Stanford MEDICINE

ENCODE

ENCODE DCC data modeling and metadata standardization vignettes

Idan Gabdank

ENCODE Data Coordination Center

GSC20

A long journey to reproducible results

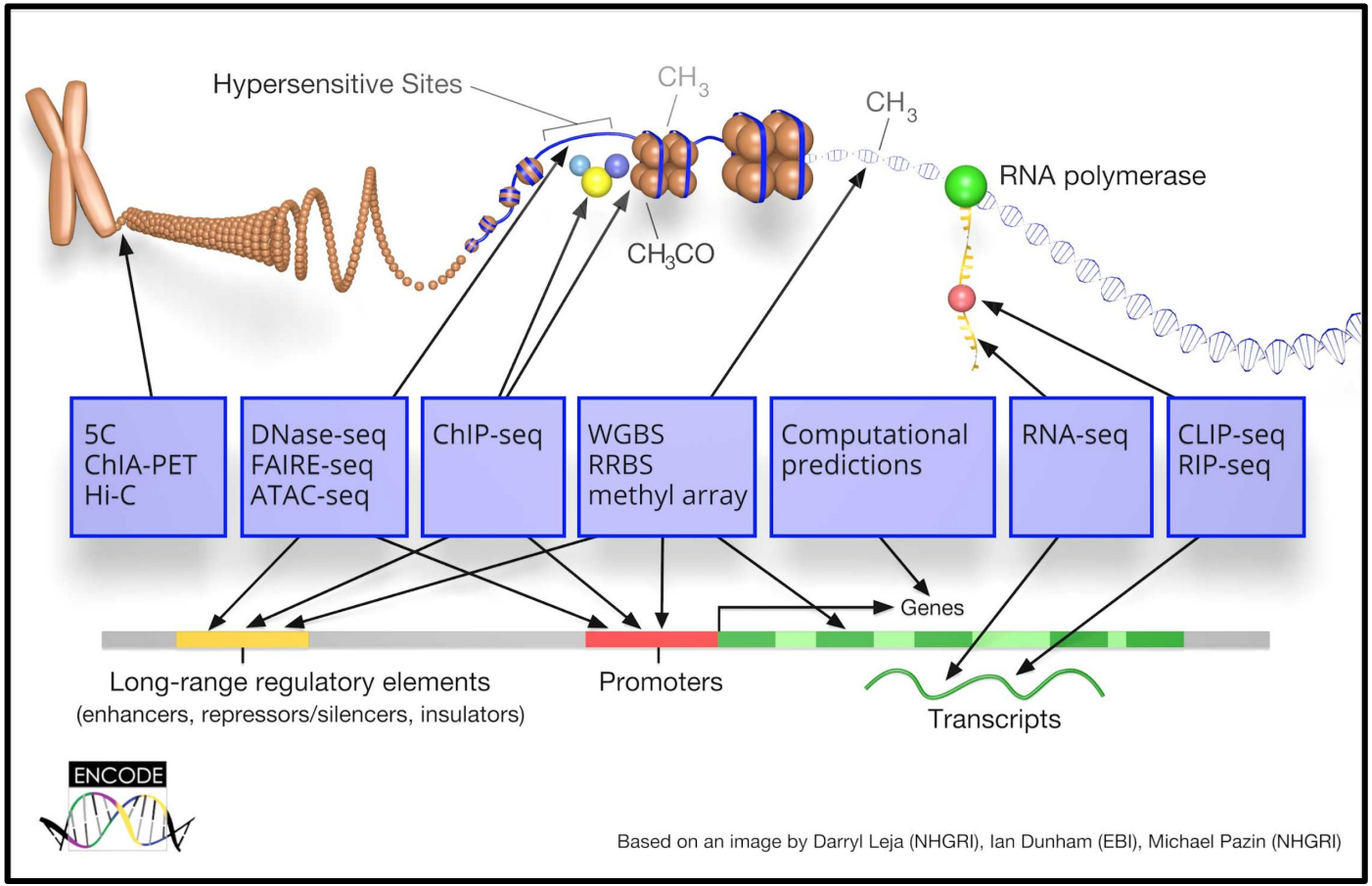
Replicating our work took four years and 100,000 worms but brought surprising discoveries, explain **Gordon J. Lithgow**, **Monica Driscoll** and **Patrick Phillips**.

Nature. 2017 Aug 22;548(7668):387-388

How do you turn irreproducible experiment into reproducible?

* methods standardization
* experimental replication
* documentation

Encyclopedia of DNA Elements



Encyclopedia of DNA Elements

National Human Genome Research Institute

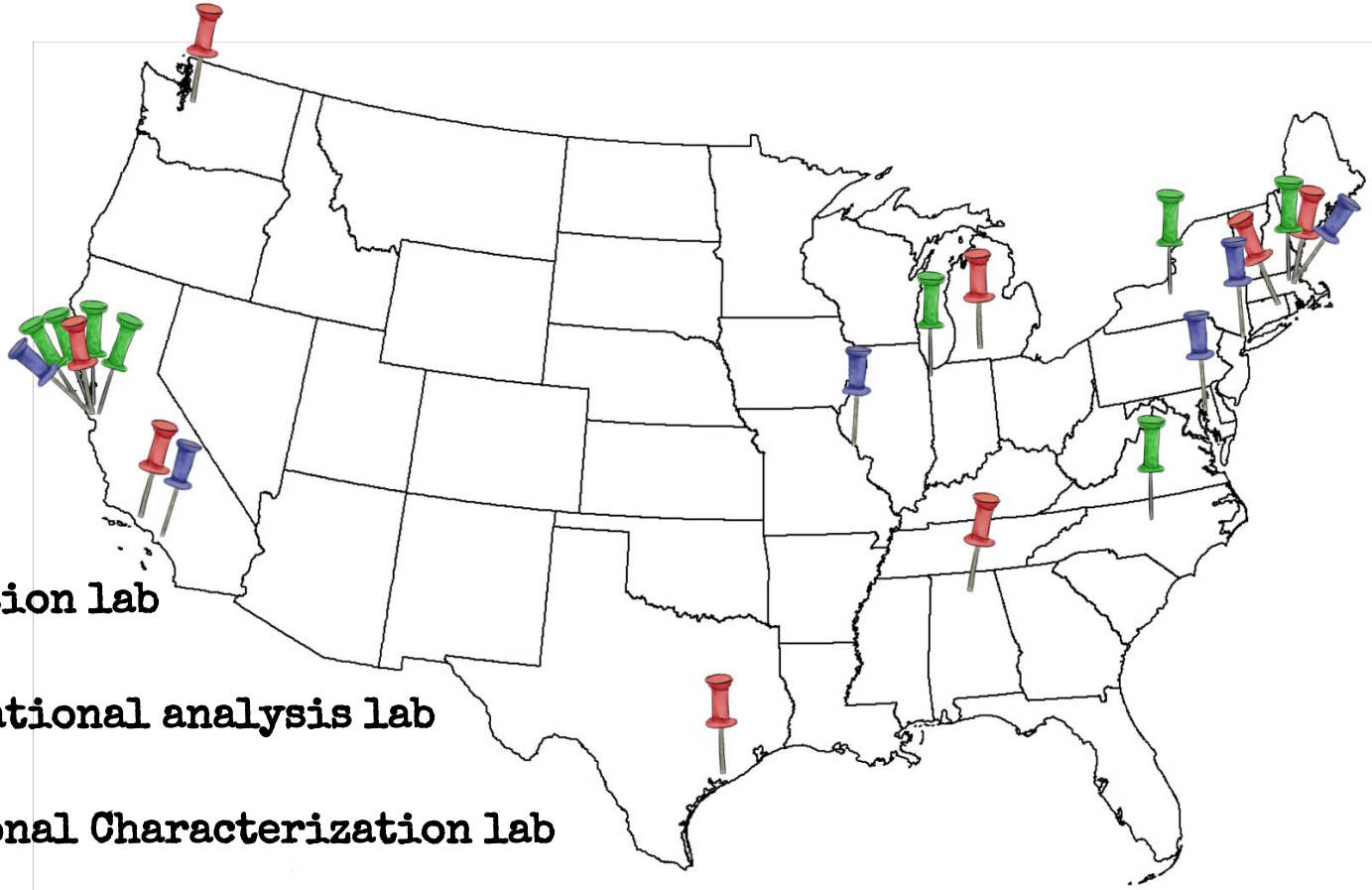
2003–2007 ENCODE

2007–2011 ENCODE PHASE 2

2012–2016 ENCODE PHASE 3

2017–2021 ENCODE PHASE 4

Encyclopedia of DNA Elements (phase IV)

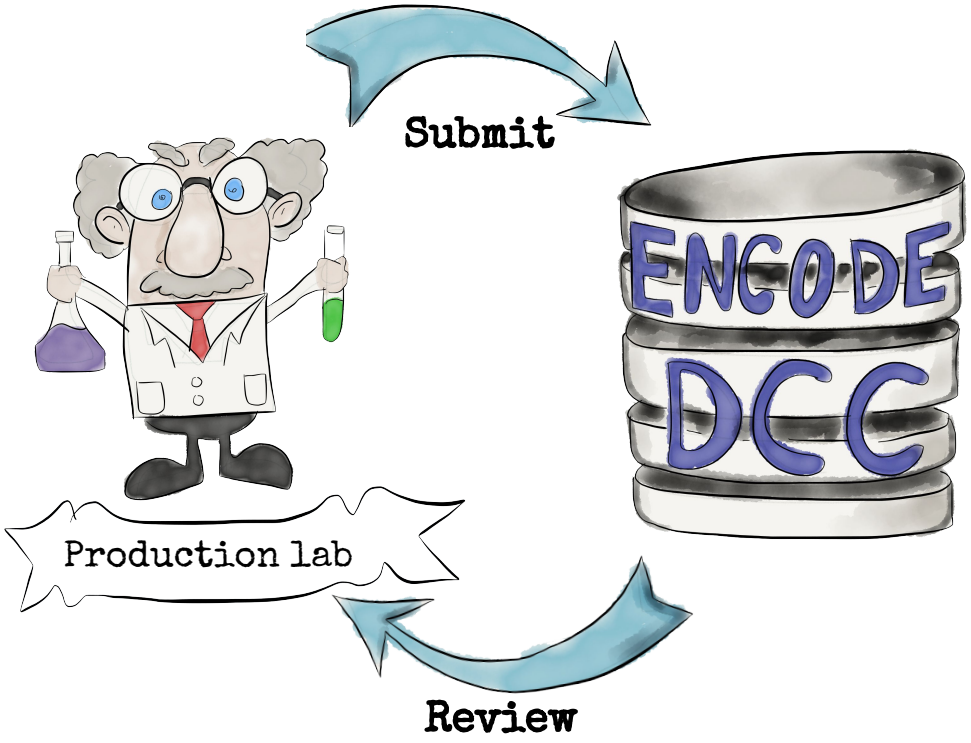


Production lab

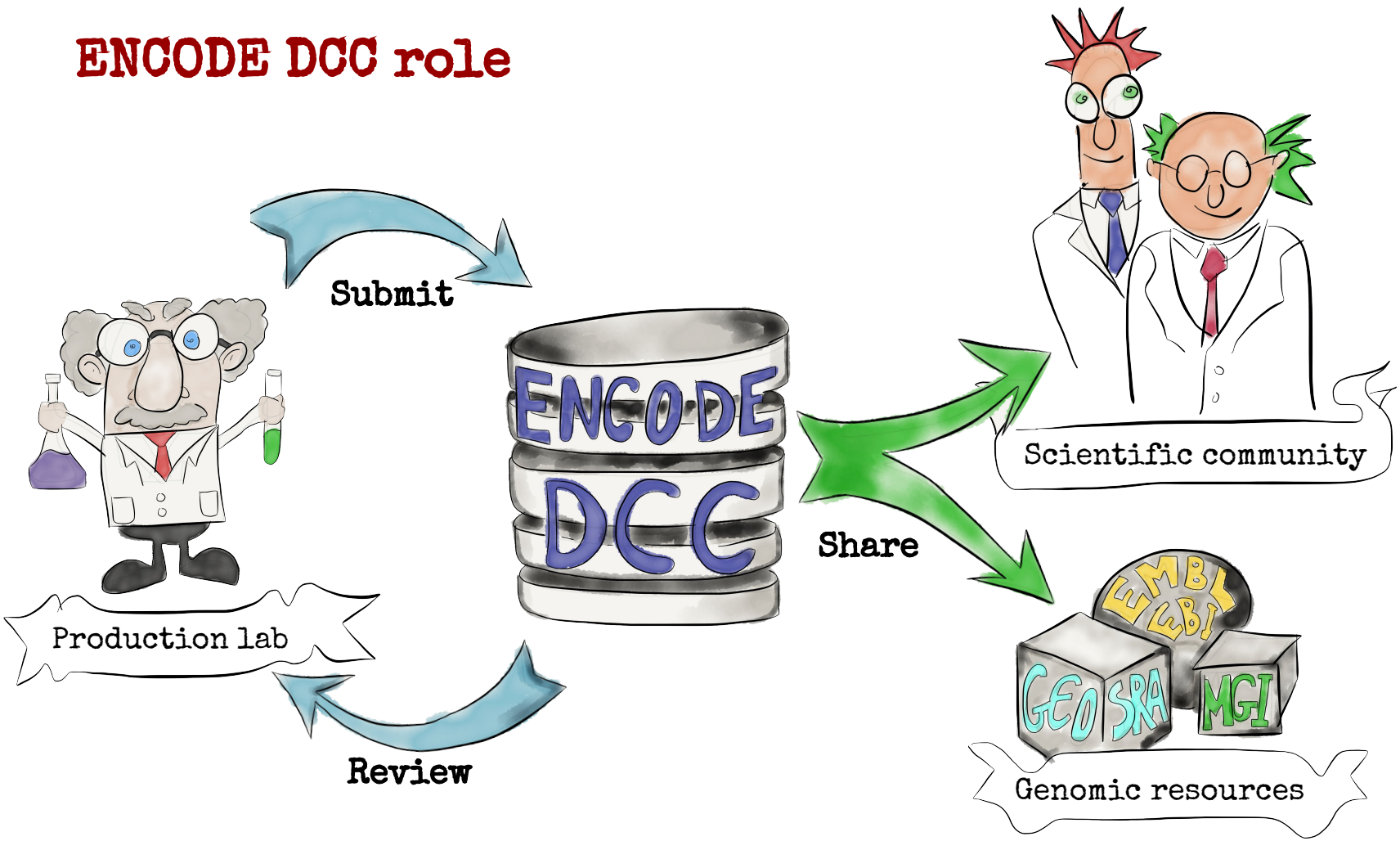
Computational analysis lab

Functional Characterization lab

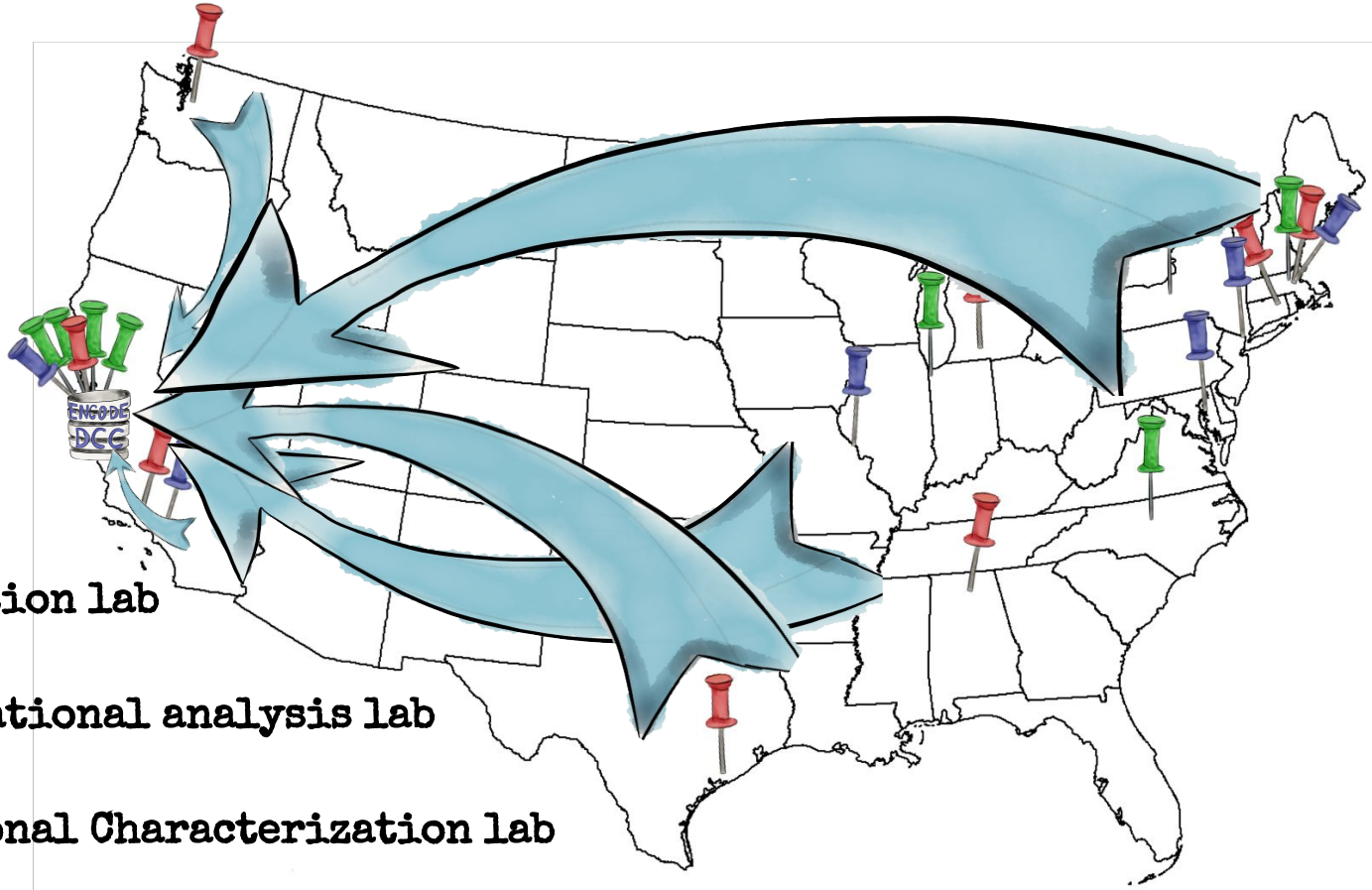
ENCODE DCC role



ENCODE DCC role



Data flow

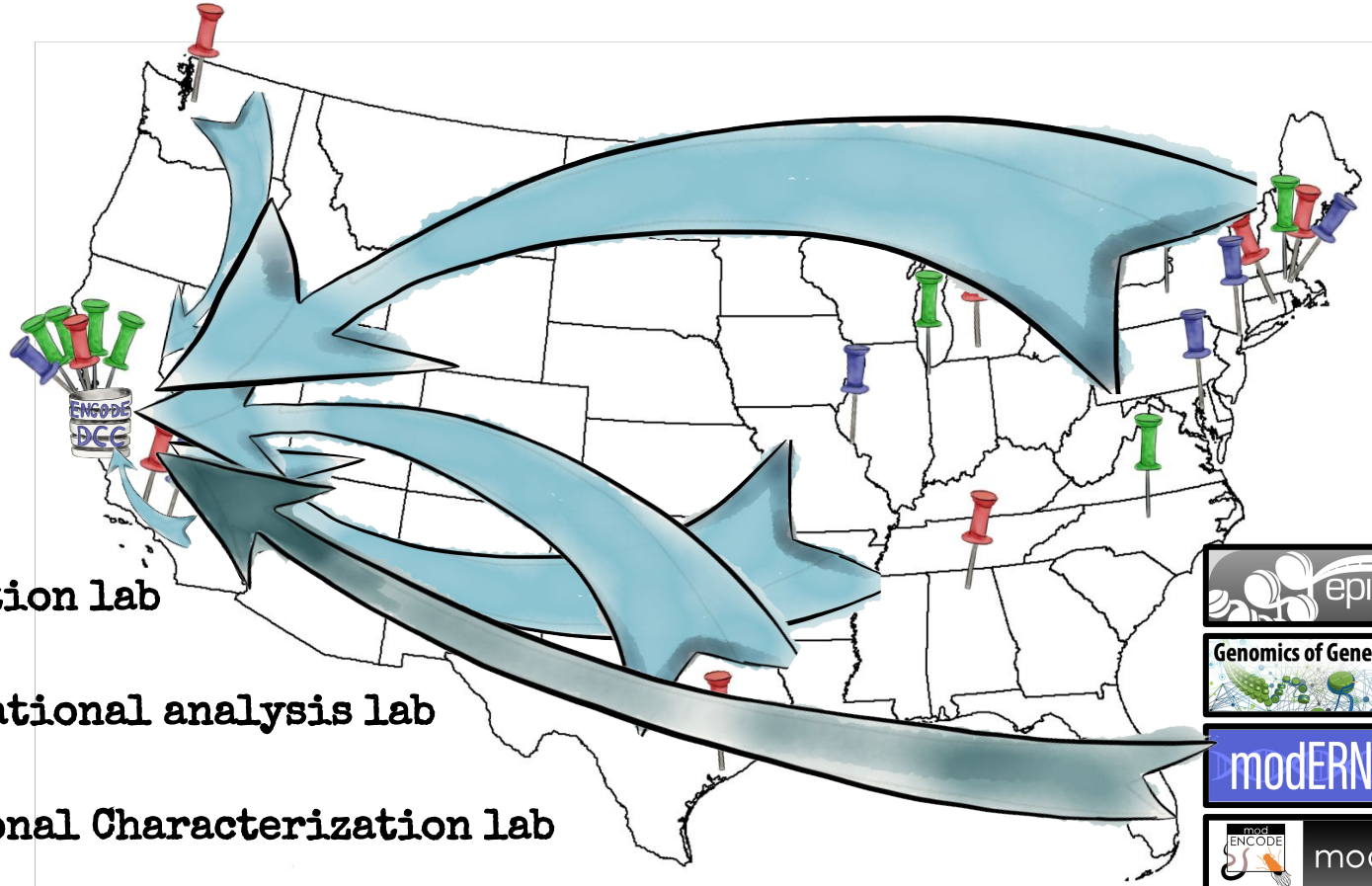


Production lab

Computational analysis lab

Functional Characterization lab

Data flow



Production lab

Computational analysis lab

Functional Characterization lab

ENCODE

ROADMAP epigenomics PROJECT

Genomics of Gene Regulation

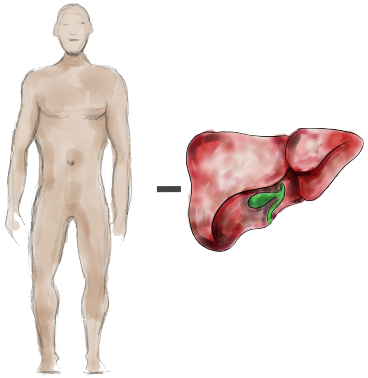
modERN

mod ENCODE modENCODE

From the lab to the portal

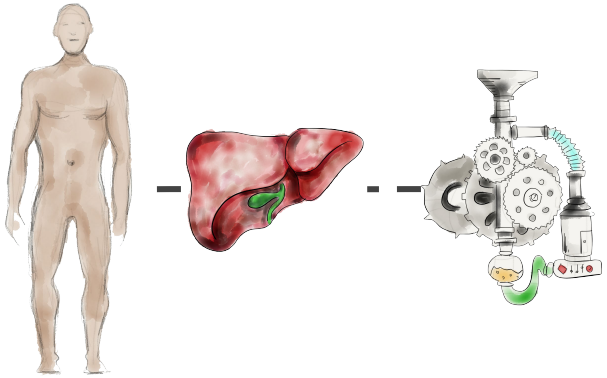
Donor

From the lab to the portal



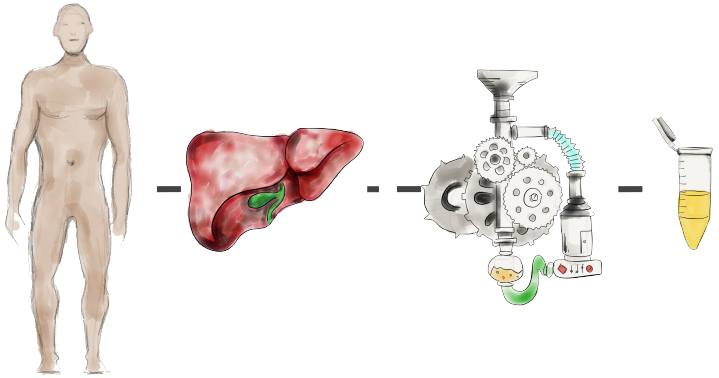
Donor Sample

From the lab to the portal



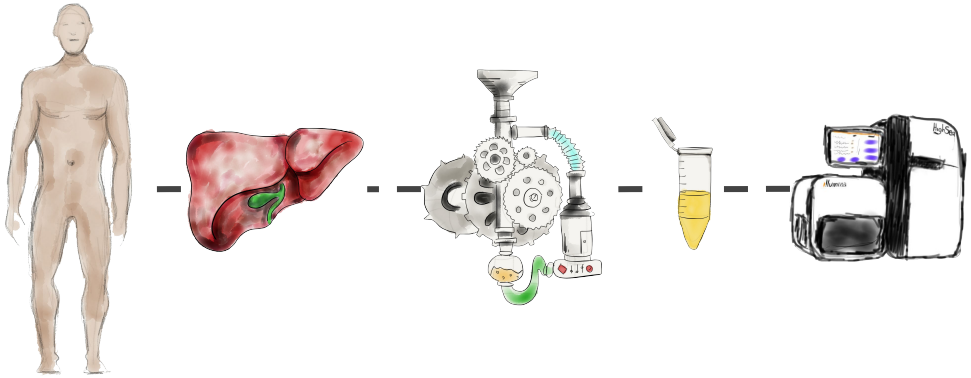
Donor Sample Experiment

From the lab to the portal



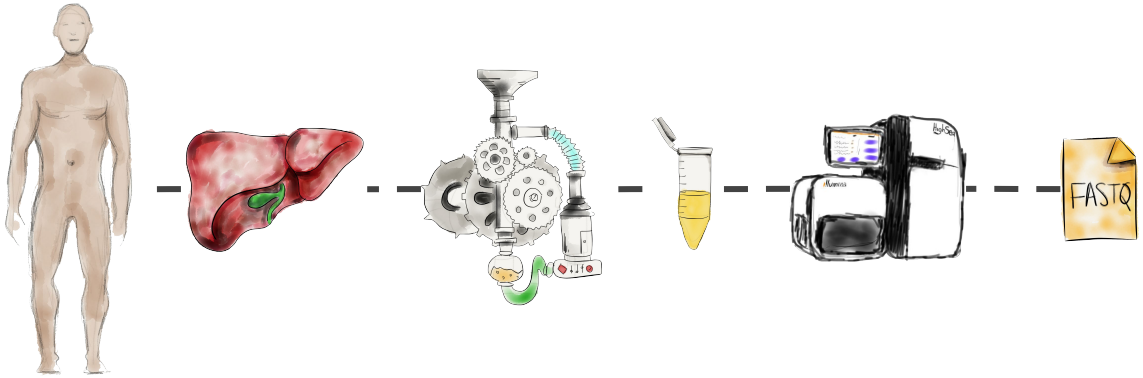
Donor Sample Experiment Library

From the lab to the portal



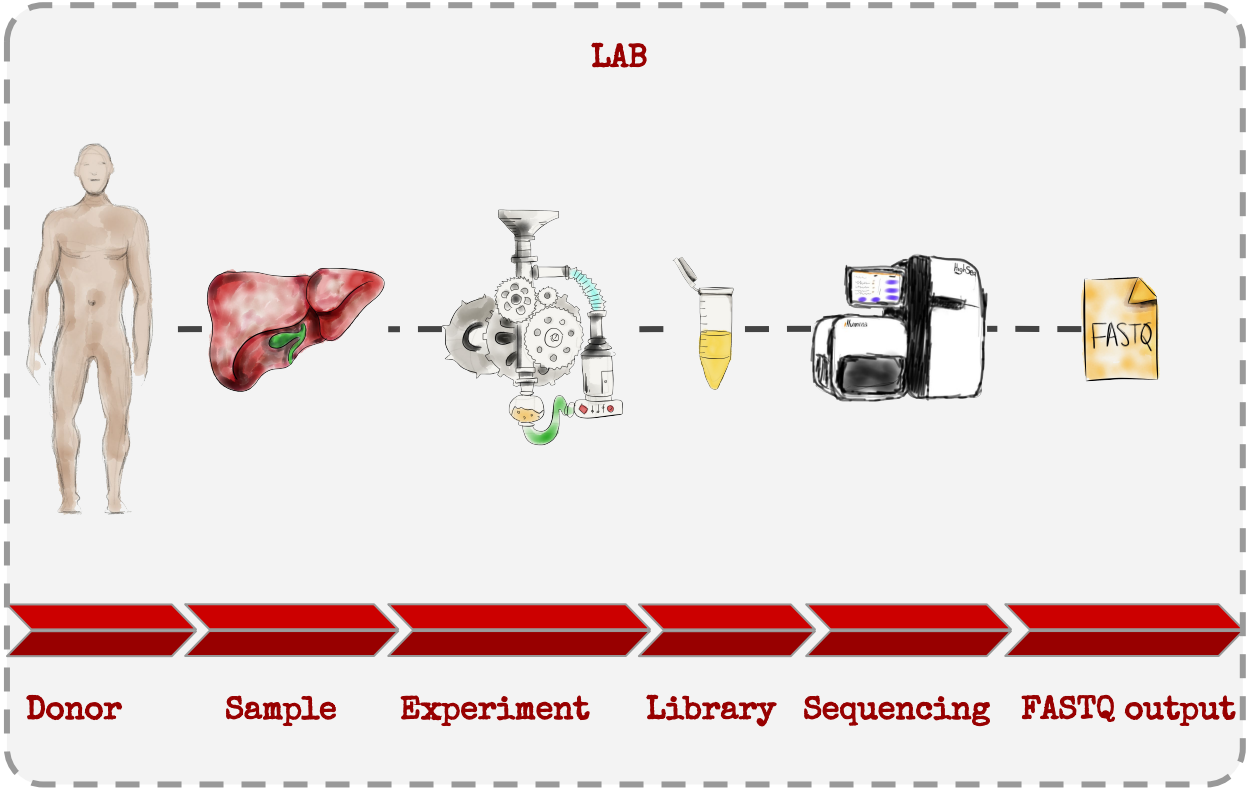
Donor Sample Experiment Library Sequencing

From the lab to the portal

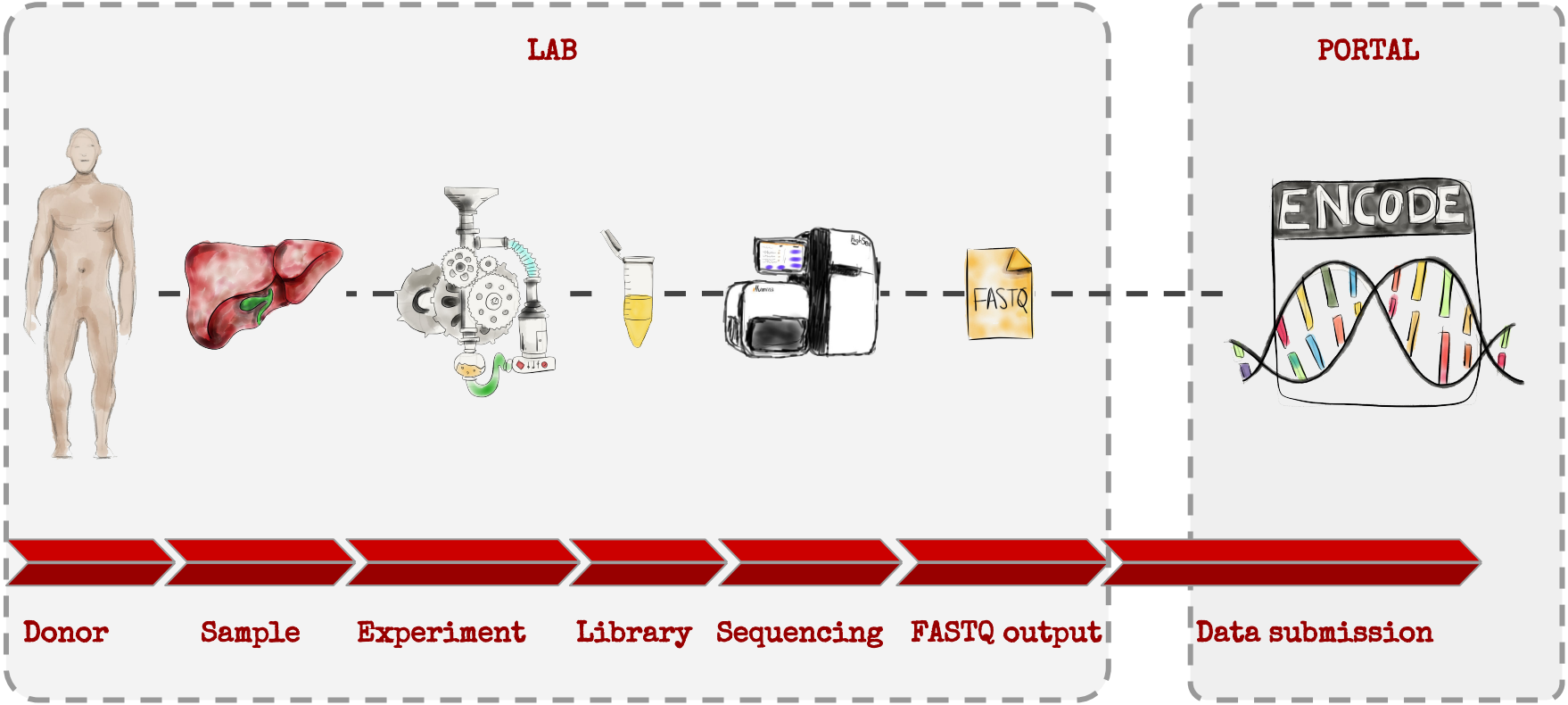


Donor Sample Experiment Library Sequencing FASTQ output

From the lab to the portal

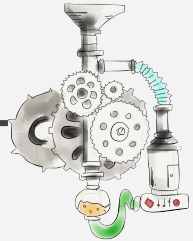


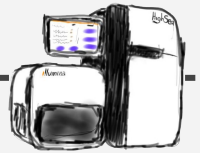
From the lab to the portal



From the lab to the portal

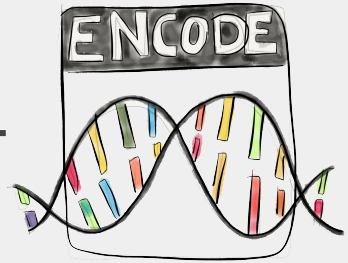
LAB





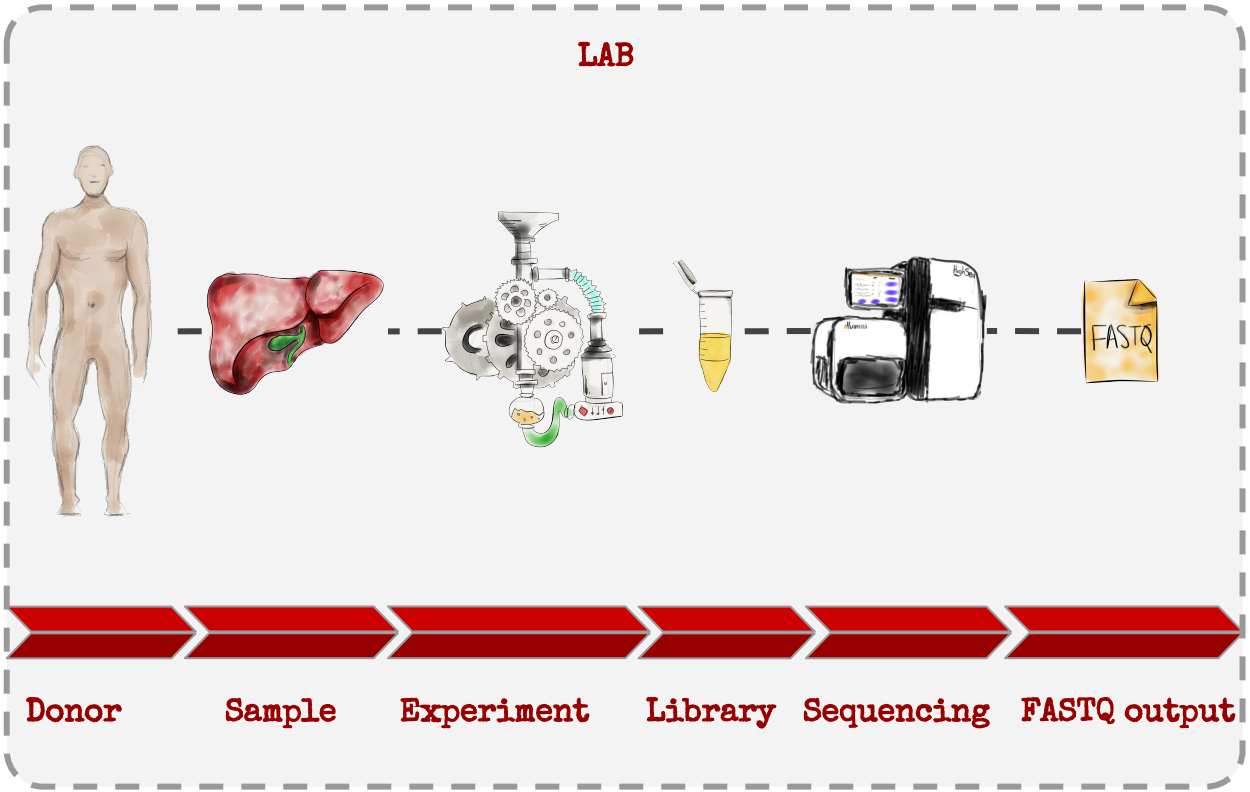
STANDARDS

PORTAL



Donor → Sample → Experiment → Library → Sequencing → FASTQ output → Data submission

Data categories (JSON objects)



Donor

* female
* 51 years old

HumanDonor.json

ENCDO271OUW

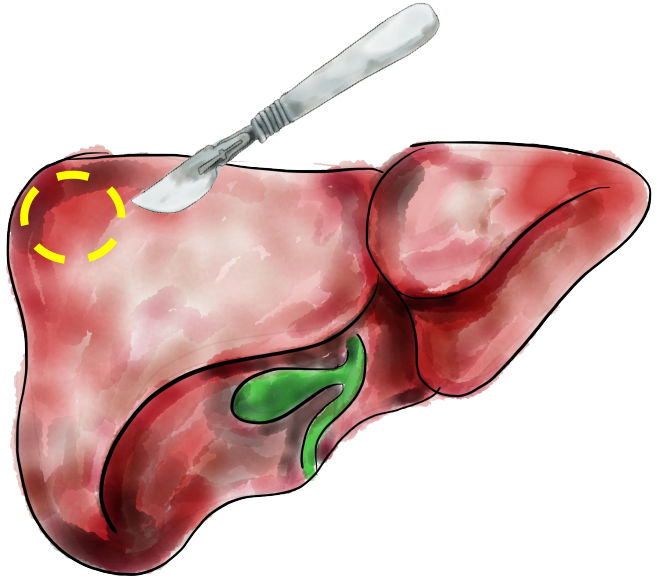
Status:
released

Donor information

Accession:	ENCDO271OUW
Aliases:	gtex:ENC-DEJ, gtex:PT-1LVAN, gtex:ENC-004, bradley-bernstein:Donor GTEX-1LVAN
Donor external identifiers:	GEO:SAMN05897787
Species:	*Homo sapiens*
Life stage:	Adult
Age:	51 year
Sex:	Female
Tags:	GTEx

Tissue sample

* right lobe of liver



Biosample.json

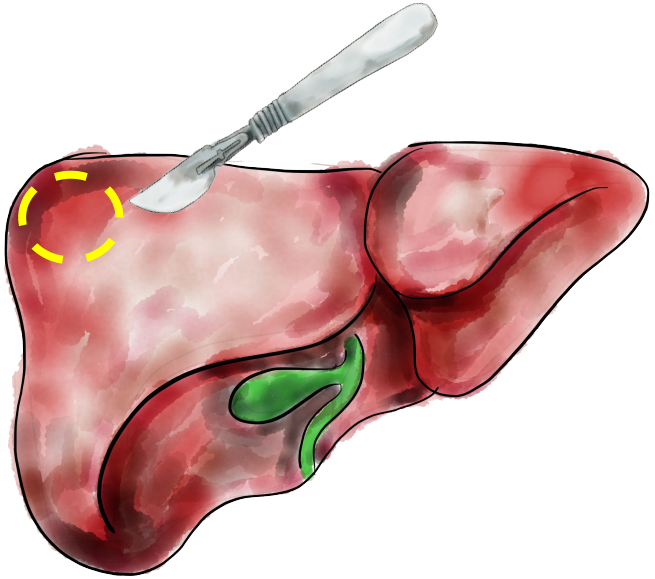
ENCBS150EDI / tissue

Status: **released**

Summary

Term name:	right lobe of liver
Term ID:	UBERON:0001114
Summary:	*Homo sapiens* female adult (53 years) right lobe of liver tissue
Parent of biosamples:	ENCBS220FFU, ENCBS688NPW, ENCBS594VBH, ENCBS796JPW, ENCBS773RMK, ENCBS021IAC, ENCBS904HZU, ENCBS290WWG, ENCBS984AKV, ENCBS682CWZ, ENCBS536THV, ENCBS405OIZ, ENCBS054RZZ, ENCBS034NFF

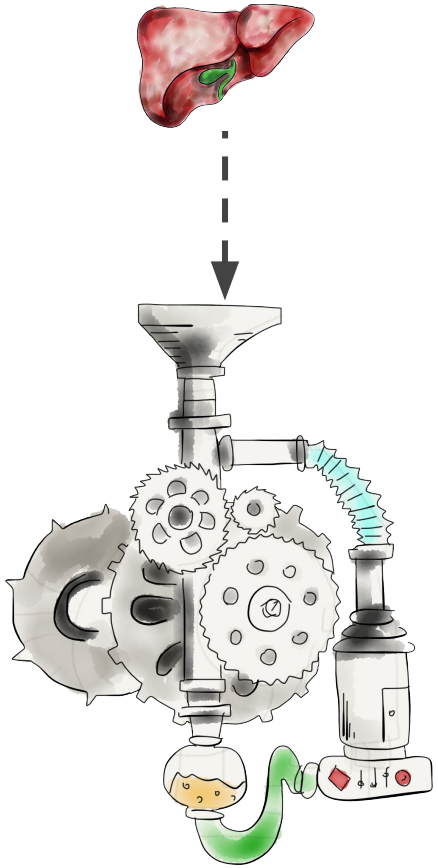
UBERON:0001114



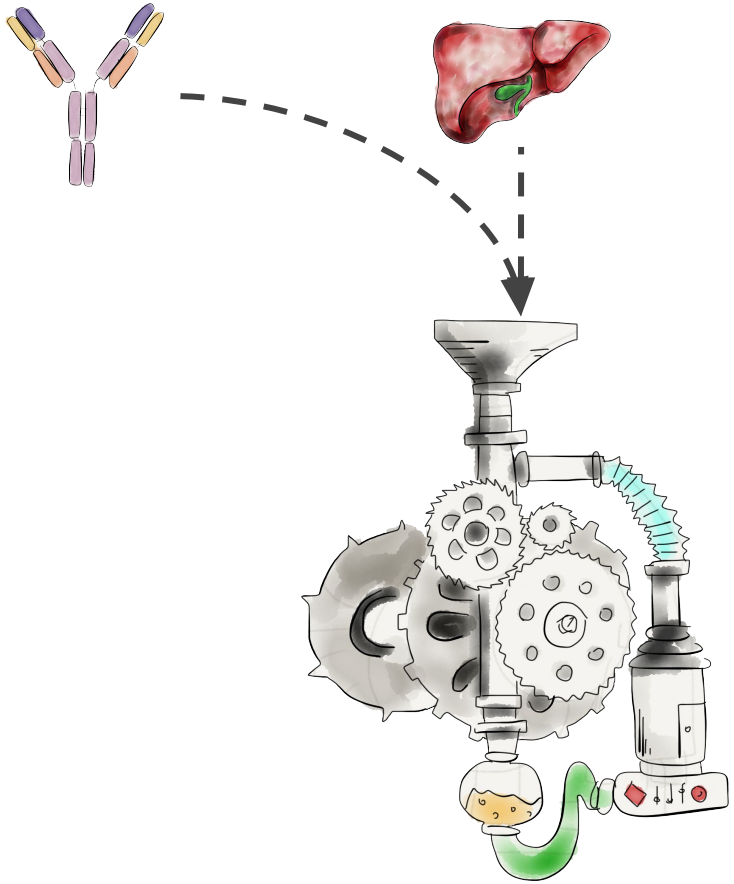
Experiment

CTCF ChIP-seq

OBI:0000716

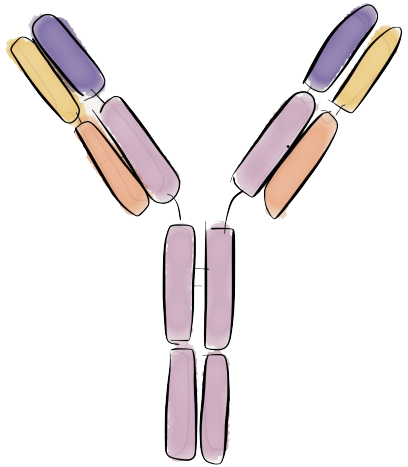


Experiment

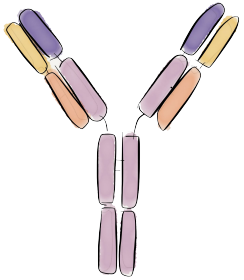


Antibody

* Vendor
* Product ID
* Lot ID



AntibodyLot.json



ENCAB635SXP

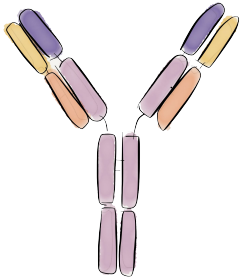
Antibody against *Homo sapiens* CTCF

Status: released | 2 | 1

Homo sapiens	at least one cell type or tissue	**Awaiting characterization**

Source (vendor):	ABclonal
Product ID:	A1133
Lot ID:	A1133
Targets:	CTCF (*Homo sapiens*)
Host:	Rabbit
Clonality:	Polyclonal
Purification:	Affinity
Isotype:	IgG
Antigen description:	Recombinant protein of human CTCF

AntibodyLot.json



ENCAB635SXP

Antibody against *Homo sapiens* CTCF

Status: released | 2 | 1

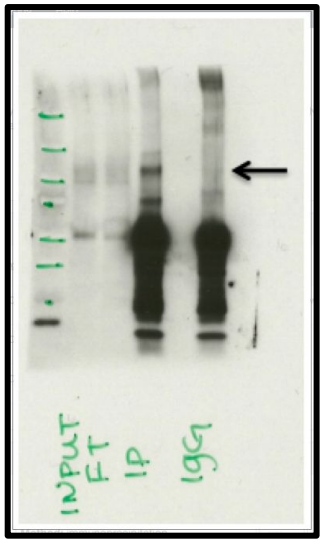
Homo sapiens	at least one cell type or tissue	**Awaiting characterization**

Source (vendor):	ABclonal
Product ID:	A1133
Lot ID:	A1133
Targets:	CTCF (*Homo sapiens*)
Host:	Rabbit
Clonality:	Polyclonal
Purification:	Affinity
Isotype:	IgG
Antigen description:	Recombinant protein of human CTCF

Antibody standards

Primary characterization	Secondary characterization

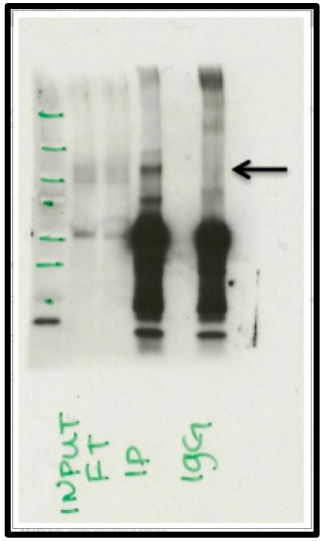
Antibody standards

Primary characterization	Secondary characterization
Western blot or IP Western 	

Antibody standards

Primary characterization

Western blot or IP Western

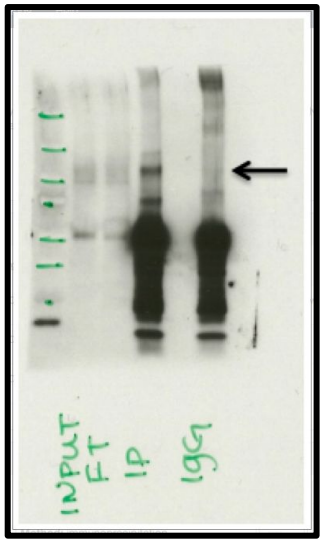
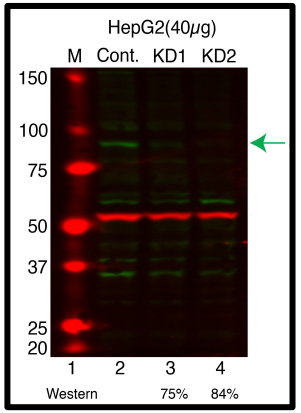


Secondary characterization

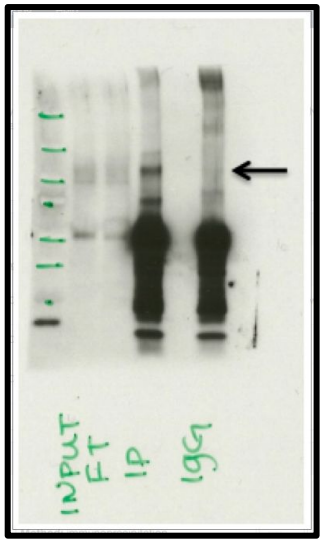
- Mass spectrometry

Spectrum	Name of Protein	Count of Peptides	Ratio(CTCF/IgG Control)
CTCF_Band A	Heterogeneous nuclear ribonucleoprotein M (Fragment)	15	NOT IN CONTROL IP
CTCF_Band A	C4A protein	13	NOT IN CONTROL IP
CTCF_Band A	Isoform 2 of Nucleolar RNA helicase 2	10	NOT IN CONTROL IP
CTCF_Band A	Prelamin-A/C	10	NOT IN CONTROL IP
CTCF_Band A	Probable ATP-dependent RNA helicase DDX5	10	NOT IN CONTROL IP
CTCF_Band A	Transcriptional repressor CTCF	10	NOT IN CONTROL IP
CTCF_Band A	Heat shock cognate 71 kDa protein	9	NOT IN CONTROL IP
CTCF_Band A	Isoform 2 of Probable ATP-dependent RNA helicase DDX17	9	NOT IN CONTROL IP
CTCF_Band A	N-acetyltransferase 10	8	NOT IN CONTROL IP
CTCF_Band A	Splicing factor 3B subunit 3	8	NOT IN CONTROL IP
CTCF_Band A	SWI/SNF-related matrix-associated actin-dependent regulator of chromatin subfamily A	8	NOT IN CONTROL IP
CTCF_Band A	Isoform Short of Heterogeneous nuclear ribonucleoprotein U	7	NOT IN CONTROL IP
CTCF_Band A	X-ray repair complementing defective repair in Chinese hamster cells 6 (Ku autoantigen,	7	NOT IN CONTROL IP
CTCF_Band A	Nucleolar protein 56 (Fragment)	6	NOT IN CONTROL IP
CTCF_Band A	ATP-dependent RNA helicase A	5	NOT IN CONTROL IP
CTCF_Band A	Interleukin enhancer-binding factor 3	5	NOT IN CONTROL IP
CTCF_Band A	Isoform UBF2 of Nucleolar transcription factor 1	5	NOT IN CONTROL IP
CTCF_Band A	X-ray repair cross-complementing protein 5	5	NOT IN CONTROL IP
CTCF_Band A	DNA topoisomerase 1	4	NOT IN CONTROL IP

Antibody standards

Primary characterization	Secondary characterization
Western blot or IP Western 	• Mass spectrometry • siRNA or shRNA against the mRNA of the target protein 

Antibody standards

Primary characterization	Secondary characterization
Western blot or IP Western 	• Mass spectrometry • siRNA or shRNA against the mRNA of the target protein • ChIP-seq data from a previously characterized antibody • • •

Antibody standards

Primary characterization √	Secondary characterization √

characterized to standards

Homo sapiens

C4-2B, HepG2, K562, RWPE2, LNCAP, GM12878, HCT116, 22Rv1, VCaP, RWPE1, HeLa-S3, Panc1, MCF-7

Antibody standards

Primary characterization √	Secondary characterization √

CHALLENGE

● **characterized to standards**

Homo sapiens

C4-2B, HepG2, K562, RWPE2, LNCAP, GM12878, HCT116, 22Rv1, VCaP, RWPE1, HeLa-S3, Panc1, MCF-7

Experiment.json

Experiment summary for ENCSR911GFJ

Status: **released** 2

Summary

Assay:	ChIP-seq
Target:	CTCF
Biosample summary:	*Homo sapiens* right lobe of liver female adult (53 years)
Biosample Type:	tissue
Replication type:	unreplicated
Description:	CTCF ChIP-seq on human right lobe of liver
Nucleic acid type:	DNA
Size range:	450-650
Strand specificity:	Non-strand-specific
Platform:	Illumina HiSeq 4000
Controls:	ENCSR336OPU

Attribution

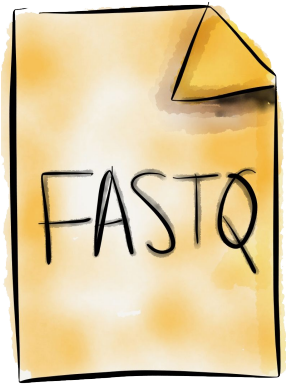
ENCODE PHASE 3

Lab:	Michael Snyder, Stanford
Award:	U54HG006996 (Michael Snyder, Stanford)
Project:	ENCODE
External resources:	GEO:GSE105829
Aliases:	michael-snyder:ChIPss-674
Date submitted:	December 6, 2016
Date released:	December 13, 2016
Tags:	GTEx

Replicates

Biological replicate	Technical replicate	Summary	Biosample	Antibody	Library
1	1	female adult (53 years) right lobe of liver tissue	ENCBS904HZU	ENCAB830JLB	ENCLB331LXY

Raw data file(s)



* file format and integrity
* data duplication prevention

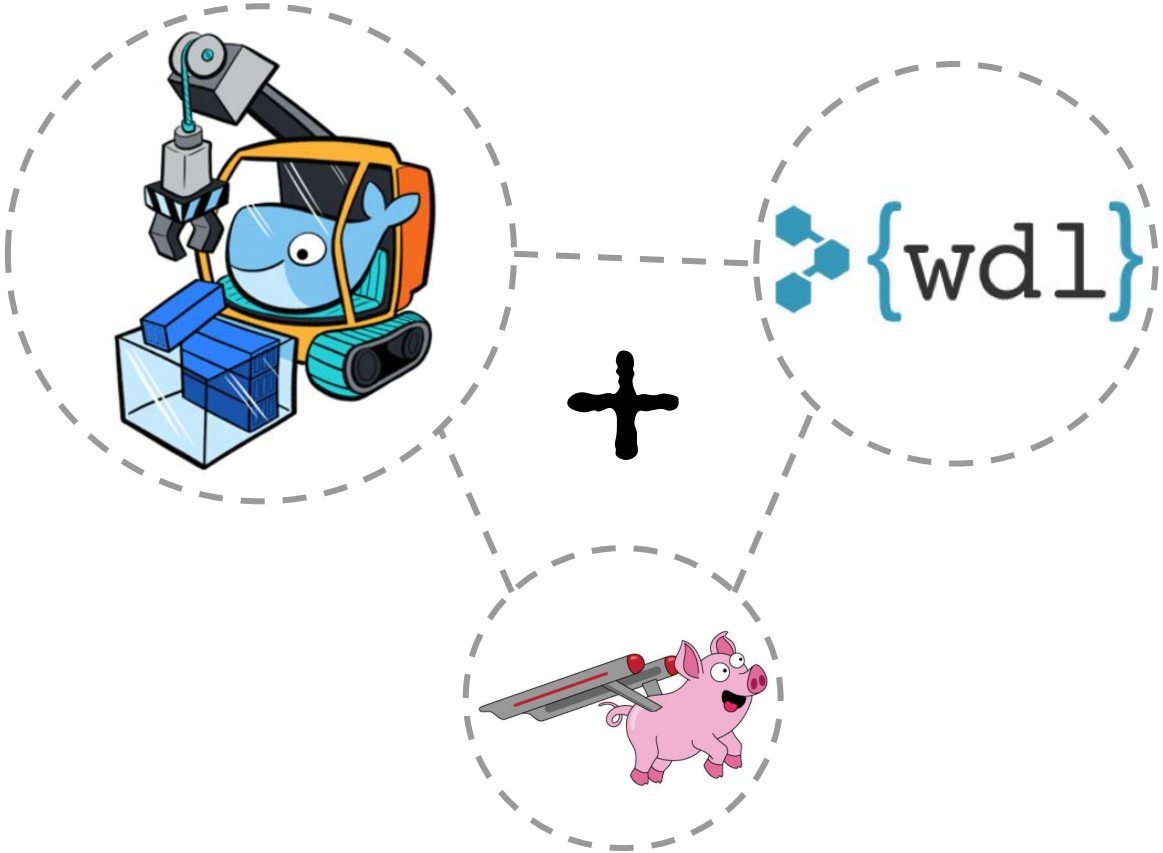
Raw data file(s)

* file format and integrity
* data duplication prevention

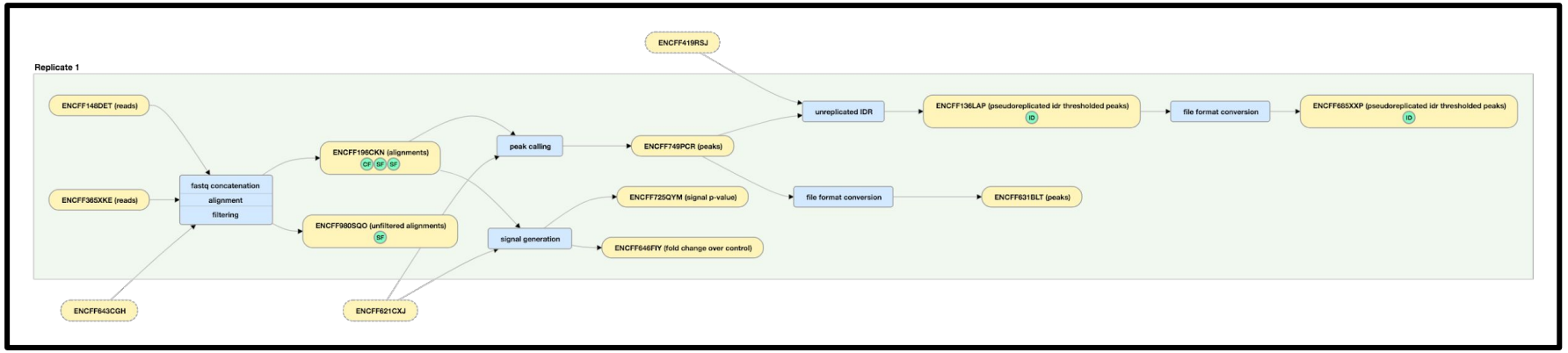
CHALLENGE

Database (Oxford). 2018 Jan 1;2018. Doi: 10.1093/database/bay008

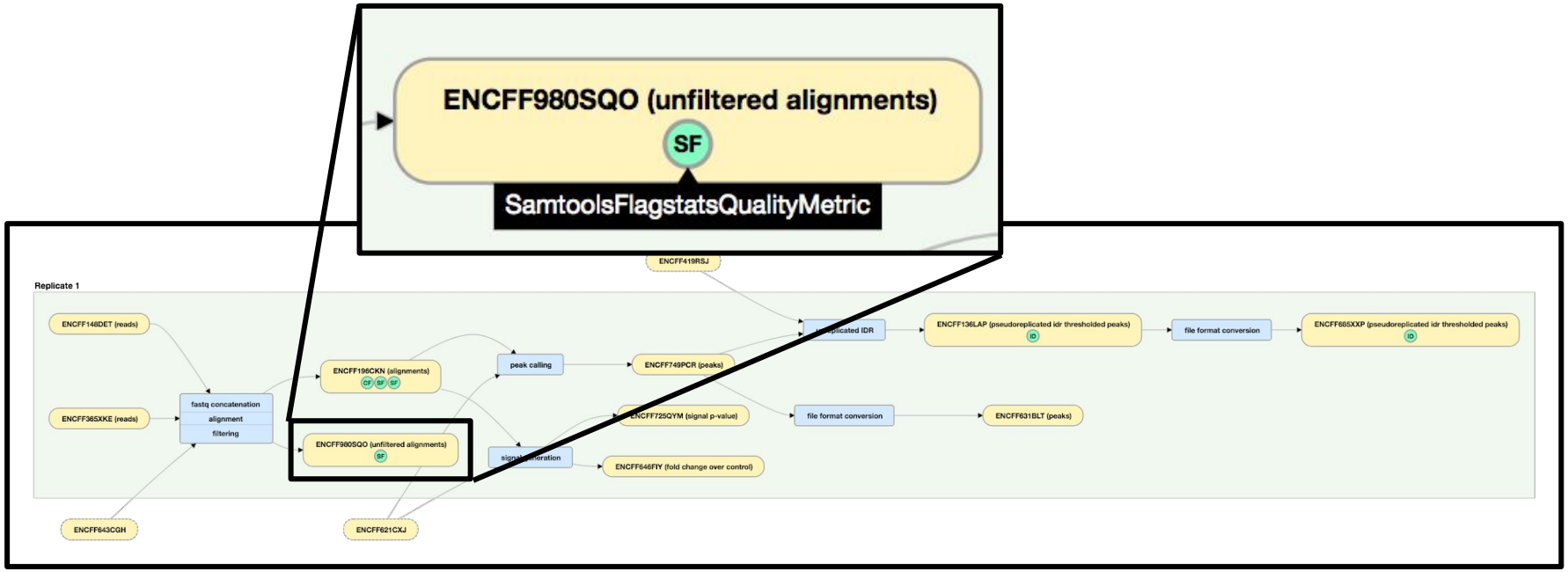
Uniform processing pipelines



Analysis file graph



QC metrics


ENCFF980SQO (unfiltered alignments)
SF
SamtoolsFlagstatsQualityMetric
Replicate 1
ENCFF148DET (reads)
ENCFF365XKE (reads)
ENCFF643CGH
fastq concatenation
alignment
filtering
ENCFF196CKN (alignments)
ENCFF980SQO (unfiltered alignments)
ENCFF621CXJ
peak calling
ENCFF749PCR (peaks)
ENCFF725QYM (signal p-value)
ENCFF646FIY (fold change over control)
ENCFF419RSJ
file format conversion
ENCFF631BLT (peaks)
ENCFF136LAP (pseudoreplicated idr thresholded peaks)
file format conversion
ENCFF685XXP (pseudoreplicated idr thresholded peaks)

Audits

* metadata integrity
* ENCODE consortium data standards

Badges

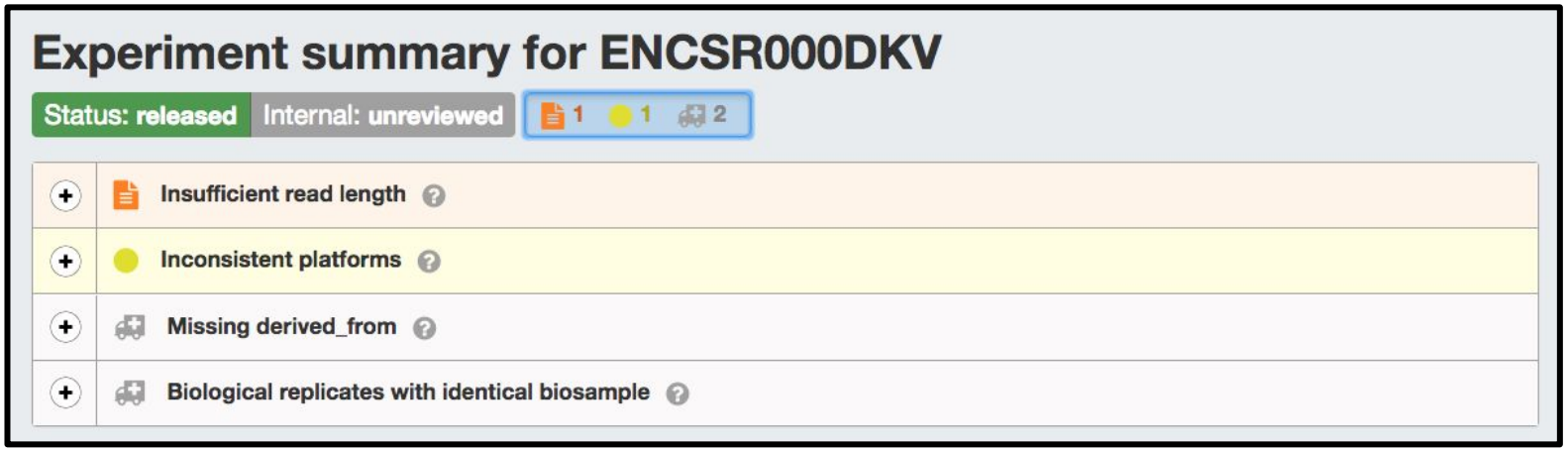
RED for ERROR 

ORANGE for NOT_COMPLIANT 

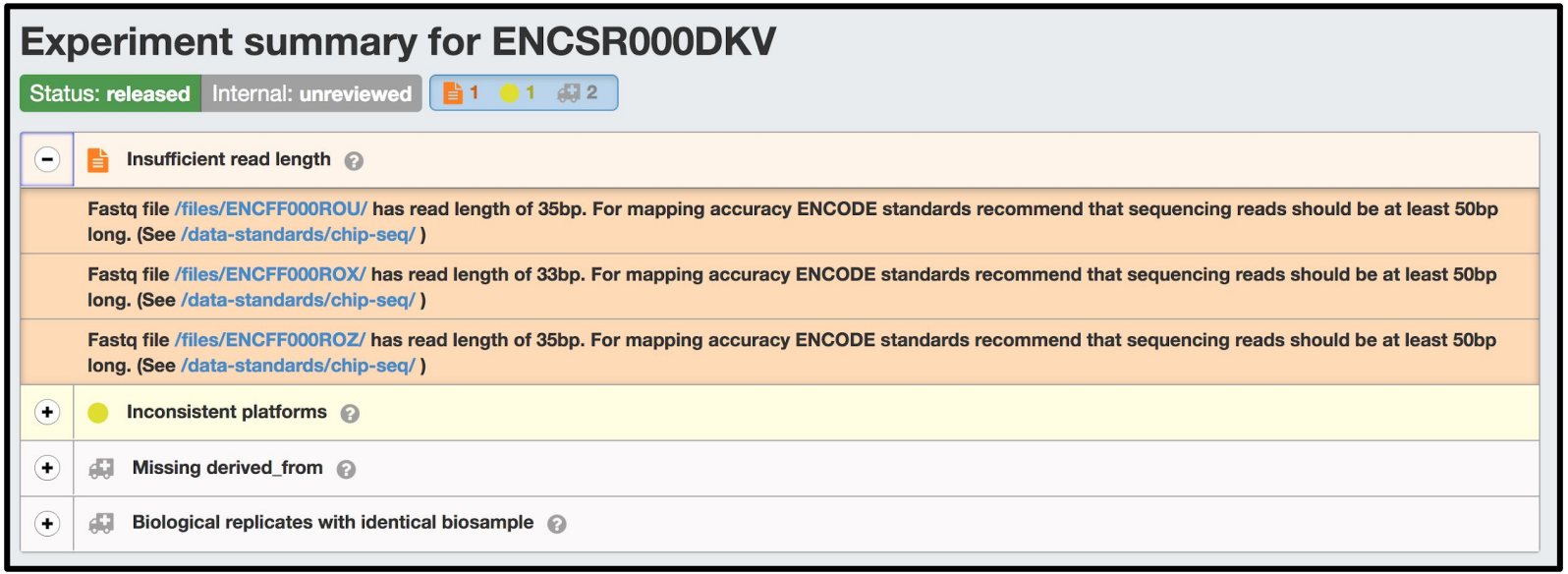
YELLOW for WARNING 

GREY for INTERNAL DCC 

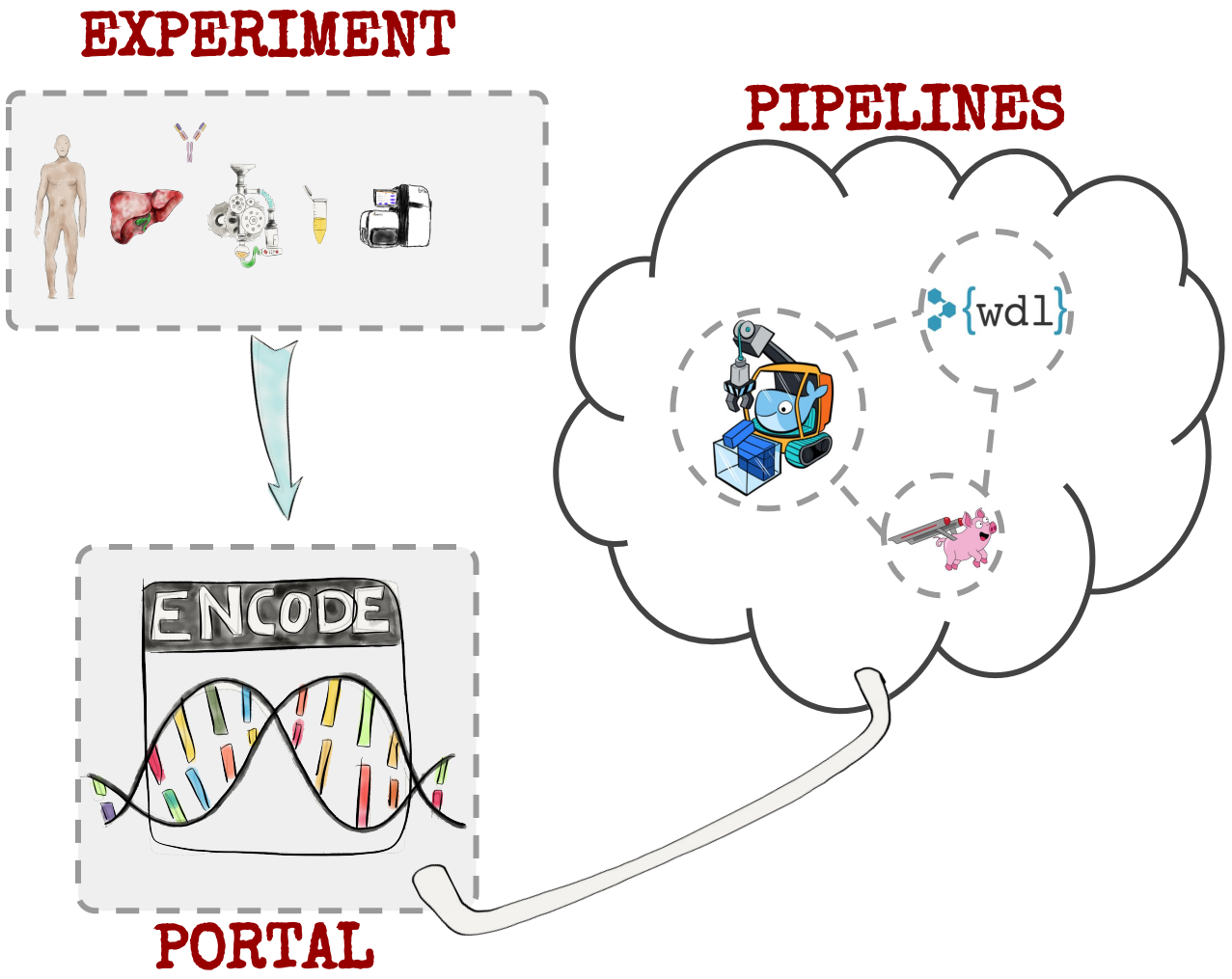
Badges


Experiment summary for ENCSR000DKV
Status: released
Internal: unreviewed
1
1
2
Insufficient read length
Inconsistent platforms
Missing derived_from
Biological replicates with identical biosample

Badges

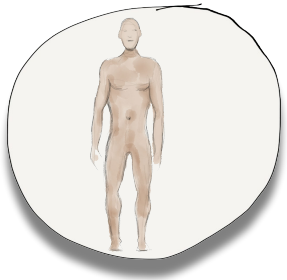

Experiment summary for ENCSR000DKV
Status: released
Internal: unreviewed
1
1
2
Insufficient read length
Fastq file /files/ENCFF000ROU/ has read length of 35bp. For mapping accuracy ENCODE standards recommend that sequencing reads should be at least 50bp long. (See /data-standards/chip-seq/)
Fastq file /files/ENCFF000ROX/ has read length of 33bp. For mapping accuracy ENCODE standards recommend that sequencing reads should be at least 50bp long. (See /data-standards/chip-seq/)
Fastq file /files/ENCFF000ROZ/ has read length of 35bp. For mapping accuracy ENCODE standards recommend that sequencing reads should be at least 50bp long. (See /data-standards/chip-seq/)
Inconsistent platforms
Missing derived_from
Biological replicates with identical biosample

Data model
Ontology
Audits (badges)
AB standards
File validation
Uniform Pipelines



Data model

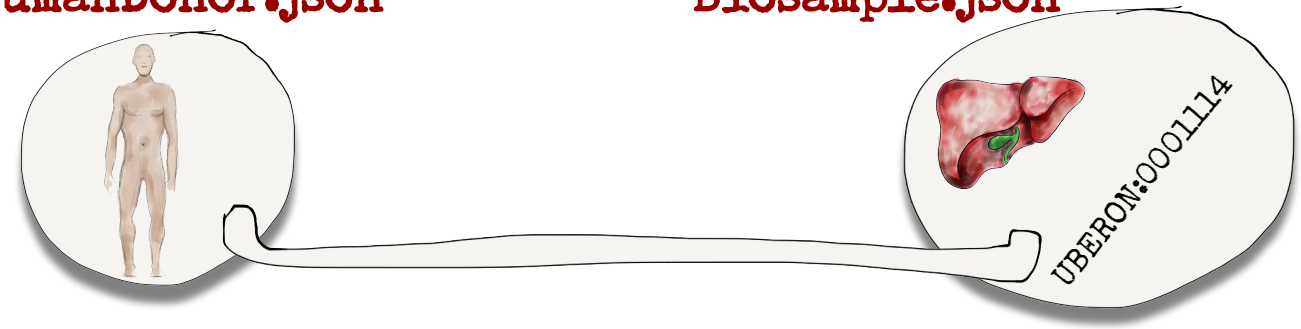
HumanDonor.json



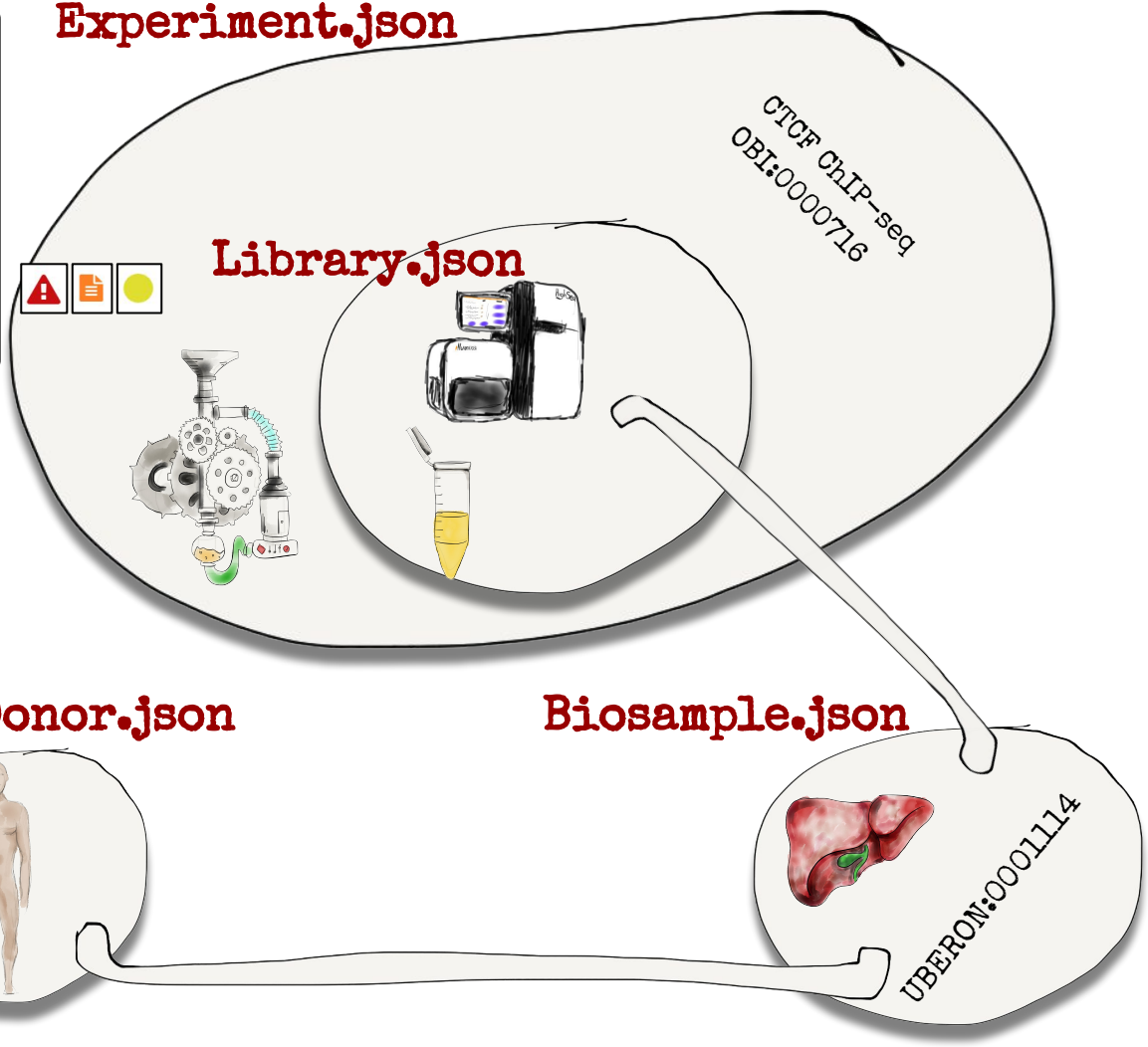
Data model
Ontology

HumanDonor.json

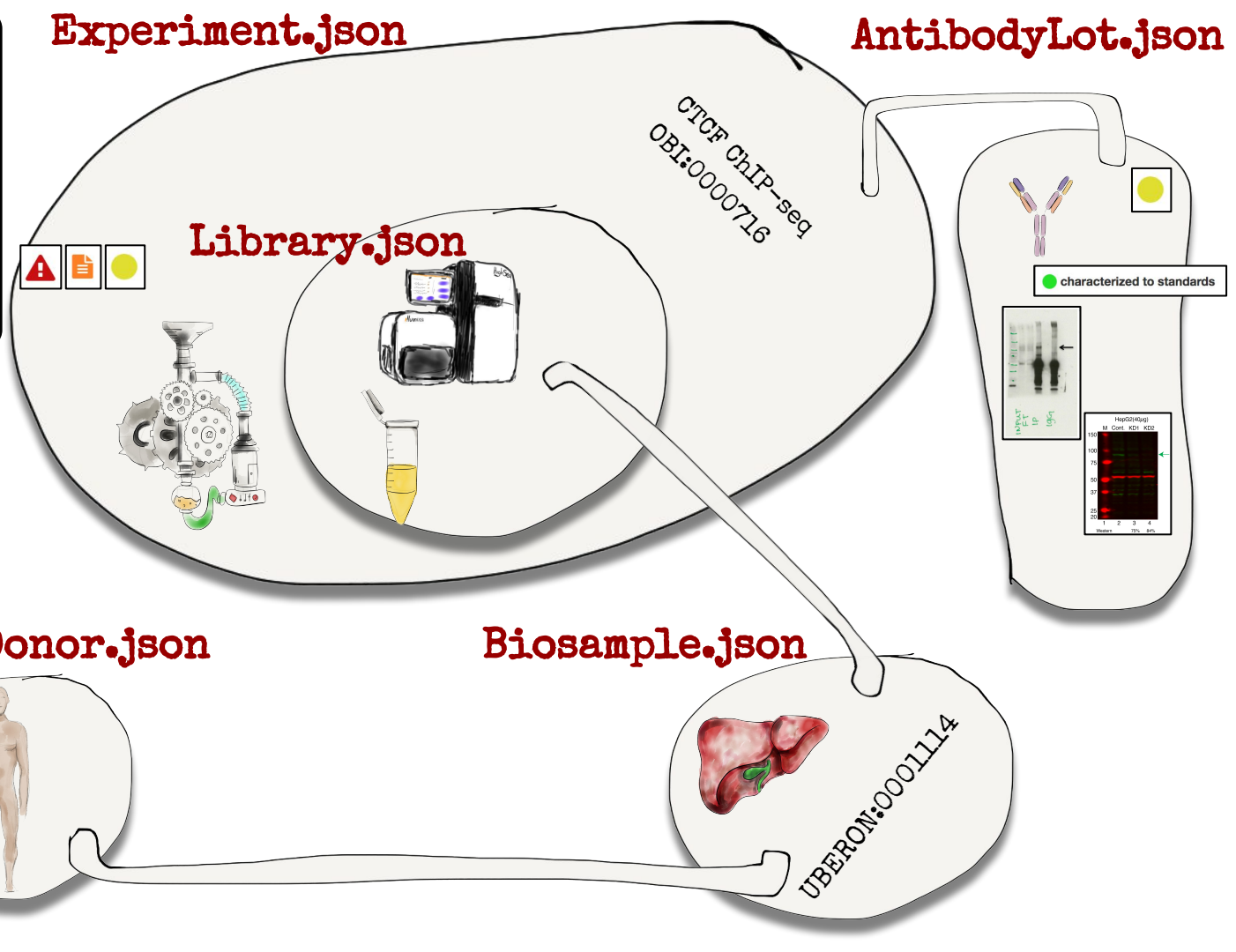
Biosample.json



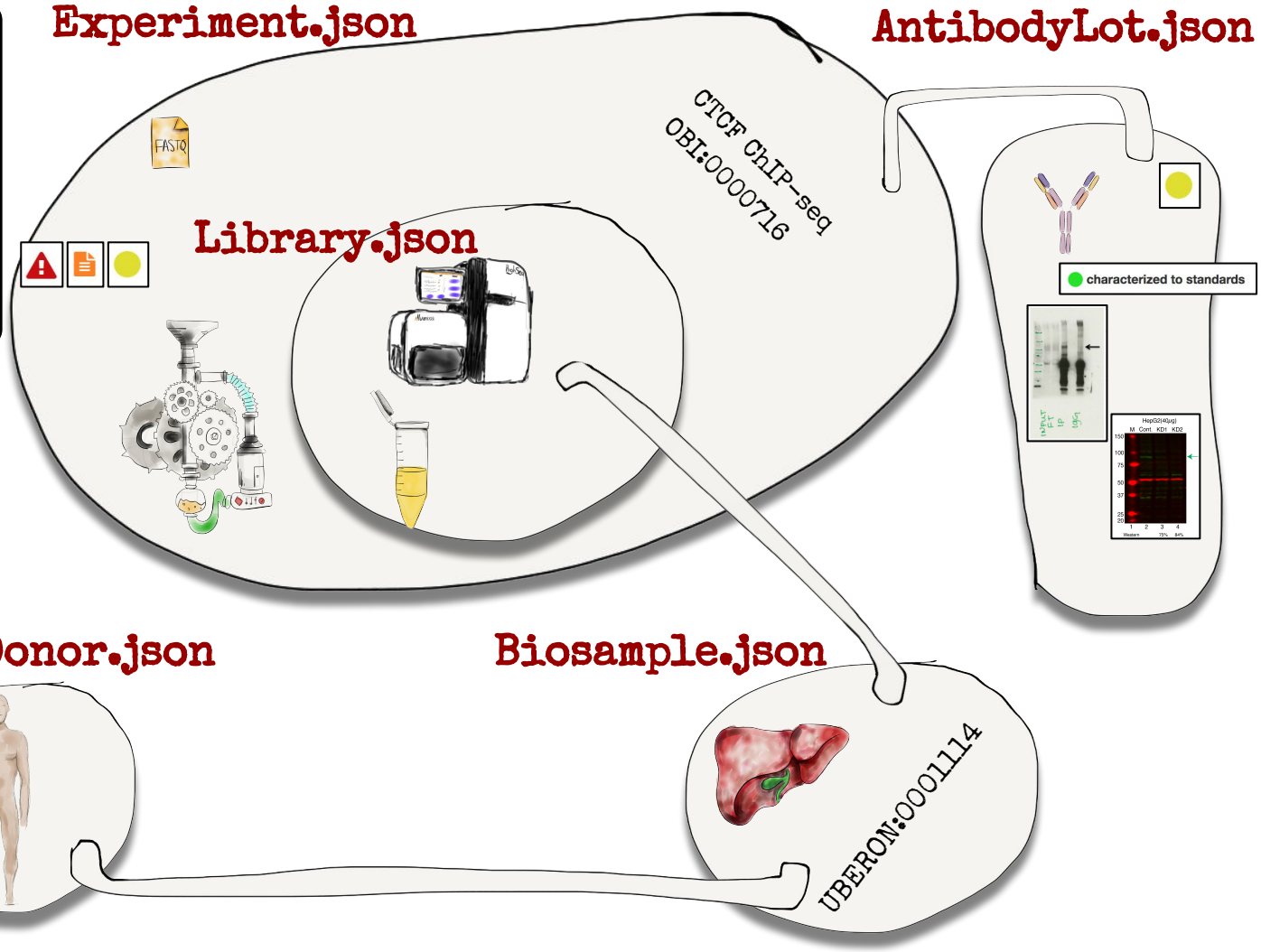
Data model
Ontology
Audits (badges)



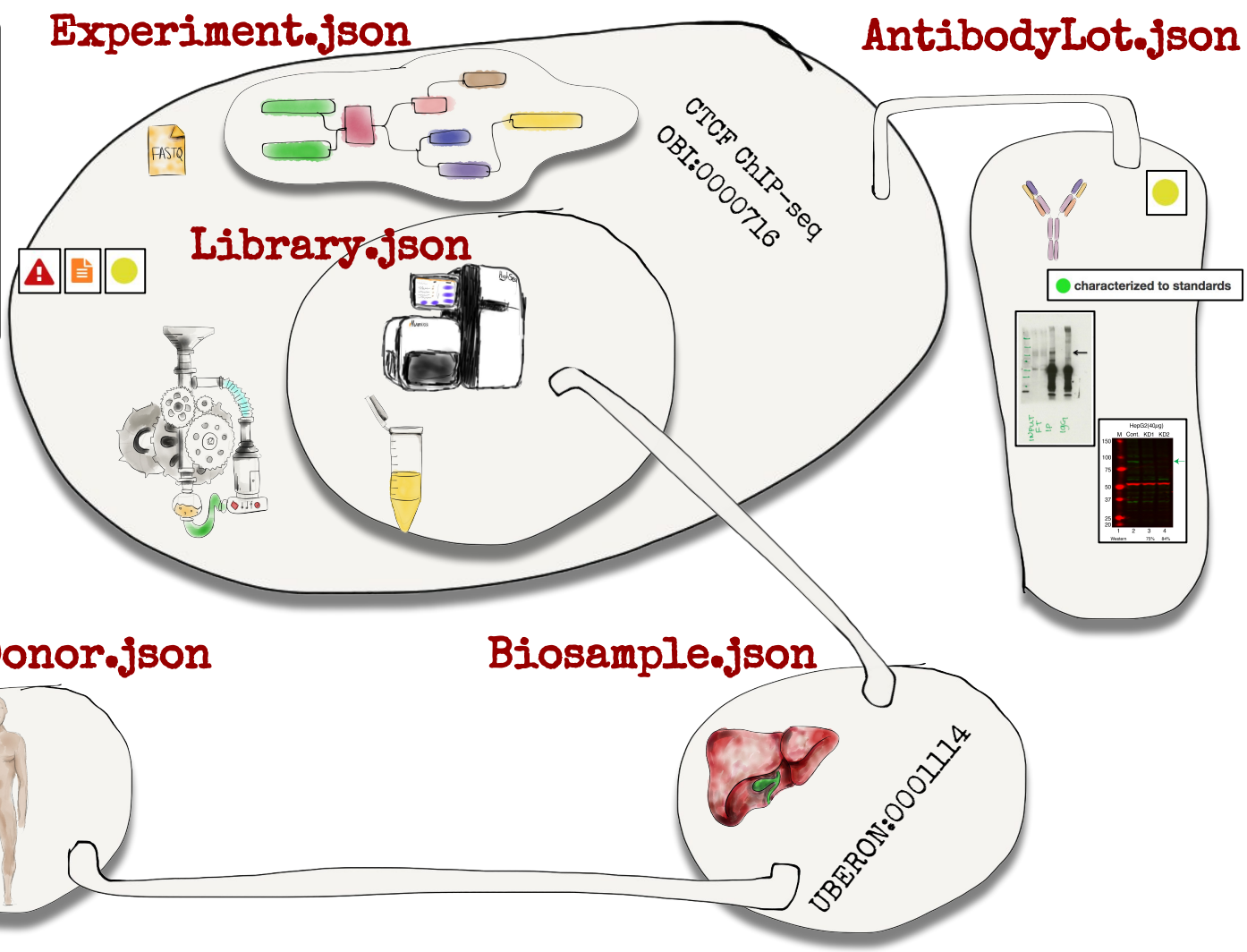
Data model
Ontology
Audits (badges)
AB standards



Data model
Ontology
Audits (badges)
AB standards
File validation



Data model
Ontology
Audits (badges)
AB standards
File validation
Uniform Pipelines

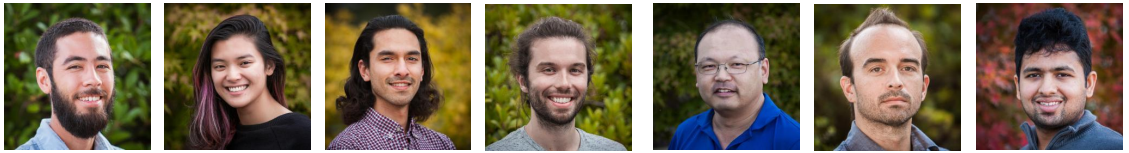


ENCODE Data Coordination Center

Mike Ben

Data wranglers **Pipeline devs.**

Jason Esther Idan Seth Otto Bek Tim



Zack Kath Nick Keenan Forrest Casey Karthik

Wrangler associates **Software devs.**

U24 HG009397 and U41 HG006992

NIH National Human Genome Research Institute

Challenges

- protocol modeling
- standardization of emerging technologies:
  - genetic modification
  - functional characterization
- community data

Challenges

Genetic modification object

- method (RNAi, CRISPR, etc.)
- purpose (activation, repression, etc.)
- category (deletion, insertion, etc.)

Genetic modification

Summary

Description:	ChIP-Seq on HepG2
Type:	insertion
Tags:	3xFLAG — C-terminal
Purpose:	tagging

Modification site

Target:	FLAG-NR5A1-human

Modification method

Technique:	CRISPR

Attribution

Lab:	Richard Myers, HAIB
Award PI:	Richard Myers, HAIB
Project:	ENCODE
Aliases:	richard-myers:SL272463-SL272464-gm

Genetic modification

To the best of our knowledge there is no large scale systematic assessment of commercial modifications, ENCODE should? devote large effort towards conducting characteriations.

future challenges

Explain the challenge to standardize the submitted chara