Introduction to TCGA data

Analyzing TCGA data in the cloud. February 28th 2018

International Agency for Research on Cancer
World Health Organization

Data location

[Genomic Data Commons (GDC) data portal]

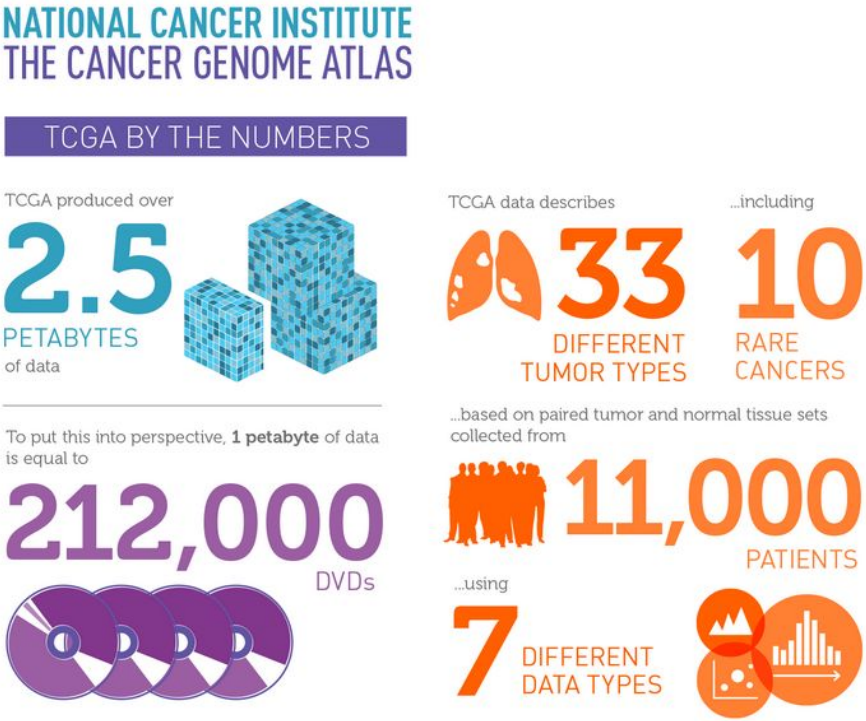
- A data sharing platform launched by the NCI
- Harmonized data using GDC Bioinformatics Pipelines

[GDC legacy archive]

- Unharmonized data
- Previously hosted in CGHub and the TCGA data portal

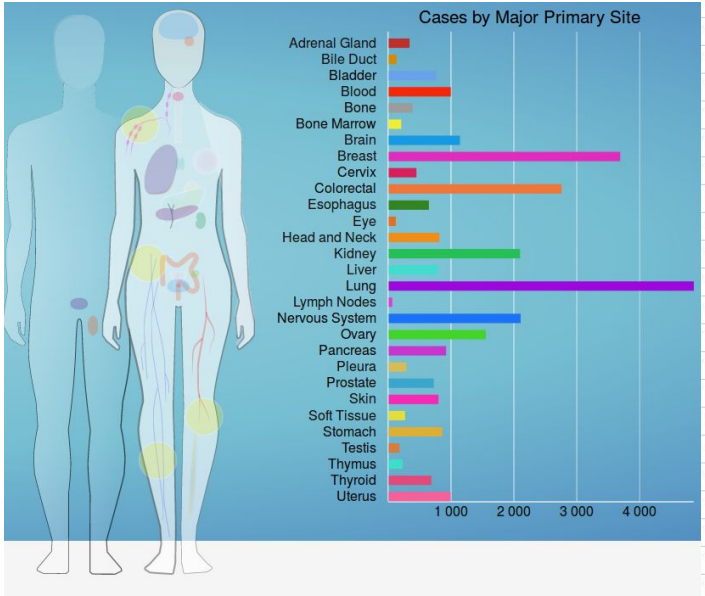
[GDC release notes]

- Data release 1.0 to 10.1
- For each release: new updates, bug fixes since last release and known issues and workarounds



Projects

- 33 different cancer types
- grouped cohorts: COADREAD, GBMLGG, KIPAN, STES
- FFPE cohort
- TARGET projects (childhood cancers)

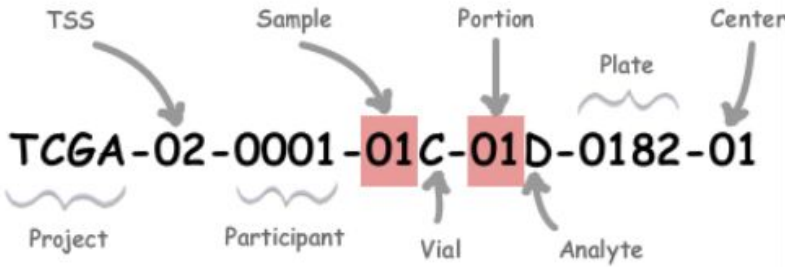


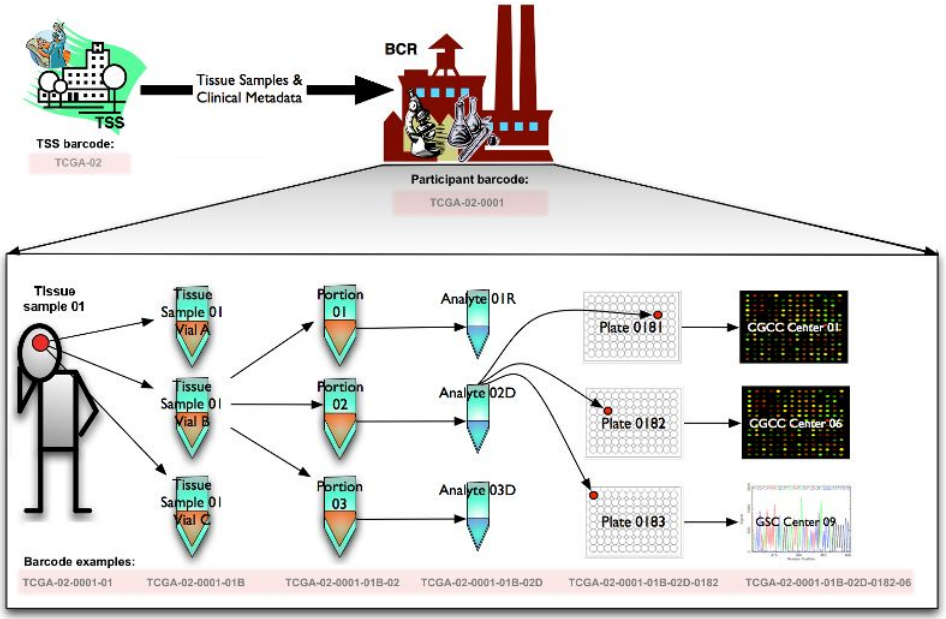
Project	Disease.Type	Primary.Site
TCGA-PCPG	Pheochromocytoma and Paraganglioma	Adrenal Gland
TCGA-ACC	Adrenocortical Carcinoma	Adrenal Gland
TCGA-CHOL	Cholangiocarcinoma	Bile Duct
TCGA-BLCA	Bladder Urothelial Carcinoma	Bladder
TCGA-LAML	Acute Myeloid Leukemia	Bone Marrow
TCGA-GBM	Glioblastoma Multiforme	Brain
TCGA-LGG	Brain Lower Grade Glioma	Brain
TCGA-BRCA	Breast Invasive Carcinoma	Breast
TCGA-CESC	Cervical Squamous Cell Carcinoma and Endocervical...	Cervix
TCGA-COAD	Colon Adenocarcinoma	Colorectal
TCGA-READ	Rectum Adenocarcinoma	Colorectal
TCGA-ESCA	Esophageal Carcinoma	Esophagus
TCGA-UVM	Uveal Melanoma	Eye
TCGA-HNSC	Head and Neck Squamous Cell Carcinoma	Head and Neck
TCGA-KIRC	Kidney Renal Clear Cell Carcinoma	Kidney
TCGA-KIRP	Kidney Renal Papillary Cell Carcinoma	Kidney
TCGA-KICH	Kidney Chromophobe	Kidney
TCGA-LIHC	Liver Hepatocellular Carcinoma	Liver
TCGA-LUAD	Lung Adenocarcinoma	Lung
TCGA-LUSC	Lung Squamous Cell Carcinoma	Lung
TCGA-DLBC	Lymphoid Neoplasm Diffuse Large B-cell Lymphoma	Lymph Nodes
TCGA-OV	Ovarian Serous Cystadenocarcinoma	Ovary
TCGA-PAAD	Pancreatic Adenocarcinoma	Pancreas
TCGA-MESO	Mesothelioma	Pleura
TCGA-PRAD	Prostate Adenocarcinoma	Prostate
TCGA-SKCM	Skin Cutaneous Melanoma	Skin
TCGA-SARC	Sarcoma	Soft Tissue
TCGA-STAD	Stomach Adenocarcinoma	Stomach
TCGA-TGCT	Testicular Germ Cell Tumors	Testis
TCGA-THYM	Thymoma	Thymus
TCGA-THCA	Thyroid Carcinoma	Thyroid
TCGA-UCEC	Uterine Corpus Endometrial Carcinoma	Uterus
TCGA-UCS	Uterine Carcinosarcoma	Uterus

All the TCGA publications are located here

Participants

TCGA barcodes

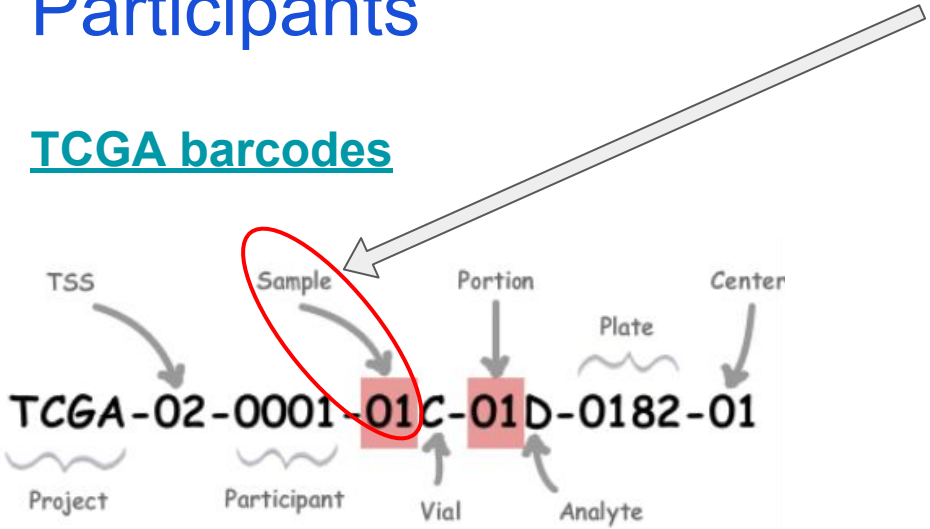




- Vial: portion of a sample
- Portion: 100-200 mg section of a vial
- Analyte: molecular specimen extracted from a portion (e.g total RNA)

Participants

TCGA barcodes



TCGA code tables

Code	Definition	Short Letter Code
01	Primary Solid Tumor	TP
02	Recurrent Solid Tumor	TR
03	Primary Blood Derived Cancer - Peripheral Blood	TB
04	Recurrent Blood Derived Cancer - Bone Marrow	TRBM
05	Additional - New Primary	TAP
06	Metastatic	TM
07	Additional Metastatic	TAM
08	Human Tumor Original Cells	THOC
09	Primary Blood Derived Cancer - Bone Marrow	TBM
10	Blood Derived Normal	NB
11	Solid Tissue Normal	NT
12	Buccal Cell Normal	NBC
13	EBV Immortalized Normal	NEBV
14	Bone Marrow Normal	NBM
15	sample type 15	15SH

Participants

Universally unique identifier: UUID

- Non-human readable ID
- Example: ebf3e73f-41a0-4ca5-b608-fe1c629e16de
- Used to avoid errors
- Each barcode is associated to one UUID
- Each file is associated to one UUID
- Each participant is associated to one UUID

Data category

- Clinical data (public files only)
- Raw sequencing Data (protected files only)
- Simple Nucleotide Variation
- Copy Number Variation
- Transcriptome Profiling
- DNA Methylation (public files only)
- Biospecimen (public files only)

Depending on the data category files are protected or public. Public files maintaining anonymity.

Number of samples per data category

Project	Disease.Type	Primary.Site	Cases	Seq	Exp	SNV	CNV	Meth	Clinical
TCGA-BRCA	Breast Invasive Carcinoma	Breast	1 098	1 098	1 097	1 044	1 096	1 095	1 097
TCGA-GBM	Glioblastoma Multiforme	Brain	617	406	166	396	593	423	596
TCGA-OV	Ovarian Serous Cystadenocarcinoma	Ovary	608	575	492	443	573	602	587
TCGA-LUAD	Lung Adenocarcinoma	Lung	585	582	519	569	518	579	522
TCGA-UCEC	Uterine Corpus Endometrial Carcinoma	Uterus	560	559	559	542	547	559	548
TCGA-KIRC	Kidney Renal Clear Cell Carcinoma	Kidney	537	535	534	339	532	533	537
TCGA-HNSC	Head and Neck Squamous Cell Carcinoma	Head and Neck	528	528	528	510	521	528	528
TCGA-LGG	Brain Lower Grade Glioma	Brain	516	516	516	513	514	516	515
TCGA-THCA	Thyroid Carcinoma	Thyroid	507	507	507	496	505	507	507
TCGA-LUSC	Lung Squamous Cell Carcinoma	Lung	504	504	504	497	504	503	504
TCGA-PRAD	Prostate Adenocarcinoma	Prostate	500	498	498	498	498	498	500
TCGA-SKCM	Skin Cutaneous Melanoma	Skin	470	470	469	470	470	470	470
TCGA-COAD	Colon Adenocarcinoma	Colorectal	461	460	459	433	458	458	459
TCGA-STAD	Stomach Adenocarcinoma	Stomach	443	443	439	441	443	443	443
TCGA-BLCA	Bladder Urothelial Carcinoma	Bladder	412	412	412	412	412	412	412
TCGA-LIHC	Liver Hepatocellular Carcinoma	Liver	377	377	376	375	376	377	377
TCGA-CESC	Cervical Squamous Cell Carcinoma and Endocervical...	Cervix	307	307	307	305	302	307	307
TCGA-KIRP	Kidney Renal Papillary Cell Carcinoma	Kidney	291	291	291	288	290	291	291
TCGA-SARC	Sarcoma	Soft Tissue	261	261	261	255	261	261	261
TCGA-LAML	Acute Myeloid Leukemia	Bone Marrow	200	191	169	149	143	140	200
TCGA-ESCA	Esophageal Carcinoma	Esophagus	185	185	184	184	185	185	185
TCGA-PAAD	Pancreatic Adenocarcinoma	Pancreas	185	185	178	183	185	184	185
TCGA-PCPG	Pheochromocytoma and Paraganglioma	Adrenal Gland	179	179	179	179	179	179	179
TCGA-READ	Rectum Adenocarcinoma	Colorectal	172	171	167	158	166	165	170
TCGA-TGCT	Testicular Germ Cell Tumors	Testis	150	150	150	150	134	150	134
TCGA-THYM	Thymoma	Thymus	124	124	124	123	124	124	124
TCGA-KICH	Kidney Chromophobe	Kidney	113	66	66	66	66	66	113
TCGA-ACC	Adrenocortical Carcinoma	Adrenal Gland	92	92	80	92	92	80	92
TCGA-MESO	Mesothelioma	Pleura	87	87	87	83	87	87	87
TCGA-UVM	Uveal Melanoma	Eye	80	80	80	80	80	80	80
TCGA-DLBC	Lymphoid Neoplasm Diffuse Large B-cell Lymphoma	Lymph Nodes	58	48	48	48	48	48	48
TCGA-UCS	Uterine Carcinosarcoma	Uterus	57	57	57	57	57	57	57
TCGA-CHOL	Cholangiocarcinoma	Bile Duct	51	51	36	51	36	36	45

Clinical data

- Categories:
  - Demographic
  - Diagnosis
  - Exposure
  - Family history
  - Follow up
  - Treatment

Description of all clinical information

- Each entry is either required, optional or preferred
- All the information are not available for each cohorts

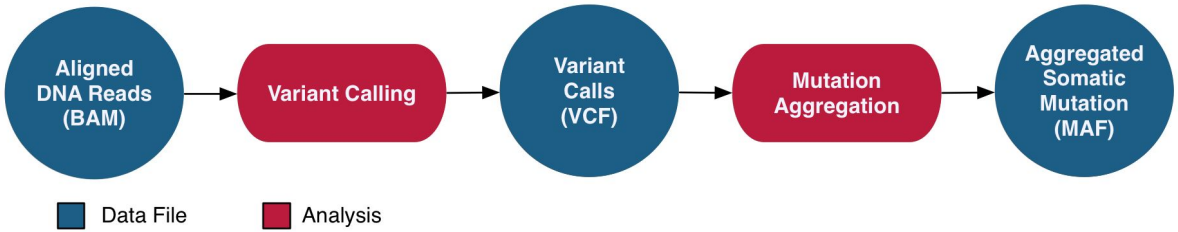
Clinical data

- Categories:
  - Demographic ethnicity, gender, year of birth, vital status ...
  - Diagnosis cell morphology, organ of origin, tumor stage, tumor grade, vital status, HPV status ...
  - Exposure cigarettes per day, alcohol history, bmi … (no required entry in this category)
  - Family history relationship age at diagnosis, relationship type … (no required entry in this category)
  - Follow up days to follow-up, m_protein, …
  - Treatment therapeutic agents, treatment outcome, … (no required entry in this category)

Description of all clinical information

- Each entry is either required, optional or preferred
- All the information are not available for each cohorts

Simple nucleotide variations



- Whole exome sequencing data
- Alignment on the human reference genome GRCh38 + viral and decoy sequences (10 types of human viral genomes)
- Variant callers used: MuSe, SomaticSniper,VarScan, Mutect2

Simple nucleotide variations

Public data

- MAF files with somatic variants: **one MAF file per project** and per variant calling

⟹ Protected MAF file after quality and germline filters

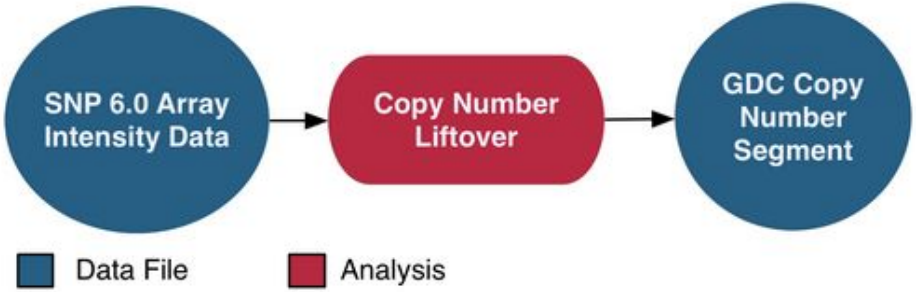
Protected data

- Tumor and normal BAM files
- Raw VCF files: one VCF per variant caller and per tumor/normal pair of BAMs (VCF format)
- Annotated VCF files
- Protected MAF files: one per project and pipeline, six more columns in comparison with the public MAF file (VCF columns)

MAF format: 1 row = 1 variant

Hugo_Symbol	Entrez_Gene_Id	NCBI_Build	Chromosome	Start_Position	End_Position	Strand	Variant_Classification	Variant_Type	Reference_Allele	Tumor_Seq_Allele2	Tumor_Sample_Barcode
GABRA6	2559	GRCh38	chr5	161690214	161690214	+	Silent	SNP	A	T	TCGA-49-4490-01A-21D-1855-08
EGFR	1956	GRCh38	chr7	55191822	55191822	+	Missense_Mutation	SNP	T	G	TCGA-49-4490-01A-21D-1855-08
TP53	7157	GRCh38	chr17	7674894	7674894	+	Nonsense_Mutation	SNP	G	A	TCGA-49-4490-01A-21D-1855-08
USH2A	7399	GRCh38	chr1	216251039	216251039	+	Missense_Mutation	SNP	C	A	TCGA-55-8205-01A-11D-2238-08
FMN2	56776	GRCh38	chr1	240257989	240257989	+	Silent	SNP	A	T	TCGA-55-8205-01A-11D-2238-08

Copy number variations



- GRCh37 probe set coordinates converted to GRCh38
- Circular binary segmentation performed using the DNACopy R package to detect chromosomal regions of equal copy number

Copy number variations

Public data

- Masked Copy Number Segment files: segments containing germline variations are removed

Protected data

- Raw intensities in the legacy archive
- Copy Number Segment files: one file per sample

Copy Number Segment file format: txt tab-delimited file

Sample	Chromosome	Start	End	Num_Probes	Segment_Mean	cohort
TCGA-4K-AA1H-01A-11D-A434-01	1	61735	98602	17	-0.7963	TGCT
TCGA-4K-AA1H-01A-11D-A434-01	1	258955	12813243	6514	0.0089	TGCT
TCGA-4K-AA1H-01A-11D-A434-01	1	12817195	12832455	4	0.9158	TGCT
TCGA-4K-AA1H-01A-11D-A434-01	1	12832654	16518437	2149	0.0084	TGCT
TCGA-4K-AA1H-01A-11D-A434-01	1	16519540	16886219	113	-0.1924	TGCT

Barcodes obtained using TCGAbiolinks package

start and end of chromosomal regions of equal copy number

Segment mean value:

log2(copy-number/2)

0 if diploid, <0 if deletion
>0 if amplification

RNA-seq

- Alignment on the GRCh38 reference genome using STAR
- Expression quantification using HT-Seq count
- Read counts normalization using two measures: Fragments Per Kilobase of transcript per Million mapped reads (FPKM) and FPKM Upper Quartile (FPKM-UQ)

RNA-seq

Public data

- Gene expression:
  - HT-Seq count
  - FPKM
  - FPKM-UQ

One file per sample

Gene expression format: txt tab-delimited file (FPKM example)

ENSG00000242268.2	0.0
ENSG00000270112.3	0.0148814840479
ENSG00000167578.15	6.60521492195

Protected data

- BAM files

FPKM: read count normalized by gene length and total protein-coding read count

FPKM-UQ: read count normalized by gene length and the 75th percentile value of the total protein-coding read count

miRNA-seq

- BWA-aln used for the alignment
- Expression quantification: pipeline developed by the University of British Columbia
- Read counts normalization: reads per million mapped reads (RPM) measure

miRNA-seq

Public data

- miRNA quantification
- Isoform quantification: informations found in the miRNA quantification file + isoform informations

Protected data

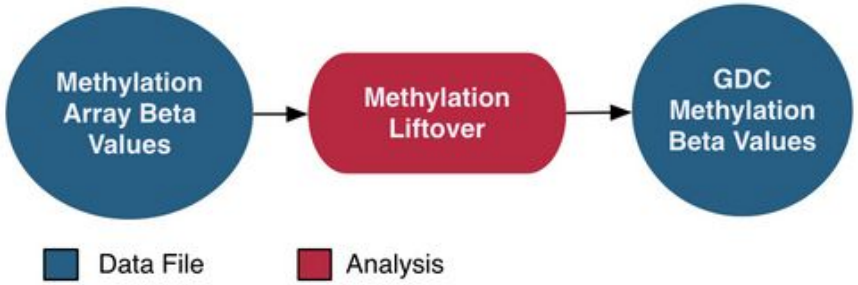
- BAM files

One file per sample

Quantification file format: txt tab-delimited file

miRNA_ID	isoform_coords	read_count	reads_per_million_miRNA_mapped	cross.mapped	miRNA_region
hsa-let-7a-1	hg38:chr9:94175938-94175962:+	1	0.234587	N	precursor
hsa-let-7a-1	hg38:chr9:94175942-94175962:+	2	0.469174	N	precursor
hsa-let-7a-1	hg38:chr9:94175961-94175983:+	2	0.469174	N	mature,MIMAT0000062
hsa-let-7a-1	hg38:chr9:94175961-94175984:+	5	1.172935	N	mature,MIMAT0000062

Methylation data



- Two arrays: Illumina Human Methylation 27 (HM27) and Human Methylation 450 (HM450)
- Measure the level of methylation at known CpG sites
- Measure: beta values (between 0 and 1)

Methylation data

Public data:

Methylation beta values table: one file per sample

Format: txt tab-delimited file

Composite.Element.REF	Beta_value	Chromosome	Start	End	Gene_Symbol	Gene_Type	Transcript_ID
cg00000029	0.410239744	chr16	53434200	53434201	RBL2;RBL2;RBL2	protein_coding;protein_coding;protein_coding	ENST00000262133.9;ENST00000544405.!
cg00000108	NA	chr3	37417715	37417716	C3orf35;C3orf35;C3orf35;C3orf35;C3orf35;C3orf35;...	lincRNA;lincRNA;lincRNA;lincRNA;lincRNA;lincRNA;lin...	ENST00000328376.8;ENST00000332506.(
cg00000109	NA	chr3	172198247	172198248	FNDC3B;FNDC3B;FNDC3B;FNDC3B;FNDC3B;FNDC3B	protein_coding;protein_coding;protein_coding;protei...	ENST00000336824.7;ENST00000415807.!
cg00000165	0.196702275	chr1	90729117	90729118	.	.	.
cg00000236	0.890144819	chr8	42405776	42405777	VDAC3	protein_coding	ENST00000022615.7

array probe ID associated with a CpG site

chr, start and end of the CpG site

genes associated to the CpG site

Thank you!