	Optimisation d'un fichier de clustering	Version 1.1
GDB_MOP_12	Génotypage	25/08/2023
Rédaction : L. LIETAR	Vérification : C. AUDEBERT	Approbation : K. LEROUX

Ce mode opératoire s'adresse à toute personne habilitée à réaliser les opérations d'analyse / transmission de résultats de génotypage


• Définitions

- GC = Gen Call score = score de qualité qui traduit la distance de l'échantillon du centre du cluster, et augmente au fur et à mesure de sa proximité du centre du cluster (de 0 à 1).
- no-call threshold = Gencall score cutoff = seuil du GenCall score en dessous duquel les génotypes ne sont pas valides.
- P10GC = valeur de Gen Call d'un échantillon pour laquelle 90% des Gen Call de tous les SNPs sont supérieurs à cette valeur, 10% y sont inférieurs.
- Call Rate = score de qualité global d'un échantillon qui traduit le nombre de SNPs génotypés dont le Gen Call est supérieur au seuil minimal spécifié, et augmente au plus ce nombre augmente (de 0 à 1).
- Gen Train score = score de qualité du cluster pour un SNP (de 0 à 1).
- Call Freq = proportion d'échantillons à chaque locus (de 0 à 1).
- Het Excess = excès d'hétérozygotie (de -1 = déficience d'hétérozygotes, à 1 = 100% hétérozygotes).
- Cluster Sep = écart entre les trois génotypes via norm θ (de 0 à 1).
- AB R mean = intensité des génotypes AB (norm R).
- AB T mean = décalage du cluster AB vers AA ou BB via norm θ (de 0 = proche de AA, à 1 = proche de BB).
- Minor Freq = fréquence de l'allèle mineur.
- P-C Errors ou P-P-C Errors = nombre d'erreurs d'héritabilité parent-enfant ou parent-parent-enfant pour un SNP.
- Rep Errors = nombre d'erreurs de reproductibilité entre deux réplicas pour un SNP.

Cette étape fait suite à la mise en œuvre de l'opération de génotypage, suite à la phase de scan des puces à ADN.


Mode opératoire

- 1) Partir de 384 génotypages minimum, et les importer dans Genome Studio en utilisant le fichier cluster standard d'Illumina correspondant au type de lame utilisé (no-call threshold = 0,15). Cliquer sur la calculatrice dans la partie samples table pour recalculer les statistiques.
- 2) Effectuer un scatter plot (nuage de points), en cliquant sur le bouton correspondant dans le Samples Table : $P10GC=f(\text{Call Rate})$, et exclure les échantillons qui possèdent un P10GC et un Call Rate faibles (outliers), afin de

	Optimisation d'un fichier de clustering	Version 1.1
GDB_MOP_12	Génotypage	25/08/2023
Rédaction : L. LIETAR	Vérification : C. AUDEBERT	Approbation : K. LEROUX

définir la population de départ pour le clustering. Par faible sont entendues des valeurs de $P10GC < 0,69$ et $CallRate < 0,98$.

- 3) Dans le tableau SNP Table, classer les données Gen Train score et sélectionner les SNPs dont le Gen Train score est inférieur à 0,75. Faire un clic droit sur la sélection et cliquer sur Cluster selected SNP. Un message apparaît ensuite demandant de mettre à jour les statistiques SNP, cliquer sur ok. Les SNPs vont être reclusterisés automatiquement en fonction des génotypes importés.
- 4) Dans le même tableau, classer les données Call Freq par ordre croissant et optimiser les clusters de manière à ce qu'ils comprennent un maximum d'individus valides.
Mettre à zéro les clusters douteux en effectuant un clic droit sur le SNP puis en cliquant sur Zero selected SNP.
- 5) Classer ensuite les données Het Excess par ordre croissant et observer les génotypes inférieurs à -0,3 et ceux supérieurs à 0,2.
Mettre à zéro les clusters douteux en effectuant un clic droit sur le SNP puis en cliquant sur Zero selected SNP.
- 6) Cliquer sur la calculatrice dans la partie samples table pour recalculer les statistiques. Certains échantillons, à l'origine outliers, ne le sont peut être plus du fait de l'optimisation des clusters. Pour le vérifier, ré-inclure les outliers et recalculer les statistiques. Les exclure de nouveau si le Call Rate n'est pas satisfaisant.
- 7) Exporter le cluster en cliquant sur file > export cluster positions > for all SNPs, dans le dossier partagé \genotypages\Fichiers_Clustering (serveur : gna2gdlabo.genesdiffusion.com), sous le nom "version de puce"_"technologie"_"espèce"_jjmmaa (ex : MDv2_XT_bovin_150121). Archiver l'ancien cluster dans le dossier \genotypages\Fichiers_Clustering_Archives.
- 8) Pour les résultats suivants qui viendraient s'ajouter au projet initial pour aboutir à un projet final complet, à partir de 1000 génotypages les clusters peuvent être de nouveau optimisés de manière plus complète :
 - effectuer un scatter plot (voir étape 2)),
 - exclure les outliers,
 - classer les données Cluster Sep par ordre croissant et optimiser les clusters dont le Cluster Sep est inférieur à 0,4,
 - classer les données Call Freq afin d'optimiser les clusters (voir étape 4)),

	Optimisation d'un fichier de clustering	Version 1.1
GDB_MOP_12	Génotypage	25/08/2023
Rédaction : L. LIETAR	Vérification : C. AUDEBERT	Approbation : K. LEROUX

- classer les données AB R mean par ordre croissant afin d'optimiser les clusters et mettre à zéro les clusters trop faibles en effectuant un clic droit sur le SNP puis en cliquant sur Zero selected SNP,
- classer les données AB T mean par ordre croissant afin d'optimiser les clusters dont le AB T mean est inférieur à 0,2 ou supérieur à 0,8,
- classer les données Het Excess par ordre croissant et observer les génotypes inférieurs à -0,3 et ceux supérieurs à 0,2 (voir étape 5)),
- classer les données Minor Freq par ordre croissant et vérifier les génotypes inférieurs à 0,1,
- classer les données Chr de manière à pouvoir observer le chromosome X, sélectionner les mâles dans le samples table, qui apparaissent en jaune dans le SNP graph. Vérifier l'absence d'hétérozygotie chez les mâles pour le chromosome X,
- si des individus partagent du pedigree, observer les données P-C Errors et P-P-C Errors afin de détecter d'éventuelles duplications/délétions et autres aberrations chromosomiques,
- si des projets incluent des répliques, observer les données Rep Errors afin d'évaluer les SNPs avec une ou plusieurs erreurs et de faire un comparatif entre répliques,
- cliquer sur la calculatrice dans la partie samples table pour recalculer les statistiques. Certains échantillons, à l'origine outliers, ne le sont peut être plus du fait de l'optimisation des clusters. Pour le vérifier, ré-inclure les outliers et recalculer les statistiques. Les exclure de nouveau si le Call Rate n'est pas satisfaisant,
- exporter le cluster en cliquant sur file > export cluster positions > for all SNPs, dans le dossier partagé \genotypages\Fichiers_Clustering (serveur : gna2gdlabo.genesdiffusion.com), sous le nom "version de puce"_"technologie"_"espèce"_jjmmaa (ex : MDv2_XT_bovin_150121). Archiver l'ancien cluster dans le dossier \genotypages\Fichiers_Clustering_Archives.

Documents associés :

GDB_FORM_05_Habilitation analyse/transmission résultats génotypage