Integrative model for individual and familial risk prediction and variant classification. Application to the MSI cancer and the Lynch Syndrom

A. Lefebvre[1] A. Duval[2] G. Nuel[1]

Context and problematic

- Microsatellite : Seq. of DNA composed of a repetition of nucleotides.
- Microsatellite instability (MSI) found in 15 % of colorectal cancers (CRC), Endometrial cancers (EC), Urothelial cancers (UC), less often in ovarian cancers (OC) and other localizations.
- Due to deleterious mutations in genes involved in the Mismatch repair (MMR) system. (MLH1, MSH2, MSH6, PMS2).
- **Inherited deleterious mutations** (Lynch Syndrome – LS) lead to predisposition earlier in life.
- **Two main issues**
  - Detecting a LS is crucial to adapt treatment and surveillance of patients
  - Next Generation Sequencing –> many Variants of Uncertain Significance (VUS) whose deleterious status is still unknown and must be determined.

State of art

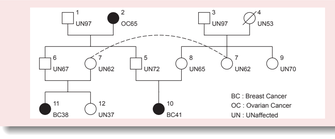
- Models computing LS risk and tumoral risk
  - MMRpro (Chen et al., 2006), $\text{PREM}_{1,2,6}$ (Kastrinos et al., 2011), MMRpredict (Barnetson et al., 2006)
- Models computing variant classification
  - InSiGHT (Goldgar et al., 2008; Thompson et al., 2013) for LS variants
  - ENIGMA (Lindor et al., 2012) for BRCA 1 and 2 variants
- **Our objective**: build a model that combines both approaches

Data structure

Variant data (var_j, V_j)

- Databases (InSiGHT)
- Functional tests

Pedigree (G_i, Unknown X_i)



Individual personal histories and pathology reports (Y_i, patho_i)

- $\{Y_i\}_{i=1,\dots,n}$ set of survival data (age at first cancer onset or censoring)
- $\{\text{patho}_i\}_{i=1,\dots,n}$ set of pathology reports :
  MSI status, IHC testing, somatic BRAF mutation, somatic MLH1 promotor hypermethylation, CRC localization, OC or UC type, ...

$$\mathbb{P}\left(V, \text{var}, X, G, Y, \text{patho}\right) = \prod_j \mathbb{P}\left(\text{var}_j | V_j\right) \mathbb{P}\left(V_j\right) \times \prod_i \mathbb{P}\left(X_i | X_{\text{pat}_i}, X_{\text{mat}_i}\right) \mathbb{P}\left(G_i | X_i\right) \mathbb{P}\left(Y_i | X_i, V_j\right) \mathbb{P}\left(\text{patho}_i | X_i, V_j, Y_i\right)$$

V : set of variants, var : databases and functional tests
X : set of true genotypes, G : set of genotyping tests
Y : set of phenotypes (survival data), patho : set of patho reports

Parameters and assumptions – Version 1

- Parameters
  - variant (allele) frequencies from InSiGHT
  - variant prior classification from InSiGHT
  - genotyping error rates
  - genetic linkage disequilibrium between MSH2 and MSH6
  - incidences per disease $D \in \{\text{CRC}, \text{EC}, \text{UC}, \text{OC}\}$, genotype X_i and sexe : piecewise constant hazard rates $\lambda_D(t|X_i)$. Figure 1 represents the incidences of CRC in MMRpro for females heterozygous carriers of a deleterious mutation in MLH1 and for females non-carriers.
- Assumptions
  - Mendelian transmission (equiprobability of pat. and mat. transmission)
  - Hardy-Weinberg equilibrium

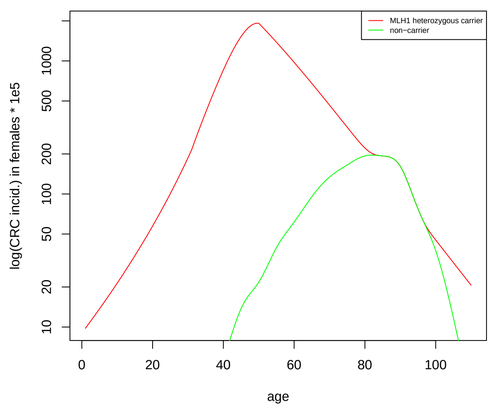


Figure 1: CRC incidences in females for heterozygous carrier of a deleterious mutation in MLH1 and for non-carriers (parameters of MMRpro (Chen et al., 2006)).

Implementation of the phenotypes Y_i

With T_i, the age at first disease onset for individual i

Survival data : $Y_i = \begin{cases} \{T_i > \tau_i\} & \text{if } i \text{ is censored at age } \tau_i \\ \{T_i = \tau_i\} & \text{if } i \text{ is affected at age } \tau_i \end{cases}$

Incidences per disease (CRC, EC, UC, OC) and per genotype: $\lambda_D(t|X_i)$

Let $\lambda_{\text{all}}(t|X_i) = \sum_D \lambda_D(t|X_i)$

- For a censored individual at age τ_i
  $\mathbb{P}(Y_i|X_i) = \mathbb{P}(T_i > \tau_i|X_i) = S_{\text{all}}(\tau_i|X_i) = \exp\left(-\int_0^{\tau_i} \lambda_{\text{all}}(t|X_i)\right)$
- For an affected individual at age τ_i with disease D
  $\mathbb{P}(Y_i|X_i) = \mathbb{P}(T_i = \tau_i|X_i) = S_{\text{all}}(\tau_i|X_i)\lambda_D(\tau_i|X_i)$

Posterior carrier probability

For the sake of simplicity we consider in the carrier risk and disease risk section:

- a single disease D.
- a single gene X associated with D with one deleterious variant and no extra latent variants. Therefore $X \in \{00, 10, 01, 11\}^n$, where $n = \#$ of individuals.
- a dominant mode of inheritence such that $\lambda_D(t|X_i = 00) = \lambda_D^0(t)$ is the basal incidence and $\lambda_D(t|X_i \neq 00) = \lambda_D^1(t)$ is the incidence for carriers.

We denote by

- $\text{ev} = \{\text{ev}_i\}_{i=1,\dots,n}$ an evidence such that $\text{ev}_i = \{G_i, Y_i, \text{patho}_i\}$ a subset of given values for individual i.
- $K_i(X_i, X_{\text{pat}_i}, X_{\text{mat}_i}) = \mathbb{P}\left(X_i | X_{\text{pat}_i}, X_{\text{mat}_i}\right) \times \mathbb{P}\left(G_i, G_i \in \text{ev}_i | X_i\right) \times \mathbb{P}\left(Y_i, Y_i \in \text{ev}_i | X_i, V\right) \times \mathbb{P}\left(\text{patho}_i, \text{patho}_i \in \text{ev}_i | X_i, Y_i, Y_i \in \text{ev}_i, V\right)$ the potential associated with i.

Using the **Bayes rule**, for any subset $X_j \in X$ (e.g. one ind. j),

$$\mathbb{P}(X_j = x_j | \text{ev}) = \frac{\mathbb{P}(X_j = x_j, \text{ev})}{\mathbb{P}(\text{ev})} = \frac{\sum_{X \setminus x_j} \prod_i \text{K}_i(X_i, X_{\text{pat}_i}, X_{\text{mat}_i})}{\sum_X \prod_i \text{K}_i(X_i, X_{\text{pat}_i}, X_{\text{mat}_i})}$$

Problematic : With $X \in \{00, 10, 01, 11\}^n \rightarrow 4^n$configurations

The sum-product algorithm (Koller and Friedman, 2009) is equivalent to the latest version of Elston-Stewart algorithm (Totir et al., 2009).
Complexity drops to $\mathcal{O}(n \times 4^{\text{TW}})$ with TW = 3 to 5 in most cases.

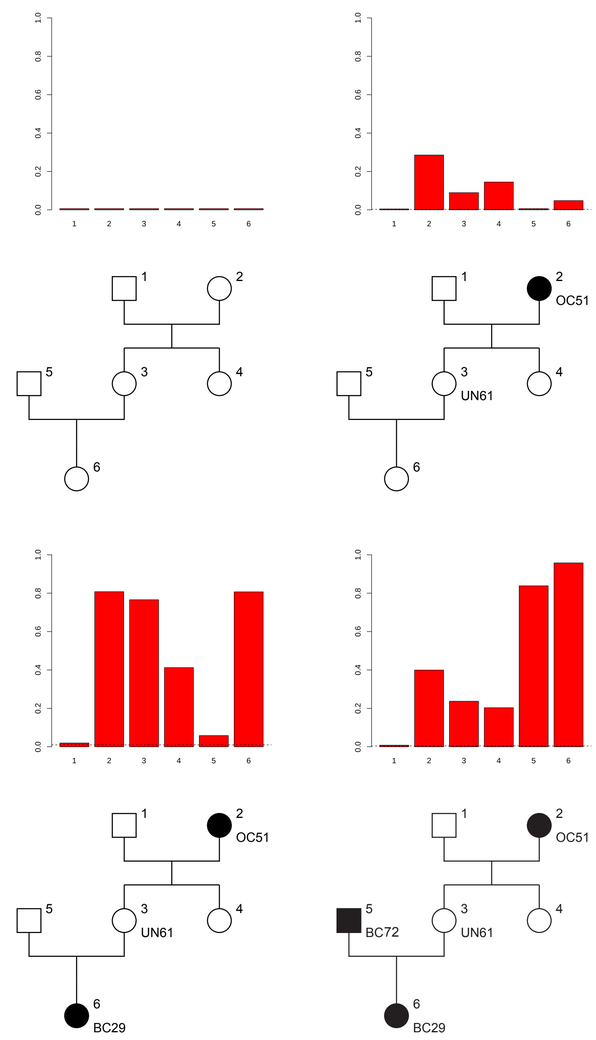


Figure 2: Probability of being a carrier for each individual of a family with evolving ev. Variant frequency and penetrance has been taken from BRCA1 deleterious variant and D = Breast Cancer.

Disease risks with competing risk of death

We denote by $\pi(\tau) = \mathbb{P}(X_i \neq 00|\text{ev})$
$S(t|\text{ev}) = \sum_{X_i} \mathbb{P}(T > t, X_i|\text{ev}) = \pi(\tau)\frac{S_1(t)}{S_1(\tau)} + (1 - \pi(\tau))\frac{S_0(t)}{S_0(\tau)}$
$\pi(t|\text{ev}) = (\pi(\tau)S_1(t)) \,/\, (S(t|\text{ev})S_1(\tau))$
$\boldsymbol{\lambda_D(t|\text{ev})} = \pi(t|\text{ev})\lambda_D^1(t) + (1 - \pi(t|\text{ev}))\lambda_D^0(t)$
$T^* = \min(T_D, T_{\text{death}})$
$\lambda_{\text{both}}(t|\text{ev}) = \boldsymbol{\lambda_D(t|\text{ev})} + \lambda_{\text{death}}(t)$ with λ_{death} from (INED, 2017)
$\mathbb{P}(T \leq t|\text{ev}) = \int_\tau^t S_{\text{both}}(u)\boldsymbol{\lambda_D(u)}\boldsymbol{du}$

Variant classification and individual risks with combined approach

- $\boldsymbol{X_i} = \{\boldsymbol{X_i^{\text{MLH1}}}, \boldsymbol{X_i^{\text{MSH2}}}, \boldsymbol{X_i^{\text{MSH6}}}, \boldsymbol{X_i^{\text{PMS2}}}\} \in \{\mathbf{0}, \boldsymbol{\nu_1}, \boldsymbol{\nu_2}\}^{\mathbf{4}}$: genotype of individual i where 0 denotes a non deleterious variant, ν_1 and ν_2 denote deleterious variants or VUS observed or latent.
- $\mathcal{V} \in \{\mathbf{0}, \mathbf{1}\}^{\mathbf{2\times 4}}$: status of each variant.
- $\text{fam} = \{\text{fam}_j\}_{j=1,\dots,N}$: set of pedigree structures and evidences related to N families.

Then for all j, $\mathbb{P}(\text{fam}_j|\mathcal{V} = \nu)$ is computed as explained in section "posterior carrier probability" and

$$\mathbb{P}(\mathcal{V} = \nu|\text{fam}) = \frac{\sum_j \mathbb{P}(\text{fam}_j|\mathcal{V} = \nu)\mathbb{P}(\mathcal{V} = \nu)}{\sum_{\nu'} \mathbb{P}(\mathcal{V} = \nu')}$$

$$\nu^{\text{MAP}} = \arg\max_\nu \mathbb{P}(\mathcal{V} = \nu|\text{fam}) \quad \text{MAP: Maximum a Posteriori}$$

$$\mathbb{P}(Y_i|\text{fam}) = \sum_\nu \mathbb{P}(Y_i|\text{fam}, \mathcal{V} = \nu)\mathbb{P}(\mathcal{V} = \nu)$$

Example of results for variant status

We consider three VUS (V_1, V_2, V_3) and five families (A,B,C,D,E). Table 1 represents the sequenced VUS per family.

	A	B	C	D	E
V_1	X			X	
V_2		X	X	X	
V_3		X			X

Table 1: Table of the sequenced variants per family A, B, C, D, E

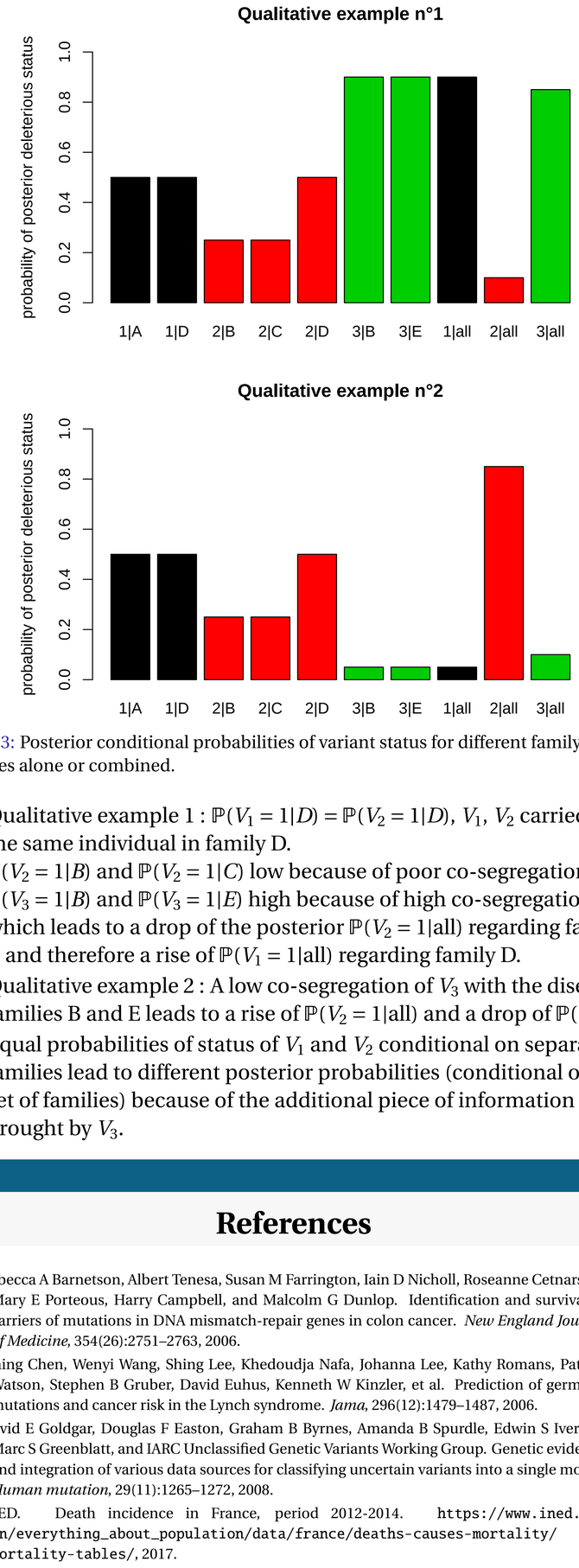


Figure 3: Posterior conditional probabilities of variant status for different family histories alone or combined.

- Qualitative example 1 : $\mathbb{P}(V_1 = 1|D) = \mathbb{P}(V_2 = 1|D)$, V_1, V_2 carried by the same individual in family D.
  $\mathbb{P}(V_2 = 1|B)$ and $\mathbb{P}(V_2 = 1|C)$ low because of poor co-segregation.
  $\mathbb{P}(V_3 = 1|B)$ and $\mathbb{P}(V_3 = 1|E)$ high because of high co-segregation which leads to a drop of the posterior $\mathbb{P}(V_2 = 1|\text{all})$ regarding family B and therefore a rise of $\mathbb{P}(V_1 = 1|\text{all})$ regarding family D.
- Qualitative example 2 : A low co-segregation of V_3 with the disease in families B and E leads to a rise of $\mathbb{P}(V_2 = 1|\text{all})$ and a drop of $\mathbb{P}(V_1|\text{all})$.
- Equal probabilities of status of V_1 and V_2 conditional on separate families lead to different posterior probabilities (conditional on the set of families) because of the additional piece of information brought by V_3.

References

Rebecca A Barnetson, Albert Tenesa, Susan M Farrington, Iain D Nicholl, Roseanne Cetnarskyj, Mary E Porteous, Harry Campbell, and Malcolm G Dunlop. Identification and survival of carriers of mutations in DNA mismatch-repair genes in colon cancer. *New England Journal of Medicine*, 354(26):2751–2763, 2006.

Sining Chen, Wenyi Wang, Shing Lee, Khedoudja Nafa, Johanna Lee, Kathy Romans, Patrice Watson, Stephen B Gruber, David Euhus, Kenneth W Kinzler, et al. Prediction of germline mutations and cancer risk in the Lynch syndrome. *Jama*, 296(12):1479–1487, 2006.

David E Goldgar, Douglas F Easton, Graham B Byrnes, Amanda B Spurdle, Edwin S Iversen, Marc S Greenblatt, and IARC Unclassified Genetic Variants Working Group. Genetic evidence and integration of various data sources for classifying uncertain variants into a single model. *Human mutation*, 29(11):1265–1272, 2008.

INED. Death incidence in France, period 2012-2014. `https://www.ined.fr/en/everything_about_population/data/france/deaths-causes-mortality/mortality-tables/`, 2017.

Fay Kastrinos, Ewout W Steyerberg, Rowena Mercado, Judith Balmaña, Spring Holter, Steven Gallinger, Kimberly D Siegmund, James M Church, Mark A Jenkins, Noralane M Lindor, et al. The PREMM1, 2, 6 model predicts risk of MLH1, MSH2, and MSH6 germline mutations based on cancer history. *Gastroenterology*, 140(1):73–81, 2011.

Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

Noralane M Lindor, Lucia Guidugli, Xianshu Wang, Maxime P Vallée, Alvaro NA Monteiro, Sean Tavtigian, David E Goldgar, and Fergus J Couch. A review of a multifactorial probability-based model for classification of BRCA1 and BRCA2 variants of uncertain significance (VUS). *Human mutation*, 33(1):8–21, 2012.

Bryony A Thompson, David E Goldgar, Carol Paterson, Mark Clendenning, Rhiannon Walters, Sven Arnold, Michael T Parsons, Walsh Michael D, Steven Gallinger, Robert W Haile, et al. A multifactorial likelihood model for MMR gene variant classification incorporating probabilities based on sequence bioinformatics and tumor characteristics: a report from the Colon Cancer Family Registry. *Human mutation*, 34(1):200–209, 2013.

Liviu R Totir, Rohan L Fernando, and Joseph Abraham. An efficient algorithm to compute marginal posterior genotype probabilities for every member of a pedigree with loops. *Genetics Selection Evolution*, 41(1):52, 2009.

Acknowledgment

This work has been funded by *Ligue Nationale Contre le Cancer* (PhD grant).

[correspondance: `alexandra.lefebvre@math.cnrs.fr`]

[1] Probability and Statistics (LPSM, CNRS 8001), Sorbonne Université, Paris; [2] Microsatellite instabiliy and cancer (CRSA, INSERM UMR 938), Saint-Antoine Hospital, Paris.