Estimating penetrance curves according to mutation in familial genetic studies in the presence of incomplete genotypes

Grégory NUEL

Stochastics and Biology Group, LPSM (CNRS 8001),
Sorbonne University, Paris, France

Séminaire MathForGenomics
May 27, 2021
IBGBI, Evry, France

Outline

1. Introduction
   - Some Recalls
   - A Fictional Genetic Study
   - Estimations from Known Genotypes
2. Expectation-Maximization
   - Principle
   - Posterior in Pedigree
   - Estimations from Unknown Genotypes
3. Advanced Stuff
   - Ascertainment Issues
   - Advanced Models
   - Sophisticated Posteriors

Outline

1. Introduction
   - Some Recalls
   - A Fictional Genetic Study
   - Estimations from Known Genotypes
2. Expectation-Maximization
   - Principle
   - Posterior in Pedigree
   - Estimations from Unknown Genotypes
3. Advanced Stuff
   - Ascertainment Issues
   - Advanced Models
   - Sophisticated Posteriors

Binary Disease

Definition:

- $Y \in \{0, 1\}$ a binary disease phenotype,
- $X \in \{\mathtt{DD} = 0, \mathtt{Dd} = 1, \mathtt{dd} = 2\}$ a bi-allelic genotype
- for all $x \in \{0, 1, 2\}$, the *penetrance* $F_x = \mathbb{P}(Y = 1|X = x)$

Mode of Inheritance:

- dominant: $F_0 < F_1 = F_2$
- recessive: $F_0 = F_1 < F_2$
- additive: $F_1 = F_0 + R$ and $F_2 = F_0 + 2R$
- multiplicative: $F_1 = F_0 \times R$ and $F_2 = F_0 \times R^2$

Time-to-event Disease

- T time before disease onset, the *hazard rate* is defined by

$$\lambda_x(t) = \lim_{\Delta \to 0} \frac{1}{\Delta} \mathbb{P}(T \in]t, t+\Delta] | T > t, X = x)$$

- phenotype is $Y = \text{UN}t = \{T > t\}$ or $Y = \text{AF}t = \{T = t\}$
- for all $x \in \{0, 1, 2\}$, the *penetrance* is now:

$$F_x(t) = \mathbb{P}(T \leqslant t | X = x) = 1 - \underbrace{\exp\left(-\int_0^t \lambda_x(s) ds\right)}_{S_x(t)}$$

- the *relative hazards* are

$$\text{RH}_1(t) = \frac{\lambda_1(t)}{\lambda_0(t)} \quad \text{and} \quad \text{RH}_2(t) = \frac{\lambda_2(t)}{\lambda_0(t)}$$

Outline

1. Introduction
   - Some Recalls
   - A Fictional Genetic Study
   - Estimations from Known Genotypes
2. Expectation-Maximization
   - Principle
   - Posterior in Pedigree
   - Estimations from Unknown Genotypes
3. Advanced Stuff
   - Ascertainment Issues
   - Advanced Models
   - Sophisticated Posteriors

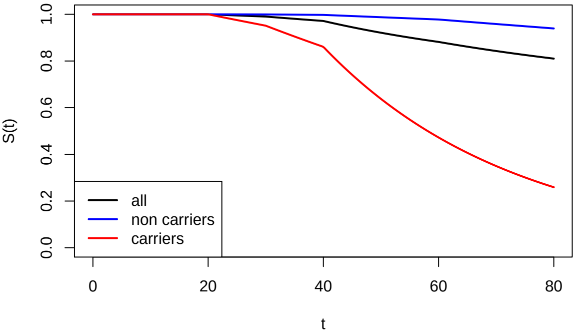
Autosomal Dominant Model

MAF $f = 0.10$ $\pi_0 = (1-f)^2$ $\pi_1 = 1-\pi_0$ $\lambda_1(t) = \lambda_2(t) = \lambda_0(t)\mathrm{RH}(t)$

$$S(t) = \pi_0 \exp\left(\int_0^t \lambda_0(u)du\right) + \pi_1 \exp\left(\int_0^t \lambda_0(u)\mathrm{RH}(u)du\right)$$

known parameter

unknown parameter

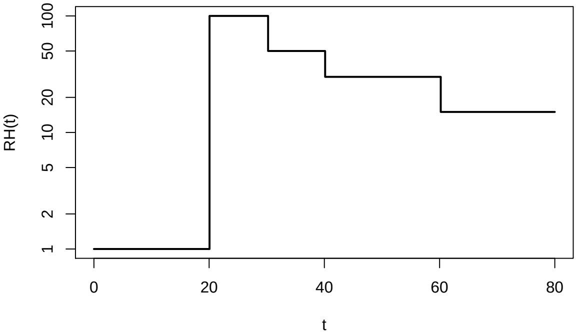


Autosomal Dominant Model

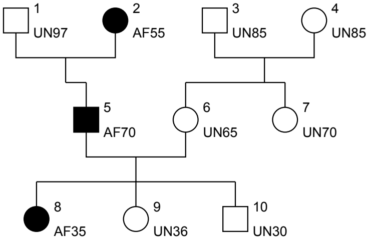
MAF $f = 0.10$ $\pi_0 = (1-f)^2$ $\pi_1 = 1-\pi_0$ $\lambda_1(t) = \lambda_2(t) = \lambda_0(t)\mathrm{RH}(t)$

$$S(t) = \pi_0 \exp\left(\int_0^t \lambda_0(u)du\right) + \pi_1 \exp\left(\int_0^t \lambda_0(u)\mathrm{RH}(u)du\right)$$

known parameter unknown parameter



Simulated Dataset



Design

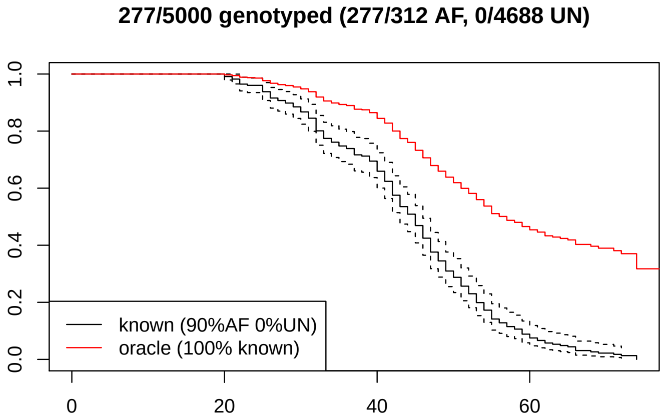
- same structure for all families
- Hardy-Weinberg for founders
- uniform censoring $\mathcal{U}([0, 80])$
- $N = 500$ families
- $n = 5000$ individuals

	unaffected	affected	total
non carrier	3985	56	4041
carrier	703	256	959
total	4688	312	5000

Outline

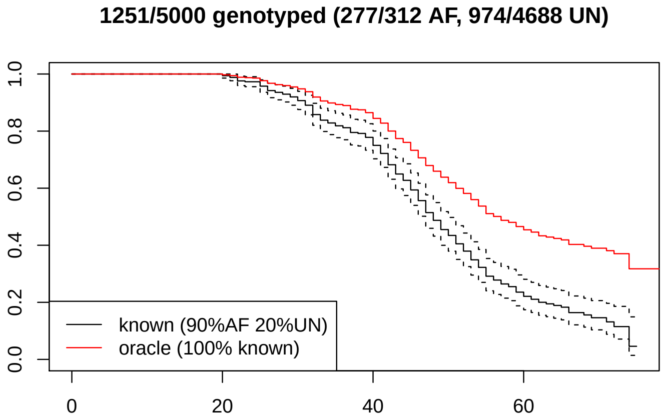
1. Introduction
   - Some Recalls
   - A Fictional Genetic Study
   - Estimations from Known Genotypes
2. Expectation-Maximization
   - Principle
   - Posterior in Pedigree
   - Estimations from Unknown Genotypes
3. Advanced Stuff
   - Ascertainment Issues
   - Advanced Models
   - Sophisticated Posteriors

Unbalanced Genotyping Scheme



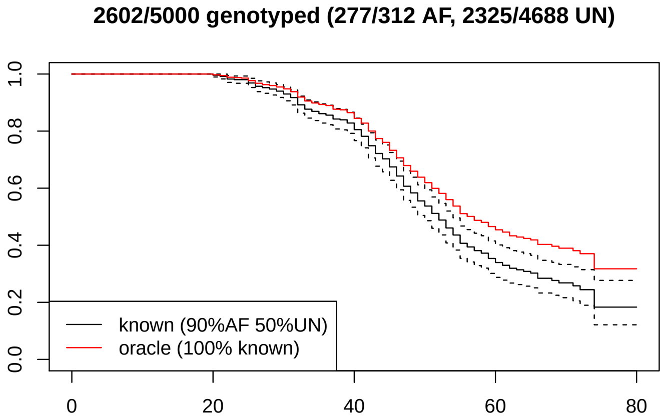
more affected genotyped than unaffected ⇒ risk of bias

Unbalanced Genotyping Scheme



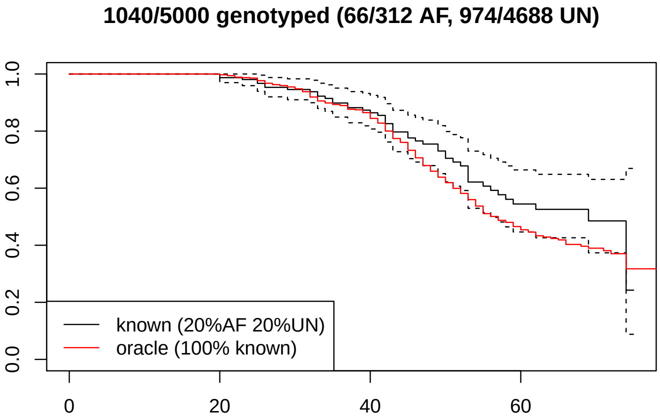
more affected genotyped than unaffected ⇒ risk of bias

Unbalanced Genotyping Scheme



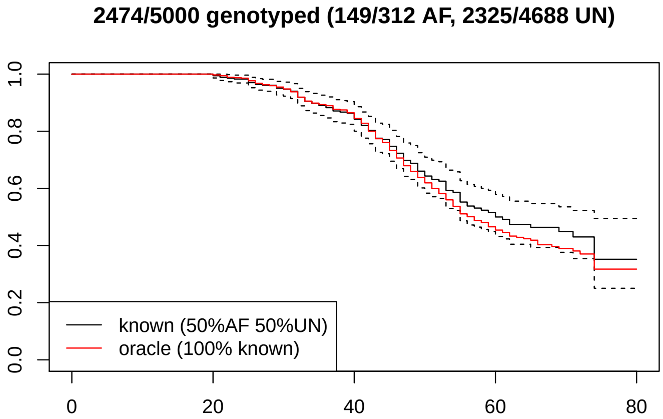
more affected genotyped than unaffected ⇒ risk of bias

Balanced Genotyping Scheme



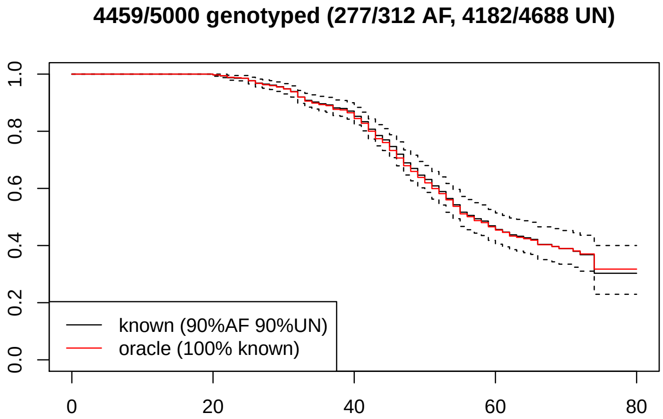
balanced genotyping scheme ⇒ no bias but unrealistic

Balanced Genotyping Scheme



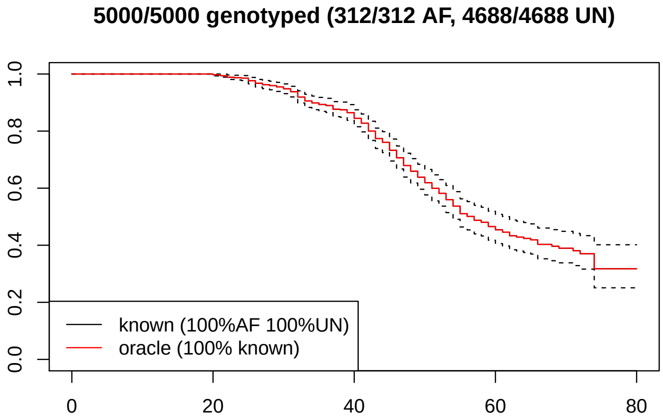
balanced genotyping scheme ⇒ no bias but unrealistic

Balanced Genotyping Scheme



balanced genotyping scheme ⇒ no bias but unrealistic

Balanced Genotyping Scheme



balanced genotyping scheme ⇒ no bias but unrealistic

Outline

1. Introduction
    - Some Recalls
    - A Fictional Genetic Study
    - Estimations from Known Genotypes
2. Expectation-Maximization
    - Principle
    - Posterior in Pedigree
    - Estimations from Unknown Genotypes
3. Advanced Stuff
    - Ascertainment Issues
    - Advanced Models
    - Sophisticated Posteriors

EM Algorithm (Dempster *et al.*, 1977)

Context: X latent variable (*e.g.* unobserved genotypes), Y observed variables (*e.g.* censored time at onset, genetic tests, etc.), θ parameter to estimate (*e.g.* penetrances, hazard rates)

$$\hat{\theta} = \arg\max_{\theta} \log \sum_{X} \mathbb{P}(X, Y|\theta)$$

EM solution: multiple imputation $X^1, \ldots, X^N \sim \mathbb{P}(X|Y; \theta_{\text{old}})$

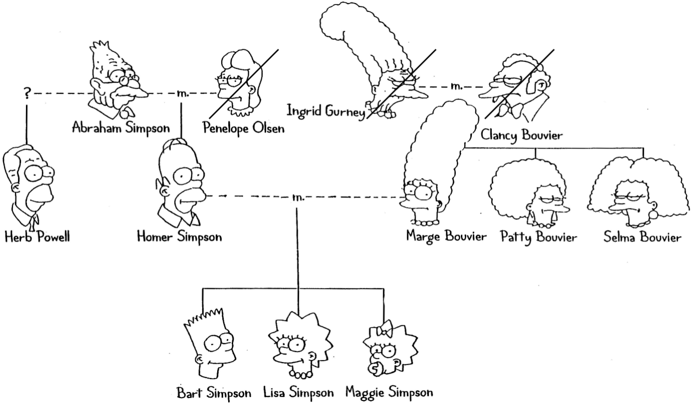
$$\frac{1}{N}\sum_{j=1}^{N} \log \mathbb{P}(X^j, Y|\theta) \xrightarrow[N\to\infty]{} Q(\theta|\theta_{\text{old}}) = \sum_{X} \mathbb{P}(X|Y; \theta_{\text{old}}) \log \mathbb{P}(X, Y|\theta)$$

$$\theta^{(\text{iter}+1)} = \arg\max_{\theta} Q\left(\theta|\theta^{(\text{iter})}\right) \quad \text{and} \quad \theta^{(\text{iter})} \xrightarrow[\text{iter}\to\infty]{} \hat{\theta}$$

Outline

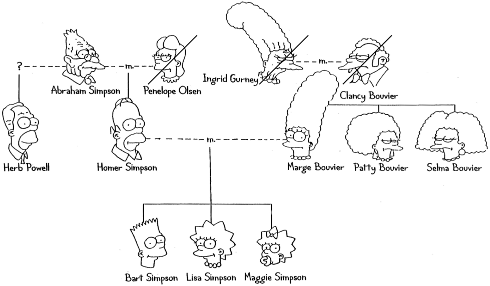
1. Introduction
   - Some Recalls
   - A Fictional Genetic Study
   - Estimations from Known Genotypes
2. Expectation-Maximization
   - Principle
   - Posterior in Pedigree
   - Estimations from Unknown Genotypes
3. Advanced Stuff
   - Ascertainment Issues
   - Advanced Models
   - Sophisticated Posteriors

Simpsons' Pedigree and Bayesian Network



1: Herb's mother, **2**: Abraham, **3**: Penelope, **4**: Ingrid, **5**: Clancy, **6**: Herb, **7**: Homer, **8**: Marge, **9**: Patty, **10**: Selma, **11**: Bart, **12**: Lisa, **13**: Maggie

Simpsons’ Pedigree and Bayesian Network



1: Herb’s mother, **2**: Abraham, **3**: Penelope, **4**: Ingrid, **5**: Clancy, **6**: Herb, **7**: Homer, **8**: Marge, **9**: Patty, **10**: Selma,

11: Bart, **12**: Lisa, **13**: Maggie

$$\mathbb{P}(X) = \mathbb{P}(X_1)\mathbb{P}(X_2)\mathbb{P}(X_3)\mathbb{P}(X_4)\mathbb{P}(X_5)$$
$$\mathbb{P}(X_6|\, X_{1,2})\mathbb{P}(X_7|\, X_{2,3})\mathbb{P}(X_8|\, X_{4,5})\mathbb{P}(X_9|\, X_{4,5})\mathbb{P}(X_{10}|\, X_{4,5})$$
$$\mathbb{P}(X_{11}|\, X_{7,8})\mathbb{P}(X_{12}|\, X_{7,8})\mathbb{P}(X_{13}|\, X_{7,8})$$

Blood Type Genetics

- ABO gene $\Rightarrow p_O = 0.60, p_A = 0.30, p_B = 0.10$
- RHD gene $\Rightarrow q_D = 0.60, q_d = 0.39, q_w = 0.01$
- This leads to a total of 12 blood phenotypes:
  A+, B+, AB+, O+, A-, B-, AB-, O-,Aw, Bw, ABw, Ow

	ABO	OO	OA	OB	AA	AB	BB
RHD		0.36	0.36	0.12	0.09	0.06	0.01
DD	0.3600	O+	A+	B+	A+	AB+	B+
Dd	0.4680	O+	A+	B+	A+	AB+	B+
Dw	0.0120	O+	A+	B+	A+	AB+	B+
dd	0.1521	O-	A-	B-	A-	AB-	B-
dw	0.0078	Ow	Aw	Bw	Aw	ABw	Bw
ww	0.0001	Ow	Aw	Bw	Aw	ABw	Bw

Simpsons' Pedigree and Bayesian Network

$$\mathbb{P}(X) = \mathbb{P}(X_1)\mathbb{P}(X_2)\mathbb{P}(X_3)\mathbb{P}(X_4)\mathbb{P}(X_5)$$
$$\mathbb{P}(X_6|\ X_{1,2})\mathbb{P}(X_7|\ X_{2,3})\mathbb{P}(X_8|\ X_{4,6})\mathbb{P}(X_9|\ X_{4,6})\mathbb{P}(X_{10}|\ X_{4,6})$$
$$\mathbb{P}(X_{11}|\ X_{7,8})\mathbb{P}(X_{12}|\ X_{7,8})\mathbb{P}(X_{13}|\ X_{7,8})$$

$$X_i \in \mathcal{G} = \{\mathrm{O},\mathrm{A},\mathrm{B}\}^2 \times \{\mathrm{D},\mathrm{d},\mathrm{w}\}^2 \quad |\mathcal{G}| = 3^2 \times 3^2 = 81$$

$$\mathrm{ev} = \{\text{Homer A+ and Bart Ow}\} \quad \mathbb{P}(X|\mathrm{ev}) = \frac{\mathbb{P}(X,\mathrm{ev})}{\sum_{X'}\mathbb{P}(X',\mathrm{ev})}$$

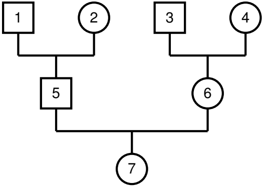
- $X = (X_1, X_2, \dots, X_{13})$ is the family genotype
- in order to compute $\mathbb{P}(\mathrm{ev}) = \sum_{X'}\mathbb{P}(X',\mathrm{ev})$
- we *just* have to sum over 81^{13} configurations

$$\mathbf{81^{13} = 6\,461\,081\,889\,226\,672\,446\,898\,176}$$

$\Rightarrow$ simply impossible !

Local computations in a simple pedigree

Idea: we consider a smaller (but similar) family, ev (evidence) still represents the available information.



- for *founders* $(1, 2, 3, 4)$ i:

$$\varphi_i(X_i) = \mathbb{P}(X_i \cap \mathrm{ev})$$

- for *offsprings* $(5, 6, 7)$
  k with parents i, j:

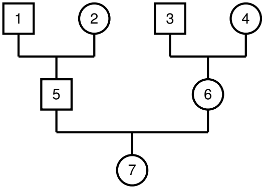
$$\varphi_j(X_i, X_j, X_k) = \mathbb{P}(X_k \cap \mathrm{ev} | X_i, X_j)$$

$$\mathbb{P}(\mathrm{ev}) = \sum_{X_1}\sum_{X_2}\sum_{X_3}\sum_{X_4}\sum_{X_5}\sum_{X_6}\sum_{X_7} \varphi_1(X_1)\varphi_2(X_2)\varphi_3(X_3)\varphi_4(X_4)$$
$$\varphi_5(X_1, X_2, X_5)\varphi_6(X_3, X_4, X_6)\varphi_7(X_5, X_6, X_7)$$

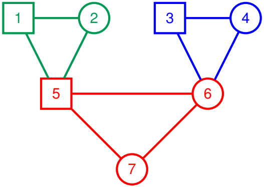
$\Rightarrow 81^7 = 22\,876\,792\,454\,961$ still too large !!

Local computations in a simple pedigree

Pedigree



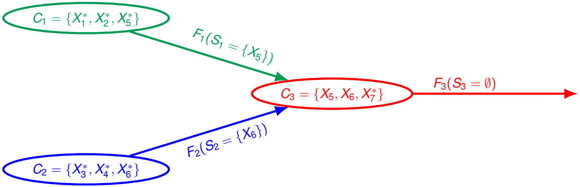
Clique decomposition



$$\mathbb{P}(\mathrm{ev}) = \sum_{X_5}\sum_{X_6}\sum_{X_7}\left\{\left(\overbrace{\sum_{X_1}\sum_{X_2}\varphi_1(X_1)\varphi_2(X_2)\varphi_5(X_1,X_2,X_5)}^{F_1(X_5)}\right)\right.$$

$$\left.\left(\overbrace{\sum_{X_3}\sum_{X_4}\varphi_3(X_3)\varphi_4(X_4)\varphi_6(X_3,X_4,X_6)}^{F_2(X_6)}\right)\varphi_7(X_5,X_6,X_7)\right\}$$

Local computations in a simple pedigree



$C_1 = \{X_1^*, X_2^*, X_5^*\}$

$F_1(S_1 = \{X_5\})$

$C_3 = \{X_5, X_6, X_7^*\}$

$F_3(S_3 = \emptyset)$

$F_2(S_2 = \{X_6\})$

$C_2 = \{X_3^*, X_4^*, X_6^*\}$

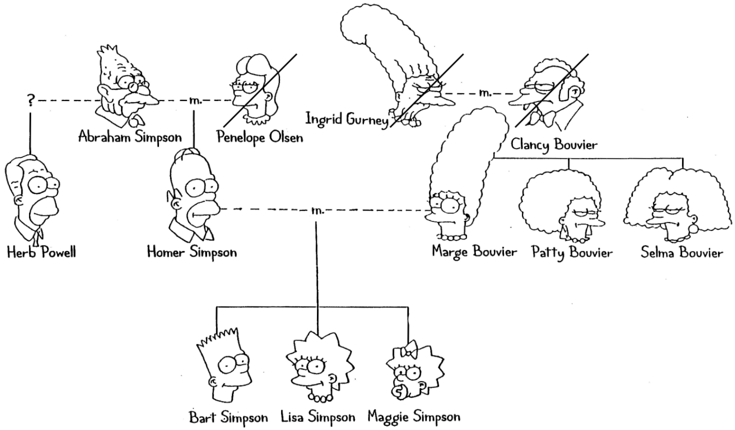
$$F_j(S_j) = \sum_{C_j \setminus S_j} \left(\prod_{i \in \mathrm{from}_j} F_i(S_i) \right) \times \prod_{X_u \in C_j{}^*} \varphi_u(X_{\mathrm{pa}_u}, X_u) \qquad F_3(\emptyset) = \mathbb{P}(\mathrm{ev})$$

Complexity:

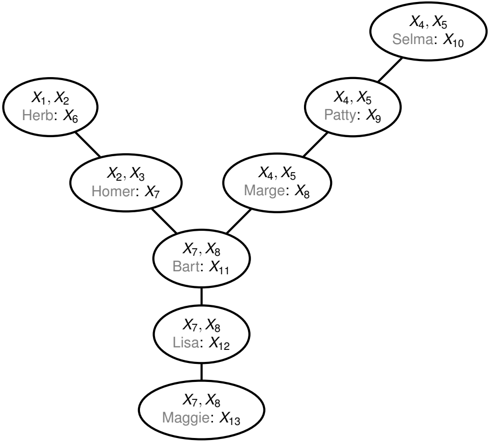
- from $81^7 = 22\,876\,792\,454\,961$
- to $3 \times 81^3 = 1\,594\,323$

Lisa: *"Much better !"* Homer: *"Woohoo !"*

Clique decomposition for the Simpsons

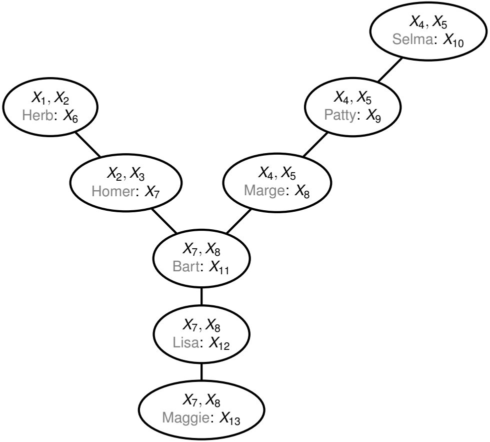


Clique decomposition for the Simpsons


X_4, X_5 Selma: X_{10}
X_1, X_2 Herb: X_6
X_4, X_5 Patty: X_9
X_2, X_3 Homer: X_7
X_4, X_5 Marge: X_8
X_7, X_8 Bart: X_{11}
X_7, X_8 Lisa: X_{12}
X_7, X_8 Maggie: X_{13}

- from $81^{13} = 6\,461\,081\,889\,226\,672\,446\,898\,176$
- to $8 \times 81^3 = 4\,251\,528$

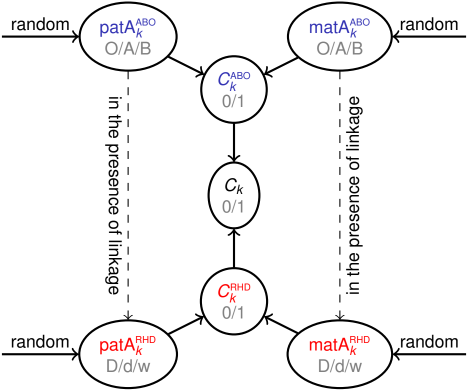
Clique decomposition for the Simpsons


X_4, X_5 Selma: X_{10}
X_1, X_2 Herb: X_6
X_4, X_5 Patty: X_9
X_2, X_3 Homer: X_7
X_4, X_5 Marge: X_8
X_7, X_8 Bart: X_{11}
X_7, X_8 Lisa: X_{12}
X_7, X_8 Maggie: X_{13}

- from $81^{13} = 6\,461\,081\,889\,226\,672\,446\,898\,176$
- to $8 \times 81^3 = 4\,251\,528$

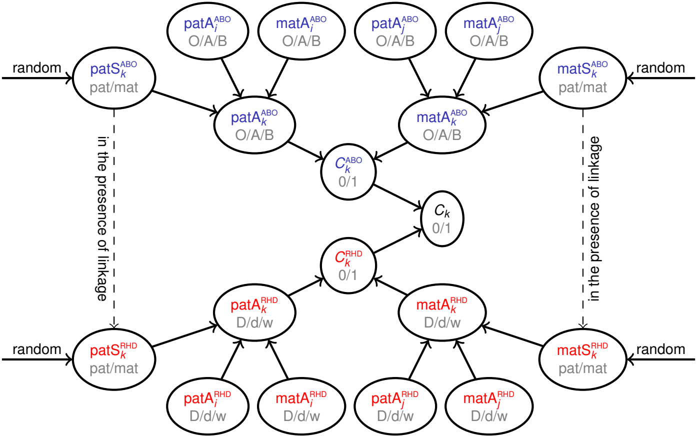
Extended Pedigree: Small Variables

For a *founder i*, instead of $X_i \in \mathcal{G}$ we have:


random
$\text{patA}_k^{\text{ABO}}$
O/A/B
$\text{matA}_k^{\text{ABO}}$
O/A/B
random
C_k^{ABO}
0/1
in the presence of linkage
in the presence of linkage
C_k
0/1
C_k^{RHD}
0/1
random
$\text{patA}_k^{\text{RHD}}$
D/d/w
$\text{matA}_k^{\text{RHD}}$
D/d/w
random

Extended Pedigree: Small Variables

For a *offspring* k (with father i and mother j), instead of $X_k \in \mathcal{G} | X_i, X_j$ we have:


$\text{patA}_i^{\text{ABO}}$ O/A/B
$\text{matA}_i^{\text{ABO}}$ O/A/B
$\text{patA}_j^{\text{ABO}}$ O/A/B
$\text{matA}_j^{\text{ABO}}$ O/A/B
random
$\text{patS}_k^{\text{ABO}}$ pat/mat
$\text{matS}_k^{\text{ABO}}$ pat/mat
random
$\text{patA}_k^{\text{ABO}}$ O/A/B
$\text{matA}_k^{\text{ABO}}$ O/A/B
in the presence of linkage
in the presence of linkage
C_k^{ABO} 0/1
C_k 0/1
C_k^{RHD} 0/1
$\text{patA}_k^{\text{RHD}}$ D/d/w
$\text{matA}_k^{\text{RHD}}$ D/d/w
random
$\text{patS}_k^{\text{RHD}}$ pat/mat
$\text{matS}_k^{\text{RHD}}$ pat/mat
random
$\text{patA}_i^{\text{RHD}}$ D/d/w
$\text{matA}_i^{\text{RHD}}$ D/d/w
$\text{patA}_j^{\text{RHD}}$ D/d/w
$\text{matA}_j^{\text{RHD}}$ D/d/w

Extended Pedigree: Small Variables

Recall on complexity:

- naive $81^{13} = 6\,461\,081\,889\,226\,672\,446\,898\,176$
- genotypes $8 \times 81^{3} = 4\,251\,528$

Small variables with the three heuristics:

- min-neighbors: the smallest clique
  $\Rightarrow$ 61154 61649 89051
- min-fill: the clique with minimum fill-in
  $\Rightarrow$ 85205 92333 92360
- weighted min-fill: the clique with minimum weighted fill-in
  $\Rightarrow$ 57530 43841 43112

The `bped` Program

`bped` is a C++ program for performing the sum-product algorithm and computing all marginal posterior distribution under the autosomal bi-allelic Mendelian model under HWE.

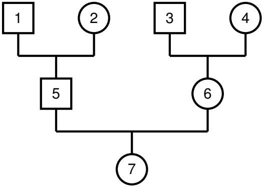
command-line : `bped file.ped file.ev [freq]`

- pedigree file (famID/indID,patID,matID)
- evidence file (famID/indID/`AA`/`Aa`/`aA`/`aa`)
- (option) allelic frequency (default $f = 0.10$)

The **ev. file** contains for each ind. $\propto \mathbb{P}(X_i = \texttt{AA}/\texttt{Aa}/\texttt{aA}/\texttt{aa} | Y_i)$

- $1/1/1/1$ is the neutral evidence (no information)
- $0/1/1/0$ is the evidence for a heterozygous carrier
- $1/0.095/0.095/0.095$ for $T_i = 67, \delta_i = 0$
- $0/0.407/0.407/0.407$ for $T_i = 38, \delta_i = 1$

`bped` Demo 1



- allelic frequency $f = 0.10$
- ev1: full neutral evidence
- ev2: $X_7 \neq \texttt{AA}$
- ev3: $X_7 \neq \texttt{AA}$ and $X_5 = \texttt{AA}$

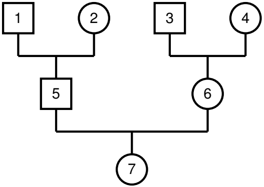
ped file:

```
1 1 0 0
1 2 0 0
1 3 0 0
1 4 0 0
1 5 1 2
1 6 3 4
1 7 5 6
```

ev1 file:

```
1 1 1 1 1 1
1 2 1 1 1 1
1 3 1 1 1 1
1 4 1 1 1 1
1 5 1 1 1 1
1 6 1 1 1 1
1 7 1 1 1 1
```

bped Demo 1

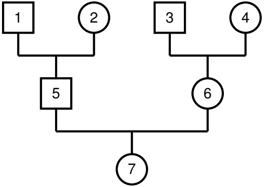


- allelic frequency $f = 0.10$
- ev1: full neutral evidence
- ev2: $X_7 \neq \texttt{AA}$
- ev3: $X_7 \neq \texttt{AA}$ and $X_5 = \texttt{AA}$

bped output for ev1:

```
1:1  0.9801  0.0099  0.0099  0.0001
1:2  0.9801  0.0099  0.0099  0.0001
1:3  0.9801  0.0099  0.0099  0.0001
1:4  0.9801  0.0099  0.0099  0.0001
1:5  0.9801  0.0099  0.0099  0.0001
1:6  0.9801  0.0099  0.0099  0.0001
1:7  0.9801  0.0099  0.0099  0.0001
```

bped Demo 1



- allelic frequency $f = 0.10$
- ev1: full neutral evidence
- ev2: $X_7 \neq \texttt{AA}$
- ev3: $X_7 \neq \texttt{AA}$ and $X_5 = \texttt{AA}$

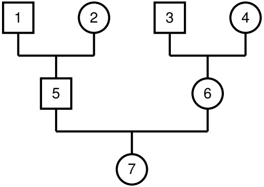
ped file:

1	1	0	0
1	2	0	0
1	3	0	0
1	4	0	0
1	5	1	2
1	6	3	4
1	7	5	6

ev2 file:

1	1	1	1	1	1
1	2	1	1	1	1
1	3	1	1	1	1
1	4	1	1	1	1
1	5	1	1	1	1
1	6	1	1	1	1
1	7	0	1	1	1

`bped` Demo 1

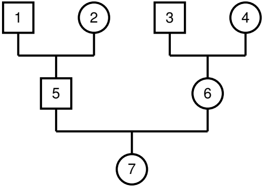


- allelic frequency $f = 0.10$
- ev1: full neutral evidence
- ev2: $X_7 \neq \texttt{AA}$
- ev3: $X_7 \neq \texttt{AA}$ and $X_5 = \texttt{AA}$

`bped` output for ev2:

```
1:1   0.736306   0.130566   0.130566   0.00256256
1:2   0.736306   0.130566   0.130566   0.00256256
1:3   0.736306   0.130566   0.130566   0.00256256
1:4   0.736306   0.130566   0.130566   0.00256256
1:5   0.492513   0.251231   0.251231   0.00502513
1:6   0.492513   0.251231   0.251231   0.00502513
1:7   0.000000   0.497487   0.497487   0.00502513
```

`bped` Demo 1



- allelic frequency $f = 0.10$
- ev1: full neutral evidence
- ev2: $X_7 \neq \texttt{AA}$
- ev3: $X_7 \neq \texttt{AA}$ and $X_5 = \texttt{AA}$

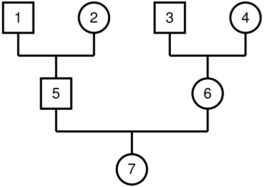
ped file:

```
1 1 0 0
1 2 0 0
1 3 0 0
1 4 0 0
1 5 1 2
1 6 3 4
1 7 5 6
```

ev3 file:

```
1 1 1 1 1 1
1 2 1 1 1 1
1 3 1 1 1 1
1 4 1 1 1 1
1 5 1 0 0 0
1 6 1 1 1 1
1 7 0 1 1 1
```

bped Demo 1

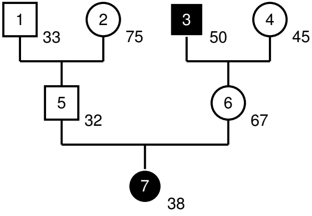


- allelic frequency $f = 0.10$
- ev1: full neutral evidence
- ev2: $X_7 \neq \mathtt{AA}$
- ev3: $X_7 \neq \mathtt{AA}$ and $X_5 = \mathtt{AA}$

bped output for ev3:

```
1:1    0.99      0.005     0.005       0
1:2    0.99      0.005     0.005       0
1:3  0.49005   0.25245   0.25245   0.00505
1:4  0.49005   0.25245   0.25245   0.00505
1:5     1         0         0          0
1:6     0       0.495     0.495      0.01
1:7     0         0         1          0
```

bped Demo 2



$$S(32) = 0.549 \quad S(33) = 0.522$$

$$S(38) = 0.407 \quad S(45) = 0.287$$

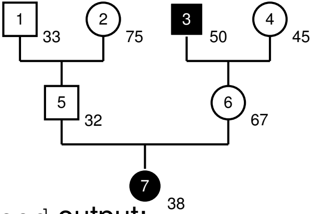
$$S(50) = 0.223 \quad S(67) = 0.095$$

$$S(75) = 0.064$$

	1	1	0	0
	1	2	0	0
	1	3	0	0
ped file:	1	4	0	0
	1	5	1	2
	1	6	3	4
	1	7	5	6

	1	1	1	0.522	"	"
	1	2	1	0.064	"	"
	1	3	0	1	1	1
ev file:	1	4	1	0.287	"	"
	1	5	1	0.549	"	"
	1	6	1	0.095	"	"
	1	7	0	1	1	1

bped Demo 2



$$S(32) = 0.549 \quad S(33) = 0.522$$

$$S(38) = 0.407 \quad S(45) = 0.287$$

$$S(50) = 0.223 \quad S(67) = 0.095$$

$$S(75) = 0.064$$

bped output:

```
1:1   0.961766   0.0189514    0.0189514    0.000331633
1:2   0.995236   0.00236164   0.00236164   4.1211e-05
1:3   0          0.495177     0.495177     0.00964697
1:4   0.988769   0.00557344   0.00557344   8.36409e-05
1:5   0.962757   0.0331219    0.0040797    4.14015e-05
1:6   0.0325306  0.959105     0.0027724    0.00559169
1:7   0          0.034096     0.96433      0.00157444
```

Outline

1. Introduction
   - Some Recalls
   - A Fictional Genetic Study
   - Estimations from Known Genotypes
2. Expectation-Maximization
   - Principle
   - Posterior in Pedigree
   - Estimations from Unknown Genotypes
3. Advanced Stuff
   - Ascertainment Issues
   - Advanced Models
   - Sophisticated Posteriors

The Method

Start from pedigree data, disease status (age, censoring), and possible extra-information (e.g. partial genotyping).

- initialization: random weights w (w_i closer to 1 for affected)
- for iter=$1, 2, 3, \dots$
  - fit a (non-parametric) survival model with weights w

    $$\texttt{fit0} = \texttt{survfit}(\texttt{Surv}(T, \delta) \sim 1, \texttt{weights} = 1 - w)$$

    $$\texttt{fit1} = \texttt{survfit}(\texttt{Surv}(T, \delta) \sim 1, \texttt{weights} = w)$$

  - write the evidence file:

    $$\begin{aligned} \text{affected:}\quad & S_0(T_i)\lambda_0(T_i)\ \text{(AA)},\ S_1(T_i)\lambda_1(T_i)\ \text{(Aa/aA/aa)} \\ \text{unaffected:}\quad & S_0(T_i)\ \text{(AA)},\ S_1(T_i)\ \text{(Aa/aA/aa)} \end{aligned}$$

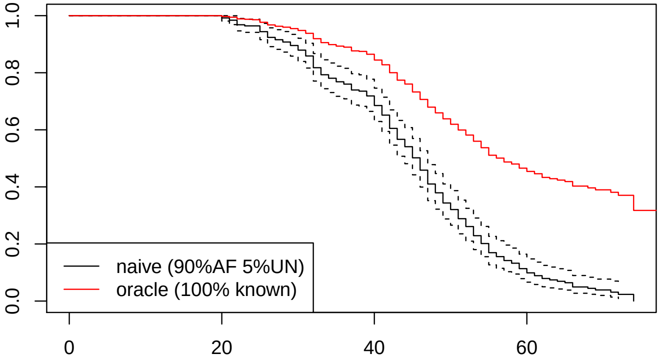
  - use `bped` to update the weights w

    ```
    bped ped ev 0.10
    ```

- output: a fitted survival `fit0`/`fit1` (including survival, confidence intervals, etc.), and post. carrier probabilities w.

Application to Our Simulated Dataset

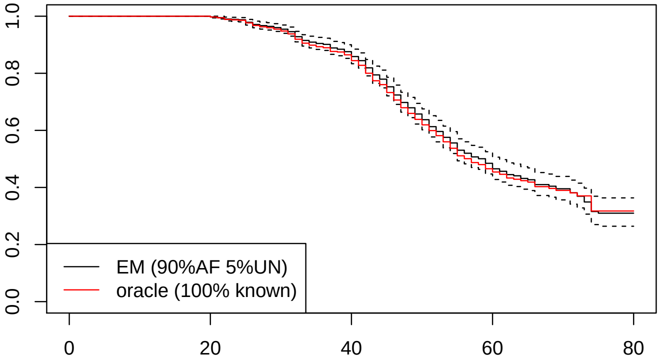
Simulated Dataset: $N = 500$ families, $n = 5000$ individuals, 312 affected, 959 carriers, 75% of affected are carriers.



Naive estimation ⇒ bias

Application to Our Simulated Dataset

Simulated Dataset: $N = 500$ families, $n = 5000$ individuals, 312 affected, 959 carriers, 75% of affected are carriers.



EM ⇒ no bias, and very close to the oracle

Outline

1. Introduction
   - Some Recalls
   - A Fictional Genetic Study
   - Estimations from Known Genotypes
2. Expectation-Maximization
   - Principle
   - Posterior in Pedigree
   - Estimations from Unknown Genotypes
3. Advanced Stuff
   - Ascertainment Issues
   - Advanced Models
   - Sophisticated Posteriors

Modified Simulation

- same model than before, but $f = 0.5\%$ instead of 10%
- $\lambda_0(t) \in (5, 200)/100000$ and $\text{RH}(t) \in (15, 100)$
- $N = 10000$ families of 10 individuals (fixed pedigree)
- ascertainement: at least one affected before age 45

	unaffected	affected	total
non carrier	97695	1310	99005
carrier	761	234	995
total	98456	1544	100000

full dataset

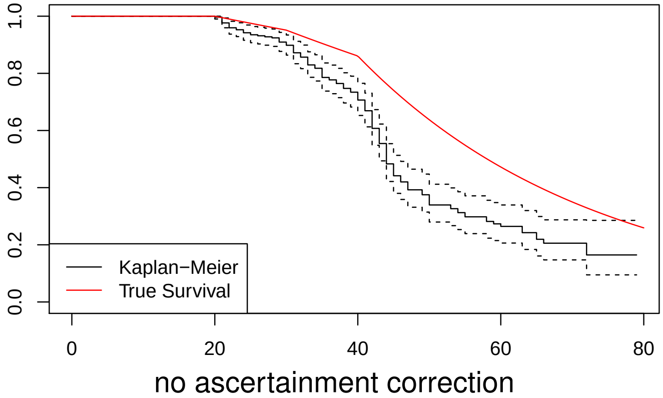
Modified Simulation

- same model than before, but $f = 0.5\%$ instead of 10%
- $\lambda_0(t) \in (5, 200)/100000$ and $\mathrm{RH}(t) \in (15, 100)$
- $N = 10000$ families of 10 individuals (fixed pedigree)
- ascertainement: at least one affected before age 45

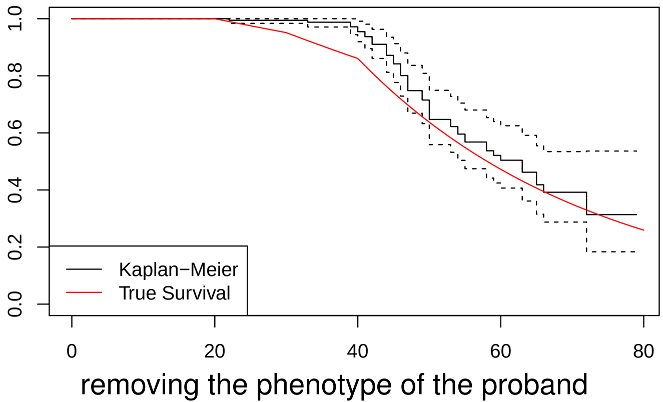
	unaffected	affected	total
non carrier	4301	442	4743
carrier	203	164	367
total	4504	606	5110

after ascertainment

Estimations with 100% Known Genotypes



Estimations with 100% Known Genotypes



Outline

1 Introduction
- Some Recalls
- A Fictional Genetic Study
- Estimations from Known Genotypes

2 Expectation-Maximization
- Principle
- Posterior in Pedigree
- Estimations from Unknown Genotypes

3 Advanced Stuff
- Ascertainment Issues
- Advanced Models
- Sophisticated Posteriors

Just in One Slide

polygenic effects (*e.g.* BOADICEA)

- latent or partially observed
- hypergeometric polygenic model
- usually discretized and approximated

familial frailty (*e.g.* Gorfine, 2013)

- Gaussian frailty shared in the family
- sum-product on a grid of frailty values
- posterior frailty distribution available

parent of origin (*e.g.* amyloid neuropathy)

- $\lambda_1^{\text{pat}}(t) = \lambda(t|X = 10)$ $\lambda_1^{\text{mat}}(t) = \lambda(t|X = 01)$
- almost impossible without EM

covariates (*e.g.* mammographic density for BC)

- effect could depend on carrier status
- how to deal with missing data

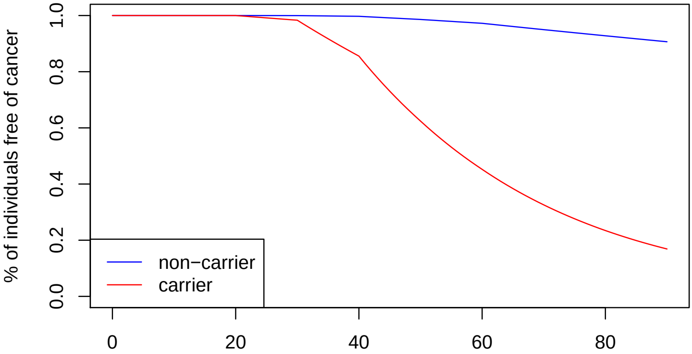
Outline

1. Introduction
   - Some Recalls
   - A Fictional Genetic Study
   - Estimations from Known Genotypes
2. Expectation-Maximization
   - Principle
   - Posterior in Pedigree
   - Estimations from Unknown Genotypes
3. Advanced Stuff
   - Ascertainment Issues
   - Advanced Models
   - Sophisticated Posteriors

Claus Model for BC/OC (Claus, 1991)

Claus' Model: dominant bi-allelic mutation, freq. $q = 0.33\%$

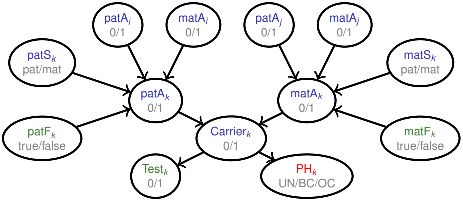
- PCH[1]: non-carrier hazard $\lambda_0(t)$, carrier hazard $\lambda_1(t)$
- male BC → BC25, OC<70 → BC25, OC≥70 → BC35



[1] Piecewise Constant Hazard with cuts in 20,30,40,50,60,70,80.

Claus Model for BC/OC

offspring model with allelic variables (inspired by Lauritzen, 2003)



$$\mathbb{P}(\text{patA}_k = a/b|\text{patA}_i = a, \text{matA}_i = b, \text{patS}_k = \text{pat}/\text{mat}, \text{patF}_k = \text{true}) = 1$$

$$\mathbb{P}(\text{patS}_k = \text{pat}) = 0.5 \quad \mathbb{P}(\text{patA}_k = 1|\text{patF}_k = \text{false}) = q$$

$$\mathbb{P}(\text{Carrier}_k = 1|\text{patA}_k = a, \text{matA}_k = b) = (a \neq 00 \text{ or } b \neq 00)$$

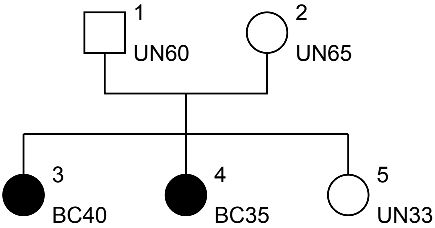
$$\mathbb{P}(\text{patF}_k = \text{false}) = 1\% \quad \mathbb{P}(\text{matF}_k = \text{false}) = 0.01\%$$

$$\mathbb{P}(\text{Test}_k = 1| \text{ Carrier}_k = 1) = 80\% \quad \mathbb{P}(\text{Test}_k = 0| \text{ Carrier}_k = 0) = 98\%$$

$$\mathbb{P}(\text{maleUN}t|\text{C}_k = a) = S_a(25) \quad \mathbb{P}(\text{femaleUN}t|\text{C}_k = a) = S_a(t)$$

$$\mathbb{P}(\text{maleBC}t|\text{C}_k = a) = S_a(25)\lambda_a(25) \quad \mathbb{P}(\text{femaleBC}t|\text{C}_k = a) = S_a(t)\lambda_a(t)$$

Simple Example

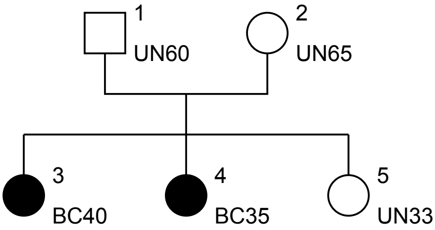


$$\pi(\cdot) = \mathbb{P}(\cdot|\text{FH})$$

Individual i	$\pi(\text{NC})$	1/1	1/2	1/3	1/4	1/5
$\pi(C_i = 1)$	–	52.1	21.2	70.0	71.6	35.4

2 founders, 3 offsprings, 32 variables, 22 cliques, complexity 396

Simple Example

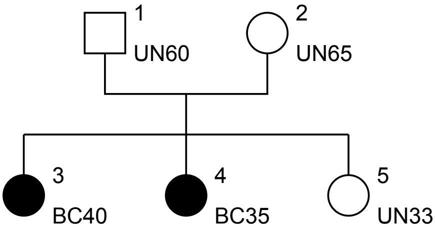


$$\pi(\cdot) = \mathbb{P}(\cdot|\mathrm{FH})$$

1/1	1/2	1/3	1/4	$\pi(\mathrm{C}_{\mathcal{J}})$
1	0	1	1	48.6
0	0	0	0	26.9
0	1	1	1	19.7
1	0	0	1	2.1
1	0	1	0	1.0
0	1	0	1	0.8

Individual *i*	$\pi(\mathrm{NC})$	1/1	1/2	1/3	1/4	1/5
$\pi(\mathrm{C}_i = 1)$	–	52.1	21.2	70.0	71.6	35.4
$\pi(\mathrm{C}_i = 1\|\mathrm{NC} = 3)$	37.4	71.4	28.6	**96.2**	**98.2**	5.6
$\pi(\mathrm{C}_i = 1\|\mathrm{NC} = 4)$	33.1	71.1	29.2	**100.0**	**100.0**	**99.8**
$\pi(\mathrm{C}_i = 1\|\mathrm{NC} = 0)$	26.9	0.0	0.0	0.0	0.0	0.0
$\pi(\mathrm{C}_i = 1\|\mathrm{NC} = 2)$	2.3	71.8	28.2	32.2	66.5	1.3
$\pi(\mathrm{C}_i = 1\|\mathrm{NC} = 5)$	0.2	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**

Simple Example

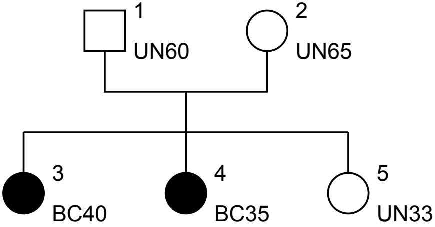


$$\pi(\cdot) = \mathbb{P}(\cdot|\mathrm{FH}, \mathsf{T}_1 = 1)$$

1/1	1/2	1/3	1/4	$\pi(\mathsf{C}_{\mathcal{J}})$
1	0	1	1	91.3
1	0	0	1	3.9
1	0	1	0	1.9
0	0	0	0	1.3
0	1	1	1	0.9
1	1	1	1	0.5

Individual *i*	$\pi(\mathrm{NC})$	1/1	1/2	1/3	1/4	1/5
$\pi(\mathsf{C}_i = 1)$	–	97.8	1.5	94.7	96.7	47.7
$\pi(\mathsf{C}_i = 1|\mathrm{NC} = 3)$	50.6	**99.0**	1.0	**96.2**	**98.2**	5.6
$\pi(\mathsf{C}_i = 1|\mathrm{NC} = 4)$	44.6	**99.0**	1.3	**100.0**	**100.0**	**99.7**
$\pi(\mathsf{C}_i = 1|\mathrm{NC} = 2)$	3.1	**99.0**	1.0	32.2	66.5	1.3
$\pi(\mathsf{C}_i = 1|\mathrm{NC} = 0)$	1.3	0.0	0.0	0.0	0.0	0.0
$\pi(\mathsf{C}_i = 1|\mathrm{NC} = 5)$	0.4	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**

Simple Example

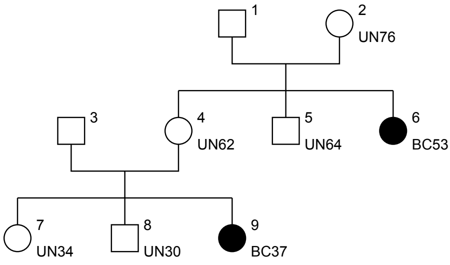


$$\pi(\cdot) = \mathbb{P}(\cdot | \mathrm{FH}, \mathsf{T}_1 = 0)$$

1/1	1/2	1/3	1/4	$\pi(\mathsf{C}_{\mathcal{J}})$
0	0	0	0	46.0
0	1	1	1	33.6
1	0	1	1	17.0
0	1	0	1	1.4
1	0	0	1	0.7
0	1	1	0	0.7

Individual i	$\pi(\mathrm{NC})$	1/1	1/2	1/3	1/4	1/5
$\pi(\mathsf{C}_i = 1)$	–	18.2	35.8	51.7	52.9	26.2
$\pi(\mathsf{C}_i = 1 \| \mathrm{NC} = 0)$	46.0	0.0	0.0	0.0	0.0	0.0
$\pi(\mathsf{C}_i = 1 \| \mathrm{NC} = 3)$	27.5	33.8	66.3	**96.2**	**98.2**	5.6
$\pi(\mathsf{C}_i = 1 \| \mathrm{NC} = 4)$	24.6	33.4	66.7	**100.0**	**100.0**	**99.9**
$\pi(\mathsf{C}_i = 1 \| \mathrm{NC} = 2)$	1.7	34.2	65.8	32.2	66.5	1.3
$\pi(\mathsf{C}_i = 1 \| \mathrm{NC} = 1)$	0.1	5.9	11.1	26.7	55.3	1.1

More Realistic Example

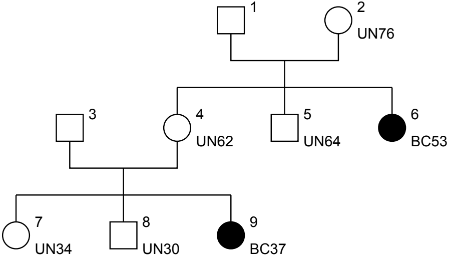


$$\pi(\cdot) = \mathbb{P}(\cdot|\text{FH})$$

Individual i $\pi(\text{NC})$	2/1	2/2	2/3	2/4	2/5	2/6	2/7	2/8	2/9
$\pi(C_i = 1)$ $-$	17.3	5.1	10.7	20.0	11.1	20.8	14.7	15.2	30.0

3 founders, 6 offsprings, 60 variables, 41 cliques, complexity 748

More Realistic Example

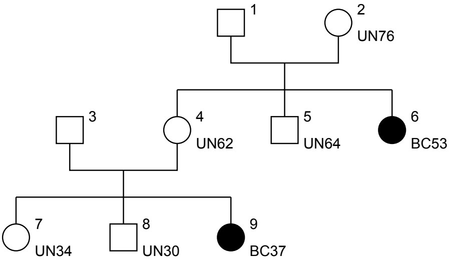


$$\pi(\cdot) = \mathbb{P}(\cdot|\text{FH})$$

2/3	2/4	2/6	2/9	$\pi(C_{\mathcal{J}})$
0	0	0	0	67.5
0	1	1	1	18.0
1	0	0	1	10.0
0	0	1	0	1.9
0	1	0	1	1.4
0	1	1	0	0.4

Individual i	$\pi(\text{NC})$	2/1	2/2	2/3	2/4	2/5	2/6	2/7	2/8	2/9
$\pi(C_i = 1)$	–	17.3	5.1	10.7	20.0	11.1	20.8	14.7	15.2	30.0
$\pi(C_i = 1\|\text{NC} = 0)$	67.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
$\pi(C_i = 1\|\text{NC} = 5)$	7.6	**77.6**	22.4	1.6	**98.5**	36.9	**93.1**	34.7	37.0	**98.1**
$\pi(C_i = 1\|\text{NC} = 6)$	7.0	**77.6**	22.6	2.1	**98.5**	68.1	**97.5**	66.0	68.3	**99.3**
$\pi(C_i = 1\|\text{NC} = 3)$	6.2	14.6	4.2	**81.1**	4.1	14.8	15.6	39.7	42.5	**83.5**
$\pi(C_i = 1\|\text{NC} = 4)$	5.5	44.0	12.6	44.2	55.7	4.4	46.6	47.5	47.8	**97.1**

More Realistic Example

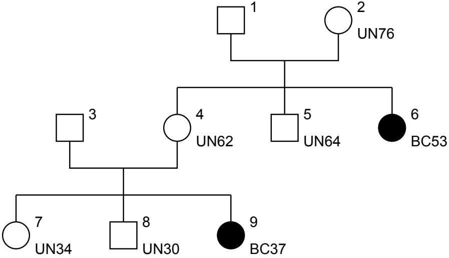


$$\pi(\cdot) = \mathbb{P}(\cdot|\mathrm{FH}, \mathrm{T}_4 = 1)$$

2/3	2/4	2/6	2/9	$\pi(C_{\mathcal{J}})$
0	1	1	1	81.8
0	0	0	0	7.7
0	1	0	1	6.4
0	1	1	0	1.7
1	0	0	1	1.1
1	1	1	1	0.8

Individual *i*	$\pi(\mathrm{NC})$	2/1	2/2	2/3	2/4	2/5	2/6	2/7	2/8	2/9
$\pi(C_i = 1)$	–	70.8	20.7	2.0	**90.9**	45.4	**84.6**	44.6	46.2	**90.2**
$\pi(C_i = 1 \vert \mathrm{NC} = 5)$	34.1	**77.6**	22.4	0.2	**100**	36.9	**93.1**	34.7	37.0	**98.1**
$\pi(C_i = 1 \vert \mathrm{NC} = 6)$	31.3	**77.6**	22.6	0.7	**100**	68.0	**97.5**	66.0	68.3	**99.3**
$\pi(C_i = 1 \vert \mathrm{NC} = 4)$	14.1	**76.2**	21.8	2.0	**98.1**	7.6	**80.5**	9.2	9.7	**95.0**
$\pi(C_i = 1 \vert \mathrm{NC} = 7)$	10.2	**77.7**	23.6	3.8	**100**	**97.4**	**99.8**	**98.9**	**99.0**	**100**
$\pi(C_i = 1 \vert \mathrm{NC} = 0)$	7.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

More Realistic Example



$$\pi(\cdot) = \mathbb{P}(\cdot|\mathrm{FH}, \mathsf{T}_4 = 0)$$

2/3	2/4	2/6	2/9	$\pi(\mathsf{C}_{\mathcal{J}})$
0	0	0	0	80.3
1	0	0	1	11.9
0	1	1	1	4.4
0	0	1	0	2.2
0	1	0	1	0.3
1	0	1	1	0.3

Individual i	$\pi(\mathrm{NC})$	2/1	2/2	2/3	2/4	2/5	2/6	2/7	2/8	2/9
$\pi(\mathsf{C}_i = 1)$	–	5.9	1.7	12.5	4.9	3.8	7.1	8.3	8.6	17.1
$\pi(\mathsf{C}_i = 1\|\mathrm{NC} = 0)$	80.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
$\pi(\mathsf{C}_i = 1\|\mathrm{NC} = 3)$	7.1	12.6	3.6	**83.9**	0.9	15.2	15.5	40.9	43.8	**83.7**
$\pi(\mathsf{C}_i = 1\|\mathrm{NC} = 2)$	4.5	20.9	5.9	73.2	0.1	1.9	24.9	1.4	1.5	70.4
$\pi(\mathsf{C}_i = 1\|\mathrm{NC} = 4)$	3.6	17.2	4.9	**79.5**	20.4	1.8	18.2	**79.5**	**79.6**	**98.8**
$\pi(\mathsf{C}_i = 1\|\mathrm{NC} = 5)$	2.0	**77.6**	22.4	7.2	**93.0**	36.9	**93.1**	34.8	37.1	**98.1**

Summary

Take-Home Messages:

- unbalanced genotyping scheme induces bias
- EM for pedigrees efficiently solves the problem
- `bped` program for posterior marginals

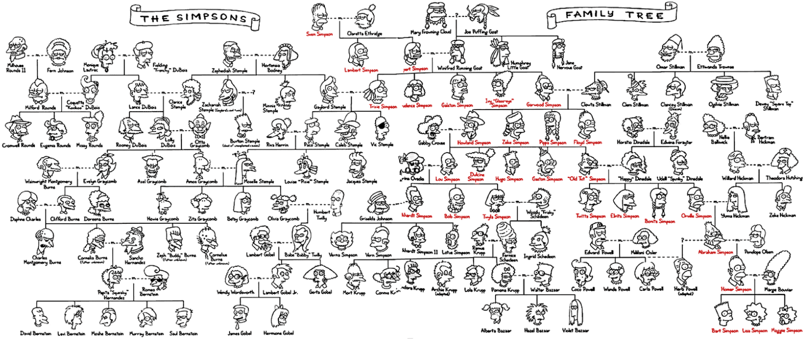
What Next:

- more sophisticated models (frailty, covariates, POO, etc.)
- tackling ascertainement (raking ?)
- clinical relevance of advanced posterior distribution

Many Human Diseases

- Cancers:
  - Breast and Ovarian: *Institut Curie*
  - MSI Cancer and Lynch Syndrome: *Saint-Antoine*
  - Li-Fraumeni: *La Pitié-Salpêtrière*
- Rare Genetic Diseases:
  - Hereditary Amyloid Neuropathy: *Henri Mondor*
  - Pulmonary Arterial HT: *Marie Lannelongue*
  - Huntington Disease: *Hôpital Saint-Anne*
- Common Disease with Genetic Factors:
  - Alzheimer Disease: *CHU Rouen*
  - Diabetes, autism, cardio-vascular, obesity, . . .

Just for Pleasure



Special Thanks:

Homer, Marge, Bart, Lisa, and Maggie Simpson
and their creator . . . Matt Groening