Haplin

Norwegian Mother and Child Cohort Study

Population (Triads = children + parents):

- Recruited more than 90.000 pregnant women, years 1999-2008
- More than 70.000 fathers participated

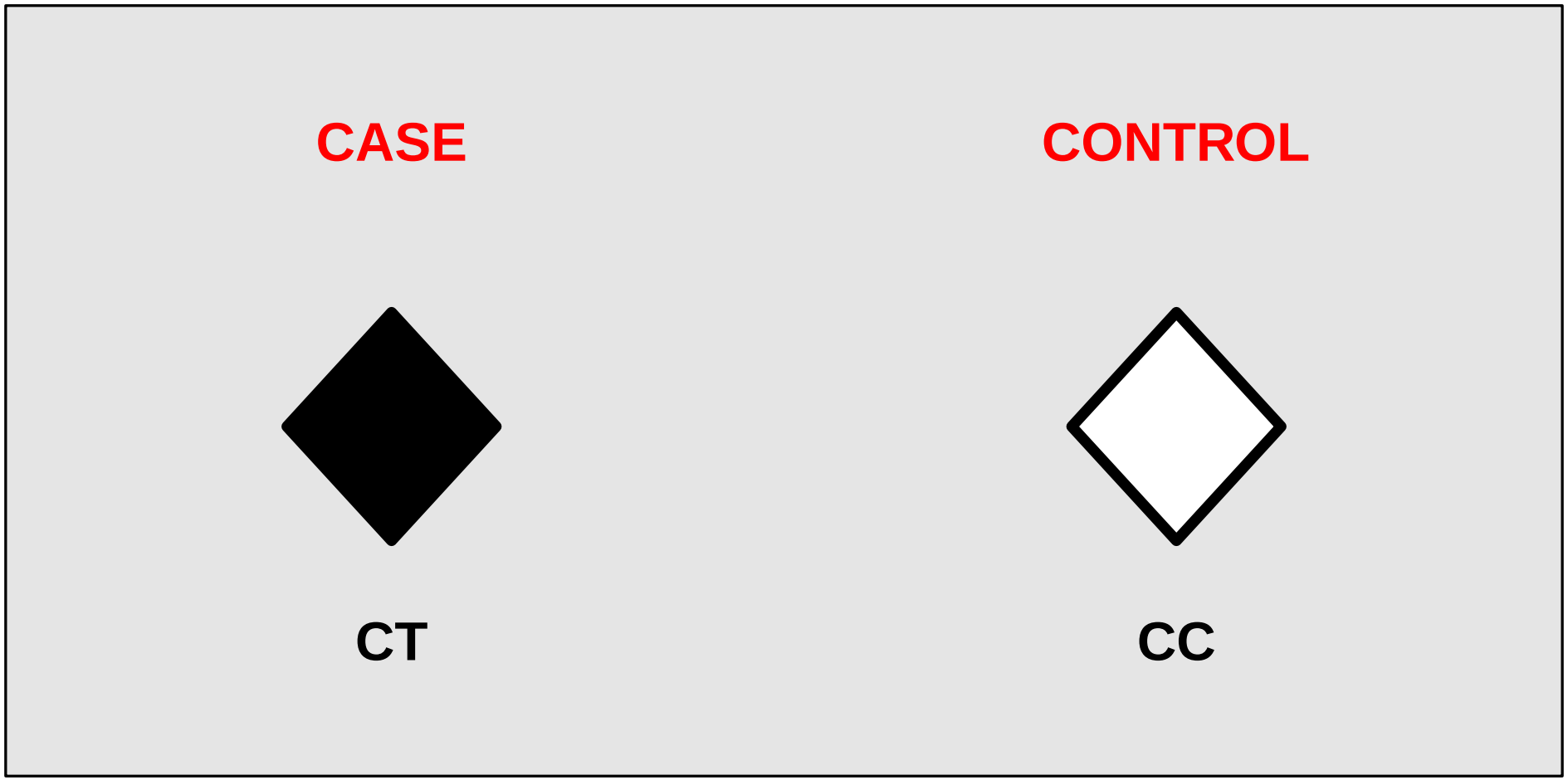
Purpose (among many):

- Identify factors that determine fetal health outcomes:
  - Cleft lip/palate
  - Preterm birth
  - Stillbirths
  - etc., etc. .....
- And other outcomes early in life, time-to-event:
  - Neurodevelopmental disorders
  - Gestational age as time-to-event (inductions as censoring)
  - etc., etc. .....

SNP genotyping:

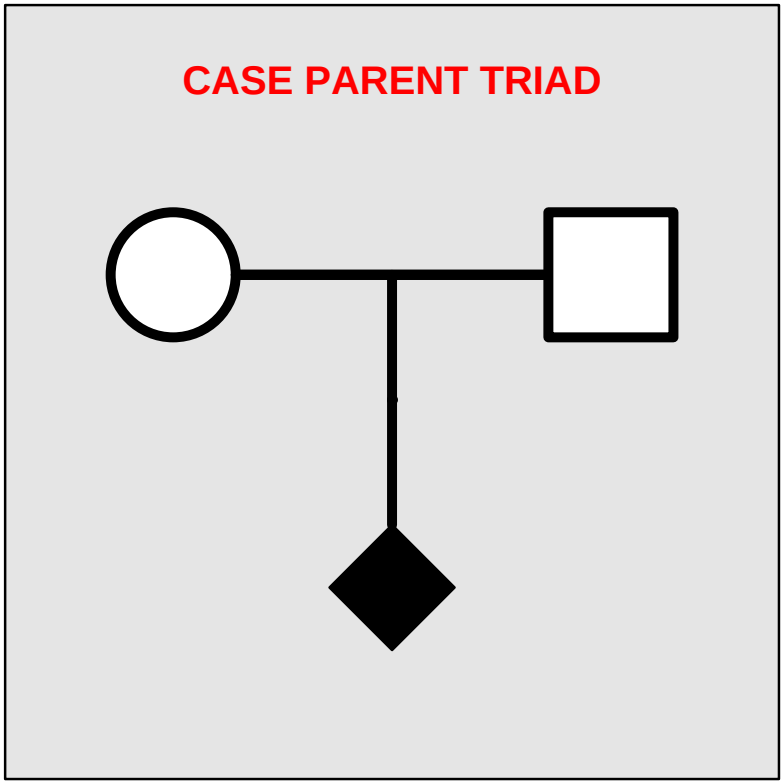
- Genome-wide association study (GWAS) genotyping
- Methylation arrays
- etc.

DESIGN: CASE-CONTROL



Genotype case children and control children

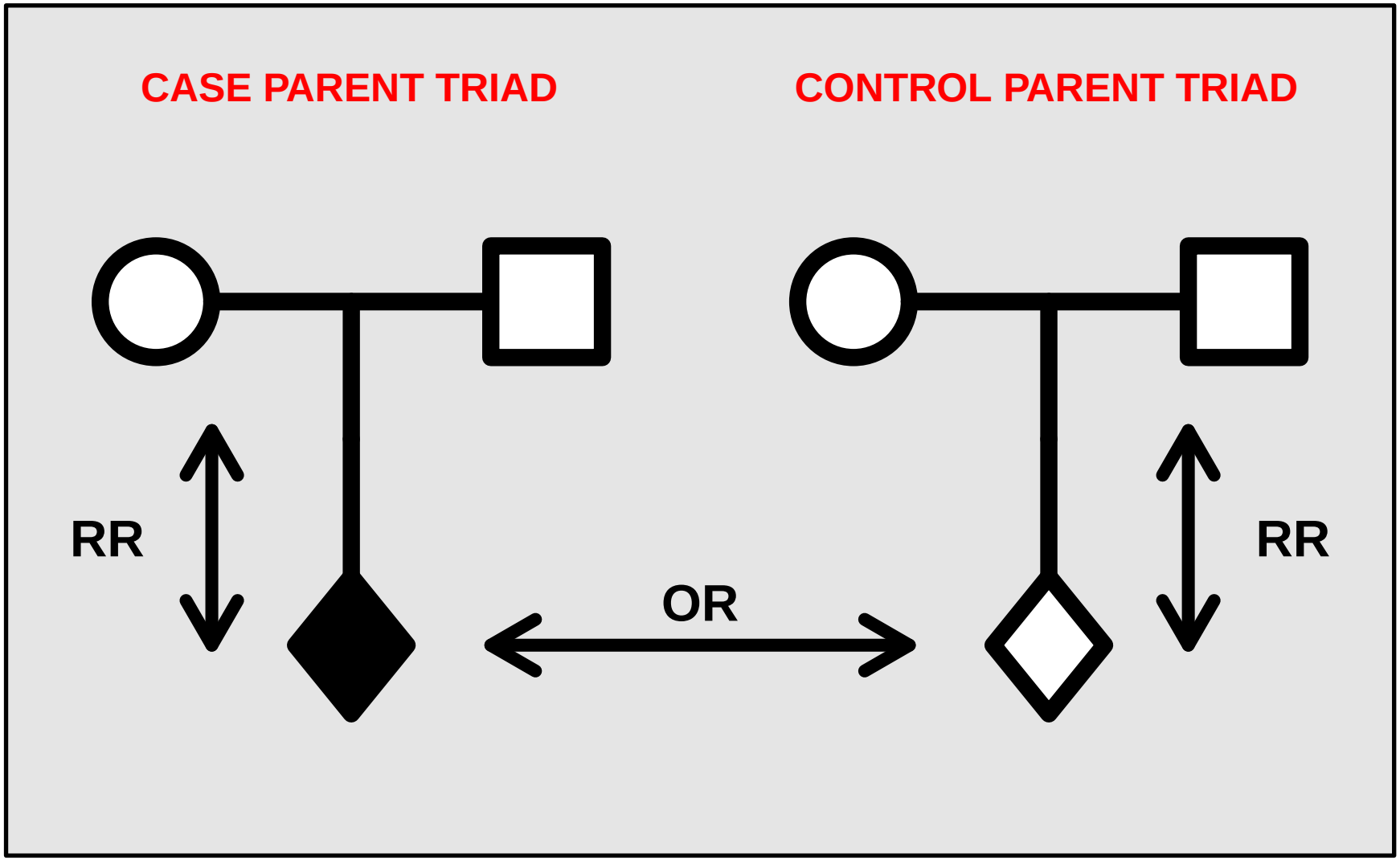
CASE-PARENT TRIAD DESIGN



- **Objective:**
  Detect association between candidate gene and disease

- Sample case child with parents

HYBRID DESIGN



HAPLIN SOFTWARE

https://folk.uib.no/gjessing/genetics/software/haplin/

Estimates association between *haplotypes* and *disease*

- Full likelihood model
- Assess all possible combinations of haplotypes
- “Reconstructs” haplotypes (using the EM-algorithm)
- “Fills in” missing data (using the EM-algorithm)
- Computes **Relative Risk** for all **fetal and maternal** haplotypes
- Computes p-values
- Combined estimation in the **hybrid design**
- Parent-of-origin, X-chromosome, Gene-environment interactions
- Parallel processing

HAPLIN SOFTWARE

Implementation

- CRAN R library, `www.r-project.org`
- Parallel processing

Data structure: small projects, candidate genes

- ASCII file in Haplin's own format (not much used now)
- Or convert ped format $\longrightarrow$ Haplin format

Data structure in `ff`: larger projects, GWAS

- Designed by Julia Romanowska (with Håkon Gjessing)
- Previous version rode piggyback on GenABEL
- Start with ped file from, for instance, `PLINK`
- Read into R with `genDataRead`, stores in `ff`
- Prepare using `genDataPreprocess`, still in `ff`
- For each Haplin run, extract from `ff`

Parallel processing

- Uses R's own `parallel` package
- More or less trivial to make parallel
- Speed is a big issue
- Roughly 2-2.5 sec per SNP (single CPU)
- 600,000 SNPs about a week(?) (single CPU)
- University of North Carolina "KillDevil" cluster about 1 hour, depending on CPU availability

EXAMPLE: A SINGLE DIALLELIC MARKER FOR CLEFT LIP/PALATE

SNP marker (MSX1-1.3) for the msx1 homeobox gene on chromosome 4

M	F	C	Frequency
CC	CC	CC	0
TC	CC	CC	0
CT	CC	TC	0
TT	CC	TC	0
CC	TC	CC	0
TC	TC	CC	0
CT	TC	TC	1
TT	TC	TC	11
CC	CT	CT	0
TC	CT	CT	0
CT	CT	TT	0
TT	CT	TT	7
CC	TT	CT	0
TC	TT	CT	7
CT	TT	TT	8
TT	TT	TT	227
			261

LIKELIHOOD SETUP: CASE TRIAD

- Poisson frequencies of triad genotype (M, F, C)
- Conditional on disease D in child

$$P((M, F, C)|D) = P(D|(M, F, C)) \times P((M, F, C)) \times \frac{1}{P(D)}$$

- Normalizing constant $\xi = \frac{1}{P(D)}$ (cannot be determined from case data)

- Population gene distribution $P((M, F, C))$
  Hardy-Weinberg, random mating, Mendelian transmission

$$P((A_i A_j, A_k A_l)) = p_i p_j p_k p_l$$

p_i are allele frequencies in background population

G. H. Hardy, Science, 1908

To the Editor of Science:

I am reluctant to intrude in a discussion concerning matters of which I have no expert knowledge, and I should have expected the very simple point which I wish to make to have been familiar to biologists. However, some remarks of Mr. Udny Yule, to which Mr. R. C. Punnett has called my attention, suggest that it may still be worth making...

Remark on Hardy-Weinberg equilibrium

- The trio design is quite well protected against population stratification
- It compares transmitted with non-transmitted alleles within the same person
- TDT-tests use this advantage fully
- Haplin assumes HWE, but it's not essential
- QC often removes the most extreme deviations from HWE
- Top hits can always be checked for HWE after-the-fact
- I.e. not likely to be an issue in Haplin analyses

LIKELIHOOD SETUP: CASE TRIAD

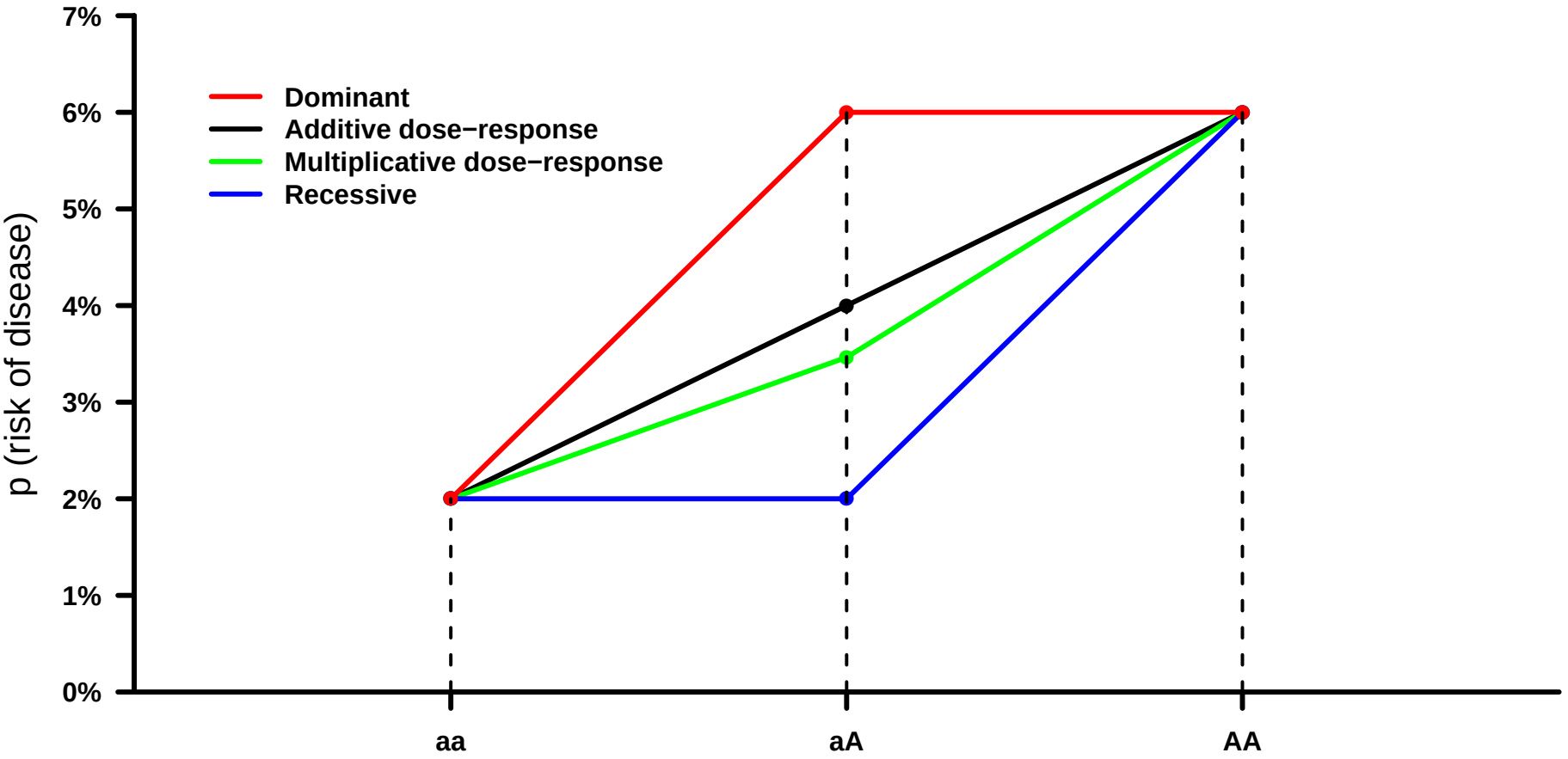
$$P((M, F, C)|D) = P(D|(M, F, C)) \times P((M, F, C)) \times \frac{1}{P(D)}$$

- Penetrance (response model)

$$P(D|(A_iA_j, A_kA_l)) = B \cdot R_jR_l \cdot R^{\star}_{jl}$$

- B is a reference level
- $R^{\star}_{jl} = 1$ when $j \neq l$
- $R^{\star}_{jj}$ measures deviation from multiplicative model for a double gene dose.

GENETIC INTERACTION MODELS: ALLELE INTERACTIONS



LIKELIHOOD SETUP: CASE TRIAD

Expected cell frequencies:

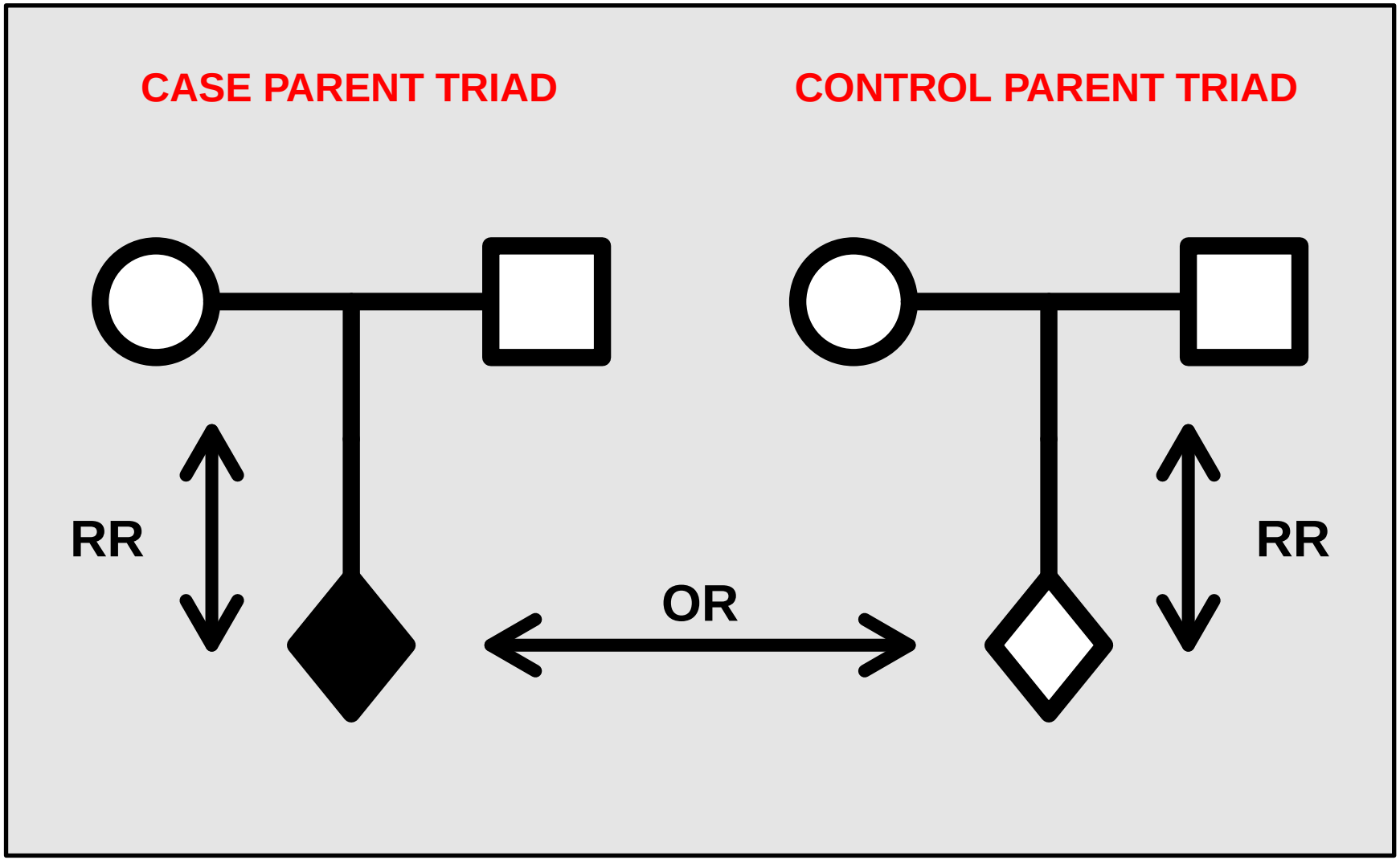
$$\begin{aligned}
\xi_{ijkl} &= n \cdot P((M,F,C)|D) \\
&= \xi' \cdot p_i p_j p_k p_l \cdot R_j R_l \cdot R^{\star}_{jl} \\
&= \xi' \cdot p_i \cdot p_j R_j \cdot p_k \cdot p_l R_l \cdot R^{\star}_{jl}
\end{aligned}$$

- Log-linear model
- Near-exact solution possible

Use EM algorithm to

- Impute randomly missing values
- Impute family members not genotyped
- Impute haplotypes

HYBRID DESIGN



LIKELIHOOD SETUP: CASE-CONTROL-TRIO (HYBRID)

Case trios:

$$P((M,F,C)|D) = P(D|(M,F,C)) \times P((M,F,C)) \times \frac{1}{P(D)}$$

Control trios:

$$\begin{aligned} P((M,F,C)|\overline{D}) =& P(\overline{D}|(M,F,C)) \times P((M,F,C)) \times \frac{1}{P(\overline{D})} \\ \approx& P((M,F,C)) \times \frac{1}{P(\overline{D})} \end{aligned}$$

Rare disease assumption: $P(\overline{D}|(M,F,C)) \approx 1$

- Still a log-linear model
- Control trios contribute (only) to allele/haplotype frequency estimation through $P((M,F,C))$