Noninvasive diagnosis of fetal aneuploidy by shotgun sequencing DNA from maternal blood

——通过DNA鸟枪法测序技术利用孕妇血液开展胎儿异倍体无创检查

# 研究背景

染色体异倍体及其他染色体畸变在新生儿中发生的概率约为千分之九，因而产前检查显得尤为重要。目前的产前检查主要分成有创产检和无创产检两类方法。

## 有创产检

主要是绒毛膜绒毛取样和羊膜腔穿刺两种方法，为染色体异常诊断的金标准，但是有创产检的方法有一定的流产风险。

## 无创产检

### 传统方法

传统的无创筛查主要依靠超声成像和产妇血清生化指标检查，但是这类方法可靠性较低。进而亟待其他高准确性的无创方法。

### 基于孕妇血液中存在游离胎儿细胞的原理

通过此方法探测胎儿基因的难度较高，目前没有实际应用。

### 基于孕妇血液中存在胎儿游离的DNA(cffDNA)的原理

#### 等位基因比例法

通过分析孕妇和胎儿在特定等位基因的突变比率来探测染色体异常的情况，但是由于受制于特定基因座多样性的原因这种方法的应用人群有限。

#### 本研究方法

基于胎儿染色体的异倍体会对孕妇血浆DNA含量有所影响这一现象，直接利用鸟枪法测序血浆DNA，通过统计分析各染色体匹配上的序列数量，判定胎儿的染色体异倍体情况。这种方法能够产生大量的短序列提高准确性，并且不需要区分DNA 是来自于孕妇还是胎儿扩大了应用人群范围。

# 研究方法

## 游离血浆DNA测序

游离血浆DNA样本取自18例孕妇血浆和1例男子血浆，另外还有一份来自该男子的全血的全基因组DNA(whole-blood genomic DNA)。取测序所得的序列片段中能够只匹配上一个染色体且最多只有一个错配的序列为有效序列，文中称为序列标签(sequence tags)，并对各染色体匹配上的序列片段加以统计分析。

### 染色体的序列标签数

匹配结果中，发现各样本中各染色体都存在着匹配上的序列片段不规则分布的情况，研究组认为这很有可能是人为因素造成的。为排除相关干扰，设置一个50kb大小的滑动窗(sliding window)分别在各个染色体上移动。统计各个滑动窗中的序列标签数，取该组序列标签数的中值代表该染色体匹配上的序列标签数。

### 序列标签密度

由于不同样本的序列标签总数不尽相同，所以在每个样本中，再取各常染色体的序列标签数的中值作为归一化常数，然后对所有的染色体的序列标签数进行归一化，得到的数值命名为序列标签密度(sequence tag density)。

*文章中虽未提及归一化的具体方法，理论上应该是：在某样本中，将某染色体的序列标签数除以该样本的归一化常数，得到的比值即为该样本中该染色体的序列标签密度。*

### 序列标签密度的修正

把不同样本（含参考全基因组DNA组）同一染色体的序列标签密度均值(mean)和标准差(standard deviation)分别和对应染色体中的GC含量作关联性分析，发现存在统计学上的显著性（均值的P＜10-9，标准差的P＜10-12），说明染色体匹配上的序列标签数和对应的GC含量有关。

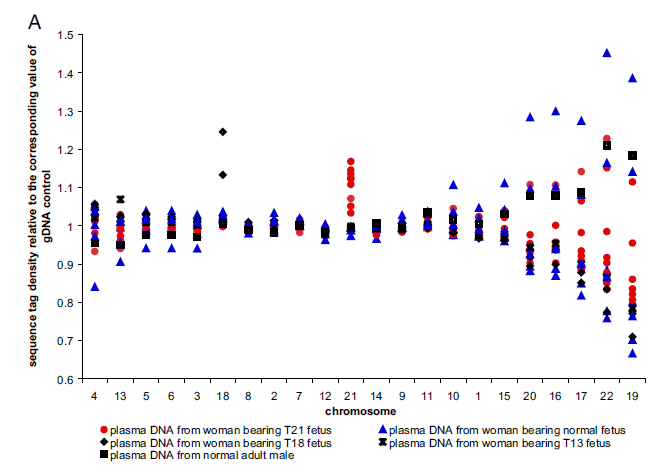
由于发现染色体匹配上的序列标签的GC含量普遍比人类基因组中的GC含量(41%)高约10%，研究组认为这种GC含量偏差是测序流程所致，所以用参考全基因组DNA对照组(genomic DNA control,数据显示全基因组DNA的GC含量不存在明显偏差)对各个样本各个染色体的标记序列密度加以修正。（*文章中未提及修正的具体方法，理论上应该是：将染色体的序列标签密度除以参考全基因组DNA组对应的染色体序列标签密度，得到的比值即为经修正的染色体序列标签密度。*）图A将各染色体上全部的序列标签密度绘制散点图，其中X轴上染色体的顺序按照其GC含量升序排列，可清晰地观察不同GC含量的染色体对应的序列标签密度的分布情况。

至此，得到了最终真正用于后续染色体异倍体检测使用的序列标签密度。

## 胎儿异倍体的检测

实际应用中，易发生染色体异倍体现象的染色体主要是21号、18号和13号。样本中确诊为异倍体胎儿有6例的T21，2例的T18，1例的T13。

利用上述修正后的序列标签密度作图，发现患病样本的非整倍体染色体的序列标签密度明显偏离正常染色体的序列标签密度。

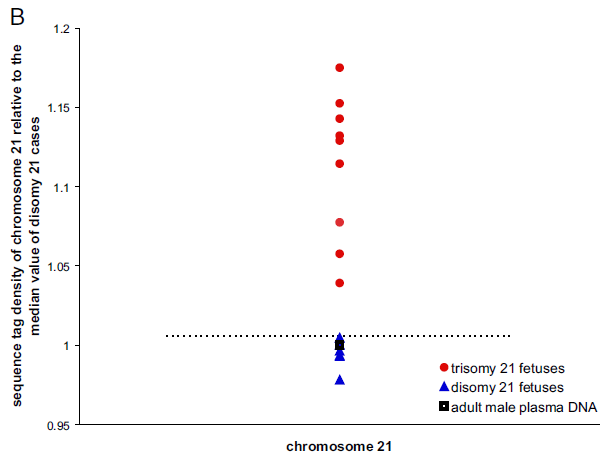


### T21的检测

从图A和图B均可看出，所有9例T21的21号染色体序列标签密度明显区别于整倍体样本的21号染色体序列标签密度。进一步计算可得，T21的21号染色体的覆盖率(coverage)为4%-18%（平均值为约11%），普遍高于正常样本中的21号染色体覆盖率。根据相关覆盖率可计算出T21中胎儿DNA比重（具体计算方法参见2.3.4）为8%-35%（平均值约为23%）。建立置信区间，将其作为判断T21和整倍体的21号染色体序列标签密度的依据。另外，如果将置信区间的边界值作为检测T21的界限值，那么胎儿DNA比重至少要为2%.

**置信区间的建立：**取样本中所有正常的21号染色体序列标签密度的中值作为参考值，将样本中所有的21号染色体的序列标签密度与参考值相比，并将所得比值作图。以正常胎儿的21号染色体序列标签密度为统计样本，画出99%的置信区间虚线。

图B结果显示，9例T21的21号染色体序列标签密度的相对值均远远大于1，且均在在设定的置信区间虚线之上；9例正常样本的21号染色体在设定的置信区间范围内。



### T18和T13的检测

虽然能用于分析的相关阳性样本较少，但是通过图A也能判断出T18的18号染色体序列标签密度（统计学上有显著差异，P＜10-7）和T13的13号染色体序列标签密度（统计学上有显著差异，P＜10-5）偏离对应的正常染色体标记密度值。

## 胎儿DNA在孕妇血浆中的比重

预测胎儿DNA在孕妇血浆中所有DNA的比重的方法主要有四种，其中基于SRY基因的TaqMan PCR预测法、基于X染色体序列标签的预测法和基于Y染色体序列标签的预测法可以预测男胎DNA在孕妇血浆中的比例；基于异倍型染色体序列标签的预测法可以预测患病胎儿DNA在孕妇血浆中的比例。

另外，研究发现胎儿DNA比重和孕龄之间存在着一定的关系。

### 基于SRY基因的TaqMan PCR预测法

测量位于1号染色体的EIF2C1基因数（一号染色体是孕妇和男胎共有的）和位于Y染色体（Y染色体只有男胎有而孕妇没有）的SRY基因数，分别代表相应染色体的数量，则男胎DNA的比重计算方法如下:

男胎DNA 比重= SRY基因数/（EIF2C1基因数/2）

由于人类细胞中所有染色体中有两条1号染色体，而男性的染色体中只有一条Y染色体，所以分母中EIF2C1基因数需要除以2。

利用本实验中的可用样本依据上述方法，计算得到的男胎DNA比重≤10%，符合过去的研究结论。

### 基于X染色体序列标签的预测法

假设胎儿DNA的比重为e，那么X染色体在男胎、女胎和孕妇中的比重分别是e、2e、2(1-e)，则怀男胎的孕妇血浆中X染色体序列标签密度=e+2(1-e)=2-e，怀女胎的孕妇血浆中X染色体序列标签密度=2e+2(1-e)=2。如果已知怀男胎的孕妇血浆中X染色体序列标签密度和怀女胎的孕妇血浆中X染色体序列标签密度（参考）的比值x，可以计算出男胎DNA的比重如下：

e=2(1-x) ← x=(2-e)/2

在本实验中，用于参考女胎的X染色体序列标签密度来自本实验中的相关样本，经计算可得本实验中的所有男胎DNA比重的平均值为19%左右（范围为8%-40%）。

### 基于Y染色体序列标签的预测法

已知男胎Y染色体在孕妇血浆中的序列标签密度和正常成年男性的Y染色体在自身血浆中的序列标签密度，则男胎DNA的比重为上述两个序列标签密度的比值。由于孕妇血浆中没有Y染色体，所以选用正常成年男性血浆中的Y染色体序列标签密度。

在本实验中，正常成年男性血浆中的Y染色体序列标签密度来自本实验的样本，经计算可得本实验中的所有男胎DNA比重的平均值为19%左右（范围为4%-44%）。

另外，将所有男胎样本通过X染色体和Y染色体计算得到的DNA比重做相关性分析，存在统计学上的显著性，说明该预测方法有一定的可信度。

### 基于异倍型染色体序列标签的预测法

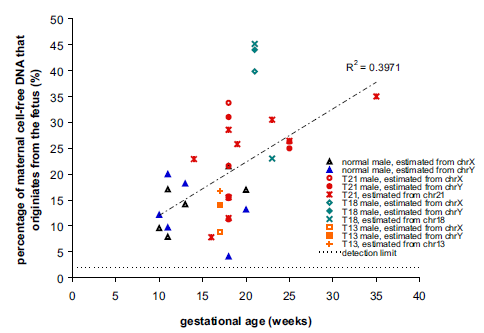
假设胎儿DNA的比重为e，那么某异倍型染色体在患病胎儿、正常胎儿和孕妇中的比重分别是3e、2e、2(1-e)，则患病的某异倍型染色体序列标签密度=3e+2(1-e)=2+e，正常胎儿对应的染色体序列标签密度=2e+2(1-e)=2。如果已知患病胎儿的异倍型染色体序列标签密度和正常胎儿的对应染色体序列标签密度（参考）的比值x，可以计算出男胎DNA的比重如下：

e=2(x-1) ← x=(2+e)/2

### 胎儿DNA比重和孕龄之间的关系

将通过上述三种基于序列标签预测方法得到的所有男胎DNA比重结果和对应的孕龄做散点图。另，由2.2.1可知，胎儿DNA比重至少要大于2%才能进行分析，本次进行分析的数据均大于2%符合要求。观察下图可得，用于分析的样本数据的孕龄基本集中在10周至25周之间。

将所有样本的胎儿DNA比重和孕龄做一元线性回归分析，其中同一样本经过不同方法预测得到的比重值取平均数作为该样本用于后续分析的胎儿DNA比重。所得斜线的P值为0.0051，R2为0.3971，结果表明胎儿DNA比重和孕龄之间存在统计学上的显著关系：随着孕龄的增加，胎儿DNA比重也相应有所增加。



## 游离血浆DNA的特点

### 游离血浆DNA片段大小的分布情况

分析发现，在所有样本中DNA片段大小的峰值和分布情况具有惊人的一致性：片段大小的峰值平均值为261bp（范围是256bp-264bp），去掉接头之后的实际长度为169bp，和一个核小体的长度相当。

为排除建库准备时的人为误差，取其中一例怀男胎的孕妇血浆DNA样本在454平台上测序。用上述中相同的方法分析，得到的序列主要长度是176bp。再结合过去的研究发现可知，大多数的游离血浆DNA片段来自于凋亡细胞。

但是将匹配上染色体Y的序列标签加以分析，得到的分布情况却与上述的分布情况不完全相符。不过，这仍然提供给我们一个重要信息：*可以尝试分析序列大小的分布情况而不仅仅分析序列大小的平均值*。

### 游离血浆DNA与核小体DNA的相同特征

由于发现游离血浆DNA主要来自凋亡细胞，所以进一步研究核小体DNA及其位点的特征是否能在血浆DNA中被发现。

核小体的一个特征是其分布和转录起始位点有关。根据此前的研究方法，以5bp的滑动窗为单位应用于RefSeq基因数据库中所有基因的2000bp(-1000bp~+1000bp)的转录起始位点序列，统计正义链和反义链上各位置的相关序列标签数。已知，正义链序列数上出现峰值表明核小体的起始，而反义链序列数上出现峰值表明核小体的结束。（相关原理及方法参看以下文献：Schones DE, et al. (2008) Dynamic regulation of nucleosome positioning in the human genome. Cell 132:887–898.）

经过统计发现，绝大多数的血浆DNA样本在转录起始位点序列附近至少出现了三组的峰值，而在任意剪切的基因组DNA序列上没有出现上述明显的现象。所以，可以看出游离血浆DNA 和核小体DNA有一定的相似特征。

# 研究结论

首先，该方法直接对孕妇血浆DNA进行测序，不仅避免了分离完整游离胎儿细胞的复杂性，而且受位点多态性限制较低而更能普遍应用。另外，由于cffDNA含量在孕妇分娩之后会在几个小时之后下降至难以侦测到的水平，所以不影响孕妇往后再次怀孕的检测。

另外，由于该方法是检测异倍型的染色体的相对含量，所以对于存在较少的非平衡易位型和染色体部分复制类型原则上也能检测出来。对于嵌合型的类型则需要进一步研究，而且受患者正常和异常细胞系的比重影响。

对于特定的染色体，匹配上的序列标签数因样本不同而明显变化，并且发现序列标签数与GC含量有着显著关系。经数据分析以及查阅他人研究（Solexa测序平台存在GC含量偏差），严重怀疑序列标签数的变化受到诸多不可控的人为因素等影响。幸运的是，较常出现异倍体现象的染色体有着较低的变化率和较高的灵敏度。

利用胎儿性染色体的序列标签数预估血浆中胎儿DNA的比重值偏高。可能的原因是Solexa平台倾向于扩增短序列片段，而血浆DNA样本中的短序列主要来自于胎儿，从而导致了相关估计值偏高。虽然分析结果与该可能原因并不十分吻合，但是由于存在人为误差以及序列标签选取不够完全等问题，所以仍然不能排除该可能性。

测序所得数据以及多项分析结果能够表明血浆中的游离DNA绝大多数来自于凋亡细胞裂解后的核小体。因为染色体上核小体的分布受到多方面因素的影响，所以需要注意的是，特定位点的标签数可能不能代表该位点所在的染色体数。

# 课题延伸

由于该方法能够充分利用到样本中的DNA片段，所以少量的样本量也能够进行有效的分析。

孕妇外周血液的提取时间应该在孕妇进行有创穿刺或绒毛提取之后多久进行，仍值得讨论。本实验是在进行有创提取样本约15-30分钟之后进行，结果显示无明显的异常。

美国妇产科医师学会目前将胎儿的染色体异倍体诊断作为产前检查必不可少的一项，而有创产检较高的流产率和传统无创产检的低灵敏度，加之大规模测序自身成本的下降，无疑都表明了该无创产检方法有着很大的应用前景。

利用大规模并行测序检查胎儿染色体异倍体的方法原理对围产期学、肿瘤学和器官移植等学科领域都有重要的启发意义。