INSTITUTE FOR ADVANCED BIOSCIENCES (IAB)

PLATEFORME EPIMED

Ekaterina Flin

Base de données EpiMed

Description du schéma de la base de données

Guide de développeur

Date de création : 20 septembre 2016

Date de modification : 23 septembre 2016

Table des matières

1 Introduction

La plateforme EpiMed (Epigénétique Médicale et Bioinformatique) est un réseau collaboratif de chercheurs, médecins et ingénieurs qui travaillent dans le domaine d'épigénétique. L'activité EpiMed a été initiée dans le but de faciliter une recherche translationnelle dans le domaine de l'épigénétique entre les équipes de recherche fondamentale de l'Institut pour l'Avancée des Biosciences (IAB) et les équipes médicales du Centre Hospitalier Universitaire Grenoble Alpes (CHU), ou d'autres centres hospitalo-universitaires (par exemple le Ruijin hospital à Shanghai). Aujourd'hui, EpiMed assure le développement de la bioinformatique pour l'ensemble des activités de l'IAB. L'activité EpiMed repose largement sur sa capacité d'analyse des données à grande échelle, essentielle à la compréhension de l'épigénome. Un effort important a été engagé pour la constitution d'une cellule d'analyses bioinformatiques et l'organisation d'une base de données interactive associant les données "omics" et cliniques et/ou biologiques, utilisable par les scientifiques et les médecins impliqués dans les projets translationnels EpiMed.

2 Données gérées par EpiMed

Les bases de données EpiMed visent à centraliser et à homogénéiser les données provenant de différentes sources pour faciliter leur traitement. La plateforme EpiMed gère les familles de données suivantes :

2.1 Données omics

Sous le terme générique "omics" s'apparentent de nouvelles technologies faisant référence à "genomics", "transcriptomics", "proteomics", etc. Ces outils permettent une analyse précoce et spécifique des effets d'une substance chimique sur l'organisme. Les données omics sont généralement issues des mesures sur les échantillons d'ADN.

On travaille couramment avec le transcriptome (mesure du niveau d'expressions des gènes). Le transcriptome est l'ensemble des ARN issus de la transcription du génome. L'analyse transcriptomique peut caractériser le transcriptome d'un tissu particulier, d'un type cellulaire, ou comparer les transcriptomes entre différentes conditions expérimentales.

Une autre technique souvent utilisée est le méthylome. La méthylation de l'ADN est un processus épigénétique dans lequel certaines bases nucléotidiques (cytosine) peuvent

être modifiées par l'addition d'un groupement méthyle. La méthylation de l'ADN agit comme un « patron » qui conditionne l'expression des gènes dans chaque cellule.

La méthylation et la transcription du génome peuvent être analysées à différents niveaux de résolution :

— au niveau du génome complet par séquençage NGS d'ADN génomique

— au niveau de régions ciblées par la technique de microarrays

D'autres mesures peuvent compléter ces techniques.

Les données omics sont structurées. Elles peuvent, par contre, être très volumineuses, notamment dans le cas d'un séquençage complet du génome.

2.2 Données descriptives des gènes

Ces données correspondent à la description des gènes dans différentes nomenclatures standards (HGNC, NCBI, UniProt, UCSC, Ensembl, UniGene, etc.). Les données sur les gènes sont structurées, avec une structure relativement complexe. Elles sont aussi assez volumineuses.

2.3 Données cliniques

Les données cliniques décrivent les échantillons utilisés dans nos études. Dans le cas général, ces données ne sont pas structurées, elles ont une petite taille et peuvent être extrêmement variables.

3 Organisation des bases de données

Les données descriptives des gènes et les données cliniques sont stockées dans des bases de données EpiMed. Les données omics étant très volumineuses, elles sont stockées directement sur le système de fichiers dans leurs formats natifs (ex. format CEL pour le transcriptome). La base de données contient uniquement les pointeurs vers ces fichiers.

Actuellement, EpiMed utilise le système de gestion de bases de données (SGBD) relationnel PostgreSQL accessible sur le serveur `epimed-db.imag.fr`.

Le SGBD définit deux possibilités de gestion de données : database et schéma. Une database peut contenir plusieurs schémas. Les databases sont isolées. C'est-à-dire qu'il n'est pas possible de croiser les données de différentes databases dans la même requête.

En revanche, on peut croiser les données qui se trouvent dans des schémas différents de la même database.

EpiMed définit deux databases :

— **epimed_prod** : base de données de production

— **epimed_dev** : base de données de développement

Chaque database contient plusieurs schémas. Dans ce document on décrit le schéma nommé **hs** (de “homo sapiens”) qui contient les données sur les gènes du génome humain.

4 Modélisation des concepts métiers

Dans la base de données EpiMed on distinguent, par convention, deux groupes de données :

— Les données en rapport avec les mesures omics : plateformes de mesure, description des gènes et des protéines, liens vers les fichiers bruts ;

— Les données en rapport avec les enregistrements cliniques : description des échantillons, diagnostic, données biopathologiques, classifications, dictionnaires ontologiques. On appelle également ces données `experimental grouping`.

Pour le premier groupe (données omics), on préfixe les tables correspondantes par **om_**. Les tables qui décrivent les données du deuxième groupe (données cliniques) sont préfixées par **cl_**. Finalement, les tables techniques nécessaires pour le fonctionnement et le suivi du des mises à jour de données sont préfixées par **tx_**.

4.1 Gène

Un gène, en génétique, est une unité de base d’hérédité qui en principe prédétermine un trait précis de la forme d’un organisme vivant. Au niveau physique, un gène est un fragment (ou Locus) déterminé d’une séquence d’ADN qui paramètre la synthèse d’un ARN donné, en prédéfinissant sa structure et, donc, celle de l’éventuelle protéine ou de l’éventuel polypeptide synthétisés à partir de cet ARN. Sur la molécule d’ADN, un gène est caractérisé à la fois par sa position et par l’ordre de ses bases azotées. Actuellement, on compte environ 25 000 gènes pour l’homme. Chaque gène est décrit par des noms et des numéros ainsi que des alias (synonymes) selon différentes nomenclatures. Notamment, la nomenclature des gènes approuvés est établie par l’organisation HGNC (HUGO Gene No-

menclature Committee) qui attribue à chaque gène un identifiant, un nom et un symbole (Figure 1).

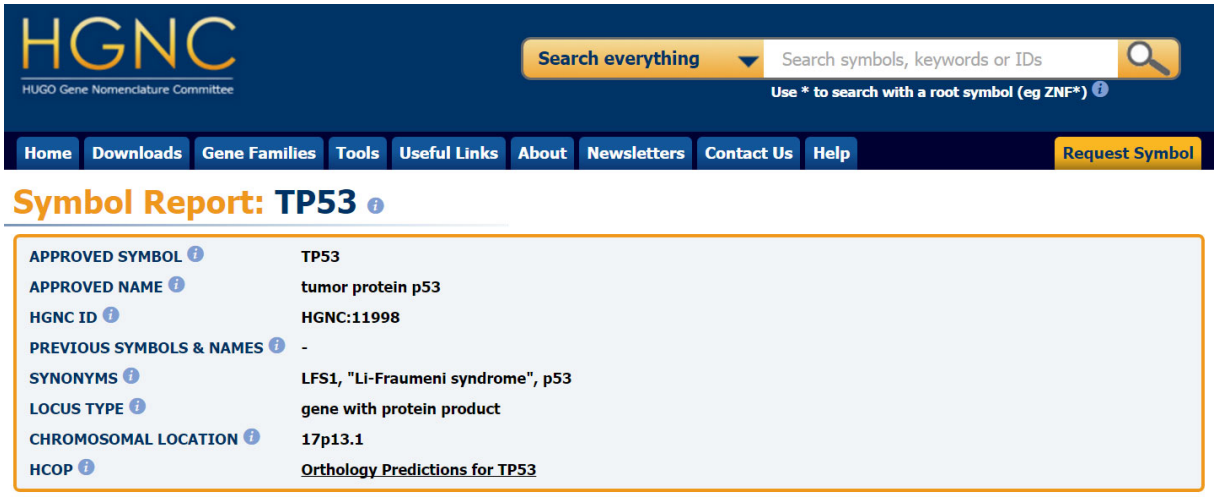

FIGURE 1 – *Description du gène TP53 selon la nomenclature HGNC.*

Il existe d'autres organisations qui établissent les nomenclatures sur les gènes. NCBI est un organisme qui décrit les gènes approuvés et non approuvés (Figure 2). La base de gènes du NCBI est plus riche que celle de l'HGNC. D'autres nomenclatures sont disponibles sur UniGene, Ensembl, Vega, UCSC, etc.

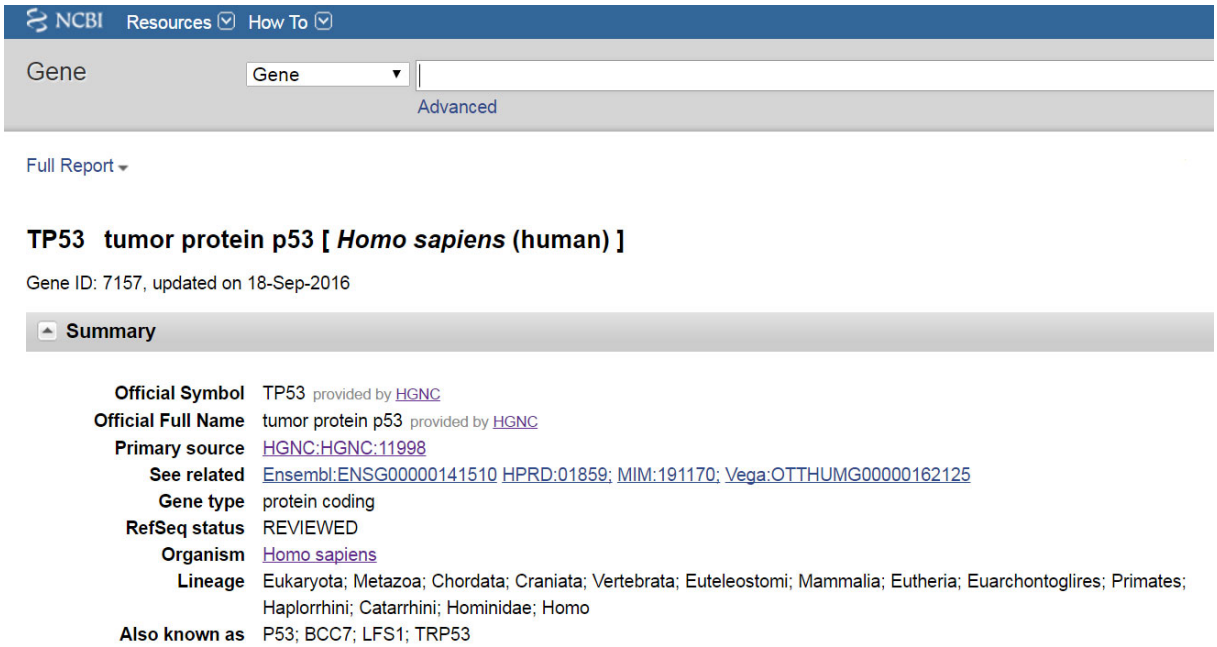

FIGURE 2 – *Description du gène TP53 selon la nomenclature NCBI.*

EpiMed utilise principalement les nomenclatures NCBI et HGNC pour décrire les

gènes. Celle de NCBI est considérée comme la nomenclature principale car elle couvre le maximum des gènes sur lesquels on travaille. La nomenclature HGNC apporte des informations supplémentaires pour les gènes approuvés.

Dans la base de données EpiMed, un gène est modélisé par la table `om_gene`. La signification de chaque attribut de cette table est décrite dans l'annexe 5.2, page 7.

5 Annexe

5.1 Table om_assembly

Contient les informations sur différents génomes de référence.

id_assembly	Identifiant unique du génome de référence selon la nomenclature standard du Genome Reference Consortium (GRC). Clé primaire. Exemple : `GRCh38`.
id_organism	Identifiant de l'organisme. Clé étrangère vers la table `om_organism`.
ucsc_code	Code UCSC correspondant à ce génome de référence. Nécessaire pour connaitre le nom de la base de données correspondante dans la nomeclature UCSC. Exemple : `hg38`.
name	Nom du génome de référence. Exemple : `Genome Reference Consortium Humain Build 38`.

Table 1 – Attributs d'un génome de référence

5.2 Table om_gene

Contient la description des gènes.

id_gene	Identifiant unique du gène selon la nomenclature NCBI (appelé `Gene ID` ou numéro `Entrez`). Clé primaire. Exemple : `7157` pour le gène TP53.
gene_symbol	Le symbole principal (nom officiel) du gène. Exemple : `TP53`
title	Titre du gène. Exemple : `tumor protein p53`
location	Position du gène sur le chromosome. Exemple : `17p13.1`
status	Statut actuel du gène. Exemple : `Approved`
type	Type du gène. Exemple : `gene with protein product`
date_modified	Date de création ou de modification du gène dans les bases sources (NCBI, HGNC). Exemple : `2016-08-25`
last_update	Date de la dernière synchronisation avec les bases sources (NCBI, HGNC) pour mettre à jour l'information sur le gène. Exemple : `2016-08-31`

Table 2 – Attributs d'un gène

5.3 Table om_gene_alias

Contient tous les alias (synonymes) pour chaque gène (Figure 3).

id_gene	Identifiant du gène pour lequel la table contient les alias du nom. Fait partie de la clé primaire. Exemple : `7157` pour le gène TP53.
alias	Alias (ou synonyme) du gène. Fait partie de la clé primaire. Exemple : `p53`.
database	La base de données source pour cet alias. Pour les synonymes et les symboles antérieurs on utilise la database par défaut `alias`. Exemple : — `ensembl` pour les identifiants dans la base Ensembl — `ucsc` pour les identifiants UCSC — `hgnc` pour les identifiants HGNC — `alias` pour les synonymes et tous les symboles antérieurs

TABLE 3 – Attributs des alias d'un gène

id_gene [PK] integer	alias [PK] character varying(50)	database character varying(50)
7157	ENSG00000141510	ensembl
7157	HGNC:11998	hgnc
7157	LFS1	alias
7157	p53	alias
7157	uc002gij.3	ucsc
7157	uc060aur.1	ucsc

FIGURE 3 – *Présentation des alias du gène TP53 (Gene ID 7157) dans la table* `om_gene_alias`.

5.4 Table om_gene_history

La table `om_gene_history` contient l'historique de modification des gènes au cours du temps. Cette table permet de créer une filiation entre deux gènes avec les identifiants différents, par exemple, dans le cas où un gène a été remplacé par un autre (Figure 4).

id_gene [PK] integer	gene_symbol character varying(50)	title character varying(255)	location character varying(100)	status character varying(100)	type character varying(100)	date_modified date	last_update date
244	ANXA8	annexin A8		Replaced by 728113			2016-08-31
728113	ANXA8L1	annexin A8-like 1	10q11.22	Approved	gene with protein product	2015-09-11	2016-08-31

FIGURE 4 – *Présentation de deux gènes (ANXA8 et ANXA8L1) avec des identifiants distincts où un des gènes (ANXA8) a été remplacé par un autre (ANXA8L1).*

Le schéma de la table `om_gene_history` peut encore évoluer pour intégrer d'autres types de relations entre les gènes, autres que le remplacement.

id_gene_before	Identifiant du gène source (gène obsolète qui a été remplacé). Fait partie de la clé primaire. Exemple : `244` dans la situation où le gène ANXA8 (Gene ID 244) a été remplacé par le gène ANXA8L1 (Gene ID 728113).
id_gene_after	Identifiant du gène destination (gène actuel pour lequel on a remplacé un gène obsolète). Fait partie de la clé primaire. Exemple : `728113` dans la situation où le gène ANXA8 (Gene ID 244) a été remplacé par le gène ANXA8L1 (Gene ID 728113).

TABLE 4 – Attributs d'une filiation entre les gènes

5.5 Table om_gene_position

Cette table décrit une position particulière sur un génome de référence. Cette position peut correspondre à un gène particulier ou pas. Les informations dans cette table proviennent essentiellement des bases UCSC et Ensembl.

id_position	Identifiant unique de la position. Clé primaire.
id_ucsc	Identifiant UCSC correspondant à cette position.
id_ensembl	Identifiant Ensembl correspondant à cette position.
id_assembly	Identifiant du génome de référence. Clé étrangère vers la table `om_assembly`.
id_gene	Identifiant du gène qui se trouve à cette position, s'il y a un gène. Une position sur le génome ne correspond pas toujours à un gène. Clé étrangère vers la table `om_gene`.
chrom	Chromosome
strand	Strand (sens de lecture sur le génome)
tx_start	Début de la région de transcription
tx_end	Fin de la région de transcription
cds_start	Début de la région complète décrite par cette position sur le génome
cds_end	Fin de la région complète décrite par cette position sur le génome
exon_count	Nombre d'exons

TABLE 5 – Attributs d'une position sur un génome de référence

5.6 Table om_organism

Décrit l'organisme auquel correspond le schéma de la base de données. Pour le schéma **`hs`** un seul organisme est décrit : homo sapiens. Cette table contient donc une seule ligne.

id_organism	Identifiant unique de l'organisme selon la nomenclature NCBI. Clé primaire. Exemple : `9606`.
name	Nom de l'organisme. Exemple : `Homo sapiens`.

TABLE 6 – Attributs d'un organisme

5.7 Table om_probe_{*identifiant de la plateforme*}

Il s'agit d'une série de tables nommés `om_probe_`{*identifiant de la plateforme*}. Une table par plateforme. Pour identifier la plateforme on utilise l'identifiant unique NCBI. Par exemple, `gpl570` (en minuscules) pour la plateforme Affymetrix Human Genome U133 Plus 2.0 Array. Chacune de ces tables contient une liste de sondes de la plateforme correspondante.

id_probe	Identifiant unique de la sonde (appelée également `probe` ou `probeset`, en anglais). Clé primaire. Exemple : `1007_s_at`.

TABLE 7 – Attributs d'une plateforme

5.8 Table om_gp_{*identifiant de la plateforme*}

Il s'agit d'une série de tables nommés `om_gp_`{*identifiant de la plateforme*}. Pour chaque table `om_probe_`{*identifiant de la plateforme*} (voir l'annexe 5.7) on crée une table de liaison `om_gp_`{*identifiant de la plateforme*} qui contient les correspondances entre les sondes de la plateforme sélectionnée et les gènes.

5.9 Table om_protein

La table `om_protein` contient une liste de protéines principales et leurs correspondances avec les gènes.

id_protein	Identifiant unique de la protéine selon la nomenclature UniProt. Clé primaire. Exemple : `P04637`.
id_gene	Identifiant du gène correspondant. Clé étrangère vers la table `om_gene`. Exemple : `7157` pour le gène TP53.

TABLE 8 – Attributs d'une protéine

5.10 Table om_protein_sequence

Contient les séquences de protéines.

id_sequence	Identifiant unique de la séquence de protéine selon la nomenclature UniProt. Clé primaire. Exemple : `P04637-3` pour l'isoforme 3 de la protéine P04637.
id_protein	Identifiant de la protéine correspondante. Clé étrangère vers la table `om_protein`. Exemple : `P04637`.
meta	Une ligne des métadonnées telle qu'elle est définie par le format FASTA dans la base UniProt. Exemple : `>sp|P04637-3|P53_HUMAN Isoform 3 of Cellular tumor antigen p53 OS=Homo sapiens GN=TP53`.
length	Longueur de la séquence, c'est à dire, le nombre de caractères dans la séquence. Exemple : `346`.
pi	Point isoélectrique (pI), appelé également potentiel hydrogène isoélectrique (pHi). Le pHi ou pI d'une molécule est défini comme le pH (potentiel hydrogène) pour lequel la charge globale de cette molécule est nulle ou, autrement dit, le pH pour lequel la molécule est électriquement neutre. Exemple : `5.64`.
average_mass	Masse moyenne de la molécule (Da). Exemple : `38500.54`.
average_mass	Masse mono-isotopique de la molécule (Da). Exemple : `38475.72`.
canonical	Un boolean qui définit si la séquence est canonique pour cette protéine. Sinon, il s'agit d'une isoforme. Exemple : `false`.
sequence	Séquence complète des acides aminées. Exemple : `MEEPQSDPSVEPPLSQETFSDLWKLLPENNVL...`
last_update	Date de la dernière mise à jour. Exemple : `2016-05-20`.

Table 9 – Attributs d'une séquence de protéine

5.11 Table tx_log

C'est une table technique qui contient les logs des scripts. Il s'agit typiquement des scripts qui mettent à jour la base de données.

last_activity	Date et heure du log. Clé primaire. Exemple : `2016-07-01 23:40:31.236`.
module	Module logiciel qui écrit le log. Exemple : `module.update.UpdateGenes`.
status	Statut actuel du processus. Exemple : `in progress`.
comment	Commentaire. Exemple : `Gene annotation has changed for 4301: current MLLT4 (Approved 2016-06-16), updated AFDN (Approved 2016-06-28)`.

TABLE 10 – Attributs des logs

5.12 Vue view_log

Une vue sur la table `tx_log` qui trie les entrées de la plus récente vers la plus ancienne. La dernière action est ainsi affichée sur la première ligne.

6 Ressources externes

Ensembl Ensembl genome browser

`http://www.ensembl.org/index.html`

HGNC HUGO Gene Nomenclature Committee

`http://www.genenames.org/`

NCBI National Center for Biotechnology Information

`http://www.ncbi.nlm.nih.gov/`

UCSC UCSC Genome Browser

`https://genome.ucsc.edu/`

UniGene `http://www.ncbi.nlm.nih.gov/unigene`

UniProt Universal Protein Resource

`http://www.uniprot.org/`