

ARTICLE

Predicting the number and sizes of IBD regions among family members and evaluating the family size requirement for linkage studies

Wanling Yang^{*,1}, Zhanyong Wang², Lusheng Wang², Pak-Chung Sham³, Peng Huang⁴ and Yu Lung Lau^{*,1}

¹Department of Paediatrics and Adolescent Medicine, The University of Hong Kong, Hong Kong, China; ²Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong, China; ³Genome Research Centre, LKS Faculty of Medicine, The University of Hong Kong, Hong Kong, China; ⁴Department of Biostatistics, Bioinformatics and Epidemiology, Medical University of South Carolina, Charleston, SC, USA

With genotyping of high-density single nucleotide polymorphisms (SNPs) replacing that of microsatellite markers in linkage studies, it becomes possible to accurately determine the genomic regions shared identity by descent (IBD) by family members. In addition to evaluating the likelihood of linkage for a region with the underlining disease (the LOD score approach), an appropriate question to ask is what would be the expected number and sizes of IBD regions among the affecteds, as there could be more than one region reaching the maximum achievable LOD score for a given family. Here, we introduce a computer program to allow the prediction of the total number of IBD regions among family members and their sizes. Reversely, it can be used to predict the portion of the genome that can be excluded from consideration according to the family size and user-defined inheritance mode and penetrance. Such information has implications on the feasibility of conducting linkage analysis on a given family of certain size and structure or on a few small families when interfamily homogeneity can be assumed. It can also help determine the most relevant members to be genotyped for such a study. Simulation results showed that the IBD regions containing true mutations are usually larger than regions IBD due to random chance. We have made use of this feature in our program to allow evaluation of the identified IBD regions based on Bayesian probability calculation and simulation results.

European Journal of Human Genetics (2008) 16, 1535–1543; doi:10.1038/ejhg.2008.116; published online 25 June 2008

Keywords: linkage analysis; simulation; family size; autozygosity mapping; remote cousins; SNP

Introduction

Classical linkage analysis often relies on genotyping of microsatellite markers. Although these markers are usually

multi-allelic and more informative than biallelic single nucleotide polymorphisms (SNPs), they are harder to genotype, and sparse compared to SNPs. Classical linkage analysis programs based on maximum likelihood estimation usually can only extract partial inheritance information.^{1–3} On the other hand, SNPs are abundant⁴ and can be readily genotyped in a high-throughput format with reduced time and cost. Although SNPs are less informative than microsatellite markers, they can be analyzed as haplotypes, which allow accurate extraction of inheritance

*Correspondence: Dr Wanling Yang or Professor Yu Lung Lau, Department of Paediatrics and Adolescent Medicine, L7-55, Laboratory Block, Li Ka Shing Faculty of Medicine, The University of Hong Kong, 21 Sassoon Rd., Hong Kong, Tel: +852-2819-9357; Fax: +852-2819-8142; E-mail: yangwl@hkucc.hku.hk, lauylung@hkucc.hku.hk
Received 22 November 2007; revised 6 May 2008; accepted 29 May 2008; published online 25 June 2008

information. Multiple studies have shown that linkage studies using high-density SNPs can extract more inheritance information than using microsatellite markers.^{2,5,6} In fact, based on an extensive simulation using Affymetrix GeneChip data for families of different structures, we have demonstrated that allele sharing status among individuals in a family can be determined with accuracy.⁷ Therefore, advances in genotyping technologies and analysis tools can help extracting full inheritance information in linkage studies.

The LOD score approach evaluates the likelihood that a genomic region is linked to the disease under study *versus* the likelihood that the two do not cosegregate. A LOD score of 3 is used as a threshold for evidence of linkage. However, it is known that a region with a high LOD score does not necessarily mean that it contains the causal mutation. For example, there could be multiple regions reaching a LOD score of 3 in a family of Mendelian inheritance, and only one region may contain the causal mutation assuming no intra-family heterogeneity. On the other hand, for a smaller family with a maximally achievable LOD score of less than 3, if high density markers are genotyped and the IBD regions among family members are unequivocally determined genome-wide, it could be sure that one of those regions contains the causal mutation. So, with the technology advancement and IBD regions determined, without doubt, an appropriate question to ask is how many such regions are expected to be identified and what would be their likely sizes. This information has implications on whether a linkage study is practical allowing reasonable amount of work for the subsequent physical mapping process.

Theoretical explorations have been carried out on the issue of IBD sharing among relatives over the past 20 years.^{8–10} However, to laypersons on statistical genetics, a simple, intuitive method on the calculation and prediction of IBD regions among family members is still lacking. Here, we have developed a computer program that simulates meiosis recombination and inheritance processes, which can be used to predict the number and sizes of IBD regions among family members. Furthermore, the program can also be used to explore the family size requirement for linkage studies, for example, in terms of the portion of the genome that can be excluded from consideration given the user-defined inheritance mode. The computer program provides a useful tool for physicians and geneticists to evaluate the sufficiency of their family collections for linkage studies and to choose the most informative members to be genotyped to minimize cost and maximize information gain.

For families of small to intermediate sizes, in addition to the true mutation region, it is inevitable that multiple regions IBD due to random chance will also be found, and it will be difficult to decide which region(s) to pursue for physical gene mapping or which one to pursue first.

Although the candidate genes in the IBD regions can be analyzed *in silico* to help the decision-making, in this study, we clearly show that there is a significant difference in the size distribution between the IBD regions containing true mutations and the regions IBD due to random chance. Our program calculates the Bayesian probability of an IBD region as being mutation-containing or not based on simulation results. It provides a useful tool to predict the most likely mutation-containing region to help users in the physical mapping processes.

Methods

Exclusion power for autosomal dominant (AD) inheritance

With high-density SNPs genotyped as genetic markers, IBD regions across the whole genome can be clearly determined. According to the inheritance mode and family structure, the portion of the genome that can be further excluded from consideration can be calculated (or simulated) for each added family member. This is referred to as 'exclusion power' for that added individual. The calculation of exclusion power is simple when at least one parent is available. For AD inheritance, availability or inference of founder genotype increases the exclusion power by two-fold. For siblings in an AD disease family, assuming full penetrance, the regions not shared IBD (roughly 50% of the genome) by two affected siblings or the regions shared IBD (also 50% of the genome) by a discordant sib pair can be excluded from consideration. With each additional sibling, roughly, a further half of the remaining genome can be excluded. However, when parent genotypes are not available, sharing status on both parental alleles has to be considered. By the Mendelian laws of inheritance, the probabilities of a sib pair sharing 0, 1 or 2 alleles IBD are $\frac{1}{4}$, $\frac{1}{2}$, and $\frac{1}{4}$, respectively. Then the first pair of siblings will only have an exclusion power of 25% of the genome, to exclude the regions they do not share at all for a concordant pair and the regions they share both alleles for a discordant pair. As each added sibling can be compared to all the other siblings pairwise, the exclusion power actually increases with each added sibling. For example, the third sibling will have an exclusion power of $\frac{5}{12}$, and it is $\frac{13}{28}$ and $\frac{29}{60}$ for the fourth and the fifth members, respectively. It can be seen that the exclusion power for each added sibling approaches, but never reaches, 50%. In the situation of AD full penetrance, unaffected siblings have the same power as the affected siblings.

For affected relatives, the further remote they are in relationship, the smaller the portion of the genome they share IBD; thus, more remote relatives confer more exclusion power (Table 1). For example, the regions shared reduce from a quarter of the genome for two first cousins to $\frac{1}{256}$ of the genome for two fourth cousins. The notion does not apply to discordant pairs or unaffected individuals.

Table 1 Exclusion power for AD inheritance and full penetrance

Relationship	Portion of the genome that two share a parental allele ^a	Portion of the genome that can be excluded in linkage studies (%)
Siblings	3/4	25
Grandparents/grand children	1/2	50
Uncle–aunt/nephew–niece	1/2	50
First cousins	1/4	75
Second cousins	1/16	93.75
Third cousins	1/64	98.44
Fourth cousins	1/256	99.61

^aAssuming founder allele unknown here, and both parental alleles have to be considered.

Exclusion power for recessive trait

For recessive trait, only the regions that the affected siblings shared both alleles need to be considered. As two siblings share both alleles in $\frac{1}{4}$ of the genome, then three quarters of the genome can be excluded from consideration, when there are two affected siblings available. Likewise, regions that a discordant pair shared both alleles can be excluded from consideration for a recessive trait, which is 25% of the genome. Therefore, unlike for dominant trait, unaffected siblings have only a third of the exclusion power of affected siblings in recessive cases.

Autozygosity mapping for recessive trait from consanguineous marriages

For rare recessive diseases, it is very hard to garner the power needed for linkage studies as only the diseased generation provides information for linkage. An advantageous exception is autozygosity mapping, in which the disease mutation derives from a common founder allele through consanguineous marriage.^{11–13} In these situations, only the autozygous allele from the parents needs to be considered for linkage, greatly reducing the region needed to be investigated. High-density SNP genotyping is especially beneficial to autozygosity mapping due to its great coverage and possibility of accurately determining the autozygous alleles. The exclusion mapping power depends on the portion of the genome shared IBD between the two parents (Table 1). For example, if the parents are second cousins to each other, they share about 1/16 of the genome IBD. Roughly, for a quarter of those regions, both alleles will pass on to a child together. So, for each affected child, only a quarter size of the parental autozygous regions needs to be considered, and a quarter of these regions for an unaffected child can be excluded. For hidden consanguineous marriages, usually the relationship between the parents is farther remote, and hence the regions shared IBD between parents are smaller, thus increasing the power of exclusion mapping. The issue for

these situations is the identification of autozygous regions, which could be in the range of a few centimorgan or smaller.

The simulation

The computer program, EcMap, is based on the simulation of the homologous recombination and inheritance processes to predict the number and sizes of the IBD regions according to given family structures and inheritance modes.

The recombination process is simulated according to an X^2 model, which is thought to best mimic the homologous recombination process.¹⁴ The simulation takes into consideration of physical interference on recombination, without considering chromatid interference. The model assumes that crossover intermediates (*C* events) are distributed along the four-strand sister chromatid bundle based on a Poisson distribution, with a rate of $2(m + 1)$ *C* events per Morgan. And every *C* event will either resolve in a crossover (*Cx*) or noncrossover (*Co*). We take an *m* value of 4 in the simulation as adopted by Broman and Weber¹⁵ and Zhao *et al.*¹⁶ When a *C* event resolves in a *Cx*, the next *m C*'s must resolve as *Co* events, and after *m Co*'s, the next *C* must resolve as a *Cx*. The leftmost *C* has an equal chance to be one of $Cx(Co)^m Cx(Co)^m$.¹⁶ The genetic distance map was obtained from HapMap and reflects average male and female genetic distances for each chromosome, with a total of 3,554 cM for all the autosomes (<http://hapmap.org>). The offspring genotypes were simulated through randomly inheriting one strand of the four-strand chromatid bundle from each parent, unless a mutation is also simulated in that region, for which the inheritance is determined by the affected status of the individuals.

The program simulates mutations in one of the founder alleles randomly. And after simulation of recombination, it forces the affected offspring to inherit the chromatid containing the mutation and the unaffected to inherit any other chromatid without the mutation. In cases of incomplete penetrance, the program analyzes the family first to identify any obligate unaffected carrier and force that individual to inherit the mutation allele. A number of unaffected are allowed to inherit the mutation allele on a random basis, and the total number of those individuals will not surpass the maximum number of unaffected mutation carriers for the family defined by the user. Similarly, the program allows a number of affected individuals not to inherit the mutation allele, and the maximum number of those individuals does not surpass the allowed number of phenocopies in a family defined by the user. When analyzing IBD regions, the program will count all the regions shared by affected plus unaffected mutation carriers taking into consideration of incomplete penetrance, but not by the unaffected or certain number of affected noncarriers taking into account of phenocopies.

The program outputs results on the following aspects: the size of the simulated IBD region containing mutation in each simulation, the number and sizes of the IBD regions not containing mutation, and the portion of the genome that can be excluded from consideration on average.

Bayesian probability prediction for an IBD region as the mutation-containing region

With certain number of simulations, the size distribution of the mutation-containing IBD regions and the size distribution of IBD region not containing mutation are two distinct populations. So, the simulation results allow the calculation of the Bayesian probability of an IBD region as mutation-containing region or an IBD region not containing causal mutation according to their sizes (cM) and simulation results:

$$P(\text{mutation}|\text{size}) = \frac{P(>\text{size}|\text{mutation})}{P(\text{mutation}) + P(>\text{size}|\text{random})}$$

$$P(\text{random}|\text{size}) = 1 - P(\text{mutation}|\text{size})$$

$P(>\text{size}|\text{mutation})$ is the area under curve for regions of certain size (cM) and above, and is calculated based on the distribution of the mutation-containing IBD region population from the simulation. $P(\text{mutation})$ would be the number of mutation-containing IBD region over all the IBD regions identified in simulations. $P(>\text{size}|\text{random})$ is calculated based on the distribution of the IBD regions without mutation. $P(\text{random})$ is the total number of IBD regions without mutation divided by the number of all IBD regions. Our program outputs a table depicting the two probabilities for region size from 1 to 100 cM, with an increment of 1 cM.

Results

Using several representative family structures, we have tested the program and analyzed the simulation results. We also discuss the implications of the simulation results on our understanding of family size requirement of linkage studies.

Simulation on an AD family assuming full penetrance

Family A in Figure 1 is used as an example of AD inheritance with full penetrance. In 10,000 simulations, the IBD regions that do not contain mutation in this family constitute 1.53% of the genome on average, similar to the calculated result. When considering both IBD regions containing causal mutation and IBD regions without mutation, they constitute 2.29% of the genome to consider for this family (Table 2). The average number of IBD region without mutation for this family is four with 10,000 simulations (Table 2). For this family structure, the IBD

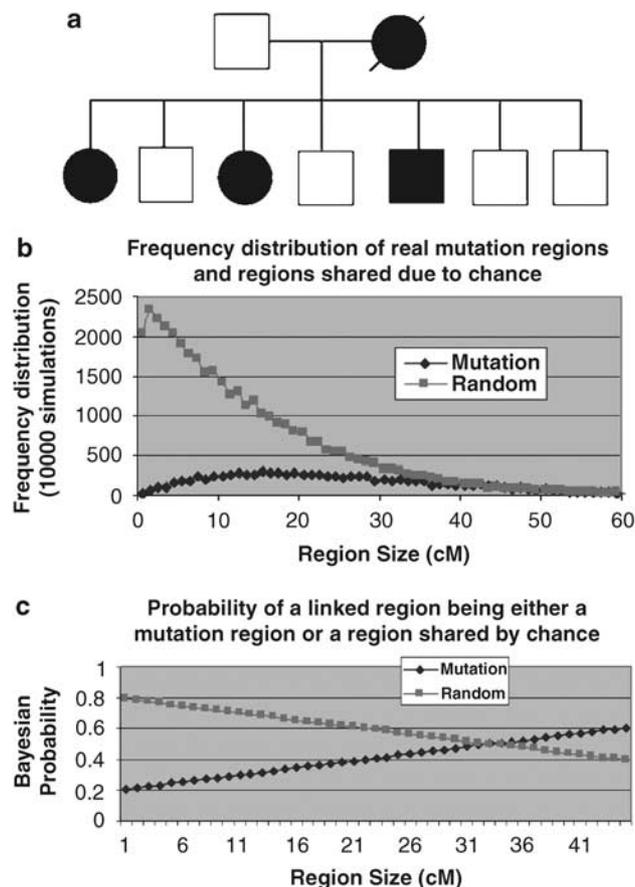


Figure 1 Simulation of linkage analysis on a family of AD inheritance with full penetrance (a) Pedigree of family A. The availability of at least one of the two parents is necessary for the simulation results. (b) Frequency distribution in size (cM) of IBD regions containing mutation (blue, diamond), and IBD regions without mutation (red, square) from 10,000 simulations. (c) Bayesian probability prediction of IBD region as being either mutation-containing (blue, diamond) or not (red, square) based on the size of the regions (X-axis) and simulation result from 1B.

regions containing mutation have an average size of 25.8 cM, and the IBD regions not containing mutation have an average size of 13.8 cM (Figure 1b). The size difference is understandable as for the mutation-containing IBD regions, once their sizes are below one Morgan, their further reduction by additional meioses will take a slower and slower pace, whereas for the IBD regions not containing mutation, the exclusion power remains steady with each added meiosis. Although the maximum achievable LOD score for this family is only 1.8, with exclusion of vast majority of the genome from consideration, and five regions on average to be considered, linkage studies could be a practical choice. Also, it is likely that among the regions that cannot be excluded, some of them could be very small, and the probability of those regions as

Table 2 Linkage analysis simulation results (10 000 simulations) for the four families in figures 1–4

	Probability of the mutation-containing region being the only IBD region (%)	Average size of the mutation-containing region (cM)	Average size of IBD regions not containing mutation (cM)	Average number of IBD regions without mutation (range)	Portion of genome can not be excluded after linkage analysis (%)
Family A (figure 1)	2	25.8	13.7	4 (0–14)	2.29
Family B (figure 2)	17	19.3	10.7	1.8 (0–8)	1.09
Family C (figure 3)	49	14.3	8.4	0.7 (0–6)	0.58
Family D (figure 4)	65	16.1	8.4	0.43 (0–7)	0.56

mutation-containing regions could be low based on our simulation results and Bayesian prediction.

Simulation on an AD family with incomplete penetrance

Here, we simulate on a family with a dominant trait but incomplete penetrance (family B in Figure 2a). In this family, individual II₅ is assumed as an obligate mutation carrier. The other four unaffected individuals (II₂, II₆, III₂, and III₅) could also be nonpenetrant mutation carriers depending on the penetrance of the mutation. We did simulation by assuming that up to two unaffected individuals could be mutation carriers in addition to II₅, and then analyzed what we can expect to achieve from a linkage analysis for this family. Figure 2 shows the distribution of the IBD regions and Bayesian probability calculation in evaluating those regions. It indicates that the average number of IBD regions not containing mutation is 1.8 in this case (Table 2). The average size of IBD regions containing mutation is 19 cM and the average size of the IBD regions without mutation is 10.7 cM. If we assume that II₅ is the only nonpenetrant carrier, then in more than 80% of the simulations, the mutation-containing region is the only IBD region to be identified in linkage analysis.

Simulation on autozygosity mapping for a recessive trait

Unless dominant inheritance cases for which multigeneration pedigrees can be analyzed, for recessive cases, one can rely on the diseased generation only. Development in genotyping technologies and analysis tools provides an opportunity to identify autozygous haplotypes.¹⁷ So autozygosity mapping provides a powerful tool to identify recessive causal mutations.

Family C is presented as a family with a recessive disease with both mutation alleles derived from a common founder mutation through consanguineous marriage. The parents of the affected children are second cousins and should share around 1/16 of the genome IBD. For a quarter of those regions, both alleles will pass on together to a

single offspring. Therefore, the total region left to be considered for this family will be $100\% \times (1/16) \times (1/4)^2 \times (3/4)^3 = 0.165\%$ on average. In 10 000 simulations from our program, the IBD regions not containing mutation constitutes 0.175% of the genome on average, which is very close to theoretical calculation of the exclusion power. In this case, the simulated mutation region constitutes about 0.4% of the genome on average (Figure 3, Table 2), adding up to 0.575% of the genome need to be considered in a physical mapping process.

Simulation on remote relatives

Linkage analysis can also be performed only using remote relatives, as high-density SNP genotyping provided an opportunity to recover IBD regions among affected remote relatives. For family D shown in Figure 4, if there is a founder mutation causal to all the three third cousins in this family, then very likely that the mutation-containing region will be the only region that is shared IBD by all three third cousins (65% of simulations, Tables 2 and 3). And the IBD region containing the mutation is likely to be larger than IBD regions not containing mutation, as shown in Figure 4b and Table 2. It does not matter whether the underlined disease is a rare disease of Mendelian inheritance or a complex disease caused by a near dominant founder mutation.

Understanding the IBD regions among remote relatives can help us design tools and strategies in both linkage analysis and association studies. Here, we have simulated families of different relationships and examined their allele sharing status in terms of the portion of the genome they share IBD and their average sizes (Table 3). It can be seen that the probability of sharing any region IBD by chance dwindles very fast when the relationship is getting more remote. An interesting phenomenon is that, if they do share a region IBD, the size of the IBD regions reduces with a slower and slower pace, as shown in Figure 5. These IBD regions are thus usually large enough to be detected by most high-density genotyping technologies, even when the relationship is hidden. This provides an advantage in both autozygosity mapping for recessive trait and for

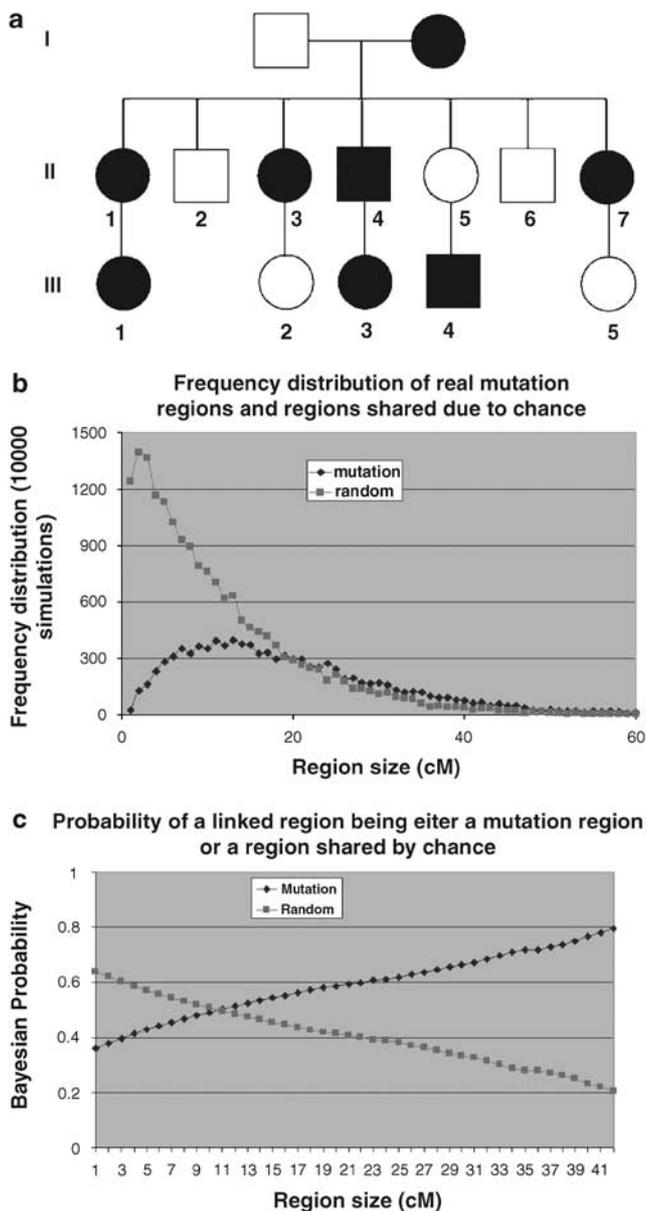


Figure 2 Simulation of linkage analysis on a family with AD inheritance but incomplete penetrance. (a) Pedigree of family B. The availability of at least one of the first generation couple is necessary for the simulation results. (b) Frequency distribution in size (cM) of IBD region containing mutation (blue, diamond) and IBD regions not containing mutation (red, square). Maximum number of nonpenetrant mutation carriers allowed for the family in the simulation is three, with individual II₅ an obligate carrier. (c) Bayesian probability prediction of IBD regions as being either mutation-containing (blue, diamond) or not (red, square) based on the size of the regions (X-axis) and simulation result from 2B.

detecting founder mutations for common complex diseases from the population level when multiple rare variant hypothesis is assumed.¹⁸

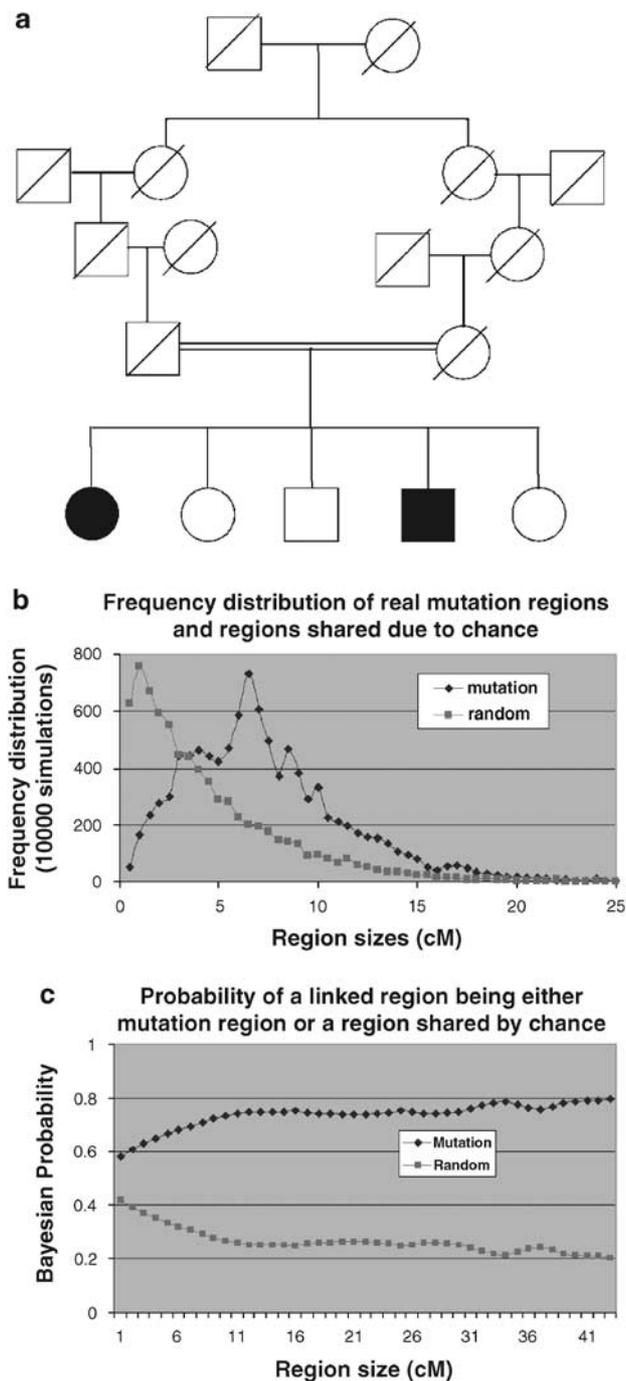


Figure 3 Simulation of linkage analysis by autozygosity mapping. (a) Pedigree of family C. The availability of the parents of the affected is not necessary for the simulation results. (b) Frequency distribution in size (cM) of autozygous IBD region containing mutation (blue, diamond) and autozygous IBD region not containing mutation (red, square) from 10 000 simulations. (c) Bayesian probability prediction of an autozygous IBD region as being either mutation-containing (blue, diamond) or not (red, square) based on the size of the regions (X-axis) and simulation results from 3B.

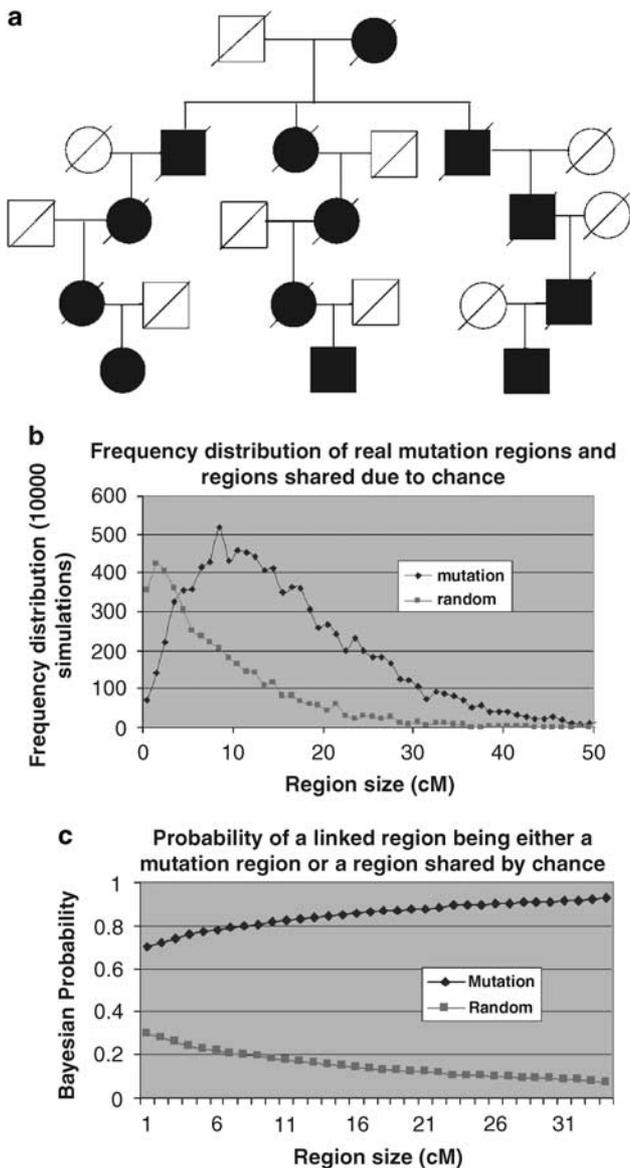


Figure 4 Simulation of linkage analysis on a family of AD with only remote relatives available. (a) Pedigree of family D. The disease could be a rare disease of Mendelian inheritance or a complex disease but suspected of a common founder mutation of near dominant effect. (b) Frequency distribution in size (cM) of the IBD region containing mutation (blue, diamond) and IBD region not containing mutation (red, square) from 10 000 simulations. (c) Bayesian probability prediction of an IBD region as being either mutation-containing (blue) or not (red) based on the size of the region (X-axis) and simulation results from 4B.

Discussion

Understanding of gene functions and their expression profiles in any detected genomic regions could greatly improve our ability to pinpoint the causal mutations. A combination of linkage and candidate gene approach in a family with merely four affected diabetic patients has led to Stoy *et al*¹⁹ to the identification of an initial insulin gene

Table 3 IBD regions among remote relatives (based on 10 000 simulations) computed in the absence of mutation

Number of individuals	Relationship (cousins)	Probability of not sharing any region IBD (%)	Average portion of the genome shared IBD (%)	Average size of regions shared IBD (cM)
2	First	0	24.5	22.6
	Second	0	6.1	15.7
	Third	0.13	1.6	12.2
	Fourth	23.7	0.4	9.9
	Fifth	63.5	0.1	8.4
	Sixth	84.9	0.03	7.2
3	First	0	6.1	15.7
	Second	8.1	0.8	11
	Third	65.2	0.1	8.4
	Fourth	92.9	0.01	6.9

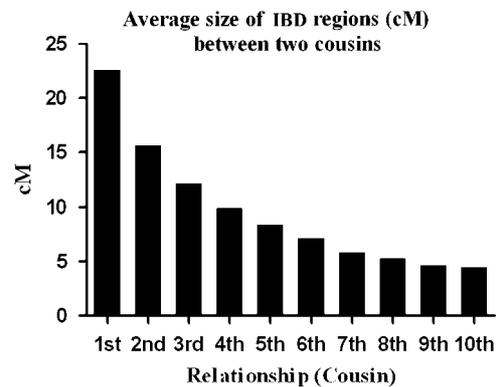


Figure 5 Average sizes of IBD regions between two cousins (by 10 000 simulations) simulated in the absence of mutations. No disease status and mutations were simulated in this case.

mutation, demonstrating the power of linkage in finding causal mutations even when the family sizes are small. Berkovic *et al*²⁰ recently reported identification of SCARB2/LIMP-2 deficiency as the cause for action myoclonus-renal failure syndrome (AMRF) using merely three cases, using a combination of autozygosity mapping, IBD mapping, and information on gene function and expression. Apparently, collection of affected families of significant sizes or multiple families with genetic homogeneity is difficult, if not impossible. And elucidating the causal mutations based on single families of smaller sizes is doable in many cases and will surely help to uncover many causal mutations for rare diseases. Even for mutations of incomplete penetrance, linkage analysis for a single family could still be practical, as demonstrated by the simulation result in family B (Figure 2).

Much improved understanding of human genomic variations and advances in genotyping technologies made genome-wide association study a reality. Similarly,

technology advancement changes the way people conduct and view linkage analysis. Whereas extracting 80% of information content was quite satisfactory with classical microsatellite genotyping,^{2,3} with high-density SNP coverage, it becomes feasible to recover the regions shared IBD by two remote cousins with near certainty.¹⁷ Here, by simulation of homologous recombination and inheritance process, we present a computer program that evaluates the number and sizes of IBD regions expected for families with diseases of Mendelian inheritance, and demonstrated the feasibility of linkage studies on single families even when the maximally achievable LOD score is less than 3. The prediction is conducted on single families, however, the information obtained can be easily combined from different families if interfamily homogeneity can be assumed.

Exclusion mapping holds much promises even when multiple IBD regions are inevitable, such as in the case of AMRF.²⁰ Even when whole genome sequencing for individual patient becomes a reality, pinpointing the real mutation from 2% of the genome would still mean a much easier task than working with the whole genome with thousands of potential mutations. This is true, especially when the mutations are not located in the coding regions of genes with known functions. Although we used rare disease families of Mendelian inheritance as examples throughout for clarity, the principle could also be applied to multiplex families of complex diseases, especially in situations of early onset and extreme subphenotypes, for which a high penetrant mutation is possible for the particular families.

Many genetic diseases are caused by known mutations in different genes (locus heterogeneity). The most striking example is retinitis pigmentosa, which may be caused by autosomal dominant, autosomal recessive, or X-linked mutations.^{21,22} Linkage studies combined with gene mapping has demonstrated that there are at least 41 loci responsible for eight X-linked forms, 13 AD forms, and 20 autosomal recessive forms of retinitis pigmentosa that are not associated with other phenotypic abnormalities.²¹ Thus, the identification of the causal mutation in any given affected family is a daunting task and could involve screening of numerous candidate genes. This poses problems to genetic counseling, prenatal diagnosis and personalized treatment for the affected families. In these situations, even when full-scale linkage studies may not be possible for the concerned families, it is still desirable to exclude majority of the genomic regions through linkage analyses based on limited family members available. This will reduce the list of genes needed to be considered and facilitate a more targeted mutation screening approach. For Mendelian diseases of unknown causes, linkage analysis could also be used to narrow down the list of potential candidate genes, therefore increasing the chances of success.

For conducting linkage analysis on single families, a few assumptions need to be fulfilled. For dominant traits, assumption of genetic homogeneity within family is essential, caution is needed, especially when the phenotypes are less typical. Autozygosity mapping is a method that greatly increases the power for identifying disease regions in recessive cases. This method is particularly useful for certain communities with a tradition of consanguineous marriages^{23,24} and when the diseases are rare in a particular population.²⁰ One alternative possibility for a recessive trait is that the disease might not arise from a common founder mutation but rather caused by compound heterozygosity or recurring mutations. It is an easy assumption for autozygosity cause for a rare allele but caution should be taken if the mutation allele is not so rare in the population in question.

Here, from simulation analysis, we have demonstrated the feasibility to conduct linkage studies on medium-sized families even when maximal LOD score cannot reach 3, with the probability that only a few regions will need to be considered, and the regions shared due to obligatory disease status would be usually larger. Even when the family size is small, we conclude that it is still possible to exclude large portion of the genome by linkage analysis. This serves well when dealing with diseases of genetic heterogeneity among families, by reducing the candidate genes need to be screened significantly. Although the calculation and the simulation presented here may not reflect the exact exclusion power in certain situations in reality,²³ the principle holds that linkage is becoming increasingly feasible for families considered too small before. On the basis of these findings, we would like to argue that in the era of association studies, linkage analysis should have a new life too – with much reduced cost but equally important potential biological gains.

Acknowledgements

WY thanks the support from the University Research Committee and LKS Faculty of Medicine, University of Hong Kong. LW's research is supported by the Research Grants Council of Hong Kong SAR, China (Project No. CityU 121207).

References

- 1 Goddard KA, Wijsman EM: Characteristics of genetic markers and maps for cost-effective genome screens using diallelic markers. *Genet Epidemiol* 2002; 22: 205–220.
- 2 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.
- 3 John S, Shephard N, Liu G *et al*: Whole-genome scan, in a complex disease, using 11 245 single-nucleotide polymorphisms: comparison with microsatellites. *Am J Hum Genet* 2004; 75: 54–64.
- 4 Carlson CS, Eberle MA, Rieder MJ, Smith JD, Kruglyak L, Nickerson DA: Additional SNPs and linkage-disequilibrium analyses are necessary for whole-genome association studies in humans. *Nat Genet* 2003; 33: 518–521.

- 5 Kruglyak L: The use of a genetic map of biallelic markers in linkage studies. *Nat Genet* 1997; **17**: 21–24.
- 6 Sawcer SJ, Maranian M, Singlehurst S *et al*: Enhancing linkage analysis of complex disorders: an evaluation of high-density genotyping. *Hum Mol Genet* 2004; **13**: 1943–1949.
- 7 Lin G, Wang Z, Wang L, Lau YL, Yang W: Identification of linked regions using high-density SNP genotype data in linkage analysis. *Bioinformatics* 2008; **24**: 86–93.
- 8 Donnelly KP: The probability that related individuals share some section of genome identical by descent. *Theor Popul Biol* 1983; **23**: 34–63.
- 9 Risch N: Linkage strategies for genetically complex traits. II. The power of affected relative pairs. *Am J Hum Genet* 1990; **46**: 229–241.
- 10 Thompson E, Basu S: Genome sharing in large pedigrees: multiple imputation of ibd for linkage detection. *Hum Hered* 2003; **56**: 119–125.
- 11 Gschwend M, Levrán O, Kruglyak L *et al*: A locus for Fanconi anemia on 16q determined by homozygosity mapping. *Am J Hum Genet* 1996; **59**: 377–384.
- 12 Lander ES, Botstein D: Homozygosity mapping: a way to map human recessive traits with the DNA of inbred children. *Science* 1987; **236**: 1567–1570.
- 13 Mueller RF, Bishop DT: Autozygosity mapping, complex consanguinity, and autosomal recessive disorders. *J Med Genet* 1993; **30**: 798–799.
- 14 Zhao H, Speed TP: On genetic map functions. *Genetics* 1996; **142**: 1369–1377.
- 15 Broman KW, Weber JL: Characterization of human crossover interference. *Am J Hum Genet* 2000; **66**: 1911–1926.
- 16 Zhao H, Speed TP, McPeck MS: Statistical analysis of crossover interference using the chi-square model. *Genetics* 1995; **139**: 1045–1056.
- 17 Woods CG, Valente EM, Bond J, Roberts E: A new method for autozygosity mapping using single nucleotide polymorphisms (SNPs) and EXCLUDEAR. *J Med Genet* 2004; **41**: e101.
- 18 Purcell S, Neale B, Todd-Brown K *et al*: PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 2007; **81**: 559–575.
- 19 Stoy J, Edghill EL, Flanagan SE *et al*: Insulin gene mutations as a cause of permanent neonatal diabetes. *Proc Natl Acad Sci USA* 2007; **104**: 15040–15044.
- 20 Berkovic SF, Dibbens LM, Oshlack A *et al*: Array-based gene discovery with three unrelated subjects shows SCARB2/LIMP-2 deficiency causes myoclonus epilepsy and glomerulosclerosis. *Am J Hum Genet* 2008; **82**: 673–684.
- 21 Daiger SP, Bowne SJ, Sullivan LS: Perspective on genes and mutations causing retinitis pigmentosa. *Arch Ophthalmol* 2007; **125**: 151–158.
- 22 Hartong DT, Berson EL, Dryja TP: Retinitis pigmentosa. *Lancet* 2006; **368**: 1795–1809.
- 23 Woods CG, Cox J, Springell K *et al*: Quantification of homozygosity in consanguineous individuals with autosomal recessive disease. *Am J Hum Genet* 2006; **78**: 889–896.
- 24 White DR, Ganesh A, Nishimura D *et al*: Autozygosity mapping of Bardet-Biedl syndrome to 12q21.2 and confirmation of FLJ23560 as BBS10. *Eur J Hum Genet* 2007; **15**: 173–178.