ELECTRONIC LETTER

SCAMP: a spreadsheet to collate autozygosity mapping projects

T Forshew, C A Johnson

J Med Genet 2004;**41**:e125 (http://www.jmedgenet.com/cgi/content/full/41/12/e125). doi: 10.1136/jmg.2004.023663

Autosomal recessive disorders are an important cause of childhood morbidity and mortality, and may reach significant frequencies in specific ethnic groups.[1] The affected progeny of consanguineous parents provide an opportunity to undertake gene mapping and positional-candidate gene analysis,[2] since it is highly likely that the disease locus is identical-by-descent from a common ancestor. The strategy of searching for regions of homozygosity in affected individuals from consanguineous families, using the methodology of autozygosity mapping, has proven to be highly effective for mapping loci and identifying autosomal recessive genes.[3] The identification of recessive disease genes enables diagnostic and carrier testing and can provide critical insights into the pathogenesis of the disease.

Most researchers who perform autozygosity mapping currently use panels of approximately 400 highly polymorphic microsatellite markers in an initial genome-wide linkage screen, to give a marker spacing of between 10 and 12 cM throughout the autosomal genome. The accurate and reliable genotyping of genetic markers is, of course, essential for the success of such mapping projects. However, we have found the subsequent collation of the large datasets generated in mapping projects to be both time consuming and prone to error if performed manually (by using, for example, ''cut and paste'' into a spreadsheet). We have therefore designed a simple but versatile Microsoft Excel spreadsheet which automatically collates and analyses genotyping data, and enables shared regions of homozygosity to be identified quickly following visual inspection of the collated data. The expected average length of autozygous regions, that are identical-by-descent around a disease locus, have been calculated by Génin *et al* to be 28 cM for affected individuals from a first cousin mating, and 22 cM from a second cousin mating.[4] In our experience, these values are good approximations to the length of homozygous regions found in real genome-wide linkage screens (fig 1), and, therefore, a marker spacing of 10 cM is adequate for a cohort of affected individuals from a first cousin mating.

We have called our spreadsheet ''SCAMP'' for spreadsheet to collate autozygosity mapping projects. The spreadsheet is available as a freeware download from the Medical and Molecular Genetics website at http://www.rch.bham.ac.uk/MMG/SCAMP.htm in the form of a Microsoft Excel workbook compressed as a .zip file or can be obtained from http://jmg.bmjjournals.com/supplemental/ as either a .zip or an .xls file. We have successfully used SCAMP in several autozygosity mapping projects[5–7] for which it provided a simple and practical solution for the analysis of genotyping data and facilitated the subsequent identification of disease loci and genes.

The key features of SCAMP are described in the following paragraphs.

I. SCAMP is based on the popular ABI PRISM Linkage Mapping Set v2.5, and contains the details of all 811 markers in panels 1–28 for the ''MD10'' configuration (coverage at 10 cM resolution) and the markers in panels 29–86 for the ''HD5'' configuration (~5 cM resolution). Marker details are collated in the ''ABI mapping panels v2.5'' spreadsheet and include panel number, chromosome number, locus and primer identifiers, dye colour, and heterozygosity value. In addition, each marker is annotated with sex-average genetic distances, based on the Marshfield and deCODE genetic maps,[8 9] and the physical distance from the deCODE genetic map. The order of the markers has been sorted by using the Excel ''data sort'' command, on the basis of chromosome number, followed by the Marshfield genetic distance. The values of genetic distance and physical location are linked to the separate spreadsheets ''Marshfield markers'' and ''deCODE markers'' using the data lookup function VLOOKUP. If required, other marker sets could be included in additional spreadsheets, although the data range for the VLOOKUP functions must be changed. The syntax for data lookup functions is easily accessible using the help files in Microsoft Excel.

II. The ''genotypes'' spreadsheet contains only the 400 ABI PRISM ''MD10'' markers which would be usual for a general genome-wide linkage screen. The markers are sorted on the basis of chromosome number and genetic distance, and the values for their genetic distances and physical locations are linked to the ''ABI mapping panels v2.5'' spreadsheet by VLOOKUP, although other spreadsheets containing marker details could be used. Figure 1 shows a small section of the ''genotypes'' spreadsheet that covers the genotyping data for chromosome 1 markers for samples 1–6.

III. Genotyping data is entered into the ''raw data'' spreadsheet. This is most easily done by saving the output from, for example, ABI PRISM ''Genotyper'' software as a

Key points

- Autosomal recessive disorders are an important cause of childhood morbidity and mortality, and may reach significant frequencies in specific ethnic groups.
- Identification of recessive disease genes enables diagnostic and carrier testing and can provide critical insights into the pathogenesis of the disease
- We have designed a simple but versatile spreadsheet which automatically collates and analyses genotyping data, and enables shared regions of homozygosity to be identified quickly following visual inspection of the collated data. The spreadsheet has been called SCAMP.
- SCAMP provides a simple and practical solution for the analysis of genotyping data and facilitates the subsequent identification of disease loci and genes.

Locus	Chr.	Panel	Marshfield genetic location (cM)	deCODE genetic location (cM)	deCODE physical location (bp)	COUNTIF HOM	Sample: 1 to 8: FAIL	1	1		2	2		3	3		4	4		5	5		6	6	
D1S468	1	2	4.22	4	3766906	1	0	2	4		5	6		3	6		1	2		5	5	HOM	5	6	
D1S214	1	2	14.04	#N/A	#N/A	4	2	2	4		3	3	HOM	6	6	HOM	0	0	FAIL	4	4	HOM	6	6	HOM
D1S450	1	1	20.61	16.61	N/A	2	2	1	1	HOM	2	5		4	6		0	0	FAIL	0	0	FAIL	4	4	HOM
D1S2667	1	1	24.68	19.88	12332845	4	0	1	1	HOM	2	2	HOM	1	1	HOM	3	4		2	2	HOM	5	6	
D1S2697	1	2	37.05	#N/A	#N/A	1	0	2	3		1	2		4	4	HOM	4	6		1	3		1	5	
D1S199	1	2	45.33	37.48	21748489	1	1	0	0	FAIL	3	5		5	9		1	3		4	7		8	9	
D1S234	1	1	55.1	44.49	NA	6	0	2	2	HOM	2	2	HOM	2	2	HOM	2	2	HOM	1	8		2	2	HOM
D1S255	1	1	65.47	58.66	43309508	5	0	6	6	HOM	6	6	HOM	6	6	HOM	6	6	HOM	6	6	HOM	6	5	
D1S2797	1	1	75.66	68.9	53929232	6	1	3	3	HOM	3	3	HOM	3	3	HOM	3	3	HOM	3	3	HOM	3	3	HOM
D1S2890	1	1	85.68	80.9	67109382	6	0	4	4	HOM	4	4	HOM	4	4	HOM	4	5		4	4	HOM	3	3	HOM
D1S230	1	2	95.31	88.12	73127176	1	0	8	9		4	4	HOM	2	5		3	4		4	5		1	2	
D1S2841	1	2	106.45	103.82	92192753	0	0	1	3		1	4		2	7		2	5		2	7		3	4	
D1S207	1	2	113.69	107.16	NA	1	0	3	5		5	9		1	3		4	7		8	9		4	4	
D1S2868	1	2	126.16	116.64	107635420	1	0	1	3		4	7		8	9		4	4	HOM	2	5		3	4	
D1S206	1	1	134.2	122.64	117752569	1	0	2	4		2	5		1	2		4	7		1	5		3	4	
D1S2726	1	1	144.38	132.11	128479059	2	1	2	6		1	8		3	3	HOM	1	2		4	4	HOM	4	5	
D1S252	1	2	150.27	139.16	135655525	1	0	4	4	HOM	2	5		3	4		4	5		1	2		2	4	
D1S498	1	2	155.89	144.94	179099580	1	0	2	7		2	3		3	5		5	8		3	4		4	5	
D1S484	1	1	169.68	157.51	190022422	1	1	3	5		0	0	FAIL	3	4		3	5		3	4		7	7	HOM
D1S2878	1	1	177.86	165.78	195390880	1	0	1	3		2	6		3	3	HOM	3	4		4	8		3	5	
D1S196	1	1	181.49	169.4	197936216	0	0	5	8		1	5		5	6		4	6		1	4		2	4	
D1S218	1	2	191.52	176.3	205630555	1	1	1	5		3	12		4	4	HOM	0	0	FAIL	1	3		3	5	
D1S238	1	2	202.73	188.55	221049296	1	0	5	9		1	3		4	7		8	9		4	4	HOM	2	5	
D1S413	1	2	212.44	194.98	232023956	1	0	4	7		8	9		4	4	HOM	2	5		3	4		4	5	
D1S249	1	1	220.65	207.07	240040126	1	0	3	6		4	5		2	4		1	2		1	4		4	4	HOM
D1S425	1	2	231.11	215.07	246595148	0	0	3	4		5	10		1	2		3	5		3	4		4	5	
D1S213	1	1	242.34	228.26	NA	0	0	2	7		2	4		4	5		3	4		4	7		2	3	
D1S2800	1	1	252.12	#N/A	#N/A	4	1	3	9		3	3	HOM	1	1	HOM	0	0	FAIL	3	4		3	3	HOM
D1S2785	1	1	266.27	257.46	278270552	0	0	3	4		1	7		2	3		3	4		2	4		2	3	
D1S2842	1	1	273.46	261.86	280760278	2	0	2	3		2	6		1	2		2	2	HOM	1	5		3	3	HOM
D1S2836	1	1	285.75	271.84	285061332	0	1	2	6		1	3		0	0	FAIL	1	4		1	4		2	4	

Figure 1 A representative section of the ''genotypes'' spreadsheet that covers the genotyping data for chromosome 1 markers for samples 1–6. The first six columns list the marker locus, the chromosome number (in this case, chromosome 1), the panel number in the ABI PRISM Linkage Mapping Set v2.5 ''MD10'' configuration, and the genetic or physical locations of the marker. Columns seven and eight in grey summarise the number of ''HOM'' and ''FAIL'' calls in the dataset (in this case, samples 1–8). SCAMP detects an interesting region of shared homozygosity, highlighted in darker grey in column seven, between markers D1S199 and D1S230. The remaining columns display genotyping data for samples 1–6 taken from the ''raw data'' spreadsheet, and a summary column to make a ''HOM'' call if both alleles are homozygous and a ''FAIL'' call if the genotyping has failed.

tab-delimited text file that can be imported into the ''raw data'' spreadsheet in the form of four continuous columns of data. Column 1 contains the marker name (which must exactly match the name in other spreadsheets), column 2 contains the sample number (in this example 1–16; see paragraph V below), and columns 3 and 4 contain the genotyping data. Column 5 contains the ROW function to return the number of the row, which is used by the data lookup functions in ''genotypes''.

IV. The ''genotypes'' spreadsheet then uses a combination of the INDEX and VLOOKUP functions to look up genotyping data in the ''raw data'' spreadsheet, and to place it in the correct position in ''genotypes''. The syntax for these functions is available from the help files in Microsoft Excel. Genotype data takes the form of two columns, with a third adjacent column to call if the alleles are either homozygous (an IF function returns a ''HOM'' value) or if the genotyping has failed (a ''FAIL'' value is returned).

V. In the freeware download, available from the University of Birmingham website, we have included typical results for two separate cohorts, each of eight affected individuals. In this example, the two sets of eight individuals are analysed separately, with the data derived from 16 rows of genotyping data for each marker in the ''raw data'' spreadsheet. The parameters of the lookup functions can be easily changed for fewer or greater number of individuals, and the existing structure of the spreadsheets can be used as a template for a customised analysis.

VI. Conditional formatting is used to colour in ''HOM'' values for samples 1–8 for the first cohort in pale green, and for samples 9–16 for the second cohort in pale cyan. (The colours refer to the formatting in the electronic form of the spreadsheet, but are rendered in shades of grey in the representative section shown in fig 1.) Summary columns on the left of samples 1–8 and samples 9–16 are coloured in grey. These use the COUNTIF function to count the number of ''HOM'' and ''FAIL'' calls for each cohort. For samples 1–8, an interesting region of shared homozygosity is detected on chromosome 1, with conditional formatting colouring in COUNTIF values of $\geqslant$5 in green. For samples 9–16, an interesting region is highlighted in cyan on chromosome 5. Conditional formatting also colours values of $\geqslant$2 in red for the COUNTIF function on the number of ''FAIL'' calls in each cohort. This shows that genotyping needs to be completed on chromosomes 17–21 for both datasets. The conditional formatting values of $\geqslant$5 out of 8 for ''HOM'' calls and $\geqslant$2 out of 8 for ''FAIL'' calls can be easily changed to suit the requirements of the mapping project, although in this example they are designed for two cohorts of eight individuals each.

ELECTRONIC-DATABASE INFORMATION

The Medical and Molecular Genetics website can be found at http://www.rch.bham.ac.uk/MMG/SCAMP.htm. The SCAMP spreadsheet is available as supplementary information from http://jmg.bmjjournals.com/supplemental/.

Authors' affiliations

T Forshew, C A Johnson, Section of Medical and Molecular Genetics, Department of Paediatrics and Child Health, University of Birmingham Medical School, Birmingham B15 2TT, UK

Conflict of interest: none declared.

Correspondence to: Colin A Johnson, Section of Medical and Molecular Genetics, Department of Paediatrics and Child Health, University of Birmingham Medical School, Birmingham, B15 2TT, UK; c.a.johnson@bham.ac.uk

Revised version received 13 July 2004
Accepted for publication 13 July 2004

REFERENCES

1 **Bundey S**, Alam H. A five-year prospective study of the health of children in different ethnic groups, with a particular reference to the effect of inbreeding. *Eur J Hum Genet* 1993;**1**:206–19.
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 **Sheffield VC**, Stone EM, Carmi R. Use of isolated inbred human populations for identification of disease genes. *Trends Genet* 1998;**14**:391–6.
4 **Génin E**, Todorov AA, Clerget-Darpoux F. Optimization of genome search strategies for homozygosity mapping: influence of marker spacing on power and threshold criteria for identification of candidate regions. *Ann Hum Genet* 1998;**62**:419–29.
5 **Aligianis IA**, Forshew T, Johnson S, Michaelides M, Johnson CA, Trembath RC, Hunt DM, Moore AT, Maher ER. Mapping of a novel locus for achromatopsia (*ACHM4*) to 1p and identification of a germline mutation in the alpha subunit of cone transducin (GNAT2). *J Med Genet* 2002;**39**:656–60.
6 **Morgan NV**, Bacchelli C, Gissen P, Morton J, Ferrero GB, Silengo M, Labrune P, Casteels I, Hall C, Cox P, Kelly DA, Trembath RC, Scambler PJ, Maher ER, Goodman FR, Johnson CA. A locus for asphyxiating thoracic dystrophy, *ATD*, maps to chromosome 15q13. *J Med Genet* 2003;**40**:431–5.
7 **Gissen P**, Johnson CA, Morgan NV, Stapelbroek JM, Forshew T, Cooper W, McKiernan PJ, Knisely AS, Klomp LWZ, Morris AAM, Abdullah MA, Wraith JE, McClean P, Lynch SA, Thompson RJ, Lo B, Quarrell OW, DiRocco M, Trembath RC, Mandel H, Karet FE, Houwen R, Kelly DA, Maher ER. ARC syndrome, a severe multisystem metabolic disorder, is caused by mutations in the *VPS33B* gene, encoding a regulator of SNARE-dependant membrane fusion. *Nat Genet* 2004;**36**:400–4.
8 **Broman KW**, Murray JC, Sheffield VC, White RL, Weber JL. Comprehensive human genetic maps: individual and sex-specific variation in recombination. *Am J Hum Genet* 1998;**63**:861–9.
9 **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.