite模块包括：

（a）第二全局平均池化层，与所述残差子模块中的（B）第一1维平均池化层相连；

（b）Reshape层，与（a）中所述的第二全局平均池化层相连，所述Reshape层输出特征图大小为 $1 \times f$ ，其中 f 为1维卷积核数量；

（c）第一全连接层，与（b）中所述Reshape层相连，所述第一全连接层输出神经元个数为 $\frac{f}{r_{SE}}$ ，其中 f 为1维卷积核数量， r_{SE} 为Squeeze-Excite模块下降速度；

（d）第二全连接层，与（c）中所述第一全连接层相连，所述第二全连接层输出神经元个数为 f ，其中 f 为1维卷积核数量；

（e）Multiply层，与（d）中所述第二全连接层、以及与残差子模块中（B）第一1维平均池化层相连。

25、根据权利要求23或24所述的方法，所述Spatial Squeeze-Excite模块包括：

a. 一个 1×1 的1维卷积层，与（B）第一1维平均池化层相连，该 1×1 的1维卷积层使用sigmoid函数作为激活函数；

b. 一个Multiply层，与之连接的是（B）第一1维平均池化层和a中所述的 1×1 的1维卷积层。

26、根据权利要求1-25任一项所述的方法，在（4）中，将所述序列特征向量与所述孕妇表型特征向量拼接获得组合特征向量。

27、根据权利要求1-26任一项所述的方法，在（4）中，所述组合特征向量 x 经标准化处理：

$$x'_i = \frac{x_i - \mu_i}{\sigma_i}$$

其中 x'_i 是标准化后的组合特征向量 x 中的第 i 个序列特征值， x_i 是组合特征向量 x 中的第 i 个序列特征值， μ_i 是组合特征向量中第 i 个序列特征值的均值， σ_i 是组合特征向量中第 i 个序列特征值的标准差。

28、根据权利要求1-27任一项所述的方法，在（4）中，所述分类检测模型为集成学习模型。

29、根据权利要求28所述的方法，所述集成学习模型为基于堆叠或多数投票的集成学习模型；所述集成学习模型优选为以下的一项或多项：支持向量机模型、朴素贝叶斯分类器、随机森林分类器、XGBoos和逻辑回归。

30、根据权利要求1所述的方法，所述染色体异常包括以下至少一项或多项：21-三体综合征、18-三体综合征、13-三体综合征、5p-综合征、染色体微缺失和染色体微重复。

31、一种构建用于检测胎儿染色体异常的分类检测模型的方法，所述方法包括：

（1）获得孕妇的游离核酸片段的测序数据和临床表型特征数据，其中所述测序数据包括若干读段，并且所述孕妇的胎儿染色体情况已知，所述孕妇的临床表型特征数据形成孕妇表型特征向量；

（2）对参考基因组染色体的至少一部分进行窗口划分，获得若干滑动窗口，统计落入所述滑动窗口内的读段，生成所述染色体序列的序列特征矩阵；

（3）利用所述序列特征矩阵和胎儿染色体情况构建训练数据集，训练所述机器学习模型以提取所述染色体序列的序列特征向量；

（4）将所述序列特征向量和孕妇表型特征向量组合形成组合特征向量，利用所述组合特征向量和胎儿染色体情况训练所述分类模型，获得训练后的所述分类检测模型。

32、根据权利要求31所述的方法，所述孕妇的胎儿染色体情况为以下的一项或多项：正常二倍体、染色体非整倍性、部分单体综合征、染色体微缺失和染色体微重复；优选地，所述染色体非整倍性包括以下至少一项或多项：21-三体综合征、18-三体综合征和13-三体综合征；优选地，所述部分单体综合征包括5p-综合征。

33、根据权利要求32或33所述的方法，所述孕妇数量大于10个，并且正常二倍体的胎儿数量与染色体非整倍性的胎儿数量之比为 $\frac{1}{2}$ 至2。

34、根据权利要求31-33任一项所述的方法，在（3）中，所述训练数据集表示为：

$$D_{seq} = \{\mathbf{z}_{i,j}^{(k)}, y^{(k)}\}_{k=1}^N,$$

$$y^{(k)} = \begin{cases} 1, & \text{样本} k \text{ 的胎儿染色体异常} \\ 0, & \text{样本} k \text{ 的胎儿染色体正常} \end{cases}$$

其中， N 表示训练样本的数量，且 N 为 ≥ 1 的整数； $\mathbf{z}_{i,j}^{(k)}$ 为训练样本 k 经标准化处理后的序列特征矩阵，且 $k \in [1, N]$ ， i 为 ≥ 1 的整数， j 为 ≥ 1 的整数；

所述染色体异常包括以下至少一项或多项：21-三体综合征、18-三体综合征、13-三体

综合征、5p-综合征、染色体微缺失和染色体微重复。

35、一种检测胎儿染色体异常的系统，包括：

数据获取模块，用于获得待测孕妇样本的游离核酸片段的测序数据和临床表型特征数据，其中所述测序数据包括若干读段，所述待测孕妇样本的临床表型特征数据形成孕妇表型特征向量；

序列特征矩阵生成模块，用于对参考基因组染色体的至少一部分进行窗口划分，获得若干滑动窗口，统计落入所述滑动窗口内的读段，生成所述染色体序列的序列特征矩阵；

序列特征向量提取模块，用于将所述序列特征矩阵输入训练的机器学习模型以提取所述染色体序列的序列特征向量；

分类检测模块，用于将所述序列特征向量和孕妇表型特征向量组合形成组合特征向量，将所述组合特征向量输入分类检测模型，获得所述待测孕妇的胎儿染色体异常情况。

36、根据权利要求35所述的系统，所述系统还包括比对模块，用于将所述测序数据的读段与参考基因组比对，获取唯一比对的读段。

37、根据权利要求35或36所述的系统，在数据获取模块中，所述游离核酸片段来自于孕妇外周血浆、孕妇肝脏和/或胎盘。

38、根据权利要求35-37任一项所述的系统，在数据获取模块中，所述游离核酸片段为游离DNA。

39、根据权利要求35-38任一项所述的系统，在数据获取模块中，所述测序数据来自超低深度测序；优选地，所述超低深度测序的测序深度为 $1\times$ 、 $0.1\times$ 或 $0.01\times$ 。

40、根据权利要求35-39任一项所述的系统，在数据获取模块中，将所述读段与参考基因组进行比对，获取唯一比对的读段（优选进行GC含量校正）；优选地，后续步骤使用所述唯一比对的读段（优选经GC含量校正的读段）进行。

41、根据权利要求35-40任一项所述的系统，在数据获取模块中，孕妇表型特征数据选自如下的一个或多个的组合：年龄、孕周、身高、体重、BMI、产检生化检验检测结果、超声检查诊断结果和血浆中胎儿游离DNA浓度。

42、根据权利要求35-40任一项所述的系统，在数据获取模块中，所述孕妇表型特征数据经异常值处理、缺失值处理和/或空值处理。

43、根据权利要求35-42任一项所述的系统，在数据获取模块中，孕妇样本的表型数据出现如下记录将被判定为异常值，并将这些异常值设置为空值：

- a. $x_{age} < 10$ 或 $x_{age} > 80$;
- b. $x_{GW} < 5$ 或 $x_{GW} > 50$;
- c. $x_{height} < 40$ 或 $x_{height} > 300$;
- d. $x_{weight} < 10$ 或 $x_{weight} > 200$ 。

44、根据权利要求42或43所述的系统，缺失值和空值使用missForest算法进行填充。

45、根据权利要求35-44任一项所述的系统，在序列特征矩阵生成模块中，所述染色体为21号染色体、18号染色体、13号染色体和/或性染色体。

46、根据权利要求35-45任一项所述的系统，在序列特征矩阵生成模块中进行：（2.1）使用长度为b的窗口以步长t对参考基因组上的长度为L的染色体序列进行重叠滑动，获得若干滑动窗口，其中b为正整数，且 $b \in [10000, 10000000]$ ，t为任意正整数，L为正整数，且 $L \geq b$ ；（2.2）统计落入每一个所述滑动窗口内的读段，生成所述染色体序列的序列特征矩阵。

47、根据权利要求35-46任一项所述的系统，在序列特征矩阵生成模块中，所述序列特征矩阵包括滑动窗口内读段数量、碱基质量和比对质量。

48、根据权利要求47所述的系统，所述碱基质量包括碱基质量的均值、标准差、偏度和/或峰值。

49、根据权利要求47所述的系统，所述比对质量包括比对质量的均值、标准差、偏度和/或峰值。

50、根据权利要求35-49任一项所述的系统，在序列特征矩阵生成模块中，所述序列特征矩阵为：

$$\mathbf{X} = (\mathbf{x}_{ij})_{h \times w}$$

其中 h 表示窗口的滑动次数， w 表示单个滑动窗口内的序列特征数量， $\mathbf{x}_{ij}$ 表示第 i 个滑动窗口中第 j 个序列特征值。

51、根据权利要求35-50任一项所述的系统，在序列特征向量提取模块中，将所述序列特征矩阵进行标准化处理。

52、根据权利要求35-51任一项所述的系统，在序列特征向量提取模块中，使用公式(I)进行所述序列特征矩阵的标准化：

$$\mathbf{z}_{i,j}^{(k)} = \frac{\mathbf{X}_{i,j}^{(k)} - \mu_{i,j}}{\sigma_{i,j}} \quad (\text{I})$$

其中， $\mathbf{z}_{i,j}^{(k)}$ 为样本 k 经标准化处理后的序列特征矩阵， $\mathbf{X}_{i,j}^{(k)}$ 表示样本 k 的第 i 个滑动窗口中的第 j 个序列特征值， $\mu_{i,j}$ 和 $\sigma_{i,j}$ 分别表示所有样本的第 i 个滑动窗口中的第 j 个序列特征值的均值和标准差。

53、根据权利要求35-52任一项所述的系统，在序列特征向量提取模块中，所述训练的机器学习模型为神经网络模型或AutoEncoder模型；优选地，所述神经网络模型为深度神经网络模型；更优选地，所述神经网络模型为基于1维卷积的深度神经网络模型。

54、根据权利要求35-53任一项所述的系统，在分类检测模块中，将所述序列特征向量与所述孕妇表型拼接获得组合特征向量。

55、根据权利要求35-54任一项所述的系统，在分类检测模块中，所述组合特征向量 $\mathbf{x}$ 经标准化处理：

$$x'_i = \frac{x_i - \mu_i}{\sigma_i}$$

其中 x'_i 是标准化后的组合特征向量 $\mathbf{x}$ 中的第 i 个序列特征值， x_i 是组合特征向量 $\mathbf{x}$ 中的第 i 个序列特征值， μ_i 是组合特征向量中第 i 个序列特征值的均值， σ_i 是组合特征向量中第 i 个序列特征值的标准差。

56、根据权利要求35-55任一项所述的系统，在分类检测模块中，所述分类模型为集成学习模型。

57、根据权利要求56所述的系统，所述集成学习模型为基于堆叠或多数投票的集成学习模型；所述集成学习模型优选为以下的一项或多项：支持向量机模型、朴素贝叶斯分类器、随机森林分类器、XGBoos和逻辑回归。

58、一种用于构建用于检测胎儿染色体异常的分类检测模型的系统，所述系统包括：

数据获取模块，用于获得孕妇游离核酸片段的测序数据和孕妇的临床表型特征数据，其中所述测序数据包括若干读段，并且所述孕妇的胎儿染色体情况已知，所述孕妇的临床表型特征数据形成孕妇表型特征向量；

序列特征矩阵生成模块，用于对参考基因组染色体的至少一部分进行窗口划分，获得若干滑动窗口，统计落入所述滑动窗口内的读段，生成所述染色体序列的序列特征矩阵；

序列特征向量提取模块，用于利用所述序列特征矩阵和胎儿染色体情况构建训练数据

集，训练机器学习模型以提取所述染色体序列的序列特征向量；

分类检测模型获得模块，用于将所述序列特征向量和孕妇表型特征向量组合形成的组合特征向量，利用所述组合特征向量和胎儿染色体情况对分类检测模型进行训练，获得训练后的所述分类检测模型。

59、根据权利要求58所述的系统，所述系统还包括比对模块，用于将所述测序数据的读段与参考基因组比对，获取唯一比对的读段。