ELECTRONIC LETTER

A new method for autozygosity mapping using single nucleotide polymorphisms (SNPs) and EXCLUDEAR

C G Woods, E M Valente, J Bond, E Roberts

J Med Genet 2004;41:e101 (<http://www.jmedgenet.com/cgi/content/full/41/8/e101>). doi: 10.1136/jmg.2003.016873

The development of a silicon chip, such as the Affymetrix 10K Xba 131, bearing sufficient oligonucleotides to analyse 10 913 single nucleotide polymorphisms (SNPs) presents a new method for seeking autosomal recessive loci.¹ This letter describes a practical strategy to analyse the data output of such an "SNP-chip" for this purpose.

Autozygosity mapping, first suggested by Lander and Botstein, is the method of choice for the discovery of autosomal recessive gene loci.² The methodology seeks homozygous regions in consanguineous families. The greater the number of affected individuals who have a shared homozygous region and the greater the size of the region, the more likely it is to harbour the mutation that causes the disease. Mueller and Bishop modelled the use of a single multi-affected family and suggested that this was the most efficient strategy to determine a disease locus, particularly given the complexities of genetic heterogeneity.³ Autozygosity mapping became practical with the discovery of multiple highly polymorphic microsatellite repeat markers spread throughout the genome.⁴ Most researchers currently use optimised panels of approximately 400 markers for an initial genome-wide screen for linkage, giving 10–12 cM coverage of the autosomal genome; a process that has led to the discovery of many recessive loci.⁵

The currently available SNP-chip detects SNPs spread throughout the genome (with the exception of the Y chromosome) and is analysed following a single hybridisation reaction with one individual's genomic DNA. The results are produced as a simple spreadsheet of the SNP allele calls. Whilst each SNP has far less power to detect a homozygous chromosomal segment than a microsatellite marker, it is both their number (10 913 SNPs are equivalent to a 3–4 cM microsatellite marker map⁶) and their ability to detect a heterozygous region, and hence exclude linkage, that suggested their potential use in autozygosity mapping. An average microsatellite marker has a 70% chance of detecting a heterozygous region, but the approximately 30 SNPs within the same region have a >99% chance of detecting heterozygosity. (The chance that one of the 30 SNPs will be heterozygous is 1–0.7³⁰, which is >0.999, and on average nine of the 30 SNPs will be heterozygous.)

We have designed the following method to adapt Affymetrix SNP-chip output for autozygosity mapping.

1. Only affected individuals from a multi-affected pedigree are analysed; in general the minimum sample analysis necessary is four from a single sibship or three from two or more sibships. Parents and unaffected siblings are not analysed.
2. The primary results from each individual's SNP-chip hybridisation are produced as a simple EXCEL spreadsheet. For each individual the results are then sorted by chromosome and genetic distance using a simple "data sort" EXCEL command. The column of sorted SNP allele calls is "copied."

Key points

- Autosomal recessive disease gene loci can be found by analysis of affected members of consanguineous families using autozygosity mapping.
- Autozygosity mapping can be performed using SNPs, particularly SNP-chips bearing thousands of SNPs.
- We have devised a method to analyse the raw output of an Affymetrix 10K SNP-chip using a freely available spreadsheet, EXCLUDEAR.
- EXCLUDEAR detects significant regions of homozygosity in one, two, three, or four affected family members.

3. The sorted data are processed using EXCLUDEAR, a freeware spreadsheet we created for this purpose. A separate EXCLUDEAR spreadsheet is available for the analysis of one, two, three, and four affected individuals (see Appendix 1).
4. There are four versions of EXCLUDEAR (AR1–4) for the interpretation of data from one affected individual (AR1), two individuals (AR2), three (AR3), and four (AR4) individuals. The sorted primary SNP allele data from step 2 is pasted into EXCLUDEAR1 cell F:19 for one individual; for two individuals into EXCLUDEAR2 cells F:19 and G:19; for three individuals into EXCLUDEAR3 cells F:19, G:19, and H:19; or for four individuals into EXCLUDEAR4 cells F:19, G:19, H:19, and I:19.
5. EXCLUDEAR first detects runs of consecutive homozygous SNP allele calls identical in all of the affected individuals analysed. It then determines if each run is of statistical significance (Table 1 and Appendix 2 explain how we derived this data). For instance, for a pair of individuals, say consanguineous cousins, a run of 12 or more homozygous SNPs in both would occur only once in 1000 analyses by chance rather than being identical by descent.
6. EXCLUDEAR lists the 10 largest homozygous SNP runs by genetic size. For each result the following are given: genetic size, chromosome, genetic location on chromosome, number of homozygous SNPs in run, number of "NoCalls" in the run and whether the result reaches statistical significance. Two graphs are generated: the first shows the genetic size versus number of homozygous SNPs for the statistically significant results; the second shows the size of all statistically significant results by chromosome.

Abbreviation: SNP, single nucleotide polymorphism.

Table 1 The probability that a run of consecutive, concordant and homozygous SNP allele calls will occur by chance

Number of individuals analysed	Number of consecutive, homozygous and concordant SNPs required to occur <1 in 1000 times by chance*	Average minimum size of the region that could be detected†‡
1 patient	46	17 cM
2 siblings	23	8 cM
3 siblings	16	6 cM
4 siblings	12	4 cM
2 unrelated	12	4 cM
2 siblings and 1 unrelated	10	3.6 cM
3 unrelated	7	2.5 cM
4 unrelated	5	1.5 cM

*See Appendix 2 for full calculations.

†Assuming a sex-averaged autosomal length of 3488 cM.⁸‡For much of the genome the coverage of SNPs is fairly uniform, but there are areas of reduced coverage (see MRC Geneservice website: <http://www.hgmp.mrc.ac.uk/geneservice/index.shtml>), particularly where 1 cM > 1 Mb and the subtelomeric regions.

- The autosomal recessive disease gene sought could be located within any of the statistically significant homozygous segments detected.
- Table 1 gives an estimate of the minimum size of a homozygous region that could be detected using this method for different family structures. EXCLUDEAR also scans for potential homozygous deletions present in all affected individuals analysed (see Table 1 and Appendix 1).
- Any regions of statistically significant homozygosity may be further analysed by conventional polymorphic microsatellite analysis. The location of the SNPs is given by reference to the Human Genome Browser, and as a DeCode genetic distance, enabling the design or selection of suitable markers.⁷

The approach outlined here is undoubtedly a simplification. For instance, the possibility of genetic interference between neighbouring SNPs is ignored. It should, however, provide a practical method for analysis of SNP-chip output to detect significant homozygous segments, and hence locate recessive gene loci. EXCLUDEAR is free and available by

Table 2 The probability that consecutive SNPs would be homozygous and concordant by chance in one, two, three and four siblings

No. SNPs	For a person		For a sib pair		For three sibs		For four sibs	
	Probability homozygous*	Ratio†	Probability homozygous‡	Ratio†	Probability homozygous§	Ratio†	Probability homozygous¶	Ratio†
1	0.700000	0.0001	0.49000000	0.0002	0.34300000	0.0003	0.24010000	0.0004
2	0.490000	0.0002	0.24010000	0.0004	0.11764900	0.0008	0.05764801	0.0016
3	0.343000	0.0003	0.11764900	0.0008	0.04035361	0.0023	0.01384129	0.0066
4	0.240100	0.0004	0.05764801	0.002	0.01384129	0.0066	0.00332329	0.028
5	0.168070	0.0005	0.02824752	0.003	0.00474756	0.019	0.00079792	0.11
6	0.117649	0.0008	0.01384129	0.01	0.00162841	0.056	0.00019158	0.48
7	0.082354	0.0011	0.00678223	0.01	0.00055855	0.16	0.00004600	2.0
8	0.057648	0.0016	0.00332329	0.03	0.00019158	0.48	0.00001104	8.3
9	0.040354	0.0023	0.00162841	0.06	0.00006571	1.4	0.00000265	35
10	0.028248	0.0032	0.00079792	0.11	0.00002254	4.1	0.00000064	144
11	0.019773	0.0046	0.00039098	0.23	0.00000773	12	0.00000015	600
12	0.013841	0.0066	0.00019158	0.48	0.00000265	35	0.00000004	2499
13	0.009689	0.0095	0.00009387	0.98	0.00000091	101		
14	0.006782	0.016	0.00004600	2.0	0.00000031	294		
15	0.004748	0.019	0.00002254	4.1	0.00000011	857		
16	0.003323	0.028	0.00001104	8.3	0.00000004	2500		
17	0.002326	0.039	0.00000541	17				
18	0.001628	0.056	0.00000265	35				
19	0.001140	0.080	0.00000130	71				
20	0.000798	0.15	0.00000064	144				
21	0.000559	0.16	0.00000031	294				
22	0.000391	0.23	0.00000015	601				
23	0.000274	0.33	0.00000007	1226				
24	0.000192	0.48						
25	0.000134	0.68						
26	0.000094	0.98						
27	0.000066	1.40						
28	0.000046	1.99						
29	0.000032	2.85						
30	0.000023	4.07						
31	0.000016	5.81						
32	0.000011	8.30						
33	0.000008	12						
34	0.000005	17						
35	0.000004	24						
36	0.000003	35						
37	0.000002	50						
38	0.000001	71						
39	0.000001	101						
40	0.000001	144						
41	0.000000	206						
42	0.000000	295						
43	0.000000	421						
44	0.000000	602						
45	0.000000	860						
46	0.000000	1228						

For explanation of footnote symbols see Table 3.

Table 3 The probability that consecutive SNPs would be homozygous and concordant by chance in families with two to four individuals in different siblings

No. SNPs	For two individuals		For two siblings and another individual		For three individuals		For four individuals	
	Probability homozygous¶	Ratio†	Probability homozygous**	Ratio†	Probability homozygous††	Ratio†	Probability homozygous‡‡	Ratio†
1	0.245000	0.0004	0.17150000	0.0005	0.08575000	0.0011	0.03001250	0.0031
2	0.060025	0.0015	0.02941225	0.0031	0.00735306	0.0125	0.00090077	0.10
3	0.014706	0.006	0.00504420	0.018	0.00063053	0.145	0.00002703	3.39
4	0.003603	0.025	0.00086508	0.11	0.00005407	1.695	0.00000081	113
5	0.000883	0.10	0.00014836	0.62	0.00000464	20	0.00000002	3764
6	0.000216	0.42	0.00002544	3.60	0.00000040	231		
7	0.000053	2	0.00000436	21	0.00000003	2689		
8	0.000013	7	0.00000075	122				
9	0.0000032	29	0.00000013	714				
10	0.00000078	118	0.00000002	4163				
11	0.00000019	480						
12	0.00000005	1961						

*The probability that a run of n SNPs are homozygous is 0.7 (the chance a SNP has either an AA or BB allele call), then multiplied by n .
†The ratio is the probability presented as 1 to y , where y is the probability that n SNPs are consecutively homozygous multiplied by $10\,913 - n$ (the total number of autosomal SNPs, minus the number of SNPs in a run).
‡The probability that a run of n SNPs are homozygous and concordant in a sib pair is 0.7 (the chance a SNP has either an AA or BB allele call) multiplied by 0.7 (rather than 0.35 , to allow for reduced informativity) for each SNP, then multiplied by n .
§As note ‡ but $0.7 \times 0.7 \times 0.7$ for each SNP, then multiplied by n .
¶As note § but $0.7 \times 0.7 \times 0.7 \times 0.7$ for each SNP, then multiplied by n .
**The probability that a run of n SNPs are homozygous in two individuals is 0.7 (the chance a SNP has either an AA or BB allele call) multiplied by 0.35 (the chance another individual has the same homozygous allele call), then multiplied by n .
***As for note ¶ but $0.7 \times 0.35 \times 0.7$, then multiplied by n .
††As for note ¶, but $0.7 \times 0.35 \times 0.35$, then multiplied by n .
‡‡As for note ¶* $0.35 \times 0.35 \times 0.35$, then multiplied by n .

download from <http://leedsdna.info/science/Autozygosity/spreadsheets.htm>

ACKNOWLEDGEMENTS

We thank Andrew Dearlove and Jo McBride at the MRC Geneservice for performing the SNP-chip analysis that led to this letter and Graham Taylor, Chris Inglehearn, Carmel Toomes, and Tim Bishop for helpful comments and discussions. We also thank the Wellcome Trust for funding this research.

Authors' affiliations

C G Woods, J Bond, E Roberts, Molecular Medicine Unit, St James's University Hospital, Leeds, UK
E M Valente, IRCCS CSS, San Giovanni Rotondo and CSS Mendel, Rome, Italy

Conflict of interest: none declared

Correspondence to: C G Woods, Molecular Medicine Unit, Clinical Sciences Building, St James's University Hospital, Beckett Street, Leeds LS9 7TF, UK; msjcgw@leeds.ac.uk

Revised version received 17 February 2004
Accepted for publication 14 April 2004

APPENDIX 1

The EXCLUDEAR program was designed using the principle of the first EXCLUDE program written by JH Edwards to detect linkage in autosomal dominant conditions; namely that after excluding regions of non-linkage any chromosomal regions that remain must contain the locus sought. This is achieved in EXCLUDEAR by the detection of heterozygous or homozygous but discordant results for each consecutive SNP. The SNPs that remain are homozygous and concordant, and when consecutive can be summed. The genetic distance from the start of the run to the end can be calculated. The minimum number of consecutive and concordant homozygous SNPs detected is set at 10 for analysis of one individual, but is reduced for versions of EXCLUDEAR analysing data from more than one person (see below). The

results are ranked by genetic distance. The 10 largest regions of SNP homozygosity are shown together with the chromosome and number of SNPs involved. SNPs for which there are no results are scored as either AA, if no individual has a result, or as any other individual being analysed, i.e. "No Call"/AB, would be scored AB/AB. This permissive approach may lead to an overestimation of the number of homozygous SNPs in a run, but as >92% of SNPs are called per analysis the effect should be small. Furthermore, when the spreadsheet assesses the significance of a result it does so for the number of homozygous SNPs minus the number of "NoCalls".

The EXCLUDEAR program is given in four versions, for the data from one, two, three, and four individuals. EXCLUDEAR2, for two individual analyses, is set to detect a minimum run size of 12 SNPs; EXCLUDEAR3 for seven SNPs; and EXCLUDEAR4 for four SNPs (see Tables 1 and 2). EXCLUDEAR will also alert to the possibility of homozygous deletions, detected as runs of "NoCall" in all individuals analysed. Graphs illustrate the major findings by SNP number and genetic size, and by genetic size on each chromosome.

APPENDIX 2

The probability that a number of consecutive SNPs would be concordant and homozygous by chance was assessed in a singleton, two, three and four siblings and a non-sibling pair, a sib pair and non-sibling, three non-siblings, and four non-siblings. The results are summarised in Table 1 and shown in full in Tables 2 and 3. They were generated by calculation of the first individual's chance of being homozygous for an SNP, which is 0.70 (0.35 for AA and 0.35 for BB, our data from six consanguineous northern Pakistani individuals), multiplied by $[10\,913 - (n-1)]$ (where $10\,913$ is the total number of autosomal SNPs on the SNP-chip and n the number of SNPs in the run being analysed), multiplied by the chance that other individuals would be concordantly homozygous. This result for one SNP is then multiplied with the results for the consecutive SNP also being homozygous in an iterative manner until a probability of <1 in 1000 was achieved. The

probability of 1 in 1000 was chosen because it is in common use as the LOD score of 3, which is regarded as significant when seeking conventional linkage.

The probability that a number of consecutive SNPs would be concordant, homozygous and identical by chance is given for the following: a singleton, two, three and four siblings and a non-sibling pair, a sib pair and non-sibling, three non-siblings, and four non-siblings. In each family situation the smallest number of consecutive homozygous SNPs to occur less commonly than 1 in 1000 by chance is shown in bold.

REFERENCES

- 1 Affymetrix Genechip Human Mapping 10K array Xba1 131 details at: http://www.hgmp.mrc.ac.uk/geneservice/DNAservices/MRCg_10Kv1_2.pdf (accessed 14 May 2004).
- 2 **Lander ES**, Botstein D. Homozygosity mapping: a way to map human recessive traits with the DNA of inbred children. *Science* 1987;**236**: 1567–70.
- 3 **Mueller RF**, Bishop DT. Autozygosity mapping, complex consanguinity, and autosomal recessive disorders. *J Med Genet* 1993;**30**:798–9.
- 4 **Weber JL**, May PE. Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. *Am J Hum Genet* 1989;**44**:388–96.
- 5 **Sheffield VC**, Stone EM, Carmi R. Use of isolated inbred human populations for identification of disease genes. *Trends Genet* 1998;**14**:391–6.
- 6 **Sellick GS**, Garrett C, Houlston RS. A novel gene for neonatal diabetes maps to chromosome 10p12.1-p13. *Diabetes* 2003;**52**:2636–8.
- 7 **Kong A**, Gudbjartsson DF, Sainz J, Jonsdottir GM, Gudjonsson SA, Richardsson B, Sigurdardottir S, Barnard J, Hallbeck B, Masson G, Shlien A, Palsson ST, Frigge ML, Thorgeirsson TE, Gulcher JR, Stefansson K. A high-resolution recombination map of the human genome. *Nat Genet* 2002;**31**:241–7. Epub.
- 8 **Collins A**, Frezal J, Teague J, Morton NE. A matrix map of humans: 23 500 loci in 850 bands. *Proc Natl Acad Sci USA* 1996;**93**: 14771–5.