遗传 Hereditas (Beijing) 2021 年 10 月, 43(10): 962－971
www.chinagene.cn

研究报告

微单倍型遗传标记的法医基因组学研究

李茜，王浩宇，曹悦岩，朱强，舒潘寅，侯婷芸，王雨婷，张霁

四川大学华西基础医学与法医学院，成都 610041

摘要： 微单倍型(microhaplotype, MH)是在一定 DNA 片段范围之内，由至少两个单核苷酸多态性位点组成的遗传标记。MH 兼具无 stutter 伪峰、多态性丰富以及扩增子较小等特点，有望成为法医学上的一种新型遗传标记。为了从全基因组维度上分析 MH 的特征，进一步发掘其应用潜能，本研究基于千人基因组计划中 105 个中国南方汉族个体的全基因组测序数据，构建了迄今为止最全面的 MH 数据集。结果表明，人类基因组中 350 bp 范围之内的 MH 位点数量共计 9,490,075 个，且微单倍型分布密度对染色体变异水平具有提示作用。从多种碱基跨度范围对 MH 的多态性分析表明，其多态性潜能可达到或者超过常用短串联重复序列位点的水平。此外，本文归纳总结了 MH 组装灵活等特点，并提出了构建微单倍型数据库的方案。

关键词： 法医遗传学；微单倍型；千人基因组计划；中国南方汉族群体

Forensic genomics research on microhaplotypes

Xi Li, Haoyu Wang, Yueyan Cao, Qiang Zhu, Panyin Shu, Tingyun Hou, Yuting Wang, Ji Zhang

West China School of Basic Medical Sciences & Forensic Medicine, Sichuan University, Chengdu 610041, China

Abstract: Microhaplotype loci (microhaplotype, MHs), defined by two or more closely linked single nucleotide polymorphisms, are a type of molecular marker within a short segment of DNA. As emerging forensic genetic markers, MHs have no stutter artefacts and higher polymorphism, and permit the design of smaller amplicons. In order to identify the markers from a genome wide perspective and explore their potential application further, we constructed the most comprehensive MH dataset to date, based on the whole genome sequencing data of 105 Han individuals in Southern China from 1000 Genomes Project. The results showed that there were 9,490,075 MH loci in the range of 350 bp in the human genome, and the distribution density of microhaplotypes suggests gene variation. Polymorphism analysis of MHs from various base spans showed that the polymorphism of MHs could reach or exceed common short tandem repeat sites. In addition, based on their flexible assembly, a scheme to build the public database of microhaplotypes was proposed.

收稿日期: 2021−05−26; **修回日期:** 2021−07−29

基金项目: 国家自然科学基金项目(编号：81571861, 81630054)资助[Supported by the National Natural Science Foundation of China (Nos. 81571861, 81630054)]

作者简介: 李茜，在读硕士研究生，专业方向：法医遗传学。E-mail: lixi1105@foxmail.com
王浩宇，在读硕士研究生，专业方向：法医遗传学。E-mail: wanghy0707@gmail.com
李茜和王浩宇并列第一作者。

通讯作者: 张霁，博士，教授，研究方向：法医遗传学。E-mail: zhangj@scu.edu.cn

DOI: 10.16288/j.yczz.21-186
网络出版时间: 2021/8/10 14:02:00
URI: https://kns.cnki.net/kcms/detail/11.1913.R.20210810.1125.001.html

Keywords: forensic genetics; microhaplotypes; 1000 Genomes; Southern Han Chinese

近些年，微单倍型(microhaplotype，MH)逐渐受到法医学领域研究人员的关注。MH 由 Kidd 实验室(美国，耶鲁大学医学院)在 2013 年首先提出[1]，是一种在几百个核苷酸以内，由两个或多个紧密连锁的单核苷酸多态性(single nucleotide polymorphism，SNP)位点组合而成的多等位基因分子标记。MH 与其他遗传标记相比具有以下特点：(1)扩增子没有 stutter 峰。微单倍型没有短串联重复结构，不会出现 stutter 峰所带来的诸如增加不平衡混合样本分析的复杂性等干扰问题[2,3]。(2)多态性通常高于 SNP。MH 具有多个等位基因，经过筛选的微单倍型拥有比 SNP 位点更高的杂合度[4]。(3)MH 为序列多态性，其检测依赖于对碱基序列的读取。二代测序技术可以在几百个碱基的 DNA 单链上进行连续测序，直接对紧密排列的 SNP 位点进行“定相”(phase)，获得真实的单倍型[5]。MH 的可检测片段长度随着测序技术的发展一直在增加，由最初定义的 200 bp 逐渐扩大到 300~500 bp[1,6,7]。而其片段长度下限，根据报道可低至 20 bp 或 70 bp[8,9]。MH 的这些特点，使其有望成为短串联重复序列(short tandem repeat, STR)位点基因分型的补充方法。

目前已有多个应用于法医 DNA 分析的 MH 体系。de la Puente 等[10]开发了包含 118 个 MH 的复合体系，由于位点的平均长度仅 51 个核苷酸，对降解的 DNA 表现出高度的敏感性。MH 的等位基因频率在不同大陆群体之间表现出差异，联合使用 118 个 MH 可以提供比常用 STR 体系更低的随机匹配概率。Oldoni 等[11]报道的 74-MH 体系在混合 DNA 分析方面表现出优势，MH 在二代测序平台检测到的等位基因覆盖度(allele coverage, AC)可以一定程度反映混合斑比例，更利于对次要贡献者的等位基因进行拆分。Wu 等[12]认为具有较高有效等位基因数(the effective number of alleles, *Ae*)的 MH 有利于在混合斑中检测到更多的等位基因，从而减少贡献者之间的等位基因共享，帮助判断贡献者个数。一些研究人员测试了 MH 体系对法医亲缘关系鉴定的适用性[13~15]。结果表明，联合使用 30~60 个 MH 在亲子鉴定和全同胞鉴别方面可优于现有 STR 或 SNP 体系，但涉及二级或更远的亲缘关系判断仍然需要添加更多的位点。上述研究均强调了开发足够数量的、多等位基因、高多态性 MH 的重要性。

根据统计，目前约有 470 个微单倍型被报道[10,12~20]，其中多数位点的 *Ae* 值在 2.0~4.0，*Ae* 达到 4.0 以上的位点有 120 个。SNP 遗传标记在人类基因组中是广泛存在的，相应的，由多个 SNP 参与定义的微单倍型的数量也是极为丰富的。相对于 MH 在全基因组中的广泛分布，目前已开发报道的微单倍型仅是其中很小的一部分。想要进一步了解 MH 的数目和属性，更好地满足个人识别、混合 DNA 分析以及亲缘关系鉴定等法医学应用的需求，需要更全面的 MH 位点信息作为支持。据此，我们从特定群体入手，以期在全基因组维度上对 MH 的特征进行分析与归纳。本研究使用的是千人基因组计划第三阶段中国南方汉族群体的遗传数据。考虑到法医学领域不同应用目的下的扩增子长度、常用测序平台的阅读长度以及位点侧翼需预留引物设计空间等因素，我们对 350 bp 范围内的微单倍型进行全面筛查，并统计了多种片段长度限制下 MH 的多态性，进一步认识和发掘这种新兴法医学遗传标记的应用潜能。

1　材料与方法

1.1　SNP 预过滤

本研究使用的全基因组测序数据下载自千人基因组计划第三阶段(GRCh37.p13)的数据库网站[21]。涉及的 105 个样本均属于中国南方汉族(Southern Han Chinese, CHS)。首先使用 VCFtools 工具对这些样本的变异检测格式(variant call format, VCF)文件进行预过滤，获取可用于后续组装微单倍型的 SNP 集合。预过滤的标准如下：(1)染色体定位在 1~22 号常染色体；(2)排除插入/缺失(insertion or deletion, InDel)变异，即在统计 MH 的分型和参数时不将 InDel 纳入考虑；(3) SNP 位点在相应群体中的最小等位基因频率(minor allele frequency, *MAF*)大于 0.01；(4)对 SNP 位点进行 Hardy-Weinberg 平衡检验，需满

足 $P>0.05$。

1.2 微单倍型的组装和过滤

本研究对于构建 MH 的要求是：获取全基因组范围内所有长度在 350 bp 以内、至少包含 2 个 SNP 的潜在微单倍型。

通过 1.1 部分的预过滤，可以在 22 条常染色体上分别获得 SNP 物理位置依次递增的预筛选集合。首先，以某一条染色体上第一个 SNP (即物理位置最小的 SNP)作为潜在 MH 的“起始 SNP”，依次纳入后续相邻的位点。然后，判断当前组合是否为满足要求的潜在微单倍型。每纳入一个 SNP，则需判断一次：如果满足要求，则将其输出；如不满足要求，则将“起始 SNP”的坐标依次向后移动，循环上述过程。当“起始 SNP”的坐标移动至该染色体预筛选集合的最后一个位点时，该染色体的检索结束。最后，对所有常染色体进行检索，并对输出的微单倍型进行编号。

该组装过程可能将规定片段长度范围内的 SNP 进行多次组合并重复输出。对于目标片段长度范围内存在的 n 个 SNP，至多可输出 $\sum_{i=1}^{n-1} i$ 种微单倍型，包含 SNP 数目最多的 MH 被称为“最长片段 MH”，其余位点被描述为“子集”。例如，在 350 个碱基跨度内存在 5 个 SNP，则至多可输出 10 种 MH，其中由全部 5 个 SNP 定义的微单倍型即是“最长片段 MH”，其余由 2~4 个连续的 SNP 定义的 MH 合称为“子集”。在之后的一些分析中，为了减少冗余数据，可能会将子集移除，得到由各目标区域内“最长片段 MH”组成的“最长片段集”。“最长片段集”与“子集”合称为“完整集”。需要注意的是，如果研究者关注的片段长度范围发生改变，那么相对应的“完整集”、“最长片段集”以及“子集”中所包含的 MH 都将发生变化。以上过程均由实验室内部基于 Python 的脚本实现。

1.3 统计学分析

上述 MH 的组装和过滤过程，仅对 SNP 在参考基因组中的物理位置(position)进行输出。之后，对于所有输出的微单倍型，按其构成提取 VCF 文件中相应 SNP 的基因分型并组装成 MH 等位基因，然后计算每个位点在中国南方汉族中的群体遗传学参数，包括杂合度观测值(observed heterozygosity, *Ho*)、个体识别概率(discrimination power, *DP*)以及有效等位基因数(*Ae*)。有效等位基因数(*Ae*)是一个经典的群体遗传学概念，它的值代表遗传标记所等价的频率相等的等位基因的个数。例如，某遗传标记的 *Ae* 值为 n，则表示该遗传标记等价于包含 n 个频率相等的等位基因，即每个等位基因的频率均为 $1/n$。通多该指标可以实现对多等位基因遗传标记的比较和排序。*Ae* 值的计算公式为：$1/\sum pi^2$，其中 pi 表示某基因座上等位基因 i 的频率[22]。

2 结果与分析

2.1 人类基因组中 MH 位点的数量

对千人基因组计划数据进行初步筛选之后，在 22 条人类常染色体上共得到 5,977,655 个 SNP 位点。按照 1.2 所述策略进行无差别组装，获取 350 bp 范围之内所有可能的 MH (“完整集”)共计 9,490,075 个。过滤子集之后，仍保留 30.47%的位点(2,891,927 个)，其中 2 号染色体的 MH 最多，22 号染色体的 MH 最少，分别为 235,330 和 40,808 (表 1)。平均每百万个碱基对(Mb)检索到大约 1000 (2,891,927/3000 Mb)个微单倍型。

图 1 以密度图的形式展示了每条染色体上 MH “最长片段集”的分布情况。一些分布特征与人类已知的变异模式相匹配：例如，在 6 号染色体主要组织相容性复合体(the major histocompatibility complex, MHC)周围观察到了极大数量的 MH；在 8q21.2 周期性新着丝粒(neocentromere)的附近[23]，也发现 MH 高密度分布区。此外，16 号染色体短臂或长臂近端粒处(16q23)的“亮黄色”区域可能提示 MH 数量高于平均水平。其余 MH 的分布相对均匀。

2.2 350 bp 范围内 MH 的统计学参数

如前所述，微单倍型标记的组装过程会将一定

表 1　SNP 及 MH 在不同染色体上的数量统计

Table 1　The number of SNPs and MHs on different chromosomes

染色体	#SNPs [a]	#MHs≤350 bp		#MHs≤150 bp		#MHs≤100 bp		#MHs≤50 bp	
		A	B	A	B	A	B	A	B
1	463,261	684,624	224,137	307,923	160,070	211,609	127,357	112,562	80,220
2	485,172	697,551	235,330	312,548	167,213	213,273	132,583	113,234	82,973
3	429,260	648,976	208,260	289,892	149,992	197,993	119,687	104,404	75,230
4	444,134	719,321	214,966	322,127	158,179	220,652	127,643	116,157	81,220
5	369,559	536,677	179,135	240,372	127,576	164,475	101,232	87,636	63,535
6	406,810	855,491	197,195	380,915	145,792	258,670	118,720	134,514	77,942
7	357,021	560,129	172,793	252,329	125,618	172,901	100,917	91,761	64,047
8	323,908	538,902	157,309	239,578	116,180	163,404	93,911	85,902	60,124
9	264,750	408,354	128,051	183,076	94,273	125,067	75,556	65,695	47,502
10	310,578	493,463	150,444	221,083	109,703	151,560	88,309	80,883	56,820
11	290,938	445,661	140,840	199,579	102,589	136,071	81,970	71,787	52,025
12	287,513	430,602	139,153	194,229	100,173	133,241	79,814	70,791	50,312
13	217,352	335,132	105,264	150,042	76,429	102,775	61,026	54,319	38,801
14	194,482	293,076	94,194	131,568	67,706	90,024	53,957	47,739	34,160
15	176,222	274,166	84,933	123,892	61,570	85,058	49,433	45,461	31,845
16	187,593	331,113	90,597	148,132	68,613	101,035	56,023	53,349	36,743
17	155,592	237,431	74,611	108,076	54,056	74,898	43,371	40,510	27,684
18	172,512	267,593	83,271	121,279	60,482	83,014	48,506	43,974	30,786
19	142,771	254,075	67,813	117,069	51,542	81,348	42,076	44,002	27,689
20	127,126	186,986	61,427	84,197	44,252	57,717	35,181	30,649	22,301
21	86,049	140,790	41,396	63,550	31,033	43,662	25,070	22,955	16,084
22	85,052	149,962	40,808	68,111	30,586	47,028	24,982	25,065	16,427
总计	5,977,655	9,490,075	2,891,927	4,259,567	2,103,627	2,915,475	1,687,324	1,543,349	1,074,470

[a]本研究在组装微单倍型过程中使用的 SNP 总数；A：当前碱基长度范围内，所有可能的 MH 数量，即“完整集”；B：当前碱基长度范围内，去除子集后潜在 MH 的数量，即“最长片段集”。

范围内的 SNP 进行重复组合和输出。为了减少冗余数据，此部分的分析只针对 350 bp 范围内、移除子集的 MH 集合(“最长片段集”)。

2.2.1　总体特征

用于定义微单倍型的 SNP 数量在 2~51 之间，其中由两个 SNP 构成的标记数量最多，占比 45.42%。观察到至少 3 个等位基因的遗传标记共计 2,494,157 个，约占 86.25%；等位基因数超过 10 的位点多达 14,133 个。有 50%的微单倍型长度范围集超过 263 bp，所有位点的平均长度是 239 bp。

根据千人基因组计划数据库中发布的“确定相位”(phased)的基因分型数据，估计微单倍型的等位基因频率信息。总的来说，微单倍型在中国南方汉族群体中，具有非常可观的遗传多态性。*Ho* 值超过 0.8 的 MH 共计 11,712 个；*DP* 值超过 0.9 的 MH 多达 21,355 个。之前一项研究提出了 *Ae* 值的阈值(*Ae* = 3)[22]，超过这一阈值的微单体型被认为具有较高的法医学应用价值。本部分共涉及 2,891,927 个微单倍型，*Ae* 值在 1.02~66.62 之间。*Ae* 值高于 3 的标记共计 199,176 个，高于 5 的标记共计 6935 个；387 个 MH 的 *Ae* 值在 10~20 之间(不包括 10)，41 个 MH 的 *Ae* 值大于 20。表 2 给出了 *Ae* 值位于前 10 的微单倍型位点信息，参与构成这些 MH 的 SNP 互不重

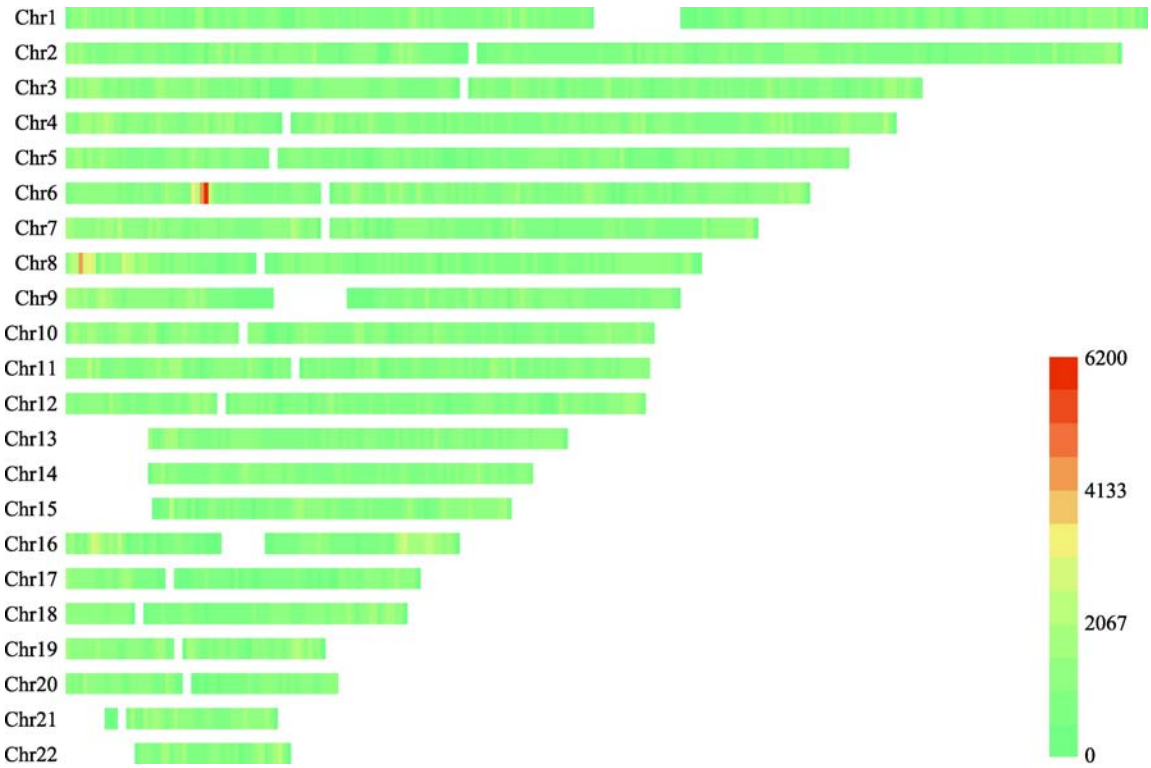


图 1　人类基因组中微单倍型遗传标记的密度分布图

Fig. 1　Density plots of microhaplotypes identified in the human genome

使用 350 bp 范围内、移除子集的 MH 数据绘制。色阶代表每 Mb 染色体序列的 MH 总数。性染色体数据未收集。

表 2　*Ae* 值前 10 的微单倍型位点信息

Table 2　The information of the top 10 *Ae* values of microhaplotypes

MH_ID	bp	*Ae*	*Ho*	*DP*	#SNPs	#Alleles	Position (GRCh37)
mh04zj0146583	346	66.6163	0.9905	0.9899	41	134	Chr4:30279658~30280003
mh01zj0675568	248	43.2353	0.9714	0.9905	15	88	Chr1:247032193~247032440
mh20zj0185187	347	32.6183	0.9999	0.9892	9	68	Chr20:62308266~62308612
mh09zj0366544	239	28.5622	0.9524	0.9883	12	60	Chr9:129479455~129479693
mh07zj0103025	323	28.3055	0.9714	0.9892	26	77	Chr7:18772264~18772586
mh04zj0352614	346	27.1218	0.9810	0.9859	22	105	Chr4:88537078~88537423
mh03zj0068937	350	24.2308	0.8762	0.9858	20	75	Chr3:11955851~11956200
mh02zj0082461	320	23.7864	0.9810	0.9874	12	56	Chr2:20701112~20701431
mh01zj0508420	337	21.7028	0.9619	0.9872	37	63	Chr1:200785797~200786133
mh04zj0474307	348	20.5307	0.9714	0.9870	11	51	Chr4:129682428~129682775

参与构成 MH 的 SNP 互不重复，Chr6 MHC 周围的 MH 没有纳入。

复，且 MHC 周围的位点没有纳入。

2.2.2　特征参数之间的关系

为了探究微单倍型遗传标记 *Ae* 值、*Ho* 值、*DP* 值、bp、构成 MH 的 SNP 数以及等位基因数之间的关系，研究者分别对每条染色体上的 MH 绘制这六个特征参数的散点图矩阵。以位点数量居中的 9 号染色体为例展示了 MH 特征参数之间的相关性(图 2，其余染色体的散点图矩阵见附图 1~21)。对角线处分别为各参数的核密度估计图，其余位置为任意两参

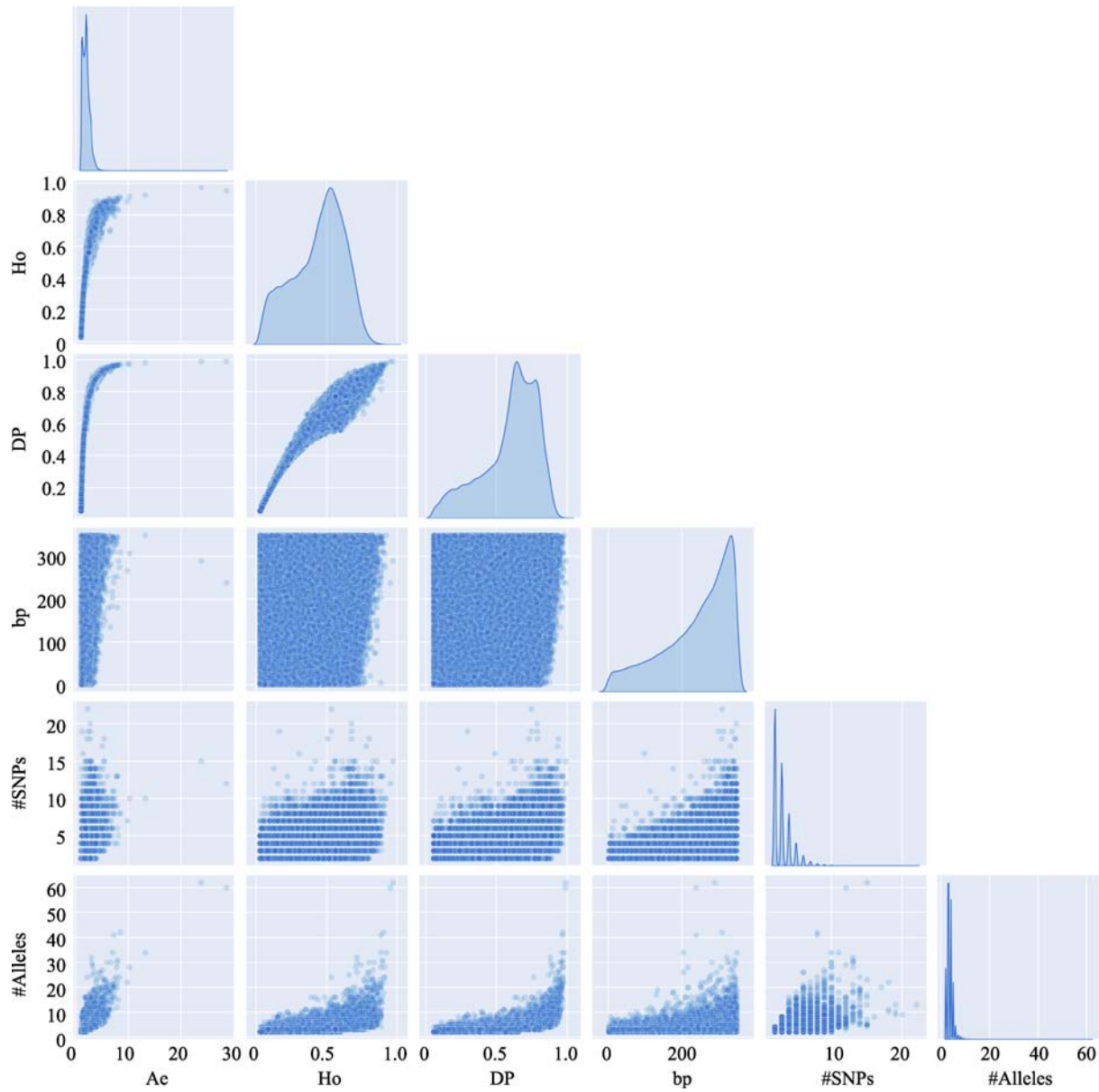


图 2　微单倍型遗传标记特征参数之间的关系

Fig. 2　Relationship among characteristic parameters of microhaplotypes

使用位于 9 号染色体、350 bp 范围内、移除子集的 MH 数据绘制(共计 128,051 个)。#SNPs：构成 MH 的 SNP 数；#Alleles：等位基因数。

数之间的散点图。核密度估计是一种从数据样本本身出发研究数据分布特征的方法，曲线下方的面积和等于 1；当存在多个波峰时，所有波峰下方的面积之和为 1。某区间所对应的曲线下面积越大，代表样本在该区间分布的概率越大。散点图直观的反映了这六个特征参数之间的关系。首先，*Ae* 值、*DP* 值、*Ho* 值三者之间具有较强的相关关系。其次，随着等位基因数的增加，*Ae* 值的最低值逐渐升高，二者存在一定的相关性。其余参数之间的相关程度均较差。

综合 22 条常染色体的 MH 数据，计算这些参数之间的成对 Pearson 相关系数(r)并绘制热图(图 3)。*DP* 值和 *Ho* 值的相关系数最高(r=0.97)；*Ae* 值与 *DP* 值和 *Ho* 值的相关系数分别为 0.85 和 0.88；等位基因数与 *Ae* 值和构成 MH 的 SNP 数呈中等程度相关；其余参数之间的相关系数均小于等于 0.4。

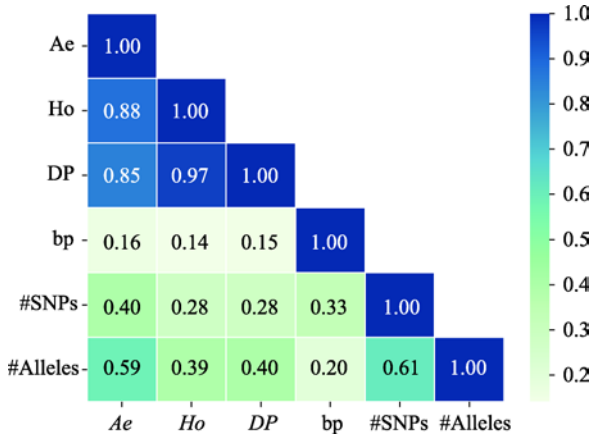


图 3　微单倍型遗传标记特征参数之间的成对相关系数

Fig. 3　Pairwise correlation coefficient between characteristic parameters of microhaplotypes

使用人类基因组 350 bp 范围内、移除子集的 MH 数据绘制(共计 2,891,927 个)。#SNPs：构成 MH 的 SNP 数；#Alleles：等位基因数。

2.3　不同片段长度范围内的 MH 的数量及 *Ae* 值分布情况

如前所述，350 bp 范围之内所有可能的 MH (即“完整集”)共计 9,490,075 个；过滤子集之后，仍保留 2,891,927 个位点(即“最长片段集”，占比 30.47%)。当将片段长度的上限分别设置为 150 bp、100 bp 和 50 bp 时，相对应的“完整集”中 MH 的数量分别为 4,259,567、2,915,475 和 1,543,349 (表 1)；移除子集之后潜在位点的数量分别减少了 50.61%、42.13%和 30.38% (图 4A)。目标区域的碱基跨度越大，可能纳入的 SNP 数目就会越多，从而产生更多的组合形式，“子集”占比也随之增高。

本研究对不同片段长度范围内的“最长片段集”微单倍型的 *Ae* 值分布情况进行了统计(表 3)。在加强碱基长度的限制之后，具有高多态性的微单倍型

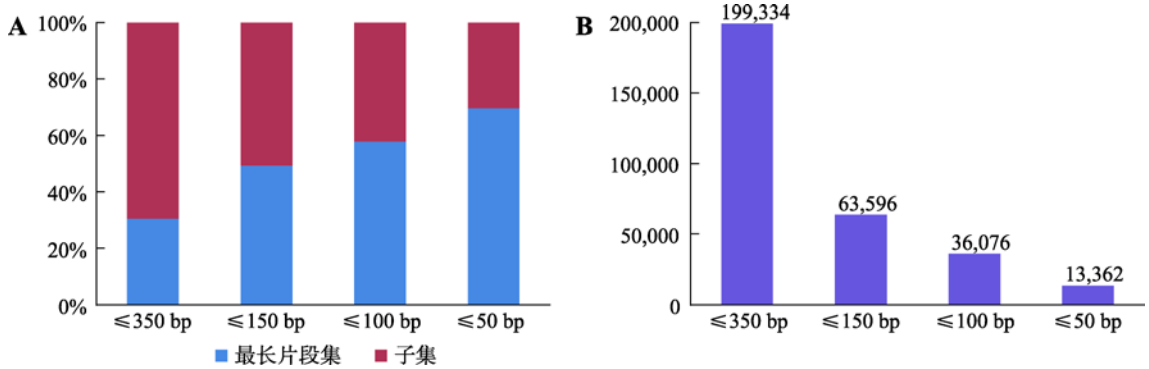


图 4　不同片段长度范围内的微单倍型遗传标记

Fig. 4　Microhaplotypes within different lengths of base pairs

A：350 bp、150 bp、100 bp、50 bp 内 MH“子集”与“最长片段集”的百分比堆积柱形图；B：不同片段长度范围内 *Ae* 值大于等于 3 的 MH 数量统计，使用“最长片段集”的 MH 数据绘制。

表 3　不同片段长度范围内微单倍型 *Ae* 值的分布

Table 3　The distribution of microhaplotypes within different lengths of base pairs

Ae	#MHs≤350 bp	#MHs≤150 bp	#MHs≤100 bp	#MHs≤50 bp
1~	1,634,742	1,345,361	1,116,786	741,819
2~	1,057,851	694,670	534,462	319,289
3~	171,637	56,715	32,419	12,211
4~	20,759	4941	2683	937
5~	6510	1875	957	213
10~	307	52	11	0
15~	80	8	5	1
20~	26	0	1	0
25~	15	5	0	0

使用相应片段长度范围内移除子集的 MH 数据绘制，即“最长片段集”。

仍然十分丰富：在 150 bp 和 100 bp 范围内，*Ae* 值大于等于 3.0 的 MH 数量分别是 199,334 和 63,596；长度降低至 50 个碱基之内时，仍有 13,362 个位点的 *Ae* 值大于等于 3.0 (图 4B)。

3　讨论

本研究使用千人基因组计划中国南方汉族群体的基因分型数据，构建了 350 bp 范围内的微单倍型标记库，展示了迄今为止最全面的人类 MH 集合，并对 MH 的特征和应用潜能有了更深刻的认识。

第一，微单倍型在人类基因组中的数量极为丰富。为了尽可能不高估 MH 的数量，本研究仅从“最长片段集”水平考虑，在 22 条常染色上共检索到 2,891,927 个位点。法医遗传学学者所熟知的 STR 基因座在人类基因组中的分布密度约 100 个/Mb[24]，相较而言微单倍型遗传标记的数量更为丰富，平均每 Mb 碱基序列检索到 1000 个 MH 位点(2,891,927/3000 Mb)。

从微单倍型密度分布图(图 1)可以观察 MH 在基因组测序数据缺失序列(gap)之外的分布情况。MH 的高密度分布区与人类基因组中一些已知的高变异区域相匹配，说明 MH 的分布密度可以一定程度体现人类基因组的变异水平。MH 的高密度分布本质上来源于 SNP 的高密度分布，这提示了微单倍型多态性来源于历史性基因突变的可能性，而 MH 多态性水平与基因重组的关系则需要在家系中进一步探究。我们建议在解决亲缘关系鉴定的问题时，对于 MH 位点的选择和使用需要慎重考虑。

第二，MH 多态性不仅优于 SNP，而且可达到甚至超过常用的 STR 基因座。MH 拥有比 SNP 位点更高的杂合度，这一观点基本被法医遗传学家所公认。其与 STR 基因座之间的比较，Oldoni 等[11,25]认为后者更具优势。本研究虽然没有考虑引物设计、位点序列与基因组对齐(BLAST)结果等因素对最终能够用于构建实验体系的 MH 位点数量的影响，但从理论上对 MH 的多态性潜能做出了评估。基于 105 个 CHS 样本的数据统计，*Ho* 值超过 0.8、*DP* 值超过 0.9 的 MH 数量分别为 11,712 和 21,355；*Ae* 达到 4.0 的位点数量也由已报到的 120 个[10,12~20]，增加至 27,697 个。更有 14,133 个 MH 的等位基因数超过 10，870 个 MH 的等位基因数超过 50，这完全超出了研究人员对于 MH 以往的印象。因此我们认为，通过筛选可以得到等位基因数和多态性都优于 STR 的微单倍型，而这样的 MH 有望在 DNA 混合物的分析中，特别是在混合斑的确认以及贡献者数量的推断方面发挥巨大优势。

第三，MH 的 *Ae* 值与 *DP* 值和 *Ho* 值之间均具有较强的线性相关关系。三者分别由不同的参数计算得到(等位基因频率、表型频率、杂合子频率)，其中 *Ae* 值与 *Ho* 值是表征遗传标记本身多态性的指标，而 *DP* 值是评价遗传标记识别不同个体效能大小的指标，三者无法直接由公式推导而进行转换。作者通过对数百万个 MH 位点的 *Ae* 值、*DP* 值和 *Ho* 值进行成对相关性分析，观察到 *Ae* 值与 *DP* 值、*Ho* 值之间具有较强的相关性(r 分别为 0.85、0.88)。这再次印证了当筛选 MH 应用于法医学领域时，以 *Ae* 值(而不计算 *DP* 值、*Ho* 值)作为主要筛选标准具有一定的合理性。此外，*Ae* 值与位点的等位基因数之间存在一定的相关性(r=0.59)，提示一些研究以等位基因数作为 MH 筛选标准具有理论依据。*Ae* 值与片段长度、构成 MH 的 SNP 数之间的相关系数不超过 0.4。这表明，虽然随着片段长度范围的增加、可纳入 SNP 数量的增多可能会丰富微单倍型位点的基因多样性，但提升效果非常有限。在评价 MH 效能之时，不能仅以片段长度或构成 MH 的 SNP 数作为标准。

第四，MH 包含大量的“子集”，这使 MH 的组装既灵活又复杂。如在一段目标碱基序列上存在 n 个 SNP，至多可组装 $\sum_{i=1}^{n-1} i$ 种微单倍型，其中 $\sum_{i=1}^{n-1} i-1$ 种均属于“子集”。根据本课题的统计结果，MH 的片段跨度越广，包含的子集数量就越多。当片段长度的上限由 50 bp 增加至 350 bp 时，相应子集占比从 30.38%增加至 69.53%。以上情况设定了 SNP 位点在特定群体中 *MAF*，如若 *MAF* 或目标群体发生变化，靶序列可输出的 MH“子集”将会变得更加复杂。也正是因为“子集”的存在，使得 MH 的拼装具有极大的灵活性。理论上，任何定义微单倍型的 SNP 只要被检测到，就可为后续个人识别或亲

缘鉴定等法医学分析提供有价值的遗传信息，即使是不完整的 MH 位点(即 MH 的“子集”)也可被充分利用，这与传统的 STR 遗传标记是截然不同的。MH 受靶片段完整性的限制更小，将这一特性与单引物延伸技术相结合，可以为降解 DNA 样本的分析提供新思路。

与此同时，由于组装“灵活性”而产生的大量子集也给 MH 数据库构建以及遗传标记频率信息共享带来挑战。随着 MH 的研究与应用越来越广泛，各科研团队由于研究目的不同，采用的位点组装标准(例如群体、*MAF*、片段长度等)也会有所差异。那么同一段靶序列可能会记录多种 MH，或者多个 MH 中包含有相同的 SNP。这会导致数据记录缺乏兼容性，不利于数据库的整合与共享。因此我们提议，除了将 MH 作为整体进行一系列信息的记录和储存之外，参与定义 MH 的 SNP 基因分型，尤其是“确定相位”(phased)的基因分型结果也应被记录在数据库之中。这样的数据储存方式，具有良好的“向后兼容性”，可以使任何公开发表的 MH 信息与之后的研究人员充分共享。

综上所述，本研究提供了一套详尽的微单倍型组装方案，证明了 MH 在人类全基因组中数量丰富，同时在不同的碱基范围尺度上揭示了 MH 多态性水平。对 MH 的特征进行了更全面的展示，并结合其特点提出构建微单倍型数据库的方案，为未来群体遗传学和法医遗传学的研究与应用提供支持。

附录：

附加材料详见文章电子版 www.chinagene.cn。

参考文献(References)：

[1] Kidd KK, Pakstis AJ, Speed WC, Lagace R, Chang J, Wootton S, Ihuegbu N. Microhaplotype loci are a powerful new type of forensic marker. *Forensic Sci Int Genet Suppl Ser*, 2013, 4(1): e123–e124. [DOI]

[2] Oldoni F, Podini D. Forensic molecular biomarkers for mixture analysis. *Forensic Sci Int Genet*, 2019, 41: 107–119. [DOI]

[3] Bennett L, Oldoni F, Long K, Cisana S, Madella K, Wootton S, Chang J, Hasegawa R, LagacéR, Kidd KK, Podini D. Mixture deconvolution by massively parallel sequencing of microhaplotypes. *Int J Legal Med*, 2019, 133(3): 719–729. [DOI]

[4] Cheung EYY, Phillips C, Eduardoff M, Lareu MV, Mcnevin D. Performance of ancestry-informative SNP and microhaplotype markers. *Forensic Sci Int Genet*, 2019, 43: 102141. [DOI]

[5] Børsting C, Morling N. Next generation sequencing and its applications in forensic genetics. *Forensic Sci Int Genet*, 2015, 18: 78–89. [DOI]

[6] Turchi C, Melchionda F, Pesaresi M, Tagliabracci A. Evaluation of a microhaplotypes panel for forensic genetics using massive parallel sequencing technology. *Forensic Sci Int Genet*, 2019, 41: 120–127. [DOI]

[7] Jin XY, Cui W, Chen C, Guo YX, Zhang XR, Xing GH, Lan JW, Zhu BF. Developing and population analysis of a new multiplex panel of 18 microhaplotypes and compound markers using next generation sequencing and its application in the Shaanxi Han population. *Electrophoresis*, 2020, 41(13–14): 1230–1237. [DOI]

[8] Cao YY, Wang QY, Zhu Q, Huang YG, Hu YH, Zhou YJ, Wang YF, Zhang J. Preliminary exploration of a novel method for the deconvolution of DNA mixtures by pyrosequencing. *Forensic Sci Int Genet Suppl Ser*, 2019, 7(1): 843–845. [DOI]

[9] van der Gaag KJ, de Leeuw RH, Laros J, den Dunnen JT, de Knijff P. Short hypervariable microhaplotypes: A novel set of very short high discriminating power loci without stutter artefacts. *Forensic Sci Int Genet*, 2018, 35: 169–175. [DOI]

[10] de la Puente M, Phillips C, Xavier C, Amigo J, Carracedo A, Parson W, Lareu MV. Building a custom large-scale panel of novel microhaplotypes for forensic identification using MiSeq and Ion S5 massively parallel sequencing systems. *Forensic Sci Int Genet*, 2020, 45: 102213. [DOI]

[11] Oldoni F, Bader D, Fantinato C, Wootton SC, Lagace R, Kidd KK, Podini D. A sequence-based 74plex microhaplotype assay for analysis of forensic DNA mixtures. *Forensic Sci Int Genet*, 2020, 49: 102367. [DOI]

[12] Wu RG, Li HX, Li R, Peng D, Wang NN, Shen XF, Sun HY. Identification and sequencing of 59 highly polymorphic microhaplotypes for analysis of DNA mixtures. *Int J Legal Med*, 2021, 135(4): 1137–1149. [DOI]

[13] Qu N, Lin SB, Gao Y, Liang H, Zhao H, Ou XL. A microhap panel for kinship analysis through massively parallel sequencing technology. *Electrophoresis*, 2020, 41(3–4): 246–253. [DOI]

[14] Sun SL, Liu Y, Li JN, Yang ZD, Wen D, Liang WB, Yan

YQ, Yu H, Cai JF, Zha L. Development and application of a nonbinary SNP-based microhaplotype panel for paternity testing involving close relatives. *Forensic Sci Int Genet*, 2020, 46: 102255. [DOI]

[15] Wen D, Sun SL, Liu Y, Li JN, Yang ZD, Kureshi A, Fu Y, Li HN, Jiang BW, Jin C, Cai JF, Zha L. Considering the flanking region variants of nonbinary SNP and phenotype-informative SNP to constitute 30 microhaplotype loci for increasing the discriminative ability of forensic applications. *Electrophoresis*, 2021, 42(9–10): 1115–1126. [DOI]

[16] Chen P, Deng CW, Li Z, Pu Y, Yang JW, Yu YF, Li K, Li D, Liang WB, Zhang L, Chen F. A microhaplotypes panel for massively parallel sequencing analysis of DNA mixtures. *Forensic Sci Int Genet*, 2019, 40: 140–149. [DOI]

[17] Voskoboinik L, Motro U, Darvasi A. Facilitating complex DNA mixture interpretation by sequencing highly polymorphic haplotypes. *Forensic Sci Int Genet*, 2018, 35: 136–140. [DOI]

[18] Kidd KK, Speed WC, Pakstis AJ, Podini DS, Lagace R, Chang J, Wootton S, Haigh E, Soundararajan U. Evaluating 130 microhaplotypes across a global set of 83 populations. *Forensic Sci Int Genet*, 2017, 29: 29–37. [DOI]

[19] Chen P, Yin CY, Li Z, Pu Y, Yu YJ, Zhao P, Chen DX, Liang WB, Zhang L, Chen F. Evaluation of the microhaplotypes panel for DNA mixture analyses. *Forensic Sci Int Genet*, 2018, 35: 149–155. [DOI]

[20] Kureshi A, Li J, Wen D, Sun SL, Yang ZD, Zha L. Construction and forensic application of 20 highly polymorphic microhaplotypes. *R Soc Open Sci*, 2020, 7(5): 191937. [DOI]

[21] 1000 Genomes Project Consortium; Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, Mccarthy S, Mcvean GA, Abecasis GR. A global reference for human genetic variation. *Nature*, 2015, 526(7571): 68–74. [DOI]

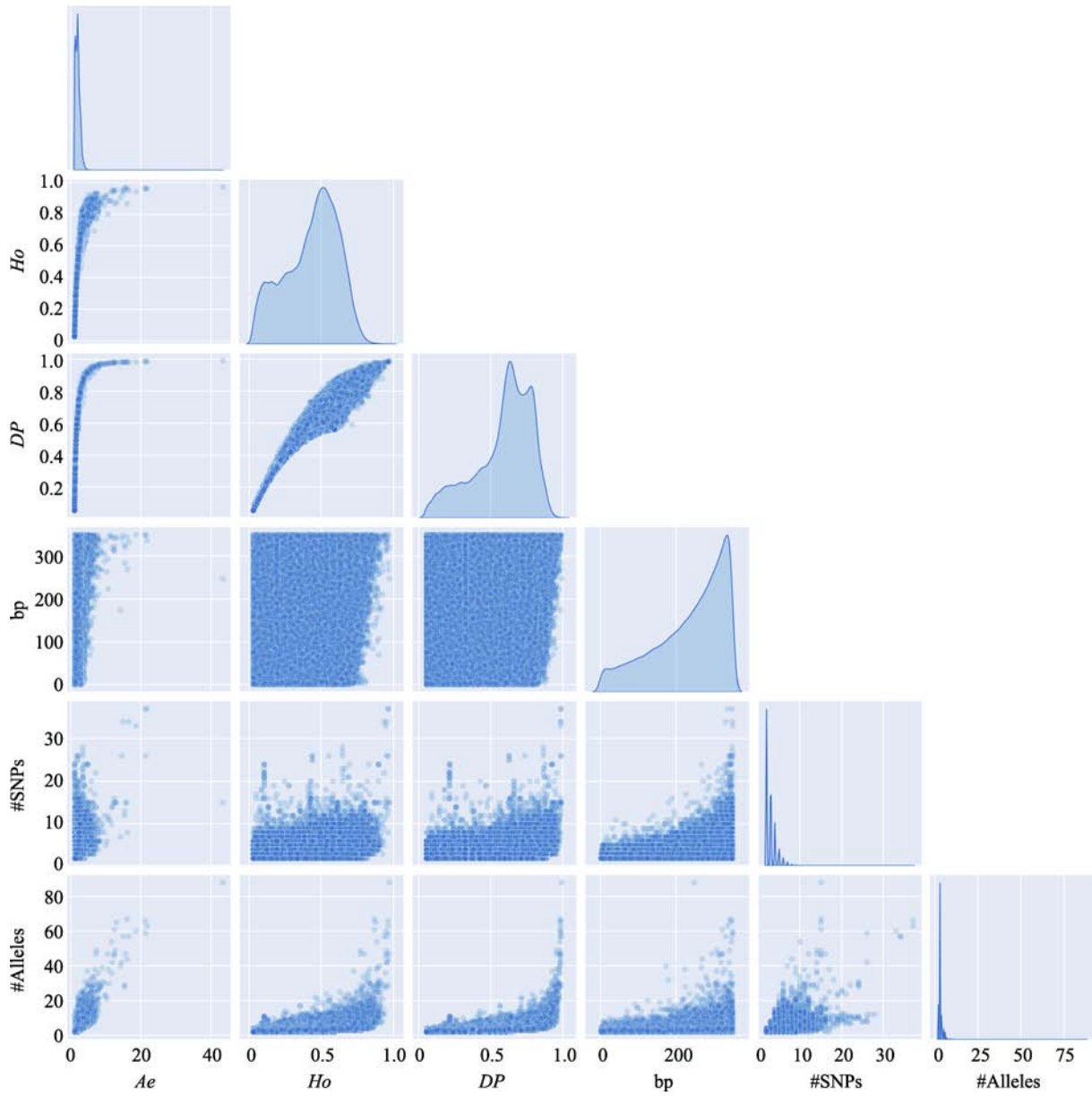
[22] Kidd KK, Speed WC. Criteria for selecting microhaplotypes: mixture detection and deconvolution. *Investig Genet*, 2015, 6(1): 1. [DOI]

[23] Logsdon GA, Vollger MR, Hsieh P, Mao YF, Liskovykh MA, Koren S, Nurk S, Mercuri L, Dishuck PC, Rhie A, de Lima LG, Dvorkina T, Porubsky D, Harvey WT, Mikheenko A, Bzikadze AV, Kremitzki M, Graves-Lindsay TA, Jain C, Hoekzema K, Murali SC, Munson KM, Baker C, Sorensen M, Lewis AM, Surti U, Gerton JL, Larionov V, Ventura M, Miga KH, Phillippy AM, Eichler EE. The structure, function and evolution of a complete human chromosome 8. *Nature*, 2021, 593(7857): 101–107. [DOI]

[24] Collins JR, Stephens RM, Gold B, Long B, Dean M, Burt SK. An exhaustive DNA micro-satellite map of the human genome using high performance computing. *Genomics*, 2003, 82(1): 10–19. [DOI]

[25] Oldoni F, Kidd KK, Podini D. Microhaplotypes in forensic genetics. *Forensic Sci Int Genet*, 2019, 38: 54–69. [DOI]

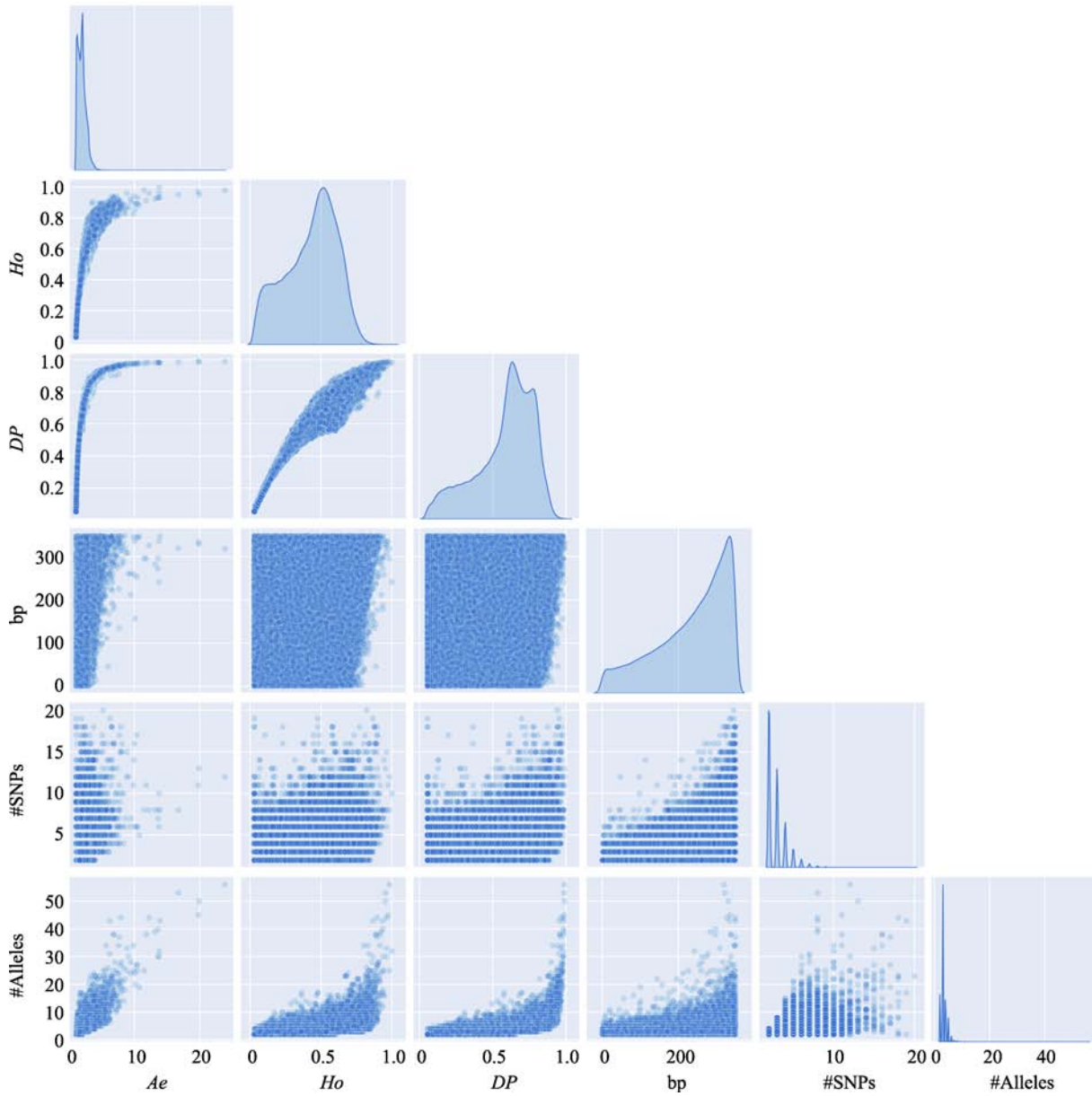
(责任编委: 朱波峰)



附图 1　1 号染色体上微单倍型遗传标记特征参数之间的关系

Supplementary Fig. 1　Relationship among characteristic parameters of microhaplotypes on chromosome 1

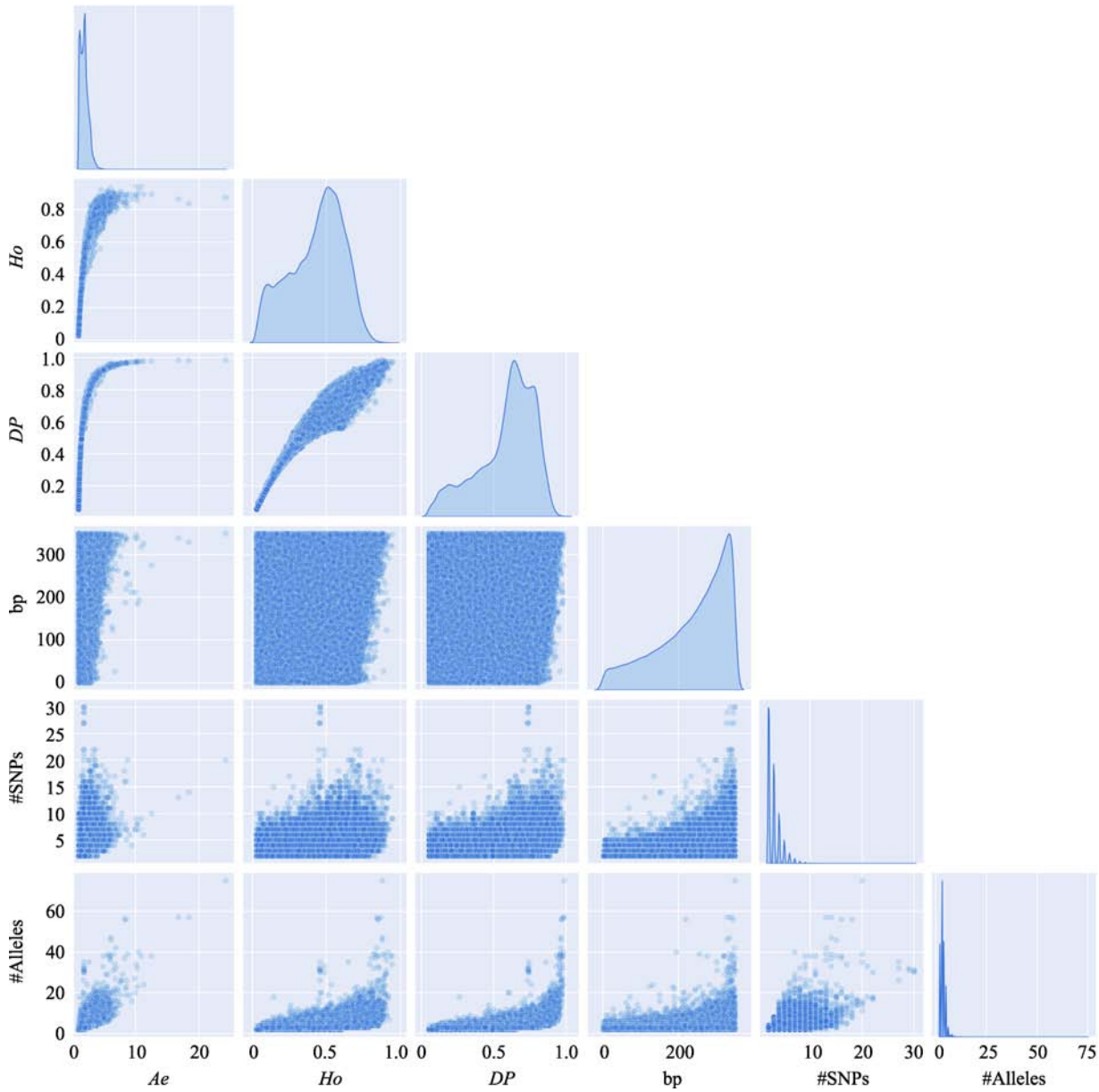
使用 1 号染色体、350 bp 范围内、移除子集的 MH 数据绘制。#SNPs：构成 MH 的 SNP 数；#Alleles：等位基因数。



附图 2　2 号染色体上微单倍型遗传标记特征参数之间的关系

Supplementary Fig. 2　Relationship among characteristic parameters of microhaplotypes on chromosome 2

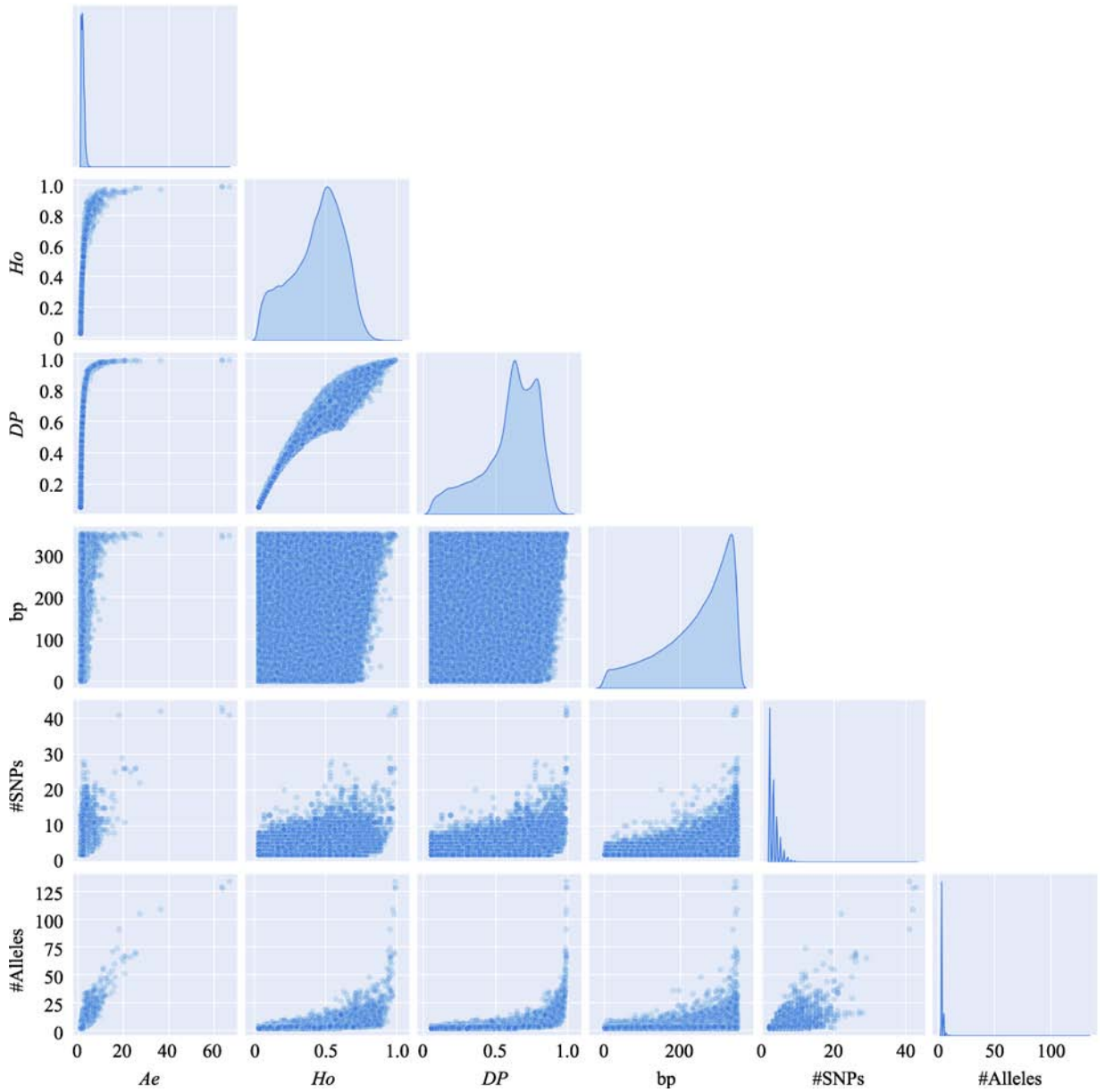
使用 2 号染色体、350 bp 范围内、移除子集的 MH 数据绘制。#SNPs：构成 MH 的 SNP 数；#Alleles：等位基因数。



附图 3　3 号染色体上微单倍型遗传标记特征参数之间的关系

Supplementary Fig. 3　Relationship among characteristic parameters of microhaplotypes on chromosome 3

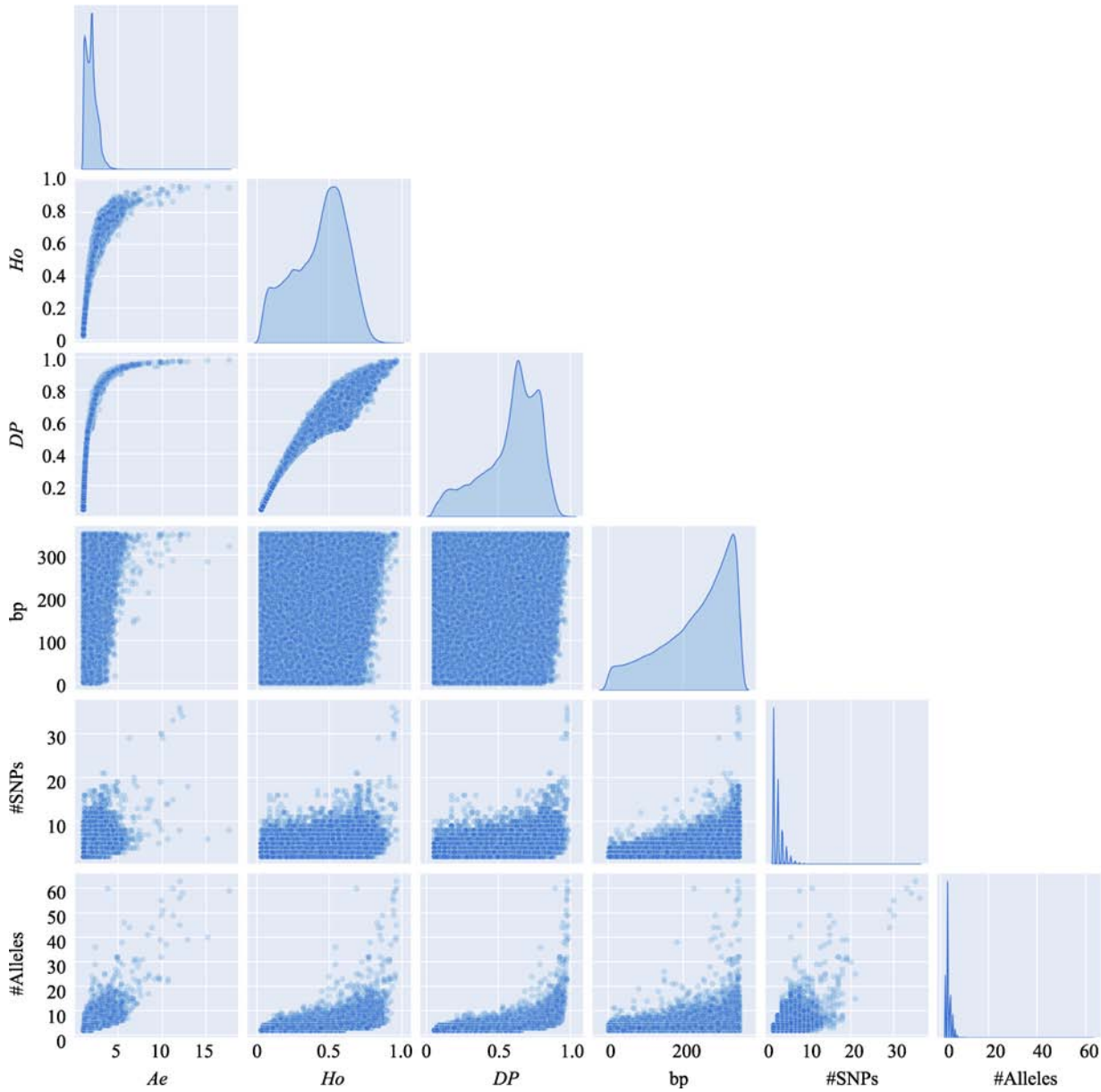
使用 3 号染色体、350 bp 范围内、移除子集的 MH 数据绘制。#SNPs：构成 MH 的 SNP 数；#Alleles：等位基因数。



附图 4　4 号染色体上微单倍型遗传标记特征参数之间的关系

Supplementary Fig. 4　Relationship among characteristic parameters of microhaplotypes on chromosome 4

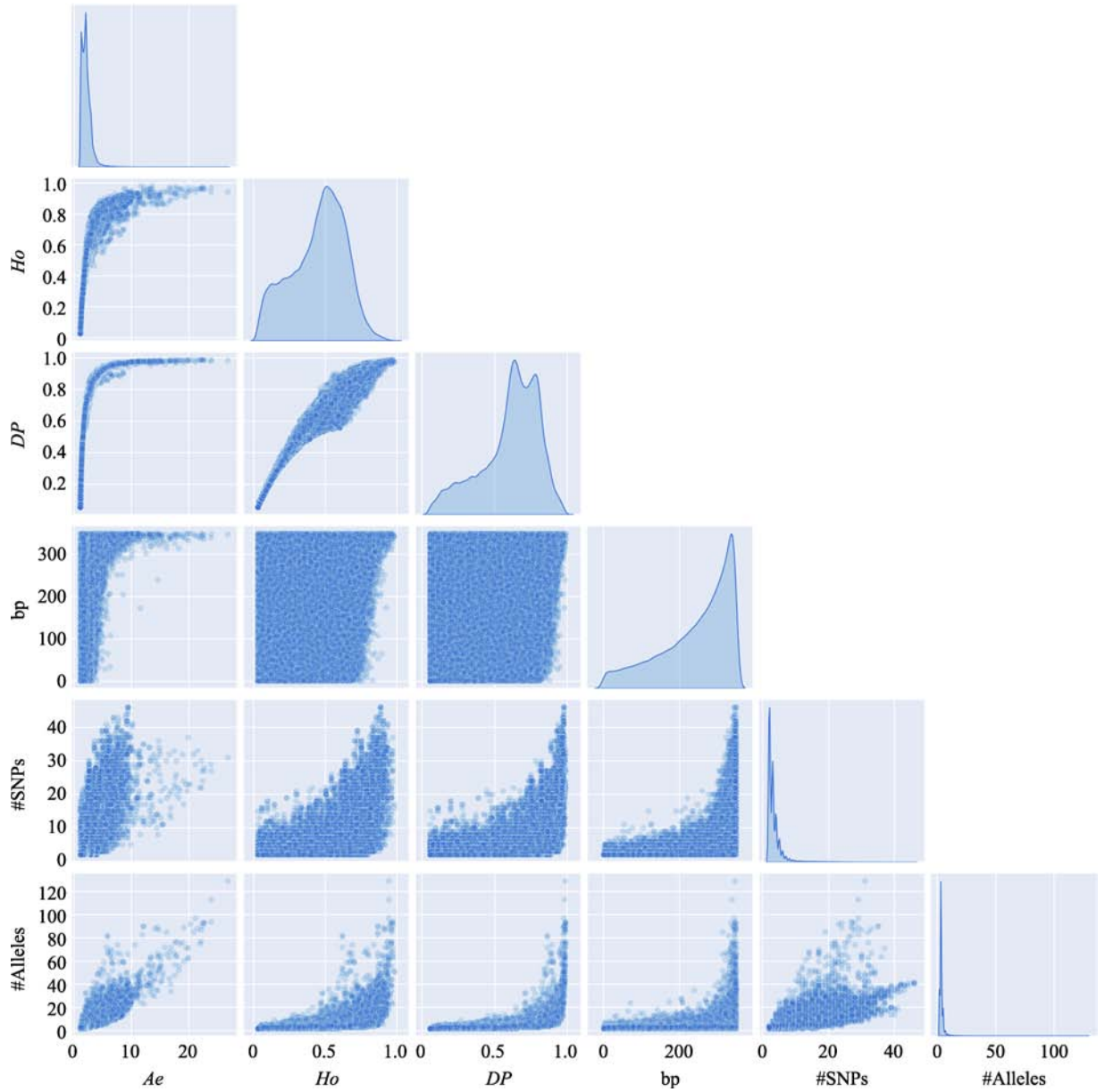
使用 4 号染色体、350 bp 范围内、移除子集的 MH 数据绘制。#SNPs：构成 MH 的 SNP 数；#Alleles：等位基因数。



附图 5　5 号染色体上微单倍型遗传标记特征参数之间的关系

Supplementary Fig. 5　Relationship among characteristic parameters of microhaplotypes on chromosome 5

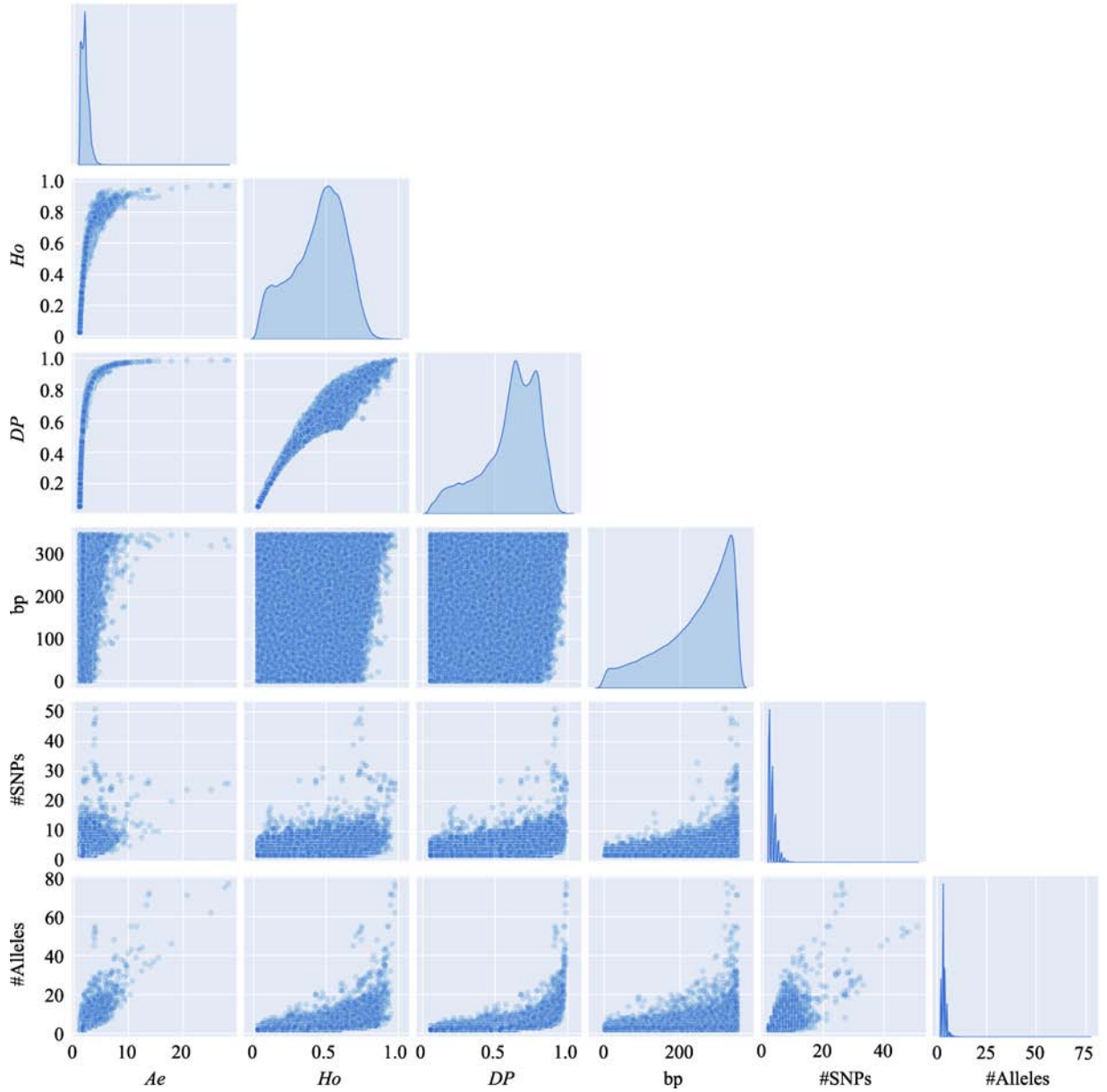
使用 5 号染色体、350 bp 范围内、移除子集的 MH 数据绘制。#SNPs：构成 MH 的 SNP 数；#Alleles：等位基因数。



附图 6　6 号染色体上微单倍型遗传标记特征参数之间的关系

Supplementary Fig. 6　Relationship among characteristic parameters of microhaplotypes on chromosome 6

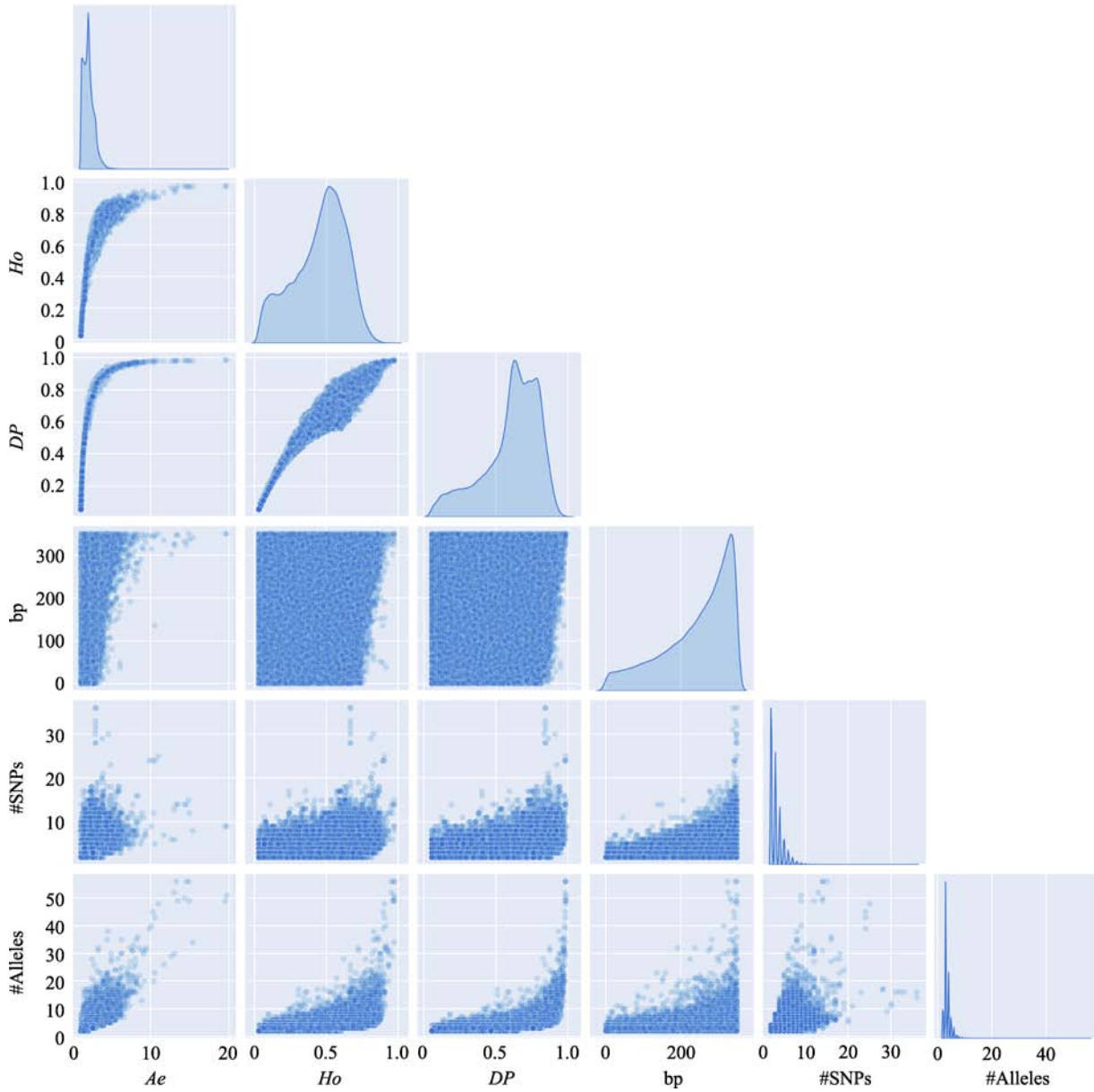
使用 6 号染色体、350 bp 范围内、移除子集的 MH 数据绘制。#SNPs：构成 MH 的 SNP 数；#Alleles：等位基因数。



附图 7　7 号染色体上微单倍型遗传标记特征参数之间的关系

Supplementary Fig. 7　Relationship among characteristic parameters of microhaplotypes on chromosome 7

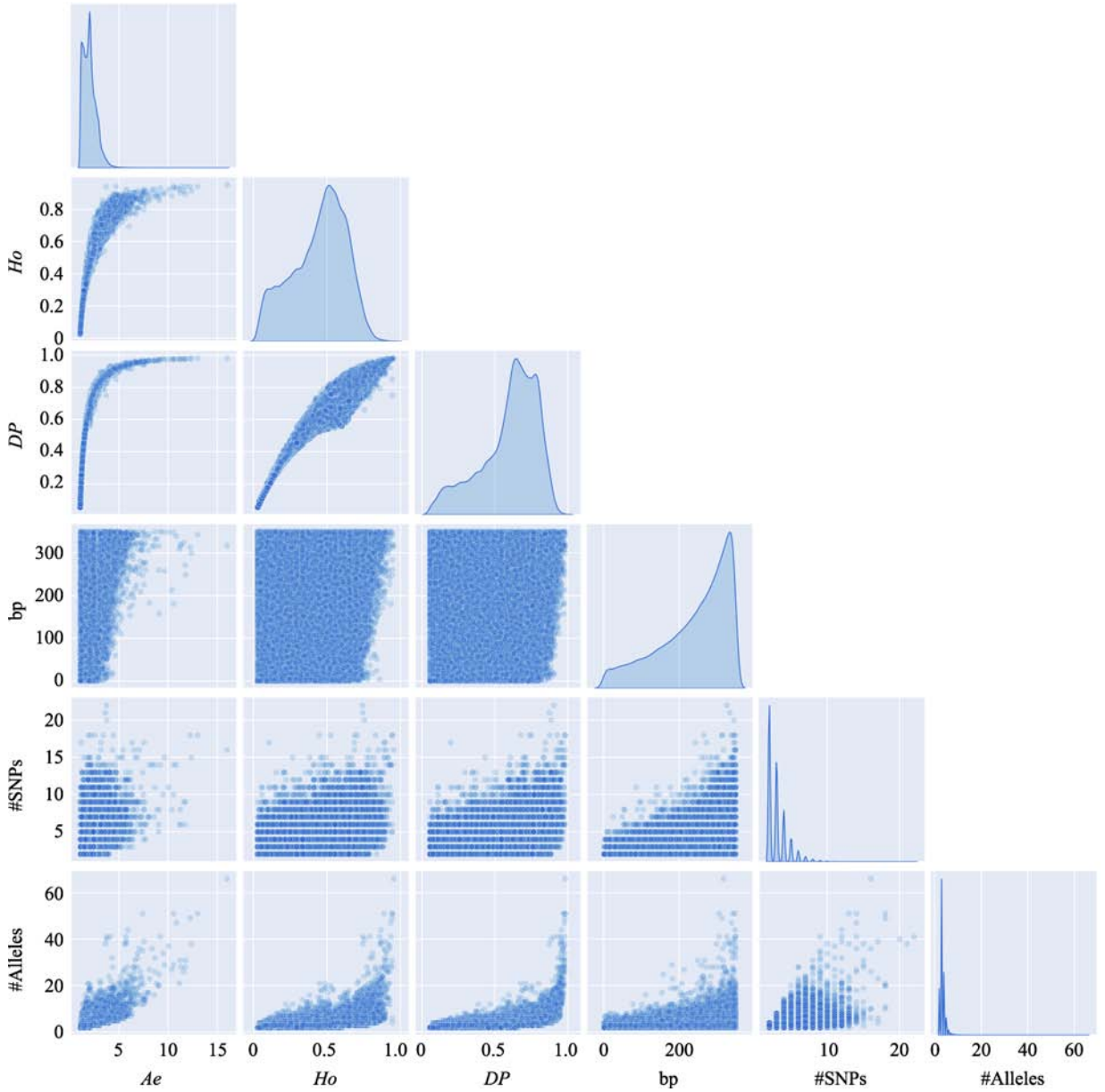
使用 7 号染色体、350 bp 范围内、移除子集的 MH 数据绘制。#SNPs：构成 MH 的 SNP 数；#Alleles：等位基因数。



附图 8　8 号染色体上微单倍型遗传标记特征参数之间的关系

Supplementary Fig. 8　Relationship among characteristic parameters of microhaplotypes on chromosome 8

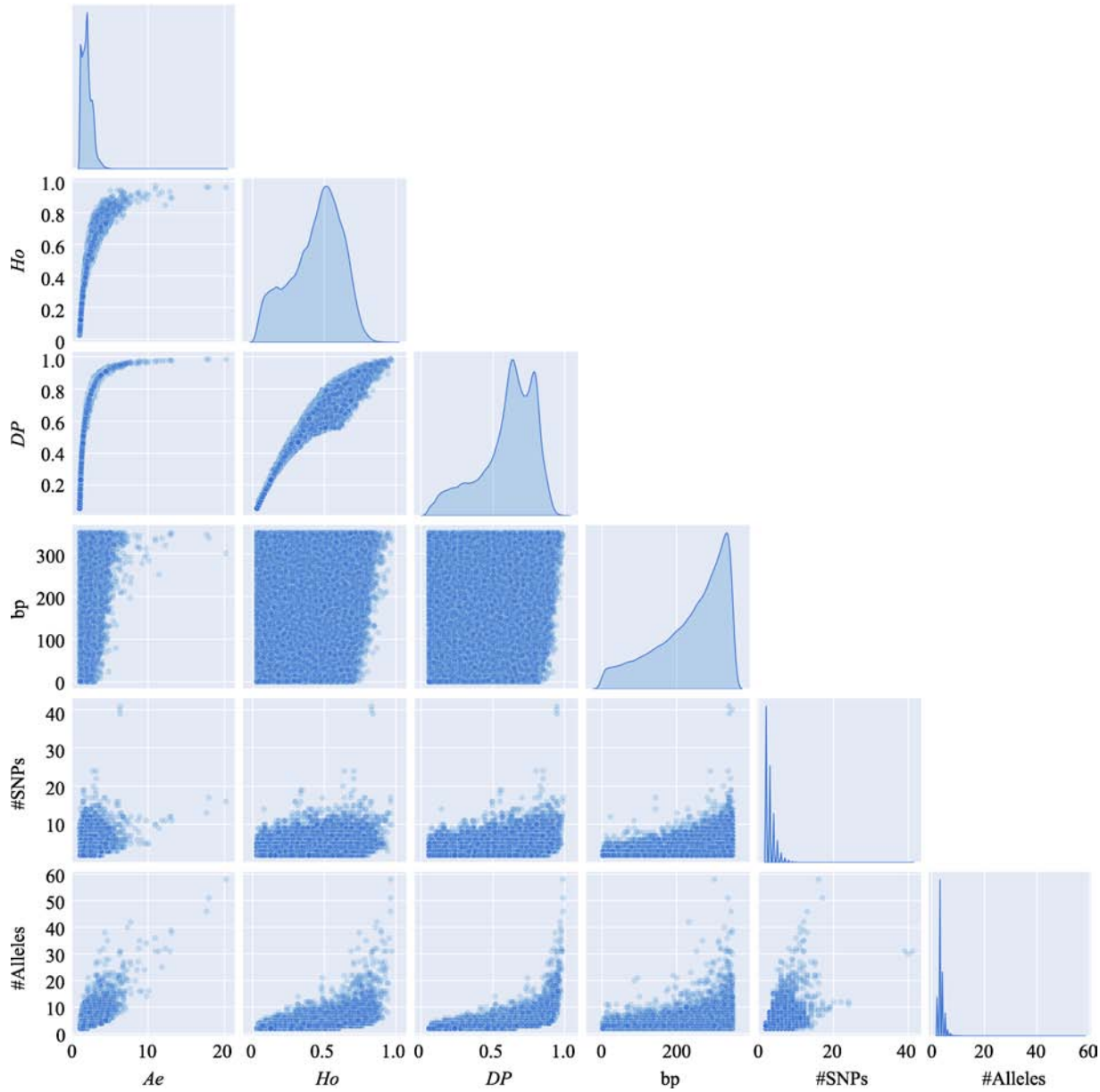
使用 8 号染色体、350 bp 范围内、移除子集的 MH 数据绘制。#SNPs：构成 MH 的 SNP 数；#Alleles：等位基因数。



附图 9　10 号染色体上微单倍型遗传标记特征参数之间的关系

Supplementary Fig. 9　Relationship among characteristic parameters of microhaplotypes on chromosome 10

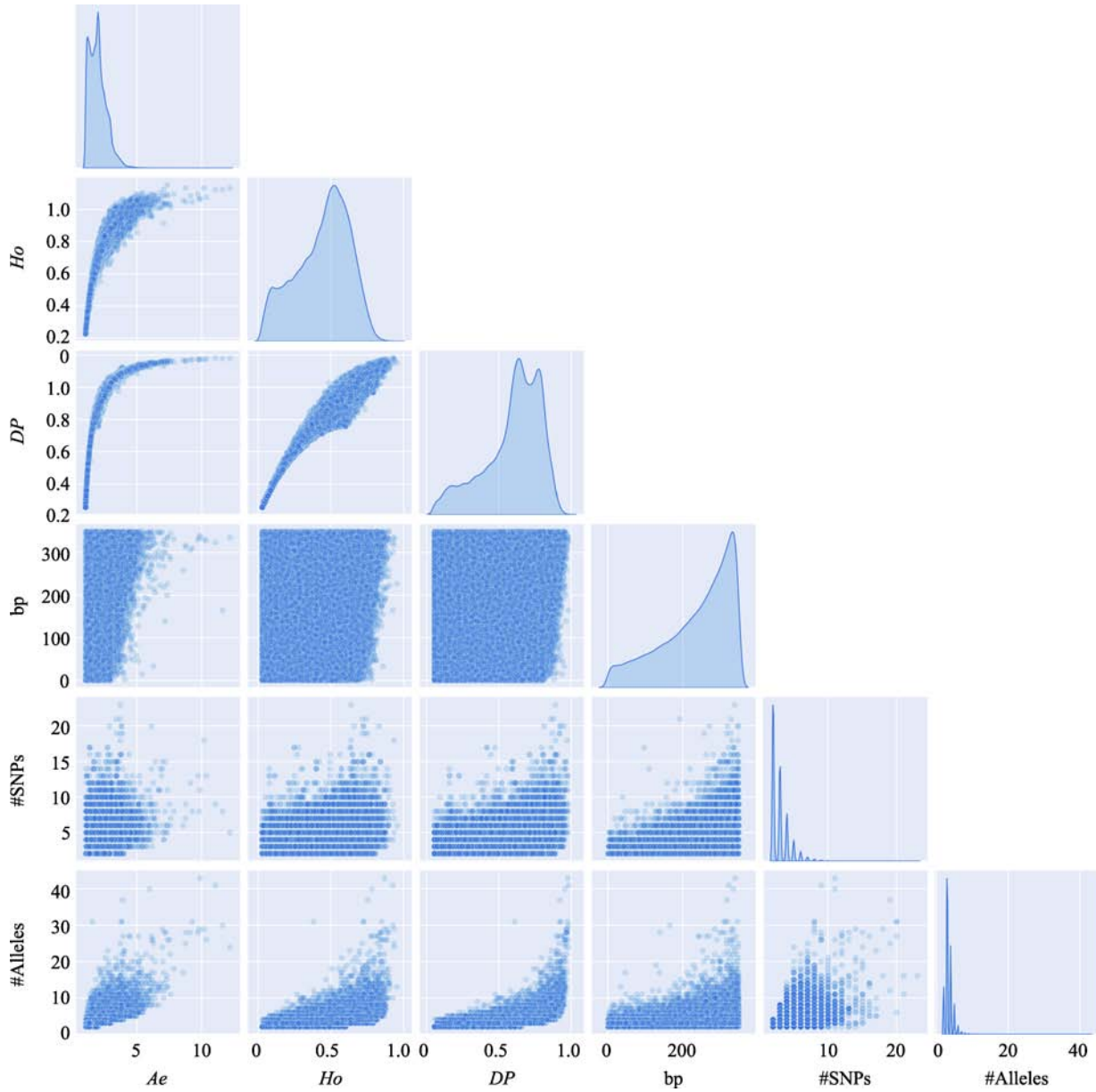
使用 10 号染色体、350 bp 范围内、移除子集的 MH 数据绘制。#SNPs：构成 MH 的 SNP 数；#Alleles：等位基因数。



附图 10　11 号染色体上微单倍型遗传标记特征参数之间的关系

Supplementary Fig. 10　Relationship among characteristic parameters of microhaplotypes on chromosome 11

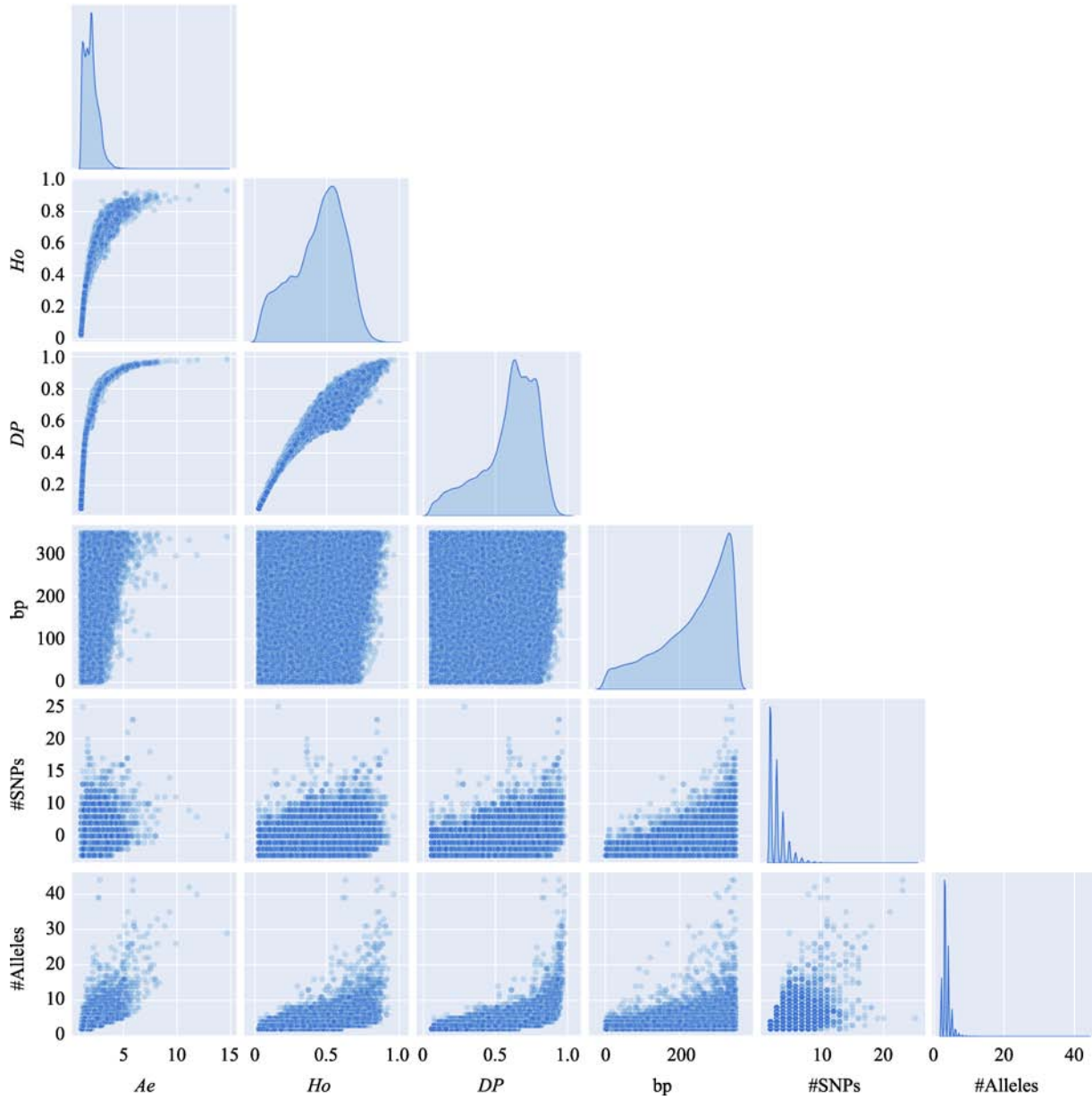
使用 11 号染色体、350 bp 范围内、移除子集的 MH 数据绘制。#SNPs：构成 MH 的 SNP 数；#Alleles：等位基因数。



附图 11　12 号染色体上微单倍型遗传标记特征参数之间的关系

Supplementary Fig. 11　Relationship among characteristic parameters of microhaplotypes on chromosome 12

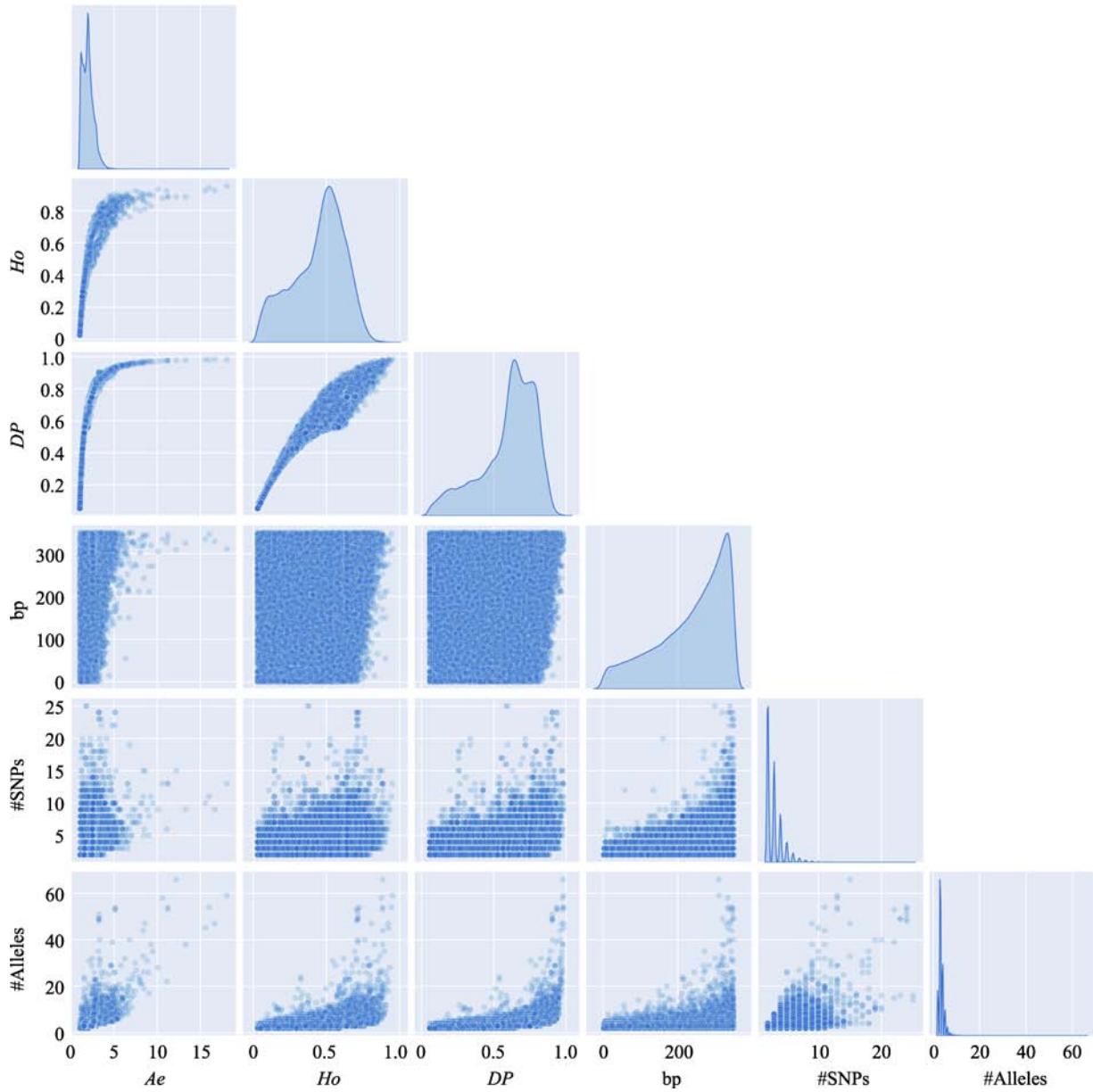
使用 12 号染色体、350 bp 范围内、移除子集的 MH 数据绘制。#SNPs：构成 MH 的 SNP 数；#Alleles：等位基因数。



附图 12　13 号染色体上微单倍型遗传标记特征参数之间的关系

Supplementary Fig. 12　Relationship among characteristic parameters of microhaplotypes on chromosome 13

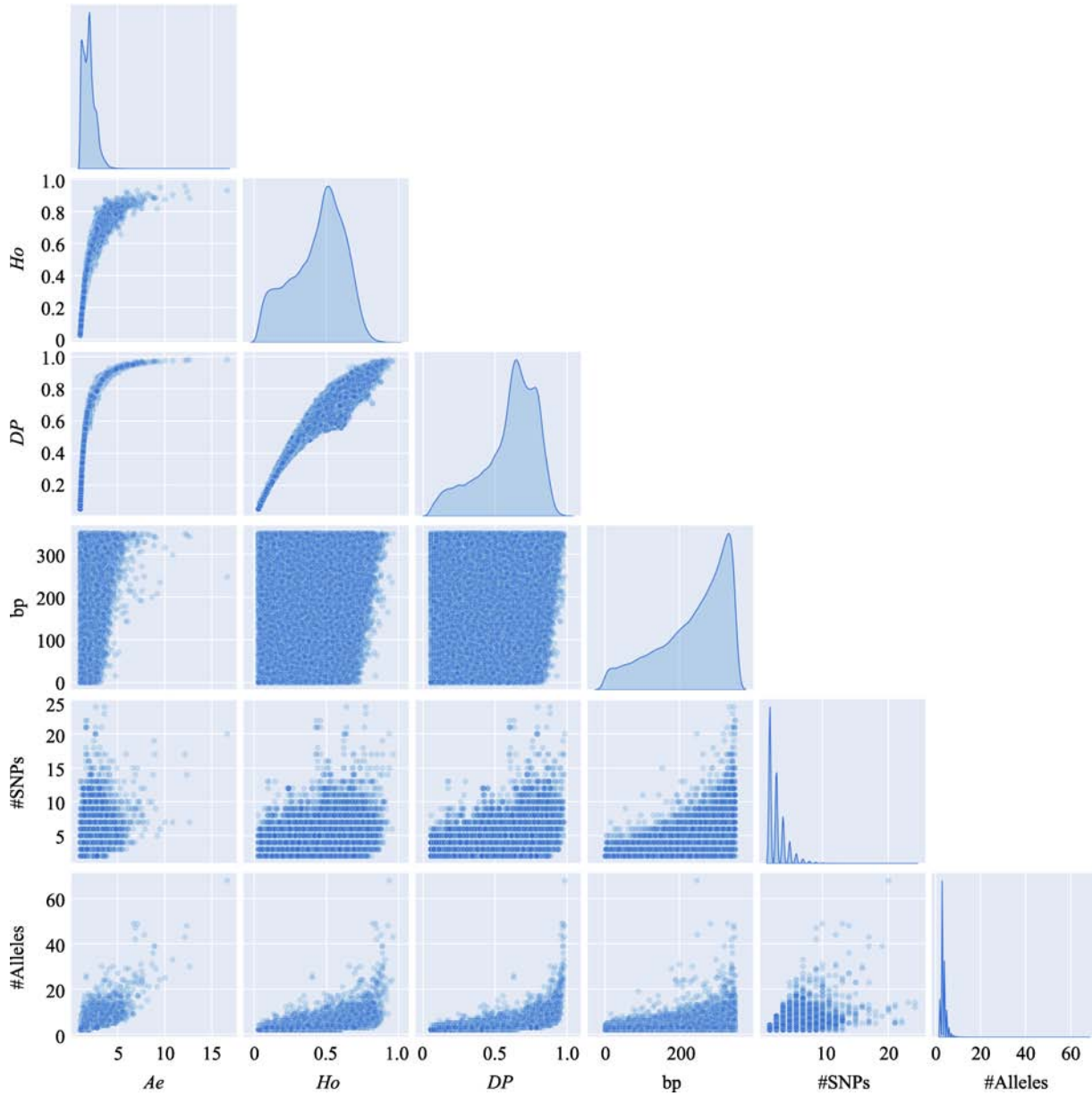
使用 13 号染色体、350 bp 范围内、移除子集的 MH 数据绘制。#SNPs：构成 MH 的 SNP 数；#Alleles：等位基因数。



附图 13　14 号染色体上微单倍型遗传标记特征参数之间的关系

Supplementary Fig. 13　Relationship among characteristic parameters of microhaplotypes on chromosome 14

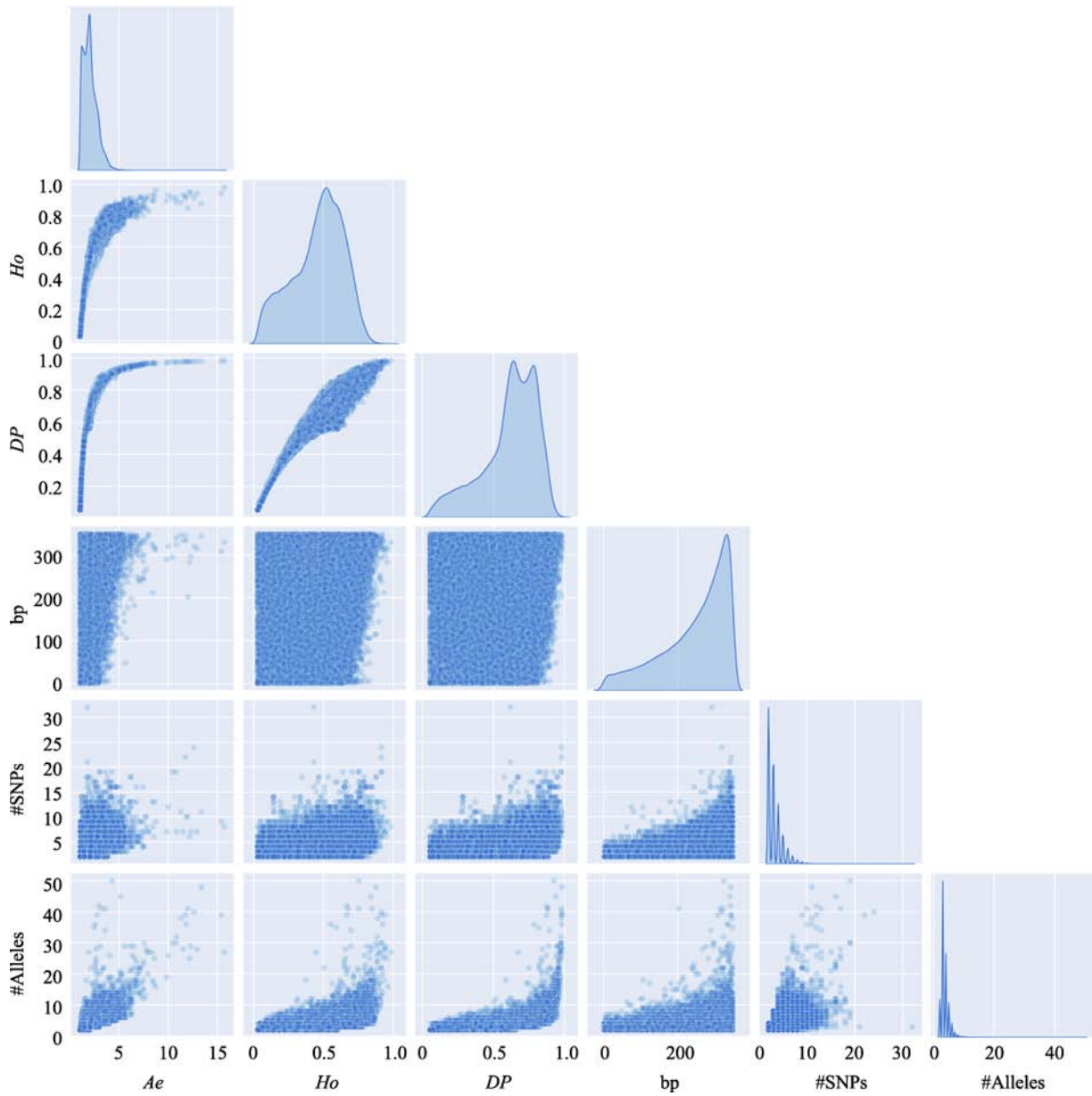
使用 14 号染色体、350 bp 范围内、移除子集的 MH 数据绘制。#SNPs：构成 MH 的 SNP 数；#Alleles：等位基因数。



附图 14　15 号染色体上微单倍型遗传标记特征参数之间的关系

Supplementary Fig. 14　Relationship among characteristic parameters of microhaplotypes on chromosome 15

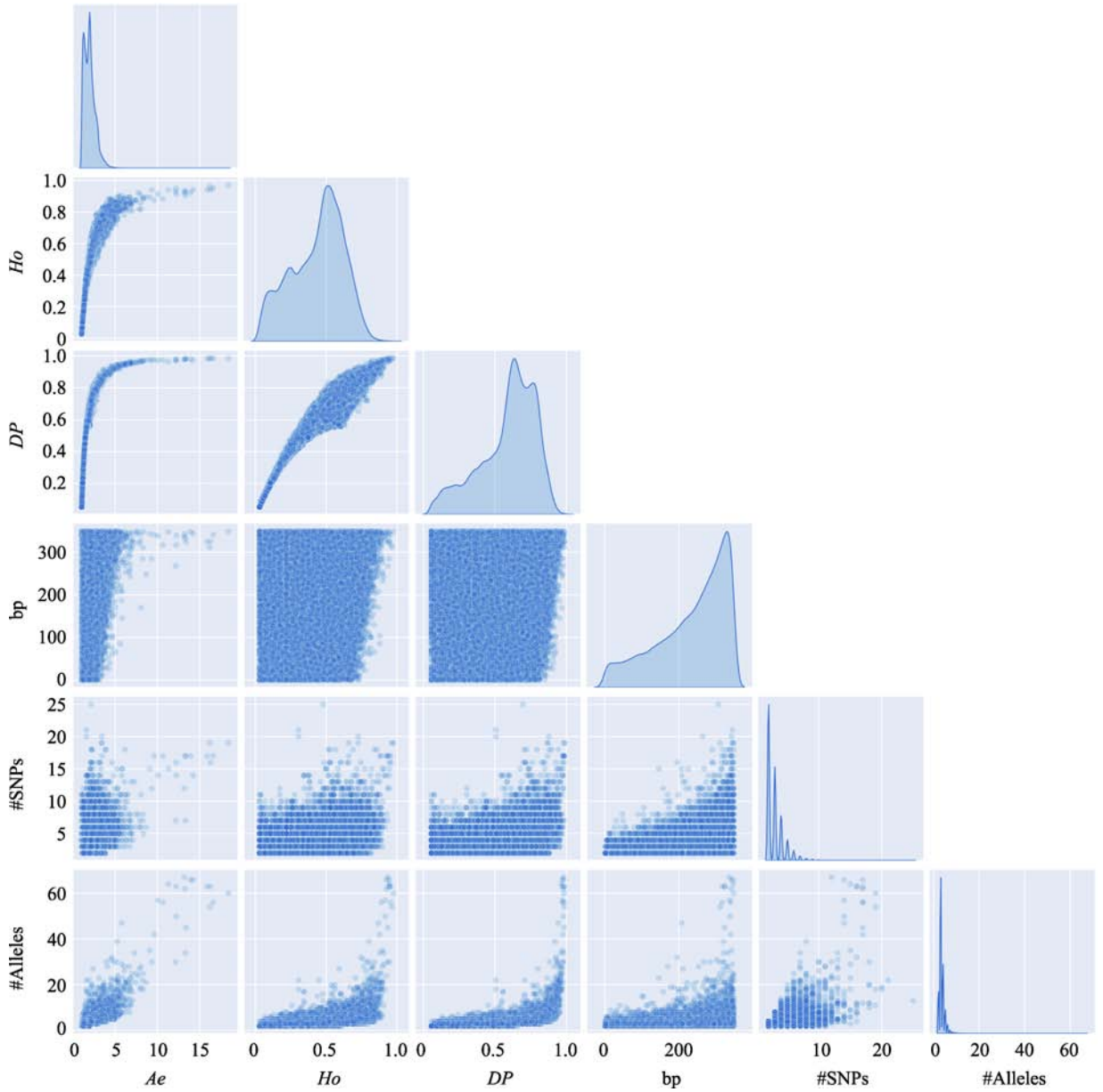
使用 15 号染色体、350 bp 范围内、移除子集的 MH 数据绘制。#SNPs：构成 MH 的 SNP 数；#Alleles：等位基因数。



附图 15　16 号染色体上微单倍型遗传标记特征参数之间的关系

Supplementary Fig. 15　Relationship among characteristic parameters of microhaplotypes on chromosome 16

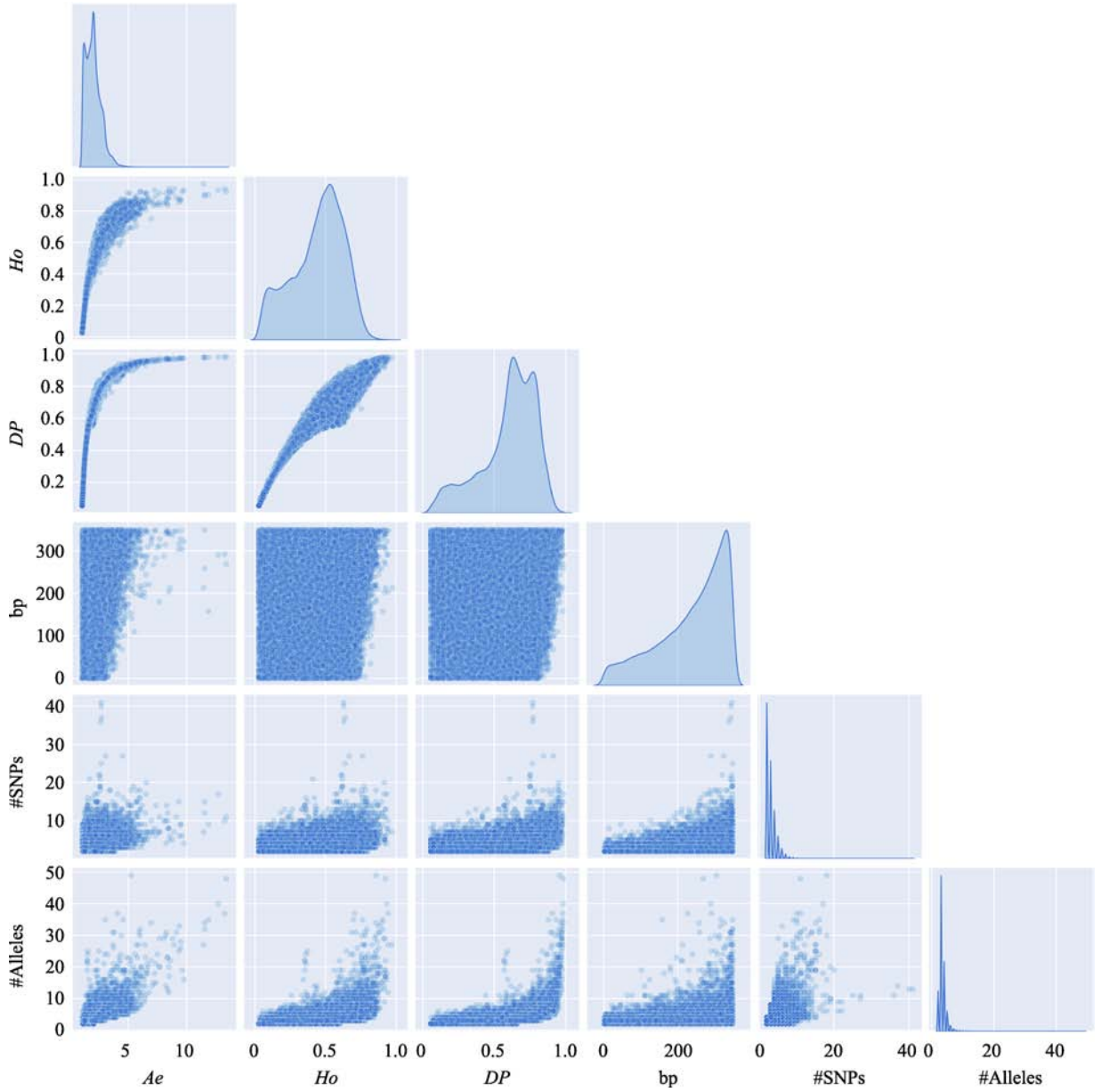
使用 16 号染色体、350 bp 范围内、移除子集的 MH 数据绘制。#SNPs：构成 MH 的 SNP 数；#Alleles：等位基因数。



附图 16　17 号染色体上微单倍型遗传标记特征参数之间的关系

Supplementary Fig. 16　Relationship among characteristic parameters of microhaplotypes on chromosome 17

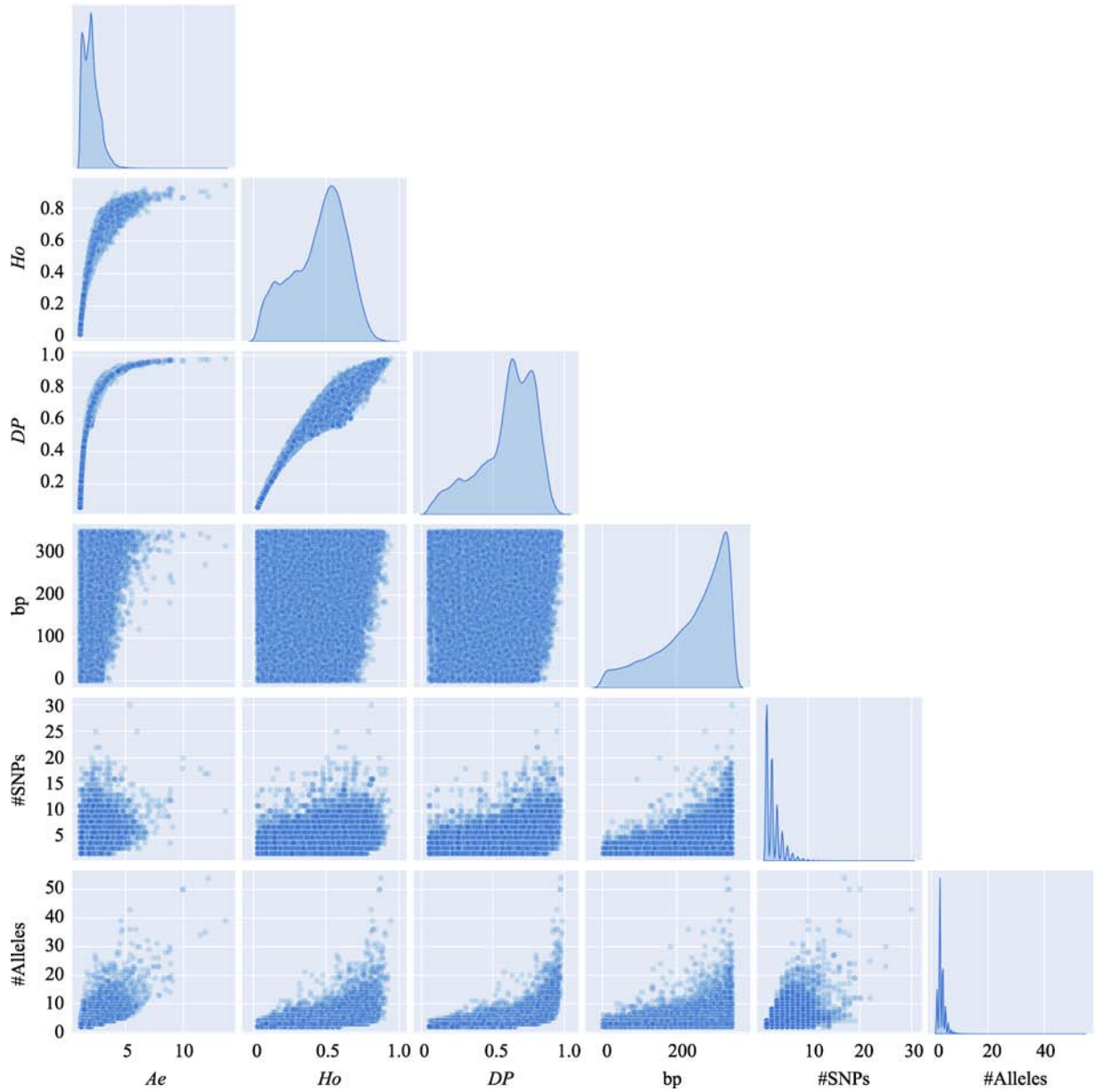
使用 17 号染色体、350 bp 范围内、移除子集的 MH 数据绘制。#SNPs：构成 MH 的 SNP 数；#Alleles：等位基因数。



附图 17　18 号染色体上微单倍型遗传标记特征参数之间的关系

Supplementary Fig. 17　Relationship among characteristic parameters of microhaplotypes on chromosome 18

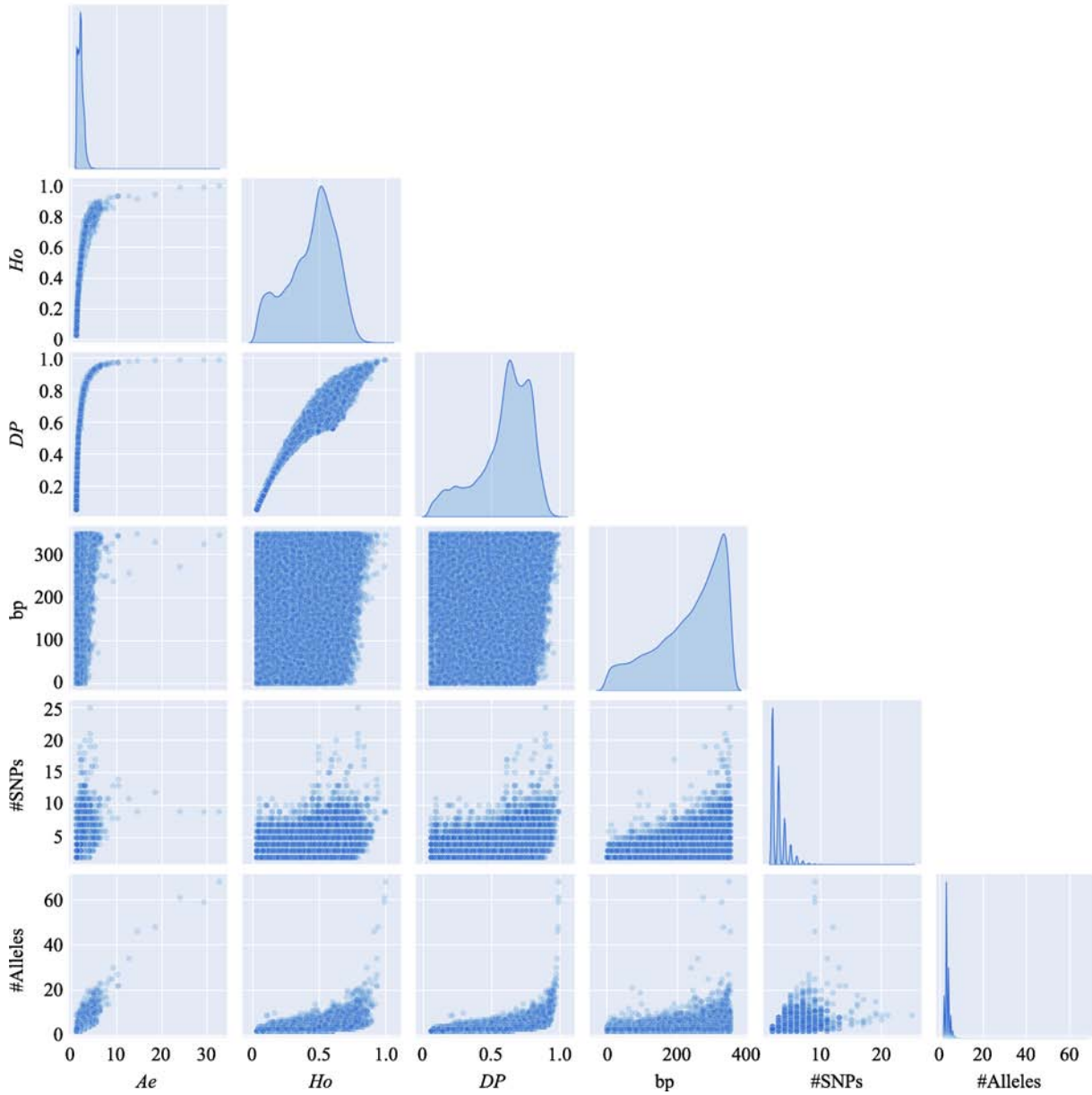
使用 18 号染色体、350 bp 范围内、移除子集的 MH 数据绘制。#SNPs：构成 MH 的 SNP 数；#Alleles：等位基因数。



附图 18　19 号染色体上微单倍型遗传标记特征参数之间的关系

Supplementary Fig. 18　Relationship among characteristic parameters of microhaplotypes on chromosome 19

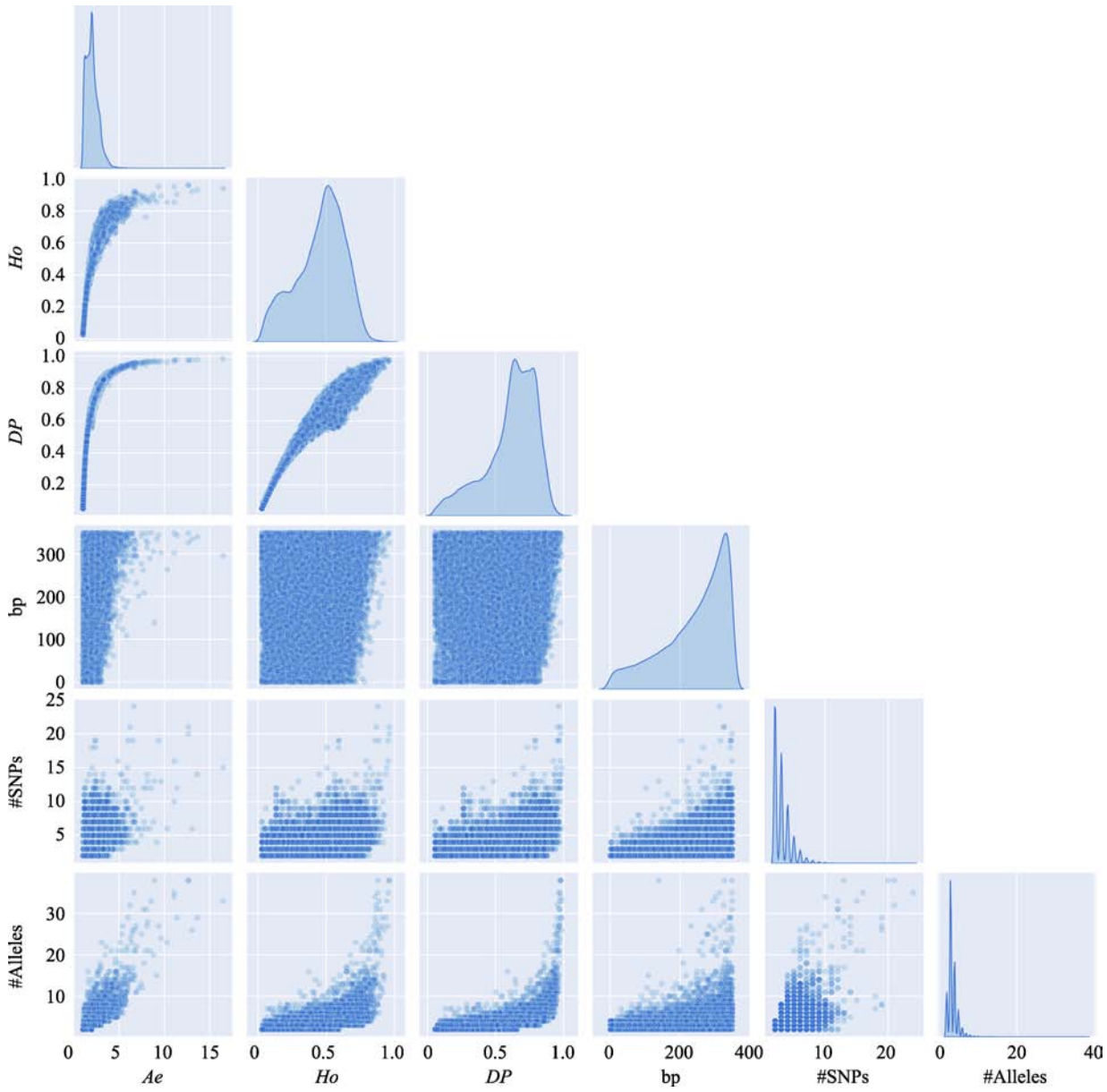
使用 19 号染色体、350 bp 范围内、移除子集的 MH 数据绘制。#SNPs：构成 MH 的 SNP 数；#Alleles：等位基因数。



附图 19　20 号染色体上微单倍型遗传标记特征参数之间的关系

Supplementary Fig. 19　Relationship among characteristic parameters of microhaplotypes on chromosome 20

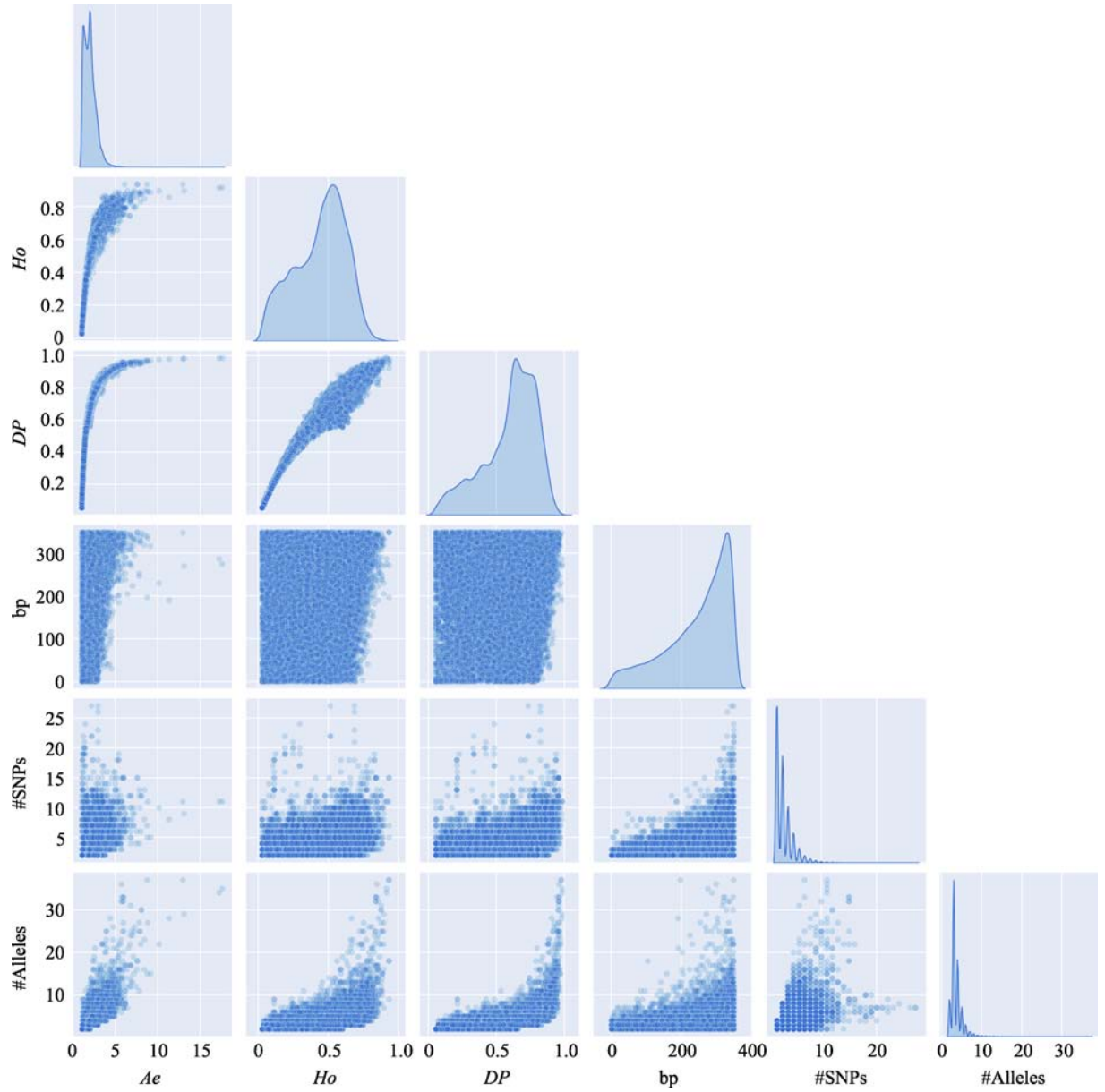
使用 20 号染色体、350 bp 范围内、移除子集的 MH 数据绘制。#SNPs：构成 MH 的 SNP 数；#Alleles：等位基因数。



附图 20　21 号染色体上微单倍型遗传标记特征参数之间的关系

Supplementary Fig. 20　Relationship among characteristic parameters of microhaplotypes on chromosome 21

使用 21 号染色体、350 bp 范围内、移除子集的 MH 数据绘制。#SNPs：构成 MH 的 SNP 数；#Alleles：等位基因数。



附图 21　22 号染色体上微单倍型遗传标记特征参数之间的关系

Supplementary Fig. 21　Relationship among characteristic parameters of microhaplotypes on chromosome 22

使用 22 号染色体、350 bp 范围内、移除子集的 MH 数据绘制。#SNPs：构成 MH 的 SNP 数；#Alleles：等位基因数。