New approaches to population stratification in genome-wide association studies

A.L. Price, N.A. Zaitlen, D. Reich and N.Patterson

Nature Reviews Genetics, Vol 11, page 459, 2010

23/09/2010

Plan de l'exposé

Plan de l'exposé

Introduction

Problème :

- Les méthodes LD (basique) font l'hypothèse que les individus sont non apparentés et ce n'est malheureusement pas le cas dans la plupart des populations.
  $\Rightarrow$ On masque les vraies associations ou on crée des faux positifs.

Différents types de stratification :

- structure de population (population structure)
- structure familiale (family structure)
- apparentements inconnus (cryptic relatedness)

Difficulté de prendre en compte la structure de population quand il existe également une structure familiale et des apparentements inconnus.
Dans cette publication, on discute des différentes méthodologies qui abordent ces problèmes.

Exemple faux positif :

Population 1 :

	*A*1	*A*2
Cas	160	160
Témoins	40	40

$\Rightarrow \chi^2 = 0$

Population 2 :

	*A*1	*A*2
Cas	160	40
Témoins	160	40

$\Rightarrow \chi^2 = 0$

Population 1 + 2 :

	*A*1	*A*2
Cas	320	200
Témoins	200	80

$\Rightarrow \chi^2 = 7.81$: Faux positif

Exemple faux négatif :

Population 1 :

	*A*1	*A*2
Cas	160	40
Témoins	40	160

$\Rightarrow \chi^2 = 144$

Population 2 :

	*A*1	*A*2
Cas	40	160
Témoins	160	40

$\Rightarrow \chi^2 = 144$

Population 1 + 2 :

	*A*1	*A*2
Cas	200	200
Témoins	200	200

$\Rightarrow \chi^2 = 0$: Faux négatif

Plan de l'exposé

Le contrôle génomique (GC)

- On estime le facteur d'inflation $\lambda_{GC} = Median(Tests)/Median(\chi^2_1)$
- $\lambda_{GC} \approx 1$ indique une absence de stratification
- $\lambda_{GC} > 1$ indique une structure de population ou familiale, ou parfois d'autres choses...

"In general this approach will not maximize power to detect true associations"

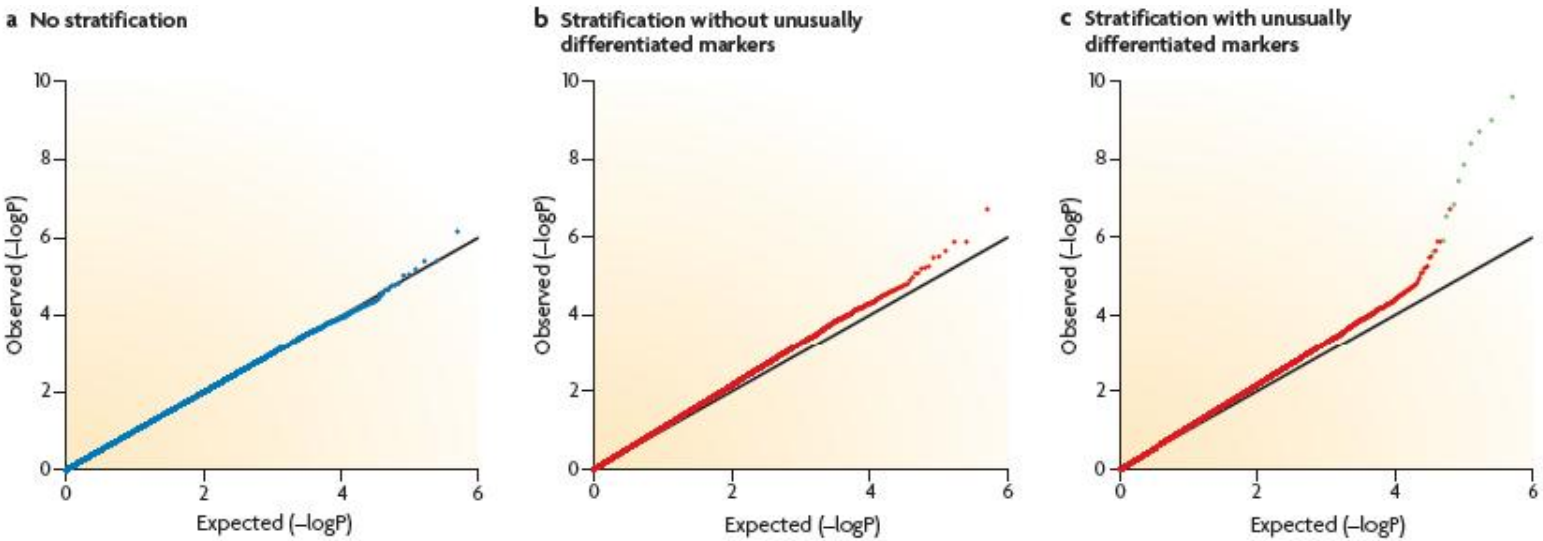

Figure 1 | **P–P plots for the visualization of stratification or other confounders.** The figure shows simulated P–P plots under three scenarios for genome-wide scans with no causal markers. **a** | No stratification: *p*-values fit the expected distribution. **b** | Stratification without unusually differentiated markers: *p*-values exhibit modest genome-wide inflation. **c** | Stratification with unusually differentiated markers: *p*-values exhibit modest genome-wide inflation and severe inflation at a small number of markers.

Plan de l'exposé

Structure d'association de *Pritchard et al. 2000*

Idée :

- Utiliser les marqueurs nuls pour mettre les individus dans des sous-populations. Ensuite, on peut faire un test d'association au sein de ces sous-populations au marqueur considéré.

Différentes étapes :

1. *Pritchard et al 2000a (STRUCTURE) :* La population qu'on étudie est réparti en K sous populations. On fait l'hypothèse que chaque individu a reçu une fraction de ses allèles des sous-populations 1 à K. La première étape consiste donc à estimer ces fractions pour chaque individu.
2. *Pritchard et al 2000b (STRAT) :* Un test de rapport de vraisemblance est mis en place avec comme H_0 : Il n'y a pas d'association entre les allèles et le phénotype (pour le locus candidat) sachant la structure estimée dans l'étape 1.

Bien uniquement si la structure de population est l'unique structure présente

Méthode par ACP de *Price 2006*

Idée :

- On utilise l'ACP sur les génotypes pour regrouper les individus sur des axes continus de variation génétique (vecteurs propres)

Commentaires :

- "it should be noted that top principal components do not always reflect population structure: they may reflect family relatedness, long-range LD or assay artefacts"
- Comme STRUCTURE, l'ACP va appliquer une plus grosse correction aux marqueurs qui présentent de fortes différences de fréquence allélique dans les populations ancestrales.
- "A limitation of the above methods is that they do not model family structure or cryptic relatedness. These factors may lead to inflation in test statistics if they are not explicitly modelled because samples that are correlated are assumed to be uncorrelated"

Family-Based association tests

Idée :

- Séparer l'effet β du SNP en deux effets intra et inter familles (β_w et β_b)

Commentaires :

- On fait généralement le test sur β_w et c'est robuste à la stratification
- FBAT, QTDT...
- On pourrait avoir plus de puissance en prenant les deux β (β_b et β_w) dans le même test mais la partie en β_b n'est pas robuste à la stratification
- Il existe un moyen d'y remedier en transformant l'information inter-familiale en "rank statistic"
- Ces approches, dans tous les cas, ont une perte de puissance liée à la décomposition de l'effet β du SNP

Modèle Mixte

- But: Modéliser les phénotypes en utilisant un mélange d'effets fixes (SNP) et d'effets aléatoires (effet polygénique)
- "Mixed models, which owe their roots to applications in animal breeding, can model population strcture, family structure and cryptic relatedness"

Population structure: a fixed or random effect?

- "However, population structure is actually a fixed effect (...) and spurious associations might result if it is modelled as a random effect based on overall covariance, particulary in the case of unusually differentiated markers"
- Modéliser en effet fixe nécessite une ACP (ou autres...)
- Son idée : mettre les deux dans le modèle, ie modèle mixte avec un effet fixe population, un effet fixe SNP et un effet aléatoire polygénique

Modelling phenotypes as fixed : Autre paragraphe ou il discute d'un modèle mixte avec y=génotype et non plus le phénotype (ROADTRIPS)

Plan de l'exposé

Simulations

Données :

	$Pop1$	$Pop2$
Cas	300	200
Témoins	200	300

- 2 populations
- 100000 SNP: 99900 avec un $F_{ST}(POP1,POP2) = 0.01$ et 100 avec une différence de fréquence allélique entre les 2 populations égale à 0.6.

2 Cas :

- Simulation 1: Tous les individus sont non apparentés
- Simulation 2: Il y a de l'apparentemment dans la population 2 (frères)

Test : Calcul de λ_{GC} pour ces méthodes:

- Test Armitage
- EIGENSTRAT (ACP)
- EMMAX, EMMAX+ACP
- ROADTRIPS

Résultats

Table 1 | **Effectiveness of different approaches for correcting for stratification**

	Simulation 1, $F_{ST} = 0.01$	Simulation 1, $\Delta = 0.6$	Simulation 2, $F_{ST} = 0.01$	Simulation 2, $\Delta = 0.6$
Armitage trend	1.40	48.4	1.57	48.3
EIGENSTRAT	1.00	1.00	1.17	1.14
EMMAX*	1.00	2.05	1.01	1.62
EMMAX* + principal components	1.00	1.02	1.01	1.01
ROADTRIPS	1.00	48.4	1.00	48.3

- Simul 1 F_{ST} : Rien
- Simul 1 Δ : structure de population élevée
- Simul 2 F_{ST} : structure familiale
- Simul 2 Δ : structure de population élevée + structure familiale

Plan de l'exposé

Conclusion

- ”Mixed models are relatively new and untested” ... mais bien sûr ! :)
- ”they seem to offer a practical and comprehensive approach for simultaneously addressing confounding due to population stratification, family structure and cryptic relatedness”
- Structure de population récente : Mixed model
- Structure de population plus ancienne : Mixed model + ACP