10/02/2015

Dogme central : flux à sens unique dans la biologie de l’ADN => ARN => protein

Génomique fonctionnelle : déclinaison du dogme à l’échelle du génome

ARN : microarrays

Limites des méthodes d’analyse sont énormes.

Côté révolutionnaire :

* Voir des choses dont on n’était pas capable avant
* Phénotype difficile à expliquer : polymorphismes humains, aborder des maladies aux causalités complexes aux facteurs environnementaux et génétiques
* Information de base est stockable et échangeable

GETEX : les gens sont morts d’accidents de voiture, etc. Mais 30% sont mortes d’accidents vasculaires cérébraux et d’attaque cardiaque. Le temps que les gens sont restés dans la salle d’opération ou sous respirateur artificiel semble avoir une influence non triviale sur l’expression des gènes.

Hypothèse : deux préparations d’ARN 🡺 deux mêmes quantités globale d’ARN

Or, ce n’est pas toujours vrai (ex avec sur-transcription globale due à Myc dans le cancer)

Poisson WT vs Poisson BMP2 muté

X-première paire de manip et Y-la seconde

Si les manips étaient parfaites, on aurait dû avoir une ligne alignée sur x=y

On peut l’expliquer en disant qu’il n’y a pas de signal important et ce qu’on mesure est du bruit de fond.

Cependant il y a des tas de gènes qui suivent la ligne et il existe des gènes régulés dans la manip 1 et pas dans la manip 2.

**Script R**

« twoD » = régression à deux dimensions (points tridimensionnels sur plan bidimensionnel)

Pas de correction possible de la griffe, quelle que soit la méthode

Log de l’intensité rouge sur log de l’intensité verte

On peut créer une variable x qui reprend les points ayant une valeur au-delà de 2.4 d’intensité 🡺 les outliers correspondent aux points sur la griffure

La quasi-totalité des outliers sont rouges, ce qui confirme que ces outliers sont liés à la griffure

Tous sauf x = !x

Pch=19 = points remplis

Pch= « + » = points représentés par un +

Censé être moins régulés dans le mutant que dans le WT. Un des trois BMP2 est exprimé, donc mauvaise effet de la mutation, pas toujours perte de fonction. Les gènes normalement non régulés par BMP2, on en constate deux qui sont sous-régulés chez le mutant, ce qui n’est pas normal.

**Comment on fait pour établir une différence d’expression entre deux conditions ?**

PTC : cancer de la papillo-thyroide

Tissus malades et sains sur la même personne sur la thyroïde

1. Ratio d’expression tumeur vs tissu sain
2. Gènes avec FC d’au moins 2 et fréquent (au moins 2/3 des patients)

VENDOR\_ID = nom de la puce affymetrix

SYMBOL = id du gène

N° = numéro du patient et N pour normal et T pour tumeur

En moyenne les gènes ne sont pas régulés, mais ceux qui le sont le sont beaucoup !!

De = differential expression

Apply = applique une fonction sur une matrice

Somme de tous les x supérieurs ou égal à 1, et envoie vrai si supérieur à 8

Or, tout cela était une illusion statistique, car les samples étaient labellés sains et tumoraux au hasard.

Il existe trois types de choix techniques :

* Qu’est-ce que ça veut dire différentiellement exprimé ?
* Distribution nulle (distribution statistique quand aucun gène est régulé)
* Choix multiple

T-test : moyenne divisé par la déviation standard

Pas satisfaisant dans le microarray, car gènes ont un sigma très petit, donc on a une variante du t-test où on ajoute une constante à la variation standard pour que le dénominateur de la fraction ne soit pas proche de zéro.

Distribution nulle : peut être estimé à partir de data mixées (bootstrap method)

Bonferroni correction : stupide

S(i) : variance globale du gène

S0 : 1% de la distribution de la différence d’expression du gène

q-value = émerge parmi les techniques de contrôles du test multiple. Fraction de Faux positifs (combien il y a d’échantillons qui ont une bonne p-value, mais qui sont faux) La distribution des p-values donnent la valeur des faux positifs. Si non-régulés, ce serait plat. Si c’est différentiellement exprimé, la distribution est biaisée.

Delta 1.5 = 11000 du transcriptome différentiellement exprimés (la moitié des gènes observés)

Avec les données trafiquées : les gènes sont tous alignées sur x=y dans qq plot 🡺 aucun gène différentiellement exprimé

**Are most published research findings false ?!**

Phase 3 : infos cliniques de phase 2 indique déjà une réussite (donne une puissance à R)

C = nombre de gènes étudiées,…

N = nombre d’études indépendantes qui traitent de la même chose

(T = associé à la maladie, F = non associé à la maladie)

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Real World | |
| Statistique |  | YES | NO |
| Significant |  |  |
| Not significant |  |  |

PPV = Nbre d’événements vrais positifs sur les événements positifs totaux

On a multiplié par 10 la croyance de l’association entre le gène Z et la shizophrénie avec la publication dans nature.

Il a démontré que le biais existe.

**Evaluation of classifiers quality**

Création d’un test statistique qui va classer le profil pour faire un diagnostic.

1. Linear classifiers : on définit des espaces de classes avec des droites, discriminants et arbres de décisions
2. Non-linéaires : courbe compliquée, neural network, support vector machines

Problème : sur-adaptation aux données 🡺 aucune appartenance à des classes (pas de classes bien définies dans le nuage de points)

Si on diagnostique un cancer grave, le fait de rater un cancer est plus grave que le fait de sur-diagnostiquer.

AUC : on peut ordonner les classifiers en calculer la zone sous la courbe ROC. Plus l’AUC est grand, plus le classifier est bon.

Aucune information venant des données d’apprentissage ne doit être présente lors de la validation.

24/02/16

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.

**Dia 24 :** Un algorithme de classification a lui-même des paramètres. Typiquement, les gens vont tenter pleins de paramètres différents et reporter le meilleur 🡺 PAS BIEN !

**Dia 25 :** On partitionne le set pour faire des cross-validations successives (principe jack-knife)

**Dia 26 :** 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

**Dia 27 :** 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

**Dimension reduction of gene expression profiles**

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

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 signigie 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

1. **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.

1. **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

1. 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

1. 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.

1. **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)

**Gene sets expression**

On cherche à étudier des groupes de gènes pour apporter un sens biologique.

On va essayer de passer du niveau gène au niveau métagène.

La pleiotropie : selon le contexte un gène a une fonction différente

Si un gène n’est pas exprimé dans un système, cela ne veut pas dire qu’un autre gène avec la même fonction n’est pas exprimé.

* L’absence d’un gène n’est pas conclusive !

Il faut que la différence d’expression soit importante pour être détecté… OR certains processus peuvent induire un changement subtil d’expression.

D’autres processus pourraient avoir un effet sur le transcriptome au complet.

La plupart des marqueurs moléculaires doivent être basés sur des dizaines de gènes !!

Une première approche pour grouper les gènes entre eux est d’utiliser l’ontologie des gènes (contraints par des règles formelles qui décrit la fonction des protéines). C’est construit du niveau le plus abstrait au niveau le plus élémentaire.

Le marqueur multigène permet de contrer le problème de mauvaise mesure à cause d’une probe mal faite,… Le marqueur est plus robuste.

Les méthodes statistiques :

Listes de gènes différentiellement exprimés entre condition A/B (et une liste bruit de fond) et on les compare avec les listes de Gene Ontology.

Ce n’est pas puissant statistiquement, car ce qui compte c’est le nombre de gènes, pas le nombre d’échantillons.

02/03/16

Précaution : contrôle négatif statistique 🡺 distribution nulle

Biais de sélection de paramètres, des gènes sélectionnés,… 🡺 erreur du classifier

Comparer les phénotypes de tissus, sans s’intéresser à une fonction particulière 🡺 levure avec des perturbations (perturbations de grandes collections de gènes)

Métagènes = expression combinée de plusieurs gènes

Les gènes n’ont pas une fonction univoque, ce qui rend l’interprétation des données difficile.

**Gene sets expression (suite)**

Grouper les gènes en fonction de leur fonction physiologique est intéressant, car on peut étudier les gènes en groupe (pour le cycle cellulaire, la conversion en énergie,…)

On utilise le groupe de gènes régulés lors d’une perturbation comme marqueur « gene set » pour l’étude et faire un lien conceptuel entre des manips in vitro et in vivo comme un genre de langage.

Méthode pour estimer l’expression différentielle d’un groupe de gènes :

* La plus utilisée, mais pas la meilleure : on sélectionne des gènes différentiellement exprimés, qu’on compare avec les gene sets d’intérêt fonctionnel. Correction de test multiple sur les p-value d’intersection entre gènes diff exprimés et gene sets.

Test hypergéométrique : pihyper ? dans r

Le nombre d’échantillons n’influencent aucunement dans ce calcul, ce qui ne va pas du tout. Puissance statistique artificielle. De plus, cela suppose que les gènes soient indépendants les uns des autres, ce qui est faux !! Groupes de gènes co-régulés en général… 🡺 surestimation grossière de la puissance statistique.

* Gene set enrichment analysis : il utilise un échantillonnage des patients, pas des gènes et éviter de mettre un treshold sur ce qui est considéré un gène différentiellement exprimé. Ils voulaient des gènes sets qui ne soient pas baser sur des connaissances scientifiques, mais sur des expériences primaires. Ligne gènes, colonnes échantillons, ordonnés selon la classe A et B et gènes arrangés selon la mesure avec laquelle ils sont associés à la classe (gène le plus exprimé dans la classe A et le plus sous exprimé dans la classe B,…) Ils marquent les gènes qui appartient à la signature q21 (par exemple) en noir en fonction du rang. Ils calculent une statistique le score d’enrichissement, ils parcourent la liste et à chaque fois qu’ils rencontrent une barre noir (hit = gene set dans cette liste) et lorsque le gène n’est pas dans la liste c’est un miss.

Trois gene set différent dans des circonstances différentes. En rouge :: les gènes globalement sont associés à la classe A. en bleu : les gènes sont reportés aléatoirement le long de la ligne 🡺 score faible. En vert : une répartition quasi aléatoire, sauf qu’il y a un peu plus de gènes é gauche qu’à droite, on détecte une association globale du gene set avec le phénotype classe A.

1. Pour chaque gene set on calcule ce score
2. Distribution nulle en distribuant aléatoirement à la classe A et B les échantillons 🡺 p-value
3. Test multiple pour prendre en fait qu’on s’intéresse à toute l’ontologie des gènes au lieu d’un seul gene set.

MsigDB 🡺 tas de gene sets d’intérêt

* CGP (chemical on genomic perturbation) (C2) = gene sets expérimentaux (5000 gene sets, sur les humains en principal)
* C3 : gènes modifiés par le même facteur de transcription et microARN
* C4 : cancers, gènes prototypiques (p53,…) et leur co-régulation
* C6 : signatures oncogéniques
* C7 : fondés sur résultats d’un consortium américain, cellules immunitaires (souris et humaines) soumises à différentes perturbations, interleukines, cytokines,…

Control : Male vs female lymphoblastoid cells lines

Ils trouvent qu’il y a deux genes sets qui sortient : liés à la bande cytogénique Yp11 et Yq11 (normal puisque les femmes n’ont pas de chromosomes Y). Avec C2, il sort des genes sets liés aux organes reproducteurs, sur les chromosomes autosomaux. C’est étonnant puisqu’il ne devrait pas avoir de différence puisque ce ne sont pas des cellules sexuelles…

Application : cytogenetic abnormalities in acute leukemia.

Scanner toutes les bandes en regardant le transcriptomes des leucémies. 24 d’origine lymphoïde >< 24 d’origine myéloïde (macrophage, dendritique)

Regarde les gene sets C1. Par exemple le chr13q14 sont souvent déletées dans les leucémies de type myéloïde. Chr14q32, gènes se retrouvent sous-régulés car les cellules cancéreuses sont sous-différenciées

Application : p53 status in cell lines

Les cellules normales ont des profils de p53 répondant à du stress, tandis que les cellules mutantes on tune surexpression de l’oncogène Ras

La plupart des gènes trouvés sont suractivés quand p53 sont suractivés, comme PIK3CA, un oncogène et deux autres impliqués dans la voie des kinases 🡺 la mutation p53 causerait l’abrogation de la réponse au stress, mais sur-active la voie PIK3CA.

Application : outcome in two lung cancer studies

Deux cohorts de patients stratifiés entre bons et mauvais prognostics, à l’aide d’échantillons tumoraux conservés pendant des années, et on observe ce qui est devenu du patient beaucoup plus tard.

Bien que les deux groupes de gènes ont une petite intersection, les gene sets ont une grande corrélation l’un avec l’autre.

Les gènes sets ne sont pas similaires pour les deux études, mais on en retrouve certains dans les deux.

The connectivity map

Utiliser les profils transcriptionnels pour connecter entre elles des perturbations biochimiques, ou les connecter à des maladies et connecter des maladies entre elles. On se fonderait entièrement sur de la biologie moléculaire, sans nommer aucun gène.

1. connecter les petites molécules

quand l’histone disparait, l’Adn est déméthylé et la transcription est possible. Ils prennent les 13 gènes répondant à l’inhibition de l’histone et regardent les expressions pour les autres perturbations et découvrent que certaines perturbations ont le même effet sur ces gènes et en déduisent qu’ils ont pour effet d’inhiber l’histone et non la méthylation.

129 gènes touchés par l’œstrogène 🡺 fulvestrant a un effet négatif sur ces gènes 🡺 antagoniste de l’œstrogène

1. Comparer des conditions médicales

Signature dérivé d’adipocytes de rats soumis à un régime induisant l’obésité et un normal.

La connectivity map les relie à des perturbations liés à PPRA gamma, impliqué dans le métabolisme des graisses.

Influence des glucocorticoides (stéroïdes suppresseurs du système immunitaire) dans le traitement de leucémies 🡺 même chose que la rapamycine (inhibiteur du gène mTOR)

**Gene expression predictors of breast cancer outcome**

On veut prévoir l’issue de la maladie des années à l’avance.

**Analyse de survie**

Voir des événements censurés : quand le patient meurt après la fin de l’étude = on n’a pas l’information (mort ou vivant) dans le temps imparti de l’étude)

Pourcentage de patient encore vivants dans la cohorte initiale 🡺 à chaque fois qu’un patient mort, la courbe descend d’un pallier et un événement censuré est marqué d’une barre mais on ne descend pas d’un pallier (courbe de Kaplan-Meier)

Ce qu’on veut, c’est une mesure intégrée de la différence de survie.

Probabilité qu’un patient soit en vie au temps t = probabilité qui ne soit pas mort durant les temps précédents

Proportionnal hazard function : hypothèse selon laquelle la probabilité de mourir est la même à tout temps t 🡺 la dérivée temportelle du log de la fonction de survie

Ce qui nous intéresse est de connaître la différence de h ratio entre les gens avec un astrocytome avec les gens qui ont un glioblastome 🡺 ebêta

Bêta représente le hazard ratio quand on augmente d’une unité la covariante dont il est le facteur

**Extremely brief introduction to cancer progression**

Hayflick limit : potential de replication limité (extrémité des télomères sont rognés peu à peu

Sanger Institute : remet en question la notion d’oncogène et de suppresseur de tumeur

* Hétérogénéité déjà présente dans les tissus normaux !! Donc les tissus cancéreux ne sont pas les seuls tissus hétérogènes

**Prediction breast cancer outcome with microarrays**

Thérapie ciblée : interfere exactement avec des protéine/gène spécifique du cancer.

Les chimiothérapies interfèrent avec les processus de prolifération. (ex : terminateurs de synthèse d’ADN, analogues nucléosidiques)

Examen des ganglions lymphatiques 🡺 si c’est le cas, tumeur agressive, chimiothérapie conseillée

Profil microarray représentant une signature transcriptionnelle. Colonnes gènes, lignes patientes.

Sous la ligne pointillée, beaucoup plus de chance d’avoir des métastases (probablement gènes en rouge)

# 14/03/16

Malédiction de la multidimensionalité = peu de points et beaucoup de paramètres 🡺 overfitting

Acute leukemia : combinaison de rapamycine et glucocorticoïdes pour combattre la résistance aux glucocorticoïdes

LINKS database pour connectivity maps, 1 million de profils d’expression et résultats pour un certain noombre de petites molécules

Analyse de COX qui quantifie la probabilité que deux courbes de survie de Kaplan Meier soient plus différentes que dû au hasard

**Prediction breast cancer outcome with microarrays (suite)**

Malgré le fait que certaines femmes n’aient pas les ganglions lymphatiques envahis ce qui est considéré un bon diagnostic, on détecte des gens avec des mauvaises signatures transcriptionnelles et une courbe de survie beaucoup plus basse que celles avec une bonne signature

Finalement on a cherché le gène qui déclenche les métastases, mais existe-t-il vraiment ? On avait un schéma où une cellule subissait une mutation qui permettait l’hyperplasie, puis l’invasion,… Chaque mutation successive amenait un clone qui battait les autres (darwinien). Cela continuait jusqu’à ce qu’on obtienne un clone capable de se disséminer.

Les 70 gènes trouvés prouvent qu’il existe des signaux transcriptionnels dans toutes les cellules de la tumeur. Au fond, s’il y a une compétition darwinienne qu’est-ce qu’il pourrait promouvoir la compétence de métastaser. Donc ce potentiel de métastaser devait exister bien avant que les cellules soient réellement capables de métastaser.

Pouvait-on apprendre la progression de tumeur vers cancer métastatiques en étudiant les différences entre les cancer à bon pronostic à ceux à mauvais pronostic ?

On regarde les tumeurs de poumons et les métastases de poumons. Ils ont découvert 128 gènes qui permettent de distinguer des cancers du poumon métastatiques à des cancers du poumon non métastatiques. Ensuite, on prend des tissus congelés dont on connait l’issue du patient (survie ou mort) 🡺 quels gènes permet de distinguer l’issue du patient ? Est-ce que ces 128 gènes aident à ce pronostic ? En effet ces gènes ont un pouvoir pronostic dans des données indépendantes à l’étude initiale. Ils arrivent à réduire ces 128 gènes à 17 (permet l’utilisation de PCR 🡺 plus facile à appliquer en hôpital) Est-ce que c’est également utilisable dans un autre organe ? Oui, dans cancer, médulloblastome et prostate, mais pas dans le cancer du sang (lymphome). Les résultats ne sont donc pas limités à un organe, ce qui est étonnant car les cancers sont considérés très spécifiques à leur localisation. Le résultat est normal pour le lymphome, car les 17 gènes ont été obtenus pour des tumeurs physiques et pas liquides (sang, lymphes,…) qui sont « naturellement » métastasiques.

**What biological processes are the outcome predictors involved in?**

Si on arrive à comprendre ce processus, on peut créer des traitements pour éviter la formation de métastases.

On a vu qu’il était possible de relier les signatures transcriptionnelles à des réponses biologiques (réponses à des perturbations,…).

Transformation de fourrier : détecte les gènes qui agissent de façon périodique

On peut attribuer les gènes avec les étapes du cycle cellulaire 🡺 si on avait aucune info a priori, on serait capable de les trouver ici.

Une tumeur ça ressemble à une blessure qui ne cicatrise jamais 🡺 processus physiologique de cicatrisation proche du mécanisme pathologique de la tumeur (inflammation, invasion cellulaire, prolifération, angiogenèse, fibrose…) Cependant dans le cancer, il n’y a pas de résolution du processus, d’où les problèmes. Pas de suivi de cette hypothèse car difficile à investiguer.

L’utilisation du sérum est un modèle in vitro de la cicatrisation 🡺 obtention de la signature transcriptionnelle de la cicatrisation.

On compare ensuite cette signature à une signature de tumeur pour voir si la cicatrisation est bien impliquée dans la progression du cancer. Le problème : un grand nombre de gènes est impliqué dans le cycle cellulaire, ce qui est d’office impliqué dans le cancer. Donc ils ont pris la précaution d’enlever les gènes associés à la prolifération qu’ils ont trouvés dans le papier où ils ont associés les gènes au cycle cellulaire 🡺 gènes core serum

Ils comparent alors des tumeurs avec des tissus normaux du même organe avec cette signature. Ils voient que les tumeurs expriment les core serum gene et pas les tissus normaux 🡺 valeur diagnostique universelle

Puis ils essaient de voir si la signature permet de distinguer les pronostics des tumeurs 🡺 pour 3 cancers c’est significativement associés à la survie/mort (ceux présentant la signature sont des tumeurs plus agressives).

La présence de cellules souches prédisent le retour de la tumeur.

D’autres parlent de l’hypoxie comme un moteur de la tumeur.

Pour les cancers du seins, plusieurs catégories de gènes :

* Associés à des marqueurs connus de la progression tumorale (prolifération,…)
* Dérivés en recherchant des gènes corrélés à des facteurs cliniques connus (index mitotique, invasion des ganglions lymphatiques,…) = grade génomique
* Associés à l’issue (profils transcriptionnels de tumeur qui ont progressé >< n’ont pas progressés et on retire les gènes qui les distingue)
* Signatures dérivées à partir d’hypothèse physiopathologiques sur les mécanismes de progression du cancer (hypothèse d’un mécanisme 🡺 on dérive une signature transcriptionnel de ce processus par in vitro/in vivo 🡺 cohortes de cancer de sein avec issue du patient et on vérifie si la signature est pertinente ou pas)

**But do these signatures have different predicting abilities ?**

Aucune de ces signatures n’ont un meilleur résultat que les 70 gènes de base.

De façon amusante, elles se trompaient toutes sur les mêmes patientes 🡺 elles se basent toutes sur le même paramètre, malgré le fait qu’elles n’ont aucun gène en commun

Breast.R

Sa = sample annotation

Ga = gene annotation

Sapply = deux arguments 🡺 function sur tous les éléments du vecteurs (= boucle for)

Attach(x) = permet de faire directement référence à x

Les patientes en l’absence de retour de la maladie sont suivies environ 9 ans.

Reccurrence = si la maladie revient

Related death = mort à cause du cancer

Surv…. S = Facteur dans R qui est le temps de survie (suivi dans + quand il y a censure de l’événement)

Variable y qu’on veut expliquer en fonction de x, z, a 🡺 utilisation du ~

S~1 🡺 regression et on veut prendre l’intercept de la regression comme base.

Chaque décroissance de la courbe représente la récurrence de la maladie + exit de l’étude. Les pointillés représentent la variance de la courbe.

S~ER 🡺 expliquer S en fonction du statut des récepteurs à l’œstrogène. Dans les années initiales on a des différences, l’absence de ER donne un mauvais pronostic, mais cet effet s’annule au fur et à mesure du temps (cependant peu d’événements au-dessus de 5 ans)

Analyse de Cox pour voir si association significative ou pas. (ph = proportional hazard)

Analyse de kaplan meier on récupère la p-value en faisant summary(coxph(s~ER))$logtest[« pvalue »] (cherche la valeur pvalue dans la colonne logtest)

S~gradehistologique 🡺 2 qui sont confondues et de mauvais pronostics et une de bon

Avec l’analyse de Cox 🡺 le grade est significativement associé à la survie 🡺 valeur pronostic au grade histologique

S~ganglions 🡺 2 courbes tout à faits séparées, le Cox révèle une p-valeur significative

S~traitement 🡺le traitement ne semble avoir aucun effet pendant les 7 premières années puis un des deux groupes survit mieux. Il y a un facteur confondant car le traitement est donné en fonction du diagnostic (présence de ER, grade histo,…)

S~tumor size 🡺les tumeurs plus petites ont un meilleur pronostic. Plus la tumeur est grande, plus il y a de candidats pour la mutation et on augmente la chance de métastases.

Modèle statistique multivarié 🡺 ~ER, taille, présence de ganglions 🡺 Quand on combine les variables, elles ne sont plus nécessairement significatives car elles sont associés entre elles. La significativité très élevée de la taille de la tumeur influence le poids des autres facteurs. Les variables continues fonctionnent mieux parce qu’on utilise le temps, lui aussi une variable continue, pour la survie. Lorsqu’on transforme la taille en valeur binaire, la significativité disparaît 🡺 la façon dont on mesure la variable est donc également très importante.

On collecte le plus de données possibles durant les études précliniques et on essaie d’étudier le traitement et le diagnostic à la fois.

Svantveer = 70 gènes

Chang = core serum

Chi 2006 = hypoxie

Carter 2006 = aberrations génomiques globales

Sortiriou = grade genomique

Glisnky = cellules souches

Intersect = gènes de la signature aussi mesurés dans le microarray

1-cor(x) = matrice de corrélation

On utilise le clustering hiérarchique pour stratifier la cohorte 🡺 binariser pour décider traitement ou pas (cutree)

Survie de s en fonction de pred 🡺 deux groupes avec une très bonne séparation

Sample(rownames(M),100) = donne 100 des lignes de M au hasard

Traiter un grade génomique en variable continue (médiane des gènes) 🡺 donnerait une meilleure significativité que la variable binaire

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

Est-ce que ces prédicteurs sont comparables ou pas ?

Présenté de la façon dont s’est déroulée la recherche.

**Part I : comment un échec à étudier la progression des cancers de la thyroïde a conduit à la découverte de la méta-PCNA**

Les cancers papillaires et anaplasiques sont très bien différenciés. Qu’est-ce qui explique leur différence d’agressivité ?

PTC = semblant de follicules >< ATC = complètement dédifférenciés

P53 peut expliquer l’agressivité… (dans cancer du sein, parfois pas la mutation de p53, mais la signature transcriptionnelle de cette mutation 🡺 même « phénotype ») On a d’abord trouvé une signature p53 muté en comparant les profils de transcription de tumeurs mammaires avec le p53 muté et le WT 🡺 espérance d’avoir une signature de p53 muté non spécifique au sein (idéaliste et naïf)

Tissus normaux vs tissus tumoraux papillaires 🡺 p53 plus dans les tissus normaux

PTC vs ATC 🡺 signature p53 est très active dans ATC, contradictoire car normalement p53 est protecteur !

Une autre hypothèse est que lors d’un stress cellulaire, p53 peut également entraîner la sénescence voire la mort cellulaire. La sénescence pourrait être un mécanisme de protection dans les cancer d’où la plus grande signature dans l’ATC qui est plus agressif.

En comparant les cellules quiescentes aux cellules sénescentes 🡺 signature transcriptionnelle de sénescence.

PTC vs ATC 🡺 même chose qu’avec p53, plus de signature de sénescence dans l’ATC, ce qui est contradictoire car le cancer est super agressif et tue les gens en quelques mois ! Donc la sénescence comme protection n’a aucune logique dans l’ATC.

Signature de la prolifération PTC vs ATC 🡺 grande prolifération dans ATC 🡺 p53 et sénescence ont pour facteur confondant la prolifération = problème de prolifération

Donc on peut mesurer la prolifération à partir du profil transcriptionnel, il suffit d’avoir un marqueur adéquat.

Super/Meta PCNA = 1% de gènes les plus positivement corrélés à l’expression de PCNA à travers un grand nombre de tissus normaux (131 gènes, exprimés quand PCNA est exprimé et pas exprimés quand PCNA n’est pas exprimé)

* Formation du fuseau mitotique
* Régulation cycle cellulaire
* Synthèse nucléosides

La plupart des gènes qui composent cet index ont leurs protéines exprimées dans le noyau.

Quand on regarde l’expression de cet index dans les TC (PTC vs ATC) 🡺 PCNA plus exprimé dans ATC 🡺 mesure l’agressivité du TC

De même, les cancers avec p53 muté ont une signature PCNA plus élevée.

YES ! Index pronostic universel !

Article avec signature pour dédifférenciation = croix quand overlap avec la signature PCNA 🡺 sur les 69 gènes de PCNA, 40 dans la signature de dédifférenciation 🡺 très similaire, ils mesuraient donc la prolifération et leur mesure de la dédifférenciation était indirecte car les cellules en prolifération sont par nature moins différenciées. D’autant plus, la différenciation est spécifique à chaque organe.

**Part II**Un article qui prétend qu’une signature des cellules souches prédit la survie du cancer du sein (BORING car ils disent tous les même chose)

Les auteurs de la signature p53 comparent avec celui de cette signature 🡺 massivement corrélée, de plus ils savent que leur signature est corrélée à la prolifération.

Facteur pronostic principal = prolifération ?

L’idée serait d’utiliser la signature pour retirer du signal pour retirer les gènes liés à PCNA et voir si ces prédicteurs fonctionneraient toujours avec ces gènes enlevés.

Cette signature de PCNA, comment se comporte-t-elle dans le cancer du sein ? Barre noire quand le gène appartient à la signature PCNA. La plupart de ces gènes sont clusterés dans le cadre mauve. Ca veut dire que ça a un sens de résumer ces gènes à un seul chiffre (médiane) = index meta PCNA.

On peut donc calculer cet index dans chaque tumeur. Et on peut corréler ça à ce qu’on veut, par exemple l’expression de gènes particuliers.

Lorsqu’on normalise les données, on peut avoir la déconvolution meta PCNA 🡺 résidus de la régression >< Chang lui a enlevé les gènes connus comme impliquer dans le cycle cellulaire. La déconvolution n’élimine aucun gène, elle enlève seulement l’expression expliquée par la prolifération PCNA. De plus, la méthode de Chang dépend de la qualité de ces données et de sa transformée de Fourrier.

Quand un gène n’est pas corrélé à meta PCNA, la transformation n’a aucun effet.

Déconvolution = pour chaque gène on calcule la distance entre l’expression du gène et la ligne de régression (échelle logarithmique, au lieu de faire un produit on fait une somme) 🡺 expression des gènes moins l’expression qui est due à PCNA = matrice dont le signal prolifératif est éliminé.

Test sur 3 grosses cohortes du cancer du sein 🡺 quand on enlève le signal prolifératif des différentes signatures pronostiques, la significativité observée disparaît

Même chose fait pour 46 signatures publiées à l’époque (hazard ratio qui quantifie de combien on augmente la survie, en bleu les signatures dans données originales et en rouge avec l’index prolifératif en moins) 🡺 tout ce qui était significatif ne l’est plus, en plus un certain nombre de signatures prétendant être normales déconvoluées n’étaient même pas significatives à la base !!

Une autre façon de voir les choses= chaque point représente un signature et on mesure la PCA sur la signature 🡺 tendance principale de cette signature là. On mesure la corrélation de cette PCA avec l’index de PCNA. On prend la valeur absolue de cette corrélation et l’axe des y = hazard ration. Plus la signature a une composante principale liée à la prolifération, plus la signature sera pronostique

Qu’en est-il des marqueurs multigènes ? Corrélation de chacun des gènes à meta-PCNA 🡺 associé à la survie en rouge et l’ensemble de la totalité des gènes en noir, les gènes pronostic en rouge ont une distribution bimodale. Si on prend des combinaisons de ces gènes, peu de chances qu’ils soient pronostiques.

Si on calcule la corrélation de chaque gène avec meta PCNA index, Vijver 2002 montre 3-5% massivement corrélées et une corrélation diffuse à travers tout le transcriptome. Avec p=0.5, plus de la moitié est corrélé. La prolifération est donc présente partout dans le transcriptome.

Etant donné que la prolifération est partout et est un facteur pronostique, toute signature est donc pronostique, non ?

On a donc cherché une signature non liée au cancer. Ils étudiaent l’effet du rire après le repas sur l’expression des gènes dans les cellules mononucléaires (globules blancs) chez les diabétiques. On leur servait un repas, le groupe traitement lisait des textes comiques et le groupe contrôle lisait des textes scientifiques. On prélevait ensuite des cellules du sang, ARN des globules blancs 🡺 microarray. Cette signature, à la surprise, est pronostique pour le cancer du sein !!

Une autre signature basée sur les fibroblastes extraits dans différentes parties du corps pour déterminer leur localisation en fonction de leur profil transcriptionnel 🡺 pronostique dans le cancer du sein aussi !

Troisième exemple sur la défaite sociale chez les souris où une lignée souffre de défaite sociale (elles ont la gueule cassée en présence de souris WT) 🡺 pronostique aussi !

Preuve qu’il y a un vrai problème avec les signatures !

On utilise la p-value d’association à la survie comme valeur pronostique de la signature (plus facile pour tester des milliers de signatures). On crée des signatures au hasard (100 gènes pris au hasard) et on mesure leur association à la survie dans le cancer du sein. La partie hachurée donne les résultats non significatifs. Le résultat dépend des cohortes, mais en général plus de 50% seront pronostiques.

Si on fait la même chose en appliquant la déconvolution, les signatures publiées sont tout à fait comparables à des signatures tirées au hasard. Tout s’écrase vers des p-values proches de 1. Attention, les signatures sont pronostiques, mais on ne peut pas impliquer que la causalité de la signature est importante dans le cancer>.

Chaque ligne correspond à une signatureplus c’est à gauche, plus c’est pronostique. En jaune représente la distribution de la valeur pronostique de milles signatures aléatoires qui ont le même nombre de gènes que la signature publiée. Plus elle contient de gènes, plus la signature a de chances d’être pronostique. Si on prend une autre hypothèse nulle, la majorité des signatures publiées ne sont pas significatives. Il y en 11 sur 41 qui sont moins bonnes que la moyenne des signatures aléatoires !

L’astuce de Chang de retirer les gènes de la prolifération a largement été utilisée. Le prof a pris la signature et a calculé la corrélation avec meta PCNA avec et sans les gènes liés à la prolifération. Malgré avoir enlevé les gènes, la signature reste très corrélée.

Si on a un bug dans mon programme et que l’expression des gènes est mélangée, on va aussi observer que la déconvolution va casser l’association entre la signature et la survie 🡺 Contrôle négatif nécessaire ! Calcul de l’index sur tumeurs permutées 🡺 pas de problème de pronostique

Il y a peut-être d’autres gènes meta ! 🡺 seulement 2 à 4 % ont le même ou un plus grand pouvoir prédictif de survie (prolifération, immunité, développement du stroma de la tumeur ont été abordés)

Revue pour certaines signatures comme ER 🡺 la prolifération n’est pas un pronostique pour les tumeurs ne présentant pas ER

Les standards statistiques utilisés sont pas adéquats, si on veut affirmer qu’une signature est pronostique (et son mécanisme est associé au cancer), le Cox ne suffit pas.

Gapminder.org 🡺 trace dynamiquement jusqu’à 4 variables simultanément.

La corrélation n’implique pas la causation ! C’est pareil avec les gènes. L’association n’implique pas la causation. Il manque également à ces signatures des contrôles négatifs.

Cette étude a des limites

* On traite ici des populations limitées de patients souffrant du cancer du sein (sous-populations pourraient avoir des signatures pronostiques significatives)
* Profilage transcriptionnel de gros blocs tumoraux, hors une tumeur n’est pas homogène (cellules dans le front invasif sont dans un environnement différent de ceux qui sont dans le cœur de la tumeur) 🡺 intérêt de stratifier les couches de tumeur

C’est un outil valable pour le pronostique, mais l’implication biologique n’est pas valable !

# 16/03/2016

Tout ce qu’on faisait en biologie avant l’avènement des big data. Comment on a transformé les manips traditionnel à des manip cribles. ON a vu comment la faculté de mesurer les phénotypes globalement nous a fait arriver à des approches différentes de l’analyse. En étudiant plusieurs gènes au lieu d’un seul, on a plus de puissance statistique et on diminue le bruit (gain de robustesse). Comparer globalement des phénotypes a également un sens, et peut être fait avec des populations entières. On a vu comment des ensembles de gènes peuvent être des marqueurs a priori 🡺 aucune limite à la combinaison pour leur donner un sens biologique.

Des expériences génomiques peuvent créer des ensembles qui sont utilisés comme marqueur avec d’autres données. Le problème est que ces méthodes ont des sources d’erreur potentielles énormes : test multiple (plus on observe choses, plus des événements aléatoires deviennent probables), malédiction de la multidimensionnalité (on peut toujours trouver un modèle pour expliquer 🡺 résolu par test d’hypothèse nulle (très difficile, car il n’existe pas de vrai bon modèle nul) et séparer les données d’apprentissage et de validation, intéressant de tester l’algorithme en le testant avec des données aléatoires), corrélation des gènes entre eux ce qui est normal car le génome est hautement structuré et il existe un énorme nombre de facteurs confondants (la causalité ne peut être établie que par des expériences, l’analyse bioinformatique pure ne permet pas de déterminer la causalité)

Le fait de rendre les données primaires publiques permet à d’autres personnes de refaire les calculs et d’apporter des critiques à l’analyse. On voudrait faire la même chose avec le code, ce qui est possible pour les microarrays, mais très difficile pour le séquençage à haut-débit. L’échange de code est compliqué.

**Introduction to high-throughput sequencing**

Coût du séquençage vs temps

Consortium a été mis en place avant une compagnie privée. Craig Venter est une espèce d’entrepeneur de génie, mais effrayant car il a breveté des EST. Son idée était de s’approprier à des fins commerciales une partie du génome humain. Il a utilisé le shotgun 🡺 morceau avec recouvrement pour séquencer les petits bouts et les reconstituer après. Le Consortium a utilisé la même méthode en voyant que ça fonctionnait bien et ont battu Craig Venter de quelques jours. Kent, un jeune programmeur, est le héros de l’histoire, il a développé le code en Perl jusqu’à en être blessé.

En 2007- 8, premier séquenceur à haut-débit 🡺 séquençage en quelques semaines et pour quelques milliers de dollars. La technologie génomique progresse plus vite que les ordinateurs (Loi de Moore). De nos jours, on est capable de séquencer un génome pour moins de 1000$. L’ambition est de séquencer le génome de la population entière 🡺 pose un problème éthique car le génome est votre identité biologique

Le séquençage :

* Production d’une librairie génomique : prendre l’ADN, amplifier, séquencer ; des séquences d’ADN reliés à des protéines spécialisées ; … Cela demande de la créativité
* Analyse des données : gros du travail en général. Avec le séquençage à haut-débit, les données sont énormes et plus complexes

Couverture = nombre de séquences à partir d’une librairie

Les données sont très dures à déplacer vu leur taille, et en plus il faut utiliser des super-serveurs pour le calcul.

La science arrive dans un domaine de big science (physique : la construction d’un moteur de particules demande des millions, mais les données sont utilisées par tout le monde), on séquence le génome de milles personnes pour trouver les polymorphismes (présents dans au moins 1% des individus),… et tout le monde peut utiliser ce genre de données. ENCODE essaie de savoir ce que fait l’ADN non codant. Il y a aujourd’hui des catalogues avec des milliers d’ARN non codants grâce au séquençage du génome.

**Example 1**

Idée : notre intestin est peuplé d’un grand nombre de bactéries. On peut extraire leur ADN (des fèces). On appelle ça la méta génomique car on séquence un grand nombre de génomes différents (Beagle avec Darwin). On séquence sans discernement des espèces qu’on a prélevées.

On peut classer les séquences qui sortent par espèces et les classer par famille. Box plots représentent chez différents individus l’abondance des espèces. (génome vs microbiome 🡺 le microbiome est au moins 1000x plus important, plus riche)

On peut aussi grouper ces séquences par fonction 🡺 fonction du microbiome dans l’intestin.

Il appliqué les méthodes PCA à ces microbiomes. Ils ont découvert trois clusters qu’ils ont analogiés aux groupes sanguins. Ces microbiomes sont présents selon l’état de santé en quantité différentes, l’origine,… Un type de microbiome est par exemple associé au diabète de type 2. On pourrait imaginer guérir des maladies en faisant des « greffes de microbiome »

**Exemple 2**

Groupe allemand qui prennent des os de Neandertal, d’en extraire l’ADN et de le séquencer. Ils ont séquencer l’ADN de momie (expérience avec des bouts de viande pourri pour voir jusqu’où on peut récupérer de l’ADN et le reconstituer). Ils ont comparé l’ADN à des ADN de signes et d’humains. Les trois génomes néandertaliens sont les courbes plus basses. Il y a une possibilité de contamination de ces os, donc il y a beaucoup de contrôles négatifs. Dans le génome humain contemporain, il y a 4 à 6 % de génome néandertalien.

En Asie, une nouvelle espèce d’hominidés dans la glace a pu être découverte par une méthode similaire avec des fragments.

En faisant des KO systématiques, ils ont créé un génome capable de faire des taches spécifiques et à le réinjecter dans une bactérie vivante, par Craig Venter. Le but est d’arriver à de la biologie synthétique adaptée à nos besoins.

**Exemple 3**

Papier de 2008. La clonalité des tumeurs est abordable grâce au séquençage à haut-débit. Il étudie des tumeurs du sang, parce que ça vient des lymphocytes qui produisent des Ig, uniques pour chaque cellule. En séquençant ces régions, on peut avoir une idée de la diversité des cellules cancéreuses. Les tumeurs étaient toutes quasi monoclonales, les autres avec un clone dominant et quelques sous-clones. Ils ont pu utiliser la phylogénétique pour avoir la progression du cancer à travers l’évolution des clones.

4 ans plus tard, un groupe anglais a pris une tumeur rénale dont ils ont séquencé plusieurs régions (exomes, partie codante du génome, 2-3% du génome). La tumeur était métastatique. Pour certains fragments ils ont trouvé les mutations (en gris) dans tous les échantillons, certains présents dans certains fragments, uniquement dans des métastases,… Ils ont créé une phylogénie qui permet d’avoir une carte des mutations en fonction de leur localisation. Ça démontre que la diversité des tumeurs solides est énorme

Le même groupe anglais a séquence les génomes complets de 21 tumeurs mammaires et une à une très grande profondeur (couverture de 200x). Ils sont capables à partir de ça, l’histoire clonale de la tumeur. Ce sont des mutations dans le génome, de duplication de fragments de ce génome. Si on une mutation et que ce fragment est dupliqué 🡺 2 fois la mutation, si l’inverse, on aura une seule fois la mutation. Ils font avec ça des calculs de durée de vie entre l’apparition de la première mutation et l’opération chirurgicale.

On présente les exomes de centaines de patients. Les lignes = gènes, colonnes = patients. On cherche les mutations chez les patients et on voit que certaines sont plus courantes que d’autres. On trouve que la plupart de ces mutations sont privées et spécifiques aux patients. Les récurrentes étaient déjà connues avant l’avènement du séquençage à haut-débit. Ça confirme l’idée que le cancer est une maladie génétique et que la diversité est énorme et donc que chaque maladie est unique. On a aussi trouvé des mutations qu’on pensait spécifique à un organe dans d’autres 🡺 espoir pour la thérapie générale, mais un peu défaitiste pour le thérapie ciblée quand on voit toutes les mutations différentes.

Il se trouve que les mutagènes ont des signatures, ils vont préférentiellement toucher certains nucléotides. Ils ont analysé les mutations en regardant quelle substitution, mais en regardant aussi le contexte génomique en 3 et 5’. Ces données ont été input dans une non negative matrix. On voit différents profils mutationnels dans différents patients. Chaque type de cancer a des types de mutation et on peut arriver à raccrocher ces mutations à un agent mutationnel (trouver la cause du cancer est accessible).

On a découvert la chromotripsis, portion de génome limité sont réduits en miettes et puis recollés dans un ordre aléatoire 🡺 mutation ponctuel dès lors capable d’être nocive. La cellule n’est pas capable de réparer son génome et est recollé n’importe comment. Phénomène souvent trouvé chez les enfants.

Une autre découverte, kataegis qui signifie averse, mutations ponctuelles massivement associées dans une portion limitée du génome. Ils ont une signature mutationnelle spécifique, reliée à « apobec » (diversifie anticorps lymphocytes B), elles éditent massivement le génome dans des zones limitées.

**\*-seq**

On peut séquencer n’importe quoi.

**Exemple 4**

Permet de décoder la structure secondaire de l’ARN. Ils se replient et forment une structure par hybridation. Ils explorent tout le transcriptome pour ce genre de structures. On traite l’Arn avec une copie qui digère le monobrin et la seconde avec une RNA doubble brin 🡺 quel ARN est apparié par comparaison des résultats.

On aligne les reads sur le génome et on voit sur quelle partie du transcriptome s’alignent les ARNs digérés par monobrin et on compare avec les zones du transcriptome sur lesquelles s’alignent les ARNS digérés par double brin. On voit un pattern de structure secondaire périodiques (3 nucléotides), on trouve les structures secondaires aux extrémités 3 et 5’.

La structure secondaire est affectée par des polymorphismes humains, notamment dans un ARN de histocompatibilité (présentation antigène aux lymphocytes T)

**Exemple 6**

Séquençage de cellules individuelles. Les cancers par exemple ont une grande diversité, on pourrait avoir des transcriptomes différents 🡺 vaisseaux sanguins, cellules immunitaires pourraient aussi se trouver dans la tumeur et brouiller les résultats. L’idée est d’isoler les séquences une par une (FACS), associer un tag (barcode) à la séquence et séquencer alors dans le désordre. Tous les ARN avec le même tag viennent de la même cellule. On perd en couverture, mais on a suffisamment pour déconvoluer les types cellulaires présents.

Transcriptome entre les cellules individuelles, calcul de la corrélation entre chaque paire de cellules. Il y a des groupes de cellules corrélées entre elles (rouge). Par projection, on observe également des clusters de cellules. Ces clusters correspondent à des types cellulaires connus. Les cellules dendritiques (nébuleuses) sont sous-divisées en 2 groupes dont on ignorait l’existence jusqu’à maintenant. On peut soumettre ces cellules à des perturbations et voir quels gènes sont perturbés dans chaque type de cellule (réponse peut être différente selon le type cellulaire).

**Alignments of RNA-seq short reads**

Données issues du transcriptome.

Transcrits déjà épissés sont séquencés durant le RNA-seq.

Les reads sont des fragments d’un transcrit, d’environ 150pb, séparés par une distance 200pb. On veut revenir vers le génome. L’intérêt du séquençage est qu’on peut partir avec zéro idée sur le transcriptome et réaligner ce qu’on a découvert sur le génome. On a découvert que l’épissage alternatif est la règle et non l’exception. Le problème est que le transcrit mature n’est pas le reflet du gène dû à l’absence d’introns.

Le transcriptome est donc plus riche que le microarray.

Un gène où on a un alignement similaire, avec un exon qui est épissé alternativement. Cet exon n’est présent que dans certaines séquences. Le RNA-seq permet donc de quantifier l’épissage alternatif et la présence de chaque transcrit.

Avec les microarray, on a l’expression à l’échelle de l’expression des gènes. On peut le faire aussi avec le RNA-seq, mais on peut aussi descendre à l’échelle des exons et même des paires de bases. On a donc une échelle du transcriptome à toutes les échelles possibles, de l’expression globale à la paire de base. On est capable de voir les polymorphismes, l’activité d’enzymes agissant sur les nucléotides comme ADAR.

On peut assimiler l’alignement comme pré-traitement qu’on fait avec le microarray. L’alignement fait, on peut étudier l’expression 3’ des gènes, l’édition de l’ARN (paire de bases modifiées après transcription), les gènes fusions (dans les cancers), les virus qui affectent l’humain (inséré dans le génome)…

Deux stratégies :

* Organisme inconnu 🡺 de novo assembling : shotgun, séquençage et on recolle
* On connait le génome 🡺 aligner sur le génome

Alignement non trivial 🡺 le génome fait 3 milliards de paire de base (bcp de copier-coller au cours de l’évolution, bcp de duplication et de mutations de gènes déjà présents) Avec des reads qui font 150pb, difficile d’associer des séquences très courtes de façon univoque à une portion unique du génome.

Les programmes d’alignements n’étaient pas suffisamment rapides pour le RNA-seq. STAR permet d’aligner un transcriptome sur 8 processeurs en moins d’une heure (mais bcp de mémoire)

**Preprocess the genome : reference indexing**

On a indexé le génome pour être capable de faire des recherches systématiques très rapides.

On cherche dans le génome des hits de la séquence qu’on veut aligner. On cherche ensuite à étendre l’alignement autour de ce hit pour maximiser le score. Cette méthode n’est pas adéquate, car pour les retrovirus endogènes ont des séquences très similaires et donc rapporter le meilleur match n’est pas ce qu’on recherche nécessairement, on cherche plutôt dan combien d’endroits différents c’est rapporté,…

Problème des pseudogènes : transcrits en ARNm, qui ont été retrotranscrits et réinsérés dans le génome, dizaines de milliers dans le génome.

Le pseudogène a en général muté (trèèèès muté, on peut les masquer grâce à cette « densité de mutation »), mais le problème est que l’alignement correct a beaucoup de gaps. Donc l’aligneur va plutôt avoir tendance à assigner le read au pseudogène plutôt que le gène d’où le transcrit provient réellement.

On voit que certains reads ont des séquences avec un tas de mutations, donc on a des problèmes d’alignements aux jonctions exoniques, ce qui provient d’artéfacts (par exemple parce que l’exon suivant et l’intron ont une séquence similaire, ce qui va faire que l’aligneur va plutôt tolérer des mismatchs plutôt que d’ouvrir un gap pour faire un alignement parfait avec l’exon beaucoup plus loin). Une solution serait de créer une database des jonctions exoniques pour éviter ces erreurs.

**Cancer genome Atlas**

Séquencer 500 tumeurs des 20 types de cancer différents, objectif accompli et même dépassé !

Pour chacune de ces tumeurs, on a l’exome (sain vs tumeur 🡺 permet de voir les mutations), nombre de copies des gènes,…

Premier niveau : données brutes

Second : données normalisées (normalisation ou alignées sur le génome de référence)

Troisième : niveau d’expression des gènes qu’on a extrait des alignements RNA-seq, matrice d’expression,…

Quartre : clustering hiérarchique, décomposition NMF,…

Privilèges d’accès pour les niveaux 1 et 2 car les séquences sont personnelles à un patient (garanti au patient de l’anonymat des données), pour y avoir accès il faut remplir des formulaires au NIH qui seront revues par un comité.

Firehose : parce que ça produit beaucoup de quantités, vomit de l’information génomique au kilomètre

Corrélation systématique entre mutation et données cliniques ☺

Fait tourner l’algorithme NMF sur les profils RNA-seq, miRNA-seq et copy number. Rapporte les clusters optimaux dans TCGA 🡺 chaque échantillon peut être accroché à un cluster particulier et les gènes qui sont fortement associés à chaque cluster

Clusters en fonction des données cliniques, méthylation,… Beaucoup de tissus fibrotiques avec la mutation braf, du coup la méthylation observée pourrait être liée à la plus grande présence de fibroblastes.

Matrice de mutation pour plusieurs types de cancers, on observe les mutations récurrentes dans les différents types. On a une table avec plein de gènes qu’ils ont organisés selon leur fonction/voie, et on peut voir la prévalence de chaque mutation dans tous les cancers humains.

Nombre de read qui porte la mutation/nombre de read total = 0 pas de mutation, 50% hétérozygote, 1 homozygote. Dans le cancer, ces résultats sont faussés 🡺 permet de voir les mutations clonales en évaluant le ratio allélique

**SIgnatures**

Ils ont compté les mutations dans chaque type de cancer et ont ordonné les cancers avec le plus de mutations et le moins de mutations.

Plusieurs signatures ont été trouvés grâce à NMF (C à gauche, T à droite,…) Vingtaine de type de signature mutationelle. Ces signatures sont explicables par des mutagènes connues (tabac,… défaut de réparation de l’ADN (courant dans cancer colo-rectal)), mais la plupart ne sont pas connues.

On peut examiner la charge de chaque signature aux mutations du patient (pour un cancer du sein c’est APOBEC, pour un autre c’est l’uv,…) 🡺 étude sur initiation du cancer

**Projet**

Le vieillissement n’est plus inévitable, on a réussi à moduler la durée de vie de certains organismes (par exemple le ver plat). Chez les organismes plus proches de nous, il est possible, en les soumettant à un régime hypo-caloriques, d’allonger leur durée de vie. On observe qu’avec un traitement à la rapamycine donné à la moitié de l’espérance de vie, on allonge cette dernière.

Syndrome de progérie (vieillissement accéléré) 🡺 grand nombre de mutations

Au fur et à mesure qu’on vieillit, le risque de cancer augmente. Un moyen de limiter ce risque serait de limiter la division cellulaire…. Donc le vieillissement nous protège car il ralentit cette division.

Est-ce que le fait de vieillir moins vote veut dire qu’on va vivre plus longtemps ?

Un bioinformaticien prétend avoir trouvé une horloge biologique mesurable grâce la méthylation de leur ADN. (300 marqueurs épigénétiques qui permet de donné l’âge chronologique)

Des gens pensent que c’est l’accumulation de cellules sénescentes qui apportent le vieillissement, mais l’article apporte une réponse négative à cette hypothèse.

On va avoir des données d’expression TGCA et les données cliniques du patient.

1. Se procurer les données

Fichier clinique avec noms échantillons et variables cliniques (âge chronologique très important)

1. On reçoit un fichier correspondant à un certain nombre de patients avec l’âge méthylation
2. Evaluer âge méthylation dans tissu normal et dans tumeur (parfois juste tumeur)

* Que vaut l’âge de méthylation ?
  + Scatter plot : Méthylation vs chronologique (ligne droite si corrélé parfaitement)
  + DNAm Tumeur vs sain (quand on a les deux)
  + Accélération de l’âge : DNAm tumeur/(chronologique ou DNAm normal) ou normal - tumeur (distribution de l’âge, préféré la distribution normale d’un point de vue statistique)

Utiliser le barcode pour faire une analyse de l’explication de l’échantillon

* Variables cliniques corrélées au DNAm ?
  + Utiliser les tests appropriés !
  + DNA mage associé à la survie ?
* Caractériser les gènes associés à cet âge de méthylation et accélération
  + GSEA : algorithme original dans R, application java, Genepattern avec interface graphique (attention, transformer les données en log de base 2 ! Valeurs doivent être maximum 20 log de base 2 de … + 1)
  + C2 :CP : gene sets fonctionnels, canonical pathways (Biocarta, KEGG, Reactome)
  + Sam, attention aux différentes options : pour traiter spécifiquement RNA-seq data
* Lien entre DNAm et mutations ? (vieillissement apporte mutations, donc le nombre de mutations avec le cancer serait-elle la raison pour l’accélération du DNAm ?)
* Lien entre DNAm et le nombre de copies ? (variations de l’intensité lumineuse consécutive dans le génome, on soustrait ce qui se passe dans le tissu adjacent du cancer et on se retrouve avec les mutations spécifiques au cancer)

Bien spécifier quels paramètres ont été utilisés, la version du programme,…

Les différentes questions clairement dans le rapport, la démarche bien expliquée, ainsi que les résultats 🡺 concis et clair !!