Hum. Genet. 47, 297—304 (1979)

Human Genetics

The Likely Region of Overlap (LRO) Method for Physical Assignment of Loci

W. R. Williams, N. E. Morton, R. Lew, and S. Yee

Population Genetics Laboratory, University of Hawaii, Honolulu 96822, USA

Summary. Numerical analysis is applied to physical assignments of loci, providing point estimates, an LRO confidence interval, and a χ^2 test of consistency whether there is a smallest region of overlap (SRO) or not. Results are given for two examples and summarized for 81 loci in man.

Deletions, duplications, and structural rearrangements of chromosomes permit physical assignment of loci. Conventionally, the smallest region of overlap (SRO) method is used, in which a locus is assigned to the smallest region common to inclusive segments. However, this method fails if the evidence is inconsistent due to misinterpretation of breakpoints, position effects, or controlling loci.

Here we develop a method of gene assignment that takes all the evidence, using efficient weights, to define a likely region of overlap (LRO), point estimates, standard errors, and tests of consistency for each locus.

Material

A catalog of inclusive segments was compiled from the literature published before 1978 (Keats et al., in press). As shown in Figure 1, an inclusive segment may be defined by either presence or absence of a gene product. Rarely an interstitial fragment or deletion is ambiguous, and then the inclusive segment is defined by consistency with other evidence. For example, in the figure the two observations AB−, BC− imply inclusive segments BD, CD.

In principle, position effects and controlling elements can cause erroneous inference, but most errors come from misinterpretation of breakpoints when only major bands were visualized. Even under exceptionally favorable conditions there is an error of part of a band for each breakpoint, and so we have taken inclusive segments to contain all of a band in which a breakpoint was inferred (e.g., p31→p21 extends from the distal edge of p31 to the proximal edge of p21).

Some authors pooled evidence from several segments, reporting only an SRO, which we have accepted as an inclusive segment. In many instances we could not determine whether two or more reports of the same inclusive segment referred to independent clones. Therefore we accepted a particular inclusive segment only once for each locus.

0340-6717/79/0047/0297/$ 01.60

A	B	C	D
Observation		Inclusive segment	
AD+		AD	
BD+		BD	
CD+		CD	
AB–		BD	
AC–		CD	
BC– or A (del BC) D+		CD or AB, whichever is consistent with other evidence	

Fig. 1. Definition of an inclusive segment. Arrow denotes locus whose gene product is present (+) or absent (–) in a segment defined by points *A, B, C, D*

The LRO method requires that inclusive segments be transformed into physical breakpoints. We did this by mensuration of banding diagrams (Paris Conference Supplement, 1975). The final data set consisted of 343 observations for 81 different genes, 77 of which could be defined by an SRO (of these 22 had only a single observation), while the remaining 4 genes had at least one inconsistent assignment.

Theory

For a given locus μ let a_i, b_i be the terminal points of the i^{th} inclusive segment ($a_i < b_i$). By heuristic analogy with a uniform density, the corresponding point estimate is

$$X_i = \frac{a_i + b_i}{2} \tag{1}$$

and the information is

$$W_i = 12/(b_i - a_i)^2 \tag{2}$$

We take W_i as an initial weight for X_i, according to which information about μ is inversely proportional to the square of the inclusive segment. However, this result is not exact, since the density of loci on the physical map is unknown; μ is a point which has no distribution except to a Bayesian (rather it is X that has a distribution), and there is some error in interpreting rearrangements, assigning terminal points, and accepting an SRO when the inclusive segments are not reported. We shall use the value of W_i suggested by a uniform distribution only to obtain a better approximation.

Now let us suppose that we have observations (a_i, b_i), $i = 1 \ldots n$. We may reasonably assume that the likelihood approaches multinormality as n increases,

$$L(X_1, \ldots X_n) \propto e^{-(1/2)\Sigma W_i (X_i - \mu)^2} \tag{3}$$

with maximum likelihood estimates

$$\hat{\mu} = \Sigma W_i X_i / \Sigma W_i \tag{4}$$

$$\text{Var}(\hat{\mu}) = \sigma^2 = 1/\Sigma W_i$$

In this asymptotic theory, goodness of fit is tested by

$$\chi^2_{n-1} = \Sigma W_i (X_i - \hat{\mu})^2 \qquad n > 1 \tag{5}$$

and the 95% confidence interval is nearly

$$\hat{\mu} \pm 2\sigma \tag{6}$$

which we define as the LRO for the locus. If desired, the LRO may be transformed from physical units back to chromosomal bands (e.g., Q3403 is 3/10 of band Q34 from its proximal edge).

Results

Improved weights that make allowance for errors in determining inclusive segments and in distribution theory were obtained from the pseudo-chi-square variable

$$Y_i = W_i (X_i - \hat{\mu})^2 \frac{n_i}{n_i - 1} \qquad n > l \tag{7}$$

which in large-sample theory should have a mean value of unity for all loci. This variable was regressed on the inclusive segment length $(b_i - a_i)$ and its square for the 55 loci with SROs and multiple segments, since the 4 loci without SROs might be expected not to have well-behaved values. Only the square of segment length contributed significantly to regression, accounting for 11% of the variation in Y (Table 1). An improved weight is

$$W_i = \frac{12}{(b_i - a_i)^2 [0.69934 + 0.000003667 (b_i - a_i)^2]} \tag{8}$$

When this refined LRO method was applied, no residual regression was significant. The mean of the new pseudo-chi-square was 0.989, close to the expected value of unity and therefore acceptable. The variance has no simple interpretation because degrees of freedom are not constant, and so variability is better judged from the χ^2 distribution.

For each locus a chi-square value was calculated by Eq. (5), with W_i defined by Eq. (8) (Table 2). Significant lack of fit (nominal $P < 0.05$) was detected in

Table 1. Regression analysis of the pseudo-chi-square variate (Y) on the square of the interval size (X)

Source of variation	*df*	Sum of squares	Mean squares
Attributable to regression	1	28.56	28.56[a]
Deviation from regression	281	221.98	0.79
Total	282	250.54	

[a] $P < 0.01$; $Y_i = 0.69934 + 0.000003667 (b_i - a_i)^2$

Table 2. Distribution of χ^2 frequencies

Probability	0–0.05	0.05–0.1	0.1–0.25	0.25–0.5	0.5–0.75	0.75–0.9	0.9–1
Inconsistent $n = 4$	2	0	2	0	0	0	0
Consistent $n = 55$	2	1	11	13	15	8	5

2/55 consistent cases (i.e., with an SRO), in good agreement with large-sample theory. Loci with inconsistent assignments (no SRO) gave 2/4 significant results. These loci are a heterogeneous class, including slight and gross discrepancies. For the two inconsistent loci with significant lack of fit we took a modified LRO,

$$\hat{\mu} \pm 2\sigma \sqrt{\chi^2/(n-1)}$$

corresponding to the observed standard deviation among inclusive segments for that locus. Other investigators might prefer to use this empirical standard deviation for all inconsistent loci, including the two that were not significantly heterogeneous.

Examples of LRO Mapping

Peptidase C (PEPC) and phosphoglucomutase-1 (PGM1) serve to illustrate mapping by the LRO method (Figs. 2 and 3). In the case of PEPC, all assignments overlap a common interstitial region (SRO) extending from band q41 to band q43. Five inclusive segments do not overlap for PGM1. The LRO point estimates

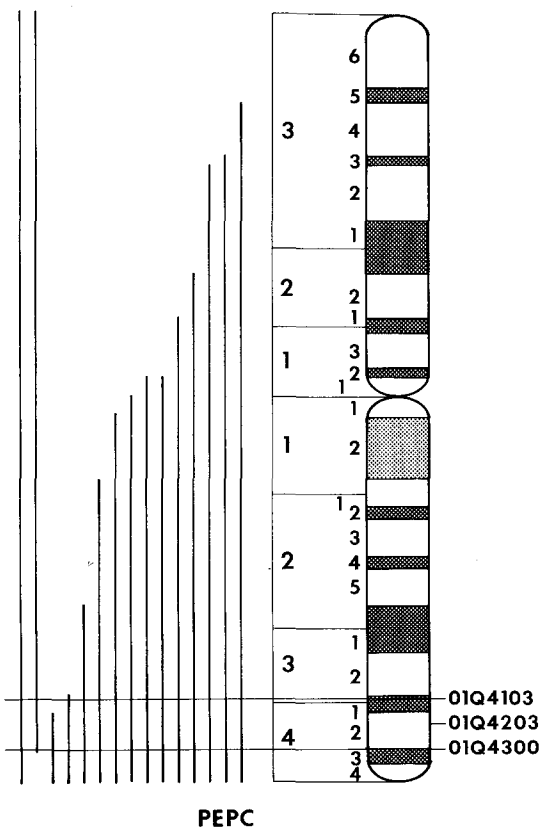

Fig. 2. Inclusive segments, LRO, and point estimate for a consistent gene, *PEPC*, on chromosome 1

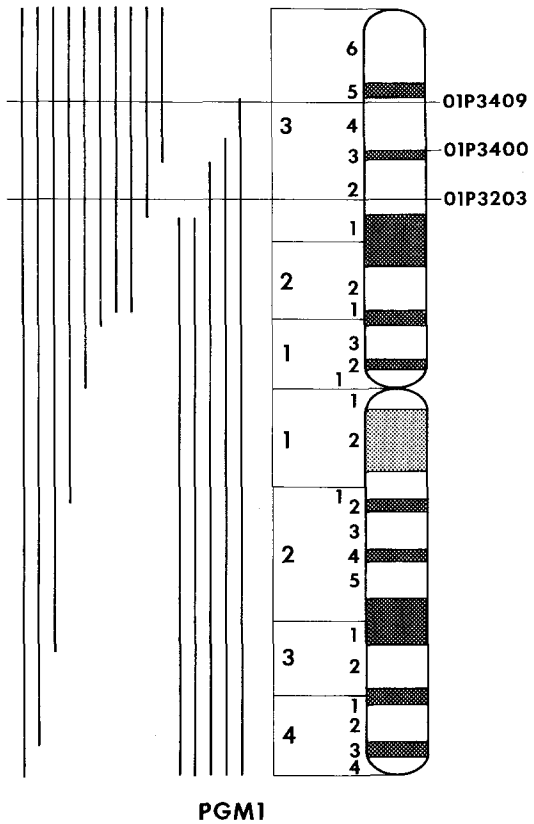

Fig. 3. Inclusive segments, LRO, and point estimate for an inconsistent gene, *PGM1*, on chromosome 1

Table 3. LRO and SRO gene assignments

Chrom.	Gene	McKusick number	*n*	LRO		SRO	
				Lower limit	Upper limit	Lower limit	Upper limit
01	*AK2*	10302	4	p3602	p3202	pter	p3200
01	*EN01*	17243	13	p3604	p3408	pter	p3600
01	*FH*	13685	12	q4100	q4300	q4200	qter
01	*FUCA*	23000	1	p3500	p3200	p3500	p3200
01	*GUK1*	13927	11	q3106	q4203	q3200	q4300
01	*PEPC*	17000	15	q4103	q4300	q4200	q4300
01	*PGD*	17220	10	p3602	p3401	pter	p3400
01	*PGM1*	17190	15	p3409	p3203	—	—
01	*RH*	11170	2	p3607	p3108	pter	p3200
01	*RN5S*	18042	4	q4109	q4306	q4200	q4400
01	*UGPP*	19175	9	q1209	q3203	q2100	q3200

Table 3 (continued)

Chrom.	Gene	McKusick number	*n*	LRO		SRO	
				Lower limit	Upper limit	Lower limit	Upper limit
01	*UMPK*	19173	10	p3409	p3108	p3300	p3200
02	*ACP1*	17150	4	p2503	p2207	p2400	p2300
02	*IDH1*[a]	14770	5	p1200	q1308	—	—
02	*MDH1*	15420	5	p2501	p2206	pter	p2300
04	*AMPRT*	17245	1	pter	q2205	pter	q2200
04	*GC*	13920	1	q1101	q2100	cen	q2100
04	*PEPS*	17025	1	pter	q2205	pter	q2200
04	*PGM2*	17200	5	p1502	q2203	p1500	q2200
05	*AVSRR*	10745	1	pter	cen	pter	cen
05	*DTS*	12615	2	q1409	q3501	q1500	qter
05	*HEXB*	14265	2	p1409	q1500	cen	q1400
06	*HLA*	1428–	4	p2306	p1106	p2400	p2100
06	*GLO1*	13875	2	p2504	p1108	p2400	p2100
06	*ME1*	15425	7	p1104	q2105	p2100	q2100
06	*PGM3*	17210	3	p1107	q2600	p2100	qter
06	*SOD2*	14746	7	q1605	q2202	q2100	q2200
07	*GUS*	25322	7	p1305	q1109	—	—
07	*HAF*	23400	2	q3409	q3608	q3500	qter
07	*JK*	11100	1	p1106	qter	cen	qter
07	*MDH2*	15410	2	p2108	q3103	pter	q3100
07	*SA7*	18552	1	pter	q1101	pter	cen
07	*SV40*	18680	1	p1106	qter	cen	qter
08	*GTR*	13830	9	p2207	p1207	p2200	p2100
09	*AC02*	10085	3	p2304	q1103	pter	cen
09	*AK1*	10300	9	q3400	q3405	q3400	q3407
09	*AK3*	10303	2	p2403	q1200	pter	cen
09	*GPUT*	23040	2	p2403	q1200	pter	cen
10	*GOT1*	13818	3	q2405	q2500	q2405	q2500
10	*HK1*	14260	1	pter	q2508	pter	q2500
10	*PP*	17903	1	pter	q2508	pter	q2500
11	*ACP2*	20095	5	p1408	q1105	p1300	cen
11	*ALA1*	15125	4	p1531	p1305	pter	p1300
11	*ALA2*	15126	4	q1303	q2209	q1300	qter
11	*ALA3*	15127	4	p1531	p1305	pter	p1300
11	*ESA4*	13322	2	p1108	q2303	cen	q2300
11	*LDHA*	15000	7	p1404	p1206	p1400	p1300
11	*SA-1*	18555	1	pter	cen	pter	cen
12	*GAPD*	13840	4	p1304	p1202	pter	p1200
12	*LDHB*	15010	6	p1209	p1109	p1300	p1200
12	*PEPB*	16990	3	q1503	q2403	q2100	q2200

Table 3 (continued)

Chrom.	Gene	McKusick number	*n*	LRO		SRO	
				Lower limit	Upper limit	Lower limit	Upper limit
12	*SHMT*	13845	1	pter	q1506	pter	q1500
12	*TPI*	19045	4	p1304	p1202	pter	p1200
13	*RB1*	18020	4	q1304	q2100	q1400	q2100
14	*NPP*	16405	6	q1102	q2205	q1200	q2200
14	*TRS*	19105	1	q1308	qter	q2100	qter
15	*BMG*	10970	4	q1406	q2108	q1400	q2200
15	*HEXA*	27280	4	q2109	q2505	q2200	qter
15	*IDH2*	14765	3	q2100	q2502	q2100	qter
15	*MANA*	15458	1	p1102	qter	cen	qter
15	*MPI*	15455	4	q2200	q2505	q2200	qter
15	*PK3*	17905	4	q2200	q2505	q2200	qter
16	*HPA*	14010	1	pter	q2302	pter	q2300
17	*GAK*	23020	3	q2102	q2209	q2100	q2300
17	*TKS*	18830	3	q2102	q2209	q2100	q2300
18	*PEPA*	16980	3	q2300	q2309	q2300	qter
20	*ADA*	10270	1	p1200	qter	p1200	qter
21	*AVP*	10745	2	q2103	q2207	q2100	qter
21	*DOWN*	—	3	q2109	q2207	q2200	qter
21	*SOD1*	14745	2	q2103	q2205	q2100	q2205
22	*DIA1*	25080	1	cen	qter	cen	qter
22	*GALB*	23050	1	q1300	qter	q1300	qter
0X	*GALA*	30150	8	q2205	q2509	q2300	q2500
0X	*G6PD*	30590	11	q2706	q2806	q2800	qter
0X	*HPRT*[a]	30800	11	q2702	qter	—	—
0X	*PGK*	31180	11	q1108	q2209	q1300	q2300
0X	*SAX*	31345	1	p1101	qter	cen	qter
0X	*XG*	31470	2	p22	q1104	pter	cen
0Y	*HY*	14315	1	pter	cen	pter	cen
0Y	*IT-Y*	—	1	cen	qter	cen	qter
0Y	*SAT*	—	1	cen	qter	cen	qter

[a] Modified LRO applied

for the *PEPC* and *PGM1* loci are q4203 and p3400, respectively. The values have been translated to cytologic bands to avoid reference to the physical map. Neither estimate is statistically unlikely, as measured by goodness-of-fit-chi-square values of 18.89 (14 *df*) for PEPC and 18.02 (14 *df*) for PGM1. The LRO for PEPC extends from q4103 to q4300, mimicking the SRO, while the likely region for the *PGM1* locus extends from p3203 to p3409. This is compatible with all but two assignments and encompasses a critical region with many breakpoints. Imprecision in defining breakpoints and possible position effects resulting from a

break near the *PGM1* locus could account for the inconsistencies in the PGM 1 data.

Results for the remaining 79 genes appear in Table 3. The LRO, SRO, and number of assignments are listed for each gene. For genes with more than one assignment, the LRO are contained within the chromosome. The data base and reference list are given with physical, cytologic, and genetic equivalences of the point estimate and LRO for each gene in Keats et al. (in press).

Discussion

Our use of distribution theory is somewhat unusual. First a uniform distribution is taken heuristically to suggest an estimate and preliminary weight, which is then refined by a least-squares procedure. The assumed normality of estimates is strictly valid only as the number of estimates increases for each locus, but in practice reasonable results require only a few observations. Normality is a versatile distribution, robustly applicable to approximations, while the weighting function may be improved as more data become available through publication of physical assignments as inclusive segments and extension to high-resolution banding. Any reasonable weighting function emphasizes small segments: the weights used here decrease more rapidly than the reciprocal of the squared segment, reflecting loss of information as the inclusive segment approaches the chromosome in length. The mean value of the pseudo-chi-square (0.989) and the acceptable fit to a χ^2 distribution tend to justify our choice of a weighting function as a good approximation.

An assessment of the SRO and LRO methods depends critically on whether erroneous assignments can be clearly recognized. If so, then the SRO method should give a consistent and correct assignment that does not depend on approximate normality, but it does not use information about segment length to give statistically efficient point estimates and confidence intervals. If erroneous assignments cannot be objectively recognized (and it is difficult to see how interpretation of breakpoints, position effects, and controlling loci could always be self-evident) and if the number of observations is not too small, the LRO method should be more reliable and its point estimate more useful in genetic mapping.

PGL paper No. 192. This work was supported by grants GM 23021 from the U.S. National Institutes of Health and 1-475 from the National Foundation.

References

Keats, B. J. B., Morton, N. E., Rao, D. C., Williams, W. R.: A source book for linkage in man. Baltimore-London: Johns Hopkins University Press (in press, 1979)

Paris Conference (1971), Supplement (1975). Standardization in human cytogenetics. Birth Defects: Original Article Series, Vol. 11, No. 9. New York: The National Foundation 1975

Received June 26, 1978 / October 9, 1978