doi: 10.7541/2020.062

基于微单体型分子标记的草鱼亲子鉴定方法

夏 雷[1,2] 石米娟[1,3] 张婉婷[1,3] 段 攸[1,2] 程莹寅[1,3] 吴 南[1,3] 夏晓勤[1,2,3]

(1. 中国科学院水生生物研究所, 武汉 430072; 2. 中国科学院科学院大学, 北京 100149;
3. 中国科学院种子创新研究院, 北京 100101)

摘要: 研究选择一种新型分子标记——微单体型用于亲子鉴定, 构建了高效的标记筛选和亲子鉴定流程, 并以草鱼(*Ctenopharyngodon idellus*) 为例评估了该亲子鉴定方法的效果。结果表明, 利用基因组重测序数据能够准确完成微单体型标记的分型, 效果和适应性明显优于传统的基于群体遗传学推断的分型; 通过信息熵的大小能够高效筛选微单体型标记组合, 3个和5个微单体型标记的亲子鉴定结果与微卫星序列(SSR)鉴定结果的一致性分别达到97.08%和99.42%。研究表明使用微单体型分子标记可以快速而准确地完成鱼类的亲子鉴定工作。

关键词: 微单体型; 亲子鉴定; 草鱼; 高通量测序

中图分类号: Q-331 **文献标识码:** A **文章编号:** 1000-3207(2020)03-0509-09

鱼类育种技术是水产养殖业可持续发展的重要保障, 目前已有大量的遗传育种手段应用于鱼类育种[1]。在各类经济鱼类的选育过程中, 所得到的子代群体往往数量庞大, 为了保证饲养环境条件的一致而将多个家系群体同池混养, 后期重建亲本与子代之间的对应关系时, 需要使用亲子鉴定技术[2]。

在通常情况下, 亲子鉴定主要根据孟德尔遗传定律, 以分子标记作为依据来判断具体的子代与亲代之间是否存在亲子关系。目前亲子鉴定主要采用微卫星(Simple Sequence Repeat, SSR)和单核苷酸多态性(Single Nucleotide Polymorphism, SNP)这两种分子标记, 其中SNP在人类亲子鉴定中应用较多[3], 而SSR主要应用于水产养殖育种[2, 4—6]。SSR具有信息含量高、多态性好的优点[7], 但标记本身的筛选过程比较繁琐, 而且后续的亲子鉴定实验也较多地依赖于人力劳动, 样本量较大时, 耗时长, 效率低; 相比之下, SNP标记则具有不易发生变异和易于分析的优点, 但其多态性较低[8], 需要使用较多的SNP标记才可达到与SSR标记同等的效果。理论上SSR标记可以直接从测序数据与参考基因组比对后所得的插入缺失(INDEL)区域中进行筛选而获得[9, 10], 但直接获得SSR的方法都受限于reads长度及碱基滑移方式, 对于重复次数相近或超过reads长度的SSR无法进行有效分型[10], 因此主要还是依靠传统的实验手段来检测。单个SNP位点的分型也可以通过多种测序技术获得, 2011年Davey等[11]对这些技术进行了总结, 包括重测序技术和RAD-seq (Restriction-site-associated DNA sequencing)等简化基因组测序技术, 其中重测序技术可以获得全基因组范围内的SNP位点, 而简化基因组测序技术只能获得部分SNP位点, 但相较而言其成本较低。在水产动物的遗传育种研究中, SSR分子标记主要用于亲子鉴定[12—14]和群体遗传多样性分析[15]等工作, SNP分子标记则主要应用于遗传连锁图谱构建[16]和GWAS分析[17]。

2013年Kidd等[18]结合上述两种分子标记的优点, 提出了微单体型(Microhaplotype, MH)的概念, 这是指长度在200 bp以内、可连锁遗传的SNP组合。作为一种分子标记, MH多态性高并且能够稳

收稿日期: 2019-05-10; **修订日期:** 2019-10-28

基金项目: 国家自然科学基金(31571275和31801055); 中国科学院战略先导专项A类项目子课题(XDA08020201)资助 [Supported by the National Natural Science Foundation of China (31571275, 31801055); the Strategic Priority Research Program of the Chinese Academy of Sciences (XDA08020201)]

作者简介: 夏雷 (1994—), 男, 博士研究生; 研究方向为水体生物信息分析。E-mail: xialei@ihb.ac.cn

通信作者: 夏晓勤 (1970—), 研究员; E-mail: xqxia@ihb.ac.cn

定遗传, 已应用于法医行业[18—20]、人类群体分析[21, 22]及祖先推断[23, 24]等。2018年, Garciafernandez等[25]以每个目标基因中的所有SNP位点作为一个单体型(Haplotype)分子标记, 即长片段的连锁遗传的SNP组合, 应用于鲷(*Sparus aurata*)的亲子鉴定。另外微单体型分子标记也曾用于分析墨绿平鲉(*Sebastes atrovirens*)的亲缘关系[26], 然而迄今没有应用于鱼类亲子鉴定。

目前获取微单体型或单体型分型的方法主要有两种: 第一种方法是根据所获得的各个SNP位点, 用群体遗传学的手段进行单体型推断; 第二种方法是直接用个体基因组的测序数据的单体型组装。用群体遗传学推断方法的常用软件HaploView[27]、PHASE[28]、SHAPEIT[29]和Beagle[30]等, 它们对群体大小的依赖性较强, 对于大样本群体可以获得较准确的分型结果[31]。然而, 在鱼类的遗传育种中亲本数量通常较小[4, 15, 32], 难以获得适用于此方法的足够数量样品, 不容易做到准确分型。基于序列组装的软件主要有HapCUT2[33]、ReFHap[34]以及FLfinder[35]等, 目前仅用于单体型的推断与分型, 而未应用于微单体型的分析。这种方法不受样本数量限制, 仅对测序深度敏感, 测序深度不足时无法获得准确的分型结果[36], 但随着高通量测序成本的迅速下降, 测序深度已不再构成瓶颈问题。

本研究在基于个体测序数据的单体型组装的基础上, 开发了一套基于微单体型分子标记的亲子鉴定流程, 其核心为微单体型标记的获取、分型与筛选, 所需要样本更少且准确率高, 适用于已有参考基因组的二倍体水产动物。为评估新方法的效果, 我们以一个草鱼群体的全基因组重测序数据为例, 以其亲子鉴定结果与SSR鉴定结果的一致性为主要指标对新方法加以验证和评估。

1 材料与方法

1.1 基于个体序列组装的微单体型获取与亲子鉴定流程

我们通过个体单体型组装软件获取微单体型, 整个亲子鉴定流程主要分为三个步骤(图 1): 第一步是获取亲本微单体型。通过与参考基因组进行比对, 获取每个亲本个体的所有SNP位点, 依据测序结果使用HapCUT2软件进行亲本个体的单体型组装, 从中选取微单体型区域, 将所有个体的区域集合起来, 获取在该基因片段上的微单体型区域。按照划分好的微单体型区域, 对每个个体进行微单体型组装及分型。然后对子代进行微单体型的分型。对于上面从亲本中得到所有微单体型区域, 为每个子代个体进行微单体型组装。将该区域上子代与亲本的全部分型结果形成列表, 依据各分型的索引构建每个个体在该微单体型下的分型结果。最后根据全部子代分型结果, 用CERVUS 3.0软件[37]进行亲子鉴定。

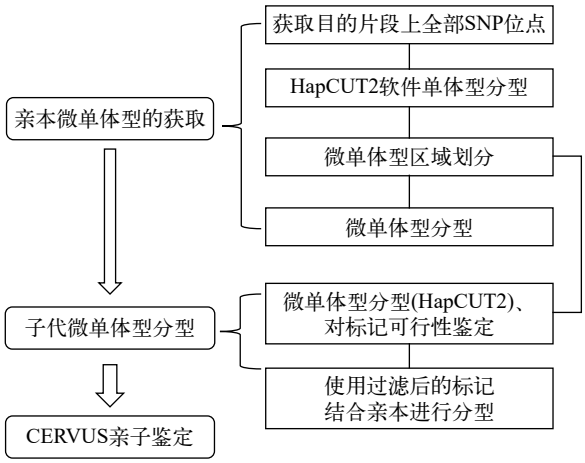

图 1 微单体型获取和亲子鉴定流程

Fig. 1 Pipeline for microhaplotype genotyping and paternity testing

1.2 微单体型区域(Microhaplotype region, MR)的划分

用HapCUT2软件处理亲本的重测序数据, 获取每个亲本个体的MR, 即一个单体型的首尾SNP位点在参考基因组上的位置。集中所有亲本个体的MR, 合并任何在位置上有重叠的MR, 构成一个序列较长的、信息量更高的MR总库。

接下来将总库中的每一个MR逐一与各亲本的MR进行比较。由于总库MR是合并所有亲本中相重叠MR所得, 那么, 每一个总库MR必然不短于单个亲本中与之有重叠的MR, 如果其非重叠区域(超出的部分)在该亲本序列中也存在杂合的SNP位点, 那么这个MR将按重叠与非重叠区域拆分成两个区域, 仅保留至少有2个SNP位点的区域作为新的MR, 这样可增加后续分析结果的准确性。最终所得的MR总库将用于后续分型研究。

1.3 微单体型分型及分析方法

依据MR总库可对各亲本进行微单体型的分型, 即针对总库中的每个MR确定各亲本的具体微单体型序列。二倍体亲本在每个MR最多只可能有2种微单体型, 这可以根据HapCUT2的分型结果筛选, 也可以参考该亲本与参考基因组比对所得的VCF文件来确定。如果由于测序错误或数据量不足, 无法明确地鉴定某MR的1—2种微单体型序列, 则舍弃该MR。剩下的各有效MR中所得到微单体型序列都可仅保留各SNP位点的碱基, 然后计算各

MR在亲本群体中的信息熵, 公式如下:

$$H = -\sum_{i=1}^{n} p_i \log_2 p_i$$

式中, n表示该MR中微单体型序列的种类数, p_i表示第i种微单体型在亲本群体中出现的频率, H表示该MR在亲本群体中的信息熵。

使用与亲本微单体型分型的方法, 也可以在每个子代中获得各MR的微单体型序列, 用于后续的亲子鉴定工作。

1.4　基于群体遗传学方法的单体型分子标记获取与分析

传统上单体型标记是通过群体遗传学分析推断而获得的, 因此本研究也使用此方法进行了单体型分子标记的分型。分析过程比较简单: 首先通过SHAPEIT软件将每个基因构建一个单体型, 再依据各分型出现顺序标注每个个体的分型情况。

1.5　使用SSR标记进行亲子鉴定的方法

我们从草鱼参考基因组[36]中筛选部分短序列重复片段作为候选SSR标记, 并设计引物对2尾父本(1#和M10)和3尾母本(2#, F13和F40)进行扩增, 扩增程序为95℃预变性10min, 95℃变性45s, 55℃退火45s, 72℃延伸1min, 循环30次, 72℃延伸10min, 4℃保存。使用ABI 3730对各标记的扩增产物进行毛细管凝胶电泳, 以条带大小为分型依据判断其多态性。最终选取5个多态性高的SSR标记进行亲子鉴定, SSR标记的相关信息见表 1。171尾子代个体也同样进行目标SSR标记扩增及分型, 然后进行亲子鉴定分析。

表 1　用于亲子鉴定的SSR标记信息

Tab. 1　Information of the SSRs used for the paternity test

标记ID Marker ID	PCR引物PCR primer (5′—3′)	重复片段 Repeat motif	信息熵 Informative index
G5010	CATTTTACTGCTTGCCTCAC CCCTTCCTTTCGCATAGA	AGAAG	2.4464
G5011	AAGCCACCAACCTCTACGA TAACAGGGATGGGATGAAAT	TTCTC	2.6464
G5012	GATGACATGGGGGTGAGTAA CAGAAAGGTAGTAAACAACGAAA	AGAGA	2.7219
G5020	CAACCCTGTTTCTGTCCTGT GCAAGCAACTGTCAACCTG	AAAGG	2.4464
G5024	ATTCCTTCCGAAATCAGTG AGAGGGAGAAAGATAAGACCA	GAGAA	2.1710

注: 信息熵计算方法与1.3中一致

Note: The informative index is calculated in the same way as 1.3

1.6　组装微单体型标记使用的数据及其处理方法

我们以草鱼基因组重测序数据(尚未发表)进行了微单体型组装, 数据来自上述五尾亲本及其171尾子代个体, 其中亲本测序深度为30X, 子代测序深度为15X, 并且评估了微单体型标记应用于亲子鉴定的结果与SSR鉴定结果的一致性。重测序数据经过了质量分析(FastQC)、与参考基因组的比对(BWA[38])等步骤, 最后利用SAMtools[39]和GATK 4.0[40]软件获取各样本的SNP位点, 作为SNP信息集。

考虑到子代个体数并不多, 我们仅选择了15个富含SNP的基因, 即*adamts*20(ADAM metallopeptidase with thrombospondin type 1 motif 20)、*brca*2 (breast cancer 2)、*dlc*1(deleted in liver cancer 1)、*gbp*(guanylate binding protein)、*lgals*9(galectin 9)、*lrp*5(LDL receptor related protein 5)、*meis2b*(meis homeobox 2)、*mrps*23(mitochondrial ribosomal protein S23)、*msi*2(musashi RNA binding protein 2)、*nos2b*(nitric oxide synthase 2)、*prtga*(protogenin A)、*rpz*4(rapunzel 4)、*snx*14(sorting nexin 14)、*thsd*4 (thrombospondin type 1 domain containing 4)和*zmym*4(zinc finger MYM-type containing 4), 以代表整个基因组来评估我们方法的效果, 这样就大大减少了计算量。从这15个基因中SNP的密度分布可以看出, 在这些基因中确实存在一些SNP较为集中的片段(图 2)。微单体型由连锁遗传的SNP位点组合构成且这些SNP位点通常距离较近, 因此这些分布较为集中的SNP可以组合成潜在的微单体型。

1.7　亲子鉴定标记组合(Paternity Test Marker Combinations, PTMC)筛选

为了获得PTMC, 我们首先依据亲本性别构建全部可能的亲本对, 为降低连锁遗传带来的影响, 随后依次挑选若干来源于不同基因的标记, 依据亲本的分型情况, 可得到每一个亲本对的后代在这些标记中的所有可能的分型组合。若存在某两个亲本对的子代分型有重合的情况, 则认为所选的标记组合无法进行亲子鉴定; 反之, 若所有亲本对的子代分型均唯一, 则认为所选标记组合可以用于亲子鉴定, 并作为可用的亲子鉴定标记组合(PTMC)输出。通过此步骤可以得到所有可能的PTMC, 从中选用平均信息熵最高的一套PTMC用于亲子鉴定可以提高结果的准确性。

基因组重组可能在子代中产生异于亲本的分型类型, 从而妨碍亲子鉴定。适当增加冗余标记可以应对这种分型矛盾。冗余标记的选择可以采用

一种简单的策略, 即从与所有已选用标记位于不同基因(或染色体)上的标记中, 选择信息熵最高的一个。重复该过程则可以选择多个冗余标记。

2 结果

2.1 基于个体序列组装的微单体型获取

我们亲本重测序数据对15个基因区域进行了微单体型的组装, 统计了每个基因上微单体型区域和SNP位点的数量, 以及第一个SNP位点到最后一个SNP位点之间的距离(表 2), 并计算信息熵作为评价微单体型区域信息量的标准。总体而言, 大部分微单体型区域的信息熵的值介于1和2之间(图 3A), 但其分布范围在各基因之中有较大的变化(图 3B), 并且信息熵值也随SNP数量的增长而呈上升趋势(图 3C)。

2.2 用单个基因的MR进行的亲子鉴定

亲子鉴定使用的软件为CERVUS 3.0[37], 参数为默认值。

由于亲本的测序深度高于子代, 有些从亲本上鉴定的MR序列在子代中没有被覆盖, 只有在所有子代中出现的MR才是有效MR。我们首先尝试将一个基因上的全部有效MR作为一套标记进行亲子鉴定, 其结果与用SSR鉴定结果的一致性见表 2。不同基因的鉴定能力差异很大, 其中*adamts*20、*brca*2、*prtga*和*snx*14鉴定的结果与SSR分析的一致性较高(>95%), 而*gbp*、*lrp*5、*meis2b*和*mrps*23的一致性较低(<50%)。亲子鉴定的效果与基因中微单体型标记的数量或信息熵之和存在一定的关系, 但

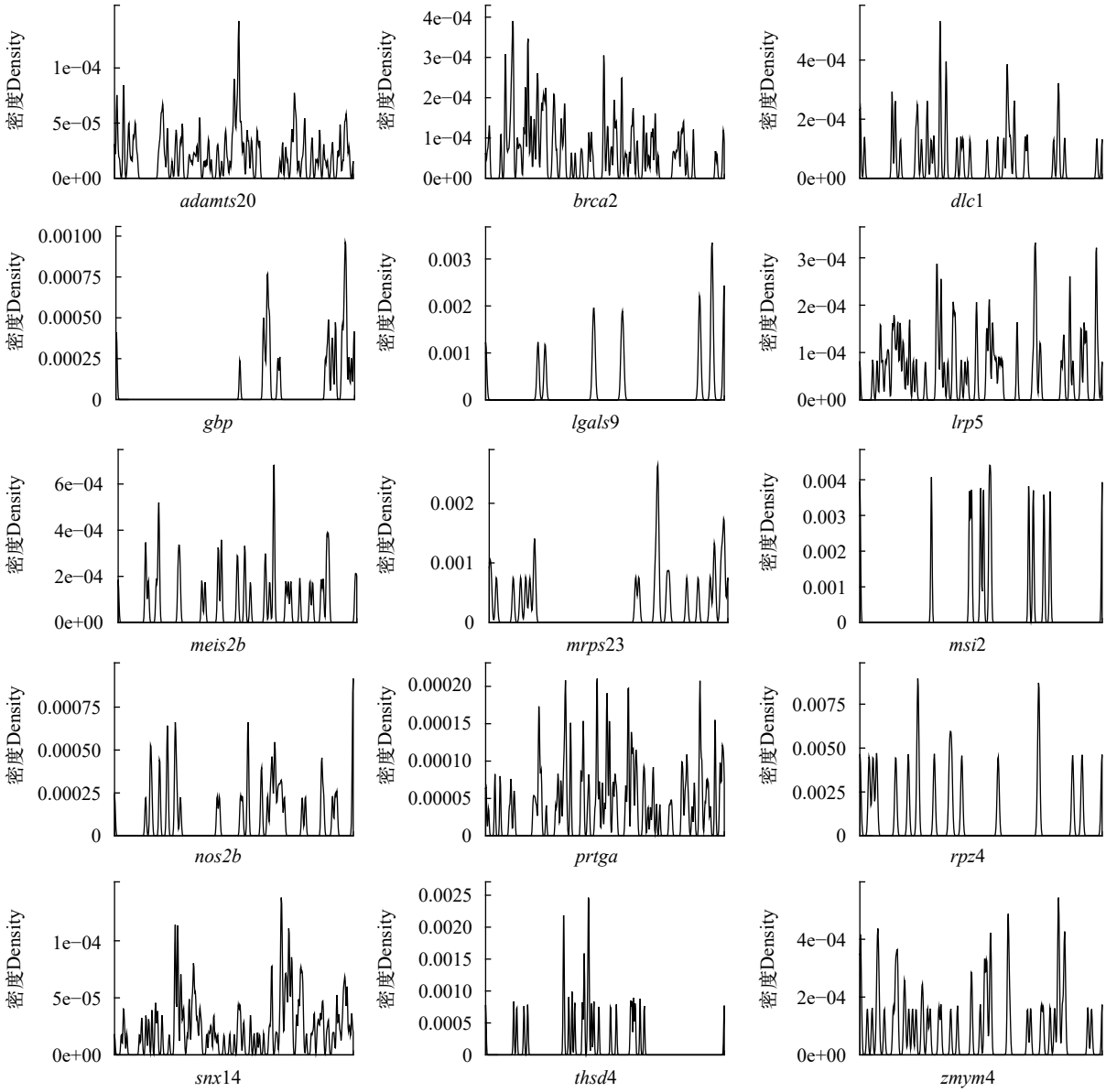

图 2 15个基因中SNP分布密度

Fig. 2 Distribution of SNPs in 15 genes

并非简单的线性关系(R^2=0.2302)。对于我们所测试的数据, 当基因中标记数或信息熵之和低于10时, 亲子鉴定的效果与之有正相关性, 但之后即达到饱和, 与SSR分析的一致性维持很高(图 4)。

2.3　MR组合的选择与亲子鉴定

如果不考虑基因重组, 根据5个亲本所有标记的实际分型情况, 理论上仅需3个标记就能分辨出它们之间任何子代的亲本, 而这样的3标记组合总共有189个。从中选择平均信息熵最高的一组作为标记进行亲子鉴定, 结果在171个子代中, 与SSR鉴定结果不同的个体仅5个, 一致率达到97.08%。再逐次增加第1个和第2个冗余标记后, 不一致个体减少为2个和1个, 一致的比率进一步提高到98.83%和99.42%(表 3)。

2.4　基于群体遗传推断的单体型分型结果

除了使用单个个体的测序数据进行微单体型组装的方法外, 还可以利用群体遗传学的方法进行单体型推断, 从而获取微单体型分型。为了比较这两种方法在亲子鉴定上的区别, 我们使用SHAPEIT软件获取了上述176尾个体在15个基因上的单体型分型情况。通过比较发现, 当某个基因中SNP数量过多时, 个体分型种类十分复杂, 不能用于亲子鉴定, 因此我们选择SNP数量较少的基因*msi2*尝试进行单体型分型。在5尾二倍体亲本中, 该基因的SNP总共出现了10种分型, 也就是说, 在5个亲本中出现了10个不同的等位基因类型。在正常情况下, 如果不考虑重组以及各种偏差, 任何两个亲本得到的后代都应该是这10个等位基因的组合, 可以被明确地鉴定其等位基因来自哪两个亲本。然而, 通过群体遗传学推断, 在171尾子代中共检出42种分型, 出现了34种亲本并不存在的分型, 并且有一个母本(2#)的两种分型还不包含在其中。

3　讨论

3.1　标记获取方法

目前标记的获取主要有三种策略: (1)从已经报道的多态性标记位点中获得[32, 41]; (2)选取适当数量(通常小于100尾)的个体数据用于测试标记多态性[12, 42, 43]; (3)使用全部亲本子代组成的群体进行标记筛选[25, 44]。本研究所采用的方法为第二种, 并且考虑到重组和变异的频率不可能很高, 子代的绝大部分标记均与其亲本一致, 因此我们用于获取标记的个体为全体亲本, 不包括子代个体。在鱼类育种研究中, 通常是很有限的亲本产生大量的子代。在亲子鉴定过程中, 我们只需要通过数量有限的亲本获取少量标记, 即可从大量的子代中获取出必要的信息, 这种做法具有快速和低成本的优点, 非常有利于大规模的亲子鉴定。

3.2　单体型/微单体型标记的分型方法比较

本研究比较了基于群体遗传推断的方法和基于个体序列组装的方法。前者使用的算法大多为EM算法或其他概率算法, 核心为参数优化, 从理论

表 2　各基因上的SNP位点与微单体型区域的数量以及亲子鉴定准确率

Tab. 2　SNP loci and MR in each gene and the accuracy of paternity testing

基因 Gene	长度 Length (bp)	SNP数 Number of SNPs	MR总数Total number of MRs	有效MR数Effective number of MRs	与用SSR鉴定结果的一致率 Consistency with SSR's results (%)	可鉴定子代占比Ratio of detectable offspring (%)
*adamts*20	51075	188	32	18	99.40	97.08
*brca*2	16330	167	23	12	98.14	94.15
*dlc*1	21420	46	9	4	58.08	97.66
gbp	16058	35	6	3	29.46	75.44
*lgals*9	5112	14	1	0	—	—
*lrp*5	18312	89	13	4	38.57	81.87
meis2b	16360	43	9	2	63.75	46.78
*mrps*23	4348	28	4	1	12.77	54.97
*msi*2	2859	13	2	0	—	—
nos2b	11347	48	9	6	91.61	90.64
prtga	25714	158	25	16	99.41	99.42
*rpz*4	1138	18	1	0	—	—
snx14	45146	192	38	26	97.14	61.40
*thsd*4	7511	30	5	0	—	—
*zmym*4	12827	59	7	2	65.09	98.83

注: 其中"—"表示由于标记信息不足以完成亲子鉴定

Note: "—" indicates the failure to perform the paternity testing due to insufficient information

本身而言, 需要大样本量来进行推断, 对于样本量较少而SNP位点较多的群体分型结果不准确[31]。依据孟德尔分离定律与自由组合定律, 理论上绝大多数子代的分型应与其对应的亲本相一致, 然而在本研究中该方法的分型结果与亲本的实际分型情况相差很大, 因此我们没有使用这些分型数据进行亲子鉴定。出现这样亲子分型差异较大的情况, 说明对于这种方法来说, 包含171尾子代个体的群体样本量依然太小, 导致进行概率计算时无法获得准确的估计, 从而产生了错误的分型结果。可见使用群体遗传学方法进行单体型分型对于样本数据量有很高的要求, 在应用中有明显的局限性。

相比之下, 利用测序数据直接进行单体型或微单体型组装的方法不仅更为准确地从亲本中得到

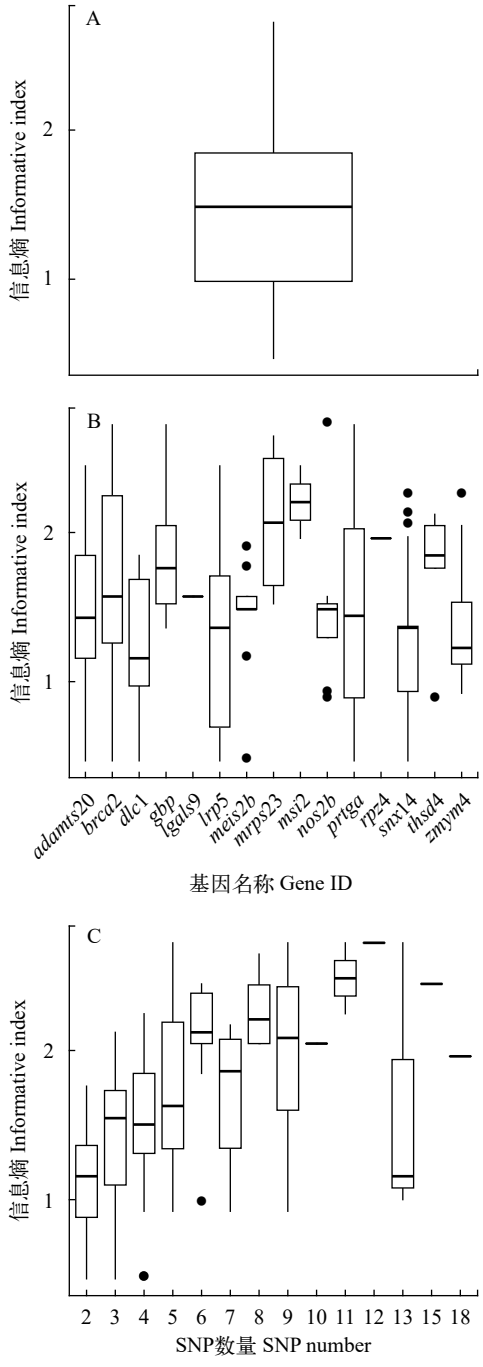

图 3 基于个体序列组装的微单体型分析结果

Fig. 3 Analysis of the microhaplotypes assembled using the sequence data of individuals

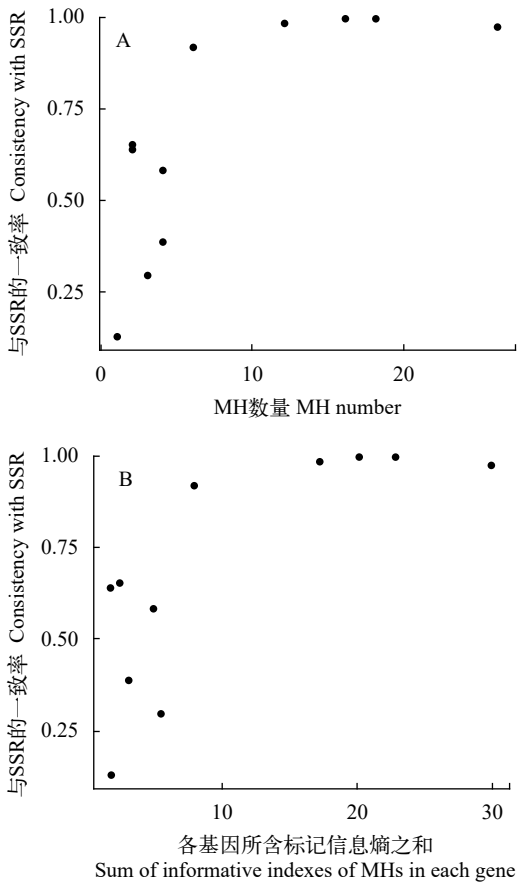

图 4 亲子鉴定一致性与MH标记数量(A)和各基因有效标记信息熵之和(B)的关系

Fig. 4 Consistency between the SSRs used in the paternity test and MH marker numbers (A) or the sum of the informative indexes of the markers within each gene (B)

表 3 所选用的MR标记与171尾子代个体的亲子鉴定

Tab. 3 MR markers adopted and paternity test results of 171 offspring

标记数量 Number of markers	标记名字 Marker ID	预测一致的子代数 Number of offspring consistent with prediction	一致率 Ratio of consistency (%)
3	prtga_14, gbp_2, brca2_1	166	97.08
4	prtga_14, gbp_2, brca2_1, adamts20_2	169	98.83
5	prtga_14, gbp_2, brca2_1, adamts20_2, snx14_5	170	99.42

了各个标记的多种分型, 也能从子代中找到这些分型, 从而顺利完成亲子鉴定过程。这种方法的主要劣势在于易受到单个位点测序深度和测序错误的影响[36]。在本研究中, 子代测序深度为15X, 在某些SNP位点上测序深度不足, 导致无法获取某些微单体型的分型。不过, 在实际应用中这也不会成为一个问题, 因为在经通过亲本的测序数据选定标记之后, 只需要在子代中扩增这些微单体型标记, 直接测序即可得到其准确的分型, 而不需要进行高成本的基因组重测序与序列组装。此外, 目前靶向捕获基因组联合二代测序技术已经得到了广泛应用[45], 同样可以用于在大量的子代群体中测序少量的标记序列。很显然, 在高通量测序技术已经普及的背景下, 基于序列组装的方法更加适合于鱼类亲子鉴定的工作。

3.3　MH在亲子鉴定中的应用

本研究比较了用MH与SSR进行亲子鉴定的结果, 并未得到完全的一致, 可能有两方面的原因。首先, 某些个体测序深度不足, 导致无法分型, 这已在上文进行讨论。其次, 各标记多态性不同, 导致亲本区分度不同。从图 2和表 2来看, 一致率低的基因往往只有极少或根本不存在SNP密集分布区域, 导致这些基因的绝大多数MH分型种类也较少, 大量亲本为纯合子, 对各亲本的区分度不足, 而目前常用的CERVUS 3.0软件倾向于选择纯合子亲本作为亲子鉴定的输出结果[46], 因此, 对于这类基因, MH亲子鉴定得到的亲本往往是纯合子, 而与真实亲本对可能有较大误差, 最终影响亲子鉴定结果的准确性。相反, 一致率高的基因大多存在若干SNP分布密集的区域且有效MR数较多, 即使单个基因近似连锁遗传, SNP密集的MH具有高多态性, 使得此基因MH标记总信息熵高, 可弥补其他MH区分度低的缺陷, 从而保留较高亲本区分度。此外, 在这些MH上亲本大多为杂合子, 因此可以极大提升亲子鉴定效果, 最终亲子鉴定一致率较高(图 4B)。最后, MH标记内部也可能存在的重组问题, 虽然这种可能性很小, 但一旦发生, 子代分型结果便无法与其真实亲本对完全对应, 从而导致亲子鉴定结果出现误差, 这种情况在SSR标记中也存在。为尽可能减少上述因素的影响, 本研究使用了高多态的分子标记及冗余标记, 其中高多态的分子标记用于提升标记区分度, 冗余标记用于矫正测序深度和重组带来的分型误差, 最终亲子鉴定一致率得到了提升(表 3)。

MH标记应用于亲子鉴定的优势在于此类标记易于获取、分型及筛选, 如果结合靶向测序就能够以极低的成本高效而准确地完成子代的分型与亲子鉴定; 其劣势则在于鉴定能力严重依赖于MH标记的多态性, 而且使用全基因组重测序数据分析时, 有些标记可能会受局部测序偏低的影响, 此外, 标记内重组也影响亲子鉴定效果。因此, 我们的方法首先根据亲本序列对目标区域的SNP密度进行评估, 筛选出高多态性的MH分子标记, 从而提升所用标记的分辨能力; 加入冗余标记则可以尽可能地减少测序深度低和标记内重组带来的影响; 如果对目标基因或片段进行靶向测序, 就能够完全排除局部测序深度低的干扰。

3.4　前景展望

虽然本研究的范例中仅从15个基因序列里筛选了亲子鉴定的标记, 但对于有参考基因组的物种, 该方法完全可以在全基因组范围内进行微单体型标记的获取与分型。通过计算这些微单体型信息熵值以筛选信息熵最高的微单体型标记, 提升标记的多态性和分辨能力, 可以用于个体鉴定、亲子鉴定, 以及更广泛的亲缘关系鉴定, 甚至还可将其应用于经济性状的关联分析中, 作为全基因组选择育种的分子标记。

参考文献:

[1] Gui J F, Zhou L, Zhang X J. Research advances and prospects for fish genetic breeding [J]. *Bulletin of Chinese Academy of Sciences*, 2018, **33**(9): 932-939. [桂建芳, 周莉, 张晓娟. 鱼类遗传育种发展现状与展望 [J]. 中国科学院院刊, 2018, **33**(9): 932-939.]

[2] Jia Z Y, Bai S S, Li C T, *et al*. Paternity Identification and genetic structure analysis of disease-resistant breeding population of german mirror carp (*Cyprinus carpio*) [J]. *Genomics and Applied Biology*, 2017, **36**(9): 3735-3741. [贾智英, 白姗姗, 李池陶, 等. 德国镜鲤抗病选育群体的亲子鉴定及遗传结构分析 [J]. 基因组学与应用生物学, 2017, **36**(9): 3735-3741.]

[3] Sobrino B, Brión M, Carracedo A. SNPs in forensic genetics: a review on SNP typing methodologies [J]. *Forensic Science International*, 2005, **154**(2-3): 181-194.

[4] Cheng W W, Yang K, Gao Y A, *et al*. Microsatellite DNA markers for parentage test of *Siniperca chuatsi* [J]. *Freshwater Fisheries*, 2016, **46**(1): 29-32, 45. [成为为, 杨凯, 高银爱, 等. 基于微卫星标记建立翘嘴鳜亲子鉴定技术 [J]. 淡水渔业, 2016, **46**(1): 29-32, 45.]

[5] Zhang D, Fu J J, Zhang L D, *et al*. The parentage analysis of bighead carp (*Hypophthalmichthys nobilis*) based on ten microsatellite markers [J]. *Genomics and Applied Biology*, 2019, **38**(7): 2949-2957. [张丹, 傅建军, 张利德, 等. 鳙基于10个微卫星标记的亲子鉴定分析 [J]. 基因组学与应用生物学, 2019, **38**(7): 2949-2957.]

[6] Tang L X, Xu Z Q, Ge J C. Application of DNA molecular markers in genetics and breeding of aquatic animals

[J]. *Journal of Aquaculture*, 2013, **34**(10): 44-48. [唐刘秀, 许志强, 葛家春. DNA分子标记技术在水产动物遗传育种中的应用 [J]. 水产养殖, 2013, **34**(10): 44-48.]

[7] Bowcock A M, Ruiz-Linares A, Tomfohrde J, *et al*. High resolution of human evolutionary trees with polymorphic microsatellites [J]. *Nature*, 1994, **368**(6470): 455-457.

[8] Gill P. An assessment of the utility of single nucleotide polymorphisms (SNPs) for forensic purposes [J]. *International Journal of Legal Medicine*, 2001, **114**(4-5): 204-210.

[9] DePristo M A, Banks E, Poplin R E, *et al*. A framework for variation discovery and genotyping using next-generation DNA sequencing data [J]. *Nature Genetics*, 2011, **43**(5): 491-498.

[10] Tang H B, Kirkness E F, Lippert C, *et al*. Profiling of short-tandem-repeat disease alleles in 12632 human whole genomes [J]. *American Journal of Human Genetics*, 2017, **101**(5): 700-715.

[11] Davey J W, Hohenlohe P A, Etter P D, *et al*. Genome-wide genetic marker discovery and genotyping using next-generation sequencing [J]. *Nature Reviews Genetics*, 2011, **12**(7): 499-510.

[12] Ren K. The parentage assignment of grass carp (*Ctenopharyngoclon idellus*) using microsatellite markers [D]. Shanghai: Shanghai Ocean University, 2013: 23-50. [任昆. 草鱼亲子亲子关系的微卫星鉴定 [D]. 上海: 上海海洋大学. 2013: 23-50.]

[13] Wen P, Zhao J, Li W, *et al*. The parentage assignment of *mauremys mutica* using multiplex PCR of microsatellites [J]. *Acta Hydrobiologica Sinica*, 2015, **39**(6): 1134-1141. [文萍, 赵建, 李伟, 等. 基于微卫星多重PCR技术的黄喉拟水龟亲子鉴定 [J]. 水生生物学报, 2015, **39**(6): 1134-1141.]

[14] Yang K, Cheng W W, Gao Y A, *et al*. Parentage analysis of F2 selective generation in mandarin fish using microsatellites [J]. *Acta Hydrobiologica Sinica*, 2015, **39**(6): 1231-1235. [杨凯, 成为为, 高银爱, 等. 翘嘴鳜F2家系选育及微卫星亲子鉴定 [J]. 水生生物学报, 2015, **39**(6): 1231-1235.]

[15] Zhu S R, Meng Q L, An L, *et al*. Microsatellite genetic analysis of gynogenetic grass carp group and two common grass carp groups [J]. *Journal of Fishery Sciences of China*, 2018, **25**(6): 1236-1244. [朱树人, 孟庆磊, 安丽, 等. 雌核发育草鱼群体及两个普通草鱼群体的微卫星遗传分析 [J]. 中国水产科学, 2018, **25**(6): 1236-1244.]

[16] Xia J H, Liu F, Zhu Z Y, *et al*. A consensus linkage map of the grass carp (*Ctenopharyngodon idella*) based on microsatellites and SNPs [J]. *BMC Genomics*, 2010(11): 135.

[17] Su J J. Correlation studies between single nucleotide polymorphisms of *TLR*7 and *TLR*8 gene and the hemorrhage disease of grass carp (*Ctenopharyngodon idella*) [D]. Xianyang: Northwest A & F University, 2017: 21-57. [苏娟娟. 草鱼*TLR*7和*TLR*8基因单核苷酸多态性与草鱼出血病的关联研究 [D]. 咸阳: 西北农林科技大学. 2017: 21-57.]

[18] Kidd K K, Pakstis A J, Speed W C, *et al*. Microhaplotype loci are a powerful new type of forensic marker [J]. *Forensic Science International: Genetics Supplement Series*, 2013, **4**(1): e123-e124.

[19] Wang H, Zhu J, Zhou N, *et al*. NGS technology makes microhaplotype a potential forensic marker [J]. *Forensic Science International: Genetics Supplement Series*, 2015(5): e233-e234.

[20] Pu Y, Chen P, Zhu J, *et al*. Microhaplotype: Ability of personal identification and being ancestry informative marker [J]. *Forensic Science International: Genetics Supplement Series*, 2017(6): e442-e444.

[21] Kidd K K, Speed W C, Pakstis A J, *et al*. Evaluating 130 microhaplotypes across a global set of 83 populations [J]. *Forensic Science International*: *Genetics*, 2017(29): 29-37.

[22] Hiroaki N, Koji F, Tetsushi K, *et al*. Approaches for identifying multiple-SNP haplotype blocks for use in human identification [J]. *Legal Medicine*, 2015, **17**(5): 415-420.

[23] Chen P, Zhu W J, Tong F, *et al*. Identifying novel microhaplotypes for ancestry inference [J]. *International Journal of Legal Medicine*, 2018, **133**(4): 983-988.

[24] Oldoni F, Hart R, Long K, *et al*. Microhaplotypes for ancestry prediction [J]. *Forensic Science International: Genetics Supplement Series*, 2017(6): e513-e515.

[25] Garciafernandez C, Sanchez J A, Blanco G. SNP-haplotypes: An accurate approach for parentage and relatedness inference in gilthead sea bream (*Sparus aurata*) [J]. *Aquaculture*, 2018(495): 582-591.

[26] Baetscher D S, Clemento A J, Ng T C, *et al*. Microhaplotypes provide increased power from short-read DNA sequences for relationship inference [J]. *Molecular Ecology Resources*, 2018, **18**(2): 296-305.

[27] Barrett J C, Fry B, Maller J, *et al*. Haploview: analysis and visualization of LD and haplotype maps [J]. *Bioinformatics*, 2005, **21**(2): 263-265.

[28] Stephens M, Donnelly P. A comparison of bayesian methods for haplotype reconstruction from population genotype data [J]. *American Journal of Human Genetics*, 2003, **73**(5): 1162-1169.

[29] Delaneau O, Marchini J, Zagury J F. A linear complexity phasing method for thousands of genomes [J]. *Nature Methods*, 2012, **9**(2): 179-181.

[30] Browning S R, Browning B L. Rapid and accurate haplotype phasing and missing data inference for whole genome association studies by use of localized haplotype clustering [J]. *American Journal of Human Genetics*, 2007, **81**(5): 1084-1097.

[31] Browning S R, Browning B L. Haplotype phasing: existing methods and new developments [J]. *Nature Reviews Genetics*, 2011, **12**(10): 703-714.

[32] Xin M M, Zhang S H, Wang D Q, *et al*. Parentage identification of polyploidy *Acipenser sinensis* based on microsatellite markers [J]. *Freshwater Fisheries*, 2015, **45**(4): 3-9. [辛苗苗, 张书环, 汪登强, 等. 多倍体中华鲟

微卫星亲子鉴定体系的建立 [J]. 淡水渔业, 2015, **45**(4): 3-9.]

[33] Edge P, Bafna V, Bansal V. HapCUT2: robust and accurate haplotype assembly for diverse sequencing technologies [J]. *Genome Research*, 2017, **27**(5): 801-812.

[34] Duitama J, Huebsch T, Mcewen G, *et al*. ReFHap: A reliable and fast algorithm for single individual haplotyping [C]//Zhang A D, Borodovsky M, *et al*. (Eds.), BCB '10 Proceedings of the First ACM International Conference on Bioinformatics and Computational Biology, 2 August - 4 August 2010, New York, USA: Association for Computing Machinery, 2010: 160-169.

[35] Zhu J, Zhou N, Jiang Y J, *et al*. FLfinder: A novel software for the microhaplotype marker [J]. *Forensic Science International*: *Genetics Supplement Series*, 2015(5): e622-e624.

[36] Rhee J, Li H L, Joung J, *et al*. Survey of computational haplotype determination methods for single individual [J]. *Genes & Genomics*, 2016, **38**(1): 1-12.

[37] Kalinowski S T, Taper M L, Marshall T C. Revising how the computer program CERVUS accommodates genotyping error increases success in paternity assignment [J]. *Molecular Ecology*, 2007, **16**(5): 1099-1106.

[38] Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform [J]. *Bioinformatics*, 2010, **26**(5): 589-595.

[39] Li H, Handsaker B, Wysoker A, *et al*. The sequence alignment/map format and SAMtools [J]. *Bioinformatics*, 2009, **25**(16): 2078-2079.

[40] McKenna A, Hanna M, Banks E, *et al*. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data [J]. *Genome Research*, 2010, **20**(9): 1297-1303.

[41] Zhang X C, Zhao J, Li W, *et al*. Multiple paternity in the cultured yellow pond turtles (*Mauremys mutica*) [J]. *Animal Reproduction Science*, 2017(183): 46-55.

[42] Liu Y, Chen Y Y, Gong Q, *et al*. Paternity assignment in the polyploid Acipenser dabryanus based on a novel microsatellite marker system [J]. *PLoS One*, 2017, **12**(9): e0185280.

[43] Wang Z. Parentage determination and genetic linkage map construction in *Sinonovacula constricta* [D]. Shanghai: Shanghai Ocean University, 2016: 20-32. [王泽. 缢蛏亲子鉴定技术及遗传连锁图谱的构建 [D]. 上海: 上海海洋大学. 2016: 20-32.]

[44] Chen L, Wang D Q, He Y F, *et al*. Parentage assignment of *Rhinogobio ventralis* using multiplex PCR of microsatellites [J]. *Journal of Agricultural Biotechnology*, 2017, **25**(9): 1526-1537. [陈亮, 汪登强, 何勇凤, 等. 基于微卫星多重PCR技术的长鳍吻鮈亲子鉴定 [J]. 农业生物技术学报, 2017, **25**(9): 1526-1537.]

[45] Yu Q X. Accurate detection for hemoglobin variants of thalassemia based on next generation sequencing target capture technology [D]. Shenzhen: Southern University of Science and Technology, 2016: 44-70. [喻秋霞. 基于二代测序靶向捕获技术准确检测地中海贫血珠蛋白基因变异 [D]. 深圳: 南方医科大学, 2016: 44-70.]

[46] Jones A G, Small C M, Paczolt K A *et al*. A practical guide to methods of parentage analysis [J]. *Molecular Ecology Resources*, 2010, **10**(1): 6-30.

A METHOD FOR PATERNITY TESTING OF GRASS CARP (*CTENOPHARYNGODON IDELLUS*) USING MICROHAPLOTYPES

XIA Lei[1, 2], SHI Mi-Juan[1, 3], ZHANG Wan-Ting[1, 3], DUAN You[1, 2], CHENG Ying-Yin[1, 3], WU Nan[1, 3] and XIA Xiao-Qin[1, 2, 3]

(1. *Institute of Hydrobiology*, *Chinese Academy of Sciences*, *Wuhan* 430072, *China*; 2. *University of Chinese Academy of Sciences*, *Beijing* 100049, *China*; 3. *The Innovative Academy of Seed Design*, *Chinese Academy of Sciences*, *Beijing* 100101, *China*)

Abstract: Paternity testing is a technique of great importance in the genetic breeding of aquatic animals. Currently, the most frequently used type of biomarker in paternity tests is microsatellites (SSRs). However, weaknesses of SSRs lie in the complicated and labor-intensive genotyping process, which leads to low efficiency when such analyses are performed on a large scale. In this study, a new type of molecular biomarker, microhaplotypes (MH), was introduced for paternity testing. For the purpose of marker screening and paternity testing, a more efficient pipeline was constructed and evaluated with data from a grass carp population. The results showed that the genotypes of the MHs can be accurately obtained from genome resequencing data with clearly improved efficiency and compatibility over conventional genotyping methods based on population genetics. It is feasible to screen highly efficient MH combinations using the informative index. The consistency with the paternity test results obtained using SSRs reached 97.08% or 99.42% when 3 or 5 MHs were used, respectively. This research suggests that MHs can be used for the rapid and accurate paternity testing of fishes.

Key words: Microhaplotype; Paternity test; Grass carp; High-throughput sequencing