Genotypage Infinium: Bilan EPGV

Génotypage Infinium

- 4 puces de Génotypage sur 4 espèces réalisées récemment:

  - Colza: Projet SeqPolyNap
  - Vigne: Projet GrapeReSeq
  - Peuplier: Projet BlackPoplar
  - Pois: Projet GenoPea

4 puces de génotypage Infinium

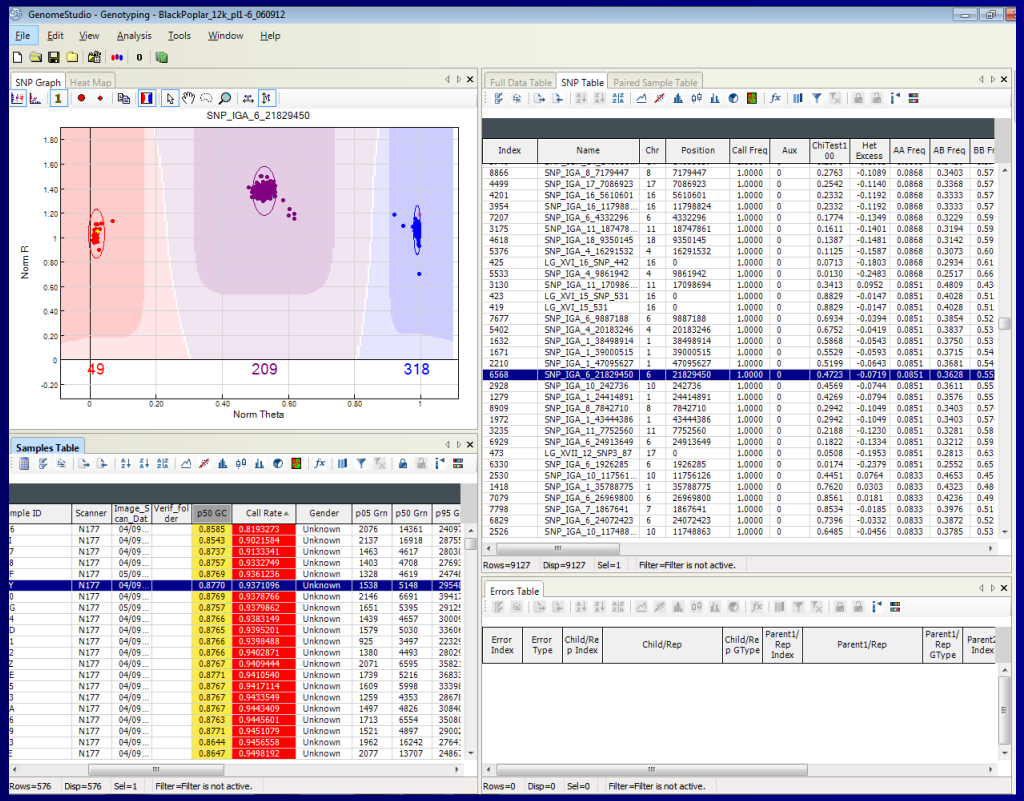
Project	NGS Re-seq Nb d'individus	Infinium Beadchip Objectif	Infinium Beadchip Nom/K billes Nb d'individus	Infinium Design Nb de SNPs	Infinium Deliver Nb de SNPs
SeqPolyNap	5	carto/seq	SeqPolyNap_20K 1440 individus	17 607	15 932 90%
GrapeReSeq	41	diversité	Grape_20K 3072 individus	20 000	18 071 90%
BlackPoplar	52	QTL/ diversité	BlackPoplar_12K 1440 individus	10 331	9 127 88%
GenoPea	16	carto/diversité	Genopea_15K 2688 individus	15 000	13 204 88%

Analyse des génotypes

Construction d'un cluster file visuel à l'aide du logiciel d'Illumina: « Genome studio »

Trois étapes:

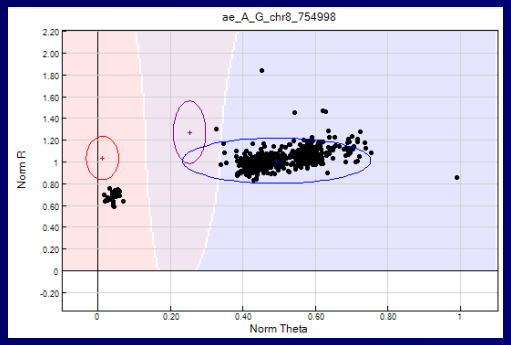
Etape 1-Clustering automatique

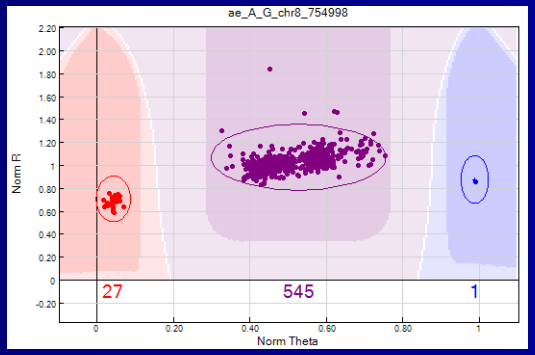


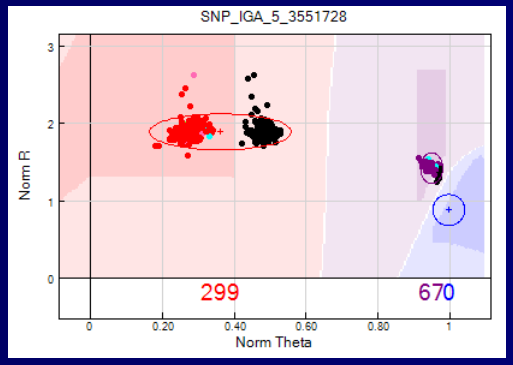
Etape 2-Elimination des échantillons de mauvaise qualité

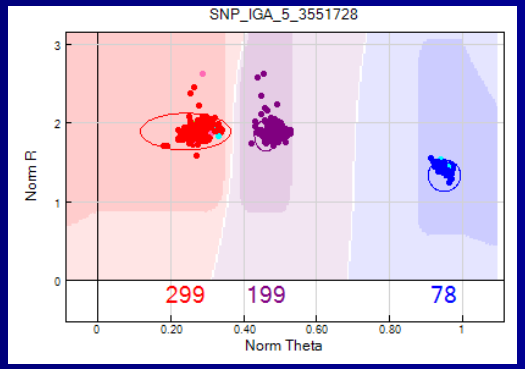
Analyse des génotypes

Etape 3- Correction visuelle du clustering







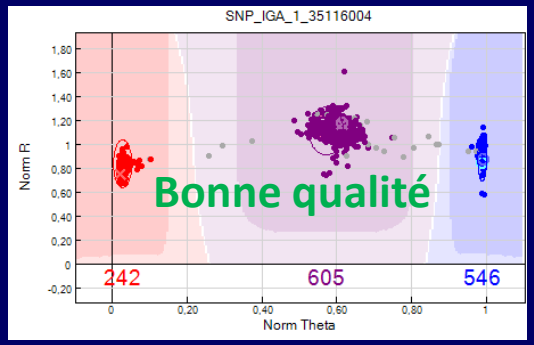


Clustering automatique Genome Studio

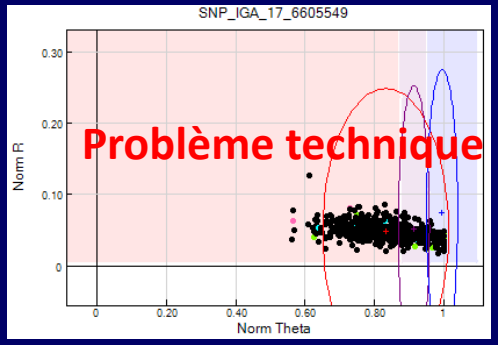
Correction visuelle du Clustering

Analyse des génotypes

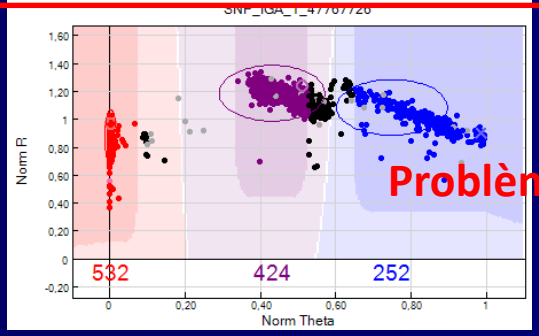
Etape 4- Classification:
Nomenclature EPGV



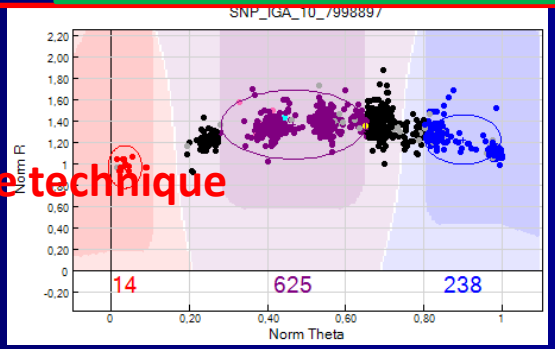
AUX 1
3 clusters bien délimités



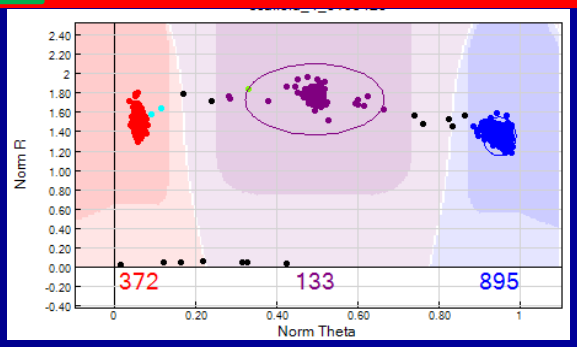
AUX 2
SNP _données manquantes



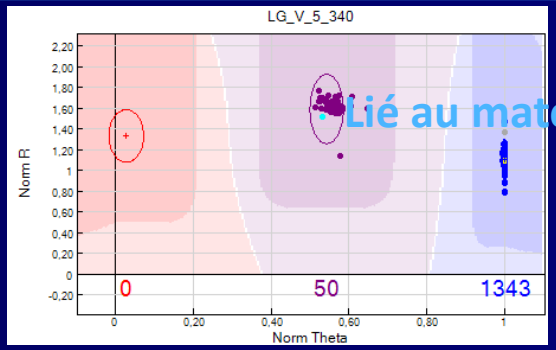
AUX 3
Clusters trop proches



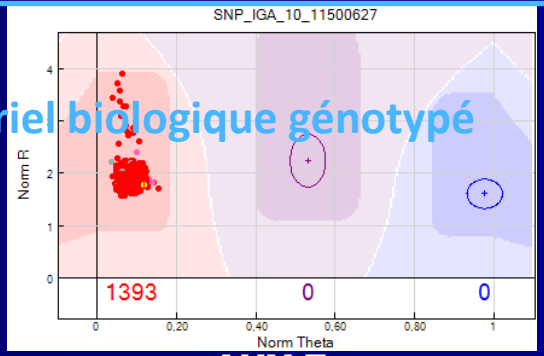
AUX 4
Plus de 3 clusters



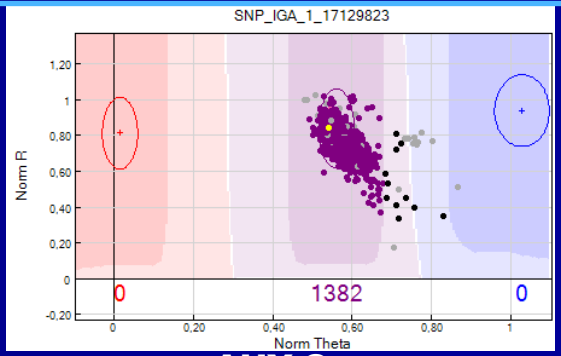
AUX 5
Contrôles manquants



AUX 6
1 cluster est manquant



AUX 7
1 seul cluster homozygote



AUX 8
1 seul cluster hétérozygote

Quelques résultats: puce vigne et peuplier

	vigne	peuplier
SNP sur la puce	18071	9127
SNP Aux 1	**12754 (70,6%)**	**7793 (85%)**
SNP Aux 6,7,8	1994 (11%)	385 (4,2%)
SNP Successful	**14748 (81,6%)**	**8178 (89,2%)**

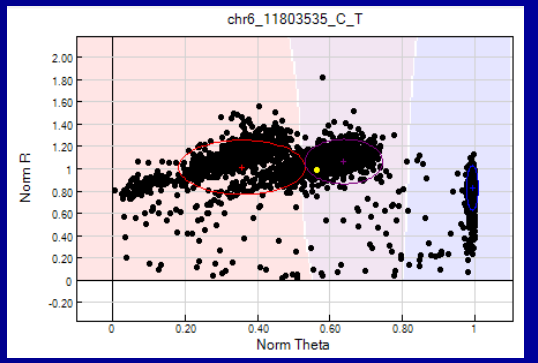
- Un cluster file global a été réalisé sur l'ensemble des individus pour chacune des 2 puces

- Pour la puce peuplier cela s'y prêtait bien car les SNPs de la puce ont été définis pour le matériel à génotyper

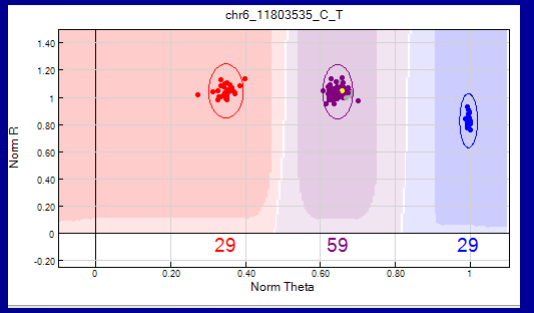
-Pour la puce vigne, les SNPs mis sur la puce ne correspondent pas forcement à la grande diversité du matériel génotypé (différents types de populations provenant de différents labos et pour différentes études). Certains SNPs sont éliminés car mauvais sur l'ensemble des individus, mais OK pour certaines populations.

→ Réaliser un cluster file par type de population

CF global



CF par pop



Quelques résultats puces colza et pois

	colza		
Population	Aviso x Aburamassari	Darmor x Bristol	Darmor x Yudal
SNP sur la puce	15932		
SNP informatifs sur la puce	9498	5020	7704
SNP Aux1	8396 (88,4%)	4430 (88,2%)	6959 (90,3%)
SNP successful	**14089 (88,4%)**		

	Pois											
Population	RIL2	RIL3	RIL4	RIL5	RIL6	RIL7	RIL8	RIL9	RIL10	RIL11	Baccara x pl	J296 x DP
SNP sur la puce	13204											
SNP Aux1	NA (dijon)	6532 (49%)	6175 (47%)	5209 (39%)	NA (dijon)	5951 (45%)	3241 (25%)	NA (dijon)	5407 (41%)	6852 (52%)	NA (dijon)	NA (dijon)
SNP Aux 6,7,8	NA (dijon)	5957 (45%)	6265 (47%)	6707 (51%)	NA (dijon)	5663 (43%)	9029 (68%)	NA (dijon)	7299 (55%)	5853 (44%)	NA (dijon)	NA (dijon)
SNP successful		12489 (94%)	12440 (94%)	11916 (90%)		11614 (88%)	12270 (94%)		12706 (96%)	12705 (96%)		
	13156 marqueurs (99%) sont notés polymorphes dans au moins une population											

Bilan Cluster file « Visuel »

- Efficace mais très chronophage:

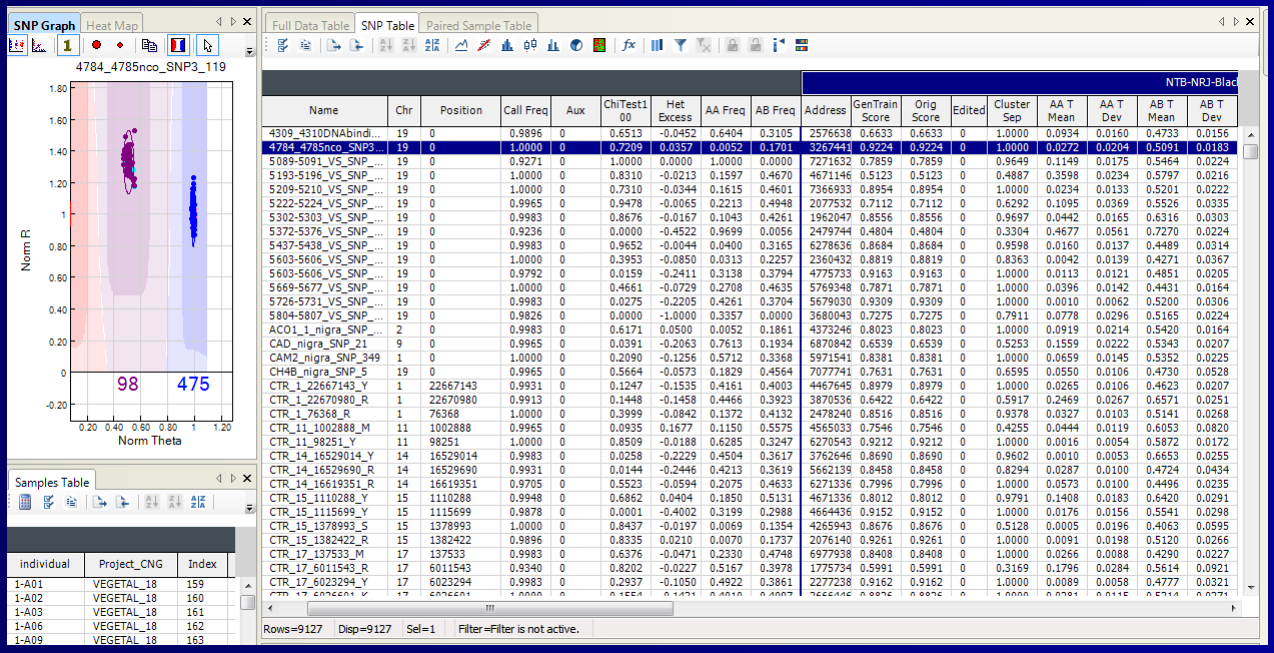
ex: 4 mois d'analyse pour la puce Vigne (plusieurs cluster file successifs) ou 2 mois pour la puce pois.

- Dépend des SNPs mis sur la puce et du matériel ayant servi à leur définition: Nécessité de faire un cluster file pour chaque type de population à génotyper.
- Visuel donc probablement dépendant de celui qui réalise le cluster file notamment pour les marqueurs de moins bonne qualité.
- Pas envisageable pour les futures puces 50K, 100k ou plus

→ Essai de création d'un cluster file automatique

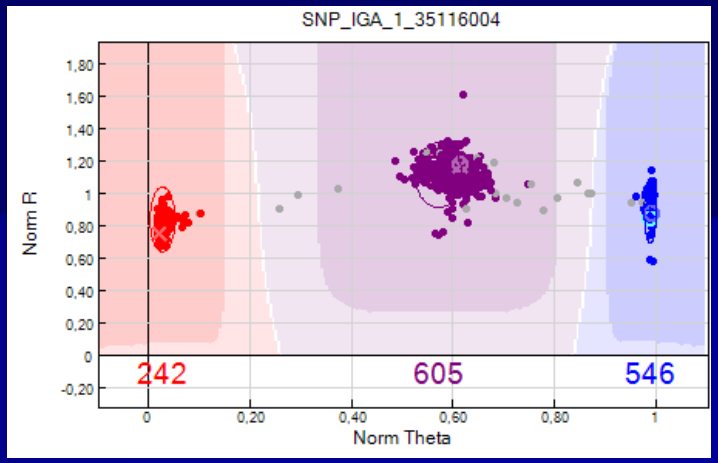
Cluster file « automatique »: objectifs

- Cluster file réalisable par un biologiste sans utilisation de programme bioinformatique.
- A partir des données de Genome Studio (SNP table), identifier les paramètres permettant d'éliminer les marqueurs SNP de mauvaise qualité (problème technique)
- Classement des SNPs en fonction des valeurs de différents paramètres



Cluster file « automatique »: nouvelle classification

Mise au point par Pauline Paul Stephen Raj



Classification des SNP	Définition	Paramètres utilisés
Class 2	Faible intensité	AA/AB/BB R mean ≤0.20
Class 3	Clusters trop proches	AA >0.2;AB <0.2/≥0.8;BB T mean <0.8
Class 4	Dispersion des individus/ centre du cluster	AA/AB/BB T Dev ≥0.05
Class 5	1 cluster est manquant	Freq AA=0, AB & BB >0; BB=0, AA & AB >0
Class 6	1 seul cluster homozygote	Freq AA>0, AB & BB=0; BB>0, AA & AB=0
Class 7	1 seul cluster hétérozygote	Freq AB >0, AA & BB=0
Class 8	Pas d'individus hétérozygotes	Freq AB=0, AA & BB >0
Class 1	3 clusters	Reste des marqueurs

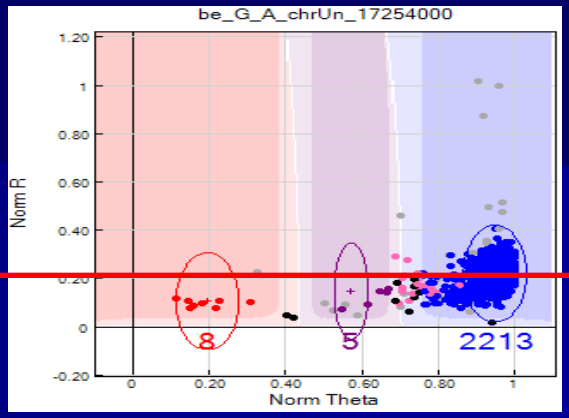
Problème technique

Relatif au matériel biologique génotypé

Bonne qualité

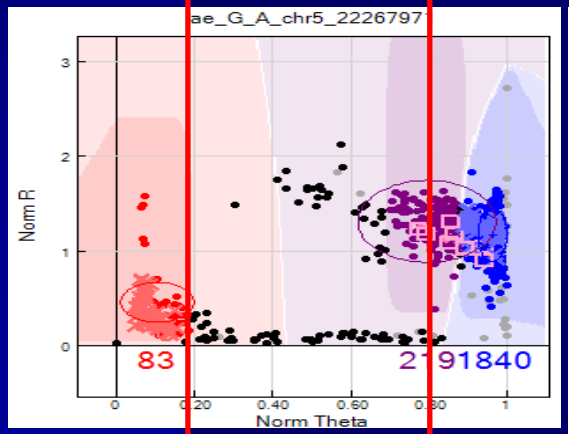
Classification automatique

Classe 2: Marqueurs faible intensité
AA/AB/BB R mean ≤0.20

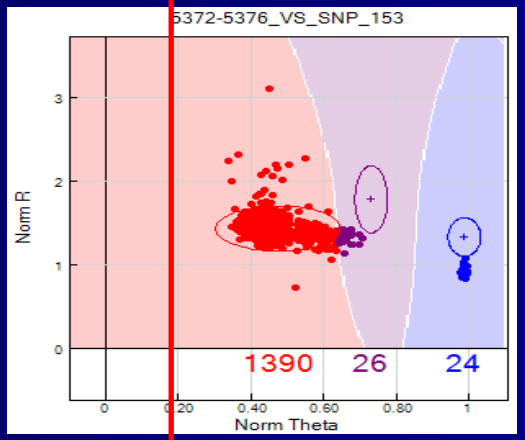


Classe 3: cluster trop proches

AB T mean ≥0.8

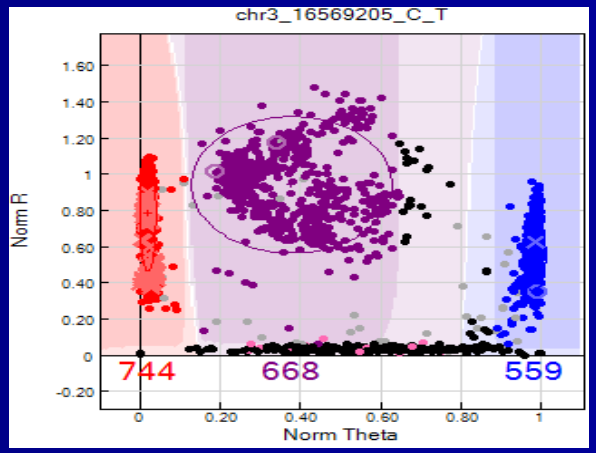


AA T mean >0.2



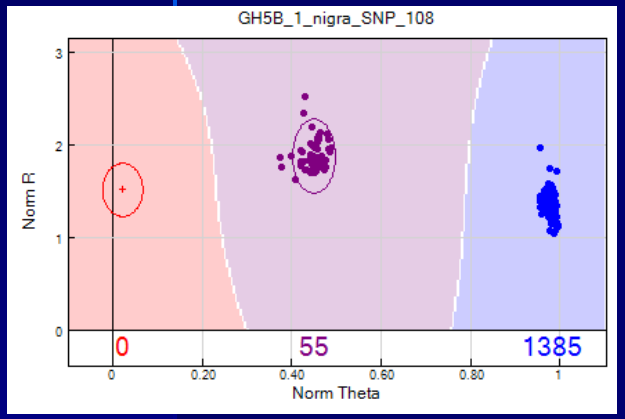
Classe 4: Multiple clusters/dispersion in T dimension

AB T Dev ≥0.05

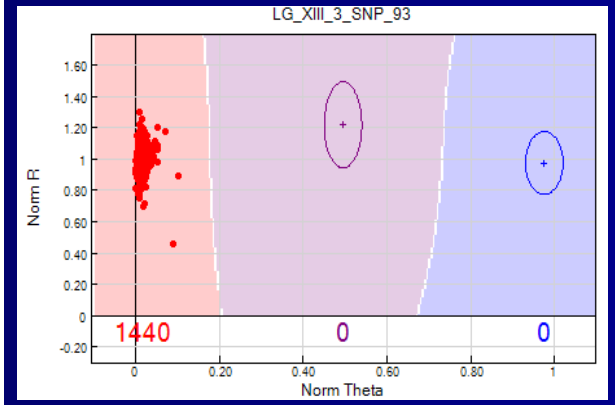


Classification automatique:

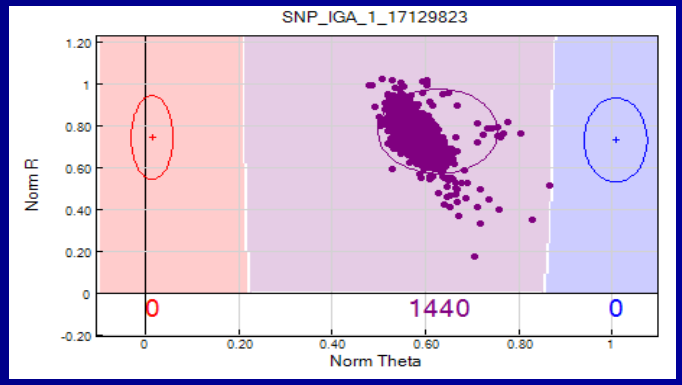
Classe 5: 1 cluster est manquant
Freq AA=0



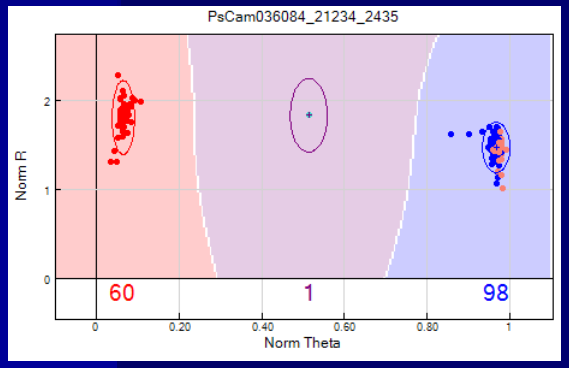
Classe 6: 1 seul cluster homozygote
Freq AA>0, AB & BB=0



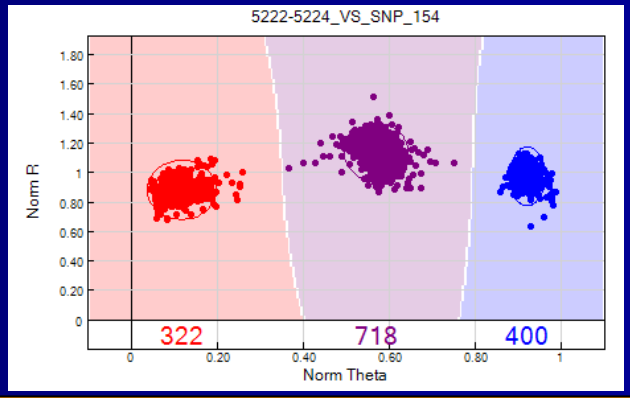
Classe 7: 1 seul cluster hétérozygote
Freq AB >0, AA & BB=0



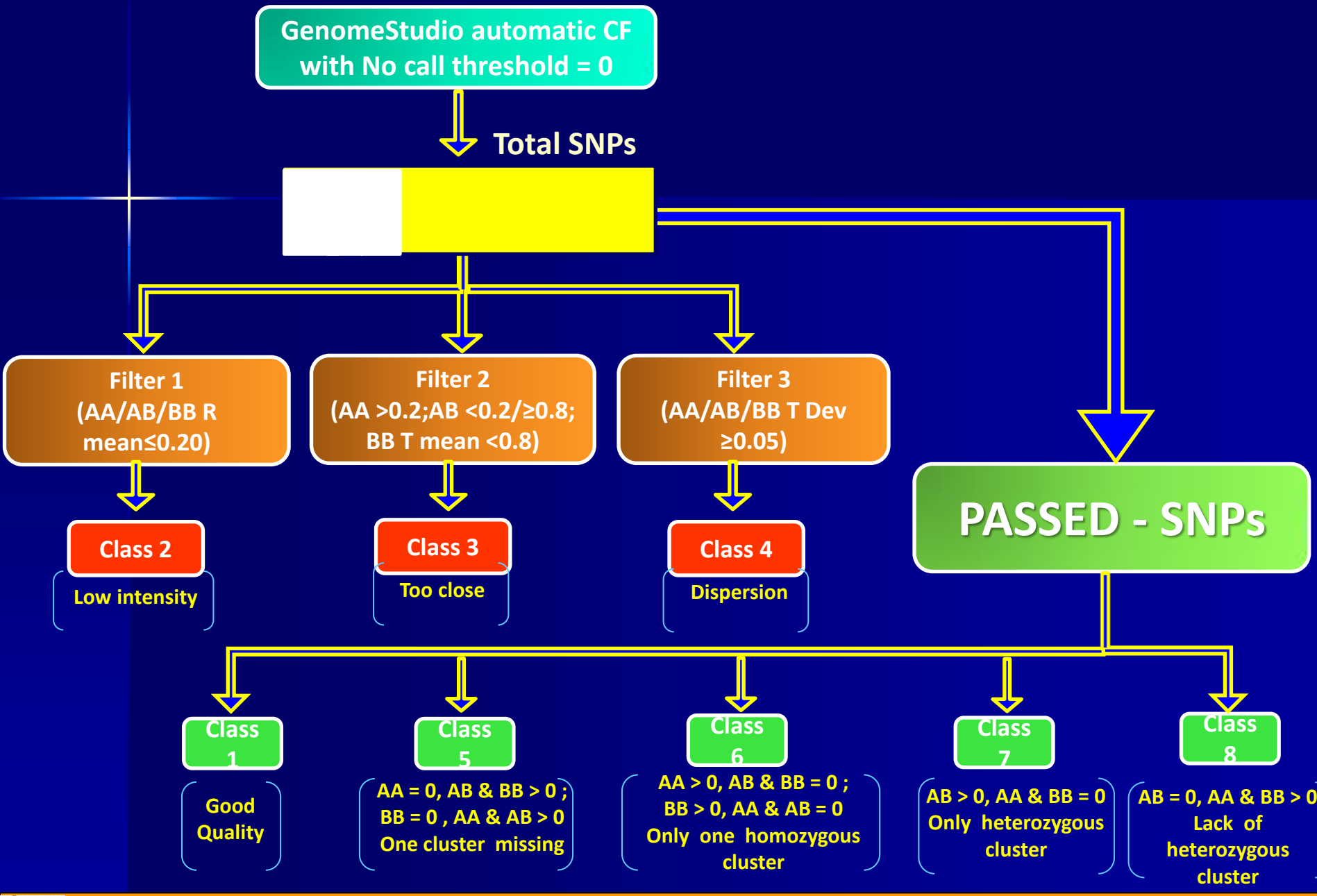
Classe 8: Pas d'individus hétérozygotes
Freq AB=0, AA & BB >0



Classe 1: 3 clusters



Workflow of the « automatic » analysis:



Comparaison CF « visuel »/ CF « automatique »: Peuplier

Peuplier : 9127 SNPs sur la puce						
Description du SNP	CF « visuel »(Aux)			CF « automatique » (classe)		
Bonne qualité	1	**7793 (85,5%)**		1	**6278 (68.8%)**	
Problème lié au matériel biologique	6	278 (3%)	385 (4.2%)	5	120 (1.3%)	120 (1.7%)
	7	101 (1.1%)		6	38 (0.4%)	
	8	6 (0.06%)		7	0 (0%)	
				8	3 (0.03%)	
Problème technique	2	159 (1.7%)	949 (10.2%)	2	87 (1%)	2684 (29.4%)
	3	337 (3.6%)		3	1007 (11%)	
	4	453 (4.9%)		4	1591 (17.4%)	

- 20% de SNPs « sans problème technique » en moins pour le CF «automatique » (classe 1_5_6_7_8) par rapport du CF « visuel » (Aux 1_6_7_8)

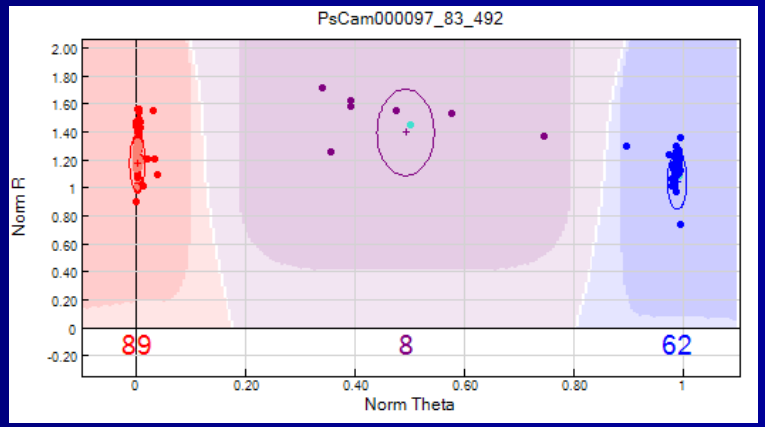
Ces SNPs se retrouvent principalement dans les classes 3 (clusters trop proches) et classe 4 (clusters trop dispersés)

→ Le clustering « automatique » est moins tolérant que le clustering « visuel »

Comparaison CF « visuel »/ CF « automatique »: Pois

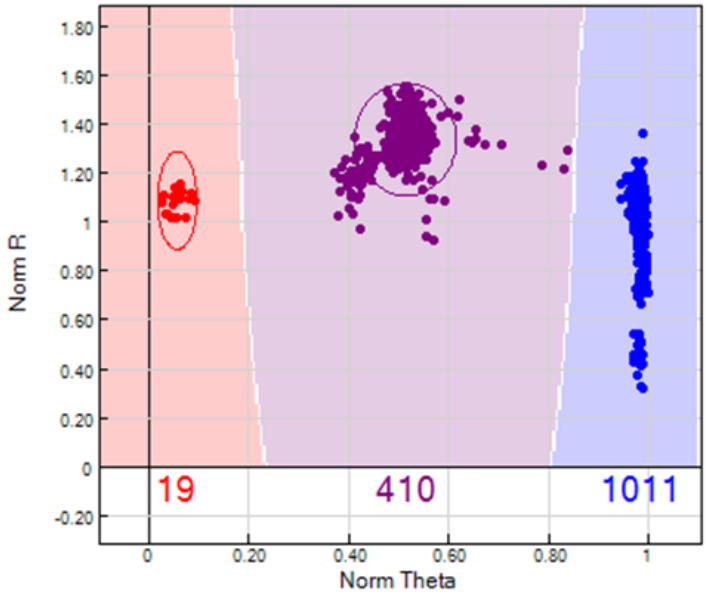
Pois (RIL4) : 13204 SNPs sur la puce						
Description du SNP	CF « visuel » (Aux)			CF « automatique » (classe)		
Bonne qualité	1	**6175 (46.7%)**		1	**3021 (22.9%)**	
Problème lié au matériel biologique	6	0 (0%)	6265 (47.4%)	5	172 (1.3%)	6130 (46.4%)
	7	6258 (47.4%)		6	5956 (45.1%)	
	8	7 (0.05%)		7	0 (0%)	
				8	2 (0.02%)	
Problème technique	2	227 (1.7%)	761 (6.2%)	2	138 (1.1%)	3753 (30.8%)
	3	284 (2.6%)		3	605 (4.6%)	
	4	250 (1.9%)		4	3010 (25.1%)	

- 20% de SNPs « sans problème technique » en moins pour le CF «automatique » (classe 1_5_6_7_8) par rapport du CF « visuel » (Aux 1_6_7_8)
- Ces SNPs se retrouvent principalement dans la classe 4 (cluster trop dispersé)



Limites actuelles du cluster file « automatique »

Impossible de classer les SNPs qui ont plusieurs nuages à une même position homozygote ou hétérozygote



SNP Table
AA R Mean = 1.0893
AA R Dev = 0.1000

Full data table
AA R Mean = 1.09
AA R Dev = 0.05
(Calculated in excel)

SNP Table
BB R Mean = 1.0573
BB R Dev = 0.1000

Full data table
BB R Mean = 1.03
BB R Dev = 0.13
(Calculated in excel)

La SNP_Table générée par Genome Studio donne des informations erronées sur la dispersion des groupes sur l'axe R :AA R Dev ou BB R Dev sont identiques.

La SNP_Table donne des informations sur l'ensemble des individus pour chaque SNP

Si on calcule la dispersion à partir de la Full_Data_table générée par Genome Studio: AA R Dev ou la BB R Dev sont différents.

La Full_Data_Table donne des informations pour chaque individu pour chaque marqueur

Limites actuelles du cluster file « automatique »

La quantité de données étant trop importante pour pouvoir faire le calcul via Excel, il est nécessaire de faire le calcul de la moyenne et de l'écart type à l'aide d'un script en cours d'écriture.

Un des objectifs de ce travail (facilité de mise en place du clustering à partir des données produites par Genome Studio) ne sera donc pas rempli.

Conclusions

- Cluster file « visuel » :
  - Donne de bons résultats
  - Prend beaucoup de temps
  - Classification assez subjective
  - Pas gérable pour des puces de très haute densité

- Cluster file « automatique »:
  - Rapide
  - Classification plus objective car liée à des paramètres mathématiques
  - Possibilité de modifier les paramètres ou d'en choisir d'autres
  - Mais actuellement:
    - Impossibilité de distinguer certaines classes de SNP car le calcul des écart-type est *a priori* faux.
    - Ecriture en cours d'un script pour faire ce calcul (impossible directement avec Excel étant donné la taille des fichiers).

Merci !