HUMAN LINKAGE AND ASSOCIATION ANALYSIS

Processing pedigree and marker data, and other practicalities

Sept 22, 2016

VC analysis example

$V_P = V_G + V_D + V_C + V_R$ …….. (broad vs narrow h^2)

0.3+0.1+0.2+0.4

$V_P = V_G + V_Q + V_D + V_C + V_R$ …(h^2_T vs h^2_Q)

0.2+0.1+0.1+0.2+0.4

$V_P = V_G + V_Q + V_D + V_C + V_R + V_{cov}$

0.2+0.1+0.1+0.2+0.1+0.3

Sources of Errors

- Pedigree errors
  - Diagnostic
  - Non-paternity
  - Unreported adoption or twin status
  - Errors in data entry
    - Gender errors
    - Pedigree structure errors
  - Sample mix-ups
- Genotyping errors
  - Errors in data entry
  - Misinterpretation of pattern on gel i.e., mistyping
  - Mutations may masquerade as errors
- Other
  - Marker allele frequencies
  - Map order
  - Map distances
  - Stratification

Data Cleaning Approaches

- Verify Mendelian inheritance
  - LINKAGE , MERLIN, and other analysis programs:
    - Most will stop with an error message when a Mendelian inconsistency is encountered
  - PedCheck
- Beyond Mendel
  - Unlikely multiple recombination events
- Verify relationships
  - Many Mendelian inconsistencies may imply incorrect relationship designations
  - PREST
  - Others, e.g., RelPair and Relative
- Estimating Allele frequencies
- Admixture

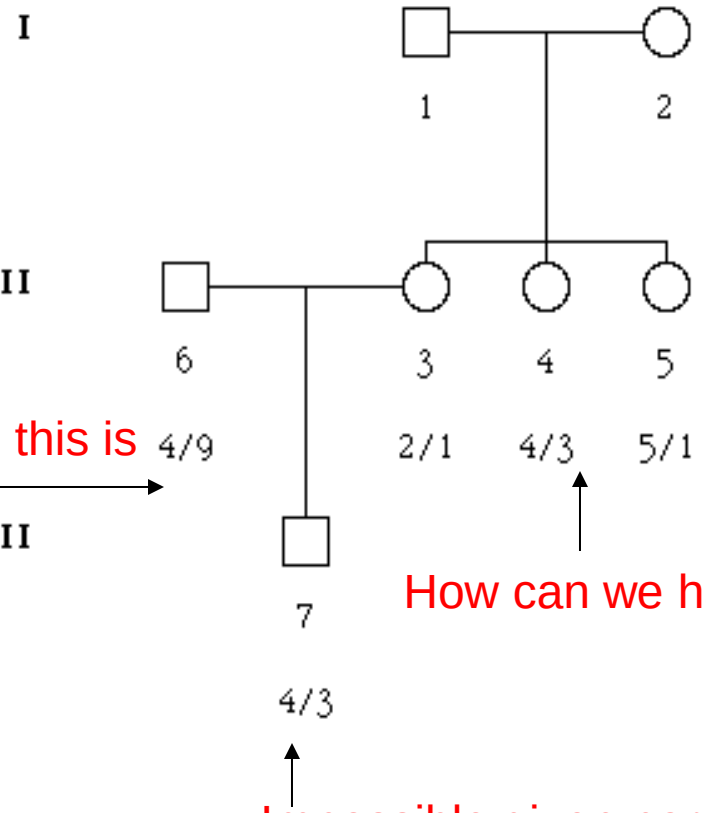
Verify Mendelian inheritance

- LINKAGE and other analysis programs:
  - Most will stop with an error message when a Mendelian inconsistency is encountered
- PedCheck (O'Connell and Weeks, AJHG 63:259-266, 1998)
  - Fast and efficient
  - Handles large data sets with hundreds of markers
  - Gives detailed diagnostic information on the source of the errors
  - Identifies the individuals involved
  - Provides 4 levels of checks: based on nuclear families (simple algorithms) to more complex and powerful checks using large pedigree (complex likelihood based algorithms).

Pedcheck: Level 1 – “Simple” errors

- Nuclear family: checks for inconsistencies between parents and offspring
  - A child and parent's alleles are incompatible.
  - More than 4 alleles in a sibship.
  - More that 3 alleles in a sibship when there is a homozygous child.
  - The allele is out of bounds
  - That in X-linked pedigrees, males are scored as homozygous.
  - A person is half-typed. This is checked because current programs cannot handle this situation.

Figure 1: Pedigree with level 1 errors



Assuming 7 alleles, this is Out of bounds

How can we have alternatives 3,4 and 5?

Impossible given parental genotypes

O'Connell & Weeks,1998

Pedcheck Level 1 Output: The nuclear family algorithm

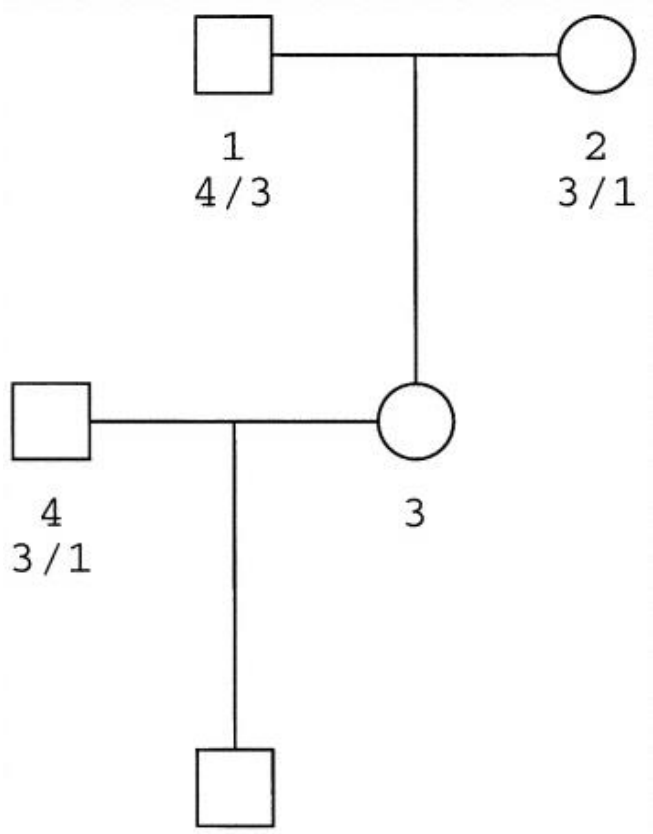
```
CHECKING FOR LEVEL 1 ERRORS
 ##### GENOTYPE ERROR: Pedigree 1   Locus  1   Name c1_m1 #####
 ERROR: Children have more than 4 alleles ( 1 2 3 4 5)

##### GENOTYPE ERROR: Pedigree 1   Locus  1   Name c1_m1 #####
 ERROR: Child 7 and Mother are inconsistent.
 ORIGINAL SCORING:
 Father 6:  9/4   Mother 3: 1/2
 Child 7:  3/4
-----------------------------------------------
! Summary of Errors: By Pedigree   !
! Pedigree 1                       !
!       marker c1_m1               !
!                                  !
! Summary of Errors: By Marker     !
! Marker c1_m1                     !
!       Pedigree 1                 !
-----------------------------------------------
 PedCheck has found 2 inconsistencies in the pedigree data.
```

Pedcheck: Level 2

- Reports errors for a nuclear family only if there were no level 1 errors. If there is a Mendelian inconsistency, it will find it.
- Uses an algorithm (genotype-elimination algorithm) that finds more subtle inconsistencies than level 1.
- Recursive (Lange-Goradia, 1987) algorithm to eliminate inconsistencies.

Figure 2: Pedigree without level 1 errors but with a level 2 error



Phase determined: the ‘2’ is from MOM
But maternal grandparents have no 2!

O’Connell & Weeks,1998

Pedcheck level 2 output

GENOTYPE ERROR: Pedigree 2 Locus 1 Name c1_m2

ORDERED GENOTYPE LISTS: Any allele greater than 4 is set_recoded.
(T) Father 4: 3|1 1|3 (2)
(U) Mother 3: 4|1 4|3 3|1 3|3 (4)
(T) Child 5: 3|2 2|3 (2)

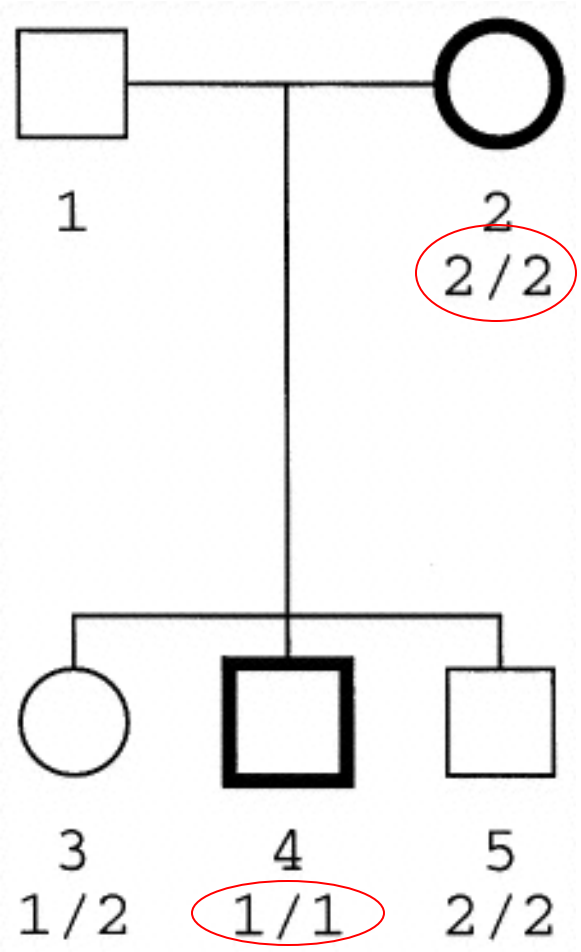
```
-----------------------------------------------
! Summary of Errors: By Pedigree    !
! Pedigree 2                       !
!       marker c1_m2                !
!                                   !
! Summary of Errors: By Marker      !
! Marker c1_m2                     !
!       Pedigree 2                  !
-----------------------------------------------
```

PedCheck has found 1 inconsistency in the pedigree data.

Pedcheck: Level 3

- This level is useful in a complicated pedigree by identifying "critical genotypes" i.e., those typed individuals who, if set to "unknown" will remove the inconsistency in the pedigree

Figure 3: Pedigree with 2 critical genotypes



O'Connell & Weeks,1998

Pedcheck level 3 output

CHECKING FOR LEVEL 1 ERRORS
##
No. 1/1 Processing pedigree 3
The pedigree has 5 persons
GENOTYPE ERROR: Pedigree 3 Locus 1 Name c1_m3
ERROR: Child 4 and Mother are inconsistent.
……..
CHECKING FOR LEVEL 2 ERRORS
……..
CHECKING LEVEL 3 ERRORS.
Untyping any person listed will result in a consistent pedigree at the given locus.
Pedigree: 3
Person 2: 2/2
Person 4: 1/1

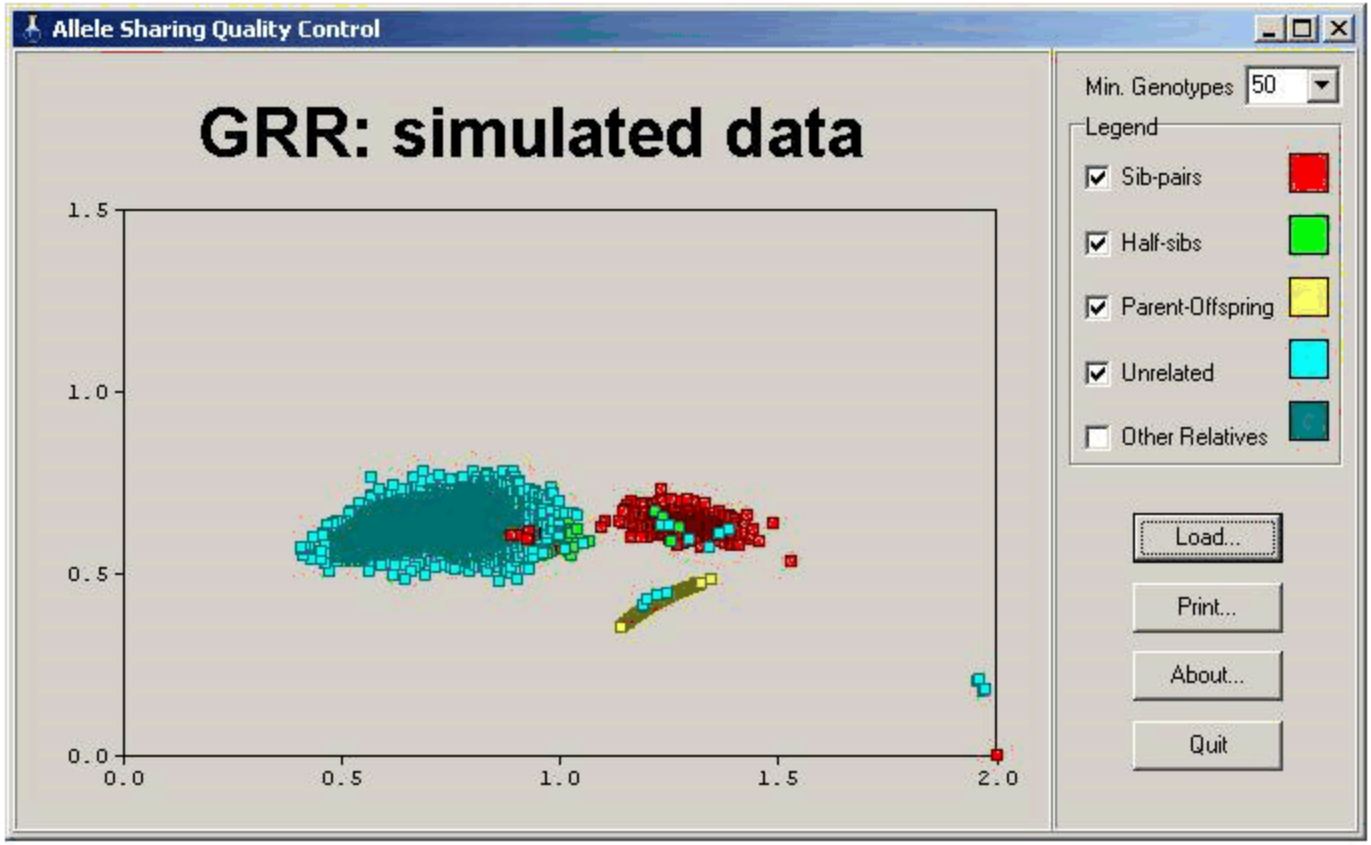
Verify relationships

- Many Mendelian inconsistencies may imply incorrect relationship designations
  - Relative [Goring and Ott, EJHG 5:69-77, 1997]
  - RelPair [Boehnke and Cox, AJHG 61:423-429, 1997]
  - SIBERROR [Ehm and Wagner, AJHG 62:181-188, 1998]
  - **GRR [Abecasis]**
  - **PREST [McPeek and Sun, AJHG 66:1076-1094, 2000; Sun et al., Human Heredity 54:99-110, 2002]**

Graphical **R**epresentation of **R**elationships

By computing the average allele sharing for any pair of individuals in a sample, across all available markers, along with the standard deviation in genome-wide IBS and plotting this mean against the standard deviation, we readily characterize full-sib from half-sib pairs, parent-offspring pairs from siblings, and unrelated individuals from relatives. Each relative class will form a distinct cluster on these plots and outliers from these clusters will represent likely pedigree errors. When including all pairings in a sample, not just pairings within families, we can further detect problems such as sample duplications or perhaps related individuals who have been presumed to be unrelated.

Just needs a .ped file!



STD IBS

Mean IBS

PREST-*plus*

Pedigree **RE**lationship **S**tatistical **T**est

- Relationships:
  1. Full-sib
  2. Half-sib
  3. Grandparents-grandchild
  4. Avuncular
  5. First cousin
  6. Unrelated
  7. Half avuncular
  8. Half first cousin
  9. Half sib plus first cousin
  10. Parent-offspring
  11. MZ twins

Relationship types as encoded in prest and expected IBD probabilities

ID	NAME	CODE	KINSHIP	IBD0	IBD1	IBD2
1	Full-Sibling	FS	0.250	0.25	0.50	0.25
2	Half-Sibling	HS	0.125	0.50	0.50	0
3	Grandparent-grandchild	GPC	0.125	0.50	0.50	0
4	Avuncular	AV	0.125	0.50	0.50	0
5	First-Cousin	FC	0.0625	0.75	0.25	0
6	Unrelated	U	0	1	0	0
7	Half-Avuncular	HA	0.0625	0.75	0.25	0
8	Half-First-Cousin	HFC	0.03125	0.875	0.125	0
9	Half-Sib+First-Cousin	HSFC	0.1875	0.375	0.5	0.125
10	Parrent-Offspring	PO	0.25	0	1	0
11	MZ-Twins	MZ	0.5	0	0	1

Prest Brief Overview

- Input files:
  - pedgree file (PLINK format)

  FID IID MA PA SEX AFF SNP1_a1 SNP1_a2 SNP2_a1 SNP2_a2
  - map file (PLINK format: chr / snp / cM genetic dist / bp dist)

  CHR SNP cM_dist basepair_dist
- 2 options:
  - Prest –file geno.ped –map geno.map
  - wped (within pedigrees)

  ./prest --file geno.ped --map geno.map --wped
  - aped (all individuals)
- Output files:
  - Prest-results

Summary of output fields:

FID1,IID1 : id of first individual

FID2, IID2 : id of second invididual

reltype : relationship type of pair, among the 11 tested relationships (see below)

commark : number of markers commonly typed in the two individuals

p.IBS0, p.IBS1, p.IBS2 : computed mean IBS sharing of 0/1/2 alleles.

IBS : mean IBS sharing

p.IBD0 , p.IBD1 , p.IBD2 : computed IBD statistics

Data cleaning with PLINK
Purcell et al., 2007

- Great for case-control – **be careful when using relateds**, particularly if not trios – needs both parents
- Ped files looks just like your MERLIN pedfile
- Map file: (chr, markername cM distance) bp
- **plink --file data –mendel**

code

- 1 AA , AA -> AB
- 2 BB , BB -> AB
- 3 BB , ** -> AA
- 4 ** , BB -> AA
- 5 BB , BB -> AA
- 6 AA , ** -> BB
- 7 ** , AA -> BB
- 8 AA , AA -> BB
- 9 ** , AA -> BB (X chromosome male offspring)
- 10 ** , BB -> AA (X chromosome male offspring)

Data cleaning with PLINK
Purcell et al., 2007

- Get pairwise IBD and then see if it conforms to what you think it should look like (remember: PLINK likes unrelateds!)

plink --file mydata –genome

Output

FID1 Family ID for first individual
IID1 Individual ID for first individual
FID2 Family ID for second individual
IID2 Individual ID for second individual
Z0 P(IBD=0)
Z1 P(IBD=1)
Z2 P(IBD=2)
PI_HAT P(IBD=2)+0.5*P(IBD=1) (proportion IBD)
IBS0 Number of IBS 0 nonmissing loci
IBS1 Number of IBS 1 nonmissing loci
IBS2 Number of IBS 2 nonmissing loci
DST IBS distance (IBS2 + 0.5*IBS1) / (N SNP pairs)
P IBS binomial test
HOMHOM Number of IBS 0 SNP pairs used in test
HETHET Number of IBS 2 het/het SNP pairs in test
RATIO R in ratio of 1:2 for IBS 0 : HETHET

Conclusions

- Cleaning data requires care and experience
- Fix pedigree errors prior to removing genotypes
- Analysts should participate in data cleaning
  - (PEDSTATS & PEDWIPE): Gross cleaning
    - Simple pedfiles, easy to make clean peds automatically; too crude
  - PEDCHECK – use the same pedigree to get detailed checks of Mendelian problems
    - Simple pedfiles, simple output, simple zero-out options; duplicates?
  - PLINK – fantastic for unrelateds
    - Very flexible file formats; caution with relateds
  - GRR – find MZ twins, duplicates, half-sibs
    - Simple ped file, graphical; not a lot of stats except IBS
  - PREST/ALTERTEST – identify incorrect relationships
    - Lots of alternative stats; now easy to use

Things to always check for…

- Biological sex: use X chromosome markers. Even better, type some Y chromosome markers.
- Call/dropout rate, marker heterozygosity, HWE.
- Relatedness – allegedly unrelated samples can have related individuals.
- Duplicates, sample swaps.
- Mendelian errors
- One strategy is to start by running pedstats and GRR. Then, depending on scope of problems, run PEDCHECK. If still too complicated, run PREST. Do not “edit” the pedigree until you are convinced you have figured all the problems out.

Things to always check for…

- *Marker* with excess missingness (1%, 5%) vs *individuals* with excess missingness.
- Deviations from HWE – always use unrelateds but remember that in certain situations, HWE deviations are expected (true association, ascertainment). In GWAS, due to multiple testing HWE p considered signif is 10^{-4} or lower.
- Allele frequency – what is too low (<1%?)?
- With related individuals – chicken-n-egg problem. Genotyping errors can look like Mendelian errors and vice versa? Which to do first?