Cleaning genotype data

Karl W Broman

Why talk about cleaning data?

- Gene hunters need high quality data
- Genetic studies are expensive, so we should ensure that the data is the best possible
- Cleaning data is tedious and requires care and experience
- Data cleaning is not always performed appropriately

Sources of errors

- Pedigree errors
  - non-paternity
  - unreported adoption/twinning
  - errors in data entry
  - sample mix-ups
- Genotyping errors
  - errors in data entry
  - misinterpretation of pattern on gel
  - mutations may masquerade as errors

The cleaning process

- Verify the subjects' sexes
- Verify relationships
  - Families with many Mendelian inconsistencies
  - Infer pairwise relationships using whole genome data [RelPair]
- Mendelian inheritance
  - Find problem genotypes
  - Identify responsible individuals [PedCheck]
- Beyond Mendel
  - Unlikely multiple recomb. events

Example data

- 2141 individuals in 400 families (mostly sibships; a handful of smallish pedigrees)
- Weber screening set 9
  - 366 autosomal markers
  - 17 on X, 4 on Y
- 803,739 total genotypes (~ 3% dropped)

RelPair

For each pair of individuals, calculate

Pr(genetic data | R)

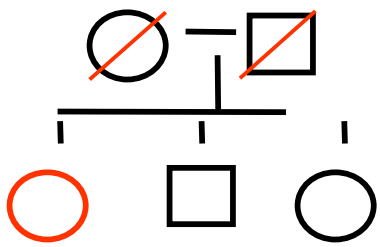
for R = MZ twins
parent/offspring
full sibs
half sibs
unrelated

Boehnke and Cox (1997) AJHG 61:423–429

Broman and Weber (1998) AJHG 63:1563–1564

Family 239

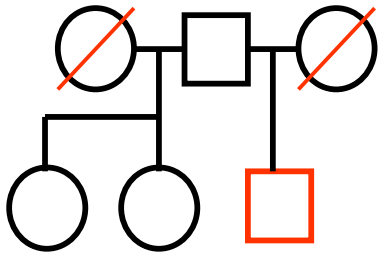
	$\log_{10}$ likelihood			IBS		
rel. pair	**full sibs**	**half sibs**	**unrelated**	**0**	**1**	**2**
1342,1343	–11	0	–27	73	230	57
1342,1344	–12	0	–34	61	252	45
1343,1344	0	–32	–110	39	174	140



1342 1343
1344

Family 403

	$\log_{10}$ likelihood			IBS		
rel. pair	**full sibs**	**half sibs**	**unrelated**	**0**	**1**	**2**
2190,2191	0	–26	–98	47	190	125
2190,2192	0	–28	–111	27	204	127
2191,2192	0	–48	–141	24	189	148



2190 2191

2192

Pedigree errors

Non-paternity (10: no data on dad; 1: maybe dad's brother)	17
MZ twins (including 1 set MZ triplets, 1 father/son)	12
Sample switch (mother/daughter)	1
Full sibs reported to be half sibs	1
Completely unrelated	2
Total	**33**

Apparent error rate (via unreported twins)

		IBS		
Fam	Individuals	0	1	2
25	126,131	0	1	348
81	**396,397**	**2**	**36**	**298**
107	537,539	0	1	346
127	628,629	0	0	357
149	742,744,745	0	3	348
173	**877,878**	**1**	**27**	**323**
212	1127,1129	0	0	361
231	1258,1259	0	2	349
251	1411,1412	0	0	359
295	1628,1630	0	0	358
330	1820,1826	0	0	351
393	2139,2140	0	4	343

Overall error rate = 0.91 %

Bad samples = 5.1 %

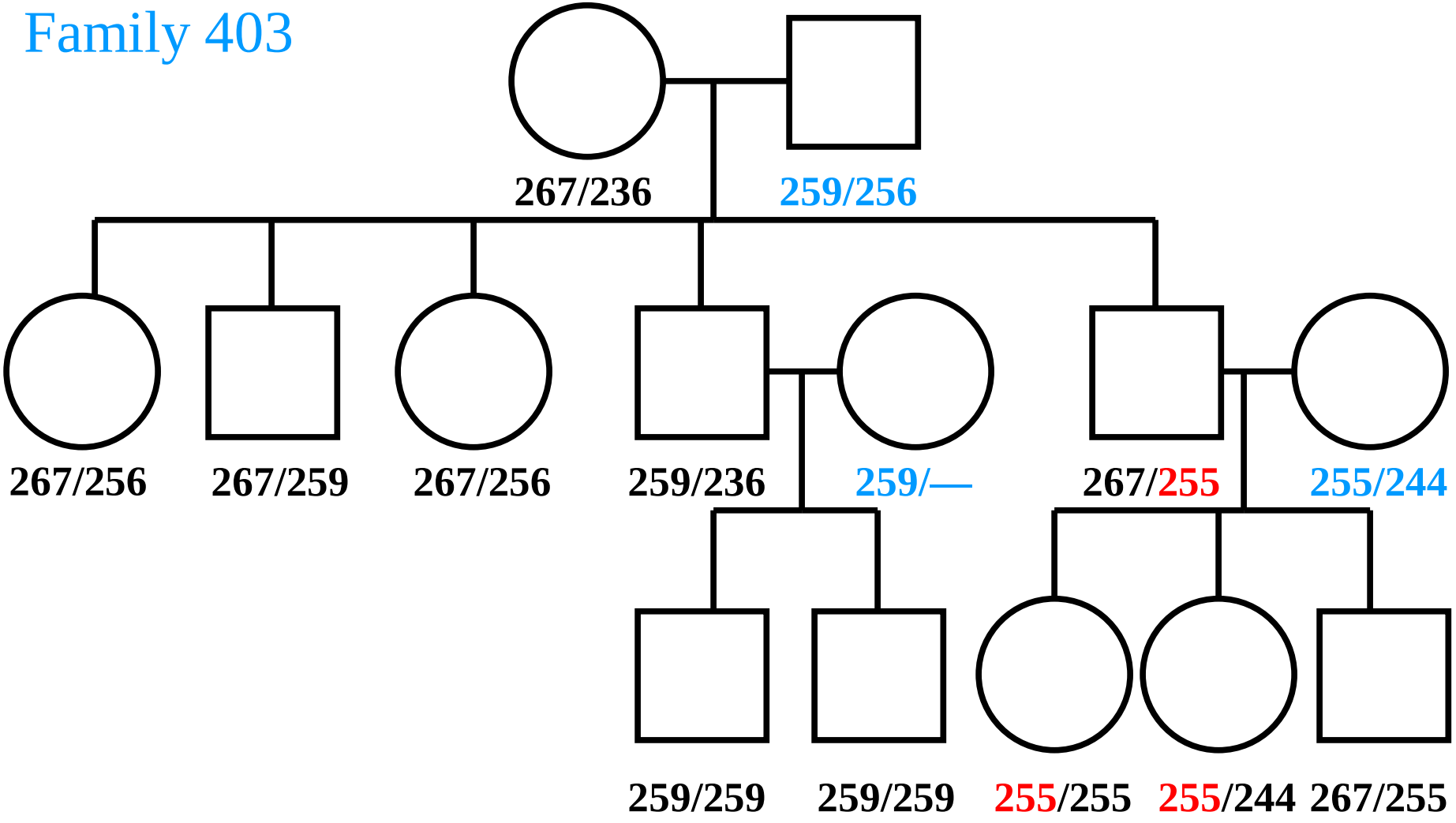
Good samples = 0.15 %

PedCheck

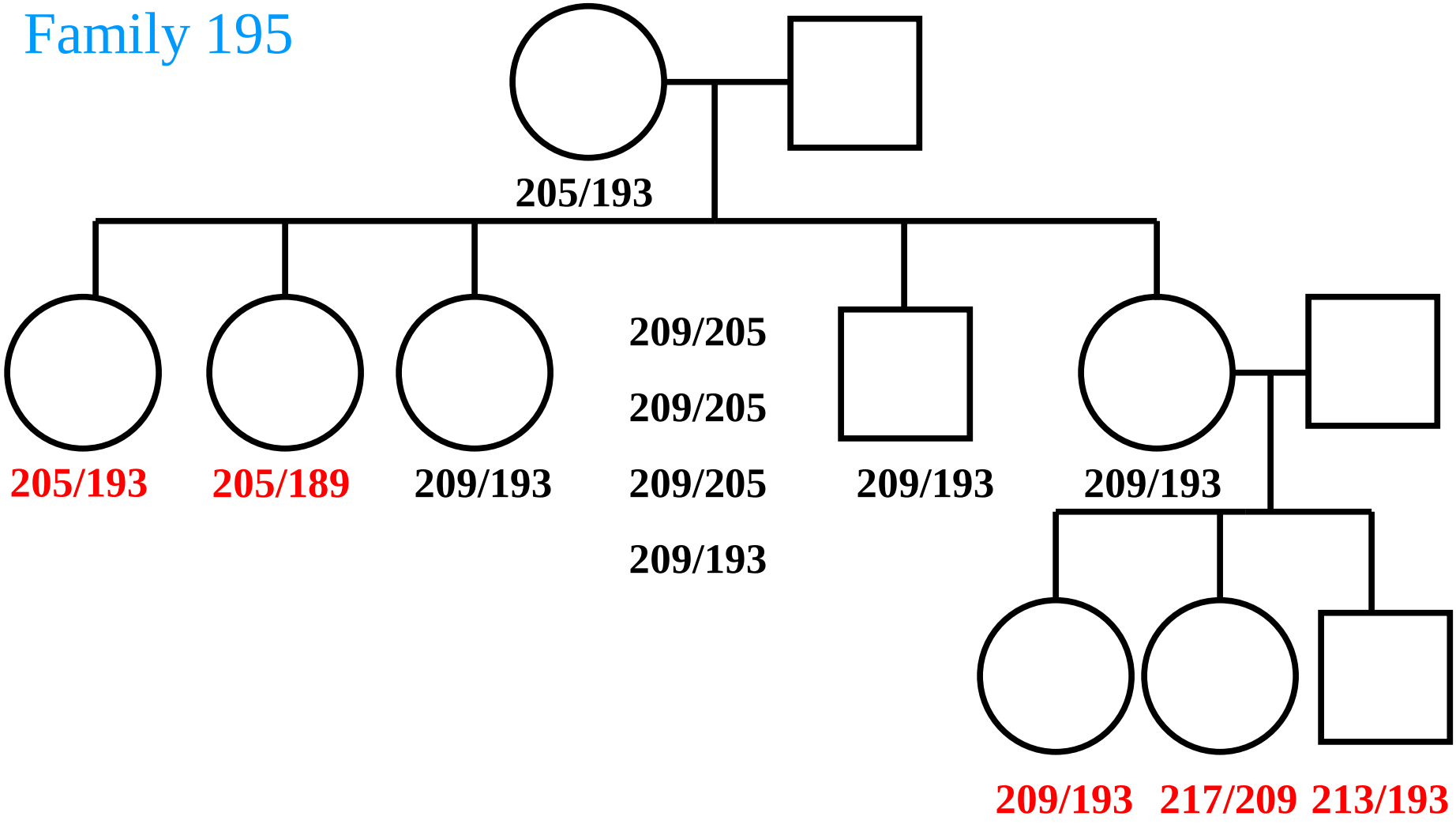
1. Check on nuclear family level
   – easy inconsistencies
2. Genotype elimination
   – find all inconsistencies
3. Determine critical genotypes
   – removing one individual's genotype eliminates inconsistency
4. For critical genotypes, calculate odds ratio to determine the most likely error.

O'Connell and Weeks (1998) AJHG 63:259–266

D1S2141
Family 403



GATA49D12
Family 195



The result of the Mendel checks

- Removed 2,216/801,848 genotypes
  (~ 0.3%)

- Mismatches in unreported twins:
  51/3,869 (~ 1.3%) prior to Mendel checks
  50/3,868 (~ 1.3%) after Mendel checks

Beyond Mendel

Individual 195-1032
chromosome 9
CRI-MAP *chrompic* output

ii--ii-ii-io-oo-
ioiiiiiooooooo-o-

15 cM

Conclusions

- Cleaning data requires care and experience
- Fix pedigree errors prior to removing genotypes
- The process could be more automated, but now requires a human brain
- Biologists should learn to program (especially perl)
- Analysts should participate in data cleaning
- Data on diallelic markers will require new techniques