Original Paper

Human Heredity

Hum Hered 2014;78:27–37
DOI: 10.1159/000360623

Received: October 29, 2013
Accepted after revision: February 15, 2014
Published online: June 21, 2014

SNP-Based Linkage Analysis in Extended Pedigrees: Comparison between Two Alternative Approaches

Aude Saint-Pierre[a, b] Yuri D'Elia[b] Marina Ciullo[c] Peter P. Pramstaller[b, d, e] Cristian Pattaro[b]

[a]INSERM U1078, Etablissement Français du Sang, Brest, France; [b]Center for Biomedicine, European Academy of Bolzano (EURAC), Bolzano, Italy – affiliated Institute of the University of Lübeck, Lübeck, Germany; [c]Institute of Genetics and Biophysics 'A. Buzzati-Traverso', CNR Naples, Naples, Italy; [d]Department of Neurology, University of Lübeck, Lübeck, Germany; [e]Department of Neurology, Central Hospital, Bolzano, Italy

Key Words
Linkage analysis · Extended pedigrees · SNP

Abstract
Background: Linkage analysis on extended pedigrees is often challenged by the high computational demand of exact identity-by-descent (IBD) matrix reconstruction. When such an analysis becomes not feasible, two alternative solutions are contrasted: a full pedigree analysis based on approximate IBD estimation versus a pedigree splitting followed by exact IBD estimation. A multiple splitting (MS) approach, which combines linkage results across different splitting configurations, has been proposed to increase the power of single-split solutions. ***Methods:*** To assess whether MS can achieve a comparable power to a full pedigree analysis, we compared the power of linkage on a very large pedigree in both simulated and real-case scenarios, using variance components linkage analysis of a dense SNP array. ***Results:*** Our results confirm that the power to detect linkage is affected by the pedigree size. The MS approach showed higher power than the single-split analysis, but it was substantially less powerful than the full pedigree approach in both scenarios, at any level of significance and variance explained by a quantitative trait locus. ***Conclusion:*** The MS approach should always be preferred to analyses based on a single split but, when adequate computational resources are available, a full pedigree analysis is better than the MS analysis. Rather than focusing on how to best split a pedigree, it might be more valuable to identify computational solutions that can make the IBD estimation of dense-marker maps practically feasible, thus allowing a full pedigree analysis.

Introduction

Large genealogies can be very informative for the genetic mapping of complex traits. Compared to studies based on nuclear families, extended pedigrees generally yield a higher statistical power to identify quantitative trait loci (QTL) through linkage analysis [1]. However, deriving exact identity-by-descent (IBD) information for such extensive pedigrees is often not computationally feasible. To overcome this limitation, Markov Chain Monte Carlo (MCMC) methods have been developed that allow approximate IBD estimation at each marker [2].

KARGER

E-Mail karger@karger.com
www.karger.com/hhe

0001–5652/14/0781–0027$39.50/0

Aude Saint Pierre
INSERM U1078
Etablissement Français du Sang
FR–29218 Brest Cedex 2 (France)
E-Mail aude.saint-pierre@inserm.fr

An alternative way is to split the pedigree into smaller subpedigrees, thus enabling a fast and exact IBD calculation. The splitting procedure, generally based on kinship coefficient and family size, produces a simpler pedigree configuration [3]. Unfortunately, there is no uniformly best way to split a large pedigree when the genetic model is unknown. This aspect is not of negligible importance because linkage results are highly dependent on how the subpedigrees are chosen [4, 5]. To overcome the possibility of an unlucky choice of the splitting parameters and to capitalize on the sensitivity of the linkage results, Bellenguez et al. [6] proposed a multiple splitting (MS) approach, which systematically evaluates linkage on multiple splitting configurations rather than on a single one. When applied to the large Hutterite pedigree, this approach helped detect a genome-wide significant locus for asthma that would not have been detected with a single-split analysis [6].

Despite these promising results, a comprehensive assessment of the advantages and disadvantages of using the MS approach in place of the full pedigree analysis has not been proposed so far. It is unclear whether the MS approach can achieve a similar power as the full pedigree analysis. This would be important information because the MS approach requires a stronger control of the multiple testing issue. Furthermore, the two methods have a very different computational burden. With the MS approach, the IBD estimation is fast, thanks to the limited size of the subpedigrees, but the whole procedure requires intensive data handling at different stages: file preparation, management of the splitting configurations, and summary of the results. On the other hand, a full pedigree analysis requires less data handling, but the IBD estimation time can grow up to unacceptable limits when the number of markers is large. This aspect is particularly relevant when dense SNP arrays are considered. In addition to association studies, SNP arrays are now also routinely used for linkage analyses, thanks to the technical advantages and reduced genotyping costs compared to microsatellites [7]. SNP arrays are also attractive because of a more straightforward alignment of linkage and association signals. Major drawbacks of SNP-based linkage analyses are the need to account for the linkage disequilibrium (LD) structure and the high computational demand caused by the array density.

In the present work, we compared the performance of the MS and full pedigree workflows in the context of a population study based on an extended pedigree with linkage assessed on a typical SNP array.

Materials and Methods

Pedigree Data

The MICROS study is a population-based study carried out in three Alpine villages located in the region of South Tyrol (Italy). The study, described in detail elsewhere [8, 9], included 1,247 genotyped individuals, all connected together through a 16-generation pedigree including 48,197 subjects [10]. In practice, the analysis of all 16 generations is rarely needed. Like often done in linkage studies, we aimed to limit our analysis to an at least 5-generation pedigree. However, starting from the genotyped individuals, there were several ways to connect them through the complete genealogy. For this reason, we developed software to connect all individuals within a set of interest by minimizing the number of those who would not be informative for linkage (*Buildped* – available upon request).

Starting from the genotyped individuals, which already spanned 3 generations, we moved 4 generations up from each one, thus including up to 7 generations. From the individuals in the oldest generations, we then moved down by including all their descendants. With such parameters, the pedigree spanned at most 7 generations from the oldest to the most recent generation. Finally, siblings that were not informative for linkage analysis were dropped. The final pedigree had a maximum of 6 generations. This pedigree was used as the basis of the simulation presented below as well as of an application to a real case. However, the pedigree complexity and inbreeding loops greatly increased the computational cost of linkage analysis. For this reason, we limited the pedigree to only include individuals from one of the three villages for the simulation. The resulting pedigree is described in table 1 and corresponds, approximately, to one third of the available data. On the other hand, for the real-case scenario, the combined data from all three villages was used, resulting in a much more complex pedigree (see table 3).

Genotype Data

Participants were genotyped on an Illumina Infinium HumanHap300 v2 SNP bead microarray. In the present analysis, all 5,450 SNPs on chromosome 22 were considered. All SNPs underwent a three-step data cleaning and pruning procedure including general quality control (QC), Mendelian-inconsistency check, and LD pruning. All these operations were performed ahead of any manipulation of the pedigree, to prevent SNP selection bias due to subject subsetting. The QC step was performed with GenABEL [11]. SNPs with a call rate <98%, a minor allele frequency (MAF) <1%, and a Hardy-Weinberg equilibrium p value $\leq 10^{-6}$ were removed as well as samples with a call rate <98%. The QC left 5,381 SNPs available for analysis. Mendelian inheritance inconsistencies were detected with PedCheck [12]. Individual inconsistent variants were replaced with missing values. Pairwise LD was estimated with Haploview 4.2 [13] using a subset of individuals sharing a pairwise kinship coefficient of ≤0.1, and then minimized with MASEL [14], based on an r^2 threshold of 0.01.

Linkage Analysis Workflows

Linkage analyses of quantitative traits were performed based on a variance components (VC) method [15]. We denote with h^2 the genetic heritability, i.e. the polygenic contribution to the trait variation, and with h^2_{QTL} the QTL-specific heritability. Linkage is evaluated by comparing the likelihood of a model incorporating

Table 1. Characteristics of the pedigree used for simulation: full pedigree and the 30 split configurations

minK	ranS	Configuration No.	Families, n	Individuals, n	Bits per family[a]	Generations per family[a]	Genotyped individuals, n[b]	Informative individuals, n[b, c]
		Full pedigree	1	598	764	5	319	318
0.0125	2–6	1	61	857	11 (0–23)	3.6 (3–5)	420 (6.7)	318 (5.2)
	3–6	2	57	817	12 (3–22)	3.6 (3–5)	417 (7.3)	312 (5.5)
	4–6	3	53	797	13 (6–22)	3.8 (3–5)	394 (7.4)	308 (5.8)
	4–7	4	47	795	15 (4–24)	3.8 (3–5)	397 (8.4)	308 (6.5)
	4–8	5	43	738	15 (6–24)	3.8 (3–5)	388 (9.0)	308 (7.2)
	6–7	6	42	773	16 (8–23)	3.9 (3–5)	376 (9.0)	289 (6.9)
0.03125	2–6	7	64	782	10 (0–20)	3.3 (2–5)	416 (6.5)	316 (4.9)
	2–7	8	54	783	12 (0–22)	3.6 (2–5)	403 (7.5)	314 (5.8)
	2–8	9	49	738	13 (0–23)	3.6 (2–5)	395 (8.1)	316 (6.4)
	3–6	10	57	767	11 (5–19)	3.6 (2–5)	398 (7.0)	307 (5.5)
	3–7	11	51	749	13 (6–21)	3.7 (3–5)	396 (7.8)	311 (6.1)
	3–8	12	48	727	13 (4–24)	3.7 (3–5)	396 (8.2)	311 (6.5)
	3–9	13	43	705	15 (4–23)	3.7 (3–5)	388 (9.0)	311 (7.2)
	4–7	14	49	711	13 (5–20)	3.6 (3–5)	378 (7.7)	299 (6.1)
	4–8	15	44	676	14 (7–22)	3.7 (3–4)	373 (8.6)	304 (6.9)
	5–6	16	49	720	13 (8–19)	3.5 (3–5)	379 (7.7)	288 (5.9)
	5–8	17	39	638	15 (8–22)	3.7 (3–5)	360 (9.2)	289 (7.4)
	6–8	18	37	629	16 (9–23)	3.8 (3–4)	339 (9.3)	280 (7.7)
0.0625	2–10	19	52	613	10 (0–24)	3.0 (2–4)	365 (7.0)	311 (6.1)
	2–11	20	52	609	11 (2–22)	3.1 (2–4)	365 (7.0)	312 (6.0)
	2–9	21	56	641	10 (0–22)	3.0 (2–4)	384 (6.7)	311 (5.5)
	3–10	22	46	595	12 (4–24)	3.1 (2–4)	356 (7.7)	302 (6.6)
	3–7	23	53	631	11 (3–20)	3.1 (2–4)	365 (6.9)	300 (5.7)
	3–8	24	51	606	11 (4–17)	3.1 (2–4)	356 (7.0)	303 (5.9)
	3–9	25	49	614	11 (4–24)	3.1 (2–4)	363 (7.4)	302 (6.2)
	4–11	26	41	570	12 (5–23)	3.2 (3–4)	341 (8.3)	292 (7.1)
	5–10	27	37	543	14 (6–22)	3.2 (3–4)	332 (9.0)	279 (7.5)
	5–8	28	40	546	13 (6–20)	3.2 (3–4)	328 (8.2)	275 (6.9)
	5–9	29	38	534	13 (5–22)	3.2 (3–4)	324 (8.5)	278 (7.3)
	6–9	30	35	518	14 (6–20)	3.8 (3–4)	320 (9.1)	266 (7.6)

minK = Minimum kinship value allowed within a subpedigree after splitting; ranS = range of informative individuals in a family (minimum–maximum).

[a] Values are mean numbers (min.–max.).

[b] Values are total numbers (mean numbers per family).

[c] Genotyped and phenotyped individuals together.

both QTL and polygenic components against a purely polygenic model. Under the null hypothesis of no linkage, the likelihood ratio test statistic is assumed to follow a mixture of χ^2 distributions with 0 and 1 degree of freedom.

The analysis was performed according to the two workflows outlined in figure 1 and described below:

Workflow 1: Full Pedigree Analysis

Since the large pedigree size did not allow exact IBD estimations, multipoint IBD was inferred at each marker location using an MCMC approximation as implemented in Loki [2]. The subsequent linkage analysis was performed with SOLAR 4.3.1 [16].

Workflow 2: MS Analysis

We followed the procedure described by Bellenguez et al. [6] and implemented in the Cilento linkage parallel pipeline (http://www.igb.cnr.it/cilentoisolates/pages/research/software.php). The pipeline makes use of Jenti [3] to split the pedigree at different values of the minimum kinship level (minK; it denotes the minimum kinship between individuals in the subgroup of related informative individuals) and range size of subpedigrees (ranS; it denotes the minimum and maximum size of the subgroups of related informative individuals). Informative individuals were those who were both genotyped and phenotyped. A diversity of pedigree configurations was generated by combining the values of minK and ranS.

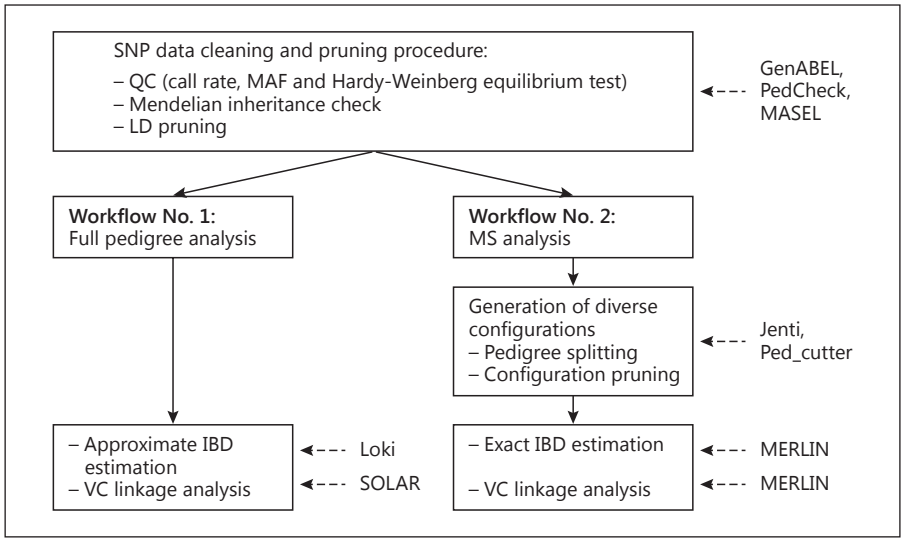

Fig. 1. The analysis workflows and related software.

We set minK to 6 possible levels: 0.0125, 0.03125, 0.0625, 0.125, 0.1875, and 0.25. The ranS parameter defined a family size range identified by a minimum and a maximum family size. The minimum value varied from 2 to 6 (5 integer values), and the maximum value varied from the minimum to 15 which corresponded to 60 ranS levels. Each combination of minK and ranS generated a single splitting configuration. The pedigree configurations maximizing the total number of informative individuals and meeting a family size constraint complexity of ≤24 bits were selected for linkage analysis (bit size = $2n - f$, with n being the number of non-founders and f being the number of founders). For each selected pedigree configuration, exact multipoint IBD estimation and VC linkage analyses were performed with MERLIN 1.1.2 [17]. At each marker location, the results of the multiple scans were summarized by taking the maximum (MaxLOD) and median (MedLOD) LOD scores across the configurations.

Simulation

To compare the power of the two workflows, we simulated 1,000 replicates of both genotypic and phenotypic values in the presence of linkage, using the pedigree of one of the three villages as template. To reduce the computational time, at each iteration, a different set of 4 consecutive markers was randomly selected from those on chromosome 22. A 5th bi-allelic QTL, with a MAF of 0.10, was simulated and placed between the 2nd and 3rd marker. We simulated a quantitative trait from a standard normal distribution. To resemble the typical situation often observed in quantitative blood parameters, we fixed $h^2 = 0.35$. Within these constraints, we let h^2_{QTL} vary from 0.05 to 0.30, resulting in a residual polygenic variance from 0.30 to 0.05. For each simulated QTL, the MS and full pedigree workflows were followed as described in figure 1, including SNP pruning, pedigree splitting, and linkage analysis.

Under the hypothesis of no linkage, the significance thresholds were derived based on the theoretical distribution of the LOD score statistic. The same theoretical distribution could not be used for the MaxLOD and MedLOD statistics in the MS approach. To this purpose, we performed 45,000 simulations under the hypothesis of no linkage. At each replicate, the marker genotypes were simulated through chromosome-wide gene dropping in the full pedigree with the Genedrop program, included in the MORGAN 3.0.3 package (http://www.stat.washington.edu/thompson/Genepi/pangaea.shtml), and the phenotype was simulated from a standard normal distribution. In this way, we obtained the null distributions of the LOD score for each configuration from which we derived the distributions of the MaxLOD and MedLOD statistics across all configurations. Significance thresholds corresponding to the two mentioned statistics were derived based on the respective 95th, 99th, and 99.9th percentiles.

Application to a Real-Case Scenario

To assess the performance of the two workflows in a real framework, we repeated the linkage analysis of serum creatinine (SCr) on chromosome 22, where we previously reported significant linkage using microsatellites, i.e. short-tandem repeats (STRs) [18]. This analysis was applied to the three villages together. This corresponds to a much more complex pedigree and a much higher number of genotyped and phenotyped individuals compared to the simulation study (online suppl. table 1; for all online suppl. material, see www.karger.com/doi/10.1159/000360623). For this reason, a new run of the splitting procedure was required. We repeated the linkage analysis using all 23 available STRs, whose map order and position were determined according to the deCODE Icelandic genetic map [for additional details, see 18]. Given that STRs are generally considered more (or at least not less) informative than SNPs for linkage analysis, we considered the STR-based results as a gold standard for comparison. In agreement with our previous study, SCr was normalized by inverse normal rank transformation and regressed on sex, age, and age^2 in a polygenic model. From the same analysis, we previously reported a LOD score of 2.68 on chromosome 22q13 using multipoint VC linkage analysis [18, table 3 therein]. The SNP-based analysis was performed using the same model as for the STRs and was based on the same SNP markers as described under 'Genotype Data' above. To maximize the comparability between STR- and SNP-based analyses, only individuals genotyped on both arrays were considered. The genotype data of individuals lacking one of the two kinds of markers was set

to missing, leaving 875 individuals available for analysis. The results presented hereafter are thus slightly different from those we presented before [18].

Previous literature suggests that SNP-based IBD estimation can be performed using MCMC sampling, but a high number of iteration is required to guarantee reliable estimates [19, 20]. For this reason, we ran Loki to reconstruct the IBD information based on 1,000,000 iterations while saving every tenth iteration and an initial burn-in of 10,000 iterations. STR and SNP map alignment was based on the Human Genome Build 36 assembly referred to the physical location of the STRs. The significance levels were assessed by 1,000 permutations of the phenotypic values within each family, as described by Churchill and Doerge [21]. Phenotypic values were permuted in the complete pedigree, i.e. before splitting, and then assigned to each individual according to the family configuration of the multiple splits.

We assessed the variability of heritability estimates according to different characteristics of the pedigree configurations. To this aim, we estimated the genetic heritability (h^2) for each splitting configuration from which we inferred the distribution of h^2 conditional on the minK levels as well as the distribution of h^2 conditional on 4 specific characteristics of pedigree configurations: the total number of individuals and the numbers of informative individuals, of genotyped individuals, and of phenotyped individuals. For each referred characteristic, the configurations were classified into 4 categories according to the quartiles of the respective distribution. We assessed the variation of h^2 over the configurations in each category. Then, we compared the distribution of h^2 between categories using a Kruskal-Wallis test. To assess the concordance of the results obtained with each individual split, we used Lin's concordance correlation coefficient [22] applied to the LOD score statistics. This allowed us to assess whether the concordance between different configurations was related to the type of markers used (SNPs or STRs). Concordance coefficients between configurations were compared by means of a Wilcoxon rank-sum test in each set of markers.

Table 2. Power to detect linkage: comparison between the two workflows at different levels of variance explained by the QTL (h^2_{QTL}), the residual polygenic effect (h^2–h^2_{QTL})[a], and the type 1 error rate (α = 5, 1, and 0.1%)

h^2_{QTL}	h^2 – h^2_{QTL}	Type of analysis	Summary statistics	Type 1 error rate, %		
				5%	1%	0.1%
5	30	Full pedigree		20.20	7.60	1.10
		MS	MaxLOD	11.70	2.80	0.30
			MedLOD	13.30	3.70	0.50
10	25	Full pedigree		43.10	22.80	7.40
		MS	MaxLOD	26.40	8.50	2.10
			MedLOD	30.10	11.60	3.00
20	15	Full pedigree		80.00	64.30	38.80
		MS	MaxLOD	57.20	33.60	13.30
			MedLOD	66.10	41.80	19.20
30	5	Full pedigree		92.80	86.60	70.10
		MS	MaxLOD	80.50	62.10	35.90
			MedLOD	86.20	72.00	46.50

[a] For each model, the proportion of variance explained by the residual component was set to 65%. Empirical power was derived at the QTL position and based on 1,000 replicates.

Results

Simulation

The pedigree used for simulation, comprising a single 5-generation family of 598 individuals (bit size = 764), is summarized in table 1. The number and type of relative pairs before splitting are reported in online supplementary table 1. The 5,381 SNPs used for simulation were characterized by a mean interSNP distance of 0.006 Mb (SD = 0.012), a median MAF of 0.25, and a median r^2 of 0.019. No SNPs were removed by the Mendelian-inconsistency check because inconsistent values were replaced with missing values. The subsequent LD pruning step selected a final subset of 176 (3.3%) SNPs, with a mean interSNP distance of 0.153 Mb (SD = 0.142), a median MAF of 0.45, and a median r^2 of 0.002.

Through the splitting procedure, we obtained 6 pedigree configurations when minK = 0.0125, 12 configurations when minK = 0.03125, and 12 configurations when minK = 0.0625, resulting in a total of 30 different configurations (table 1). Compared to the complete pedigree, the 30 configurations were characterized by a higher number of genotyped individuals but a lower number of informative individuals (online suppl. fig. 1). This is due to the splitting procedure which led to the duplication of some individuals, whose genotypes, but not phenotypes, were retained to improve IBD estimation. While the total number of informative individuals decreased with increasing minK values, the mean proportion of informative individuals per family slightly increased with the minK value. This could be expected because lower minK values result in larger pedigrees of more distantly related informative individuals, with a consequent inclusion of more noninformative ancestors in their genealogy.

We derived empirical significance thresholds for the LOD score statistic, under the null hypothesis of no linkage, for the full pedigree analysis, for each single pedigree configuration, and for the MaxLOD and MedLOD summary statistics accounting for multiple testing (online suppl. table 2). The empirical thresholds estimated in the full pedigree analysis and in each single pedigree configuration were close to the expected asymptotic values. These results showed that no inflation of the linkage statistic,

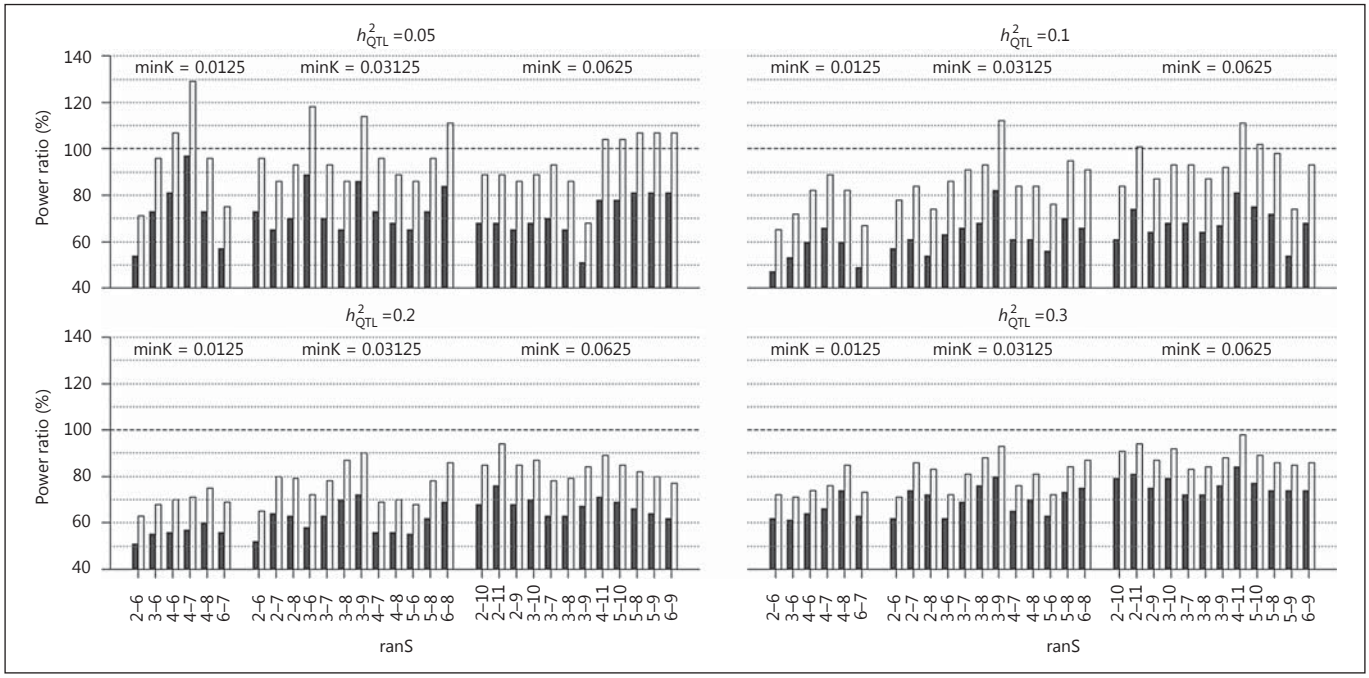

Fig. 2. Ratio between the power of each single configuration and that of the MS analysis when the MedLOD (dark gray) or MaxLOD (light gray) summary statistics are considered. The results are displayed for α = 1% at different h^2_{QTL} levels and are grouped by minK and ranS values, following the order in table 1.

due to residual LD between SNPs, remained in our data after the SNP selection procedure. The significance thresholds for the MaxLOD statistic were 1.443, 2.072, and 2.940 at α = 5, 1, and 0.1%, respectively, and those for the MedLOD statistic were equal to 0.338, 0.669, and 1.209 at α = 5, 1 and 0.1%, respectively. Power estimates based on these thresholds are reported in table 2.

As expected, the power increased with the proportion of variance explained by the QTL. For all the investigated genetic models, the full pedigree analysis results had a higher power than those of the MS analysis. When α = 1% and h^2_{QTL} = 20%, the power to detect linkage was 64.3% in the full pedigree analysis, while it was only 33.6 and 41.8% for the MaxLOD and MedLOD statistics, respectively. The power gain of the full pedigree analysis over the MS approach increased with decreasing values of h^2_{QTL} (online suppl. fig. 2). Figure 2 shows the ratio between the power of single-split analyses over that of the MS analysis. When α = 1% and h^2_{QTL} = 5%, the power of single-split analyses was 51–97% of the power of a MS analysis based on the MedLOD statistics (median = 72%) and 68–129% of the power of a MS analysis based on the MaxLOD (median = 95%). Similar power losses were observed at h^2_{QTL} = 10% (median power ratio = 64% for the MedLOD and 87% for the MaxLOD), at h^2_{QTL} = 20% (median power ratio = 63% for the MedLOD and 79% for the MaxLOD), at h^2_{QTL} = 30% (median power ratio = 73% for the MedLOD and 85% for the MaxLOD). When considering the MedLOD statistics, which had higher power than the MaxLOD, the MS approach was always more powerful than the single-split analysis. No clear power reduction trends have been observed at different minK, ranS, and h^2_{QTL} levels. Similar results were obtained at different α levels.

Real-Case Scenario

We also compared the full pedigree and MS workflows in a real-case scenario. The characteristics of the complete pedigree are shown in table 3. The number and type of relative pairs before splitting are described in online supplementary table 1. A total of 37 families, spanning from 2 to 6 generations and totaling 2,219 individuals, were informative for linkage, i.e. included at least two informative individuals. Ninety percent of the individuals were clustered in only 3 families: one family with 175 founders (693 bits; mean kinship ± SD = 0.09 ± 0.08), one with 238 founders (880 bits; mean kinship ± SD = 0.08 ± 0.07), and one with 144 founders (764 bits; mean kinship ± SD = 0.07 ± 0.07). A total of 1,067 individuals (43.9% males)

Table 3. Characteristics of the pedigree used in the real-case scenario: full pedigree and the 23 split configurations

minK	ranS	Configuration No.	Families, n	Individuals, n	Bits per family[a]	Generations per family[a]	Genotyped individuals, n[b]	Informative individuals, n[b, c]
		Full pedigree	37	2,219	67 (0–880)	2.6 (2–6)	875 (23.6)	847 (22.9)
0.0125	2–5	1	214	2,259	8 (0–24)	3.3 (2–5)	1,026 (4.8)	812 (3.8)
	2–6	2	202	2,208	8 (0–23)	3.2 (2–5)	1,020 (5.0)	821 (4.1)
	3–5	3	191	2,211	8 (0–20)	3.4 (2–5)	990 (5.2)	776 (4.1)
	3–6	4	171	2,112	9 (0–20)	3.4 (2–5)	978 (5.7)	777 (4.5)
	4–5	5	174	2,042	9 (0–21)	3.4 (2–5)	925 (5.3)	732 (4.2)
0.03125	2–6	6	204	2,028	7 (0–21)	3.1 (2–5)	1,011 (5.0)	812 (4.0)
	2–7	7	191	2,023	8 (0–20)	3.2 (2–5)	996 (5.2)	810 (4.2)
	3–6	8	178	1,996	9 (0–24)	3.3 (2–5)	969 (5.4)	772 (4.3)
	3–7	9	167	1,917	9 (0–21)	3.3 (2–5)	945 (5.7)	770 (4.6)
	3–8	10	157	1,822	10 (0–24)	3.3 (2–5)	924 (5.9)	770 (4.9)
	4–6	11	154	1,790	9 (0–24)	3.3 (2–5)	879 (5.7)	717 (4.7)
	5–7	12	127	1,564	10 (0–22)	3.3 (2–5)	779 (6.1)	651 (5.1)
	6–7	13	112	1,433	11 (0–24)	3.3 (2–5)	708 (6.3)	589 (5.3)
0.0625	2–7	14	207	1,685	6 (0–16)	2.7 (2–5)	937 (4.5)	802 (3.9)
	2–9	15	197	1,683	7 (0–23)	2.7 (2–5)	927 (4.7)	801 (4.1)
	2–10	16	195	1,652	7 (0–24)	2.7 (2–5)	917 (4.7)	804 (4.1)
	2–11	17	195	1,682	7 (0–24)	2.7 (2–5)	935 (4.8)	806 (4.1)
	3–7	18	170	1,596	8 (0–24)	2.9 (2–5)	887 (5.2)	739 (4.3)
	3–8	19	165	1,565	8 (0–22)	2.9 (2–5)	862 (5.2)	743 (4.5)
	3–9	20	163	1,566	8 (0–21)	2.9 (2–5)	864 (5.3)	745 (4.6)
	3–10	21	161	1,546	8 (0–22)	2.9 (2–5)	864 (5.4)	747 (4.6)
	4–9	22	138	1,389	9 (0–20)	2.9 (2–5)	784 (5.7)	681 (4.9)
	4–10	23	136	1,373	9 (0–24)	2.9 (2–5)	781 (5.7)	680 (5.0)

minK = Minimum kinship value allowed within a subpedigree after splitting; ranS = range of informative individuals in a family (minimum–maximum).

[a] Values are mean numbers (min.–max.).

[b] Values are total numbers (mean numbers per family).

[c] Genotyped and phenotyped individuals together.

were phenotyped for SCr. The mean SCr level was 0.88 mg/dl (SD = 0.17), it was higher in males (0.98 ± 0.15 mg/dl) than in females (0.80 ± 0.17 mg/dl). The age distribution was similar between males (46 ± 17 years) and females (46 ± 18 years). Similar characteristics were observed in the 847 informative individuals.

The SNP cleaning and pruning step left a subset of 325 SNPs (6.0%) on chromosome 22 with an average marker spacing of 0.081 Mb (SD = 0.091). The median MAF was lower before (0.26) than after the cleaning procedure (0.41). Most of the LD was removed with the SNP selection procedure by MASEL (median $r^2 = 0.002$). The selected SNPs had a mean heterozygosity of 0.418.

Of the 37 families composing this pedigree, only 4 had a bit size of 24 or more (including the 3 largest families described above). These 4 families were the only ones requiring pedigree splitting. The other 33 families were smaller than 24 bits and therefore remained unchanged across all splitting configurations. We obtained a total of 23 configurations: 5 with minK = 0.0125, 8 with minK = 0.03125, and 10 with minK = 0.0625 (table 3). Similar to what was observed in the simulation study, the MS configurations in the real-case scenario were also characterized by a higher number of genotyped and a lower number of informative individuals as compared to the full pedigree. The number of genotyped and informative individuals decreased when the minK values increased (online suppl. fig. 3).

The genetic heritability of SCr in the full pedigree was $h^2 = 47.43 \pm 0.06$. For the sake of comparison, we also estimated the heritability within each single split using SOLAR. The median h^2 across the multiple splits was 57.69,

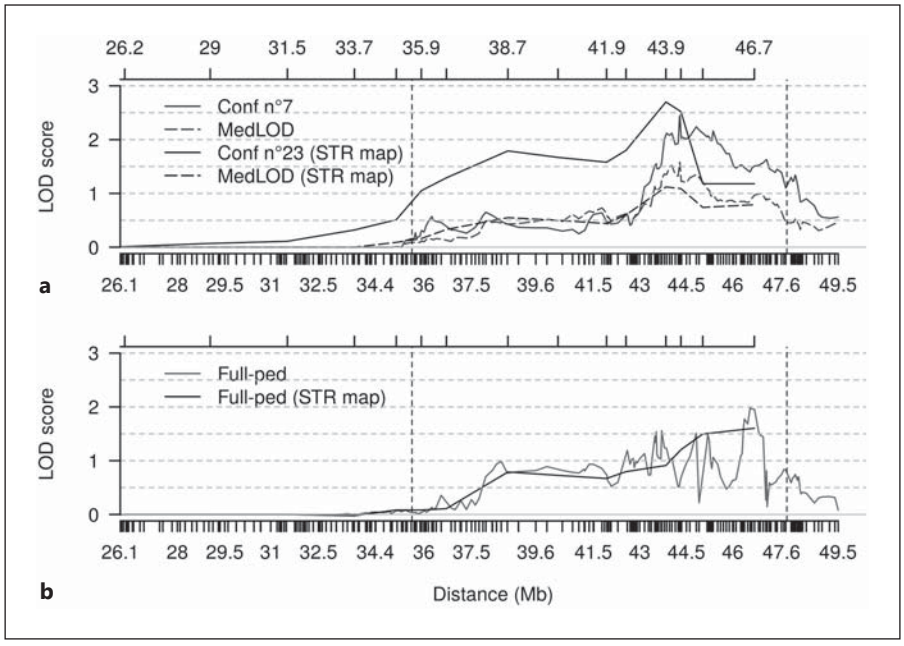

Fig. 3. Linkage analysis of SCr on chromosome 22 based on the MS approach (**a**) and full pedigree analysis (**b**). STR and SNP positions are shown on the top and bottom horizontal axes, respectively. In **a**, solid and dashed lines identify the best and median configurations, respectively. Vertical dashed lines indicate the reference 1-LOD region from our previous analysis [18]. Full-ped = full pedigree.

but the results varied depending on the minK and ranS values: h^2 varied from 44.75 ± 0.09 when minK = 0.03125 and ranS = 6–7 to 71.47 ± 0.09 when minK = 0.03125 and ranS = 4–6. Overall, the mean heritability slightly decreased when the minK value increased (online suppl. fig. 4), but no significant effect was observed. On the other hand, h^2 varied with the quantile distribution of individuals (online suppl. fig. 5). It increased when the number of individuals increased even though the variation was not significant across quantiles (Kruskal-Wallis test p value = 0.067, 0.353, and 0.520 for the total group of individuals, genotyped individuals, and informative individuals, respectively). Sex, age, and age^2 accounted for 29.39% of the total phenotypic variance in the full pedigree and varied from 28.29 to 30.73% in the MS.

Results of the SNP-based MS and full pedigree analyses compared to the STR gold standard are shown in figure 3. These new results were in concordance with our previous ones [18]. The MaxLOD of the STR analysis was always included within the 1-LOD linkage region (from 35.59 to 47.79 Mb) in Pattaro et al. [18]. With the MS approach (fig. 3a), the MaxLOD of the STR-based analysis was 2.70 (p = 0.009) and was observed at 43.85 Mb with configuration No. 23 (black solid line). A peak at the same position (LOD score 1.12, p = 0.014) was observed with the MedLOD statistic (black dashed line). The SNP-based MS analysis identified a similar pattern to STRs for both the MaxLOD and MedLOD statistics which showed linkage peaks at 44.31 Mb, with LOD scores of 2.45 (p = 0.052) and 1.58 (p = 0.06), respectively.

In the full pedigree analysis (fig. 3b), the MaxLOD of the STR-based analysis was observed at 46.73 Mb (LOD score = 1.62, p = 0.024). The MaxLOD across the chromosome of the SNP analysis was higher than that obtained with STRs and was detected at a similar position (LOD score = 1.99 at 46.59 Mb, p = 0.026). Another four not significant peaks, located between 43.54 and 45.22 Mb, were clearly identified with LOD scores >1.50.

Despite the visual impression, it is worth noting that the full pedigree approach outperformed the MS approach in terms of statistical significance. This was expected because the MS analysis undergoes a multiple testing penalty to account for the different configurations. On the other hand, the calculation time of the MS analysis was significantly shorter than that of the full pedigree analysis. The analysis was run on an Opteron 8356 Quad Core (2.3 GHz) CPU and took 28 h with the MS workflow and 120 h with the full pedigree workflow, i.e. the MS workflow was more than 4 times faster than the full pedigree one.

By focusing on the MS workflow, we additionally assessed the variability of linkage results when the analysis is based on STRs rather than on SNPs. Figure 4 shows the LOD score distribution for each subpedigree configuration with both kinds of markers. In each panel we highlighted all configurations obtained with the same minK

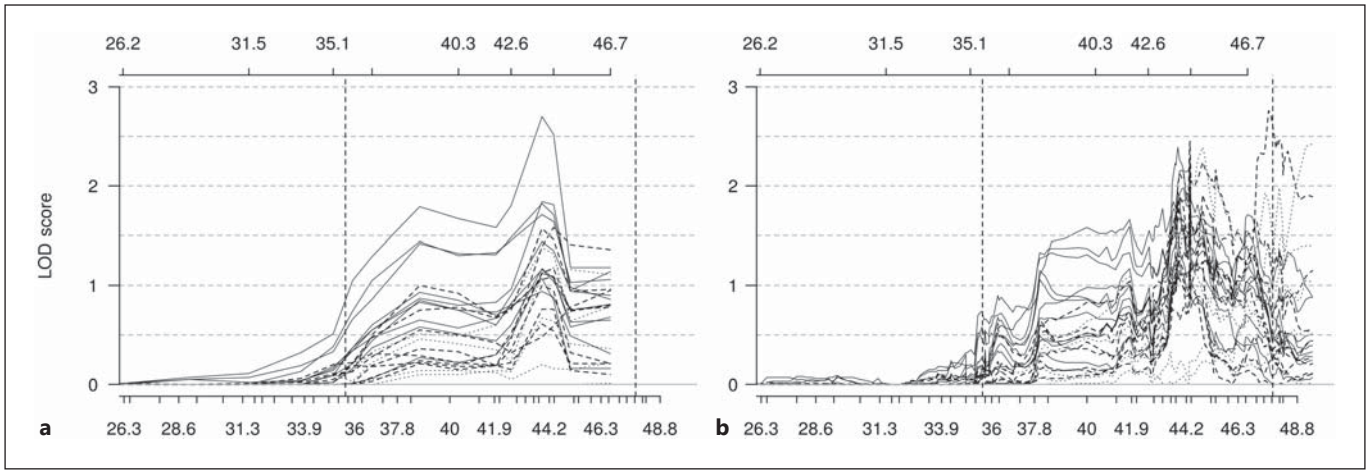

Fig. 4. Linkage analysis of SCr on chromosome 22 with the MS analysis to highlight the variability between the results from each pedigree configuration based either on STR (**a**) or SNP (**b**) arrays. The STR and SNP positions are shown on the top and bottom horizontal axes, respectively. Dashed lines indicate the results obtained from pedigree configurations based on a minK value of 0.03125, dotted lines correspond to minK = 0.0125, and solid lines to minK = 0.0625. The two vertical dashed lines identify the reference 1-LOD region from our previous analysis [18].

value (at varying ranS values). The LOD score distributions for each particular minK are highlighted in color in online supplementary figure 6. The STR-based results showed a smoothed pattern compared to the SNP-based results, due to the lower marker density. Moreover, while in the STR-based analysis there seemed to be a trend associating the LOD score magnitude with the minK values (the LOD scores were higher for all configurations obtained with a minK value of 0.0625), this was not the case on the SNP map. In fact, the LOD score of the SNP-based analysis showed maximal peaks under any value of the minK coefficient. In figure 4, the STR-based results seemed more similar across the configurations than the SNP-based results did. But this was just a visual impression. In fact, the concordance level between the different splitting configurations was slightly higher for the SNP-based (0.69 ± 0.25) than for the STR-based analysis (SD = 0.67 ± 0.28). However, the difference was not significant (Wilcoxon rank-sum test p value = 0.59).

Discussion

We compared the power of two alternative approaches to identify the presence of linkage when VC linkage analysis is concerned. Results from the simulation revealed that, when computationally feasible, the analysis of the complete pedigree should always be preferred to the MS approach. This result was confirmed by a real-case analysis. However, when splitting is not avoidable for computational reasons, the MS approach was definitely more powerful than considering a single split. The power increase was so relevant that using a single split can come to the false conclusion of an absence of linkage. Even if this approach requires a higher time management than using a single split, the increased complexity should not discourage from applying the MS approach.

More specifically, our simulation showed that a substantial amount of linkage information is lost when the pedigree is broken into subpedigrees. The advantage of the full pedigree analysis remains also when the results from the multiple splits are combined through summary statistics, such as the MaxLOD or the MedLOD at each marker across the configurations. Of these two summary statistics, the MedLOD guarantees higher power than the MaxLOD statistic. The advantage of the full pedigree over the MS analysis is especially remarkable if one considers that IBD-sharing estimates in the full pedigree must be inferred stochastically through MCMC methods [2], whilst the splitting approach can exploit exact IBD estimation.

A limitation of the MS approach is that there are no significance thresholds for the summary statistics that are universally valid, since they depend on the number of configurations. Significance can be determined using permutation methods, as we did in this study, or with

more sophisticated methods, like for example the local false discovery rate [23].

MCMC methods extend the feasibility of linkage analysis with regard to the complexity of a pedigree that can be handled while leaving it intact and with regard to the number of loci that can be analyzed jointly. However, this increased complexity comes at a cost of increased computational time. In the full pedigree analysis, the MCMC-based method for IBD estimation was more time intensive than the MS approach. Our analyses were based on a very stringent r^2 statistics cutoff for LD pruning, leading to a relatively small subset of SNPs. Our results revealed that the MCMC-based method for IBD estimations would not be feasible when the number of markers is too large. On the other hand, a further reduction of the number of markers, by using even stricter LD thresholds, would wash out the power to an unacceptably lower level. An alternative approach to reduce the computational time of IBD estimations when exploiting SNP arrays has been proposed by Day-Williams et al. [24]. The authors, aiming to infer relatedness directly from whole genome data on SNPs rather than using pedigree information, developed fast algorithms to estimate the kinship coefficient and the IBD matrix. This IBD matrix can then be used in classical linkage software.

Our application to a real-case scenario consistently demonstrated the advantage of the full pedigree approach over the MS approach, which in turn is more advantageous than a single-split approach. Using SNP markers, we could replicate a previously reported QTL for SCr [18]. Due to the higher marker density on the SNP map relative to the microsatellite map, the linkage region was narrowed down. However, the empirical comparison between MS and full pedigree analyses highlighted that different conclusions could be the result of considering one of the two approaches, with the maximum signal observed approx. 3 Mb apart between the two analyses. Caution should be taken when aiming to refine a linkage region for further analysis.

To validate our procedure, both workflows were applied to a STR-marker panel. Our results were in concordance with our previous analysis [18]. As expected, due to the lower marker density, the STR-based results showed a smoother pattern compared to the SNP-based results. In the STR-based analysis and conversely to the SNP-based analysis, the MaxLOD were observed for only one magnitude of minK. This would suggest that for low minK coefficients, increasing the marker density might add some information in regions where the STR-marker panels are poorly informative for linkage. We noted that, despite the visual impression, the concordance level between the configurations was slightly higher for SNPs than for STRs. However, these results are based on a single analysis and cannot be generalized to other contexts.

A limitation of our simulation and analyses is that they were based on a complex genealogy with a specific structure that can be different from pedigrees from other studies. However, even if the variety of human genealogies and disease models should prevent from systematic rules, our general conclusion seems to hold more generally. In fact, our results are congruent with those reported by Bellenguez et al. [6]. The MS is a powerful approach to detect linkage that could be missed if only a single split is considered. Our results also confirmed that linkage analysis on the full pedigree is more powerful than performing linkage on smaller broken pedigrees [1]. Critical for the reconstruction of the IBD information in the extended pedigree is the choice of software used. We cannot exclude that other software may provide more accurate IBD results than Loki. It has been shown, for example, that the MORGAN package can also achieve highly accurate results when using large and dense SNP sets [25].

Overall, there is a trade-off between increasing the pedigree complexity and the calculation time needed for IBD-sharing estimations. For practical applications, when the pedigree is complex and the number of markers is large, it might be computationally too cumbersome to run a genome- or even a chromosome-wide linkage analysis without splitting the pedigree. Solutions to overcome such limitations include the possibility to break the analysis into smaller chromosomal segments (chunks) and to join the results after the calculation, by paying attention to the region(s) crossing the chunk boundaries. Another option is to parallelize the calculation on multicore machines or to exploit the possibilities offered by cloud computing [26]. Finally, a two-stage approach might also be considered, i.e. to apply the MS approach for chromosome-wide scans and the full pedigree analysis to refine selected regions. Given the lower power of the MS approach, this solution would require running the MS analysis based on liberal thresholds. Regions of interest could then be followed up, even at higher marker density, with the full pedigree approach using stricter significance levels. However, this two-stage approach should undergo additional evaluation to assess its effective power and to quantify the percent of false-positive regions which may be submitted to follow-up.

Acknowledgments

We thank Fabio Marroni (Istituto di Genetica Applicata, Udine) for the helpful discussion and Daniel Taliun (European Academy of Bolzano) for his analytical support. The study was supported by the Ministry of Health of the Autonomous Province of Bolzano and by the South Tyrolean Sparkasse Foundation.

Disclosure Statement

The authors declare that there are no conflicts of interest.

References

1 Dyer TD, Blangero J, Williams JT, Göring HH, Mahaney MC: The effect of pedigree complexity on quantitative trait linkage analysis. Genet Epidemiol 2001;21(suppl 1):S236–S243.
2 Heath SC: Markov chain Monte Carlo segregation and linkage analysis for oligogenic models. Am J Hum Genet 1997;61:748–760.
3 Falchi M, Fuchsberger C: Jenti: an efficient tool for mining complex inbred genealogies. Bioinformatics 2008;24:724–726.
4 Chapman NH, Leutenegger AL, Badzioch MD, Bogdan M, Conlon EM, Daw EW, Gagnon F, Li N, Maia JM, Wijsman EM, Thompson EA: The importance of connections: joining components of the Hutterite pedigree. Genet Epidemiol 2001;21(suppl 1):S230–S235.
5 Ciullo M, Bellenguez C, Colonna V, Nutile T, Calabria A, Pacente R, Iovino G, Trimarco B, Bourgain C, Persico MG: New susceptibility locus for hypertension on chromosome 8q by efficient pedigree-breaking in an Italian isolate. Hum Mol Genet 2006;15:1735–1743.
6 Bellenguez C, Ober C, Bourgain C: A multiple splitting approach to linkage analysis in large pedigrees identifies a linkage to asthma on chromosome 12. Genet Epidemiol 2009;33:207–216.
7 Evans DM, Cardon LR: Guidelines for genotyping in genomewide linkage studies: single-nucleotide-polymorphism maps versus microsatellite maps. Am J Hum Genet 2004;75:687–692.
8 Pattaro C, Marroni F, Riegler A, Mascalzoni D, Pichler I, Volpato CB, Dal Cero U, De Grandi A, Egger C, Eisendle A, Fuchsberger C, Gögele M, Pedrotti S, Pinggera GK, Stefanov SA, Vogl FD, Wiedermann CJ, Meitinger T, Pramstaller PP: The genetic study of three population microisolates in South Tyrol (MICROS): study design and epidemiological perspectives. BMC Med Genet 2007;8:29.
9 Marroni F, Grazio D, Pattaro C, Devoto M, Pramstaller P: Estimates of genetic and environmental contribution to 43 quantitative traits support sharing of a homogeneous environment in an isolated population from South Tyrol, Italy. Hum Hered 2008;65:175–182.
10 Gögele M, Pattaro C, Fuchsberger C, Minelli C, Pramstaller PP, Wjst M: Heritability analysis of life span in a semi-isolated population followed across four centuries reveals the presence of pleiotropy between life span and reproduction. J Gerontol A Biol Sci Med Sci 2011;66:26–37.
11 Aulchenko YS, Ripke S, Isaacs A, van Duijn CM: GenABEL: an R library for genome-wide association analysis. Bioinformatics 2007;23:1294–1296.
12 O'Connell JR, Weeks DE: PedCheck: a program for identification of genotype incompatibilities in linkage analysis. Am J Hum Genet 1998;63:259–266.
13 Barrett JC, Fry B, Maller J, Daly MJ: Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 2005;21:263–265.
14 Bellenguez C, Ober C, Bourgain C: Linkage analysis with dense SNP maps in isolated populations. Hum Hered 2009;68:87–97.
15 Amos CI: Robust variance-components approach for assessing genetic linkage in pedigrees. Am J Hum Genet 1994;54:535–543.
16 Almasy L, Blangero J: Multipoint quantitative-trait linkage analysis in general pedigrees. Am J Hum Genet 1998;62:1198–1211.
17 Abecasis GR, Cherny SS, Cookson WO, Cardon LR: Merlin – rapid analysis of dense genetic maps using sparse gene flow trees. Nat Genet 2002;30:97–101.
18 Pattaro C, Aulchenko YS, Isaacs A, Vitart V, Hayward C, Franklin CS, Polasek O, Kolcic I, Biloglav Z, Campbell S, Hastie N, Lauc G, Meitinger T, Oostra BA, Gyllensten U, Wilson JF, Pichler I, Hicks AA, Campbell H, Wright AF, Rudan I, van Duijn CM, Riegler P, Marroni F, Pramstaller PP, EUROSPAN Consortium: Genome-wide linkage analysis of serum creatinine in three isolated European populations. Kidney Int 2009;76:297–306.
19 Daw EW, Heath SC, Lu Y: Single-nucleotide polymorphism versus microsatellite markers in a combined linkage and segregation analysis of a quantitative trait. BMC Genet 2005;6(suppl 1):S32.
20 Hinrichs AL, Bertelsen S, Bierut LJ, Dunn G, Jin CH, Kauwe JS, Suarez BK: Multipoint identity-by-descent computations for single-point polymorphism and microsatellite maps. BMC Genet 2005;6(suppl 1):S34.
21 Churchill GA, Doerge RW: Empirical threshold values for quantitative trait mapping. Genetics 1994;138:963–971.
22 Lin LI: A concordance correlation coefficient to evaluate reproducibility. Biometrics 1989;45:255–268.
23 Ruggiero D, Dalmasso C, Nutile T, Sorice R, Dionisi L, Aversano M, Bröet P, Leutenegger AL, Bourgain C, Ciullo M: Genetics of VEGF serum variation in human isolated populations of Cilento: importance of VEGF polymorphisms. PLoS One 2011;6:e16982.
24 Day-Williams AG, Blangero J, Dyer TD, Lange K, Sobel EM: Linkage analysis without defined pedigrees. Genet Epidemiol 2011;35:360–370.
25 Wijsman EM, Rothstein JH, Thompson EA: Multipoint linkage analysis with many multiallelic or dense diallelic markers: Markov chain-Monte Carlo provides practical approaches for genome scans on general pedigrees. Am J Hum Genet 2006;79:846–858.
26 Silberstein M, Weissbrod O, Otten L, Tzemach A, Anisenia A, Shtark O, Tuberg D, Galfrin E, Gannon I, Shalata A, Borochowitz ZU, Dechter R, Thompson E, Geiger D: A system for exact and approximate genetic linkage analysis of SNP data in large pedigrees. Bioinformatics 2013;29:197–205.