



Design of the PGRN mini-GAW semi-synthetic dataset

Revised February 8, 2005

Simulation Team: Michelle Carrillo, Nancy Cox, Warwick Daw, Ross Lazarus, Michael Province (chair), and Marylyn Ritchie

Specific Aims:

1. To find the SNPs and/or haplotypes associated with a specific pharmacogenetic complex trait, using a semi-synthetic dataset. The genotypes are from a real study, but the phenotypes, covariates and treatment variables have been simulated.
2. To attend the PGRN min-GAW and present your own analyses of these data to the PGRN.
3. To compare various analytical methods of pharmacogenetic gene discovery by various investigators.

Background and Significance:

Tenureitis is a disease characterized by an obsessive drive for meaningless job security. Symptoms include: repeated, manic episodes of impulsive, vaguely thought-through research; a compulsive pursuit of ever more expensive and ambitious grants; and a paranoid world view in which every other scientist is perceived as a rival and a threat¹. The CDC reports that an academic environment puts subjects at a 1000-fold increased risk for tenureitis². One proposed mechanism has been exposure to ivy, or perhaps chalk dust³. It is estimated that tenureitis is responsible for more than 99.44% of the toxic pollution of prestigious scientific journals with thousands of useless, red-herring publications each year, costing billions of dollars for post-docs to debunk⁴⁻⁶. Curiously, once frank tenure has been reached, about half the patients immediately stop all aberrant behavior, turn lethargic, and begin to calcify into petrified wood. The remaining patients progress to the more serious condition, nobelaureatenvy (sometimes called Lander's Syndrome⁷), in which they seem to develop a complete tolerance for the highs formerly associated with grant awards, prestigious publications, and invited platform presentations. At this stage, the patients are no longer satisfied with controlling their own careers and those of their subordinates but begin to suffer delusions of grandeur and feel they must be in control of all other scientists on the planet as well. They create impossibly ambitious and fantastically expensive "Projects", "Initiatives", and "Roadmaps" which suck up all available research dollars like a massive black hole. Amazingly, nobelareatenvys are quite persuasive and often succeed in conning the scientific community into committing vast resources towards their harebrained schemes. Examples include the "Human Gnome Project" (complete resequencing of human by garden-gnome hybrid crosses); the "QuackMap Project" (sonic mapping of the migratory patterns of ethnically diverse duck populations) and the "NIN RoadyMap" (addresses and phone numbers of all of the groupies for the rock group Nine-Inch-Nails—actually, this one is quite useful). Nobelareatenvy patients who fail to reach their goal of complete world domination often abruptly leave academia, found their own biotech companies, and then bail out with all the cash before the bubble bursts. Most wind up soliciting investments via e-mail, posing as the son of the Finance Minister of Zimbabwe. Treatment options for tenureitis are limited but include isolation, massive doses of alcohol, and lobotomy⁸. Anecdotal success has been reported with electroshock and/or assignment of administrative duties⁹. Under no circumstances should the afflicted be invited to tell a joke.

The evidence for any genetic influences whatsoever on tenureitis is spotty, indirect, and wholly unconvincing¹⁰. Therefore, the genetics must be complex. It probably involves a lot of neat sounding stuff, like gene by environment interactions, epistasis, context dependencies, system-biologic pathways and other impressive buzz words which provide us with plenty of excuses to continue to get funding even if we don't find anything.

Study Design:

557 (unrelated) untenured patients were recruited into the study who met DSM-IV criteria for tenureitis: [either a MEDLINE search of their CVs confirms at least 100 peer reviewed publications which are meaningless drivel (identifiable from the telltale adjectives “novel” or “nearly-significant”); or they had committed at least one homicide of a scientific rival -- preferably a former colleague]¹. SNP genotyping was done in three specific candidate genes and their flanking regions. For one region, a random subset of 206 of these 557 patients were selected, on whom dense with the idea that a smaller subset of LD tag SNPs could be identified for genotyping in the complete set of 557.

Treatment:

All patients were treated with massive doses of alcohol every night for 2 years, and re-assessed to see how much they loosened up.

Genotypes:

SNPS in three candidate genes were typed on these subjects: a) the *beta-just-regurgitate receptor gene*, which is thought to promote production of excessive review manuscripts⁴⁻⁶, b) the *PI-hat-is2big4em gene*, which is believed to play a role in the submission of unrealistically ambitious research grants³, and c) the *NRA-magnum357 gene*, which is suspected to be involved in the elimination of scientific rivals¹². For each region, SNPs were typed in the gene itself and in the intergenic flanking regions. For the *beta-just-regurgitate receptor* region, 74 SNPs were typed, for the *PI-hat-is2big4em* region 86 SNPs were typed, and for the *NRA-magnum357 region*, 28 SNPs were typed for a total of 188 SNPs per subject. For the *beta-just-regurgitate gene* region, LD patterns were identified and tag SNPs were defined (and a few additional ones added to ensure adequate coverage) in the random set of 206 subjects, resulting in 20 SNPs selected to be done on the entire set of 557 subjects.

Phenotypes:

In addition to the above genotypes, information was collected on all 557 subject on 3 primary response phenotypes: 1) the number of new publications in the treatment period and especially how many and what percent were drivel; 2) the total grant fund dollars awarded as PI in the treatment period, and 3) whether the subject committed “rivalside”, i.e. elimination of a scientific rival, during the treatment period (actual convictions only). Although all subjects were given a standard amount of regular alcohol treatment, the average serum alcohol level was measured over the treatment period. Adverse events were collected, primarily consisting of DUI arrests. Baseline covariates and known/suspected risk factors to tenureitis were also measured once prior to treatment, including sex, age, exposure to ivy and or chalkdust, rank, degree, years since terminal degree, and previous publication record prior to treatment. Details of all variables are given at the end of this document.

Replications

Via the miracle of cloning, 100 replication datasets were generated, keeping the real data portion (genotypes) fixed each time, but drawing phenotypes, side-effects, and their covariates randomly according to a specific causative generating model. Some replications will naturally run “hot” for the gene effect and be easy to detect that gene, while others will run “cold” and be difficult to detect.

Answers:

By default, participants in the PGRN min-GAW will be blinded as to the generating model and the final answers. We will follow the GAW rules and allow people to ask for the answers from the beginning if they prefer. But they must let everyone know if they analyzed the data blinded or unblinded at the time of their workshop presentation. To get the answers, contact Mike Province at mike@wubios.wustl.edu.

There is at least one fairly obvious gene signal, and at least one more subtle one, and maybe others? Keep looking (the truth is out there....).

References:

1. [DSM-IV - Diagnostic and Statistical Manual of Mental Disorders](#) 303.3 Tenuritus an obsessive drive for meaningless job security—how to spot it.
2. Oblivious IM, Obvious UR, Obsequious ET (2004) Risk factors for acute tenureitis at coastal universities: the CDC See-Da-Sea Study Academic Journal of Occupational Health Psychology, Vol. 9, No. 4, 296–305.
3. Leaha P, Vader D, Kanobi O1, D2 R2, PO C3 (3006) The use of time travel to improve the false discovery rate in genome scans. Science, -10: 21-25.
4. Armbruster D, Potts T, Wirth J, Calvin S (2003) Abuse of scientific integrity by duplicate publication increases chances of tenure. Academic Journal of Vanity Press, 44: 225-227.
5. Armbruster D, Potts T, Wirth J, Calvin S (2003) Tenure chances are increased via duplicating publications: scientific integrity abuse. Journal of Academic Vanity Press, 227: 44-225.
6. Armbruster D, Potts T, Wirth J, Calvin S (2003) Scientific integrity abuse increases chances of tenure when publications are duplicated with just a trivial change in the titles and sending them to different journals. Nature Reviews the Journal of Vanity Academic Press, 227: 44-225.
7. Kim JI-II, Kim AA, Kim AB, Kim AC, Kim AD, Kim AE, ..., Kim ZZ, and Lander E (2004) The North Korean Investigation of Tenuritus Officialitus (NOKIT-Off Study): provides support for the Universal Human Sequencing Project: definitive proclamations from our glorious leader (as well as Kim Jong III), which should put to rest any further questions on the whole subject Pyongyang Journal of Ultimate Truth and Happiness 3225:106-110.
8. Winfrey OP, Graham SG, McGraw PH (2002) Extreme tenure makeovers and awesome cornbread stuffing recipes Journal for Journalistic Journalism, 15:201-305.
9. Driftwood OP (2001) 101 neat things to do with electroshock therapy, Volts and Amps, 26:119-203.
10. Li RU, Lee MI, Lea RE, (2001) A single (double?) case study of Siamese twins reared apart who both failed to reach tenure on the exact same date: strong, irrefutable evidence for the genetic nature of type II tenureitis academiatiatentuiatolja. NEJM 1056:221-227.
11. Collins J, Fabio F, Clinton WJ (2004) A novel gene for romance writing is nearly significant as he held her close to his manly chest. Nature, 333:333-333.
12. Heston C, Stalin J, Kahn G, Hun AT (1999) Use of the Second Amendment of the US Constitution to attain tenure: Darwinian academic dynamics ultimately strengthens the gene pool and will cleanse the weak, gutless liberal bias that infests our universities Guns and Ammo, 23: 111-115.

Appendix: Semi-Synthetic Dataset File Formats

Datasets:

Each of the complete set of files are provided in TWO different formats for you to choose. The content of each set is identical, and only the format is different. One set of files are EXCEL spreadsheets, and the other are ASCII TAB-separated formatted files. Each file is a simple data matrix. The first ROW of each file gives the variable names. Subsequent rows are variable values. Each COLUMN is a variable.

MAP File:

1 record/SNP (189 records=Header row + 188 SNPs) This dataset gives the gene regions and the locations of each SNP in each region.

Variables (in the order listed in the file):

- Region = Number of the gene region (1, 2, or 3). These are sufficiently far apart as to be independent (e.g. different chromosomes). All 3 are AUTOSOMAL chromosomal regions
- SNP = Names of the SNP variables (these correspond to the variable COLUMNS in the GENO dataset). The SNP names are all of the form RjSNPi for snp number "i" in region "j". Thus R2SNP25 is SNP 25 in region 2. SNPs are numbered consecutively for each region.
- DIST = Location (distance) of the SNP in BASE-pairs, relative to the FIRST SNP in that region. So RjSNP1 is always DIST=0 for each region j=1,2,3.
- GENE = Name of the gene region (beta-just-regurgitate receptor; NRA-magnum357; or PI-hat-is2big4em)
- RANDONLY = 1 indicates these are the SNPs in gene region which were done ONLY on the 206 randomly selected subjects; 0 indicates all other SNPs (were attempted to be typed on all subjects)

GENO File:

1 record/person (558 records=Header row + 557 subjects) This dataset gives the individual SNP genotypes for each person in each of the 3 gene regions.

Variables (in the order listed in the file):

- ID = ID number of the subject (a random integer assigned to each subject) These match 1-1 to the corresponding ID numbers in the 100 replications of the phenotype datasets TENUREITUS1, ..., TENURITUS100.
- RjSNPi = Names of the SNP variables (these correspond to the ROW values in the MAP dataset of the column variable "SNP"). The SNP names are all of the form RjSNPi for snp number "i" in region "j". They are in order of region and then SNP number within region, i.e. in the order R1SNP1, R1SNP2, ... R1SNP78, R2SNP1, R2SNP2, ... R2SNP28, R3SNP1, R3SNP2, ..., R3SNP86. The individual subject values for each SNP are all coded 11, 12, 22 or *blank* corresponding to homozygote, heterozygote, homozygote and missing. (NOTE: alleles 1 and 2 are coded arbitrarily, and do **NOT** denote either allele frequency, nor wildtype, nor presumed function).
- Random = indicator of the 206 randomly selected subjects who were more densely genotyped in gene region 1 (1=in random group, 0=not in random group)

TENUREITUSn Files:

1 record/subject (558 records=Header row + 557 subjects). There are 100 of these files, corresponding to the 100 simulation (replications) of the "cloning" experiment. The files are labeled TENUREITUS1, TENUREITUS2, ..., TENUREITUS100. Each of these contain the

phenotypes, treatments, side effects and covariates of for each of the .557 subjects. Note that there is only ONE SET OF GENOTYPES for all subjects (GENO file above), but 100 sets of outcomes, responses, phenotypes, treatments, covariates, side-effects, etc. So you will have to merge each of the 100 phenotype files with the single genotype file to do 100 different analyses.

Identifiers:

- IREP = Replication number (1-100). This is the same value for all records in that file and should correspond to the name of the file itself (to make sure you are in the right file!)
- ID = ID number of the subject (a random integer assigned to each subject) These match 1-1 to the corresponding ID numbers in the GENO dataset. The same ID numbers are used over and over again in each of the 100 files.

Phenotypes:

- NPUBS = number of new publications in 2 year treatment period (response)
- NDRIVEL = number of new filler publications in 2 year treatment period (response)
- PCTDRIVEL = % of new pubs which are drivels = NDRIVEL/NPUBS (response)
- RIVALSIDE = destruction of a scientific rival in treatment period (1=Yes/0=No) (response)
- GOTGRANTS = total grant fund dollars (millions) awarded as PI in treatment period (response)

Treatment:

- ALCOHOL = average serum alcohol level during 2 year treatment (treatment, but could also be pharmacokinetic response)

Adverse Events:

- DUI = # arrests for driving under the influence in the treatment period (side effect)

Baseline Covariates/Exposures:

- SEX = 1=Male, 2=Female (# X chromosomes) (covariate)
- AGE = Years (covariate)
- IVY = exposure to ivy for at least 1 year during schooling (1=Yes/0=No) (exposure)
- CHALKDUST = mg of chalkdust in lungs prior to treatment period (exposure)
- RANK = Rank prior to treatment (Instructor, Assistant Professor) (covariate)
- DEGREE = Ph.D, M.D., MD/PhD (covariate)
- DEGYEARS = # years from terminal degree to start of treatment period (covariate)
- PRENPUBS = number of publications prior to treatment period (covariate)
- PRENDRIVEL = number of filler publications prior to treatment period (covariate)
- PREPCTDRIVEL = % of pubs which were drivels = PRENDRIVEL/PRENPUBS (covariate)

If you have any questions regarding these datasets, contact Mike Province at mike@wubios.wustl.edu