# Prediction of Autozygous Regions in Families

Agilent Technologies' GeneSpring GT 1.0 software implements a novel algorithm to identify regions of autozygosity in affected individuals within families. This tool has demonstrated its ability to identify regions associated with recessive diseases in inbred families[i]. Searching for autozygous regions is most useful in isolated populations with few founders and a limited number of generations between the founders and the affected individuals under study. As such, we refer to these populations as being "recently inbred."

Autozygosity refers to the state of a genetic variation in which the two alleles in an individual are homozygous, as a result of being inherited from a common ancestor carrying the same allele. Autozygous alleles are commonly described as being identical by descent (IBD). It is important to differentiate autozygosity from allozygosity, in which two alleles at a given loci are homozygous by recombination events.

## *Methodology*

This algorithm is an extension of the technique described by Broman and Weber[ii] that addresses multiple individuals simultaneously. It consists of two principal steps:

1. Computing LOD scores for each variation per affected individual. This score indicates how likely the observed value for a variation results from autozygosity as opposed to being drawn from a known population.
2. Computing LOD scores for regions. LOD scores are summed for contiguous variations (on the same chromosome). Regions are made from the contiguous variations that have high LOD scores.

***Step 1: Calculation of Autozygosity LOD Scores:***

The method used to calculate LOD scores compares the null hypothesis that alleles for a variation are not autozygous for the affected people, and compares this to the assumption that the alleles are autozygous from a common ancestor. The analysis makes the assumption that the alleles in question are in Hardy-Weinberg equilibrium, and requires the following input data:

1. allele frequency estimates from genotypes of other individuals in the population
2. an estimate of the error rate $\varepsilon$, and random sampling

The likelihood of observing the measured values under each assumption (autozygous and non-autozygous alleles) can then be computed by multiplying together the observed probabilities for everyone measured in the inbred population. We can now generate a likelihood of observing what was observed given that the segment containing this allele was autozygous (without the allele being known) by making a weighted (by allele frequency) sum of the likelihoods given the particular autozygous assumptions. The LOD score can be calculated by dividing this by the probability of being not autozygous.

| ***Observed*** | ***Autozygous A*** | ***Autozygous B*** | ***Autozygous C*** | ***Not autozygous*** |
|---|---|---|---|---|
| Affected AA | $(1-\varepsilon)+\varepsilon p_A^2$ | $\varepsilon p_A^2$ | $\varepsilon p_A^2$ | $(1-\varepsilon)p_A^F p_A^M+\varepsilon p_A^2$ |
| Affected AB | $2\varepsilon p_A p_B$ | $2\varepsilon p_A p_B$ | $2\varepsilon p_A p_B$ | $(1-\varepsilon)(p_A^F p_B^M+p_B^F p_A^M)+2\varepsilon p_A p_B$ |
| Parent AA | $(1-\varepsilon)p_A/(2-p_A)+\varepsilon p_A^2$ | $\varepsilon p_A^2$ | $\varepsilon p_A^2$ | $p_A^2$ |
| Parent AB | $(1-\varepsilon)2p_B/(2-p_A)+2\varepsilon p_A p_B$ | $(1-\varepsilon)2p_A/(2-p_B)+2\varepsilon p_A p_B$ | $2\varepsilon p_A p_B$ | $2p_A p_B$ |

Table 1. **Likelihoods of observing allele pairs AA or AB in individuals autozygous and non-autozygous for the alleles A, B, and C.** The symbols $p_A$, and $p_B$, represent the probability of encountering the alleles A and B in the population, and ε is the combined rate of genotyping errors and mutations. By symmetry, each other possible combination of alleles can be taken from this table. More complex pedigree information (such as grandparent genotype information) could be used with straightforward modifications of this table. Note that the first two rows reduce to table 1 in reference ii if $p_X^M=p_X^F=p_X$ and you do an allele frequency distribution weighted sum of the autozygous X columns to get the autozygous column in [ii].

Formally, if O is the set of genotype measurements believed to come from a single founder, o is a genotype in O, and Pr(o | autozygous X) and Pr(o | not autozygous) come from table 1:

$$Pr(O \mid autozygousi) = \prod_{o \in O} Pr(o \mid autozygousi)$$

$$Pr(O|not\ autozygous) = \prod_{o \in O} Pr(o|not\ autozygous)$$

$$Pr(O|autozygous) = \sum_i p_i\, Pr(O|autozygous\, i)$$

$$LOD = \log_{10} \frac{Pr(O|autozygous)}{Pr(O|not\ autozygous)}$$

### *Step 2: Calculation of Regional LOD Scores*

In order to calculate a regional LOD score, this method makes the assumption that adjacent alleles are mutually independent (i.e., no linkage disequilibrium). This is clearly an incorrect assumption for dense SNP maps, and can lead to reported regional LOD scores being higher than reasonable; however the results are still frequently still useful as the highest scoring region(s) are good candidates for investigation by other methods. This allows us to define the odds ratio for a region as the product of all of the individual odds ratios. Note that this is equivalent to adding all of the LOD scores within a defined region. The following method is used to find consecutive regions in which the sum of the LOD score is highest.

Note that these regional LOD scores calculated from a dense SNP map are not comparable to traditional multipoint LOD scores used for measuring genetic linkage in a sparse map. Because linkage disequilibrium is significant in a dense SNP map the normal multipoint independence assumptions are incorrect, and the regional autozygosity LOD scores tend to be significantly higher than low-density-linkage LOD scores. Do not be surprised if you find more than one region with a LOD score above 5. In practice, these regional LOD scores are still useful, even if they are not conclusive, as they are more sensitive than the single point scores.

## *Implementation*

In GeneSpring GT 1.0, three pieces of data are created for each autozygosity analysis:

- a list containing each measured variation and its respective point LOD score (calculated in step 1 above)
- a list containing each measured variation and its regional LOD score (calculated in step 2 above). Note that variations within the same region are given an identical LOD score.
- a list containing the location of contiguous segments of DNA where the regional LOD scores are above some user defined threshold

Taken together, these lists can be used to generate images that simplify the search for potentially autozygous regions. In figure 1, black bars represent sequence regions where the regional LOD scores were higher than 5.

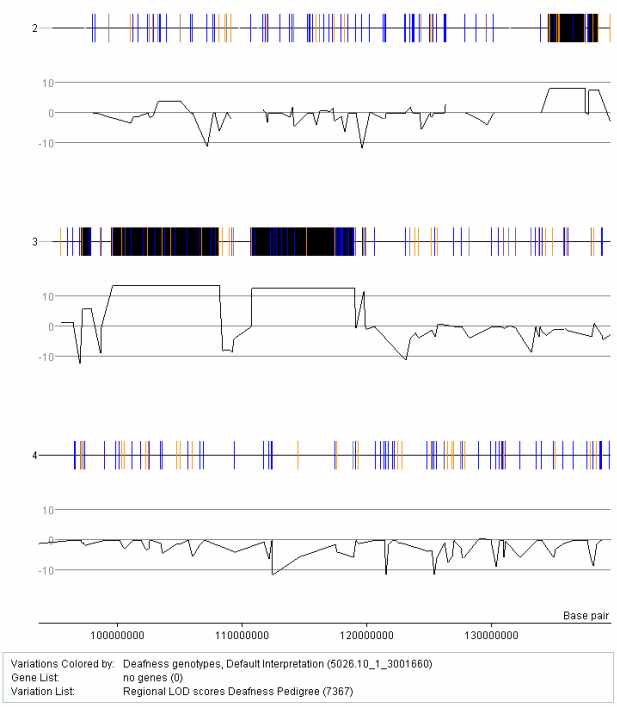


**Figure 1. Graphical representation of potentially autozygous regions.** This image, generated by GeneSpring GT, depicts regions in chromosomes 2, 3, and 4 that are highly likely to be autozygous. Each measured variation is represented as a vertical line crossing the chromosome. Black rectangles represent segments where the regional LOD score was higher than 5. Below each chromosome is a graph depicting the regional LOD score for each variation.

### *Application to Inbred Populations*

This technique is very powerful for recessive traits in inbred populations with a common ancestor with a small enough number of generations since the ancestor that several of the measured variations around the mutation site will have passed to each of the affected descendents. It will not work with non-recessive traits, although you could try it with close-to-recessive traits. It will not work well with multiple ancestors with the disease allele (though if you have two or more distinct inbred populations you could analyze them separately and then add the resulting LOD scores).

---

[i] Stephan, D.et al. (2004) “Title Here” *Manuscript in Preparation*.

[ii] Karl W. Broman and James L. Weber (1999). Long Homozygous Chromosomal Segments in Reference families from the Centre d'Étude du Polymorphisme Humain. *Am. J. Hum. Genet. 65*:1493-1500.

[i] Puffenberger et al. (2004) “A High-density SNP Genome Scan Identifies *TSPYL* Loss-of-function as Causitive of Swarey Syndrome” *Manuscript in Preparation*.



sig_support@agilent.com | Main 866.744.7638