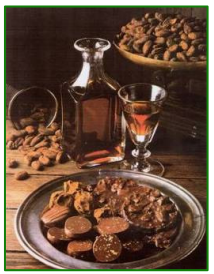
Recherche de SNP et génotypage haut débit de populations de cacaoyers à l'aide du système Illumina Golden gate

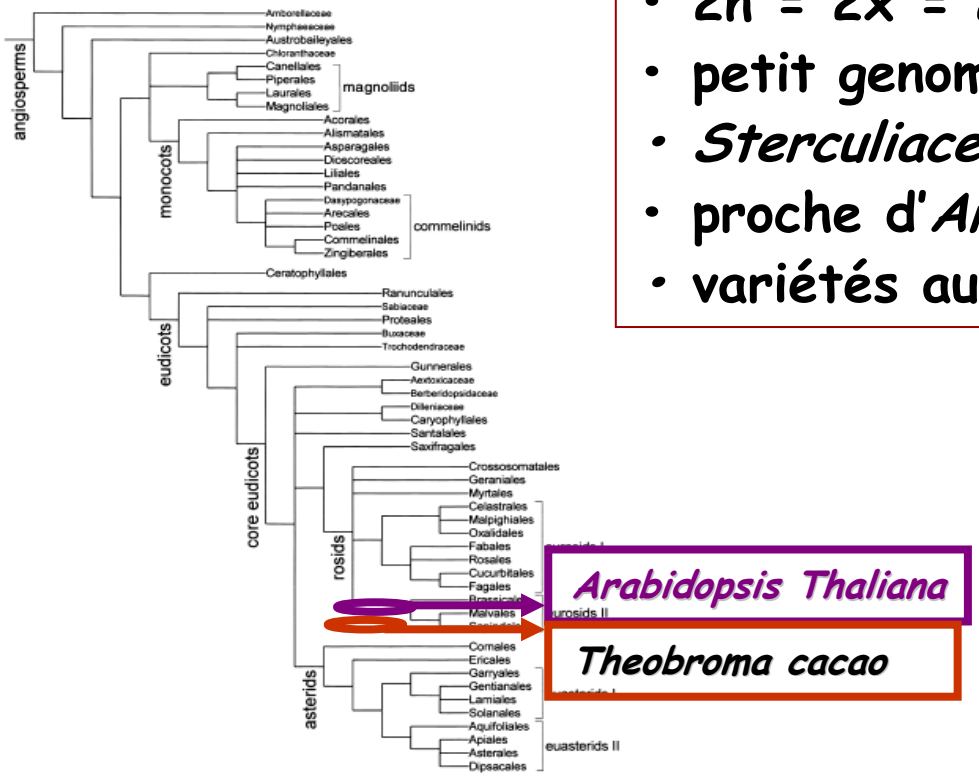


ACAACTGC
ACAACTGC
ACAACTGC
ACGACTGC
ACGACTGC
ACAACTGC
ACAACTGC
ACAACTGC
ACGACTGC
ACGACTGC

C. Lanaud, M. Allegre, X. Argout, O. Fouet, M. Boccara, A. Bérard, D. Brunel

Theobroma cacao L.

- *Theobroma cacao* L.
- originaire d'Amérique du Sud
- 2n = 2x = 20
- petit genome: 400 Mb
- *Sterculiaceae --> Malvaceae*
- proche d'*Arabidopsis*
- variétés autogames et allogames



Angiosperm Phylogeny Group (APGII, 2003)

T.cacao L.: culture et problèmes de sélection

• étapes de fermentation, séchage et torrefaction

20 à 40 fèves/cab.

Production de cacao:

- 3 millions de tonnes /an
- produit principalement par des petits planteurs (70% en Afrique)
- 14 millions de planteurs impliqués dans la production de cacao

Importantes pertes de production dues aux maladies (30% au total)

- *Phytophthora, Moniliophthora perniciosa, Moniliophthora roreri*
- recherche d'une résistance durable par cumul de différentes sources de résistance

Caractères de qualité: "bulk" ou "cacao fins"

- Stratégique pour certains pays
- liés à la variété (Criollo/Trinitario, Nacional)

Approches intégrées conduites pour connaître les déterminants génétiques et moléculaires de ces caractères et les sélectionner

Objectifs du projet SNP

Produire un grand nombre de marqueurs SNP définis dans des gènes ayant une fonction putative et **répartis sur tout le génome** (choix du kit Golden gate - 1536 SNP), et les utiliser, en particulier, pour l'étude des résistances aux maladies et des qualités du produit.

- **Etablir une cartographie génétique dense** pour:

- l'identification et/ou la cartographie fine de QTL (analyses QTL classiques ou études d'association)
- la recherche de gènes candidats (co-localisations avec QTL/eQTL)
- la sélection assistée par marqueurs
- l'appui à un projet de séquençage du génome complet de cacaoyer

- **Etudier la diversité génétique de l'espèce avec des marqueurs fonctionnels**

- structure de la diversité, DL, ……

Production de marqueurs SNP

- **Collection de 149650 EST** récemment produite en collaboration avec le Genoscope *(Argout et al., 2008)* et dans le cadre d'un projet internationnal:

- **56 banques cDNA** produites à partir d'un panel d'organes différents soumis ou non à des stress biotiques et abiotiques, et à partir de 2 génotypes principaux correspondant à 3 origines génétiques contrastées:

SCA6: Forastero de haute Amazonie (Pérou)
ICS1: Trinitario: hybride entre Criollo et Forastero de basse Amazonie (Brésil)

- **Set unigenes de 48594 sequences**
- **61,4% sequences** uniques montrent une similarité avec des gènes d'autres espèces

- **Databases:** **Tropgene DB, ESTtik** (http://esttik.cirad.fr), **cocoagen DB**

Recherche de SNP dans les EST

Difficultés :

- Distinguer les variations alléliques des variations de séquence de gènes paralogues
- Repérer les erreurs liées au séquençage

QualitySNP :

Nouvelle méthode de détection de SNP basée sur une reconstitution par haplotype

(Cette méthode mathématique a démontré son efficacité chez la pomme de terre, le poulet et l'humain).

Méthode : 3 filtres :

1 : *Recherche de tous les SNPs* potentiels et identification des variations inter / intra génotypes.

2 : *Reconstitution par haplotype* pour détecter des SNPs fiables. Identification des paralogues et des SNPs « faux positif » dus aux erreurs de séquençage.

3 : *Attribution d'un score de confiance* pour chaque SNP, basé sur la redondance de séquence et la qualité. Détection des *SNPs synonymes et non-synonymes*.

Recherche de SNP dans les EST

Paramètres :

- Contigs >=4 (minimum 2 copies par allèle) et <=100 membres = 4847 contigs sélectionnés
- Stringence du contigage augmentée pour avoir des contigs avec des blocs d'alignement.
- Lancement des scripts de recherche de SNPs avec une limitation à chaque extrémité, souvent étant une zone de faible qualité.

Résultats :

- 16 639 SNP potentiels
- 5 218 SNP véritables (avec indels) (après filtre haplotype)
- 4150 SNP (1834 contigs) envoyés pour score au CNG/Illum.

Validation d'un set de 1536 SNP

- Analyse par Illumina → score pour chaque SNP
- Elimination des SNP avec scores < 0,4
- Intégration donnés dans AMIGO (GO)
- Vérification manuelle de la qualité du SNP (pas de SNP situé à -60pb, pas d'anomalie de type polyA ou délétion…)

Choix de 1 SNP par contig ayant une annotation GO

↓

1477 SNP

Ajout de 59 SNP pour obtenir un set de 1536 SNP

- Ajout de SNP dans des gènes d'intérêt (SNP multiples) (défense, qualité, gènes cytoplasmiques)

Populations à génotyper par le système Illumina Golden gate

- **1274 individus à génotyper (14 plaques x 96):**

10 plaques de 96 ind déjà génotypées - 4 autres à terminer

- **2 Populations de cartographie:**

- Descendance de référence UPA 402 x UF676; 264 ind.
    - UPA402: Forastero, origine haute Amazonie (Pérou)
    - UF676: Trinitario, hybride entre Criollo et Forastero de basse Amazonie (Brésil)
- Descendance F2 :(SCA6x ICS1) (coll.EST); 200 individus

- **Collection de ressources génétiques:** environ 800 individus issus de la collection internationale de Trinidad et de prospections récentes faites au Pérou

- études de diversité génétique, DL etc…
- études d'associations: résistance aux maladies (Phytophthora/balai de sorcière), qualité, etc….

Difficultés rencontrées (génotypage SNP)

- **Extraction d'ADN:**

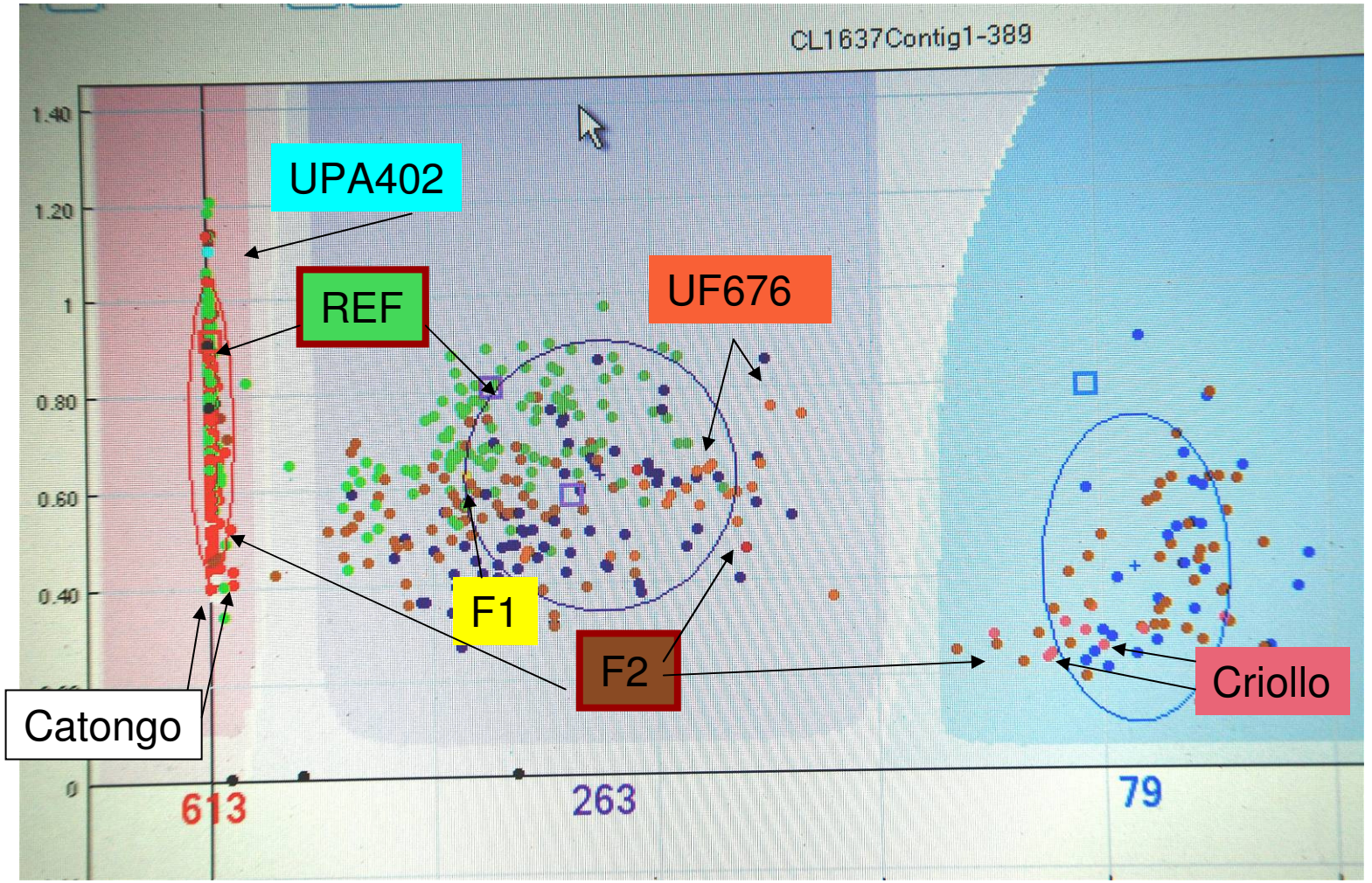
Ne pas négliger **le temps nécessaire pour préparer des ADN de qualité et à la bonne concentration!**

- Chez le cacaoyer, présence de polyssacharides qui gênent les purifications. Les étapes de purification et contrôle qualité avant envoi au CNG ont été particulièrement longues et laborieuses.

- Travail avec du matériel végétal qui ne pousse pas en France : les feuilles arrivent peu fraîches, ce qui complique encore l'obtention d'un ADN de très bonne qualité.

- **Interprétation des données:** beaucoup d'erreurs dans le positionnement des nuages

Exemples de nuages bien positionnés

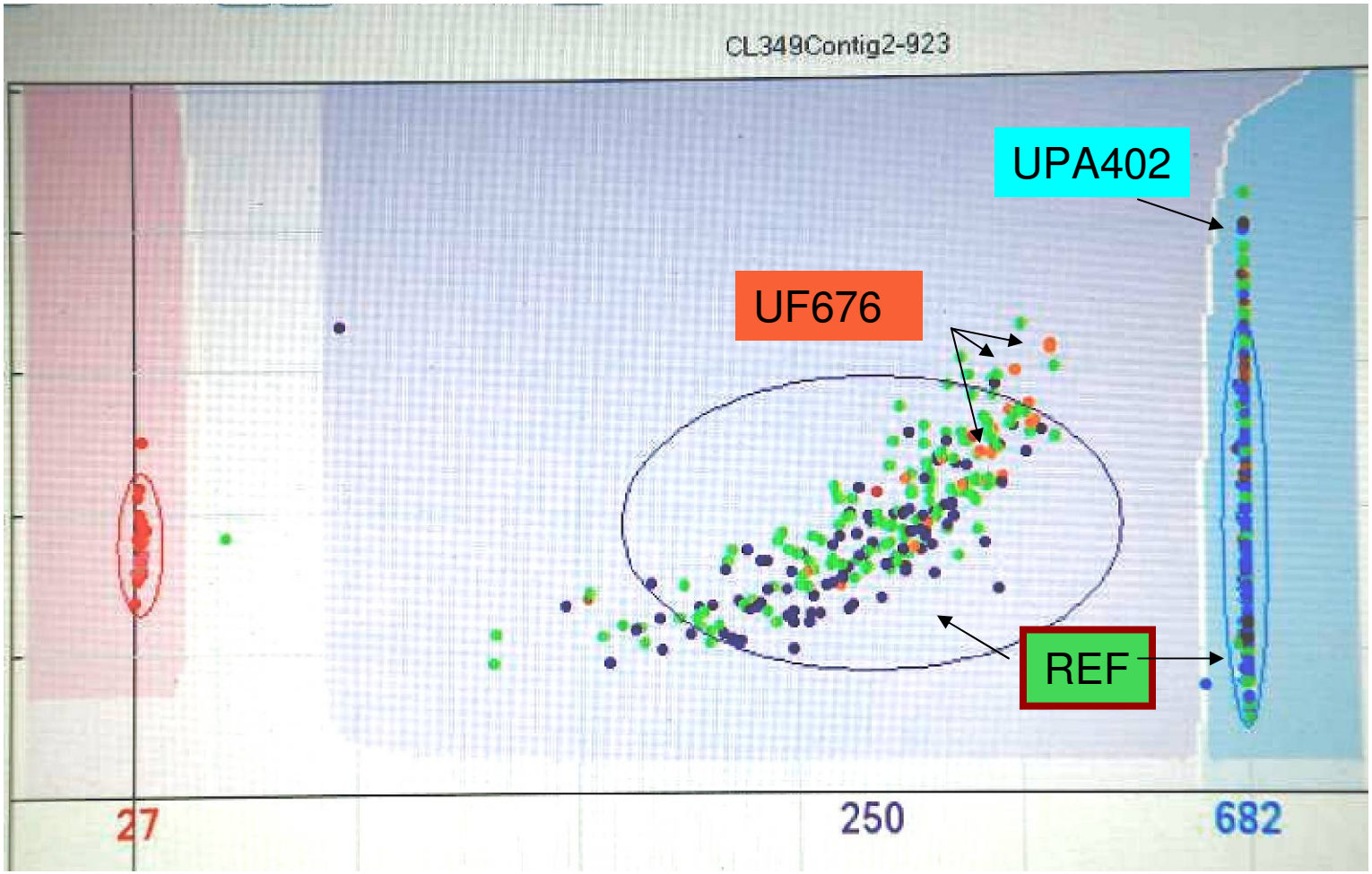


2 génotypes homozygotes

- Catongo
- Criollo

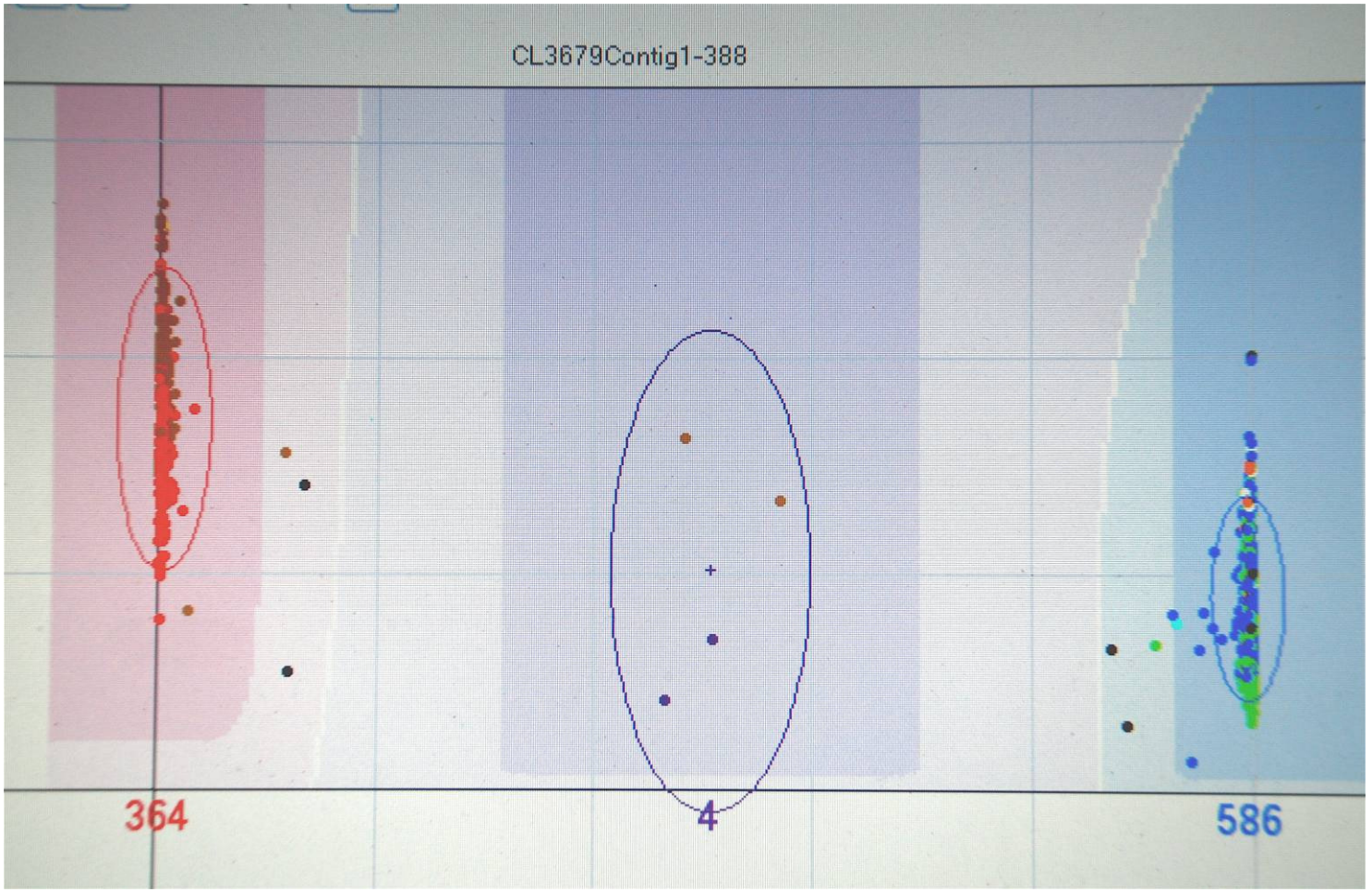
- Pop REF: UPA402 x UF676
- UPA402
- UF676
- Population F2: (SCA6 x ICS1)
- F1: Sca6 x ICS1

Exemples de nuages bien positionnés



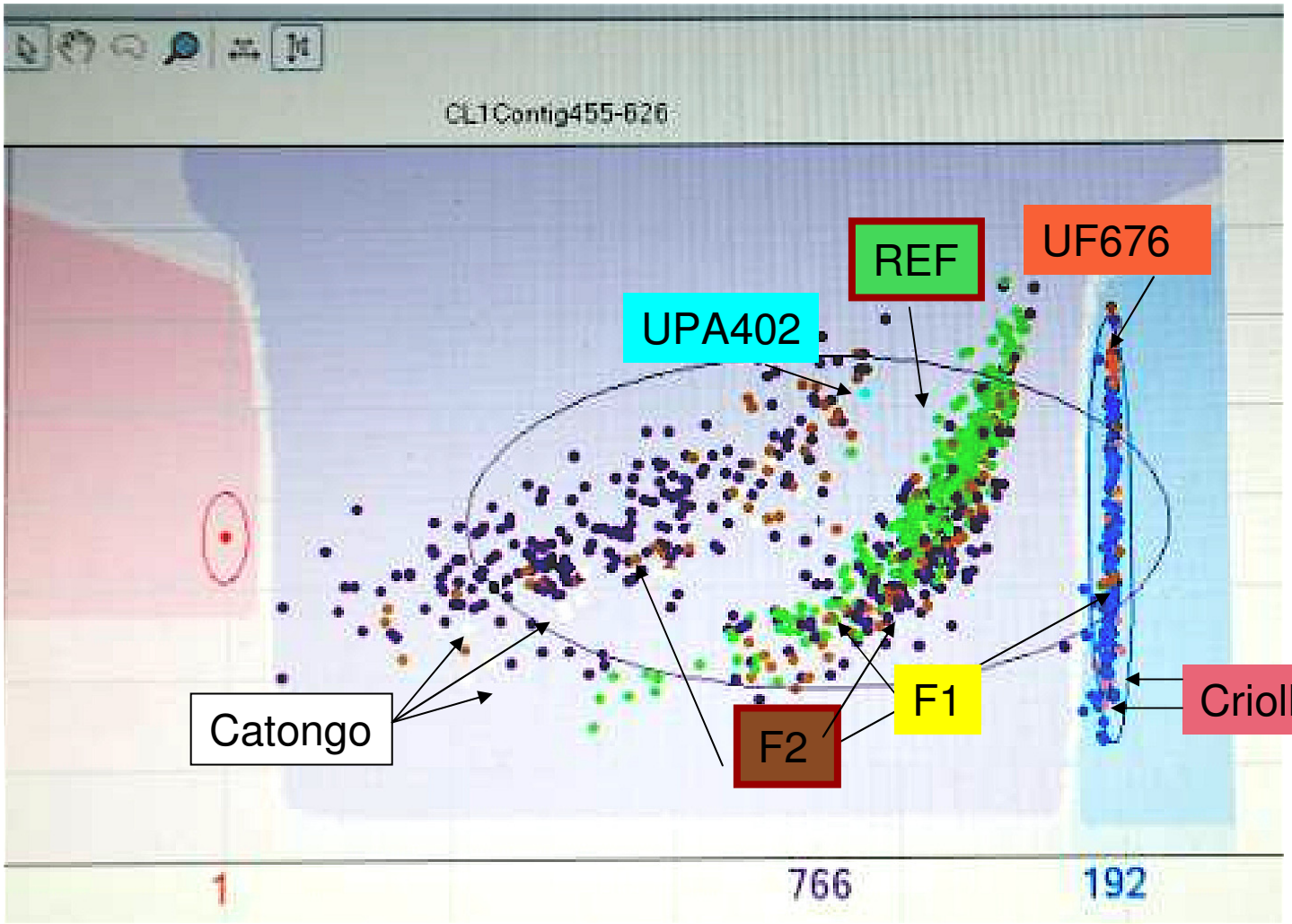
- Population REF: UPA402 x UF676
- UPA402
- UF676

Exemples de nuages bien positionnés



- 2 nuages de points
- gène mitochondrial

Exemples de nuages mal positionnés

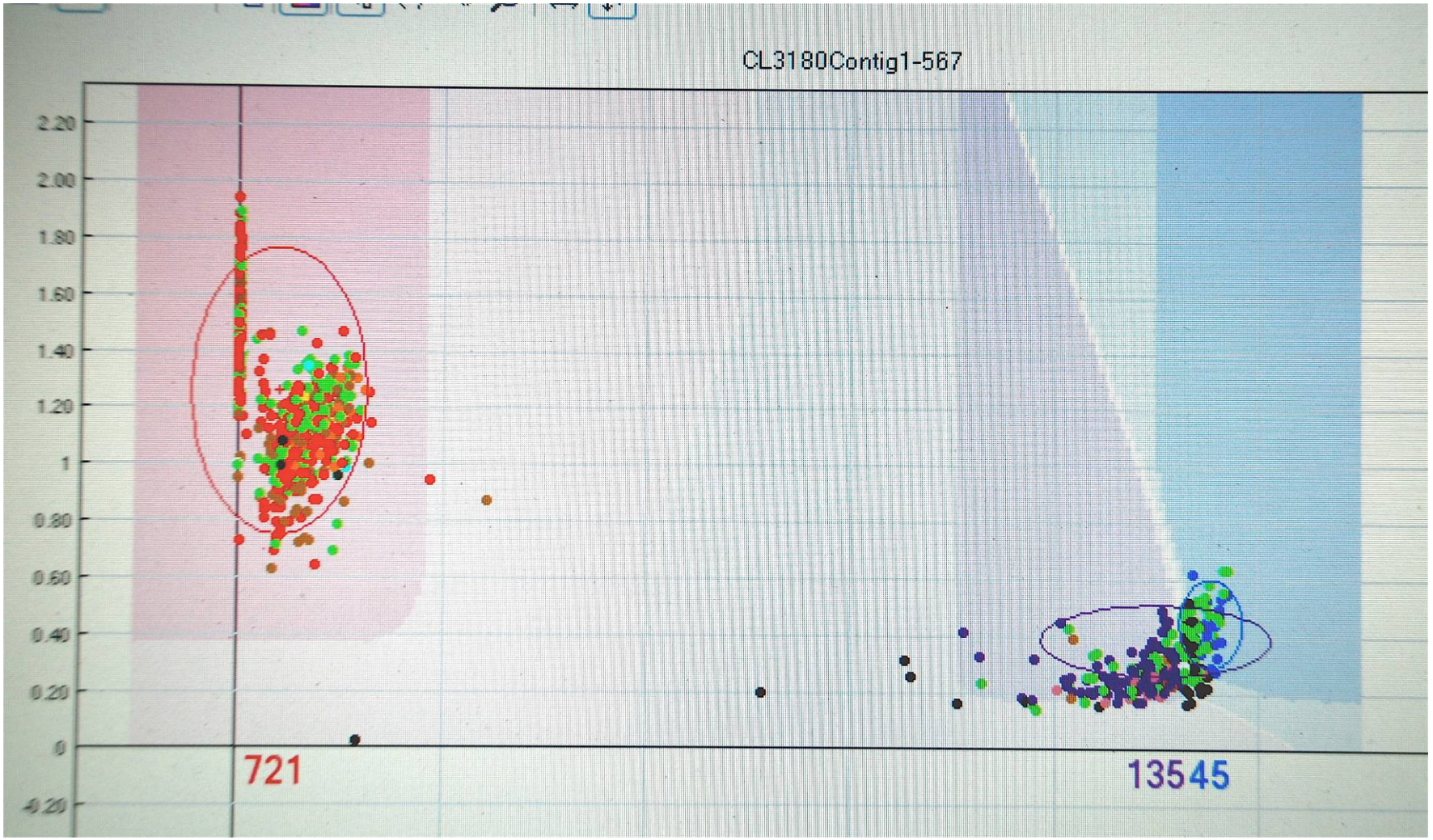


2 génotypes homozygotes

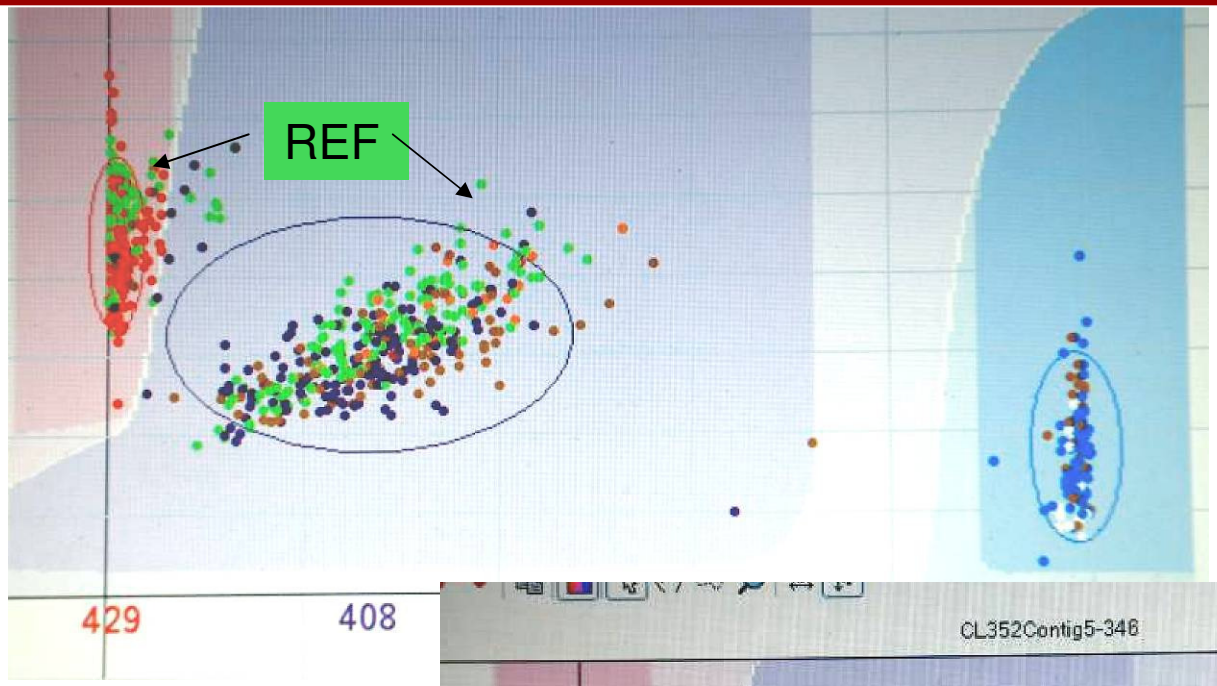
- Catongo
- Criollo

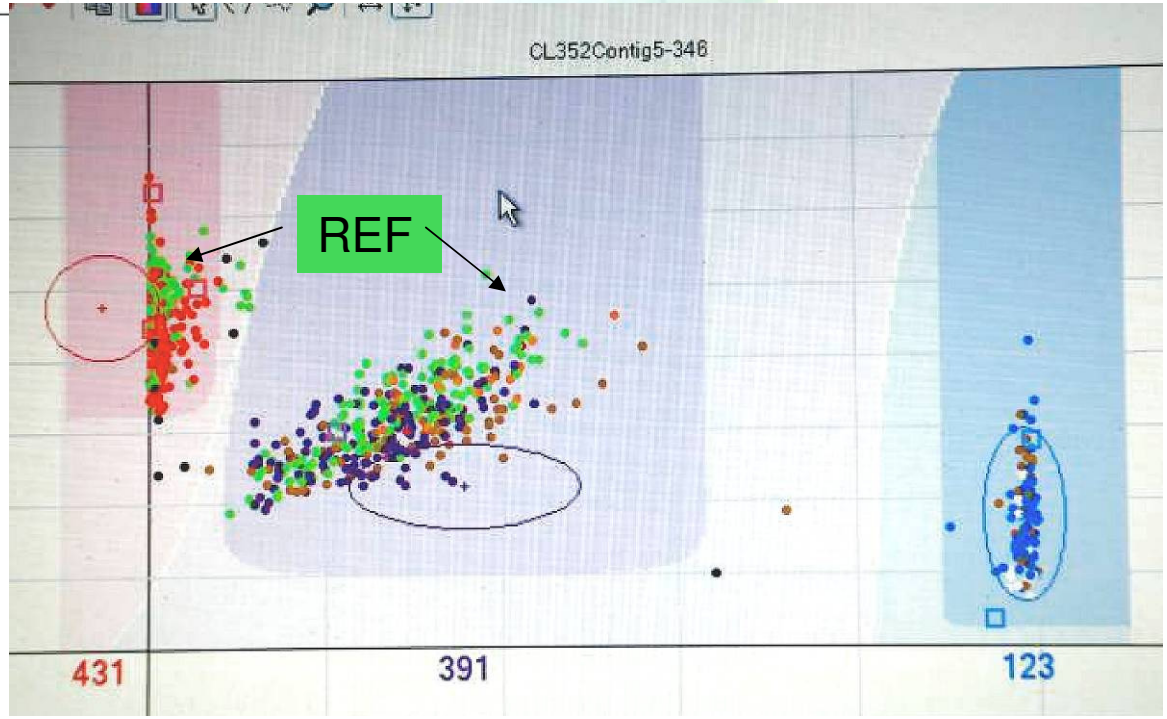
Mal positionné mais récupérable en partie

Exemples de nuages mal positionnés

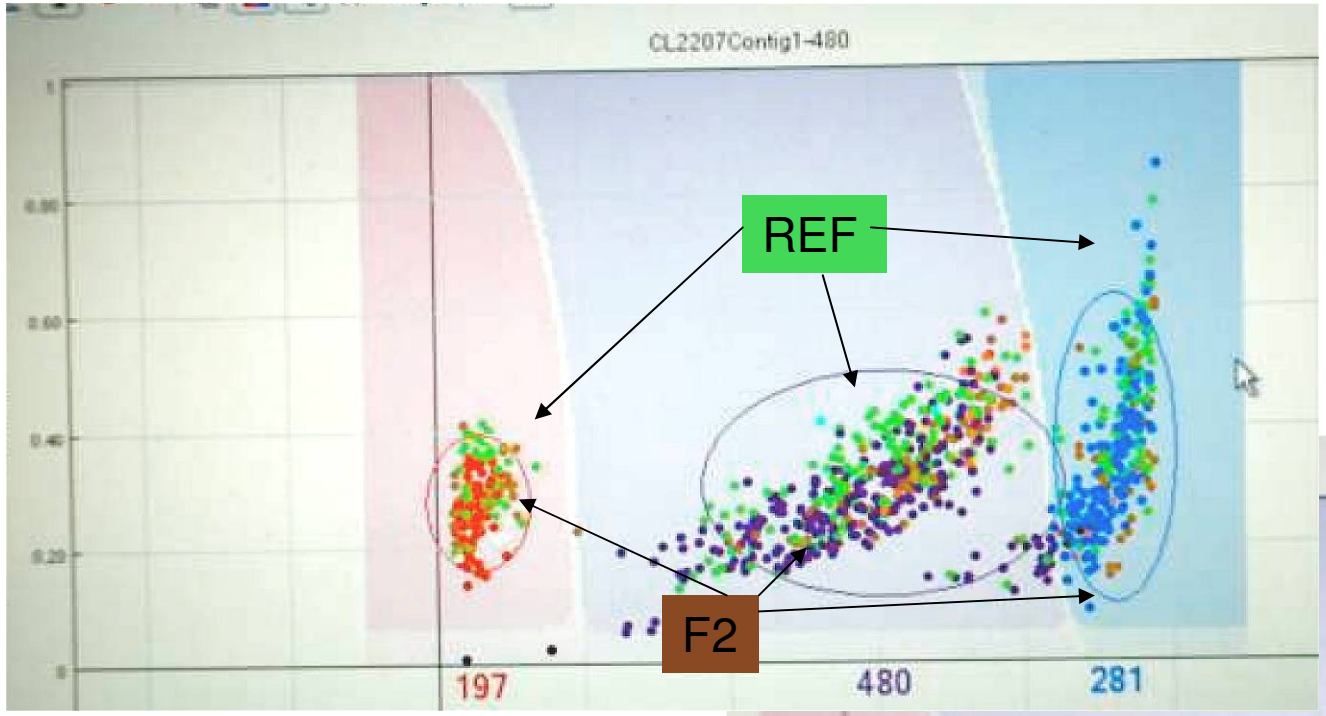


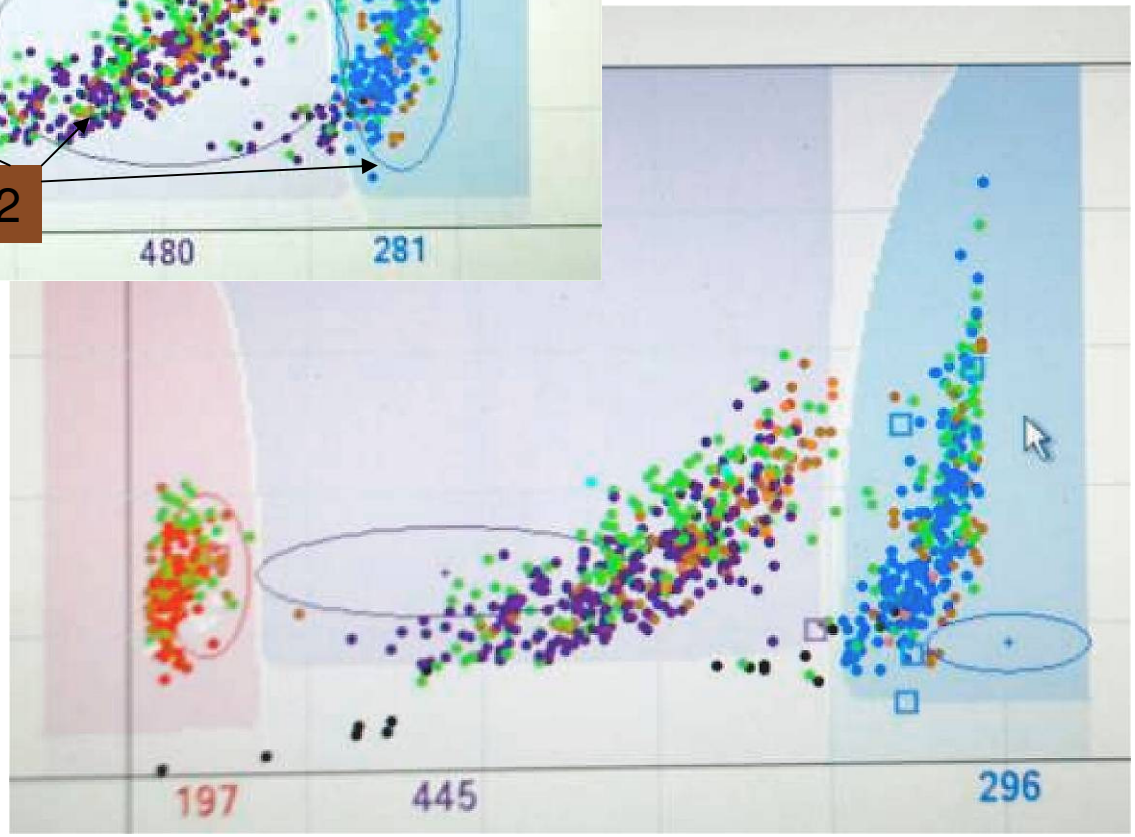
Exemple de repositionnement



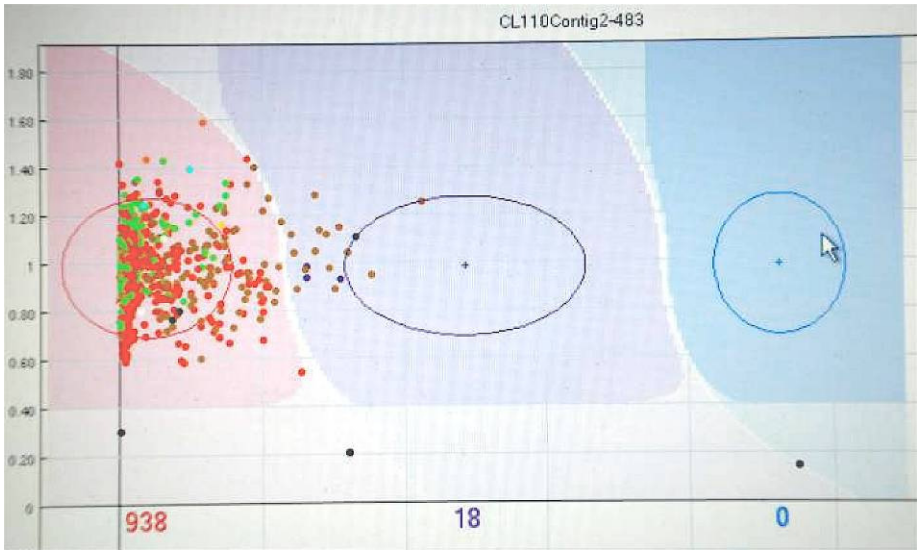


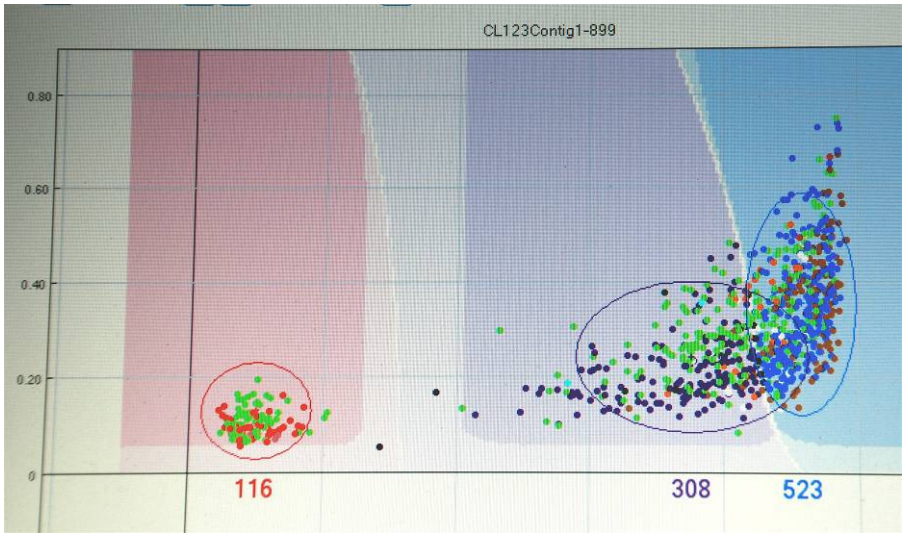
Exemple de repositionnement

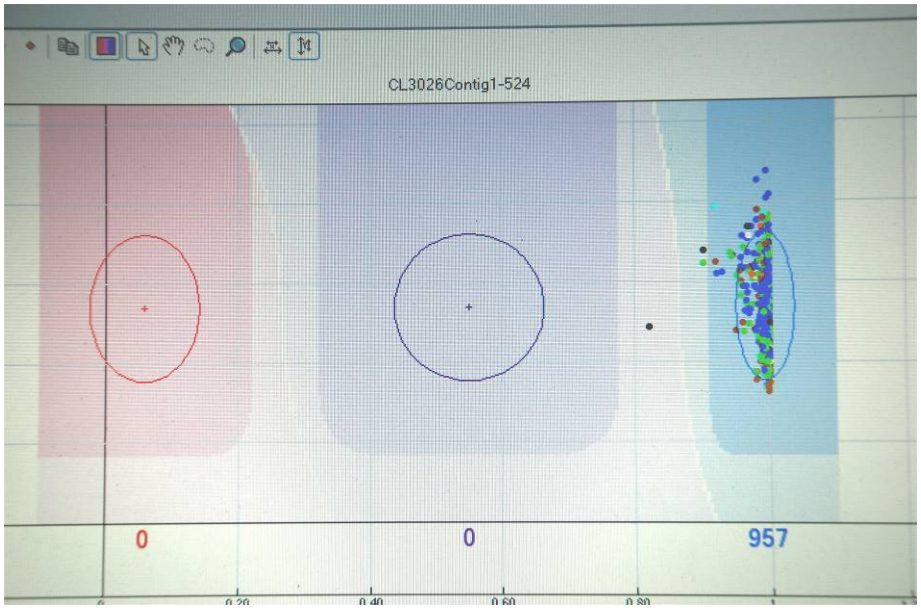


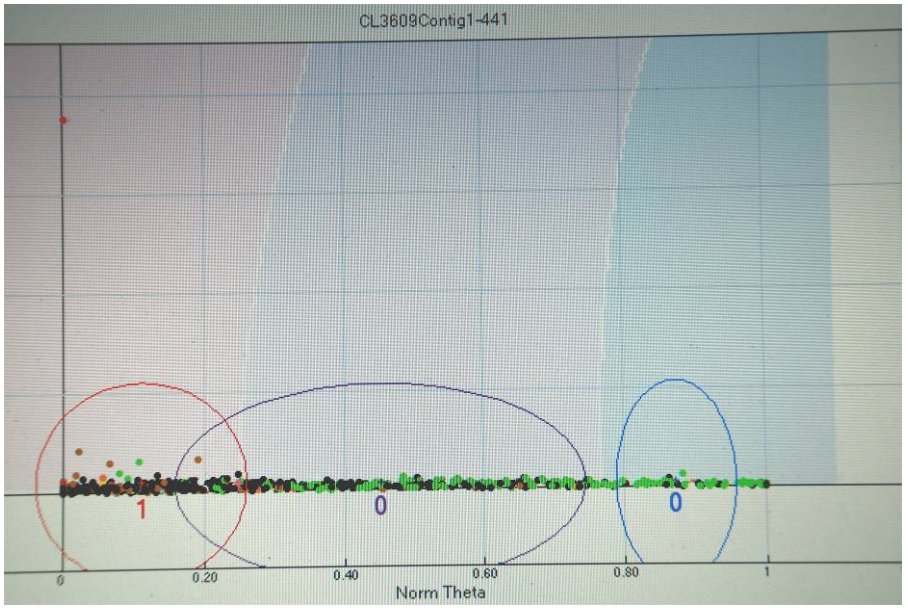


Exemples de nuages mal positionnés et/ou inexploitables









Bilan

- Sur les 1536 marqueurs (génotypage de 900 individus):

- **842 sont polymorphes (55%) et 737 cartographiables**

- **435 apparaissent monomorphes, soit 28%**

- étonnant vu les critères adoptés pour identifier les SNP

- pourraient correspondre à une mauvaise séparation des nuages ou mauvaise amplification d'un des 2 allèles, ou à des gènes dupliqués

- **259 ne sont pas exploitables (en bas ou nuages mal séparés) soit 17%**

- **Données manquantes**:

- obligation d'être très stringent dans la position des nuages pour ne pas générer d'erreurs de génotypage
- élimination de zones entières de certains nuages, d'où **un certain nombre de données manquantes**.
- exemple : 6,5% de données manquantes pour la population de cartographie de référence.

Conclusion

- beaucoup d'erreurs générées par le logiciel Illumina dans le positionnement des nuages de points

- nécessité d'inclure beaucoup de témoins dans l'analyse pour le repositionnement correct des nuages de points homozygotes/hétérozygotes: témoins homozygotes et hétérozygotes; la disponibilité de populations de cartographie en ségrégation est l'idéal pour mieux définir les nuages de points.

- marqueurs uniquement utilisables après un tri sévère des marqueurs à retenir, à faire manuellement, marqueur par marqueur

- Production de données robustes qui permettent un grand nombre d'analyses génétiques

Partenaires du projet

CIRAD

M. Allegre
X. Argout
D. Clément
O. Fouet
C. Lanaud
T. Legavre
AM. Risterucci
X. Sabau
JM. Thevenin

CEPLAC — K. Gramacho (Brazil)

CNRA — M. Tahi (Côte d'Ivoire)

THE UNIVERSITY OF THE WEST INDIES
AT ST. AUGUSTINE, TRINIDAD AND TOBAGO

M. Boccara (Trinidad)
CRU/CIRAD

INRA cea

D. Brunel
A. Bérard
A. Bolland
M. Foglio

USDA UNITED STATES DEPARTMENT OF AGRICULTURE

D. Zang (USA)
L. Motilal

GENOSCOPE
Centre National de Séquençage

P. Wincker
C. Da Silva

Avec le soutien financier de

USDA UNITED STATES DEPARTMENT OF AGRICULTURE

la Région Languedoc Roussillon

Merci de votre attention