The intersectional genetics landscape for human

Andre Macedo* and Alisson M. Gontijo*

Chronic Diseases Research Center (CEDOC), NOVA Medical School | Faculdade de Ciências Médicas, Universidade Nova de Lisboa, Lisbon, Portugal.

* Correspondence: andre.macedo@nms.unl.pt (A.M.) and alisson.gontijo@nms.unl.pt (A.M.G.)

Supplementary Materials

This file contains Supplementary Figures S1-4, Supplementary Tables S1-S4, and Supplementary Data S1. The latter is available at https://github.com/AndreMacedo88/VEnCode.

Supplementary Figures

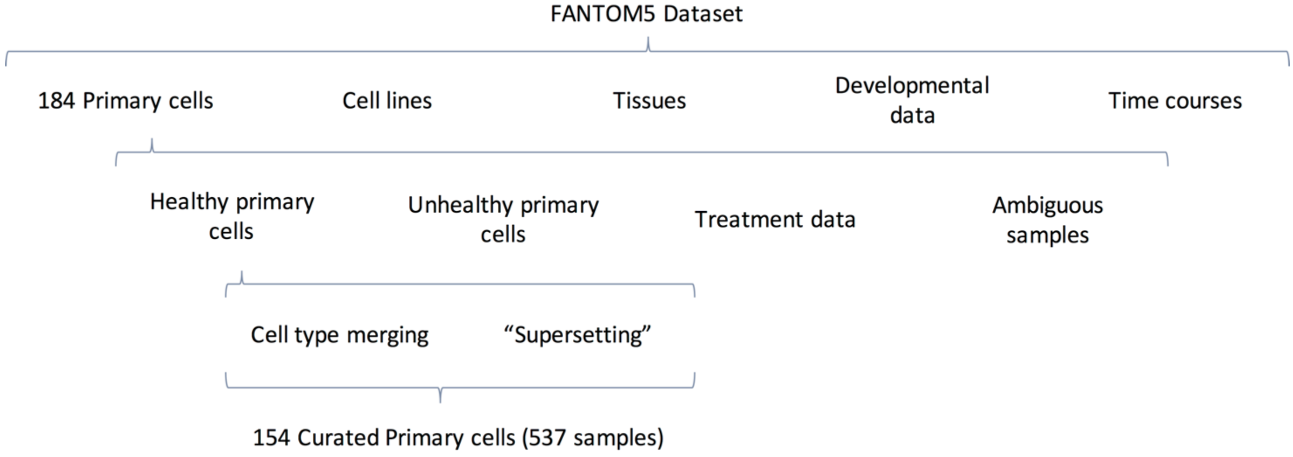


Figure S1. Pipeline for FANTOM5 data preparation and curation. Further details on cell type merging, “supersetting”, and excluded primary cells are provide Supplementary Table S1.

1. Calculate E_{raw} from various samples of inactive promoters

Vary number cell types →

RE activity

RE1 0 0 0 0 0 0 0 0 0 0 0 …
RE2 0 0 0 0 0 0 0 0 0 0 0 …
… 0 0 0 0 0 0 0 0 0 0 0 …
0 0 0 0 0 0 0 0 0 0 0 …
0 0 0 0 0 0 0 0 0 0 0 …

Vary *k*

2. Calculate correlation best-fitting curve

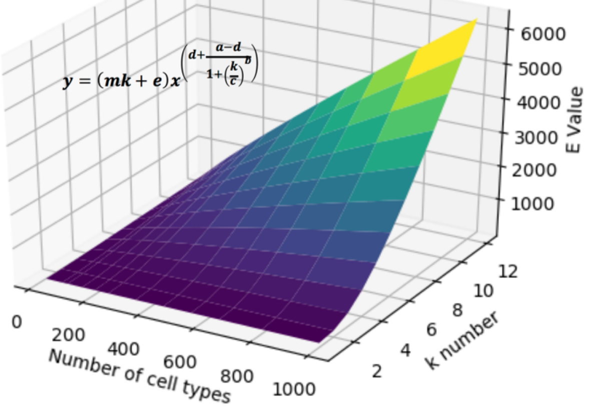
Vary number cell types: $y = sk^h$

Coefficient variation with k:

$$s = mk + e \qquad R^2 = 0.9993$$

$$h = d + \frac{a - d}{1 + \left(\frac{k}{c}\right)^b} \qquad R^2 = 1$$

Coefficients					
a	b	c	d	m	e
-164054,1	0,9998811	6,08895E-06	1,00051	0,9527	-0,1131



3. Generate function that generates expected best *E*

$$y = (mk + e)x^{\left(d + \frac{a-d}{1+\left(\frac{k}{c}\right)^b}\right)}$$

Figure S2. Generating the function that returns the best possible *E* (*E_{best}*). We generated E_{raw} (as described in Figure 6) for simulated data reflecting a best-case-scenario for a VEnCode – where all *k* REs are inactive in the non-target cell – varying the number of cell types and *k* in the dataset (1.). By changing cell type number ranging from 20 to 1000, we calculated the best fitting curves that predicted E_{raw} and generated the general equation $y = sk^h$ (2.). However, *s* and *h* depend on the number of REs (*k*). So, varying *k* from 1 up to 10, we obtained the best fitting equations that explain this variation (2.). With this data we then generated a function that best accounts for the variation of both *k* and cell type number in the dataset (3.). The effectiveness of this equation can be seen in Table S2.

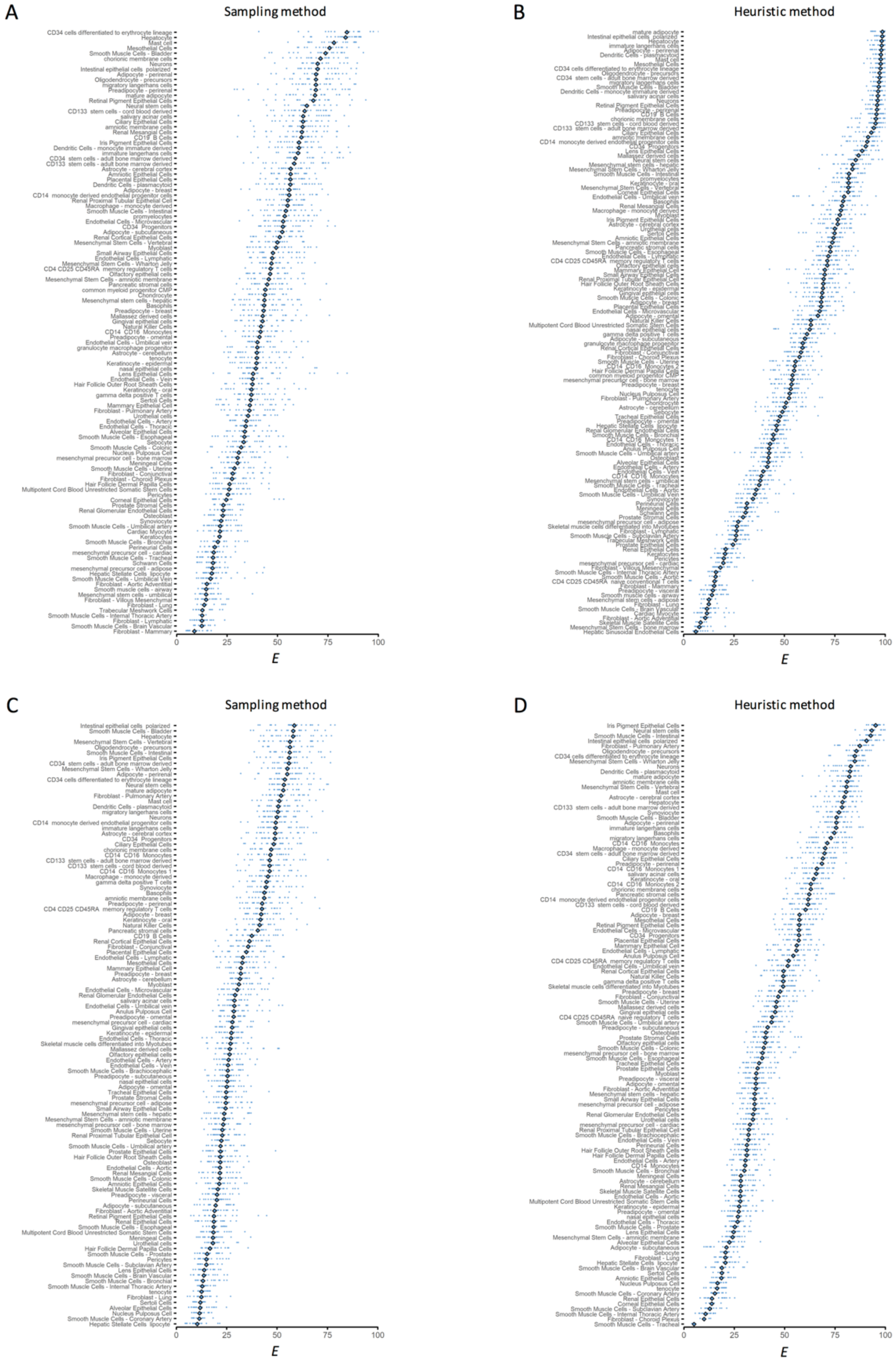


Figure S3. *E* value variation by cell type. 5 to 20 VEnCodes at *k* = 4 were obtained for every possible primary cell type and their *E* values were determined as described in Figure 6. **A, B.** Results for VEnCodes generated using the promoter dataset. In (**A**) the sampling method (see Figure 3) was used to obtain VEnCodes and determine *E* for 114 cell types. In **(B)**, the heuristic method (see Figure 4) was used for the same purpose, allowing us to analyze *E* for up to 20 VEnCodes for 131 cell types. **C, D.** Results for VEnCodes generated using the enhancer dataset. **C.** *E* value distribution of VEnCodes obtained with the sampling method for 112 primary cell types. **D.** Similarly, 117 cell types had 5-20 enhancer VEnCodes retrieved using the heuristic method and their *E* was determined.

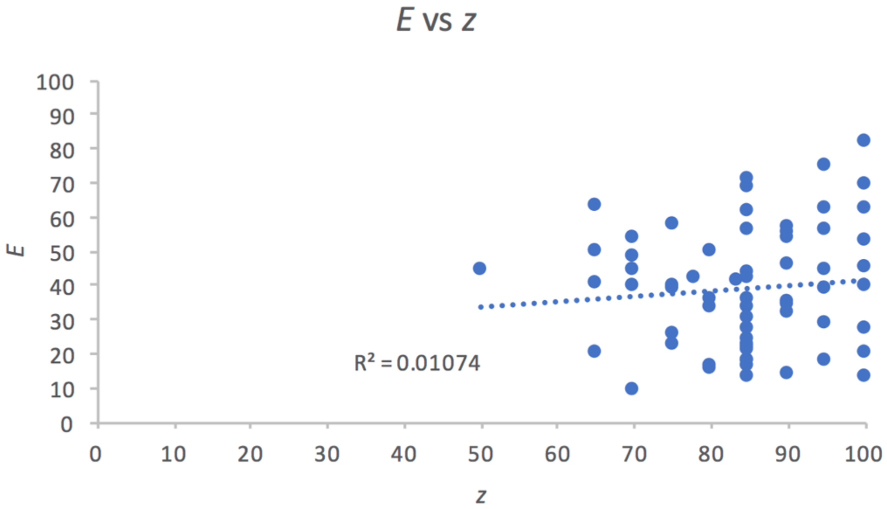


Figure S4. *E* versus *z* scores. Plotted is the *E* and *z* scores for each cell type of a list of 64 cell types with three donors and which we managed to retrieve both *E* and *z* values. Also plotted is the linear regression attempt to model the relationship between these two values.

Supplementary Tables

Table S1. List of curated primary cell types and merged, supersets and excluded categories used in this study. The list of curated cell types contains 154 primary cell types, encompassing a total of 537 cell line samples. Merging was done with the rationale of turning a vast data on diverse cell conditions into biologically relevant cell types. Merged cell types are then used as curated cell types and the original "cell type" data is not accessed independently. On the other hand, a superset cell type means that the superset data includes the subset data, but each subset is still included in the curated cell type list used in this study. The excluded cell type category lists the data not used at any point throughout the study.

Curated cell types (154 Primary cell types)
Adipocyte - breast
Adipocyte - omental
Adipocyte - perirenal
Adipocyte - subcutaneous
Alveolar Epithelial Cells
Amniotic Epithelial Cells
amniotic membrane cells
Anulus Pulposus Cell
Astrocyte - cerebellum
Astrocyte - cerebral cortex
Basophils
Bronchial Epithelial Cell
Cardiac Myocyte
CD133+ stem cells - adult bone marrow derived
CD133+ stem cells - cord blood derived
CD14+ monocyte derived endothelial progenitor cells
CD14+ Monocytes
CD14+CD16- Monocytes
CD14+CD16+ Monocytes
CD14-CD16+ Monocytes
CD19+ B Cells
CD34 cells differentiated to erythrocyte lineage
CD34+ Progenitors
CD34+ stem cells - adult bone marrow derived
CD4+ T Cells
CD4+CD25+CD45RA- memory regulatory T cells
CD4+CD25+CD45RA+ naive regulatory T cells
CD4+CD25-CD45RA- memory conventional T cells
CD4+CD25-CD45RA+ naive conventional T cells
CD8+ T Cells
Chondrocyte
chorionic membrane cells
Ciliary Epithelial Cells
common myeloid progenitor CMP
Corneal Epithelial Cells
Dendritic Cells - monocyte immature derived
Dendritic Cells - plasmacytoid
Endothelial Cells - Aortic
Endothelial Cells - Artery
Endothelial Cells - Lymphatic
Endothelial Cells - Microvascular
Endothelial Cells - Thoracic
Endothelial Cells - Umbilical vein
Endothelial Cells - Vein
Eosinophils
Esophageal Epithelial Cells
Fibroblast - Aortic Adventitial
Fibroblast - Cardiac
Fibroblast - Choroid Plexus
Fibroblast - Conjunctival
Fibroblast - Dermal
Fibroblast - Gingival
Fibroblast - Lung
Fibroblast - Lymphatic
Fibroblast - Mammary
Fibroblast - Periodontal Ligament
Fibroblast - Pulmonary Artery
Fibroblast - skin
Fibroblast - Villous Mesenchymal
gamma delta positive T cells
Gingival epithelial cells
granulocyte macrophage progenitor
Hair Follicle Dermal Papilla Cells
Hair Follicle Outer Root Sheath Cells
Hepatic Sinusoidal Endothelial Cells
Hepatic Stellate Cells (lipocyte)
Hepatocyte
immature langerhans cells
Intestinal epithelial cells (polarized)
Iris Pigment Epithelial Cells
Keratinocyte - epidermal
Keratinocyte - oral
Keratocytes
Lens Epithelial Cells
Macrophage - monocyte derived
Mallassez-derived cells
Mammary Epithelial Cell
Mast cell
mature adipocyte
Melanocyte
Meningeal Cells
mesenchymal precursor cell - adipose
mesenchymal precursor cell - bone marrow
mesenchymal precursor cell - cardiac
Mesenchymal stem cells - adipose
Mesenchymal Stem Cells - amniotic membrane
Mesenchymal Stem Cells - bone marrow
Mesenchymal stem cells - hepatic
Mesenchymal stem cells - umbilical
Mesenchymal Stem Cells - Vertebral
Mesenchymal Stem Cells - Wharton Jelly
Mesothelial Cells
migratory langerhans cells
Multipotent Cord Blood Unrestricted Somatic Stem Cells
Myoblast
nasal epithelial cells
Natural Killer Cells
Neural stem cells
Neurons
Neutrophil
Nucleus Pulposus Cell
Olfactory epithelial cells
Oligodendrocyte - precursors
Osteoblast
Pancreatic stromal cells
Pericytes
Perineurial Cells
Placental Epithelial Cells
Preadipocyte - breast
Preadipocyte - omental
Preadipocyte - perirenal
Preadipocyte - subcutaneous
Preadipocyte - visceral
promyelocytes
Prostate Epithelial Cells
Prostate Stromal Cells
Renal Cortical Epithelial Cells
Renal Epithelial Cells
Renal Glomerular Endothelial Cells
Renal Mesangial Cells
Renal Proximal Tubular Epithelial Cell
Retinal Pigment Epithelial Cells
salivary acinar cells
Schwann Cells
Sebocyte
Sertoli Cells
Skeletal Muscle Cells
Skeletal muscle cells differentiated into Myotubes - multinucleated
Skeletal Muscle Satellite Cells
Small Airway Epithelial Cells
Smooth muscle cells - airway
Smooth Muscle Cells - Aortic
Smooth Muscle Cells - Bladder
Smooth Muscle Cells - Brachiocephalic
Smooth Muscle Cells - Brain Vascular
Smooth Muscle Cells - Bronchial
Smooth Muscle Cells - Carotid
Smooth Muscle Cells - Colonic
Smooth Muscle Cells - Coronary Artery
Smooth Muscle Cells - Esophageal
Smooth Muscle Cells - Internal Thoracic Artery
Smooth Muscle Cells - Intestinal
Smooth Muscle Cells - Prostate
Smooth Muscle Cells - Pulmonary Artery
Smooth Muscle Cells - Subclavian Artery
Smooth Muscle Cells - Tracheal
Smooth Muscle Cells - Umbilical artery
Smooth Muscle Cells - Umbilical Vein
Smooth Muscle Cells - Uterine
Synoviocyte
tenocyte
Trabecular Meshwork Cells
Tracheal Epithelial Cells
Urothelial cells

Table S1. Continued.

Merged data	
Merged cell types	**Original cell types**
CD14+ Monocytes	CD14+ monocytes - mock treated CD14+ monocytes - treated with BCG CD14+ monocytes - treated with B-glucan CD14+ monocytes - treated with Candida CD14+ monocytes - treated with Cryptococcus CD14+ monocytes - treated with Group A streptococci CD14+ monocytes - treated with IFN + N-hexane CD14+ monocytes - treated with lipopolysaccharide CD14+ monocytes - treated with Salmonella CD14+ monocytes - treated with Trehalose dimycolate (TDM) CD14+ Monocytes
CD19+ B Cells	CD19+ B Cells (pluriselect) CD19+ B Cells
CD4+CD25+CD45RA- memory regulatory T cells	CD4+CD25+CD45RA- memory regulatory T cells expanded CD4+CD25+CD45RA- memory regulatory T cells
CD4+CD25+CD45RA+ naive regulatory T cells	CD4+CD25+CD45RA+ naive regulatory T cells expanded CD4+CD25+CD45RA+ naive regulatory T cells
CD4+CD25-CD45RA- memory conventional T cells	CD4+CD25-CD45RA- memory conventional T cells expanded CD4+CD25-CD45RA- memory conventional T cells
CD4+CD25-CD45RA+ naive conventional T cells	CD4+CD25-CD45RA+ naive conventional T cells expanded CD4+CD25-CD45RA+ naive conventional T cells
CD8+ T Cells	CD8+ T Cells (pluriselect) CD8+ T Cells
Chondrocyte	Chondrocyte - de diff Chondrocyte - re diff
Fibroblast - skin	Fibroblast - skin dystrophia myotonica Fibroblast - skin normal Fibroblast - skin spinal muscular atrophy Fibroblast - skin walker warburg
Mast cell	Mast cell - stimulated Mast cell
Melanocyte	Melanocyte - dark Melanocyte - light
Neutrophil	neutrophil PMN Neutrophils
Prostate Epithelial Cells	Prostate Epithelial Cells (polarized) Prostate Epithelial Cells

Table S1. Continued.

Curated supersets		Excluded cell types
superset	subset	
CD14+ Monocytes	CD14+CD16- Monocytes	mesenchymal precursor cell - ovarian cancer left ovary
	CD14+CD16+ Monocytes	mesenchymal precursor cell - ovarian cancer metastasis
		mesenchymal precursor cell - ovarian cancer right ovary
CD4+ T Cells	CD4+CD25+CD45RA- memory regulatory T cells	Osteoblast - differentiated
	CD4+CD25+CD45RA+ naive regulatory T cells	Peripheral Blood Mononuclear Cells
	CD4+CD25-CD45RA- memory conventional T cells	Whole blood (ribopure)
	CD4+CD25-CD45RA+ naive conventional T cells	

Table S2. Normalized *E* (*E* = E_{raw}/E_{best}) values for a range of number of cell types in data and *k* REs used to generate a VEnCode. Values were generated, as described in Figure 6, for a intraindividual robust best-case possible VEnCode and normalized using the function described in Figure S2. Thus, the values reflect the average *E* expected for the most intraindividual robust VEnCodes.

		Number of Cell types											
		20	**80**	**100**	**154**	**200**	**250**	**350**	**450**	**550**	**650**	**800**	**1000**
k	**1**	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	2	99.5	98.8	96.5	96.1	97.9	100.0	97.2	97.3	97.4	100.0	100.0	100.0
	3	96.3	98.9	97.7	96.7	98.6	97.9	95.8	98.6	95.2	97.1	99.5	95.9
	4	99.2	98.4	97.3	98.4	97.6	98.4	99.1	98.3	98.5	98.6	98.7	97.0
	5	99.5	98.9	99.1	99.1	97.7	99.0	99.0	97.2	98.8	98.5	99.1	99.0
	6	98.5	99.9	99.4	99.3	98.5	99.7	99.3	99.4	99.1	99.0	99.1	99.8
	7	100.0	99.9	100.0	98.9	99.5	99.9	100.0	99.3	99.7	99.2	98.0	100.0
	8	99.9	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	99.6
	9	100.0	100.0	100.0	99.9	100.0	100.0	100.0	99.9	100.0	100.0	100.0	100.0
	10	100.0	100.0	100.0	99.7	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0

Table S3. List of curated cancer cell types used in this study. The list contains 158 biologically relevant cell types, encompassing a total of 274 cancer cell line samples.

Curated cancer cell types
acantholytic squamous carcinoma cell line:HCC1806
acute lymphoblastic leukemia (B-ALL) cell line
acute lymphoblastic leukemia (T-ALL) cell line
acute myeloid leukemia (FAB M0) cell line
acute myeloid leukemia (FAB M1) cell line
acute myeloid leukemia (FAB M2) cell line
acute myeloid leukemia (FAB M3) cell line
acute myeloid leukemia (FAB M4) cell line
acute myeloid leukemia (FAB M4eo) cell line
acute myeloid leukemia (FAB M5) cell line
acute myeloid leukemia (FAB M6) cell line
acute myeloid leukemia (FAB M7) cell line
adenocarcinoma cell line:IM95m
adrenal cortex adenocarcinoma cell line:SW-13
adult T-cell leukemia cell line:ATN-1
alveolar cell carcinoma cell line:SW 1573
anaplastic carcinoma cell line:8305C
anaplastic large cell lymphoma cell line:Ki-JK
anaplastic squamous cell carcinoma cell line:RPMI 2650
argyrophil small cell carcinoma cell line:TC-YIK
astrocytoma cell line:TM-31
b cell line:RPMI1788
B lymphoblastoid cell line: GM12878 ENCODE
basal cell carcinoma cell line:TE 354.T
bile duct carcinoma cell line
biphenotypic B myelomonocytic leukemia cell line:MV-4-11
bone marrow stromal cell line:StromaNKtert
breast carcinoma cell line
bronchial squamous cell carcinoma cell line:KNS-62
bronchioalveolar carcinoma cell line
bronchogenic carcinoma cell line:ChaGo-K-1
Burkitt lymphoma cell line
carcinoid cell line
carcinosarcoma cell line:JHUCS-1
cervical cancer cell line
cholangiocellular carcinoma cell line:HuH-28
chondrosarcoma cell line:SW 1353
choriocarcinoma cell line
chronic lymphocytic leukemia cell line:SKW-3
chronic megakaryoblastic cell line:MEG-01
chronic myeloblastic leukemia cell line:KCL-22
chronic myelogenous leukemia cell line
clear cell carcinoma cell line
colon carcinoma cell line
cord blood derived cell line:COBL-a untreated
diffuse large B-cell lymphoma cell line:CTB-1
ductal cell carcinoma cell line
embryonic kidney cell line: HEK293/SLAM untreated
embryonic pancreas cell line
endometrial carcinoma cell line:OMC-2
endometrial stromal sarcoma cell line:OMC-9
endometrioid adenocarcinoma cell line:JHUEM-1
epidermoid carcinoma cell line
epithelioid sarcoma cell line
epitheloid carcinoma cell line: HelaS3 ENCODE
Ewing sarcoma cell line:Hs 863
extraskeletal myxoid chondrosarcoma cell line:H-EMC-SS
fibrosarcoma cell line:HT-1080
fibrous histiocytoma cell line:GCT TIB-223
gall bladder carcinoma cell line
gastric adenocarcinoma cell line
gastric cancer cell line
gastrointestinal carcinoma cell line:ECC12
giant cell carcinoma cell line
glassy cell carcinoma cell line:HOKUG
glioblastoma cell line
glioma cell line:GI-1
granulosa cell tumor cell line:KGN
hairy cell leukemia cell line:Mo
Hep-2 cells mock treated
hepatic mesenchymal tumor cell line:LI90
hepatoblastoma cell line:HuH-6
hepatocellular carcinoma cell line: HepG2 ENCODE
hepatoma cell line:Li-7
hereditary spherocytic anemia cell line:WIL2-NS
Hodgkin lymphoma cell line:HD-Mar2
keratoacanthoma cell line:HKA-1
Krukenberg tumor cell line:HSKTC
large cell lung carcinoma cell line
large cell non-keratinizing squamous carcinoma cell line:SKG-II-SF
leiomyoblastoma cell line:G-402
leiomyoma cell line
leiomyosarcoma cell line:Hs 5
lens epithelial cell line:SRA
liposarcoma cell line
lung adenocarcinoma cell line
lung adenocarcinoma papillary cell line:NCI-H441
lymphangiectasia cell line:DS-1
lymphoma malignant hairy B-cell cell line:MLMA
malignant trichilemmal cyst cell line:DJM-1
maxillary sinus tumor cell line:HSQ-89
medulloblastoma cell line
melanoma cell line
meningioma cell line:HKBMM
merkel cell carcinoma cell line
mesenchymal stem cell line:Hu5/E18
mesodermal tumor cell line:HIRS-BM
Epithelioid mesothelioma cell line
Sarcomatoid mesothelioma cell line
Biphasic mesothelioma cell line
mixed mullerian tumor cell line:HTMMT
mucinous adenocarcinoma cell line:JHOM-1
mucinous cystadenocarcinoma cell line:MCAS
myelodysplastic syndrome cell line:SKM-1
myeloma cell line:PCM6
myxofibrosarcoma cell line
neuroblastoma cell line
neuroectodermal tumor cell line
neuroepithelioma cell line:SK-N-MC
neurofibroma cell line:Hs 53
NK T cell leukemia cell line:KHYG-1
non T non B acute lymphoblastic leukemia cell line:P30/OHK
non-small cell lung cancer cell line:NCI-H1385
normal embryonic palatal mesenchymal cell line:HEPM
normal intestinal epithelial cell line:FHs 74 Int
oral squamous cell carcinoma cell line
osteoclastoma cell line:Hs 706
osteosarcoma cell line
pagetoid sarcoma cell line:Hs 925
pancreatic carcinoma cell line:NOR-P1
papillary adenocarcinoma cell line:8505C
papillotubular adenocarcinoma cell line:TGBC18TKB
peripheral neuroectodermal tumor cell line:KU-SN
pharyngeal carcinoma cell line:Detroit 562
plasma cell leukemia cell line:ARH-77
pleomorphic hepatocellular carcinoma cell line:SNU-387
prostate cancer cell line
rectal cancer cell line:TT1TKB
renal cell carcinoma cell line
retinoblastoma cell line:Y79
rhabdomyosarcoma cell line
sacrococcigeal teratoma cell line:HTST
schwannoma cell line:HS-PSS
serous adenocarcinoma cell line
serous cystadenocarcinoma cell line:HTOA
signet ring carcinoma cell line
small cell cervical cancer cell line:HCSC-1
small cell gastrointestinal carcinoma cell line:ECC10
small cell lung carcinoma cell line
small-cell gastrointestinal carcinoma cell line:ECC4
somatostatinoma cell line:QGP-1
spindle cell sarcoma cell line:Hs 132
splenic lymphoma with villous lymphocytes cell line:SLVL
squamous cell carcinoma cell line:EC-GI-10.CNhs11252.10463-106H4
squamous cell carcinoma cell line:JHUS-nk1.CNhs11749.10646-109A7
squamous cell carcinoma cell line:T3M-5.CNhs11739.10616-108G4
squamous cell lung carcinoma cell line
synovial sarcoma cell line:HS-SY-II
T cell lymphoma cell line:HuT 102 TIB-162
teratocarcinoma cell line
testicular germ cell embryonal carcinoma cell line
thymic carcinoma cell line:Ty-82
thyroid carcinoma cell line
transitional cell carcinoma cell line
tridermal teratoma cell line:HGRT
tubular adenocarcinoma cell line:SUIT-2
Wilms tumor cell line
xeroderma pigentosum b cell line:XPL 17