Chapter 1

DESPAIR

DESPAIR is a program to help in designing linkage studies for searching the whole autosomal genome. Originally created for a study comprising affected pairs of relatives of a particular type, the latest version of DESPAIR has been modified to further incorporate discordant relative pairs into the study. The program can be used to determine, for specified power and significance level, the optimal two-stage study design – i.e., how many pairs of relatives should be studied, how many equally spaced markers should be used initially, and what criterion should be used to specify the markers around which further searching should be done. Alternatively, the program will calculate either the number of relative pairs required for a given number of first-stage markers, or the number of markers required for a given number of relative pairs.

1.1 Limitations

1.1.1 Theoretical Limitations

The method used assumes that independent pairs of relatives of a single particular type (full sibling, half-sibling, grandparent-grandchild, avuncular, or first cousin) are being sampled. Only three levels of interference are considered, corresponding to Haldane's mapping function (no interference), Kosambi's mapping function (moderate interference), and Morgan's linear mapping function (extreme interference). The spacing between markers is not allowed to be less than one tenth of a centimorgan, nor as much as one morgan, and markers are assumed to be in linkage equilibrium. Two test statistics are allowed for in the cases of sibling pairs, but only one (that based on the mean test) is implemented for designs that use both affected and discordant pairs.

1.2 Theory

It is well understood that linkage of a putative disease locus to a polymorphic marker can be conducted through a study design of affected pairs of relatives, and this is usually the most powerful sampling strategy for binary traits (Blackwelder and Elston, 1985; Risch, 1990). However, recent research shows that, under certain situations, using discordant relative pairs can be as powerful as, or even more powerful than, using affected relative pairs. Moreover, combining discordant with

affected relative pairs provides a more valid and reasonable study from both a theoretical and practical point of view (Guo and Elston, 2000). Specifically, linkage can be studied by typing pairs of relatives and examining the proportions of the pairs sharing 0, 1, or 2 alleles identical by descent (IBD) at the marker locus. The test for linkage in DESPAIR is based on either the proportion of pairs sharing 0 alleles IBD or the mean proportion of marker alleles shared IBD, which depend on the type of relative pair.

Denote the expected values of either of these proportions under the null hypothesis of free linkage π_0. If there is linkage, the expected values are $\pi_0 + \delta_c$ and $\pi_0 - \delta_d$, corresponding to a design using affected relative pairs alone and a design using discordant relative pairs alone, respectively; δ_c and δ_d are the expected deviations respectively for affected pairs and discordant pairs due to linkage. Both these measures depend not only on the type of relative pair, but also on the recombination fraction θ between the marker and disease loci. In addition, δ_c depends on the relative recurrence risk of disease, due to the disease locus, to first degree relatives of affected persons:

$$\lambda = \frac{Pr(\text{first degree relative of affected person is affected})}{Pr(\text{random member of population is affected})}$$

and δ_d depends on the corresponding relative non-recurrence risk ratio for an affected-unaffected first degree relative pair:

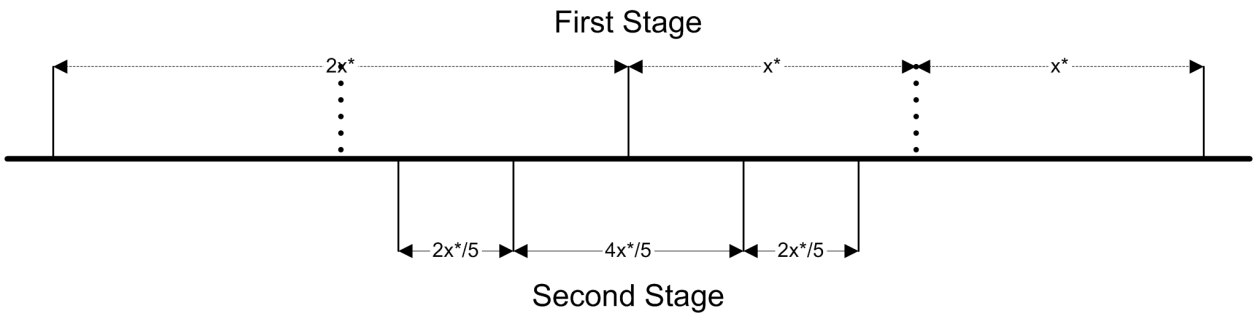
$$\lambda^- = \frac{Pr(\text{first degree relative of affected person is unaffected})}{Pr(\text{random member of population is unaffected})}.$$

Each of these relative risks, often called risk ratios, can be to either a parent/offspring (λ_o,λ_o^-) or to a full sibling (λ_s,λ_s^-).

If several disease loci act multiplicatively, the relative risk is the product of λ's, one for each locus. For a study design that combines affected relative pairs with discordant relative pairs, the test statistic is based on the notion that, in the presence of linkage, affected relative pairs are expected to share a larger proportion of marker alleles IBD, whereas discordant relative pairs are expected to share a smaller proportion of alleles IBD. The difference in the proportion of alleles shared IBD between affected pairs and discordant pairs is quantified by Δ, a weighted difference in the deviations of the mean proportions from π_0. Δ equals zero under the null hypothesis of no linkage, and is greater than zero when linkage is present. The values of Δ can be expressed as a function of θ, λ, λ^-, and the ratio (r_p) of the number of affected relative pairs to the number of discordant relative pairs that are sampled. Values of $\pi_0 + \delta_c$ were given by Risch (1990), and values of $\pi_0 - \delta_d$, and Δ were given by Guo and Elston (2000), for five relative pairs: full sibling, half sibling, avuncular, grandparent-grandchild, and first cousin.

The test based on the proportion sharing 0 alleles IBD and the mean test give identical results except in the case of full sib pairs. The test based on the proportion sharing 0 alleles IBD is not implemented for designs using both concordant affected and discordant full sib pairs.

Assume that at a first stage, *m* fully informative markers, equally spaced along an autosomal genome *M* morgans long, are determined on *n* pairs of relatives of a particular type. For each marker, a one-sided test is performed at the α^* significance level to decide whether the sample proportion of alleles shared IBD deviates significantly from π_0, suggesting linkage. Around each marker suggesting linkage at the first stage, a further 2*k* fully informative markers are tested for linkage at a second stage, assuming that these are placed, *k* on either side of the first stage marker, to span in an optimal manner the interval of interest suggested by the significant first-stage marker (see Figure 1.1).



Genetic map segment (thick horizontal line) showing the placement of three first-stage markers (solid vertical bars above the line) spaced 2x* = M/m Morgans apart, and 2k = 4 second-stage markers (vertical bars below the line) spaced 2x = 2x*/(2k+1) Morgans apart. Dotted vertical bars indicate midway points between the first-stage markers.

Figure 1.1: Stage-1 and Stage-2 Marker Placement

Assume that the study is designed to have power 1 - β of detecting a disease locus with relative risk ratio λ at a significance level α at the second stage, and that there are actually *d* such disease loci present. Finally, assume that the cost of recruiting a person into the study is *R* times the cost of determining one marker on one person. Under these assumptions, if at most one first stage marker is linked to any disease locus, the expected cost of the study is proportional to

$$2n\{R + m + 2k[\alpha^* m + (1 - \beta)d]\} \tag{1.1}$$

However, because there may be more than one first-stage marker linked to the disease locus, the total expected cost is more appropriately reflected by

$$C = 2n\{R + m + 2k[\alpha^*(m - \sum_{i=1}^{d} l_i) + \sum_{i=1}^{d}\sum_{j=1}^{l_i}(1 - \beta_{ij})]\} \tag{1.2}$$

where l_i is the number of first stage markers linked to disease locus *i*, and $1 - \beta_{ij}$ is the probability that 2*k* second stage markers are typed around marker *j* that is linked to disease locus *i* (Ziegler et al. 2001). In this revised version of DESPAIR, which implements cost function (1.2), users have the option to input a maximum distance (g) between any disease locus and a "linked" marker. Then significant results obtained within *g* morgans from any disease locus are considered to be successes, and any outside that range are considered to be false positives. By making the distance *g* small in comparison to the distance between first stage markers, for a large number of markers cost function (1.2) approaches cost function (1.1), which was the function used in the original version of DESPAIR.

Given α, β , λ, *R*, *d*, *g*, *M*, the type of relative pairs, and the type of data (affected relative pairs, discordant relative pairs, or both discordant and affected pairs: for the latter two cases, λ^- must also be specified; and for the last one case, the ratio (r_p) of the number of affected to the number

of discordant relative pairs to be sampled must also be specified), DESPAIR finds the values of *m*, *n*, and α^* that minimize this expected cost for different mapping functions (linear, Kosambi's, and Haldane's), and for values of *k* from 0 (a one-stage design) to a specified maximum value of *k*, subject to the limitation *M* < *m* < 1000M (i.e., the markers must be spaced less than one morgan apart, and must be no closer than one tenth of a centimorgan apart). There is an option (c) to include the cost of screening the population to find the desired sample (the cost of screening is taken to be the same as the cost of recruiting), in which case the user must also enter the proportion of the screened population (r_s) that becomes the final sample.

It is assumed that *n* is large enough, in determining the test criterion corresponding to α^* and β, that the distribution of the proportion of pairs sharing 0 alleles IBD or the mean proportion of marker alleles shared IBD is normally distributed. However, in the case of α, which is typically much closer to zero, there is the option of using either this same approximation assumption (the approximate method), or exact binomial distribution probabilities (the exact method, not implemented for the case where the sample includes both affected and discordant pairs).

To allow for less than fully informative markers, a value of the polymorphism information content (PIC), which measures the markers informativeness (assumed to be the same for all markers), can be specified. This is converted by the program to the corresponding type-of-pair-specific LIC value (Guo and Elston, 1999; Guo et al. 2002). Similarly, a fraction *h*, heterogeneity, can be specified that represents the proportion of the sample pairs affected due to causes other than segregation at the linked locus (in this case one would typically specify a large value for λ and/or a small value of λ^-).

Further details of the method are given in the references.

1.3 Running the Program

DESPAIR can be run by clicking on the DESPAIR GUI link on the S.A.G.E. website

http://darwin.cwru.edu/despair/

and inputting one or more sets of parameters for which the sample size (numbers of affected sib pairs and/or number of markers) is desired. The parameters may be specified as follows:

`parameter`	**Explanation**
`relative_pair_type`	Specifies relative pair type.
	Value Range: **S** (full siblings), **G** (grand-parental), **A** (avuncular), **H** (half siblings), **C** (first cousins)
	Default Value: S
	Required: Yes
	Applicable Notes: None
`concordance_type`	Specifies the phenotypic concordance status (adb) of the observations.
	Value Range: **A** (affected relative pairs), **D** (discordant relative pairs), **B** (both)
	Default Value: A
	Required: Yes
	Applicable Notes: None
`method`	Specifies the analysis method to be used.
	Value Range: **A** (approximate), **E** (exact)
	Default Value: A
	Required: Yes
	Applicable Notes: 1
`significance`	Specifies the statistical significance level α.
	Value Range: (0, 1)
	Default Value: 0.000101
	Required: Yes
	Applicable Notes: 2
`power`	Specifies the statistical power level $1 - \beta$.
	Value Range: (α, 1)
	Default Value: None
	Required: Yes
	Applicable Notes: None

`test_statistic`	Specifies the type of test statistic to be employed (mp).	
	Value Range	**M** (mean statistic) **P** (proportion statistic)
	Default Value	**M**
	Required	Yes
	Applicable Notes	3
`offspr_recurrence_risk`	Specifies λ_o, the locus-specific relative recurrence risk ratio of disease for an offspring of the affected person.	
	Value Range	(1, $+\infty$) for ARPs (0, 1) for DRPs
	Default Value	None
	Required	Yes
	Applicable Notes	3
`offspr_nonrecurrence_risk`	Specifies λ_o^-, the locus-specific relative nonrecurrence risk ratio of disease for an offspring of the affected person.	
	Value Range	(0, 1) for ARPs (1, $+\infty$) for DRPs
	Default Value	None
	Required	Yes
	Applicable Notes	3
`sib_recurrence_risk`	Specifies λ_s, the locus-specific relative recurrence risk ratio of disease for a sibling of the affected person.	
	Value Range	(1, $+\infty$) for ASPs (0, 1) for DSPs
	Default Value	None
	Required	Yes
	Applicable Notes	3
`sib_nonrecurrence_risk`	Specifies λ_s^-, the locus-specific relative nonrecurrence risk ratio of disease for a sibling of the affected person.	
	Value Range	(0, 1) for ASPs (1, $+\infty$) for DSPs
	Default Value	None
	Required	Yes
	Applicable Notes	3
`cost_ratio`	Specifies the ratio (R) of the cost of recruiting a person to the cost of performing one marker assay.	
	Value Range	(0, $+\infty$)
	Default Value	None
	Required	Yes
	Applicable Notes	None

`num_loci`	Specifies the number (d) of disease loci being analyzed.	
	Value Range	{1, 2, 3, ...}
	Default Value	1
	Required	Yes
	Applicable Notes	None
`genome_length`	Specifies the length (M), in morgans, of the underlying genome.	
	Value Range	{1, 2, 3, ...}
	Default Value	36
	Required	Yes
	Applicable Notes	None
`linked_distance`	Specifies the maximum distance (g), in morgans, between any disease locus and a "linked" marker.	
	Value Range	$(0, +\infty)$
	Default Value	0.4
	Required	Yes
	Applicable Notes	None
`pic`	Specifies the value of the polymorphism information content (PIC) used to constrain marker selection.	
	Value Range	(0, 1]
	Default Value	1
	Required	Yes
	Applicable Notes	None
`heterogeneity`	Specifies the heterogeneity proportion (h) of sample pairs affected due to causes other than segregation at the linked locus.	
	Value Range	[0, 1)
	Default Value	0
	Required	Yes
	Applicable Notes	None
`screening_cost`	Specifies option (c) to include the cost of screening the population to obtain desired pairs.	
	Value Range	**Y** (include the cost) **N** (do not include the cost)
	Default Value	**N**
	Required	Yes
	Applicable Notes	None
`screened_proportion`	Specifies proportion of collected samples in the screened population (r_s)	
	Value Range	(0, 1]
	Default Value	1
	Required	Yes
	Applicable Notes	4

`conc_disc_ratio`	Specifies the ratio (r_p) of concordantly affected to discordant relative pairs to be sampled. Value Range: $(0, +\infty)$ Default Value: None Required: Yes Applicable Notes: 5
`num_stage_one_markers`	Specifies the number (m) of first-stage markers to be used. Value Range: $\{M+1, M+2$ ..., 1000M$\}$ Default Value: None Required: No Applicable Notes: 6
`num_stage_two_markers`	Specifies the maximum value for the number of markers (k) to be typed, during the second stage, on each side of the markers found to be significant during the first stage. Value Range: {0, 1, 2, ...} Default Value: None Required: Yes Applicable Notes: None
`num_pairs`	Specifies the number of relative pairs (n) to be analyzed. Value Range: {1, 2, 3, ...} Default Value: None Required: No Applicable Notes: 6

Notes

1. The `method` parameter is not applicable for sample data comprising both affected pairs and discordant pairs; only the approximate method (**A**) is implemented for such data.
2. The default value for α corresponds to a lod score of 3 if the `method` parameter is set to **A** (approximate).
3. The parameters `offspr_recurrence_risk` and `offspr_nonrecurrence_risk` are used by the proportion test for linkage, while the parameters `sib_recurrence_risk` and `sib_nonrecurrence_risk` are used by the mean test.
4. When the value of the `screening_cost` parameter is set to **N**, the `screened_proportion` parameter will be ignored.
5. When the value of the `screening_cost` parameter is set to **N**, or the `concordance_type` parameter is set to either A or D, the `conc_disc_ratio` parameter will be ignored. In other words, the `conc_disc_ratio` parameter is applicable only when the `concordance_type` parameter is set to **B**.
6. The user may specify a value for either `num_stage_one_markers` or `num_pairs`, but not both. If a value for either one of the parameters is specified, the other will be determined by the program. If neither parameter is specified, the program will determine both.

1.4 Output

DESPAIR produces a Standard Output File that includes:

- Title, version, and date of the program for each problem
- Control values specified by user
- For each k = 0, ..., max k, and for each mapping function, tabulation of optimal values of m and n with corresponding α*, cost (in units of the cost of typing one marker on one person), and the first and second stage marker spacings in centimorgans

For an example output, see the end of this document.

1.4.1 Error Messages

DESPAIR has an error checking routine. Values of any parameter that are out of bounds are not allowed. When an error is detected during the analysis, DESPAIR will identify the error and display the error message associated with it. The error messages that may be displayed are as follows:

- `The following fields were set to values out of bounds: <FIELD LIST>`
- `The exact test is not implemented for the case in which both concordant and discordant pairs are available.`
- `The test based on the proportion sharing 0 alleles i.b.d. is not available. The above results are for the mean test.`