Pacific Symposium on Biocomputing 13:190-200(2008)

GATHERING THE GOLD DUST: METHODS FOR ASSESSING THE AGGREGATE IMPACT OF SMALL EFFECT GENES IN GENOMIC SCANS*

MICHAEL A PROVINCE, PH.D.

AND

INGRID B BORECKI, PH.D.

Division of Statistical Genomics Box 8506
Center for Genome Sciences
Washington University School of Medicine
4444 Forest Park Blvd
St. Louis, MO, 63108, USA

Genomewide association scan (GWAS) data mining has found moderate-effect “gold nugget” complex trait genes. But for many traits, much of the explanatory variance may be truly polygenic, more like gold dust, whose small marginal effects are undetectable by traditional methods. Yet, their collective effects may be quite important in advancing personalized medicine. We consider a novel approach to sift out the genetic gold dust influencing quantitative (or qualitative) traits. Out of a GWAS, we randomly grab handfuls of SNPs, modeling their effects in a multiple linear (or logistic) regression. The model’s significance is used to obtain an iteratively updated pseudo-Bayesian posterior probability associated with each SNP, which is repeated over many random draws until the distribution becomes stable. A stepwise procedure culls the list of SNPs to define the final set. Results from a benchmark simulation of 5 quantitative trait genes among 1,000, in 1,000 random subjects, are contrasted with marginal tests using nominal significance, Bonferroni-corrected significance, false discovery rates, as well as with serial selection methods. Random handfuls produced the best combination of sensitivity (0.95) specificity (0.99) and true positive rate (0.71) of all methods tested and better replicability in an independent subject set. From more extensive simulations, we determine which combinations of signal to noise ratios, SNP typing densities, and sample sizes are tractable with which methods to gather the gold dust.

1. Introduction

The Gold Rush of the 1840s and 50s produced a flood of prospectors in the American West. Fortified by dreams of easy discovery and driven by a desire for great fame and wealth, the only thing that separated the bold visionaries from

* This work is partially supported by NIH grants HL087700, HL088655 DK068336, and AG023746.

the reckless fools was the luck of where they staked their claims. Today, a "Genetic Gold Rush" is taking place. Scientists compete with one another to be the first to find novel complex disease genes. Using the increasingly affordable technology of Genome-Wide Association Scan (GWAS) SNP chips, new scientific prospectors are becoming inspired by the early successes such as macular degeneration[1] and obesity[2]. Like the initial 1844 discovery of gold at Sutter's Mill, these first GWASes have generated much excitement and high expectation that all such searches will be simple, straightforward, and lucrative. With talk of "low hanging fruit" there is much optimism that the dream of personalized medicine may actually be around the corner.

While the first GWASes appear to have discovered some new signals, these do not appear to explain a large portion of the variance in the target traits that our heritability estimates would indicate are in the genome. For example, the FTO gene identified by Frayerling et al., (2007)[2] is homozygous in only 16% of adults (in Caucasian populations) and increases risk for obesity by 1.67-fold (approx 3kg of weight on average). As the total heritability for BMI is roughly in the 50% range, the FTO gene is clearly only one small "gold nugget" in the entire genomic treasure. Over 40 years experience with painstaking candidate gene work in humans would indicate that there may be many genes of small effect for complex traits. For instance, the AGT gene has been studied since the 1980s as a strong candidate for hypertension, but has been estimated in large populations to explain only 0.1% of the variance[3] which may explain why some studies have detected an effect while others have not. Many genomic scanning techniques will fail to find such small effect size genes, which individually are trivial, but which in the aggregate may explain a substantial part of the variance, especially when epistatic interactions are considered[45]. Instead of concentrating only on finding the relatively few gold nuggets, perhaps we should consider ways to gather many small effect "gold dust" variants, not because any one grain is by itself important, but because in the aggregate, a pouch of gold dust can be very valuable.

2. Methods

We consider several alternative methods for "gathering the gold dust" in a genomic scan.

2.1. *Univariate Screening*

The traditional method for identifying signals in a GWAS is to screen one SNP at a time, and choose the most significant SNPs. In the univariate screening category, we considered 3 variations, choosing:

1. N = all SNPs that are nominally significant (α= 0.05)
2. B = all SNPs that are significant at the Bonferroni level ($\alpha = 0.05/M$)
3. F = all SNPs that are significant using the False Discover Rate.

Intuitively, if there are no epistatic interactions between SNPs, and all signals operate additively and independently, then there is little information to be gained about the impact of any one SNP on the phenotype by considering other SNPs simultaneously. In that case, we might expect that some kind of univariate screening procedure which evaluates the marginal effects of each SNP would give the most efficient and powerful estimates of the genomic signals. However, if the actions of genetic signal variants are more complex, if there are epistatic interactions, genes which down and up regulate the action of other genes, AND/OR logic "gate-keeper" variants which are required to be present for other SNPs to have any appreciable impact, particular haplotypic combinations that increase or decrease phenotypic risk, or even environmental factors that potentiate and reveal a set of related genes in a pathway, then we may miss such signals in univariate screening, since we only examine the marginal impact of each SNP. In that case, we might prefer a method which examines SNPs in combination, using a multivariate modeling approach.

2.2. *Random Handfuls*

The random handful approach is a pseudo-Bayesian algorithm, in which we serially update information that a SNP is a signal for a phenotype, based upon the results of randomly drawn multivariate models predicting that phenotype. Let $\mathfrak{I} = \{ S_i \mid i = 1,2, \ldots, M\}$ be the set of SNPs in a GWAS, of size "M." Let H^* be the set of true signal SNPs for a fixed phenotype Y. SNPs in H^* are either themselves causative functional variants (which would be the most extraordinary luck), or more likely, they are SNPs in LD with causative variants. Then $(\mathfrak{I} \setminus H^*)$ is the set of noise SNPs for Y.

Let $P_0[S_i \in H^*]$ be any prior probability density function (p.d.f.) on $\mathfrak{I}$. If we wish to remain agnostic about the genetic causes of Y, then the initial priors will be flat for all SNPs, i.e. $P_1[S_i \in H^*] = (1/M)$. If we wish to incorporate prior knowledge about the genetic architecture of Y, either from some other association or linkage scan, or from a microarray experiment, or from biological knowledge of the genetic pathways, then we may suitably choose some other prior $P_1[S_i \in H^*]$.

Let $H \equiv \{S_j\}$, be a set of SNPs of a fixed size (say L), drawn at random from $\mathfrak{I}$, which we will call a "random handful". Let M(H) denote the multivariate model predicting Y from H. If Y is a continuous phenotype (e.g. lipid level), M(H) is a multivariate regression model:

$$E[Y \mid H] = \alpha^H + \Sigma_j (\beta^H{}_j S_j) \qquad (1)$$

If Y is a categorical outcome (e.g. diabetes), then M(H) is a multivariate logistic regression model:

$$\text{Prob}[Y \mid H] = 1 / (1+ \exp(\alpha^H + \Sigma_j (\beta^H{}_j S_j)) \qquad (2)$$

If Y is a survival outcome (e.g. age-at-onset of hypertension), then M(H) is a multivariate proportional hazards model:

$$Y(t)\, d\, \Lambda(t, H) = Y(t) \exp(\alpha^H + \Sigma_j (\beta^H{}_j S_j))\, d\, \Lambda_0(t, H) \qquad (3)$$

where $\Lambda_0(t, H)$ is the baseline hazard function

In all three cases, it is trivial to include additional fixed effect covariates into these models, such as age, sex, diet, lifestyle, exposures, etc. by simply adding more linear combination terms, i.e. the extended model becomes $(\alpha^H + \Sigma_j (\beta^H{}_j S_j) + \Sigma_j \gamma_j X_j)$ for covariates X_j. Even when they are not the main focus of our research, proper modeling of important covariates can reduce unexplained variance and therefore increase our power to detect gene signals (e.g. see Province et al., 2004[6]). However, as incorporation of such covariates is simple in both the random handful and the univariate screen approaches, in order to keep the notation crisp and to maintain our focus, we will ignore this complication here and concentrate on the case of no non-genetic covariates.

Based upon the results of any model M(H) predicting Y, we can update information on the probability that each SNP S_i in H is amongst the signals for Y (i.e. that $S_i \in H^*$). If M(H) predicts Y well, then it is more likely that H contains some signal SNPs, so we raise the posterior probability that each S_i in H is a signal. Conversely, if M(H) predicts Y poorly, then the component SNPs in H are less likely to be signals, so that $S_i \notin H^*$ is more probable and we lower the probability that it is a signal. As we randomly pick many random handfuls of SNPs and evaluate each of those multivariate models, any given SNP S_i will be sampled many times in the context of many other background SNPs (many different Hs), so we get better and better estimates of the probability that S_i is a signal. At stage 1, we use the intial prior $P_1[S_i \in H^*]$. With each new random handful, H_k, we serially update the posterior probabilities for each SNP S_i, so that the posterior from the kth random handful becomes the prior for the next (k+1)th random handful, as depicted in Figure 1.

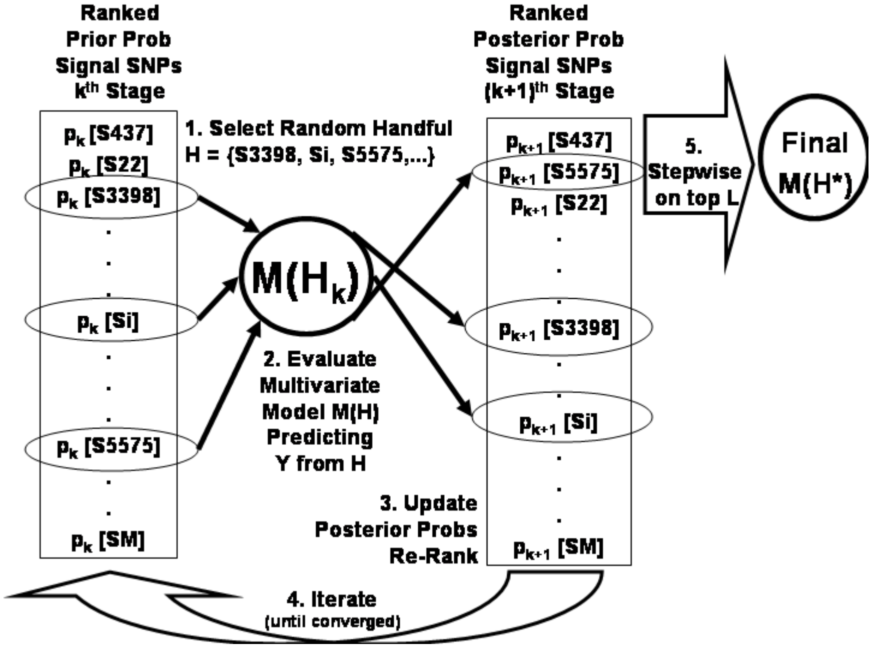

Figure 1. Random Handful Algorithm Given prior probability rankings at stage k that SNPs are signals : 1) we randomly select a handful of SNPs, Hk of size L. 2) Next, we evaluate the multivariate model M(Hk) predicting the phenotype, Y, from Hk. 3) We update the posterior probabilities at stage (k+1) to get new rankings that SNPs are signals. 4) Posterior probabilities now become the Priors for the next Hk+1 random handful. The procedure is terminated when the top L handful of SNPs is consistently and stably ranked at the top in successive updates. 5) Finally, a standard stepwise algorithm is applied to the (now stable) top L set, to select the final significant independent set of SNPs, forming the multivariate model M(H*).

Formally, at stage k, if $P_k[S_i \in H^*]$ is the current prior that S_i is a signal, then given the results of the k^{th} multivariate model $M(H_k)$ for a random handful of SNPs containing S_i, the posterior probability that S_i is a signal SNP is:

$$P_{k+1}\left[S_i \in H^* \mid M(H_k)\right] = \frac{P\left[M(H_k) \mid S_i \in H^*\right] P_k\left[S_i \in H^*\right]}{P\left[M(H_k) \mid S_i \in H^*\right] P_k\left[S_i \in H^*\right] + P\left[M(H_k) \mid S_i \notin H^*\right] P_k\left[S_i \notin H^*\right]} \qquad (4)$$

We approximate $P[M(H_k) \mid S_i \in H^*]$ by the power of the multivariate model $M(H_k)$ to detect SNP $S_i \in H^*$ since this is the probability under the alternative. In calculating this power, we assume all other SNPs in H_k are random noise, so that all of the R^2 from the multivariate model $M(H_k)$ is due to the single SNP S_i. Of course, this assumption may not be correct since H may contain many other signal SNPs as well, and we could weight the various possibilities by their corresponding probabilities (again making certain assumptions about that distribution) to obtain a “better” estimate of the desired conditional probability. However, there are 2^L possible ways for the L SNPs in H_k to be distributed as either noise or signal SNPs, which is too many to exhaustively evaluate. For the purposes of this algorithm, we are not as much interested in the exact value of the posterior probability that a SNP is a signal, as we are in approximately ranking the SNPs according to those probabilities. Again, we need not spend an inordinate amount of effort to ensure we have gathered every tiny last grain of gold, so long as we wind up with a pretty valuable pile that is mostly gold in the end.

We also use the current k^{th} stage prior distributions $P_k[S_i \in H^*]$ in selecting which SNPs to include in the corresponding random handful H_k to evaluate for the next stage. We form the cumulative distribution of the priors across all M SNPs (normalizing by the sum) and then select H_k via importance sampling according to this distribution. In this way, the SNPs with the currently highest probabilities of being a signal, are more likely to be included in the model $M(H_k)$. This strategy has two advantages. First, if the current probabilities are accurate, we are efficiently refining the posterior probabilities for the most likely signal SNPs, and thus likely to converge more quickly to the correct set of signal SNPs H*. Second, if the higher probability for a SNP S_i is inaccurately high for some reason (e.g. by the luck of the draw, it has always accidentally been in the company of other signal SNPs in each previous H_j in which it was evaluated, even though it itself is not a signal), then preferentially sampling such a SNP in H_k gives us the opportunity to correct the probability for S_i with an additional model.

We approximate $P[M(H_k) \mid S_i \notin H^*]$ as the product of the overall p-value for the multivariate model $M(H_k)$ times the type-III SSQ p-value for the particular SNP S_i in the model (i.e. the p-value for the test of the independent contribution of S_i to Y). This is also not strictly speaking correct, as the model is not independent of one of its components. But again, we prefer a simple approximation to a more complex solution at this stage. This approximation is intuitively appealing, as we would like to include call a SNP as a signal if both the multivariate model which contains it is significant as well as if its conditional test of independent contribution is significant.

We continue iterating until the algorithm consistently ranks the top L SNPs from one stage to the next. Then we take these top L SNPS as potential signals. The final step is to do a traditional stepwise model, selecting only the significant SNPs from the top L. This insures that each SNP in the final random handful model is making an independent contribution to prediction of the phenotype.

The Random Handfuls algorithm is written as a SAS MACRO (SAS™)[7]. Within each iteration, multivariate regression is done using PROC REG and power is calculated using PROC POWER.

2.3. *Simulation*

To evaluate the performance of the methods, we conducted several Monte Carlo simulation experiments in SAS™. As large scale simulations can take excessive amounts of CPU time, our initial experiments have been relatively small, to allow us to explore a more broad space of conditions. In the first series of experiments, we simulate N=1,000 unrelated subjects on which we conduct an M=1,000 SNP scan, with 5 signal SNPs, each of which explains 2% of the variance of Y (a quantitative trait), which therefore has a total heritability of 10%. We generate SNPs without LD. The second simulation is an extension of the first, in which we add 5 pairs of epistatically interacting SNPs (none of which have any main effects). Each of these interactions has a heritability of 2%, for a total trait heritability of 20% for all main effects and interactions. For each condition, we generate 100 replications and analyze with all 4 methods.

3. Results

Results from the simulations are shown in Table 1. We tabulate the average performance across all replications, of each of the 4 methods (Nominally significant, Bonferroni significant, FDR significant, and Random Handfuls algorithm) for finding polygenic SNPs. We can classify each of the M=1,000 SNPs as real signals (including all main effect SNPs as well as each pair of SNPs involved in any interaction) vs. the noise. Thus, in each replication we can calculate agreement statistics for each screening method to capture the signals.

Not surprisingly, when there are no interactions (top half of Table 1), selecting all nominally significant SNPs at $P<0.05$ is highly sensitive (99%.), and the specificity runs at 95%, which corresponds well to the expected 5% false positives. However, the true discovery rate is only 9%, since most positives will be false. Selecting only Bonferroni or FDR significant SNPs trades much of the sensitivity (now down to 66% and 78%, respectively) for increased specificity and much improved True discovery rates (99% and 96%, respectively). There is considerably less noise in our final answer when we correct for multiple comparisons. The random handful algorithm competes well in this regard, having the high sensitivity and specificity of the Nominal criteria with a much higher True Discovery Rate. We also calculate the Kappa statistic, which is the amount of agreement beyond that expected by chance (perfect agreement yields

Table 1 Results from Monte Carlo Simulation
Comparing 4 Methods for Detecting Small Gene Polygenic Aggregate Effects
(N = 1,000 Subjects, K = 100 Replications)

M = 1,000 SNPs Total h^2 = 10%
H* = 5 Additive signal SNPs (h^2g=2% each)

Method	Sensitivity	Specificity	True Discovery Rate	Kappa	Training R^2	Test R^2
Nominal	0.99	0.95	0.09	0.16	0.26	0.01
Bonferroni	0.66	0.99	0.99	0.77	0.08	0.07
FDR	0.78	0.99	0.96	0.85	0.09	0.08
Random Handfuls	0.95	0.99	0.71	0.81	0.12	0.07

M = 1,000 SNPs Total h^2 = 60%
H* = 5 Additive SNPs (h^2g=2% each)+ 5 Epistatic Interaction SNP-pairs (h^2g=10% each)

Method	Sensitivity	Specificity	True Discovery Rate	Kappa	Training R^2	Test R^2
Nominal	0.44	0.95	0.12	0.16	0.37	0.06
Bonferroni	0.27	0.99	0.99	0.42	0.17	0.09
FDR	0.28	0.99	0.93	0.42	0.14	0.08
Random Handfuls	0.34	0.99	0.89	0.49	0.12	0.09

For each of the 4 methods (Nominal, Bonferroni, FDR and Random Handfuls) , we tabulate average agreement statistics (sensitivity, specificity, true discovery rate and Kappa) across all replications, quantifying the agreement of that method to the true classification of SNPs into signals vs. noise (signals are all main effect SNPs as well as any SNP involved in an epistatic interaction). We also show the percent of variance explained (R2) from the sum of risk variants across all SNPs chosen in the corresponding model, both in the original Training dataset (on which the SNPs were originally selected) as well as on an independent Training dataset of equal size.

Kappa=1, while chance agreement corresponds to Kappa=0). Kappa is poor for the Nominal selection method, and in the 75%-85% range for each of the Bonferroni, FDR and Random Handful methods.

We also evaluated the amount of variance explained by the selected SNPs in both the original training dataset (on which the models were developed) as well as on an independent test set of equal size. Since the true heritability in each dataset is 10%, we can see that the Nominal model overfits to noise in the training dataset, producing an R2=26%. Of course, such a model does not reproduce well in an independent test dataset, explaining only 1% of the variance. Each of the Bonferroni, FDR and Random Handful methods run much closer to the expected 10% of explained variance. Thus, the Random Handful method is as good as or sometimes exceeds the operating characteristics of the other methods, when there are no epistatic interactions.

When there are epistatic interactions without main effects (bottom half of Table 1), the pattern of results is similar. None of the methods was very successful in capturing the epistatic interacting SNPs without main effects (including the random handfuls algorithm). However, the random handfuls gave scored at least as good and sometimes better than the other algorithms across all measures of performance.

4. Discussion

Many complex traits (such as obesity, diabetes, heart disease, cancer, etc.) have heritabilities in the 30-50% range. If all of the variance in a 50% heritable trait were due to large effect genes explaining 10% of the variance or more, there would only be 5 such in the genome, and the current statistical methods for genomic scans would easily find these. However, suppose there are only two such large effect genes, explaining 20% of the variance between them, and the remaining 30% of the variance is due to 30 polygenes, each of which explains only 1% of the variance, or worse, 300 polygenes, each of which explain only 0.1% of the variance (as is the order of magnitude in our AGT – hypertension example[3]). Then the current methods which concentrate only on the marginal effects of variants will almost surely fail to find any gold dust.

It is disappointing that the Random Handfuls method was not successful in detecting interactions without main effects. We are currently incorporating a refinement of the algorithm to explicitly find such effects. There are too many possible interactions to consider all of them in a brute force way. For instance, if there are L main effects then there are

$$\binom{L}{2} = \frac{L!}{(L-2)!2!} \tag{5}$$

possible 2-way interactions. If we formally test for each of these in each random handful of size L, we can easily fit to noise. We are examining the utility of a two stage test within each random handful iteration to handle this problem. In the first stage of each iteration k, we add a single variable which is the sum of all 2-way interaction terms that can be made from the L main effects in the current random handful H_k, and test for its significance. If it is not significant, we stop at stage 1 and consider only main effects as before. If the aggregate interaction term is significant then we move to stage 2 and potentially add all pair-wise interactions in a stepwise algorithm. The conditional probabilites in the Bayesian formula for each SNP S_i are the most significant p-values of either the main effect test or of any pair-wise interaction. The idea is that the noise interaction terms will tend to cancel out, so that the aggregate interaction term provides a good 1 d.f. screen to test whether any interactions are present. If they are, then we test for which ones should be incorporated.

Our random handfuls method is similar in spirit and philosophy to Bayesian averaging and model selection methods, in which a lot of genetic work has been done recently (e.g. Viallefont et al., 2001[8]; Blangero et al., 2005[9]; among others). We do not claim that our current algorithm is the best or most optimal method for finding the "gold dust" genes. Much refinement of the technique is possible and under development. But we do believe that methods in this vein can be useful to mine the gold for complex traits, and that more investigators should consider novel ways to find the aggregate effects of small effect genes, instead of fixating on the few gold nuggets.

Acknowledgments

This work is partially supported by NIH grants HL087700, HL088655 DK068336, and AG023746.

References

1. Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, Haynes C, Henning AK, SanGiovanni JP, Mane SM, Mayne ST, Bracken MB, Ferris FL, Ott J, Barnstable C, Hoh J. Complement factor H polymorphism in age-related macular degeneration. *Science*. 2005 Apr 15;**308(5720)**:385-9.
2. Frayling TM, Timpson NJ, Weedon MN, Zeggini E, Freathy RM, Lindgren CM, Perry JR, Elliott KS, et al. The Wellcome Trust Case Control Consortium; Hattersley AT, McCarthy MI. A Common Variant in the FTO Gene Is Associated with Body Mass Index and Predisposes to Childhood and Adult Obesity. *Science.* 2007 Apr 12; [Epub ahead of print] PMID: 17434869
3. Province MA, Boerwinkle E, Chakravarti A, Cooper R, Fornage M, Leppert M, Risch N, Ranade K. Lack of association of the angiotensinogen-6 polymorphism with blood pressure levels in the comprehensive NHLBI Family Blood Pressure Program. National Heart, Lung and Blood Institute. *J Hypertens*. 2000 Jul;**18(7)**:867-76.PMID: 10930184
4. Borecki IB, Province MA, Ludwig EH, Ellison RC, Folsom AR, Heiss G, Lalouel JM, Higgins M, Rao DC. Associations of candidate loci angiotensinogen and angiotensin-converting enzyme with severe hypertension: The NHLBI Family Heart Study. *Ann Epidemiol*. 1997 Jan;**7(1)**:13-21. PMID: 9034402
5. Ludwig EH, Borecki IB, Ellison RC, Folsom AR, Heiss G, Higgins M, Lalouel JM, Province MA, Rao DC. Associations between candidate loci angiotensin-converting enzyme and angiotensinogen with coronary heart

disease and myocardial infarction: the NHLBI Family Heart Study. *Ann Epidemiol*. 1997 Jan;**7(1)**:3-12. PMID: 9034401

6. Province MA, Rice TK, Borecki IB, Gu C, Kraja A, Rao DC. A Multivariate and multilocus variance components method, based on structural relationships to assess quantitative trait linkage via SEGPATH. *Genet Epidemiol.* 2003 Feb; **24(2)**:128-38. PMID: 12548674
7. SAS Institute, Inc. SAS/STAT© Users Guide Version 6, Fourth Edition, *SAS Institute Inc*., Cary, NC, 1989, 846p.
8. Viallefont V, Raftery AE, Richardson S. Variable selection and Bayesian model averaging in case-control studies. *Stat Med.* 2001 Nov 15;**20(21)**:3215-30.PMID: 11746314
9. Blangero J, Goring HH, Kent JW Jr, Williams JT, Peterson CP, Almasy L, Dyer TD. Quantitative trait nucleotide analysis using Bayesian model selection. *Hum Biol*. 2005 Oct; **77(5)**:541-59. PMID: 16596940