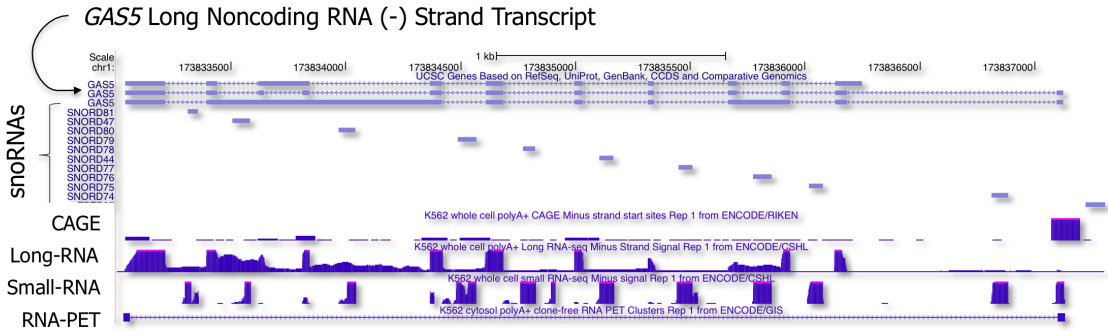
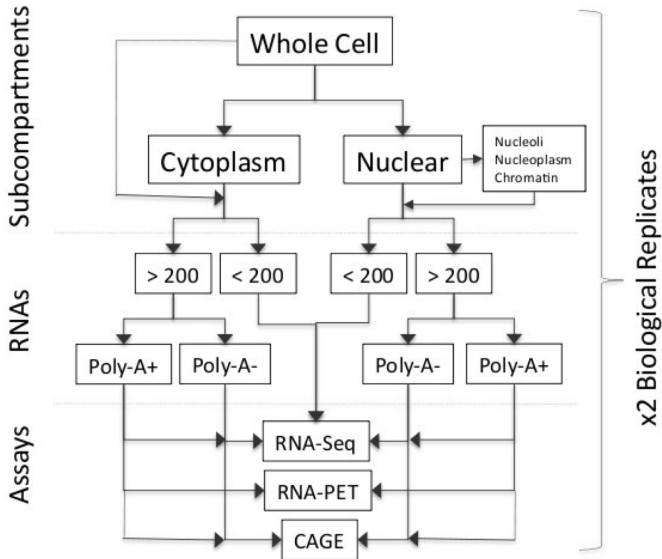


# Human ENCODE Transcriptome Datasets

All Data is Available Now for Analysis

# General Workflow



# The Cell Lines

Tier	Name	Description	Primary: (y/n)	Tissue derived from	Cell Type	Replicate 1		Replicate 2		Same Donor	Source
						Age	Gender	Age	Gender		
1	K562	chronic myelogenous leukemia	n	blood	leukemic lymphoblast	53	f	53	f	y	ATCC
1	GM12878	EBV transformed lymphoblastoid	y	blood	B lymphocyte	?	f	?	f	y	Coriella
1	H1 ES Cells	embryonic stem cells	y	yes	yes	0	m	0	m	y	Cell Dynam
2	HepG2	hepatocellular carcinoma	n	liver	carcinoma	15	m	15	m	y	ATCC
2	Hela-S3	cervical carcinoma	n	cervix	carcinoma	31	f	31	f	y	ATCC
2	HUVEC	Umbilical vein endothelial cells	y	umbilical vein	endothelial	?	m	?	m	y	Lonza
2	A549	lung carcinoma	n	lung	carcinoma	58	m	58	m	y	ATCC
2	CD14	CD14+ monocytes	y	blood	monocytes	32	f	32	f	y	J. Stam
2	CD20	CD20+ B cells	y	blood	B cells	27	f	43	f	n	J. Stam
2	IMR90	Fetal Lung Fibroblasts	y	lung	fibroblasts	16 wk	f	16 wk	f	y	ATCC
2	MCF7	Mammary Gland Adenocarcinoma	n	mammary gland	adenocarcinoma	69	f	69	f	y	ATCC
2	SK-N-SH	Brain neuroblastoma	n	brain	neuroblastoma	4	f	4	f	y	ATCC
3	NHEK	Epidermal Keratinocytes	y	epidermis	keratinocytes	?	f	?	f	y	Lonza
3	AGO4450	Fetal Lung Fibroblasts	y	lung	fibroblast	12 wk	m	12 wk	m	y	J. Stam
3	SK-N-SH RA	SK-N-SH stimulated with retinoic acid	n	brain	neuroblastoma	4	f	4	f	y	J. Stam
3	HSMM	Skeletal Muscle Myoblasts	y	skeletal muscle	myoblasts	?	?	-	-	-	Bernstein
3	NHLF	Lung Fibroblasts	y	lung	fibroblasts	?	?	?	?	y	Bernstein
3	HMEC	Mammary Gland Epithelial	y	mammary gland	epithelial	?	f	-	-	-	Crawford
3	BJ	Skin fibroblast foreskin	y	epidermis	fibroblast	newborn	m	newborn	m	y	J. Stam
3	CD34+ Mobilized	CD34+ progenitor cells	y	blood	progenitor	23	m	-	-	-	J. Stam
3	SkMC	Skeletal Muscle Cells	y	skeletal muscle	muscle	64	f	21	m	n	Promocell
3	NHDF	Dermal Fibroblasts	y	temple/breast	fibroblast	55	f	44	f	n	Promocell
3	HVMF	Villous Mesenchymal Fibroblasts	y	placenta villous	fibroblasts	newborn	m	newborn	f	n	Promocell
3	HOB	Osteoblasts	y	calcenous bone	osteoblasts	62	m	56	f	n	Promocell
3	HAoAF	Aortic Adventitial Fibroblasts	y	Aorta	fibroblasts	47	m	24	f	n	Promocell
3	HAoEC	Aortic Endothelial Cells	y	Aorta	endothelial	22	f	55	m	n	Promocell
3	HCH	Chondrocytes	y	cartilage	chondrocytes	64	m	56	f	n	Promocell
3	HFDPC	Follicle Dermal Papilla Cells	y	epidermis (scalp)	follicle dermal papilla	47	f	70	f	n	Promocell
3	HMEpC	Mammary Gland Epithelial	y	mammary gland	epithelial	23	f	-	-	-	Promocell
3	hMSC-AT	Mesenchymal Stem Cells from Adipose Tissue	y	adipose	mesenchymal stem	42	f	37	f	n	Promocell
3	hMSC-BM	Mesenchymal Stem Cells from Bone Marrow	y	bone marrow	mesenchymal stem	57	m	60	f	n	Promocell
3	hMSC-UC	Mesenchymal Stem Cells from Umbilical Cord	y	umbilical cord	mesenchymal stem	newborn	f	newborn	m	n	Promocell
3	HPC-PL	Pericytes from Human Placenta	y	placental	pericytes	newborn	m	newborn	f	n	Promocell
3	HPIEpC	Human Placental Epithelial Cells	y	placental	epithelial	newborn	f	newborn	m	n	Promocell
3	hMNC-CB	Mononuclear Cells from Cord Blood single donor	y	blood	mononuclear	newborn	f	newborn	m	n	Promocell
3	hMNC-PB	Mononuclear Cells from Perip Blood single donor	y	blood	mononuclear	42	m	52	f	n	Promocell
3	HWP	White Preadipocytes subcutaneous	y	adipose	white preadipocytes	65	m	62	f	n	Promocell
3	NHEM_f_M2	Epidermal Melanocytes	y	epidermis	melanocytes	3	m	1	m	n	Promocell
3	NHEM_M2	Epidermal Melanocytes	y	epidermis	melanocytes	55	m	52	f	n	Promocell
3	HSaVEC	Saphenous Vein Endothelial Cells	y	saphenous vein	endothelial	52	m	48	m	n	Promocell

? = metadata not collected

- = data not collected

# Data Matrix

Cellular Fraction	Assay	Whole Cell								Nuclear								Cytoplasm													
		Long-Seq				Small-Seq				CAGE	PET	Long-Seq				Small-Seq				CAGE	PET	Long-Seq				Small-Seq				CAGE	PET
		RiboDep	RiboDep	RiboDep	TAP	CIP-TAP	Untreated	N/A Poly-A +	N/A Poly-A +			RiboDep	RiboDep	RiboDep	TAP	CIP-TAP	Untreated	N/A Poly-A +	N/A Poly-A +			RiboDep	RiboDep	RiboDep	TAP	CIP-TAP	Untreated	N/A Poly-A +	N/A Poly-A +		
RNA Type	Total	Poly-A+	Poly-A-	Total	Total	Total			Total	Poly-A+	Poly-A-	Total	Total	Total			Total	Poly-A+	Poly-A-	Total	Total	Total	Poly-A+	Poly-A+							
1	K562																														
1	GM12878																														
1	H1 ES Cells																														
2	HepG2																														
2	HeLa-S3																														
2	HUVEC																														
2	A549																														
2	CD14																														
2	CD20																														
2	IMR90																														
2	MCF7																														
2	SK-N-SH																														
3	NHEK																														
3	AGO4450																														
3	SK-N-SH RA																														
3	HSM1																														
3	NHLF																														
3	HMEC																														
3	Bj																														
3	CD34+ Mobilized																														
3	SkMC																														
3	NHDF																														
3	HVMF																														
3	HOB																														
3	HAoAF																														
3	HAoEC																														
3	HCH																														
3	HFDPC																														
3	HMEpC																														
3	hMSC-AT																														
3	hMSC-BM																														
3	hMSC-UC																														
3	HPC-PI																														
3	HPIEpC																														
3	hMNC-CB																														
3	hMNC-PB																														
3	HWP																														
3	NHEM.f M2																														
3	NHEM M2																														
3	HSaVEC																														

White = yes data

Black = no data

Most done in replicates

Also, additional early generation datasets lacking replicates (CAGE on Poly-A- and tiling array) not shown here but available at the DCC

# File Types

- Primary: FASTQ
  - 99.9% of data done on Illumina. Some early data on 454 and Solid.
  - One file per replicate, i.e. replicates not merged
- Secondary/Mapped: .BAM and .bigwig.
  - One file per replicate, i.e. replicates not merged
- Tertiary/Elements: .BED and .GTF
  - Replicates merged -> element called -> scored against each replicate -> npIDR
  - One file per merged replicates
  - All element calls w/ npIDR score included, analyst needs to threshold as desired.

All available for download at UCSC

# Element Files

- Long RNA-Seq Elements
  - Contigs (Felix Schlesinger, [schlesin@cshl.edu](mailto:schlesin@cshl.edu))
  - Junctions (Alex Dobin, [dobin@cshl.edu](mailto:dobin@cshl.edu))
  - Cufflinks Genes & Transcripts (Alex Dobin, [dobin@cshl.edu](mailto:dobin@cshl.edu) and Chenghai Xue, [xuec@cshl.edu](mailto:xuec@cshl.edu))
  - FluxCapacitor Gencode Quantifications (Sarah Djebali [sarah.djebali@crg.edu](mailto:sarah.djebali@crg.edu) )
- Small RNA-Seq Elements
  - Contigs (Wei Lin, [wlin@cshl.edu](mailto:wlin@cshl.edu))
  - Gencode exon quantifications (Wei Lin, [wlin@cshl.edu](mailto:wlin@cshl.edu))
- CAGE Elements
  - CAGE Peaks (Timo Lassmann, [timolassmann@gmail.com](mailto:timolassmann@gmail.com))
  - HMM Peaks (Timo Lassmann, [timolassmann@gmail.com](mailto:timolassmann@gmail.com))
  - FluxCapacitor Gencode Quantifications (Sarah Djebali [sarah.djebali@crg.edu](mailto:sarah.djebali@crg.edu) ) ?????
- RNA-PET
  - Di-tags: Oscar Luo ([luojh@gis.a-star.edu.sg](mailto:luojh@gis.a-star.edu.sg))

All files at the DCC and have README describing column values.

Analysts encouraged to contact data provider (above) to inquire about element processing and files/formats.

# Where Can I Get Data

- CRG Dashboard
  - [http://genome.crg.es/~jlagarde/encode\\_RNA\\_dashboard/](http://genome.crg.es/~jlagarde/encode_RNA_dashboard/)
- DCC (go to Preview Browser)
  - <http://encodeproject.org/ENCODE/index.html>
- GEO
  - <http://www.ncbi.nlm.nih.gov/geo/info/ENCODE.html#ENCODE>