Analysis of functional and comparative genomics data

Q2 2016

03/02/2016

1 Overview of functional genomics

Evaluation = projet

Dogme central de la biologie = les cellules contiennent le même ADN mais ont des fonctions différentes. Ceci est possible car les cellules n'ont pas le même ARN. C'est un flux à sens unique de l'ADN vers les protéines

Génomique fonctionnelle = déclinaison du dogme

Génome (commun à toutes les cellules) → transcriptome (spécifique à la cellule) → protéome (spécifique à la cellule)

Le cours est axé essentiellement sur le transcriptome. Les protéines sont assez bien représentées par le transcriptome "omic" fourni des infos sur le génome. Plusieurs méthodes microarray ou séquençage cf slides

microarray suppose qu'on connaît le gène → biaisé. On n'a pas ce soucis avec le séquençage

Révolution de Louis Pasteur possible grâce à la découverte/invention du microscope.

03/02/2016

2 Preprocessing of dual-channel microarray Data (1995)

Sl1 : Pipeline d'analyse → commence in vitro et fini in silico

Certains nombre étapes :

— marquage
— hybridation
— scan des lames
— quantification des spots
— prétraitements
— analyse données

Sl2 : Marquage : cellules cancéreuses et cellules de contrôle (tissus sains). rétrotranscription → choisi couleurs ≠

Sl3 : hybridation : labeled cDNA . lame exposée au cDNA et voit quels gènes sont exprimés (1 spot correspondant à 1 gène)

Sl4 : scan : introduit dans scanner capte onde laser utilisés → image avec points lumineux allant de rouge au vert en passant par jaune

Sl5 : quantification : quantifie le nombre de spot, programme de reconnaissance d'image (fourni avec scanner). 2 matrices (rouge et vert) on a aussi bruit de fond pour rouge et vert → 4 matrices pour quantification des gènes

Sl6 : prétraitement : enlever bruit de fond ...

Sl7 : analyse créative commence à l'analyse

Sl8 : correction bruit de fond : même en l'absence d'expression, un peu fluorescent → on mesure quelque chose

Sl9 : spots 1 e génération fluorescence des spots moins forte bruit de fond → intensité négative → impossible

Sl10 : hypothèses :

— localisation physique du spot n'influe pas sur l'expression du gène
— niveau régulation est constant dans la tumeur ou dans la cellule de contrôle → informatif = ratio rouge-vert et comparaison

Sl11 : MAplot : chaque point = un gène, abscisse = ratio rouge-vert. Ligne rouge = régression non linéaire. Expression non centrée sur 0 → plus de fluorescence sur une couleur que l'autre. Si il est plus exprimé, plus de chance d'être dans le rouge que dans le vert → artefact ⇒ normalisation (on garde distance entre lignes rouge et les points et on redresse la ligne rouge)

Sl12 : Myc = facteur de transcription → pas de cible spécifique. Amplifie la transcription.

Sl13 : correction de biais print-tips : Mécanique → sujet à l'usure et la quantité de liquide déposé non contrôlé. Les têtes d'impression ont des propriétés ≠. Les boites ont des étendues ≠ → normalisation uniformise

Sl14 : fluorofore s'incorpore ≠ → on ne peut pas corriger par le calcul → on répète la manip en échangeant les fluorescences. On fait la moyenne en inversant le signe d'une des 2 manip

Sl15 : filtre → enlève outlire

Sl16 : approche normalisation s'applique à grandes séries de microarray. On peut choisir de normaliser la variance pour l'uniformiser et les rendre plus comparables

Sl17 : pas de consensus sur comment faire en pratique. Pas vraiment de données références.

Sl18 : analyse ne doit pas dépendre du bricolage de la normalisation → si oui pb. Bonne pratique = vérifier au max sur des données publiées → utilité des données free access. Base de données : Geo et array expressed #Package Marray sur R, emacs

exemple : swirl microarray poissons (refont la manip plusieurs fois). Cf code R

3 Preprocessing of affymetrix microarray data

Sl1 : technologie brevetée.

Sl2 : double canaux = imprimé avec pointe ou jet d'encre. Affymetrix différent. S'inspire de la fabrication des microprocesseurs (impression par couche avec masque → Photolithographie). Attache des p/b d'ADN puis amorce en mettant masque puis réaction qui ajoute nucléotide. Av : haute densité oligo et versatilité énorme (masque généré de manière informatique). Mais oligos très petit : max 25 p/b ce qui est très peu par rapport génome humain et sa répétitivité → pb de spécificité

Sl5 : probset = oligo du gène qu'on spot sur microarray. 2 versions : wild type et contrôle (mutation du nucléotide central). Séquence avec mutation va moins bien s'hybrider. Manque de spécificité compensée par redondance. 25 oligo de 25 p/b *2

Sl6 : comment exprimer àpd de 50 chiffres ?

Sl7 : à généré beaucoup de littérature. Développement logiciel normalisation libre. 4 catégories : traitement un par un (MAS5), multi-array, propriétés des séquences, données publiques dans base de données

Sl8 : mas5 assez mauvaise méthode. Multi-array et spécialement RMA. Abscisse expression réelle et ordonnée fluorescence des sondes. $\neq$ assez énorme entre les sondes mais $\neq$ constantes. On peut mesurer le niveau de base. Correspond au niveau d'affinité.

Sl9 : microarray sur lequel on hybride toujours le même ARN et 16 gènes pour lesquels on change la concentration. Compétition d'algorithme. 2 critères : accuracy et precision. Compromis entre accuracy et precision. MAS5 difficile de voir signal de bruit de fond. Les méthodes les plus précises sont multi-array.

Sl10 : même graphique mais sans MAS5. RMA et GCRMA bonnes méthodes (bon compromis)

Sl11 : on a plus qu'un seul canal. PCR fold change de 2 → microarray plus sensible. Donne un seul microarray et analyse avec universel

Sl12 : biais spécifiques de l'étude. Géré avec RMA mais pas avec fRMA

Sl13 : overfitting : il est possible que les méthodes soient adaptées pour le dataset et améliorations non adaptées pour des data réelles.

Sl14 : résultats solides ne doit pas dépendre de la méthode qu'on utilise.

03/02/2016

4 Detection of regulated genes in microarray data

Remind : Dogme central = flow à sens unique de l'ADN vers l'ARN puis les protéines. A chaque étape réalisation dogme on a des technologies (microarray, chip . . .). Révolution (3 choses) : on peut voir des choses qu'on ne pouvait pas avant, . . . , information de base est stockable et échangeable (grâce à nature informatique des données)

Sl1 : données d'un groupe en Pologne → comparer cancer papillon de la tyroïde (tumeur sur une aile et prélève aussi autre aile pour avoir tissu sain)

Sl2 : étendre l'idée de la PCR (généralisation)

Sl3 : liste des tops gènes

Sl4 : gènes associés au remodelage de la matrice cellulaire. Cancer doit réaménager pour qu'il puisse se déplacer

Sl5 : dédifférenciation de la fonction thyroïdienne. Les gènes sont en général sous régulé

Sl6 : nuak régulateur négatif

Sl7 : associé à d'autres types de cancer

Sl8 : illusion stat de 2 ordres (on trouve des gènes régulé et on tape dans pubmed pour le cancer et on trouve toujours des articles)

Sl9 : qualité de mesure bonne → ce qui n'est pas juste c'est la méthode stats. Contrôles expérimentaux ne servent à rien → on doit faire des contrôles stats

Sl10 : test multiple

Sl11 : signification stats → pb que ce qu'on observe soit du au hasard et rien qu'au hasard (p-value). Pas d'équivalence entre signifiant stats et biol

Sl13 : test statistique : but = écarter le hasard comme explication triviale des observations. Etabli $\neq$ entre plusieurs catégories

Sl15 : t-test moyenne divisée par déviation standard (pas bon pour biologie → déviation standard petite). Variante : t-test modéré

Sl16 : distribution nulle = hypothétique. Estimer en appliquant random sur les données. On peut toujours calculer la distribution nulle

Sl17 : bonferroni suppose que les données sont indépendantes (un peu stricte et trop conservative) FDR le plus utilisé

Sl21 : pas distribution normale

Sl23 : boite = échantillon. On fait plusieurs permutations

Sl24 : pas de gènes qui défini taille de la tumeur (pas significativement corrélé)

Sl25 : problème des faux positifs (erreur de type 1).

exemple de code : Summary(z) → nombre gène differentiellement exprimé

Au dessus ou en dessous ligne Sam plot = différentiellement exprimé

Avec les données random → aligné

10/02/2106

5 Are most published research findings false ? !

Sl8 : cf calcul papier

Sl13 : plus on augmente le biais moins mon étude est informative.

Sl14 : losange = valeurs synthétisée. Pour les non publiées facteur de risque n'existe pas, non indexé existe mais pas fort, indexé fort présent

Sl15 : forest plot.

Sl16 : 1 personne travaillant seule et rapportant un résultat positif à plus de poids qu'une personne travaillant avec 50 personnes

Sl18 : problème de reproductibilité

10/02/2016

6 Evaluation of classifiers quality

Sl2 : 2 tâches : training puis test (ou validation)

Sl3 : 2 grandes familles de classifier (linéaire et non linéraire). Linéaire trace droite qui sépare les 2 classes.

Sl4 : non linéaire : sépare avec un courbe. Possibilité de classification supérieure

Sl5 : overfitting = sur adaptation aux données. Se produit quand grand nombre de variable et petit nombre échantillon (typique microarray) → curse of dimensionality

Sl9 : seuil de détection

Sl10 : AUC (Area Under the Curve)

Sl11 : données test et validation doivent être totalement séparées.

Sl12 : validation croisée → évalue classifier (quand peu de données)

24/02/2016

Chaque étape du génome jusqu'à sa manifestation phénotypique (ARN, protéine,...) peut être étudiée grâce à la génomique fonctionnelle.

En général, on compare des conditions. La normalisation va partir du fait que la quantité de gènes est la même pour les deux.

Plus on cherche, plus on a de chances de trouver quelque chose ⇒ BIAIS

Distribution nulle = contrôle négatif de statistique ⇒ mesurer quand il n'y a pas d'effet

Définir une procédure pour corriger les p-values (probabilité d'un faux positif ⇒ false discovery rate, q-value ; bonferroni ⇒ brutal)

Validation des classifiers (suite)

On veut déduire d'un groupe de gènes classifier un profil.

Le problème fondamental : sur-ajustement et dimensionnalité

Très important d'ajuster les paramètres d'un classifier et de séparer les échantillons utilisés pour l'apprentissage et ceux utilisés pour valider le classifier.

Sensibilité : découvrir un cancer quand on présente un cancer

Spécificité : ne pas appeler n'importe quel tissu un cancer si ce n'est pas un cancer

On devrait intégrer la sélection des gènes dans le machine learning, comme une partie intégrante de l'apprentissage, au lieu d'introduire un biais en choisissant nous-mêmes les gènes à étudier.

Sl24 : Un algorithme de classification a lui-même des paramètres. Typiquement, les gens vont tenter pleins de paramètres différents et reporter le meilleur ⇒ PAS BIEN !

Sl25 : On partitionne le set pour faire des cross-validations successives (principe jack-knife)

Sl26 : est-ce qu'on est capable de prévoir si on a des métastases dans les ganglions afférents selon le cancer de la thyroïde primaire ? On construit des classifiers sur les données et quelle est la probabilité d'obtenir une erreur de 60% en utilisant des données permutées ? Une erreur de 60% n'est pas significative, parce que la chance de l'obtenir par hasard est très grande.

À droite : plus le dataset est petit, plus la variance d'estimation d'erreur est grande ⇒ plus on a de probabilité que quelqu'un qui rapporte une haute précision de son classifier pour ce cancer a des chances de se tromper et d'avoir quelque chose de non significatif

Sl27 : CONTRÔLE NEGATIF !! Qu'est-ce qu'on mesure quand il n'y a rien à mesurer ? Impliquer cette mesure et voir ce qu'il mesure

24/02/2016

7 Dimension reduction of gene expression profiles

Comment calculer et comparer ces phenotypes globaux ? on parle de tous les gènes simultanément.

7.1 Hierarchical clustering

On a l'expression de chaque gène dans chaque tumeur.

La matrice avant prétraitement ⇒ tapis aux différentes teintes, pas informatif. Tout l'art est d'analyser les données pour donner un sens à cette matrice d'expression . L'apprentissage non-supervisé ⇒ dit à la machine de trouver toute celle qqch

Supervisé ⇒ on connait la structure et on demande à l'algorithme de trouver des traits à cette structure

Claster hiérarchique : non supervisé

Un exemple : ensemble de données avec des cancers thyroïdiens avec des pathologies tumorales (adénomes autonomes A/V). De lui-même, ils les agrègent par similarité et les sépare selon ces deux groupes.

On réorganise alors les colonnes et les lignes de la matrice ⇒ heatmap montre les groupes de gènes sur/sous-exprimés dans les groupes de mêmes profils ⇒ caractéristique/propriété globale des échantillons

Comment ça marche ? On a un groupe d'échantillons dans un espace à deux dimensions. On va du simple vers le global. On commence par apparier les échantillons les plus proches (calcul des distance entre les pairs d'échantillons), puis on recommence en regroupant à chaque itération les éléments les plus proches.

1. Qu'est-ce que ça veut dire que deux échantillons sont proches ? Dans un espace à 20.000 dimensions ? = distance
2. Quand on agglomère un cluster, un échantillon,... comment on définit la dsimilarité de ces deux clusters ? = linkage

La distance : euclidienne (mais sensible à l'amplitude des variations, par exemple les gènes de structure sont plus exprimés que d'autres), la corrélation (similarité sur une échelle invariante) est préférée

Si on connait x, on connait y (par exemple les niveaux d'expression) ⇒ le nuage de point correspond à une corrélation de 1

Ou alors on peut avoir une corrélation parfaite, mais inversée. C'est également invariant d'échelle (malgré les échelles différentes, cela ne change rien à la corrélation)

L'absence de corrélation ne signifie pas l'absence de relation. C'est linéaire, donc ça détecte une relation représentée par une droite. Ici, on voit bien une relation sinusoïdale, mais il n'y a pas de corrélation.

Pour les linkages : single (on prend les éléments les plus proches et on les relies, peu utilisé), complete (distance maximale entre les éléments les plus éloignés des clusters, peu employé), average (moyenne des distance entre tous les élements de chaque clusters, le plus utilisé), ward (cherche à minimiser la variance à l'intérieur de chaque cluster, utilisé par le prof, donne des clusters de taille équilibrée et compacts)

3 applications du clustering

1. Profiler des levures prises dans différents états, puis clustering hiérarchique sur les résultats. Le profil d'expression global est présenté à gauche. On observe des groupes de centaines de gènes co-régulés. Première fois qu'on a accès à l'expression globale ⇒ découverte des vagues de co-régulation des gènes. Matrice hautement structurée ⇒ effet cohérent sur l'ensemble des gènes selon l'état de la levure
2. Fibroblastes humains, un groupe ne faisait rien et le second les exposait au sérum. On regarde l'expression globale et il notait ce qui était régulé. Les cellules exposées au sérum proliféraient beaucoup plus. Les gènes plus régulés étaient associés à la prolifération, mais également des gènes liés à la réponse immunitaire, à l'angiogenèse, de la réorganisaition du cytosquelette,... Des gènes qu'on observe impliqués dans la cicatrisation. Le sérum induisait l'expression de gènes auxquels on ne s'attendait pas et implique le rôle du sérum dans la cicatrisation.
3. Tumeurs du sein in-vivo profilés. On a sélectionnés les 500 gènes les plus variables dans les microarrays ⇒ gènes « intrinsèques ». La classification du cancer du sein était limité avant l'article : tumeurs exprimant le récepteur d'œstrogène et celui ne l'exprimant pas (notamment celui exprimant ERBB2 ⇒ mauvais pronostic). On a découvert dans l'article une classification encore plus complexe. À l'intérieur des ER-, ERBB2 et basales (couche basale de la tumeur exprime des protéines basales du cancer du sein, ajd appelé triple négatif). Les tumeurs de ces différentes classes ont des pronostics différents. Le profil moléculaire sans a priori a permis une classification objective des données.

!! Les clusters sont produits malgré tout !! ⇒ pas de mesure claire de la qualité d'un clustering

Il n'y a pas de degré d'appartenance à une classe (discrète) Mauvaise utilisation du clustering : pour confirmer la classification supervisée, utilisé avec des gènes choisis avant

7.2 Multidimensional scaling

On conserve les distances dans l'espace de basses dimensions. On a des sphères en trois dimensions, reliées par des ressorts. Le nuage de boule est déterminé par la tension de chaque ressort. Le scaling va écraser la boule de sphères, tout en minimisant la tension des ressorts (la distorsion des distances trouvées dans l'espace à multi-dimensions vers l'espace à deux dimensions = stress)

Cancer en noir et rouge, tumeur bénignes en bleu (nodules froid, dans la thyroïde sous –actives, sans production d'hormones) et vert (nodules chauds, produisent d'hormones thyroïdiennes en excès). Stress de 20%, la réduction de la dimension a du sens car les verts sont groupés ensemble et forme un cluster compact dans cet espace réduit.

7.3 Principal components analysis (PCA)

On conserve la même idée qu'avec la méthode précédente.

Les deux projections sont différentes, malgré que c'est le même objet. Dans la deuxième, la projection projette les points, on maximise l'étalement des points sur la surface (la variance). La PCA va trouver la projection des données, va trouver le vecteur où la projection des données a une variance maximale, recommence avec un vecteur orthogonal à la première qui va également maximise les données,... etc

Rouge cancer causé par Tchernobyl et noir cancer de patients de français. On n'a pas l'air de savoir les séparer...

Décomposition des composants en variance expliquées possible uniquement parce que les vecteurs sont orthogonaux.

Multidimensional scaling = optimisation linéaire où on cherche à minimiser la tension des ressorts

PCA = formule mathématique d'algèbre linéaire où on trouve « le meilleur à tous les coups »

La transformation consiste en une rotation combinée à une translation. ⇒ produit matriciel

Chaque visage est le résultat d'une combinaison linéaire des composants.

H = parallèles aux gènes

W = composants principaux

1. On prend différents groupes ethniques humains, est-ce que les gènes sont exprimés différemment ?
   Script R cheung
   $Pca\$dev^2/sum(pca\$dev^2)$ ⇒ donne le pourcentage de variance expliquée par la composante
   Etant donné qu'on prend deux européens pris au hasard, quelle est la probabilité qu'un chinois soit plus proche du point de vue transcriptome que les deux européens ?
   Multi Scaling ⇒ gradient européens japonais/chinois, chinois
   Matrice de corrélation ⇒ distribution de la corrélation globale (dens)
   Permet de voir les distances entre les sous-groupes.
   Effet de batch dû à la différence de date du processing des profils
2. Probing l'hétérogénéité géographique des populations humaines
   3-5% de la variation génétique est lié à l'ethnicité ⇒ PCA
   Clustering ⇒ plus on ajoute du k, des groupes spécifiques apparaissent
3. Microarrays SNP : history human population
   On est capable de dessiner un arbre de similarité génétique des individus globaux.
   Plus un groupe est arrivé tardivement sur un territoire, moins il est génétiquement varié

7.4 Non-negative factor Matrice expression

On impose une autre contrainte au principe PCA ⇒ au lieu d'avoir des vecteurs orthogonaux, on a des vecteurs qui sont positifs (signification physique, qui représentent des parties d'objets)

H = métagènes

W = coefficients dans les échantillons

Cet algorithme peut être utilisé pour la classification semi-supervisée (on choisit le metagène, puis on laisse faire)

S'il y a k clusters, on va avoir k metagènes représentatifs de ces clusters

On va construire une matrice de connectivité qui va relier les échantillons entre eux (=1 si dans le même cluster). On fait plusieurs runs et on fait la moyenne des matrices obtenues ⇒ valeur entre 0 et 1

Corrélation cophenetique = Distance dans l'espace des gènes comparé à la distance dans le dendrogramme (on veut le plus proche de 1)

02/03/2106

8 Gene set expression

Meta gène : expression combinée de plusieurs gènes

On a trouver des distinctions : gènes ou groupes de gènes → fondement biologique de ces différences ?

Gène n'ont pas une fonction univoque → pléiotropie

La présence d'un gène peut être informative mais pas l'absence (redondance de fonction pour plusieurs gènes différents. Peut-être qu'un autre joue son rôle)

Concentration sur un groupe de gène augmente statistique

Comment définir les groupes :

— ontologie des gènes : vocabulaire contrôlé par des humains (classer par fonction biochimique, par localisation ou par fonction physiologique)

— in vitro (gene set). On peut faire un lien entre des manip in vitro et in vivo.

Sl13 : méthodes pour estimer les groupes de gènes : première méthode : sélection de gènes différentiellement exprimés puis comparaison liste avec les gene set → calcul de l'intersection. Correction pour test multiple (on teste généralement des milliers de gène

Sl14 : calculer la probabilité d'avoir cette distribution avec test chi2 ou hypergéométrique

Sl15 : le nombre d'échantillon n'intervient qu'en amont. N très grand mais artificiel. Ne prend pas en compte une notion de puissance statistique pertinente. Et hypergéométrique et chi2 suggère hypothèse indépendance mais faux ici. Faire attention quand on utilise les sites (rentrer des listes de gènes au hasard)

Sl16 : méthode GSEA (Gene Set Enrichment Analysis). Veulent méthode utilisant échantillonnage des patients. Pas de treshold d'expression de gène.

Sl17 : ordonne les échantillons selon classe A et B + ordonne les gènes dans la mesure où ils sont associé aux classes (en haut surexprimé dans A). Marque les gènes associés à 1Q21.

Sl18 : calcule score d'enrichissement (ES). Parcours la liste de gauche à droite. Chaque rencontre avec une barre noire, ajout de Phit (R = somme de toutes les corrélations de tous les gènes)

Sl20 : 3 gene sets dans conditions différentes. 1e : On voit que majorité des gènes sont assoicés à la classe A. 2e : pas d'association particulière. 3e : intermédiaire, répartition quasi aléatoire. Pas mal de gène entre 0 et 10 000 et un peu moins entre 10 000 et 20 000. GSEA présente intérêt ici car détecte association faible avec A.

Sl21 : permutation aléatoire et on refait les calculs plusieurs fois pour avoir la distribution nulle. Calcul de p-value et ajustement de type q-value pour test multiple. Maintenant puissance statistique dépend nombre individu.

Sl22 : raffinement supplémentaire. Gene set significatif en fonction de sa taille. Taille optimal.

Sl23 : MSigDB : compilation de gene sets d'intérêt. Plusieurs collections de gènes. 7 grandes catégories. C2 : recouvre réactome, Kegg, ... base de données gérées par des humains en fonction des articles. C2-CGP = gene sets expérimentaux.

Sl24 : exemple cytobande : appliqué C1Trouve 2 gene sets qui sortent → lié au chromosome Y (Yp11 et Yq11). Appliqué C2 : gene set lié phénotype sexuel. Surprenant car ne vient pas de cellules sexuelles.

Sl25 : combinaison

Sl26 : scanne ALL (leucémie lymphoïde) et AML (leucémie miéloïde), prenne gene set qui sont les bandes chromosomiques.

Sl27 : facteur de transcription suractivé en cas de stress, choc, ... Si soumet choc thermique, il y a des capteurs et suractive p53.

Sl28 : font tourner GSEA C2. Dans wildtype, trouvent gène de réponse au stress, signature lié radiation, hypoxie, choc thermique. Gene set lié à RAS qui ressort.

Sl29 : p3KCA et d'autres sont retrouvés lorsque p53 est muté ⇒ cause abrogation réponse au stress et active voie p3KCA

Sl30 : pronostique cancer poumon. Prennent 2 cohortes de patients stratifiées par bon et mauvais (guéri ou revenu). Trouve gènes corrélés à ce pronostique. A peu près 10% de gènes pronostiques qui sont les même (or ils doivent étudier la même chose !)

Sl31 : crée 2 nouveaux gene sets. Bien que les 2 groupes aient un petit nombre de gène en commun, ils sont significativement bon

Sl32 : pas exactement les même gene set qui sortent de manières significative mais grand recoupement. Suggère que analyser des groupes de gènes que des gènes seuls est beaucoup mieux

Sl33 : connectivity map (ceux qui ont inventé GSEA). 164 petites molécules approuvées par le MDIA. Connecter des perturbations biochimiques avec des maladies, ou des maladies entre elles. Si on a une signature transcriptionelle, on peut voir à travers la DB son expression.

Sl35-36 : connecter des petites molécules entre elles. Prennent 13 gènes de Glaser. Voit que d'autres inhibiteurs de HDAC apparaissent. Confirmation que les perturbations consistent inhibition HDAC. Effectivement connecté 2 petites molécules. Lien fonctionnel sans apriori de ce qu'ils font.

Sl37 : estradiol = agoniste. fulvestrant = antagoniste de l'oestrogène. Voit que la query permet aussi de sortir des antagonistes. Si on ne savait pas que c'était un antagoniste, on pourrait grâce à cette technologie le trouver.

Sl38 : profil transcriptionnel utilisé pour réduire les expériences sur les animaux. Si on a un produit qu'on soupsconne d'être toxique, on compare son profil avec un qu'on connait déjà.

Sl40 : étude de la résistance au glucocorticoïde (suppresseur de la réponse immunitaire). On ne connait pas le mécanisme.

Sl41 : connectivity map sors Rapamicyn.

Sl42 : c'est suractivation de mTor qui explique la résistance aux glucocorticoïdes.

Sl45 : pas un test très puissant. Pas mal d'incertitudes sur la taille du gene set (pas vraiment traité).

NB : supprimer Sl46

Sl47 : LINCS : on étudié très grand nombre de profiles. Avec 1000 gènes ils capturent 80% de la variance d'expression.

02/03/2016

9 Gene expression predictors of breast cancer outcome

Etant donné le profil transcriptionnel voir si ça va revenir ou pas.

9.1 Analyse de survie

Sl3 : event = mort du patient. Données censurées = plus de nouvelles du patient (quand le patients quitte l'étude sans qu'on sache si il survit ou si il est vivant).

Sl8 : Hypothèse forte que peu importe le temps, la probabilité de mourir est la même.

Sl12 : hasard ratio different en fonction de l'unité de mesure qu'on choisi

9.2 Extremely brief introduction of cancer progression

Sl14 : progresse lentement. Met un certain temps

Sl15 : ce qui compte c'est le temps après l'exposition, pas l'âge de la personne

Sl15 : pour le cancer de la peau effet de l'âge à l'exposition mais pas énorme

Sl16 : progression cancer colon peut être vue car passe par adénome avant de devenir cancéreux (peu coexister)

Sl17 : gros intestin tapissé de vilosités. Au centre des vilosités = crypte contenant des cellules souches. Régénération rapide du colon.

Sl21 : Hayflick limit.

Sl23 : notion de progression évolue beaucoup avec le séquençage

9.3 Predicting breast cancer outcome with microarrays

Sl30 : traitements adjuvant = chimiothérapie, surgery, radiothérapie. Chimiothérapie → injecte terminateur de chaine. Ressemble à nucléotide. Peuvent se lié mais rien ne peut se lier.

Tamoxifène (thérapie hormonale cancer du sein)

Sl31 : les microarry ont été utilisées pour prédire la finalité du cancer (voir si le traitement hormonal permettra d'augmenter l'espérance de vie ou non)

Sl33 : 1 ligne = 1 patiente. Les patients en bas, surexprimme des gènes et ont plus de chances de métastaser donc on leur donne le traitement adjuvant. Biais dans l'étude. Chaque cancer est unique. Un même traitement ne convient pas à tous.

14/03/2016

Remind : Base de données links.

Analyse de survie pas inventée dans le domaine de la biologie, dans les entreprises.

Cancer se développe lentement et par étapes

Métastase : arrive à pénétrer dans le sang et infecte d'autres organes. Ce sont ces métastases qui tuent les patients.

4 types traitements : chirurgie, rayon = traitements locaux. Chimiothérapie (attaque les cellules qui prolifèrent vite. Donc entre autre les cellules cancéreuses mais aussi les cheveux ou les intestins → effets secondaires) et traitements hormonaux.

Sl34 : D = pas les glandes lymphatiques envahis. Avec un marqueur classique, ils sont considérées comme ayant un bon pronostique. Signature transcriptionelle (ST) permet de raffiner encore plus.

Sl35 : se posent la question d'existence de gènes spécifiques. Séries de mutations amenant un clone supplantant les autres cellules. Si mutation, dans une cellule qui se réplique après. Or, signal global dans l'ensemble du transcriptome tumoral → mutation se trouve dans l'ensemble des cellules tumorales.

Sl38 : tumeur primaire en noire et métastase en rouge. découverte de 128 gènes permettant de distinguer les cancer du poumon métastatiques des non métastatiques. Existent cancer solide (on peut le retirer par chirurgie) et cancer liquide. pour les cancers liquides, pas vraiment de notion de métastases car déjà dans le sang. Utilise cancer liquide comme contrôle négatif.

Sl39 : possèdent des tissus congelés et savent si le patient à survécu ou non. Gènes ont pouvoir de pronostiques. Arrivent à réduire les 128 gènes à seulement 17 donc plus besoin de microarray.

Sl40 : téléchargent cohortes d'autres cancers solides que celui du poumon. Prognostiques bon pour les 3 solides et non pour le cancer du sang. Signifie que gènes communs pour 4 cancers (prostate, sein, poumon et blastome). Pas une seule manip in vivo ou in vitro dans le papier. Uniquement basé sur des données réutilisées (in silico).

9.4 What biological processes are outcome predictors involved in ?

Sl41 : si on arrive à comprendre le système métastatique on pourrait trouver un moyen de contourner et empêcher ça. Pas de curiosité malsaine des chercheurs.

Sl42 : possible d'associer ST au système biologique (ex : réponse au sérum)

Sl43 : chaque ligne représente un gène. Transformation de fourrier (oscillation). Vert = sous-exprimé et rouge = sur-exprimé. Certains gènes ont expression périodique basé sur cycle cellulaire.

Sl44 : Contrôle absolu de l'expérience donc savent exactement dans quelle phase du cycle on est. Permet de voir quel gène est impliqué dans quelle phase du cycle cellulaire.

Sl45 : H de Dovrac : tumeur ressemble à cicatrice qui ne cicatrise jamais. Tous les deux implique prolifération cellulaire/fibroblaste important, invasion tumorale, fibrose, angiogenèse (irigé par sang). Différence principale est qu'il y a une résolution du processus de la cicatrice (quand elle est finie, stoppe le processus).

Sl46 : sérum = modèle in vitro pour la cicatrisation. Reprennent manip du sérum de '99. Trouvent gènes caractéristiques d'une exposition au sérum. Prennent la signature pour voir si correspond dans les cancer. Complication = associé au cycle cellulaire et donc à la prolifération. Prennent donc la précaution d'enlever les gènes associés à la prolifération. Se retrouvent alors avec le 'Core Serum' (= base de réponse au sérum - la prolifération)

Sl48 : vert = tissu normaux. Valeur diagnostique universel et pronostique.

Sl52 : H : hypoxie permet de créer environnement qui promeut le cancer. Présence cellules hypoxique permet prédire présence cancer.

4 cat principales :

— proliferation-related

— grade-related : dérivées de marqueurs cliniques connus pour mauvais diagnostique

— outcome-related

— hypothesis-drivent : dérivé apd d'H phisio-pathologiques

Sl58 : 100 aine signatures publiées mais aucune meilleure que les 70 gènes de 2002. Comparaison des signatures entre elles montrent qu'elles se trompent sur les même patients → font les mêmes erreurs. Quasi aucuns gènes en commun, H biologiques différentes mais pronostique identique.

script : sapply applique une fonction sur un vecteur. sa notation échantillon.

Surv = survival = temps de survie suivie de + quand censure événement et rien si événement

survfit trace kaplan meier

barre verticale = patiente qui quite l'étude et baisse de la courbe = une patiente qui rechutte.

survfit(s er) : essaie de définir s en fonction du récepteur d'oestrogène : au delà de 5 ans peu d'événement donc le fait que s'annule n'est pas vraiment significatif.

coxph : analyse de cox pour voir si significatif ou non. Result = non significatif (p-value = 0,552)

récupération p-value lors d'analyse de kaplan meier = summary(coxph(s er))$logtest["pvalue"]

3 grades : 1 économie de thérapie adjuvante, 2 brouillard totale et 3 d'office traitées.

survfit : 3 courbes de KM. une de meilleurs pronostiques et deux autres de moins bon. On ne sait pas encore quel grade correspond à quel courbe.

valeur pronostique pour le grade 1

x[x==3] <- 2 : mélange grade 2 et 3 . on obtient quelque chose d'encore plus significatif

Traitement semble ne pas avoir d'effet quand on trace KM car donné en fonction du diagnostique (= confondant).

Taille tumeur pronostique ? les plus petites ont meilleurs pronostique

analyse multivariée : quand combine les variables, celles qui étaient significatives, ne le sont plus spécialement. Ce la veut dire qu'elles sont associées entre elles. Si on en regarde une, on regarde implicitement les autres. Les combiner n'ajoute pas d'infos. La façon dont on mesure est très importante.

on pourrait réduire à médiane continue en prenant médiane plutôt qu'une variable binaire.

14/03/2016

10 Gene expression predictors for breast cancer outcome reflect programs related to proliferation

Est-ce que les prédicteurs sont comparables ou pas ?

10.1 how failed attempts to understand thyroid cancer agriveness ...

Sl3 : thyroid normale = boules (folicule). PTC formes foliculaires. ATC = amas cellules informes.

Sl4 : certains cancers n'avaient pas p-53 mais ST. Utilisation astuce de Chang pour avoir nouvelle signature. La plupart des gène sont d'un côté du diagramme. Séparation nette entre tissus cancéreux et normaux. Signature p-53 plus exprimée dans PTC que dans tissus normaux. Comparaison ATC - PTC, plus dans ATC. Contre intuitif car ATC plus agressif.

Sl5 : quiescente, proliférescente, sénescente. H : scénescence = protège contre le cancer. Expression de marqueurs scénescence plus élevé dans PTC que dans normaux. Plus élevé aussi dans ATC que PTC or contre-intuitif car agressif donc le patient n'est pas protéger car va trop vite.

Sl7 : PCNA = marqueur de la prolifération. Noyaux en train de se diviser en brun. Calcule la fraction de noyaux en cycle.

Sl8 : metaPCNA = les 1% des gènes les plus corrélés à PCNA. 131 gènes. Exprimé quand PCNA est exprimé. Index convaincant.

Sl9 : PCNA permet de mesurer l'agressivité du cancer.

Sl10 : relation inverse entre prolifération et différentiation. Différentiation différente pour chaque organe.

10.2 part2

Sl17 : comment se comporte metaPCNA dans les cancers du sein ? barre noire quand gène fait partie des 1% de gènes les plus corrélés. sens de résumer l'expression les 130 gènes à un seul chiffre (metaPCNA).

Sl19 : déconvolution est fondamentalement différente de l'astuce de Chang. Chang enlève de l'info mais ici pas. Chang dépend de Fourier.

Sl21 : application log 2 sur les données. Lumière exponentiel

Sl29 : est-ce que toutes les signatures ne sont pas pronostiques ? Recherche de signature non associé au cancer. Trouve article japonais sur signature rire postprandial ⇒ signature pronostique pour cancer du sein. Autre =localisation fibroblaste ⇒ pronostique. 3e exemple : effet défaite sociale ⇒ pronostique aussi ! :o On a donc un problème avec les signatures.

Sl30 : étude non plus de 3 signatures aléatoires mais de 10 000 aléatoires. Prend comme ST et mesure association avec cancer du sein

Sl31 : EG plus de 50% sont pronostiques.

Sl32 : même chose avec déconvolution : tout s'écrase vers des p-values proches de 1 (0 en log)

Sl41 : pronostique = prolifération, , dvlpmt strauma

Sl43 : transcriptome pronostique que dans ER

Sl47 : limites de l'étude : 1 traite population limitée de patients du cancer du sein et toutes ensembles. 2 pas qualité pronostique des tumeurs qui sont mises en cause mais leur interprétation biologique

16/03/2016

11 Introduction to high-troughput sequencing (next generation sequencing)

Remind : Combinatoire infinie de la génomique fonctionnelle. Mais beaucoup d'erreurs ! Plus on observe de choses, plus des choses aléatoires deviennent probable. Toujours tester sur des données aléatoires. Souvent, le bon modèle nul n'est pas toujours évident. Variables souvent corrélées entre elles car le génome et les séquences ADN sont corrélées. Causalité ne peut pas être faite par de la bio-informatique pure car besoin d'expériences pour induire des perturbations et observer les résultats. Echange de données (publique) permet à d'autres de refaire les calculs et de critiquer. Echange de code difficile pour le séquençage à haut débit car déplacement des données lourd et il faut un bon appareillage informatique pour les faire tourner.

Sl2 : y = cout séquençage génome humain. Shotgun sequencing : casse le génome en petit morceaux et les séquences avec une certaine redondance. Les morceaux se recouvrent et permettent de reconstituer la séquence entière. Cout du séquençage du génome énorme au début. 2008 révolution technologique : séquenceur à haut débit → permet baisse du coup. Depuis le cout ne fait que baisser. Loi de Moore d'application avant 2007. Après 2007, la technique génomique progresse plus vite que les ordinateurs. Séquençage des anglais (1000 genomes project) → permettra de faire des avancées en médecine mais pb éthique car le génome est quand meme très personnel.

Sl3 : séquençage va bien au delà du génome. Production d'une librairie (amplification, ...) → séquençage → analyse des données. Séquençage est devenu industriel

Sl5 : la science arrive dans un moment de Big Science. Grands centres produisent des infrastructures de données (cf séquençage 1000 génome).

Sl6 : application des technologies de séquençages. Intestin peuplé des bactéries. Echantillonage des espèces présentes dans notre organisme → meta génomique (on séquence plusieurs génomes). 2004 premier papier sur la meta génomique par Graig Venter. Pas possible de caractériser chaque bactéries individuelles mais fait grandes catégories. Diversité de notre microbiome est inimaginable. Facteur 1/1000 par rapport avec notre génome. On peut aussi les classer par catégories fonctionnelles.

Sl7 : Application PCA sur les microbiomes. 3 clusters des individus. Analogies avec les groupes sanguins. Pas tous identiques. Certains ont une prédominance d'un certains groupe de microbiome.

Sl8 : trouve qu'un type est associé au diabète de type2. Idées de greffe de microbiome pour guérir certaines maladies.

Sl9 : prendre des os de Neanderthal, en extraire l'ADN et le séquencer. Comparer aux génomes de singes ou au génome humain actuel. Possibilité de contamination par de l'ADN contemporain car manipulé par humain. Détermine que dans le génome européen contemporain il y 4 à 6% qui vientt de Neandertal. Graig Venter à réussi à injecter un génome artificiel dans une bactérie. L'idée est de créer des bactéries de toutes pièces.

Sl10 : papier de Pieter Campbell. Etude de tumeurs du sang car viennent de lymphocytes. Hémunoglobuline différent pour chaque partie du corp. Séquence cette partie caractéristique plutôt que tout le génome. Reconstitution d'un arbre phylogénétique de la tumeur.

Sl11 : séquençage de fragments distant physiquement d'une tumeur rénale. Patient métastatique, post mortem. Pour chaque fragment trouve les mutations (en gris). Certaines présentes dans tous les échantillons, d'autres uniquement dans les métastases, certains dans un seul fragment, ... Permet de nouveau de faire une arbre phylogénétique de la tumeur. Diversité génétique des tumeurs solides est énorme. Tous les scénarios sont possibles. Thérapie ciblée devient plus difficiles que ce qu'on pensait.

Sl12 : séquençage de génome de tumeurs mammaires. Evolution clonale (= mutation, duplication) permet de refaire l'histoire de la tumeur et des mutations.

Sl13 : ligne = gène et colonne = patient. On cherche les mutations chez quel patient. Catalogue complet des mutations codantes. La plupart des mutations sont privées (propre au patient) et les mutations fréquentes étaient déjà connues. Renforce l'idée que le cancer est une maladie génétique et différente d'une personne à l'autre. Certaines mutations qu'on pensait propre à un type de cancer se révèlent dans d'autres cancer. Débat pour savoir si on doit séquencer encore plus de génome.

Sl14 : On peut trouver dans le génome des historique des infections (fossiles des virus). Les différents mutagènes ont une signature, ils ne vont pas faire muter n'importe quelle partie/zone. Différents type de cancers ont différents types de mutations et donc on peut les associer à différents agents connus ou non. Permet de déterminer la cause du cancer.

Sl16 : phénomène kataegis(averse) = nouveau

Sl17 : on peut générer des codes barres ADN

Sl18 : séquençage permettant d'analyser la structure secondaire des ARN.

Sl22 : trouve des polymorphismes humains importants qui affectent la structure secondaire. Conséquences médicales importantes

Sl23 : séquençage de cellules individuelles. Nouveau et va exploser dans les quelques années à venir. Peut être que la moyennes des cellules ne sont pas représentatives des cellules individuelles. Pour l'instant beaucoup dans le sang. Tag sur les ARN puis séquençage normal. Lors de l'analyse, on sait que les ARN portant un certain tag appartiennent à un individu.

Sl24 : matrice de corrélation entre les transcriptomes. Plus c'est rouge plus c'est corrélé. Voient aussi des clusters de cellules dans l'espace réduit à 2 dimensions.

Sl25 : Les cellules dentritiques sont sous-divisées en groupes de cellules. Soumission des cellules à des perturbations pour voir quels gènes sont perturbés. Voir effet antigène sur les cellules individuelles. Fold défini par "pas de perturbation".

16/03/2016

12 Alignment of RNA-seq short reads

Sl2 : fragments = 150 pb
Sl4 : technologie non biaisées comparé aux microarray. Epissage alternatif = règle et pas exception.
Sl5 : on peut déduire la structure du gène grâce au séquençage et à l'alignement.

13 The cancer genome Atlas

14 Projet

Notion d'age et de vieillissement. Idée que le vieillissement est inévitable est en train de changer. On a pu augmenter la durée de vie de certains organismes (C. Elegans d'un facteur 7, chez la souris augmentation espérance de vie d'un tier).

Idée est de sous alimenter les souris mais pas bien pour l'homme. Autre idée est de prendre de la Rhapamycine.

Vieillissement = ralentissement métabolique = protection contre le cancer car limitation des duplications (limite la transmission des mutations).

Différence entre age biologique et chronologique.

Travail de Steve Horvath : existe-t-il une horloge biologique mesurable ? Il travaille sur la methylation. Trouve un ensemble de 300 marqueurs épigénétiques permettant de définir l'age biologique (avec précision de 3-4 ans). Valable pour n'importe quel organe.

Q1.3. : Est ce que si la personne a un foie jeune, son cancer sera jeune aussi ?

Q1.4. : Nom TCGA = nomenclature difficile. Prendre barcode. (wiki.nci.nih.gov/display/TCGA/TCGA+Barcode)

Q2 : Calculer toutes les corrélations entre age clinique et DNAm age. 2. Age methylation avec la survie utiliser kaplan meier.

Q3 : 1.GSEA package initial ou app java ou interface graphique ou autre algo (cat gène C2 : CP). 1 et 2 comme question 2. Sam Options pertinentes (utilise RNA seq). GSEA transformer en log 2 ajout log(1+donnee) (si s'arrête à 20 et histo avec longue queue non).

bien documenter en donnant les versions, les paramètres qu'on à utiliser, ...

Le plus court le mieux : concis et précis