微卫星序列

微卫星（英语：Microsatellite，亦称为简单重复序列（Simple Sequence Repeats，SSRs）或**短串联重复序列**（英语：short tandem repeats，**STRs**））是多型性的一种类型。指两个或多个核苷酸重复排列，且不同的重复序列相邻的形式，其中某些 DNA 序列(长度从一个到六个或更多的碱基对不等)被重复，通常为5-50次。，常见于非编码的内含子中。(二核苷酸微卫星、三核苷酸微卫星、四或五核苷酸微卫星都属于微卫星，即单重复mono-，双重复di-，三重复tri-，四重复tetra-，五重复penta-和六重复hexanucleotide，每种类型的总数随着重复单元大小的增加而减少。人类基因组中最常见的 str 是二核苷酸重复序列。另一方面，根据重复序列的结构，可将 str 分为完全重复序列(简单重复序列)和不完全重复序列(复合重复序列)，前者只包含一个重复单元，后者由不同组成的重复序列组成。)

微卫星和它们较长的近亲小卫星一起被归类为 VNTR (串联重复序列的可变数目variable number of tandem repeats) DNA。“卫星” DNA 这个名称指的是早期的观察，即离心试管中的基因组 DNA 可以从伴随的重复 DNA 的“卫星”层中分离出一个突出的体 DNA 层。它们被广泛用于癌症诊断、亲属关系分析(特别是亲子鉴定)和法医鉴定。它们还用于遗传连锁分析，以定位一个基因或突变负责给定的特征或疾病。微卫星也被用于测量亚种、群体和个体之间的群体遗传学。

> 目前用于法医分析的微型卫星都是四核苷酸或五核苷酸重复序列，因为这些重复序列提供了高度的无差错数据，同时短到足以在非理想条件下保存降解。即使是更短的重复序列也会受到 PCR 断断续续和优先扩增等伪影的影响，而更长的重复序列会受到生物可分解添加物的影响更大，并且 PCR 扩增效果也会更差。
>
> 法医学的另一个考虑是，必须尊重个人的医疗隐私，以便选择非编码、不影响基因调节的法医 str，而且通常不是可能涉及亨廷顿氏病等三联体扩增疾病的三核苷酸 str。

微卫星常被法医遗传学家和遗传系谱学学家称为短串联重复序列(short tandem repeats, str)，或者被植物遗传学家称为简单序列重复序列(simple sequence repeats, SSRs)。

非编码区的微卫星可能没有任何特定功能，因此可能无法进行选择对照; 这使它们能够不受阻碍地在代代之间积累突变，并产生可用于 DNA 指纹鉴定和识别目的的变异性。其他微卫星位于基因的调控侧翼或内含子区，或直接位于基因的密码子——在这种情况下微卫星突变可导致表型变化和疾病，特别是在脆性 x 综合征和亨廷顿氏病等三联体扩增疾病中。

由于 SSRs 通常位于启动子、非翻译区甚至编码序列中，这种突变可以直接影响基因功能的几乎任何方面。某些三联体重复序列的突变扩增是导致一些遗传性神经退行性疾病的原因，但 SSR 等位基因也可能导致大脑和行为特征的正常变异。

> 短串联重复序列(Short tandem repeats，str)和可变数目串联重复序列(variable number tandem repeats，VNTRs)又称微卫星(micro-and minatellites)，在操作上分别定义为 DNA 长度为1－6 bp 和≥7 bp 的串联重复单位。这些串联重复序列之间的突变率可能比基因组的独特部分高几个数量级，在 STRs 中每代$10^{-6到10}$-2个核苷酸不等。一个特定位点的突变率可能变化很大，而最长和最纯的串联重复序列通常定义了最不稳定的 STRs 和 VNTRs。因此，长期以来，strs/vntr 一直被认为是基因组中最多态的标记之一。它们也是基因组不稳定性的重要来源，这种不稳定性与几种人类疾病有关，包括重复扩增障碍，因为它们有通过复制滑移、 DNA 修复或非等位基因同源重组扩增的趋势。

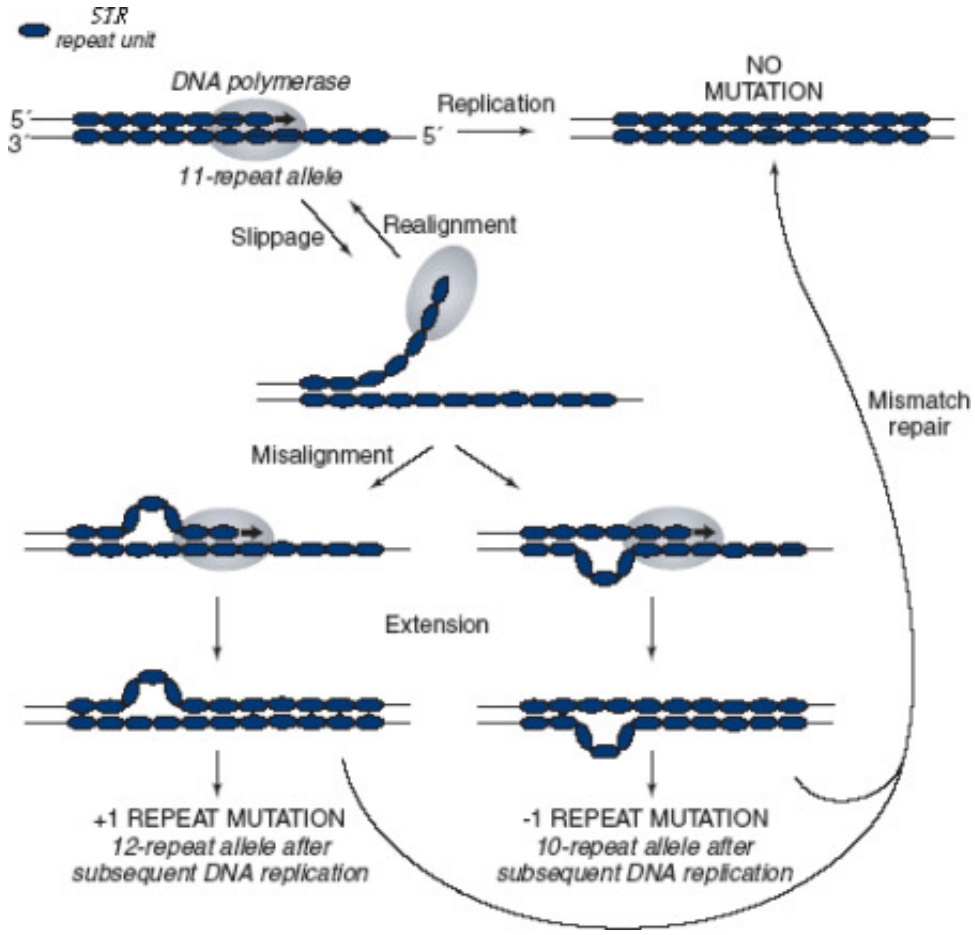
微卫星的突变机制

与点突变只影响单个核苷酸不同，微卫星突变会导致整个重复单位的增加或减少，有时会同时发生两个或更多的重复。因此，微卫星位点的突变速率预计将不同于其他突变速率，如碱基置换率。微卫星突变的真正原因是争论不休的。

微卫星序列突变可能是由于减数分裂中的不等交换（Unequal crossing over in meiosis）。这是一个众所周知的产生大块卫星 DNA 的机制。它与同源染色体之间重复单位的交换有关。然而，这一过程涉及不同的染色体，因此在 STR 突变中起着限制性作用。然而，这种机制可能是 STR 多步骤突变的原因 \citep{huang2002mutation}。

微卫星序列突变可能是由于反向转位机制（Retrotransposition mechanism）。这一机制推测，丰富的碱基转录因子是由逆转录本的3’延伸产生的，类似于 mRNA 的多聚腺苷酸化。有证据表明，最常见的人类STR与富含A的成分和转座因子之间存在联系。然而，转座因子的高密度并不总是与 STRs 的高密度相吻合。还需要进一步的研究来阐明它是否是 STR 突变的真正机制。

微卫星突变的原因除了点突变外，还有可能是由DNA聚合酶滑移（replication slippage）导致的。这也被称为 DNA 滑脱，聚合酶滑脱，或滑脱链错配。DNA 聚合酶是在复制过程中负责阅读 DNA 的酶，在沿着模板链移动时会滑落，并继续在错误的核苷酸上。DNA 聚合酶滑移更有可能发生在重复序列(如 cgcg)被复制时。由于微卫星是由这样的重复序列组成的，DNA 聚合酶在这些序列区域中可能会以较高的速率产生错误。一些研究已经发现证据表明滑动是微卫星突变的原因。通常，每个微卫星的滑移大约每1000代发生一次。因此，在基因组的其他部分，重复 DNA 的滑脱性变化比点突变更常见3个数量级。大多数滑移只会导致一个重复单位的改变，不同等位基因长度和重复单位大小的滑移率也不同，在不同的物种中也是如此。



但其滑移率与 STR 的表观突变率不一致。体外实验表明，DNA 滑移发生率非常高。但在活体内，大部分 DNA 环通过错配修复系统被识别和消除。研究表明，一个功能性错配修复系统可以降低 STR 突变率100到1000倍。因此，观察到的 STR 突变率取决于滑动率和修复系统纠正不匹配的效率 \citep{fan2007brief}。

位点间的突变率存在明显的差异。Chakraborty 等表明，在人类非致病性 STR 基因座中，二核苷酸重复序列的突变率最高，而四核苷酸 STR 基因座的突变率低50% 。然而，疾病相关的三核苷酸重复突变率超过正常值四至七倍。

目前已经发展了多种方法来阐明 STRs 的突变率，如家族方法\citep{huang2002mutation}、生物模型方法、种群方法\citep{di1998heterogeneity}和生殖细胞方法。家族方法是最直接的方法，其中突变率和突变类型都可以在从父母到后代的 STR 传播过程中直接检测\citep{huang2002mutation}。在生物模型方法中，将 STR 克隆到载体中，在宿主体内进行繁殖，从而评价 STR 突变的自发率，并估计各种因素对 STR 突变的影响。利用种群方法，可以检测到 STRs 的共同进化起源，并且可以追溯到许多代 \citep{hastbacka1992linkage}。STR 突变率也可以通过生殖细胞方法直接分析生殖细胞，特别是精子 \citep{holtkemper2001mutation}。

研究者调查了来自53个多世代家系的630名受试者在362个常染色体二核苷酸微卫星位点的97个突变事件的模式和特征。一个大小依赖的突变偏见(其中长等位基因偏向于收缩，而短等位基因偏向于扩展)被观察到。突变的大小(重复数在突变过程中改变)和突变的方向(收缩或扩展)与标准化等位基因大小之间存在显著的负相关关系。与早期在人类身上的发现相比，研究中的大多数突变事件(63%)是涉及多个重复单位变化的多步骤事件。突变率与重组率无相关性。研究数据表明，微卫星位点的突变动力学比广义逐步突变模型更为复杂\citep{huang2002mutation}。

微卫星序列突变对蛋白的影响

在哺乳动物中，20% 至40% 的蛋白质含有由短序列重复序列编码的重复序列。大多数基因组蛋白质编码部分的短序列重复序列都有三个核苷酸的重复单位，因为这个长度在突变时不会引起帧移位。每一个三核苷酸重复序列被转录成相同氨基酸的重复序列。在酵母中，最常见的重复氨基酸是谷氨酰胺、谷氨酸、天冬氨酸和丝氨酸。

这些重复片段的突变可以影响蛋白质的物理和化学性质，并可能产生逐渐和可预测的蛋白质作用的变化。例如，Runx2基因群体重复区的长度变化导致了家养犬(Canis familiaris)面部长度的差异，这与较长的序列长度和较长的面部长度有关。这种联系也适用于更广泛的食肉动物种类。HoxA13基因内多丙氨酸片段的长度变化与人类的手足生殖器综合症有关，这是一种发展障碍。其他三重复序列的长度变化与人类的40多种神经系统疾病有关，特别是脆性 x 综合征和亨廷顿舞蹈病等三重复序列扩展性疾病。复制滑移的进化变化也发生在更简单的生物体中。例如，微卫星长度变化在酵母表面膜蛋白中很常见，提供了细胞特性的快速进化。具体来说，FLO1基因的长度变化控制着对底物的粘附程度。短序列重复还能使致病细菌的表面蛋白发生迅速的进化变化，这可能使它们跟上宿主的免疫变化。真菌(粉色面包霉菌)中短序列重复序列的长度变化控制着其生物钟周期的持续时间。

微卫星序列对基因调控的影响

启动子和其他顺式调控区内微卫星长度的变化可以在世代间快速改变基因表达。人类基因组在调控区包含许多(> 16,000)短序列重复序列，为许多基因的表达提供了“调节旋钮”。通过改变启动子间距，细菌 SSRs 的长度变化可以影响流感嗜血杆菌菌毛的形成。二核苷酸微卫星与人类基因组顺式调控区的丰富变异有关。田鼠抗利尿激素1a 受体基因控制区的微卫星影响它们的社会行为和一夫一妻制水平。

在尤文氏肉瘤中，一个点突变产生了一个扩展的 GGAA 微卫星，这个微卫星结合了一个转录因子，这反过来又激活了驱动癌症的 EGR2基因。此外，其他 GGAA 微卫星可能影响与尤文肉瘤患者临床预后有关的基因的表达。（似乎GGAA微卫星序列可以调节增强子活性，也就是说，微卫星序列可以调节转录元件的活性？）

分析方法

重复 DNA 不容易使用下一代 DNA 测序方法。因此，微卫星通常通过传统的 PCR 扩增和幅度测定来进行分析，有时接着进行 Sanger DNA 测序。

> 在法医学中，分析是通过从感兴趣的样本细胞中提取细胞核 DNA，然后通过聚合酶链式反应扩增提取 DNA 的特定多态性区域来完成的。一旦这些序列被放大，它们就可以通过凝胶电泳或者毛细管电泳来解析，这样分析师就可以确定有问题的微卫星序列有多少个重复。

Detecting short tandem repeats using WGS

数据库

法医学数据库：

英国：SGM + 系统

美国：FBI CODIS Core STR Loci

澳大利亚：NCIDD

RAD-测序——ddRADseq数据

> ddRAD，是 RAD 测序协议的一种变体，用于 SNP 的发现和基因分型。在这种变化中，采用二次限制消化代替碎块剪切，提高了粒度选择步骤的可调性和准确性。
>
> 优势：不需要参考基因组；与全基因组测序相比，相对便宜；基因组覆盖度可以通过选择不同的限制性内切酶来调整。

生信分析工具

lobSTR

一种在个人基因组中分析 STRs 的新方法。利用信号处理和统计学习的概念来避免间隙对齐，并解决 STR 调用中的特定噪声模式。

lobSTR 实现提供了一个完整的解决方案，它采用原始测序数据并报告每个异常 STR 位点上的等位基因。程序的输入是一个或多个 FASTA/FASTQ 或 BAM 格式的序列库。输出是 BAM 格式 STR 读取的对齐方式，以及自定义制表符分隔文本格式中每个 STR 位点最可能的等位基因。lobSTR 支持多线程处理。

http://jura.wi.mit.edu/erlich/lobSTR/

http://lobstr.teamerlich.org/ （usage and download）

原文： https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3371701/

输入：fastq/fasta/bam（最好用bwa）格式的数据

输出：bam 格式的 STR reads 以及自定义制表符分隔文本格式中每个 STR 位点最可能的等位基因

在线教程：arvados https://dev.arvados.org/projects/arvados/wiki/LobSTR_tutorial

可视化：pybamview；samtools tview；UCSC; IGV

Tantem Repeat Database

地址：http://tandem.bu.edu/trf/trf.html

用户名：jinsy.liu@gmail.com

密码：prgu3556

可以提交genbank编号或上传数据进行在线分析，然后获得一份输出结果：

Indices	Period Size	Copy Number	Consensus Size	Percent Matches	Percent Indels	Score	A	C	G	T	Entropy (0–2)
1--61	5	12.2	5	76	23	58	39	60	0	0	0.97
2--62	2	33.0	2	75	15	60	40	59	0	0	0.98
11876--11935	27	2.2	27	93	0	102	16	25	25	33	1.96
12474--12839	48	7.6	48	91	0	561	25	11	29	34	1.90
13000--13113	45	2.5	45	81	5	140	48	5	33	13	1.64
13001--13128	15	8.5	15	79	8	118	46	5	33	14	1.67
14791--14821	13	2.4	13	94	0	53	58	9	12	19	1.62
24308--24367	27	2.2	27	93	0	102	16	25	25	33	1.96
24690--24780	21	4.3	21	79	5	103	2	16	26	54	1.53
25165--25278	45	2.5	45	81	5	140	48	5	33	13	1.64
25166--25293	15	8.5	15	79	8	118	46	5	33	14	1.67
25401--26304	135	6.7	135	90	2	503	20	14	32	31	1.93
25673--27105	135	10.6	135	95	0	2230	20	14	32	32	1.93
31123--31152	2	15.0	2	92	0	51	50	3	0	46	1.18
76836--76864	15	1.9	15	100	0	58	20	20	37	20	1.94
76983--77049	24	2.8	24	80	8	66	52	1	37	8	1.42
77502--77546	3	15.0	3	95	0	81	64	0	35	0	0.94
77577--77602	12	2.2	12	100	0	52	34	42	15	7	1.75
99945--99976	14	2.3	14	100	0	64	37	12	12	37	1.81
100371--100414	18	2.4	18	88	0	61	4	34	22	38	1.75
100465--100516	27	1.9	27	92	0	86	25	38	7	28	1.83
101479--101511	15	2.2	15	94	0	57	24	24	9	42	1.83

（该结果列出了STR开始的位点等信息）

STRScan

它使用贪婪算法对下一代测序(NGS)数据进行目标 STR 分析。测试了来自 Venter 基因组测序和全基因组测序千人基因组计划的数据。结果表明，STRScan 可以在目标集中描绘出 lobSTR 遗漏的20% 以上的 str。

> 用于在下一代测序(NGS)数据中对 STRs 进行剖析。在这里，我们采用了一种针对性的方法来进行 STR 分析: **我们尝试只研究用户定义的 STR 基因座子集，而不是在整个基因组范围内挖掘所有 STR (这是 lobSTR 或 STR-fm 的目标)**，这种方法对于法医或基因检测特别有用，因此避免了耗时的全基因组读取映射程序。因此，我们的方法不受参考基因组中以线性 DNA 序列表示的 STR 基因座序列比较的限制，可以采用微调比对算法进行 DNA 序列中的 STR 鉴定。除了挖掘全基因组测序数据外，我们的方法还可以直接应用于目标 STR 样本 NGS 数据中的 STR 分析，经过 STR 浓缩，或者特定 STR 基因座的 PCR 扩增(例如用于身份鉴定或基因检测)。

在 STRScan 中，每个 STR 基因座由一个或多个重复单元的串联拷贝以及重复单元之间的上游、下游和中间序列组成的模式表示，这些模式可以从生物体(如人类)的参考基因组序列中构建出来。

缺陷：从传统的下一代测序技术(例如，Illumina 公司的全基因组测序序列测定器)获得的短片段可能不适合用于 STR 的定向分析: 在全基因组测序数据中，只有少量的短片段能够被识别出来支持常见的 STR 面板(例如 y 染色体和 CODIS)。另一方面，相对较长的读取自 Illumina miSeq，可能达到500-600bps 的长度，相当于文特尔基因组数据集中的桑格测序读取长度，对于目标 STR 分析更加敏感(如表1所示)。结合特定 STR 基因座的目标扩增，miSeq 测序可以在 DNA 取证中获得满意的 STR 分型灵敏度，在基因疾病筛查中获得目标 STR 分型灵敏度。

http://darwin.informatics.indiana.edu/str/

原文：https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-017-1800-z

STR-FM

STR-FM，使用基于侧翼的映射的微卫星分析，一个计算流水线，可以检测短读数据的 STR 等位基因的全谱，可以适应新兴的读映射算法，并可以应用于异种遗传样本(如，肿瘤，病毒，和细胞器的基因组)。

原文：https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4417121/

HipSTR

HipSTR，一种新的基于单倍型的方法，可以从 Illumina 测序数据进行稳定的 STR 基因分型和分期。

原文：https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5482724/

安装与教程：https://hipstr-tool.github.io/HipSTR/

RepeatSeq

使用贝叶斯模型选择引导一个经验派生的错误模型，其中包括序列和读取特性。

原文：https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3592458/

安装与教程：https://github.com/adaptivegenome/repeatseq

STRaitRazor

STRait Razor，这是一个用于描述大规模并行处理机测序数据中短串联重复序列(short tandem repeats，str)单倍型特征的程序。

原文：https://pubmed.ncbi.nlm.nih.gov/28605651/

STRmix

STRmix是专家法医软件，可以解决以前无法解析的混合 DNA 图谱。

STRmix包含一个功能，允许软件直接在数据库中匹配混合的 DNA 图谱。对于那些没有嫌疑人，而且在一份样本中有来自多个捐献者的 DNA 的案件来说，这是一个重大进步。

STRsearch

STRsearch 不仅可以通过计算 STR 区域的重复模式和 INDELs 来确定等位基因，而且还可以将 MPS 结果转换为标准的 STR 命名法(数字和字母)。

原文：https://hereditasjournal.biomedcentral.com/articles/10.1186/s41065-020-00120-6

安装与教程：https://github.com/AnJingwd/STRsearch

GangSTR

原文：https://academic.oup.com/nar/article/47/15/e90/5518310

数据集

Analysis of tandem repeat variability in a three-generation nuclear family

https://www.ncbi.nlm.nih.gov/bioproject/PRJNA229524

原文章：https://academic.oup.com/nar/article/42/9/5728/2903185#119400513

研究流程/方法：

HapMap trio

https://www.ncbi.nlm.nih.gov/variation/news/NCBI_retiring_HapMap/

微卫星序列是否受表观遗传修饰/调控

发现了一个来自于人类 x 连锁的视网膜色素变性2 RP2启动子的 CpG 岛的外源性串联 GAAA 重复序列 230-bp，其5meCpG 状态与 XCI 相关。

AR CAG 和 RP2 GAAA多态性与正确识别母系来源的染色体优先失活（可以区分Xa和Xi）\citep{machado20145}

微卫星序列在T细胞里的突变方向

微卫星序列在干细胞中是否更稳定

微卫星序列可以用于鉴定细胞谱系、干细胞\citep{barallon2010recommendation}。微卫星分子遗传标记技术是目前为数不多的几种可用的 DNA 分析技术之一，目前正被提议用于人类细胞系、干细胞和组织的常规鉴定(认证)。与同工酶分析、核型分析、人类白细胞抗原分型等方法相比，该技术的优势在于，只要对适当数量和类型的基因座进行评估，STR 分析可以在个体水平上建立身份\citep{nims2010short}。

微卫星序列可以用于检测异基因造血干细胞移植后受者血液中的共者血细胞百分比，这一结果可以用于监测移植后进化，对预测移植物排斥反应和疾病复发的风险是有用的\citep{dumache2018chimerism}。

发现人类STR似乎容易富集在Alu元件附近。

来自父母哪一方的微卫星序列更稳定?

应用于人类身份鉴定的STR：

STR Loci	Chromosomal Location	Repeat Motif	Allele Range[a]	PCR Product Sizes in Identifiler Kit (dye label)
CSF1PO	5q33.1	TAGA	6–15	305–342 bp (6-FAM)
FGA	4q31.3	CTTT	17–51.2	215–355 bp (PET)
TH01	11p15.5	TCAT	4–13.3	163–202 bp (VIC)
TPOX	2p25.3	GAAT	6–13	222–250 bp (NED)
VWA	12p13.31	[TCTG] [TCTA]	11–24	155–207 bp (NED)
D3S1358	3p21.31	[TCTG] [TCTA]	12–19	112–140 bp (VIC)
D5S818	5q23.2	AGAT	7–16	134–172 bp (PET)
D7S820	7q21.11	GATA	6–15	255–291 bp (6-FAM)
D8S1179	8q24.13	[TCTA] [TCTG]	8–19	123–170 bp (6-FAM)
D13S317	13q31.1	TATC	8–15	217–245 bp (VIC)
D16S539	16q24.1	GATA	5–15	252–292 bp (VIC)
D18S51	18q21.33	AGAA	7–27	262–345 bp (NED)
D21S11	21q21.1	[TCTA] [TCTG]	24–38	185–239 bp (6-FAM)
D2S1338	2q35	[TGCC] [TTCC]	15–28	307–359 bp (VIC)
D19S433	19q12	AAGG	9–17.2	102–135 bp (NED)
Amelogenin (sex-typing)	Xp22.22 Yp11.2	Not applicable	Not applicable	X = 107 bp (PET) Y = 113 bp (PET)

The 13 core STR loci used for the U.S. national DNA database are shown in bold font. See www.cstl.nist.gov/biotech/strbase/multiplx.htm for information on other commercially available STR kits.
[a]Ranges are calculated from kit allelic ladders (see Figure 1) and do not represent the full range of alleles observed in world populations. A more complete allele listing of these short tandem repeat (STR) loci is available at www.cstl.nist.gov/biotech/strbase/str_fact.htm.

在从父母到子女的传播过程中，重复序列的长度有时会扩大。大多数重复序列没有可察觉的功能，但有些重复序列的数量超过某一特定阈值时有可能成为致病性。

利用STR构建细胞发育树

大致思路：

收集同一个体的不同时期/组织细胞测序数据；

lobSTR call STR；

> 原则上，任何突变信息都可能有助于谱系树的重建。我们关注 MS 即 微卫星 突变的原因如下:
>
> (1) MS 滑移突变，即在 MS 中插入或删除重复单位，被认为是在 DNA 复制过程中发生的，因此与细胞分裂耦合;
>
> (2) MS 突变发生率相对较高，并提供了广泛的选择范围;
>
> (3) MS 突变被认为是在不同的位点上独立发生的，通常不影响表型，而且由于大多数突变发生在非编码基因组序列中，因此不可能通过体细胞选择突变
>
> (4) MS 在人和小鼠中比较丰富
>
> (5) 错配修复基因突变的动物在所有组织中的错配修复基因[32,33]都表现出非常高的突变率，可用于实验和分析。

利用获得的STR和Nei's DA distance (1983)公式计算细胞间遗传距离：

$$D_A = 1 - \sum_{\ell} \sum_{u} \sqrt{X_u Y_u} / L$$

Nei's DA distance比较适用于微卫星，其计算过程中的假设是：

其他计算遗传距离的方法还包括：

Absolute distance – the distance between two samples（i and j is the average absolute differences between their number of repeats in all alleles which were analyzed in both samples) :

$$D(i,j) = \frac{1}{L} \sum_{l \in \{L\}} \left| A_i^l - A_j^l \right|$$

Normalized Absolute distance – the distance between two samples (i and j is the average normalized absolute difference between their number of repeats in all alleles which were analyzed in both samples): $D(i,j) = \frac{1}{L} \sum_{l \in \{L\}} \left| \frac{A_i^l}{\sum_{l \in \{L\}} \left| A_i^l \right|} - \frac{A_j^l}{\sum_{l \in \{L\}} \left| A_j^l \right|} \right|$

Euclidean distance – the distance between two samples: $D(i,j) = \sqrt{\sum_{l \in \{L\}} \left(A_i^l - A_j^l \right)^2}$

"Equal or not" distance – the distance between two samples (i and j is the number of alleles that differ between the two identifiers): $D(i,j) = \frac{1}{L} \sum_{l \in \{L\}} 1\left\{ \left(A_i^l - A_j^l \right) \neq 0 \right\}$

根据计算出的遗传距离构建n个细胞间的n x n遗传距离矩阵；

利用UPGMA算法构建细胞发育树（rooted tree）；

统计学检验：

Hardy–Weinberg expectations

> 对象应该是一群能进行繁殖的population

检测数据的整体显著性

> Holm-Bonferroni Method
>
> The **Holm-Bonferroni Method** (also called Holm's Sequential Bonferroni Procedure) is a way to deal with familywise error rates (FWER) for multiple hypothesis tests.

对于population，FST value and RST value

> *F*-statistics: Weir & Cockerham (1984) **Values** can range from 0 to 1. **High FST** implies a considerable degree of differentiation among populations.
>
> *R*ST: Slatkin (1995). Rst is a pairwise population genetic distance that is analogous to Fst, but that takes into account differences in the number of repeats between microsatellite alleles (allele size). 该检验表明等位基因大小是否提供了更多的种群分化信息。Significant tests on Rst values are expected if populations had diverged for a sufficiently long time and/or if populations exchanged migrants at a rate similar or inferior to the mutation rate (Hardy et al., 2003).
>
> Differences between *F*ST/*R*ST values for different loci combinations were tested using the Wilcoxon rank sum test and the median test.

对于population，Migration indices (*Nm*)

> 用 RST 代替 FST
>
> 应用公式 *F*ST=1/(4*Nm*+1) Wright (1969)

参考文献

```
@article{barallon2010recommendation,
  title={Recommendation of short tandem repeat profiling for authenticating human cell lines, stem cells, and tissues},
  author={Barallon, Rita and Bauer, Steven R and Butler, John and Capes-Davis, Amanda and Dirks, Wilhelm G and Elmore, Eugene and Furtado, Manohar and Kline, Margaret C and Kohara, Arihiro and Los, Georgyi V and others},
  journal={In Vitro Cellular \& Developmental Biology-Animal},
  volume={46},
  number={9},
  pages={727--732},
  year={2010},
  publisher={Springer}
}

@article{dumache2018chimerism,
  title={Chimerism Monitoring by Short Tandem Repeat (STR) Markers in Allogeneic Stem Cell Transplantation.},
  author={Dumache, Raluca and Enache, Alexandra and Barbarii, Ligia and Constantinescu, Carmen and Pascalau, Andreea and Jinca, Cristian and Arghirescu, Smaranda},
  journal={Clinical laboratory},
  volume={64},
  number={9},
  pages={1535--1543},
  year={2018}
}
```

```
@article{nims2010short,
  title={Short tandem repeat profiling: part of an overall strategy for
reducing the frequency of cell misidentification},
  author={Nims, Raymond W and Sykes, Greg and Cottrill, Karin and Ikonomi,
Pranvera and Elmore, Eugene},
  journal={In Vitro Cellular \& Developmental Biology-Animal},
  volume={46},
  number={10},
  pages={811--819},
  year={2010},
  publisher={Springer}
}

@article{huang2002mutation,
  title={Mutation patterns at dinucleotide microsatellite loci in humans},
  author={Huang, Qing-Yang and Xu, Fu-Hua and Shen, Hui and Deng, Hong-Yi and
Liu, Yong-Jun and Liu, Yao-Zhong and Li, Jin-Long and Recker, Robert R and
Deng, Hong-Wen},
  journal={The American Journal of Human Genetics},
  volume={70},
  number={3},
  pages={625--634},
  year={2002},
  publisher={Elsevier}
}

@article{di1998heterogeneity,
  title={Heterogeneity of microsatellite mutations within and between loci, and
implications for human demographic histories},
  author={Di Rienzo, Anna and Donnelly, Peter and Toomajian, Chris and Sisk,
Bronwyn and Hill, Adrian and Petzl-Erler, Maria Luiza and Haines, G Ken and
Barch, David H},
  journal={Genetics},
  volume={148},
  number={3},
  pages={1269--1284},
  year={1998},
  publisher={Genetics Soc America}
}

@article{hastbacka1992linkage,
  title={Linkage disequilibrium mapping in isolated founder populations:
diastrophic dysplasia in Finland},
  author={H{\"a}stbacka, Johanna and de la Chapelle, Albert and Kaitila, Ilkka
and Sistonen, Pertti and Weaver, Alix and Lander, Eric},
  journal={Nature genetics},
  volume={2},
  number={3},
```

```
  pages={204--211},
  year={1992},
  publisher={Nature Publishing Group}
}

@article{holtkemper2001mutation,
  title={Mutation rates at two human Y-chromosomal microsatellite loci using small pool PCR techniques},
  author={Holtkemper, Ulrike and Rolf, Burkhard and Hohoff, Carsten and Forster, Peter and Brinkmann, Bernd},
  journal={Human molecular genetics},
  volume={10},
  number={6},
  pages={629--633},
  year={2001},
  publisher={Oxford University Press}
}

@article{huang2002mutation,
  title={Mutation patterns at dinucleotide microsatellite loci in humans},
  author={Huang, Qing-Yang and Xu, Fu-Hua and Shen, Hui and Deng, Hong-Yi and Liu, Yong-Jun and Liu, Yao-Zhong and Li, Jin-Long and Recker, Robert R and Deng, Hong-Wen},
  journal={The American Journal of Human Genetics},
  volume={70},
  number={3},
  pages={625--634},
  year={2002},
  publisher={Elsevier}
}

@article{fan2007brief,
  title={A brief review of short tandem repeat mutation},
  author={Fan, Hao and Chu, Jia-You},
  journal={Genomics, proteomics \& bioinformatics},
  volume={5},
  number={1},
  pages={7--14},
  year={2007},
  publisher={Elsevier}
}
```