Précision et biais des estimations de l'heritabilité avec des SNPs

Andrès Legarra
(raconté par Hélène Gilbert)

- la notion de parenté / parenté génomique
- l’héritabilité estimée par Yang et al sur la taille chez l’homme
- est ce une erreur d’estimation ?

Common SNPs explain a large proportion of the heritability for human height

Jian Yang[1], Beben Benyamin[1], Brian P McEvoy[1], Scott Gordon[1], Anjali K Henders[1], Dale R Nyholt[1], Pamela A Madden[2], Andrew C Heath[2], Nicholas G Martin[1], Grant W Montgomery[1], Michael E Goddard[3] & Peter M Visscher[1]

first:

la parenté (génomique):

cet inconnu

Version 2

Messages

- Le LD décrit la parenté (+ ou – ancienne)
- La parenté crée le LD
  - Ce sont deux faces de la même chose – aux mutations près

- Donc ce que l'on fait avec la parenté, on peut le faire avec des SNPs
  - Evaluation génétique
  - Estimation de l'héritabilité
    - ➔ J'ai (Andrès) cherché quelle précision d'estimation pour des individus « non apparentés »
  - Corrections lors des études qui impliquent individus « corrélés » (i.e. GWAS appliqué à un ensemble de familles nucléaires)
    - Tout cela en relation avec un article publié dans Nature Genetics

Mesures de l'apparentement

- Coefficient de parenté f_{ij} (Malécot coefficient, « kinship ») : se définit comme
  - probability(IBD) pour 2 allèles pris au hasard chez deux individus
  - excès d'allèles partagés par rapport à l'équilibre de H-W (Wright; *can be negative* !!)

→ Message : IBD = proxy de l'identité IBS (inconnue) au gène d'intérêt

+ Coefficient d'apparentement est généralement positif (si fondateurs sont considérés comme non apparentés et équilibre de H-W)
+ Mais il n'y a pas de besoin d'imposer cette contrainte

Estimation de l'apparentement à l'aide de SNP

Rationale for estimating genealogical coancestry from molecular markers

Genetics Selection Evolution 2011, 43:27

Toro, M.A., García-Cortés, L.A., Legarra, A.

ETSIA UPM, Ciudad Universitaria 28040 Madrid, Spain.
INIA, Ctra. Coruña Km 7.5 28040 Madrid.
INRA, UR 631 SAGA, F-31326 Castanet Tolosan, France.

Si g_{ik} est la fréquence (= gene content/2) de l'allèle au SNP *k* chez l'individu *i*

	Individu i	*Individu j*	g_{ik}	g_{jk}	$g_{ik}\,g_{jk}$	$(1-g_{ik})\,(1-g_{jk})$
Locus 1	**AA**	**AA**	**1**	**1**	**1**	**0**
Locus 2	**Bb**	**Bb**	**0.5**	**0.5**	**0.25**	**0.25**
Locus 3	**Cc**	**CC**	**0.5**	**1**	**0.5**	**0**
.	.	.	.	.	.	.
.	.	.	.	.	.	.
Locus L	**mm**	**MM**	**0**	**1**	**0**	**0**

=2/4

Identité entre individu *i* et individu *j*, pour *L* locus (*k* = 1,…,*L*)

$$f_{M(i,j)} = \frac{1}{L}\sum_{k} \left[g_{ik}\,g_{jk} + (1-g_{ik})(1-g_{jk}) \right]$$

moléculaire

2) Covariance moléculaire

Si g_{ik} est la fréquence (= gene content/2) de l'allèle au SNP *k* chez l'individu *i*

	Individu i	*Individu j*	g_{ik}	g_{jk}
Locus 1	**AA**	**AA**	**1**	**1**
Locus 2	**Bb**	**Bb**	**0.5**	**0.5**
Locus 3	**Cc**	**CC**	**0.5**	**1**
.	**.**	**.**	**.**	**.**
.	**.**	**.**	**.**	**.**
Locus L	**mm**	**MM**	**0**	**1**
		$\bar{g}_i$	**2/4**	**3.5/4**

Fréquence allélique moyenne intra-individu

$$\bar{g}_i = \frac{1}{L}\sum_k g_{ik}$$

$$Cov_{M(i,j)} = Cov(g_{i\ldots}, g_{j\ldots}) = \frac{1}{L}\sum_k (g_{ik} - \bar{g}_i)(g_{jk} - \bar{g}_j)$$

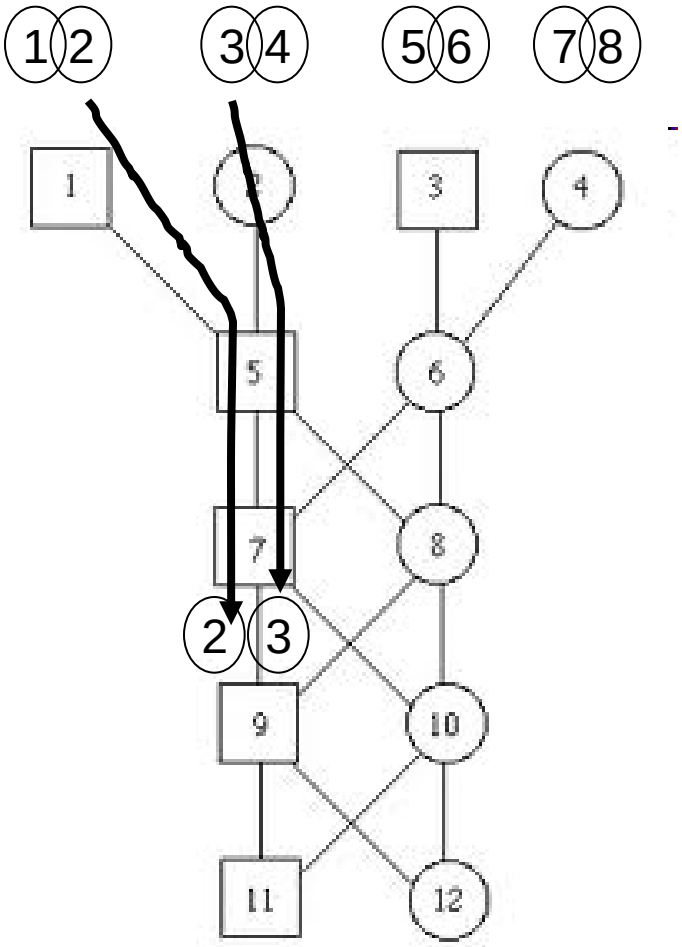
Equivalences

- Malécot : suppose 2N allèles fondateurs

- Si on génotype l'individu 9
- *Alors*,
  - Parenté moléculaire = coefficient d'apparentement de Malécot (IBD)

SNP sont bialléliques

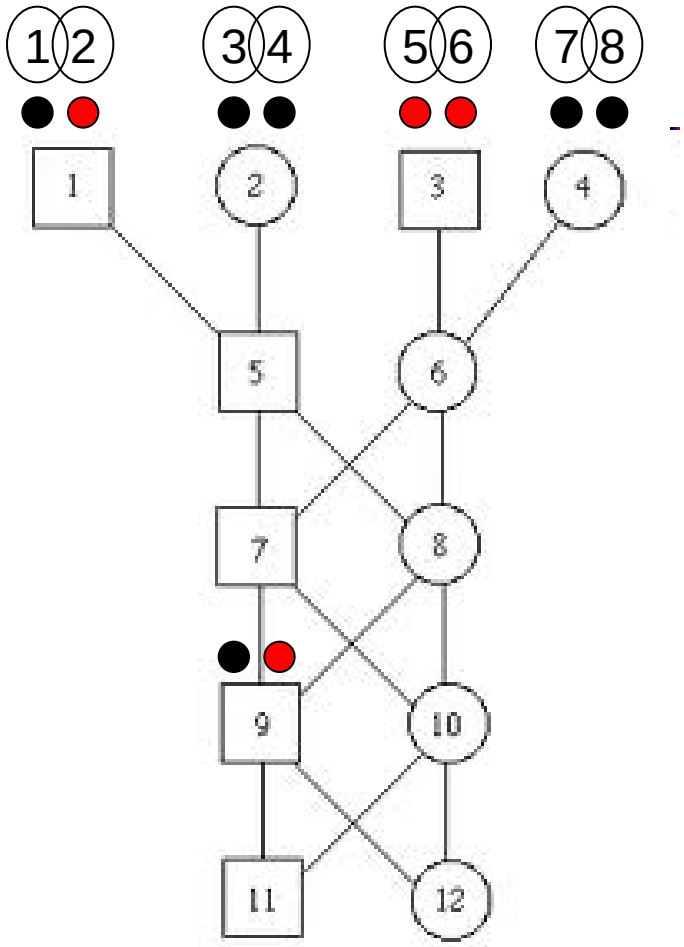
➔ Quel effet sur ces équivalences?



Avec des SNPs…

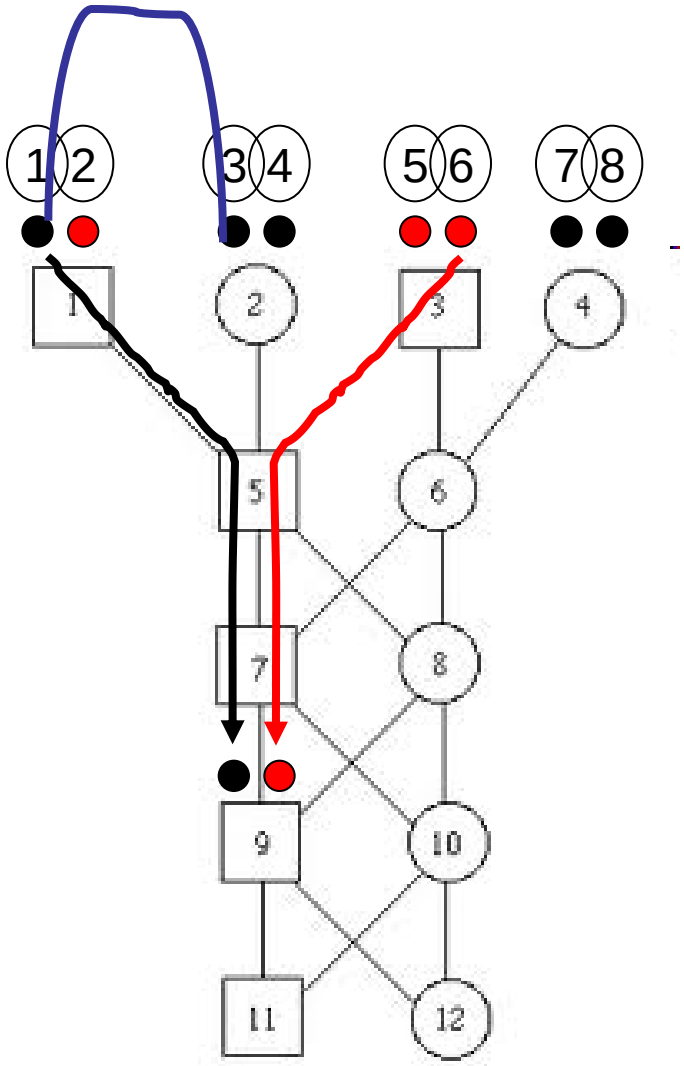
- Si on assigne un état A ou a à chacun des 2N allèles fondateurs au hasard avec une probabilité p et q=1-p
- Si on génotype l'individu 9

→ Que peut on dire de l'allèle hérité par l'individu 9 parmi les 8 allèles fondateurs?



Avec des SNPs…

- L'apparentement moléculaire entre deux individus *i* et *j* est la probabilité que deux allèles soient identiques (alike in state) f_{Mij},
  - Soit parce qu'ils sont identiques par descendance
  - Soit parce qu'ils se ressemblent dans la population fondatrice



$$f_{M_{ij}} = p^2 + q^2 + 2pqf_{ij}$$

En termes de formules (Cockerham, 1969)

- On peut montrer que, *en espérance*,

$$E\ Cov_{Mij} = f_{ij}\,pq$$

Covariance moléculaire

apparentement

$$E\ f_{Mij} = p^2 + q^2 + 2\,pq f_{ij}$$

Apparentement moléculaire

apparentement

- En d'autres termes
  - $Cov(g_i,g_j) = f_{ij}/pq$

$f_{ij} = A_{ij}\ /2$

→ Les estimateurs de l'apparentement f_{ij} à partir de l'information des SNPs peuvent être obtenus aisément

Autres estimateurs , exemples de VanRaden's (2008) **G**'s

fondateurs

Not averaged within-individual but (possibly) within loci

1st ⇨ $$\hat{f}_{VR1ij} = \frac{1}{L}\frac{\sum g_{ik} - p_k \quad g_{jk} - p_k}{\sum p_k \quad 1 - p_k}$$

allelic frequencies are « fixed » (not random)

2nd ⇨ $$\hat{f}_{VR2ij} = \frac{1}{L}\sum \frac{g_{ik} - p_k \quad g_{jk} - p_k}{p_k \quad 1 - p_k}$$

numerically unstable if p ~ 0

Copied or cited by: Astle & Balding, Aulchenko, etc etc

Message

- Des estimateurs de la parenté « vraie » peuvent être construits comme une forme linéaire de scores de génotypes aux SNPs

- Ces estimateurs sont identiques à une certaine forme d'analyse d'association par régression multiple « tous SNPs »

  (comme on verra …)

Common SNPs explain a large proportion of the heritability for human height

Jian Yang[1], Beben Benyamin[1], Brian P McEvoy[1], Scott Gordon[1], Anjali K Henders[1], Dale R Nyholt[1], Pamela A Madden[2], Andrew C Heath[2], Nicholas G Martin[1], Grant W Montgomery[1], Michael E Godd Peter M Visscher[1]

- Ou: La « missing » heritability a toujours été présente

Missing heritability

- La variation des SNP associée à la taille chez l'homme n'explique qu'une très petite fraction de l'héritabilité estimée sans information marqueurs

- Explication la plus probable:
  - Beaucoup de variabilité et peu de puissance

Yang et al

- Modèle mixte pour estimer l'héritabilité
- Trouvent une valeur inférieure à leur attente
- Expliquent cela par le fait que la majorité des QTL affectant le caractère doit avoir une MAF <0.1

Yang et al, modèle

- Tous les SNP ont un effet : vecteur **a**

$$\mathbf{y} = \sum z_k a_k + \mathbf{e} = \mathbf{Za} + \mathbf{e}$$

Régression multiple sur tous SNPs en même temps

- Ce qui ressemble à $\mathbf{g} = \sum z_k a_k = \mathbf{Za}$

  (valeur génétique = somme des effets SNP)

- En standardisant

$$\text{Var}(\mathbf{g}) = \mathbf{ZZ'}\ \sigma^2_u / k = \mathbf{G}\sigma^2_u$$

$$\hat{f}_{VR1ij} = \frac{1}{L} \frac{\sum g_{ik} - p_k \quad g_{jk} - p_k}{\sum p_k \ 1 - p_k}$$

C'est la même chose (à la notation près)

Méthodes

- Estimation de l'héritabilité par un REML sur les effets SNPs, appliqué à une population « non apparentée », avec une matrice d'apparentement génomique
- Apparentement estimé avec une formule équivalente à VR1, à une correction de la diagonale près (non utilisée désormais)

$$\hat{f}_{VR1ij} = \frac{1}{L} \frac{\sum g_{ik} - p_k \quad g_{jk} - p_k}{\sum p_k \quad 1 - p_k}$$

$$A_{jk} = \frac{1}{N} \sum_i A_{ijk} = \begin{cases} \frac{1}{N} \sum_i \frac{(x_{ij} - 2p_i)(x_{ik} - 2p_i)}{2p_i(1 - p_i)}, & j \neq k \\ 1 + \frac{1}{N} \sum_i \frac{x_{ij}^2 - (1 + 2p_i) x_{ij} + 2p_i^2}{2p_i(1 - p_i)}, & j = k \end{cases}$$

- Individus « non apparentés » : apparentements entre -0,025 et 0,025
  → Est-ce que ce n'est pas la source du problème?

Résultats

- Estimation de $h^2 = 0.45 \pm 0.08$
- Estimation pedigree habituelle autour de 0.8

→ Pourquoi?

Est-ce que l'« apparentement » est un « vrai » apparentement?

- Hypothèse : SNP ne donnent pas l'information d'apparentement réalistes pour l'étude de ce caractères car ce ne sont pas les QTL eux même
  - Quel impact si les MAF des QTL sont < à celles des SNP?

  → Apparentements sous-estimés

Compare Aij estimés avec les SNP de faibles MAF **A*** et celles avec les N SNP **A**

$$A^*_{jk} = \begin{cases} \beta A_{jk}, & j \neq k \\ 1 + \beta(A_{jk} - 1), & j = k \end{cases} \qquad \beta = 1 - \frac{(c + 1/N)}{\mathrm{var}(A_{jk})}$$

c>0 si MAF SNP and QTL sont différentes

Résultats avec SNP faibles MAF

- $h^2 = 0.84 \pm 0.16$

- Sont ils contents de ce résultat ?

This does not prove that the causal variants have MAF < 0.1, but it shows that if this were the case, they could explain the estimated heritability of height (~0.8).

- Pour Andrès c'est un sous-produit de la manipulation, une propriété mathématique qui ne démontre rien sur les fréquences des QTLs.

En parallèle

- En bovin lait, poule, porc, nous estimons des h^2 correctement avec la même méthode *sans* manipulation
  - mais les animaux sont assez apparentés
  - donc l'estimation plus précise

→ Cherche les biais et précisions des estimateurs de l'héritabilité avec individus « non apparentés »

Ou « A la recherche des formules d'autrefois »

Données

- Individus « non apparentés »
  - suppose
    - *s* pseudo-familles de
    - *n* pseudo-cousins liés par
    - *r*, coefficient corrélation entre individus(parenté).
  - Si estime l'héritabilité:

→ Quel biais?

→ Quelle précision de l'estimateur?

Formules pour le biais et la précision (standard error) d'estimations de corrélations intra-classes de Ponzoni & James (1978; TAG) pour le biais et Falconer & MacKay (p.182 in "my" Spanish edition) pour les s.e.

Corrélations intra-classes

- h² étant considérée comme une corrélation

$$t = Cor\ \ y_{ij}, y_{ik}\ \ = \frac{Cov\ \ y_{ij}, y_{ik}}{\sqrt{Var\ \ y_{ij}\ Var\ \ y_{ik}}} = \frac{Cov\ \ u_{ij}, u_{ik}}{\sigma_y^2} = \frac{r\sigma_u^2}{\sigma_y^2} = rh^2$$

$$\hat{h}^2 = \hat{t} / r$$

- Biais pour l'estimation de h^2 :

$$E\left[\hat{h}^2 - h^2\right] = \frac{1}{r}\frac{-2\left[1-t\right]\left(t+\frac{1-t}{n}\right)\left(t+\frac{1-t}{sn}\right)}{s-1}$$

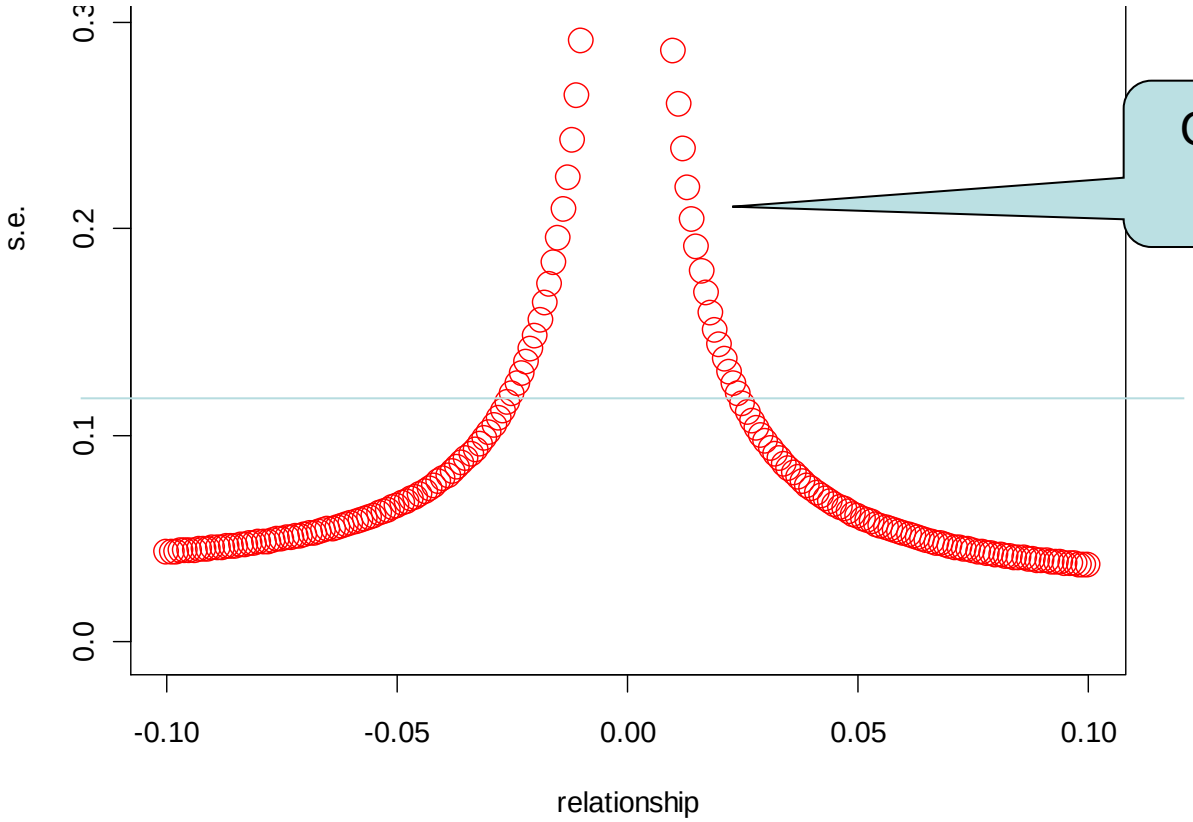
Negligible except for very small « parenté » (<0.001)

- L'erreur standard est :

$$s.e.\left[\hat{h}^2\right] = \frac{1}{r}\left\{\frac{2\left[1+\left(n-1\right)t^2\right]\left[1-t^2\right]}{n\left(n-1\right)\left(s-1\right)}\right\}^{1/2}$$

Graphiquement

- 63 familles de 63 pseudo-cousins liés par *r « parenté »*, et caractère de h^2=0.8 (taille chez l'homme)



Et alors?

- Dans leur étude ils ont trouvé un écart type (théorique) de l'estimateur de 0.08
  - Trop peu?
    - Selon formule le s.e. est ~0.12
    - Pas si different que ça
  - La s.e. publiée est correctement évaluée (d'après le code):
    - Obtenue à partir des dérivées secondes de la vraisemblance, suite à l'algorithme

Test

- L'h^2 a une allure de corrélation, donc on peut utiliser la *z*-transform de Fisher pour tester si

  $h^2 = 0.45 (\pm 0.08)$ est différente de 0.8

  ➔ $p<0.01$

(Andrès) Conclusion

- Papier très intéressant
- Ils ont raisond e dire que l'héritabilité n'est pas manquante et que le modèle mixte l'estime correctement
- L'utilisation d'individus « non apparentés « provoque des problèmes d'estimation
  - énorme écart type: on ne peut (doit) pas tirer trop de conclusions

+ SNP ne permettent pas de retrouver toute la variabilité causale, mais pas seulement à cause des MAF (effets très petits, épistasie, hétérogénéité des locus à effets entre individus)