Novel approach for IBD segment filtration of raw ancIBD output

Background

The current implementation of SNP density filtration in the ancIBD framework is a solid general approach to eliminate the majority of false positive identity-by-descent (IBD) segments and retain most of the true positives from raw IBDs. On the one hand, the single global threshold allows easy control of the sensitivity and specificity; on the other hand, it does not account for the potential differences in the marker composition within and between different chromosomal regions and marker sets. Accordingly, for each marker set, it has to be fine-tuned and validated on synthetic data to achieve an optimal balance on sensitivity and specificity. In the ancIBD manuscript, the 220 SNP/cM threshold was validated for the 1240K Allen Ancient DNA Resource (AADR) marker set for the >8cM IBD segments, showing a good compromise for sensitivity and specificity. This can be also seen if we compare the number of raw IBD and the filtered IBD segments that meet the density criterion in a large cohort of samples (Table 1).

Table 1 Per chromosome raw and filtered counts of > 8cM IBD segments in a large cohort of mixed Carpathian Basin and Sarmatian period individuals.

chr	length (cM)	marker count	raw IBD count	IBD count (density)
1	284,26	89079	13950	1214
2	268,82	94167	1503	1166
3	223,26	77600	1040	952
4	214,2	68714	1131	896
5	204,05	69350	1048	932
6	191,72	75811	976	975
7	187,15	59833	908	746
8	168	61091	1156	836
9	166,14	50628	922	783
10	180,91	58760	1205	975
11	158,22	54795	797	781
12	174,59	53881	998	763
13	125,51	38991	79427	781
14	118,6	36279	782573	441
15	141,34	34407	29958	450
16	134,03	34372	758	530
17	128,5	29289	758	436
18	117,55	33899	631	468
19	107,73	18440	631	113
20	108,21	29053	678	463
21	63,64	16031	9950	286
22	72,44	15792	669362	159

In our study we co-analyzed approximatively 1400 individuals from numerous cemeteries where we had a considerable number of close relatives indicated by kinship

analysis. Thus, it is expected that in such a large cohort, where true relatives and also ~1 million pairwise relations exist between mostly unrelated individuals, we will have a considerable number of true IBD segments and also a considerable number of false hits in the raw unfiltered data. As shown in Table 1 the number of raw IBD in a large cohort of individuals has very large variations between the chromosomes, while the filtered IBD counts have much less variations, showing that the density-based raw IBD filtration overall has good specificity. The uneven raw IBD distribution indicates that the local SNP density must also greatly vary within and between chromosomes in the 1240K marker set.

However, the original manuscript also emphasizes that, in the case of smaller than 8 cM long IBD segments, in addition to the density criterion, it is recommended to apply masking of specific genomic locations to avoid an excessive number of false positive IBD segments. This suggests that while the density criterion is a good general approach for longer IBD segments there are other factors that also influence the accuracy of the IBD sharing analysis.

Despite the good specificity of true IBD filtration with the applied length and density criteria, the density criterion can also lead to unexpected behavior. For example, an IBD segment of 8cM with sufficient SNP density will be identified as a true IBD, however, the same IBD segment extending into a less densely (marker density) represented genome region could lead to it being filtered out, as the larger IBD segment's mean SNP density could fall below the 220 SNP/cM threshold. This contradicts our expectation, since in case we have statistically enough markers to prove that two individuals share IBD, then having more markers that also comply with the shared IBD state, this should not be an exclusion criterion.

An extremity of the 1240K marker set is chromosome 19 where the overall SNP density is only ~171 SNP/cM for the whole chromosome, therefore it falls below the recommended 220 SNP/cM threshold. Accordingly, for all parents and offspring who by definition share their entire chr19 as a single IBD segment, this IBD segment will be removed by the density criterion. On the other hand, smaller IBD segments that fall into denser marker regions of the same chr19 where the local SNP density exceeds the 220 SNP/cM threshold can still be identified. Consequently, in this case, although the density method controls the specificity very well, it lacks sensitivity for genome regions with lower SNP density.

In theory, IBD segments are randomly distributed across the genome; therefore, the number of IBDs should correlate well with the genetic map length of the chromosomes. The considerable number of true IBD segments indicated by the density criterion in our experiment is sufficiently large that small stochastic variations should not significantly influence the expected distribution. To test and visualize our null hypothesis we plotted the number of filtered (true) IBD segments and the length of chromosomes (Figure 1).

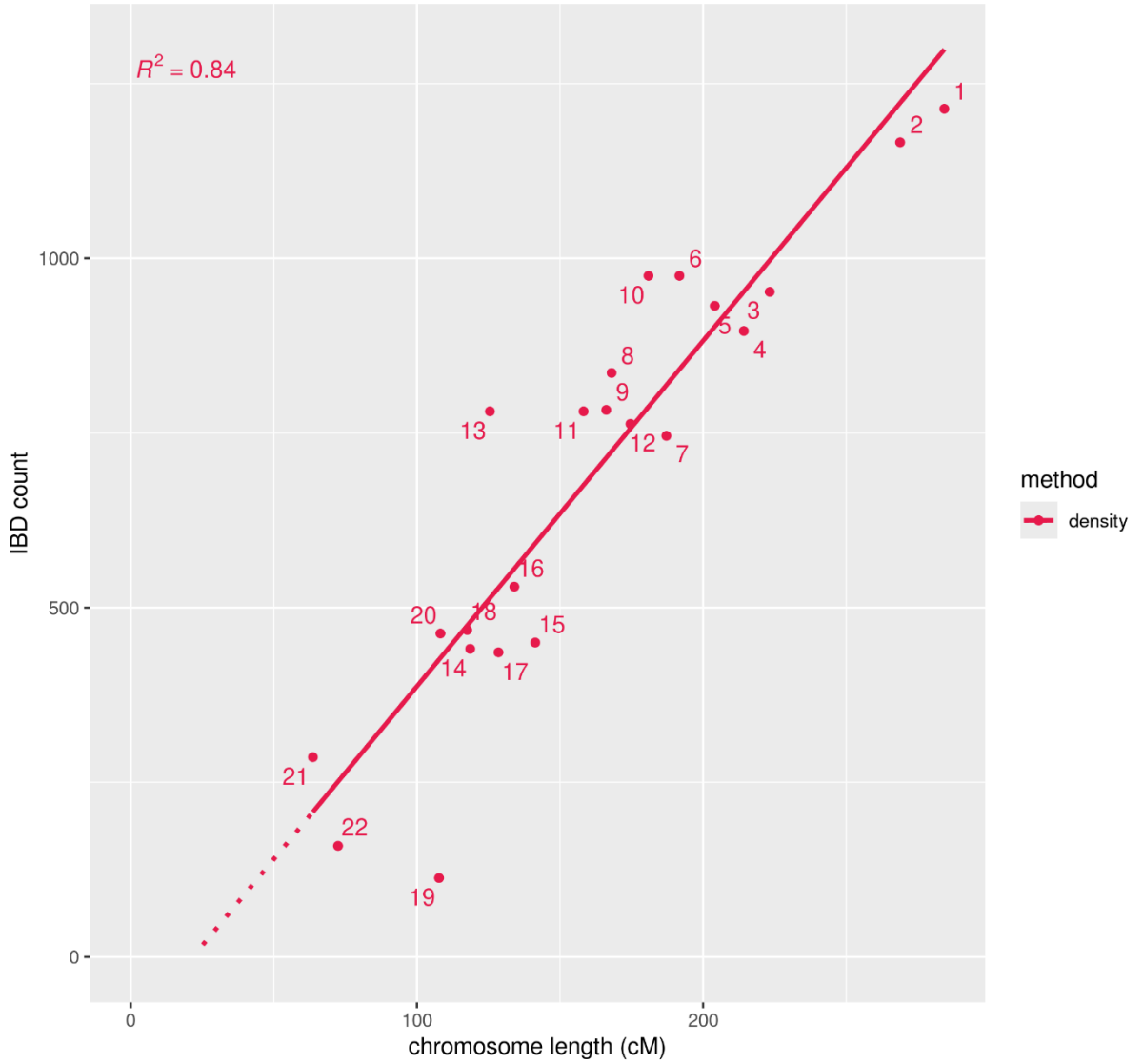

Figure 1 IBD count per chromosome indicated by the SNP density method in a large cohort of samples.

Consistent with our null hypothesis, there is a significant correlation (R=0.84) between the observed count of indicated true IBD segments and chromosome length, suggesting that the density criterion is effective in the majority of cases. While the density filtration method and the 8cM length criteria removes majority of false positive IBD segments, the experimental IBD counts of the density-based filtration also illustrates some anomalies compared to the null hypothesis. Notably chromosomes such as chr13, chr10 and chr6 exhibits significantly higher than expected IBD counts likely due to excess of false positive IBD segments that were not excluded by the density criterion. In contrast there are also chromosomes that have less than expected IBD counts (chr19 is a prominent example), indicating a potential lack of sensitivity. The positive intercept of the linear fit along the x-axis further corroborates the suboptimal sensitivity, indicating that in some chromosomes an excess of true positive IBD segments were excluded.

Experimental distribution of the raw IBD segments within the genome

Assuming the null hypothesis that the true IBD segments are randomly dispersed within the genome, the IBD coverage defined as the number of true IBD segments crossing any genome position should follow a Poisson distribution. In practice, we only have sparse genotype information at the positions of the markers, and the IBD segments are defined as a subset of sequential markers with their unique genotypes. Accordingly, the experimental IBD count (coverage) at any marker position can be defined as the number of IBD segments that includes the particular marker. Due to the false positive IBD segments in the raw data the distribution of the experimental IBD count of the markers deviates from the Poisson distribution. However, according to our null hypothesis, the subset of markers present in the true IBD segments should exhibit a Poisson distribution for this metric. Our results show, that the majority of markers follow the expected Poisson distribution for IBD count (Figure 2).

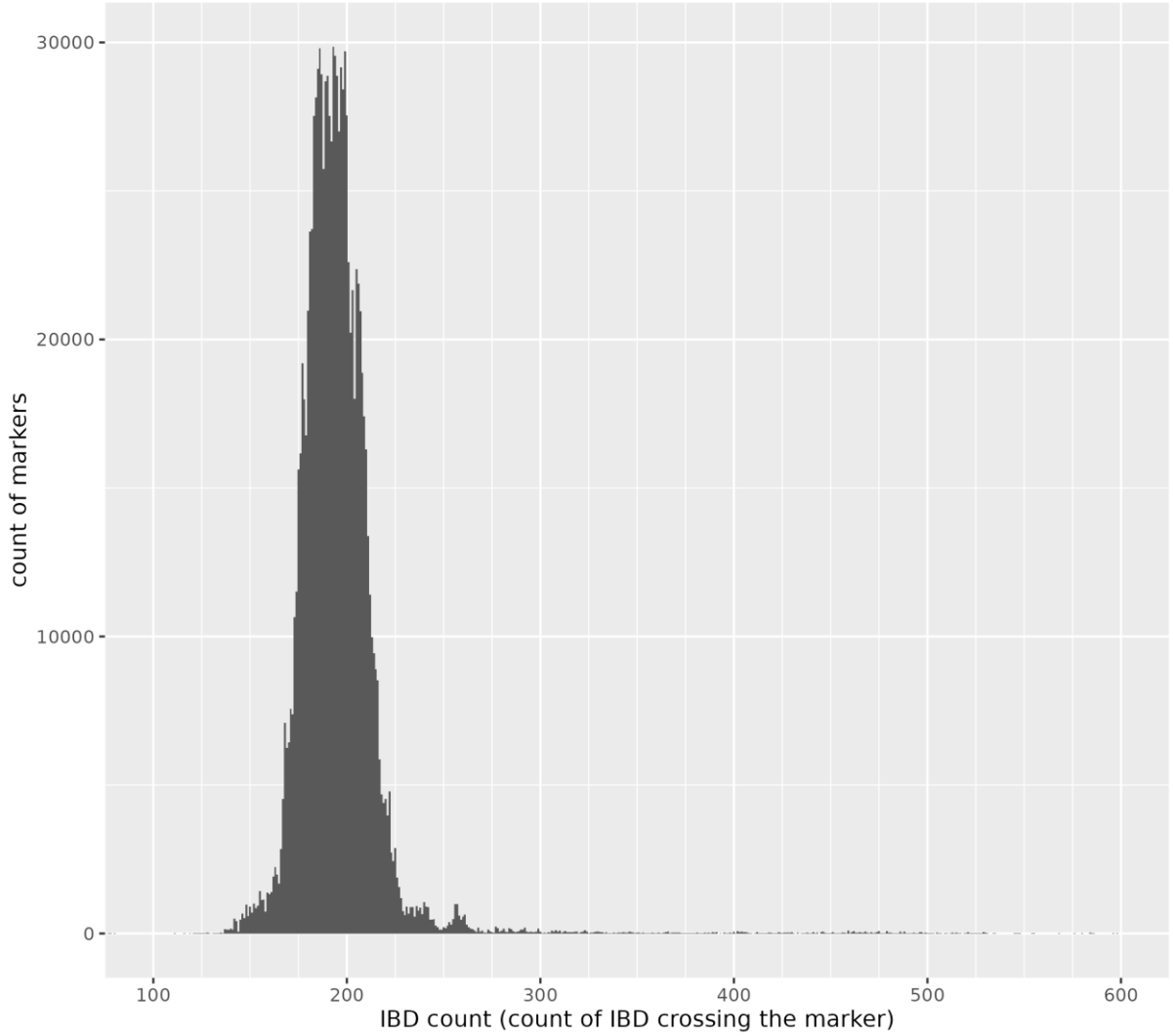

Figure 2 Histogram of the number of 1240K AADR markers included in a particular number of IBD segments (IBD count) based on experimental raw IBD data.

A small fraction of the markers has higher than expected IBD counts. Since the very long right tail of the distribution is hard to visualize in Figure 2 we also present the percentile table of IBD counts of the 1240K markers (Table 2). The percentile data indicates, that less than ~3% of the 1240K marker set has IBD counts that are much higher than those expected from the Poisson distribution with a median of 193 IBD count. Only about 0.3% to 0.5% of the markers have several magnitudes higher IBD counts, ranging from about 100,000 to 800,000.

These counts correspond to the case that 10% to 80% of all sample pair combinations share raw IBD segments at these particular locations, thereby accounting for the majority of the false positive raw IBD segments.

Table 2 The distribution of markers in the 1240K AADR marker set based on the metric of the count of IBDs that include the particular marker.

percentile	IBD count
1,0%	155
5,0%	171
10,0%	176
25,0%	184
50,0%	193
75,0%	204
90,0%	213
95,0%	222
96,0%	227
97,0%	240
98,0%	263
99,0%	1091
99,1%	2627
99,2%	6180
99,3%	9742
99,4%	11441
99,5%	29309
99,6%	79017
99,7%	79018
99,8%	660260
99,9%	777376

The proportion and the distribution of the affected markers indicates that the problematic regions are confined to a small fraction of the genome where large fraction of unrelated samples seemingly share IBD with each other. These genomic regions mainly overlap with the mask regions described in the ancIBD manuscript (**Table 3-4**).

Table 3 The "mask" regions according to the ancIBD manuscript

CHR	start (cM)	end (cM)	length (cM)
1	147,421	163,079	15,658
2	0,014	26,182	26,168
4	0,341	11,535	11,194
8	0,0004	31,511	31,5106
10	62,682	77,491	14,809
14	111,711	120,2	8,489
15	0,005	43,484	43,479
17	54,307	63,702	9,395
18	0,1604	25,465	25,3046
19	0,002	29,543	29,541
19	74,533	107,7316	33,1986
21	0,861	21,041	20,18
22	1,723	23,842	22,119

Table 4 The "mask" regions based on the experimental IBD count (genome locations exceeding the median IBD count with greater than 6SD).

CHR	start (cM)	end (cM)	length (cM)	mean IBD count	SNP/cM
1	147,119	164,242	17,123	3927,4	192,8
2	2,479	10,178	7,699	357,5	145,6
8	21,257	30,052	8,795	443,6	171,6
10	63,722	74,055	10,333	410,7	127,8
13	0,614	10,978	10,364	58760,7	224,3
14	2,477	16,408	13,931	470279,9	179,0
15	14,099	25,337	11,238	18001,4	103,8
21	1,896	12,078	10,182	7852,9	181,2
22	2,492	20,169	17,677	398523,2	135,5

According to experimental data, the majority of mask regions have low marker density in the 1240K marker set. In general, both the ancIBD mask track and our experimental data suggest that the problematic regions correspond mainly with the telomeric and centromeric regions of chromosomes. However, not all chromosomes are equally affected, and these regions also exhibit large variations in the extent of false positive IBDs and the spatial distribution of markers (Figure 3).

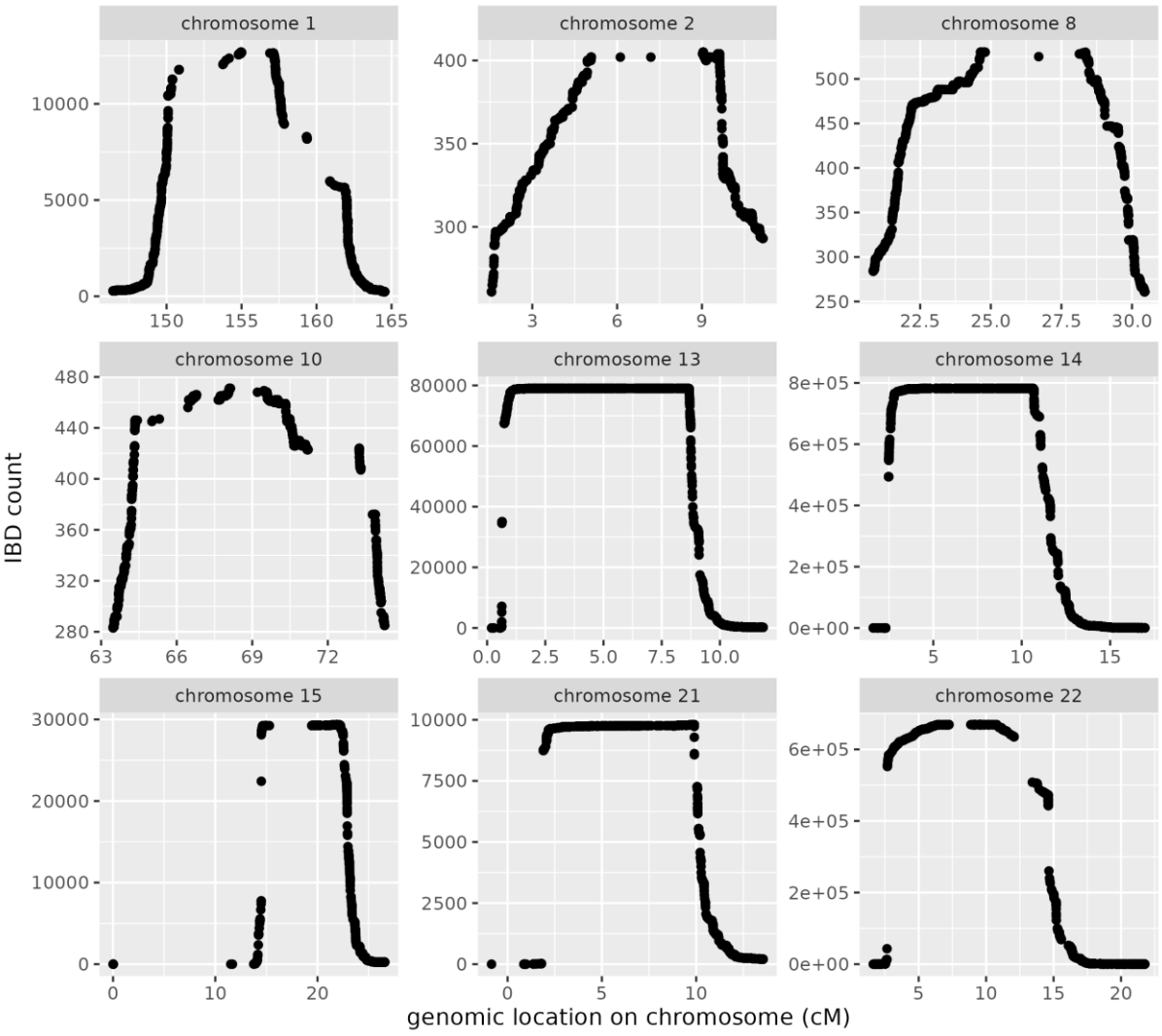

Figure 3 The experimental IBD count distributions at the low confidence genomic regions (mask regions).

The higher incidence of low-confidence regions near the telomeric and centromeric regions is likely the result of multiple factors. In the telomeric and centromeric regions, the density of unique ancestry-informative markers is generally lower because of the higher fraction of repetitive sequences. In case of low-coverage ancient sequences, the genotypes that are the basis of IBD analysis are inferred by imputation. Consequently, the accuracy of the imputation also plays an important role. Since imputation uses the spatial genotype context to infer the most likely haplotype(s), and in these regions the sufficient length of flanking context is partially missing, it could also attribute to the lower accuracy in these regions. The observed patterns suggest that the differences between the distribution and the density of markers in the 1240K marker set, and also the imputation accuracy are mainly responsible for the observed differences between the accuracy of IBD analysis within and between the chromosomes.

Although the low-confidence regions suggested by our analysis overlap with the mask track provided in ancIBD, we also see some notable differences. According to the experimental IBD count data, approximately 80000 sample pair combinations share IBD on the telomere of chromosome 13 which is several magnitudes higher than the expected 193 IBD count. The excess of IBD count indicates a low confidence region despite the greater than 220 SNP/cM density that is not included in the ancIBD mask track. An additional difference is that the mask track of the ancIBD manuscript includes the centromere of chromosome 14 while the experimental data suggest that only in the telomere do we have much higher than expected IBD sharing compared to the null hypothesis (Figure 3).

We have to note that in the ancIBD manuscript the locations of the mask regions were determined using the experimental data of a large cohort of Eurasian Ancestry individuals, while our data set included mainly Carpathian basin individuals and Sarmatian related individuals out of the Carpathian basin. In theory masking is only recommended below 8cM IBD length, thus we expect that the ancIBD mask track would include additional low-confidence regions where the specificity of density filtration drops. However, the additional low confidence region even at 8cM and the observed differences between the optimal mask areas of the two sample cohorts suggest that the uncertainty of IBD detection could also depend on the genome structure of the analyzed samples.

Although the global marker density in the panel largely correlates with the density of the actual markers that are informative in the analyzed cohort, in a smaller fraction of genomic regions where markers do not represent all populations equally, differences in the analyzed populations could influence the distinguishing power of the density-based single global threshold filtration method. Consequently, the optimal mask track could be somewhat different based on the populations being analyzed, as we expect that these false positive IBD areas correlate with the actual marker informativity representing the analyzed populations in the given region. For example, if our marker set in a specific genomic region mainly contains markers that have genotype, and haplotype variability in only African people, then this region will be appropriate for analyzing individuals with African ancestry. However, despite of the overall SNP density in this region, individuals of non-African ancestry will have a smaller genotype and haplotype variability, leading to a greater chance of sharing an identity-by-state (IBS) with other individuals of non-AFR ancestry. This is especially true as we decrease the length of the analyzed IBD segments. Furthermore, this also implies that for different marker sets (with their unique marker composition and distribution), the problematic regions could be different and the mask track has to be optimized individually for the optimal sensitivity and specificity.

Marker informativity based filtration

To address the aforementioned issues, we propose a new approach for filtration based on the experimental distribution of raw IBD segments. According to our null hypothesis, the

IBD segments confined to these problematic genomic regions with unexpectedly high raw IBD counts are likely false. Consequently, the distribution of the experimental raw IBD segments can be used to indicate the 'mask areas' specific for the analyzed cohort. We defined IBD count of a marker as the number of IBD segments a particular marker is included. As we expect Poisson distribution for the true IBD segments, the ratio of the expected (true) and observed raw IBD counts (true + false) of each marker reflects the likelihood that the marker indicates true IBD segments. Based on this, we can calculate a marker informativity for each marker by the following formula:

$$M_I(i) = \frac{\boldsymbol{med}(I_c)}{I_c\,(i)}$$

Where $M_I(i)$ is the marker informativity of marker *i*, $\boldsymbol{med}(I_c)$ is the median of all marker IBD counts and $I_c(i)$ is the IBD count of marker *i*.

Since each IBD segment includes multiple markers, we can also calculate the sum of marker informativity a metric for each IBD segments that correlates with likelihood that the particular IBD segment represents a true IBD by the following formula:

$$I_s(j) = \sum_{i \in M(j)} M_I(i)$$

Where $I_s(j)$ is the IBD informativity score of IBD segment j and $M(j)$ denotes all markers included in IBD segment j. Similarly, to the density method, a single threshold can be used to test against the calculated probability of all IBD segment to assess whether the raw IBD segment is likely false or true positive. However, unlike the global density threshold, the proposed metric is based on empirical probabilities that indicate whether the underlying markers within the IBD segment represent a true IBD segment. Consequently, this approach is less susceptible to the limitations posed by specific marker sets or population-specific regions that exhibit low accuracy in imputation and/or IBD analysis, leading to better sensitivity and specificity of IBD filtration.

The global 220 SNP/cM density threshold validated on the 1240K AADR marker set for IBD segments larger than 8cM corresponds to the IBD informativity score of 8×220. Therefore, the evaluation of raw IBD segments with the score criterion in the 1240K marker set, could be done with the following test.

$$I_s(j) = \begin{cases} > 8 \times 220, & true\ IBD \\ \leq 8 \times 220, & false\ IBD \end{cases}$$

This threshold would result in a very similar performance to density-based method for the majority of genome regions (approximatively 97% of the markers) where the raw IBD counts follow the Poisson distribution. This is due to the fact that in case of a greater than 8cM IBD segment with density of greater than 220 SNP/cM, when the underlying markers are within a genomic region with a median IBD coverage it is granted that the IBD informativity score will be greater than 8×220 and therefore the IBD will be classified true just like with the density model.

As highlighted by the ancIBD manuscript, the specificity and accurate measurement of the lengths of the IBD segments are vital in IBD analysis to mitigate the risk of drawing incorrect conclusions based on invalid IBD segments or misinterpreted close relationships. Accordingly, the implementation of our method consists of two consecutive steps.

The first step involves calculating the experimental IBD informativity scores from the raw IBD distribution. Our formula effectively reduces the scores of IBD segments that fall within low-confidence areas and improves the exclusion of false positive raw IBDs leading to better specificity. In the subsequent step, we also test whether the ends of the identified IBDs extend into a low marker-informativity "mask" area. In such a case, we truncate the end of the IBD extending into this low-confidence region where the markers are likely inaccurate, and only include the IBD in the final results in case the remaining high-confidence part of the IBD is still longer than applied length threshold. This second step ensures that IBD lengths are not over-estimated at low confidence genomic regions.

The notable differences against the density models are the following:

- If a large portion of the markers falls within a region where the empiric IBD count is high indicating a likely false positive “mask region”, then these markers will have <<1 marker informativity and the IBD informativity score can go below the 1760 (8×220) threshold signaling, that even in case the global marker density in the region is higher than 220 SNP/cM, many of these markers are likely not indicating a true IBD segment. Furthermore, the end of IBDs extending into low-confidence regions are truncated, and the IBD segment is only kept, if the remaining high-confidence part is still larger than the length threshold. Hence, no manual masking is required while we expect improved specificity and a conservative length estimate around the low-confidence regions.
- The other notable difference is that the density method may drop out IBDs spanning larger genomic segments that have a general low density of SNPs, while the IBD informativity score method can still indicate these as valid IBD if sufficiently large number of markers indicates that the IBD is shared between the two individuals and these makers do not fall within the problematic “mask” regions with low probability of true IBD. Consequently, the score method will not throw out an 8cM IBD segment if it extends into a lower density but non-problematic genomic region, and it can also identify IBDs at generally low marker density areas if the supporting marker informativity is sufficiently high. Hence, we expect an improved sensitivity.

Comparison of the density and marker informativity methods

To assess the feasibility of the proposed method, we used a large cohort of samples from the Carpathian basin (approximatively 1300 individuals, 50+ cemeteries) and approximately 100 publicly available samples (that are source of some migrations into the Carpathian basin from the same period but outside of the Carpathian basin). Our data contained a large number of proven true relatives including first, second, and more distant relatives, thus we expected considerable amount of true IBDs in the cohort, while also most of the sample pair combinations should be between individuals that are unrelated, providing us enough false positive raw IBD segments to filter out. We included only shotgun WGS data for all of the samples (no capture data). We excluded individuals with lower than 0.5x mean genome coverage or greater than 4% contamination suggested by ANGSD X contamination or Schmutzi mtHG contamination analysis. The common markers of 1KG Phase 3 data (~78 million positions) were imputed by GLIMPSE2 using the phased genotypes of the 1KG phase 3 individuals as reference.

We used the official ancIBD workflow, to restrict sites to the 1240K marker set with the 8cM length threshold for IBD identification. Using the *hapBLOCK_chroms()* function, ancIBD indicated a large number of raw IBD segments (~**1.6M** > 8cM raw IBD segments in all chromosomes). We used the official 220 SNP/cM density method and the proposed IBD informativity score method to filter the raw IBD segments and compare the results. As

expected, both methods filtered out most of the false positive IBD segments, while retaining a very similar number of true IBDs (Table 3)

Table 3 True IBD counts filtered by density and IBD informativity score method

method	IBD count
density	15146
score	16390

Furthermore, the two methods were largely concordant as ~92% of the indicated true IBD segments were the same. The score method rejected 662 IBD segments indicated by the density method while it also identified 1906 additional IBDs (Table 4).

Table 4 Concordance between density and IBD informativity score methods.

type of IBD	IBD count
concordant	14484
exclusive to density	662
exclusive to score	1906

To test against the random distribution null hypothesis, we also collected the number of true IBDs per chromosome indicated by the two methods (Table 5).

Table 5 Per chromosome stats on the true IBD counts by the density and IBD informativity score methods

CHR	length (cM)	marker count	IBD count (score)	IBD count (density)
1	284.26	89079	1247	1214
2	268.82	94167	1201	1166
3	223.26	77600	1023	952
4	214.20	68714	1010	896
5	204.05	69350	1011	932
6	191.72	75811	975	975
7	187.15	59833	866	746
8	168.00	61091	780	836
9	166.14	50628	836	783
10	180.91	58760	879	975
11	158.22	54795	792	781
12	174.59	53881	917	763
13	125.51	38991	545	781
14	118.60	36279	468	441
15	141.34	34407	514	450
16	134.03	34372	634	530
17	128.50	29289	611	436
18	117.55	33899	589	468
19	107.73	18440	402	113
20	108.21	29053	572	463
21	63.64	16031	282	286
22	72.44	15792	336	159

We plotted the number of true IBD segments and the length of the chromosomes to visualize the correlation expected from our null hypothesis (Figure 4). Our results show that the IBD informativity score method has a better correlation (R=0.95) between chromosome

length and IBD counts, suggesting that the distribution of indicated IBD segments is closer to the random distribution as assumed from our null hypothesis. Furthermore, the intercept on the x-axis is closer to the expected 0, suggesting that the score method also has slightly improved sensitivity, without gross error on the specificity.

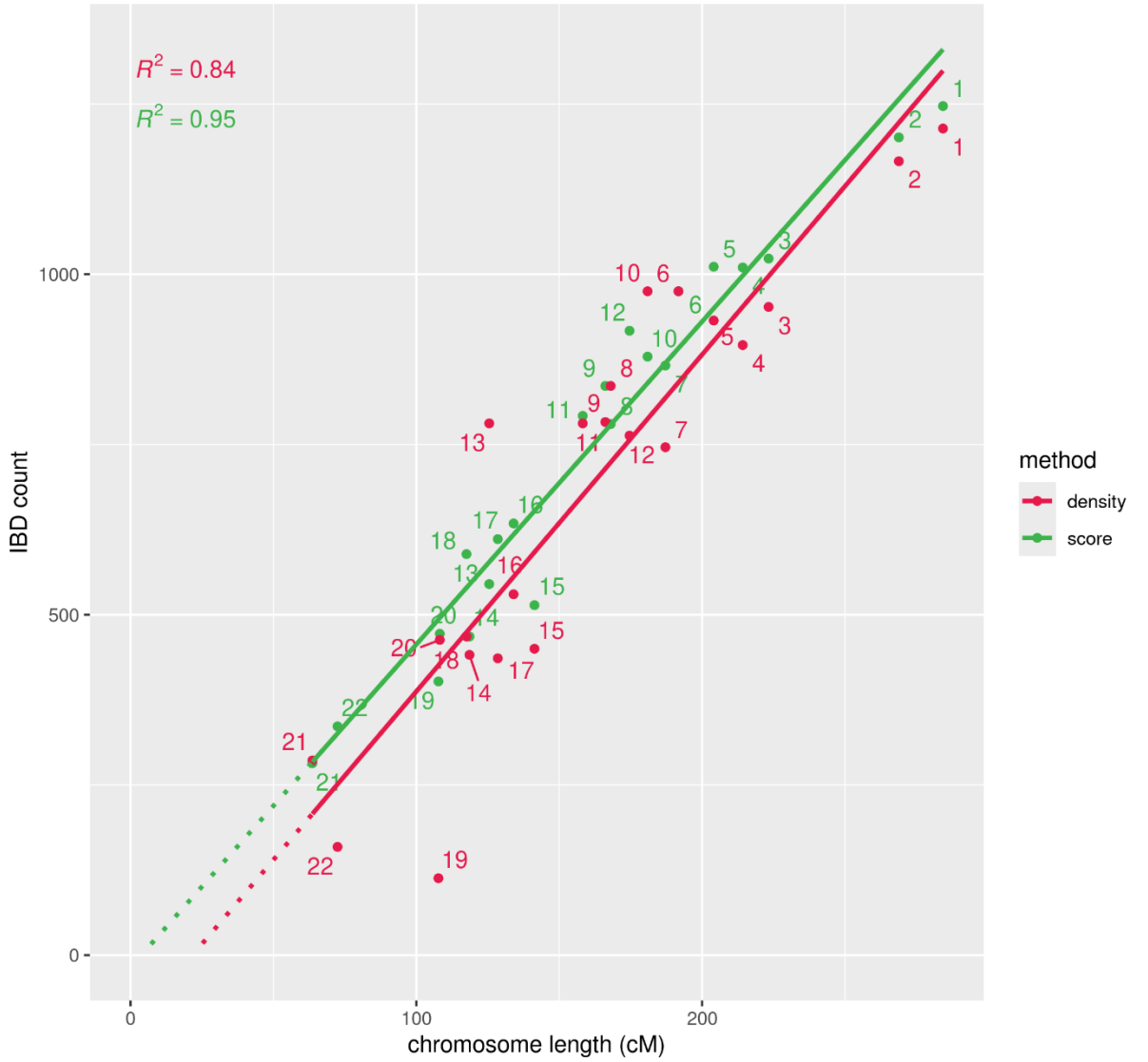

Figure 4 IBD count per chromosome indicated by the SNP density and the IBD informativity score methods in a large cohort of samples.

Unsurprisingly, most of the extra IBD segments indicated by the density method and excluded by the score method was on chromosomes (13, 10) where we had an excess of IBD compared to the expectation based on the chromosome length as seen in Figure 4. These excluded IBDs were concentrated around mask areas, where a large number of unrelated individuals shared IBD and had marker densities slightly above the 220 SNP/cM density threshold (Table 6).

Table 6 Example IBDs at chromosome 13 excluded by the IBD informativity score method.

start	end	start (M)	end (M)	length (marker)	length (M)	CHR	ID1	ID2	score	density (SNP/cM)
78	2345	0,0074	0,1075	2267	0,100111	13	RUN2_ind1	RUN12_ind1	136,73	226,45
78	2345	0,0074	0,1075	2267	0,100111	13	RUN2_ind1	RUN15_ind1	136,73	226,45
73	2377	0,0064	0,1096	2304	0,103246	13	RUN2_ind2	RUN22_ind1	157,55	223,16
78	2383	0,0074	0,1098	2305	0,102427	13	RUN2_ind3	RUN22_ind2	161,83	225,04
78	2507	0,0074	0,1147	2429	0,107308	13	RUN2_ind4	RUN4_ind1	272,81	226,36
78	2343	0,0074	0,1074	2265	0,100095	13	RUN2_ind5	RUN19_ind1	135,61	226,29
176	2598	0,0092	0,1188	2422	0,109577	13	RUN2_ind5	R11118.SG	369,5	221,03
73	2446	0,0064	0,1127	2373	0,106332	13	RUN2_ind6	RUN3_ind1	213,51	223,17
73	2376	0,0064	0,1095	2303	0,103196	13	RUN2_ind6	RUN18_ind1	156,85	223,17
78	2629	0,0074	0,1226	2551	0,115238	13	RUN2_ind6	RUN18_ind1	404,21	221,37
74	2376	0,0064	0,1095	2302	0,103148	13	RUN2_ind6	RUN20_ind1	156,85	223,17
78	2458	0,0074	0,1128	2380	0,105427	13	RUN2_ind7	RUN13B_ind1	224,44	225,75
78	2457	0,0074	0,1127	2379	0,105385	13	RUN2_ind7	RUN19_ind1	223,49	225,74
78	2420	0,0074	0,1119	2342	0,104523	13	RUN2_ind8	RUN13A_ind1	190,76	224,07
78	2420	0,0074	0,1119	2342	0,104523	13	RUN2_ind8	RUN22_ind1	190,76	224,07

Since this region had much higher number of IBD segments than expected compared to our null hypothesis, marker informativity was low. Consequently, their IBD informativity score was only between in the range of 136-404 significantly below the 1760 applied threshold, although the SNP density within these IBD segments was slightly above the overall 220 SNP/cM density threshold. Despite the sufficient marker density, the low IBD informativity scores mean that this area is a "mask" area, as also shown in Figure 3.

The other example where the density method significantly deviated from the null hypothesis was chromosome 10 (Figure 4). In Table 7 we present a few examples of the IBD segments excluded by the IBD score method, all found within the low confidence centromeric region of chromosome 10 (Figure 3).

Table 7 Example IBDs from chromosome 10 excluded by the IBD informativity score method.

start	end	start (M)	end (M)	length (marker)	length (M)	CHR	ID1	ID2	score	density (SNP/cM)
18598	20717	0.6194	0.7002	2119	0.0808	10	RUN12_ind1	RUN10_ind1	1359.5	262.2
18526	20848	0.6180	0.7049	2322	0.0869	10	RUN2_ind1	RUN2_ind1	1477.5	267.1
18670	20944	0.6208	0.7063	2274	0.0855	10	RUN13B_ind1	R10620.SG	1396.8	265.9
18702	20945	0.6216	0.7063	2243	0.0847	10	RUN13B_ind1	R10631.SG	1371.7	264.8
20454	23051	0.6765	0.7907	2597	0.1142	10	RUN12_ind1	RUN18_ind3	1821.9	227.5
18670	22109	0.6208	0.7616	3439	0.1408	10	RUN12_ind2	RUN18_ind4	2188.2	244.2
17380	22173	0.5844	0.7643	4793	0.1799	10	RUN12_ind2	RUN22_ind1	3419.7	266.4
17309	20419	0.5822	0.6650	3110	0.0827	10	RUN14_ind1	RUN8_ind1	2420.6	375.8
17691	20684	0.5982	0.6969	2993	0.0987	10	RUN15_ind2	RUN5_ind1	2164.3	303.2
21073	24566	0.7109	0.8225	3493	0.1116	10	RUN15_ind2	RUN21_ind1	2889.8	313.1
21127	24767	0.7324	0.8267	3640	0.0943	10	RUN17_ind1	RUN1_ind1	3040.3	386.1
20963	24253	0.7070	0.8157	3290	0.1087	10	RUN18_ind1	RUN6_ind1	2664.8	302.7
18670	22119	0.6208	0.7631	3449	0.1423	10	RUN18_ind1	RUN22_ind1	2197.0	242.4
17681	23343	0.5964	0.7982	5662	0.2019	10	RUN19_ind1	RUN21_ind1	4158.1	280.5
18528	21422	0.6181	0.7412	2894	0.1231	10	RUN19_ind2	MJ-43_noUDG.SG	1773.3	235.2

In case of chromosome 10, the core of the low confidence area between ~61-70 cM has a general high mean marker density, with some larger gaps between markers within the region and also around 71 cM (Figure 3). All the indicated IBD segments within this region have greater than 220 SNP/cM density; therefore, they were kept by the density filtration method. On the contrary, shorter segments of IBD that only span the low confidence area were excluded due to low IBD informativity scores. However, in most cases, after truncating the low-confidence regions, the length of the high-confidence part of the IBD fell below the threshold, leading to exclusion by the score method.

In summary, our result suggests that the differences between the two methods are mainly due to that fact that a small fraction of genomic regions has excessive raw experimental IBD counts, signaling that the region has low confidence in IBD analysis. While the majority of such regions also fall within the low SNP density areas, a tiny fraction of these region had above than 220 SNP/cM density leading to significant number false positive IBD segments with the density approach.

In contrast to the suboptimal specificity observed in certain chromosomes, Figure 4 also illustrates that the density method has a lack of sensitivity in a few chromosomes, with chromosome 19 being the most notable example. As described previously, chromosome 19 has a low overall marker density (~171 SNP/cM). Unsurprisingly, a large portion of newly indicated IBDs are on chromosome 19 where the density model excluded many IBDs due to low global SNP density. Since we had many first relatives and even sample duplicates in our cohort, we could verify that known IBDs that were excluded due to the global low SNP density of chr19 were still identified with the probability method (Table 8).

Table 8 Example IBDs from chromosome 19 between known relatives and sample duplicates that were indicated by score but excluded by the density criterion.

start	end	start (M)	end (M)	length (marker)	length (M)	CHR	ID1	ID2	score	density (SNP/cM)
11	18440	0,0008	1,0773	18429	1,0766	19	RUN10_ind10	RUN13A_ind1	16081,7	171,18
16	18440	0,0012	1,0773	18424	1,0762	19	RUN10_ind7A	RUN10_ind7B	16068,5	171,2
8403	18440	0,5250	1,0773	10037	0,5523	19	RUN10_ind7A	RUN18_ind15	9042	181,73
188	5902	0,0303	0,4536	5714	0,4233	19	RUN10_ind7A	RUN8_ind19	4462,82	135
13732	18440	0,7514	1,0773	4708	0,3260	19	RUN10_ind7A	RUN8_ind19	4229,24	144,43
8403	18440	0,5250	1,0773	10037	0,5523	19	RUN10_ind7B	RUN18_ind15	9042	181,73
178	5875	0,0281	0,4514	5697	0,4233	19	RUN10_ind7B	RUN8_ind119	4444,11	134,59
13732	18440	0,7514	1,0773	4708	0,3260	19	RUN10_ind7B	RUN8_ind119	4229,24	144,43

To assess the validity of the IBDs that are unique to a specific filtration method, we also compared the IBD segments suggested exclusively by either method with the consensus IBD segments, which were suggested by both methods. We calculated the number of IBDs that were between sample pairs already indicated by the consensus IBD segments. We also calculated the number of IBDs between newly indicated sample pairs that were not indicated by any IBD within the consensus IBD segments (Table 9).

Table 9 Concordance of the sample pairs indicated by the method exclusive IBDs with the sample pairs indicated by the consensus IBD segments (indicated by both methods).

IBD suggested by only	IBD between already indicated sample pairs	IBD between newly indicated sample pairs	percent of already indicated	percent of newly indicated
density	74	588	11,18%	88,82%
score	1186	720	62,22%	37,77%

Our analysis shows that a large portion (~90%) of the IBD segments indicated by the density method only and excluded by the score method indicate a connection between the sample pairs that did not share IBD within the consensus IBD segments. While 62% of IBDs proposed by only the score method indicate IBD sharing between sample pairs that already share IBD within the consensus IBD segments. The ratio also suggests that the score method excludes likely false negative IBD segments, as most excluded IBDs are between random unrelated people indicated in genome regions with unexpectedly high IBD counts where the local SNP density is slightly above the 220 SNP/cM threshold. On the other hand, majority of the IBD segments indicated exclusively by the score method are between sample pairs that also

share IBD according to the consensus of the two methods, suggesting that these IBD segments are likely true positives that only fell out due to the SNP density criterion in the genome regions with lower local SNP density.

Lastly, we also plotted the inferred IBD count and cumulative IBD length for the greater than 12 cM long IBD segments indicated by the score method (Figure 6) to compare the distribution with the empirical distribution published in the original ancIBD manuscript based on 4248 Eurasian individuals (https://www.nature.com/articles/s41588-023-01582-w/figures/3).

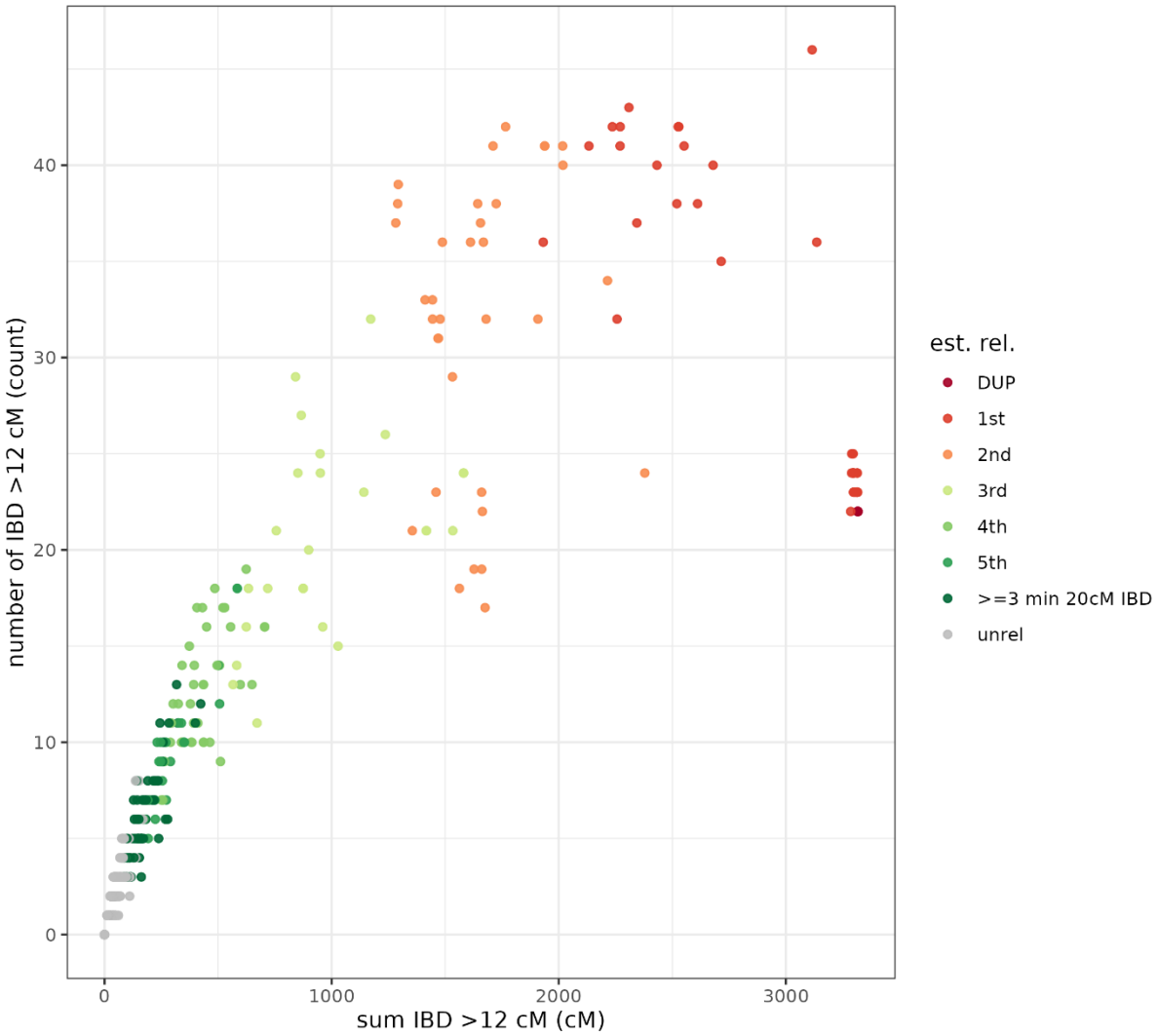

Figure 6. Inferred IBD among pairs of 1,388 ancient Carpathian basin and Sarmatian ancestry related individuals. The plot visualizes both the count (y axis) as well as the summed length (x axis) of all IBD >12cM long. Kinship relations were estimated with the correctKin tool using the same 1240K AADR marker data used in the IBD analysis.

We also applied the five-state HMM for haploid IBD sharing analysis where the length and count distribution of the parent-offspring and full siblings are markedly different. The plot shows that both distribution based on imputed experimental low coverage ancient data are very similar. However, the IBD count of the 39 known sample duplicates and 60 parent-offspring indicated by the score method are within a much closer range (22-25 respectively) to the expected IBD count of 22 (sharing all haploid autosomal chromosome) than the distribution of IBD count of parent-offspring relations indicated by the density-based method (20-35 respectively, as seen in Figure 3 of the ancIBD manuscript). As the data points representing

these relations are very closely aligned on Figure 6, it cannot be seen that in the majority of cases the score method indicated the expected 22 IBD count, and only a smaller fraction of cases this value deviated from this (Table 10).

Table 10 The count of >12 cM IBD shared between 39 sample duplicates and 60 parent-offspring, where the expected number of shared IBD is the 22 autosomal chromosome.

indicated chromosome count	number of cases
22	72
23	15
24	9
25	3

Furthermore, the range of cumulative IBD lengths of the parent-offspring is in a narrower range (3300.7-3356.6 cM) range compared to the density-based plot (~3200-3400cM respectively) presented in the ancIBD manuscript. These figures also indicate that the score method offers improved sensitivity and specificity compared to the density approach.

Visualization of the TP and FP IBD segments based on the two metrics

To compare the distribution of consensus and method exclusive data, we plotted the distribution of the true positive IBD segments (TP consensus, density, score method), the suggested false positives (FP consensus, density, score method), the true positives suggested by only the density method, the true positives suggested by only the score method, and finally the IBDs that were excluded in the score method after truncating the low-confidence part of the IBD extending into a mask area. We present both the distribution of SNP density and IBD informativity scores versus the length distribution of the IBD segments in the 9 IBD groups (Figure 7, 8).

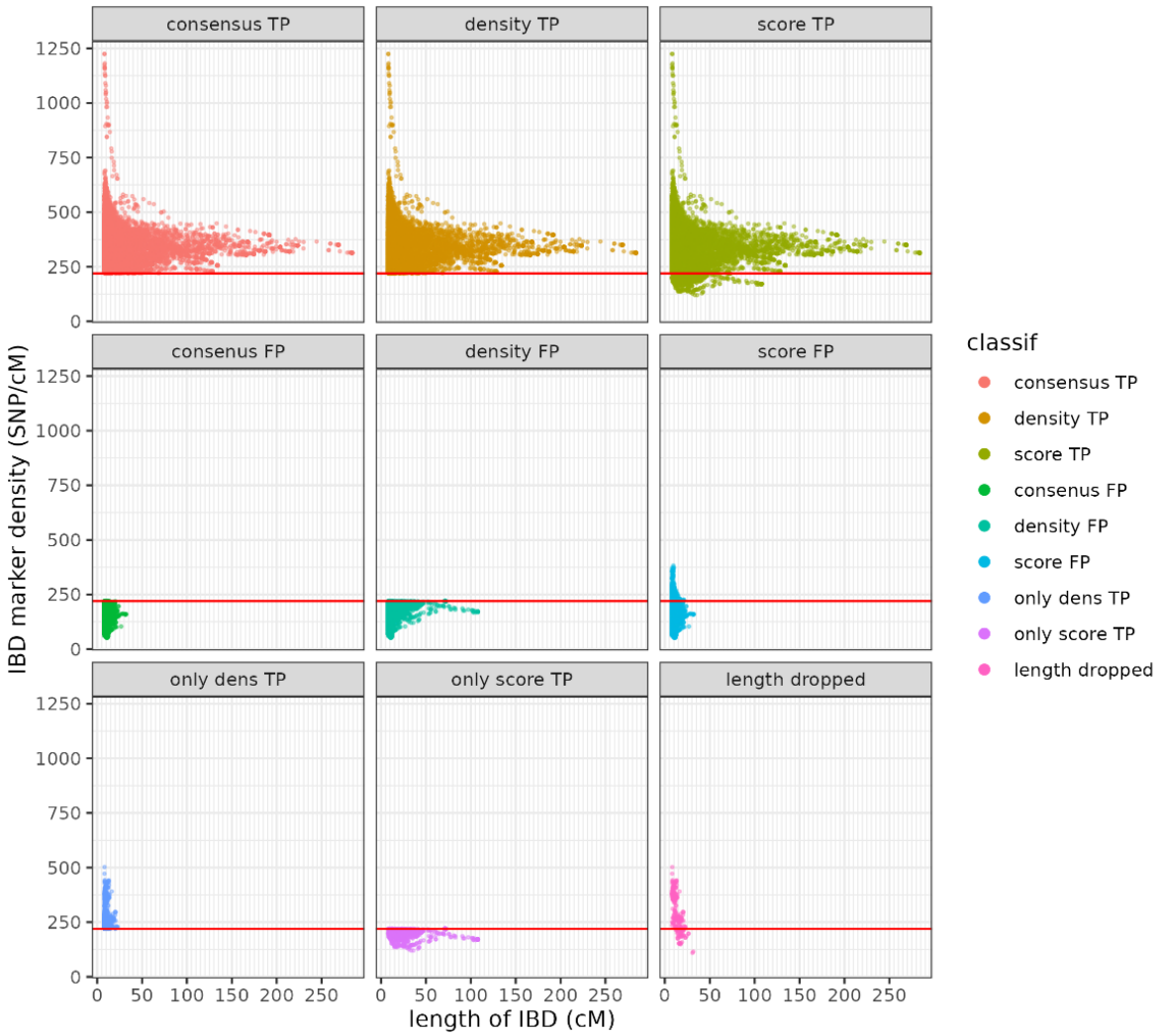

Figure 7 The SNP density/length distribution of IBDs classified as: true positive (TP) or false positive (FP) by both method (consensus), density method, score method; the TPs of only the density method, only the score method; and the SNP density/length distribution of the length dropped IBDs of the score method.

We must note that in our cohort there are ~1.6M raw IBDs. Compared to this, the true positive IBDs indicated by either method are only a tiny fraction (15-16K), while the method exclusive IBDs are even smaller fraction (662 and 1906). Consequently, although it cannot be properly visualized in Figure 7, almost all of the 1.6 million raw IBDs are in the consensus FP category as excluded by both methods. In the case of the indicated false positive IBD segments, Figure 7 shows that compared to the density method distribution, the consensus distribution is

much more similar to the score method distribution. We also see that the score method FPs include some very short IBDs with high SNP density. According to the IBD informativity formula, this can only happen in the case of IBD segments that consists largely of markers with low marker informativity. Therefore, these short IBDs must fall into genomic regions where a much higher number of individuals share IBDs than expected by the random distribution of true IBD segments. It is important to note that the TPs exclusive to the density method indicate small IBD segments only, with most segments marginally above the 220 SNP/cM density threshold. In contrast, the true positives identified exclusively by the score method are generally longer IBD segments, some even exceeding 100 cM length.

The IBD informativity score versus the length distribution of the IBD segments for the same 9 IBD groups is shown in Figure 8.

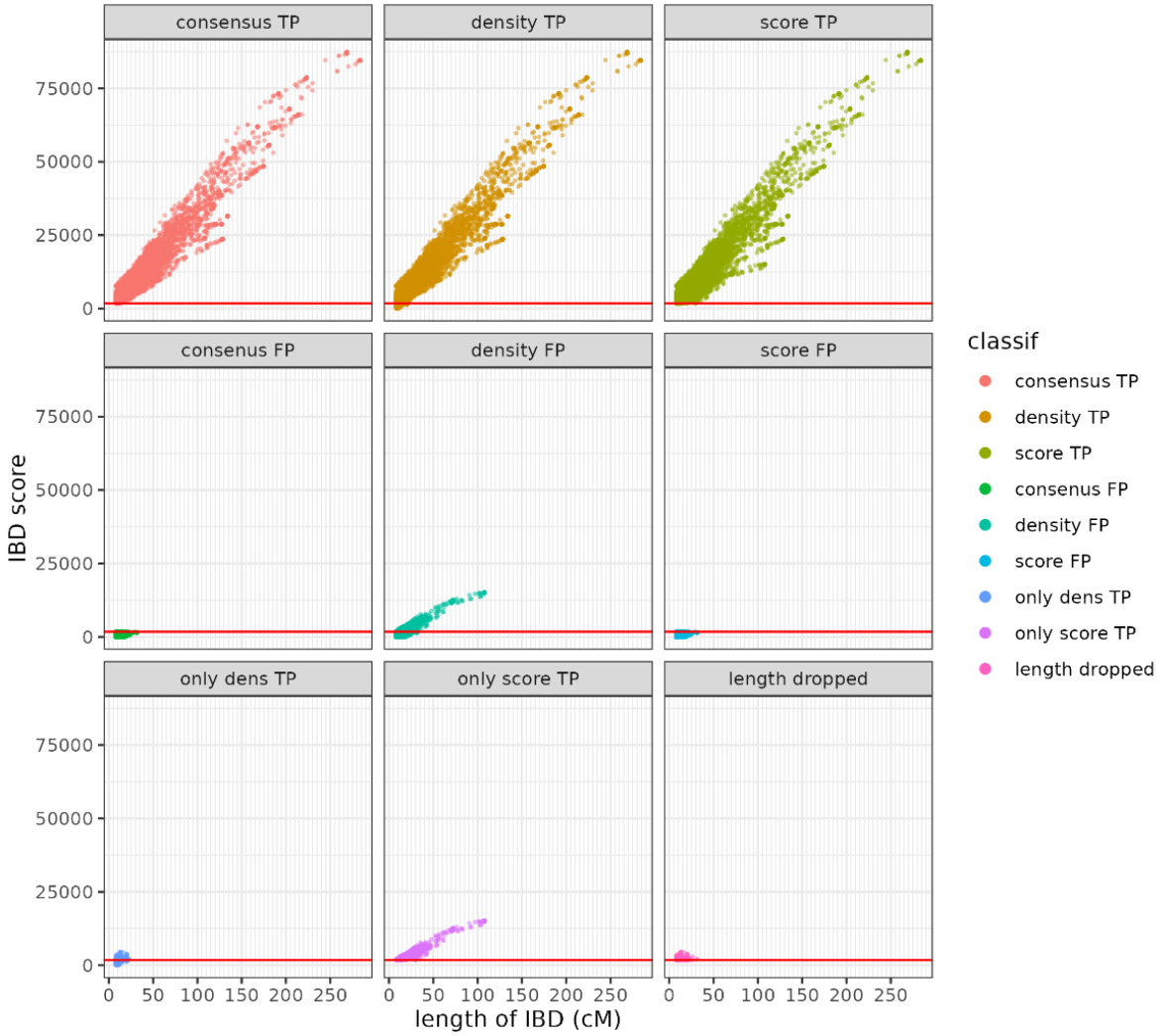

Figure 8. The IBD informativity score/length distribution of IBDs classified as: true positive (TP) or false positive (FP) by both method (consensus), density method, score method; the TPs of only the density method, only the score method; and the IBD informativity/length distribution of the length dropped IBDs of the score method.

As we can see in Figure 8 the IBD informativity score differentiates TP and FP IBDs on a much larger scale. Most FPs included in the consensus FP group have <1760 IBD informativity score and are confined to a very small distinct area in the plot. On the other hand,

most consensus TPs have magnitudes of higher IBD informativity scores and are more distinct from consensus FPs compared to the SNP density plot (Figure 7). Similarly, as seen in Figure 7 the distribution of consensus FP IBD segments is more similar to the distribution of the IBD score method compared to the SNP density method. Furthermore, the distribution of density FP is much more similar to the consensus TP compared to the consensus FP, indicating that these IBDs were falsely rejected by the global density threshold. On the contrary, the density TPs are all low-score IBDs, many of which are also included in the length dropped group, indicating that while the marker density within the IBD segment is above the 220 SNP/cM threshold, these IBD segments are mainly overlapping or extending into low-confidence mask areas and their IBD informativity score or their high-confidence IBD length is below the applied thresholds.

Our result shows that the distribution of the consensus TP and FP IBD segments aligns more closely with the distributions indicated by the IBD informativity score than with those derived from the density method. The score method displays reduced overlap between the distributions of consensus FP and consensus TP IBD segments. This suggests, that the currently used 1760 IBD informativity score threshold could be likely fine-tuned and a more stringent threshold could be applied to improve specificity without substantial compromise on the sensitivity.

Conclusion

Unlike the density method, which treats each marker as equally informative for identifying true IBD segments, our approach assesses the likelihood that the underlying markers indicate a true IBD segment. The formula applied for the calculation of the IBD informativity score mitigates the manual “masking” of low confidence genomic areas in an automatic manner, largely independent of the length of the IBD, the specific set of markers and the differences due to population structure of the analyzed individuals. Our analysis suggests that the IBD informativity score method improves the sensitivity and specificity of the raw IBD filtration compared to the density approach. These improvements together could potentially allow for more robust filtration of raw IBDs and analysis of shorter IBD segments.